Ronald DeVore
Angela Kunoth
Editors

Multiscale, Nonlinear and Adaptive Approximation

Springer

Multiscale, Nonlinear and Adaptive Approximation

Ronald A. DeVore • Angela Kunoth
Editors

Multiscale, Nonlinear and Adaptive Approximation

Dedicated to Wolfgang Dahmen on the Occasion of his 60th Birthday

Springer

Editors
Ronald A. DeVore
Department of Mathematics
Texas A&M University
College Station, TX 77840
USA
rdevore@math.tamu.edu

Angela Kunoth
Institut für Mathematik
Universität Paderborn
Warburger Str. 100
33098 Paderborn
Germany
kunoth@math.uni-paderborn.de

ISBN 978-3-642-03412-1 e-ISBN 978-3-642-03413-8
DOI 10.1007/978-3-642-03413-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009935698

Mathematics Subject Classification (2000): 41-XX, 65-XX

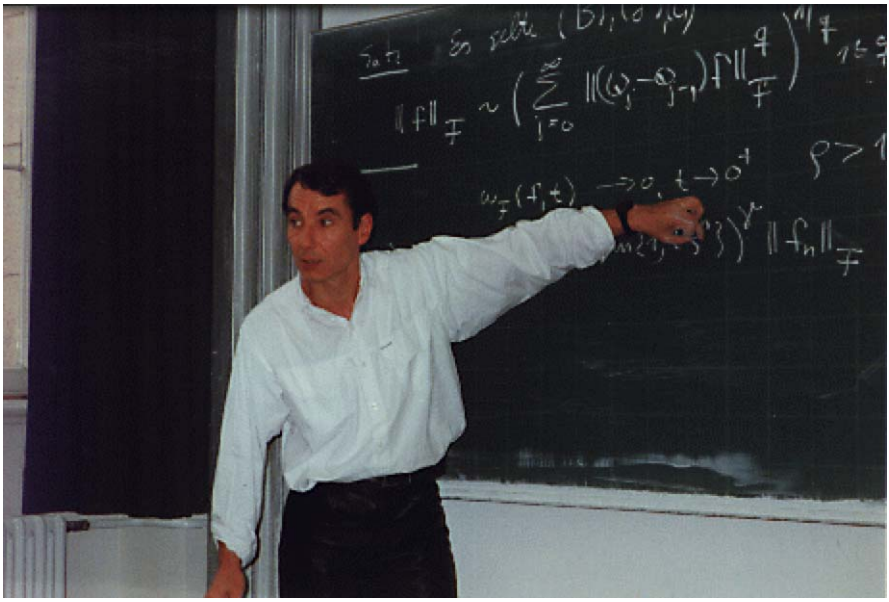
Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

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

Dedicated to Wolfgang Dahmen on the Occasion of his 60th Birthday

Preface



On the occasion of his 60th birthday in October 2009, friends, collaborators, and admirers of Wolfgang Dahmen have organized this volume which touches on various of his research interests. This volume will provide an easy to read excursion into many important topics in applied and computational mathematics. These include nonlinear and adaptive approximation, multivariate splines, subdivision schemes, multiscale and wavelet methods, numerical schemes for partial differential and boundary integral equations, learning theory, and high-dimensional integrals.

College Station, Texas, USA — *Ronald A. DeVore*
Paderborn, Germany — *Angela Kunoth*
June 2009

Acknowledgements

We are deeply grateful to Dr. Martin Peters and Thanh-Ha Le Thi from Springer for realizing this book project and to Frank Holzwarth for technical support.

Contents

Introduction: Wolfgang Dahmen's mathematical work 1
Ronald A. DeVore and Angela Kunoth
1 Introduction 1
2 The early years: Classical approximation theory 2
3 Bonn, Bielefeld, Berlin, and multivariate splines 2
3.1 Computer aided geometric design 3
3.2 Subdivision and wavelets 4
4 Wavelet and multiscale methods for operator equations 5
4.1 Multilevel preconditioning 5
4.2 Compression of operators 5
5 Adaptive solvers 6
6 Construction and implementation 7
7 Hyperbolic partial differential equations and conservation laws . . . 8
8 Engineering collaborations 9
9 The present 9
10 Final remarks 10
Publications by Wolfgang Dahmen (as of summer 2009) 10

The way things were in multivariate splines: A personal view 19
Carl de Boor
1 Tensor product spline interpolation 19
2 Quasiinterpolation 20
3 Multivariate B-splines 21
4 Kergin interpolation 23
5 The recurrence for multivariate B-splines 25
6 Polyhedral splines 27
7 Box splines 28
8 Smooth multivariate piecewise polynomials and the B-net 31
References 34

On the efficient computation of high-dimensional integrals and the approximation by exponential sums 39
Dietrich Braess and Wolfgang Hackbusch
1 Introduction 39
2 Approximation of completely monotone functions by exponential sums 41
3 Rational approximation of the square root function 43
3.1 Heron's algorithm and Gauss' arithmetic-geometric mean 43
3.2 Heron's method and best rational approximation 44
3.3 Extension of the estimate (19) 48
3.4 An explicit formula 49
4 Approximation of $1/x^\alpha$ by exponential sums 50
4.1 Approximation of $1/x$ on finite intervals 50
4.2 Approximation of $1/x$ on $[1,\infty)$ 51
4.3 Approximation of $1/x^\alpha$, $\alpha > 0$ 54
5 Applications of $1/x$ approximations 55
5.1 About the exponential sums 55
5.2 Application in quantum chemistry 55
5.3 Inverse matrix 56
6 Applications of $1/\sqrt{x}$ approximations 58
6.1 Basic facts 58
6.2 Application to convolution 58
6.3 Modification for wavelet applications 60
6.4 Expectation values of the H-atom 60
7 Computation of the best approximation 62
8 Rational approximation of $\sqrt{x}$ on small intervals 63
9 The arithmetic-geometric mean and elliptic integrals 65
10 A direct approach to the infinite interval 67
11 Sinc quadrature derived approximations 68
References 73

Adaptive and anisotropic piecewise polynomial approximation 75
Albert Cohen and Jean-Marie Mirebeau
1 Introduction 75
1.1 Piecewise polynomial approximation 75
1.2 From uniform to adaptive approximation 77
1.3 Outline 79
2 Piecewise constant one-dimensional approximation 80
2.1 Uniform partitions 81
2.2 Adaptive partitions 83
2.3 A greedy refinement algorithm 85
3 Adaptive and isotropic approximation 87
3.1 Local estimates 88
3.2 Global estimates 90
3.3 An isotropic greedy refinement algorithm 91

3.4 The case of smooth functions. 95
4 Anisotropic piecewise constant approximation on rectangles 100
4.1 A heuristic estimate 100
4.2 A rigourous estimate 103
5 Anisotropic piecewise polynomial approximation 108
5.1 The shape function 108
5.2 Algebraic expressions of the shape function 109
5.3 Error estimates 111
5.4 Anisotropic smoothness and cartoon functions 112
6 Anisotropic greedy refinement algorithms 116
6.1 The refinement algorithm for piecewise constants on rectangles 119
6.2 Convergence of the algorithm 121
6.3 Optimal convergence 124
6.4 Refinement algorithms for piecewise polynomials on triangles 128
References 134

Anisotropic function spaces with applications 137
Shai Dekel and Pencho Petrushev
1 Introduction 137
2 Anisotropic multiscale structures on $\mathbb{R}^n$ 139
2.1 Anisotropic multilevel ellipsoid covers (dilations) of $\mathbb{R}^n$. 139
2.2 Comparison of ellipsoid covers with nested triangulations in $\mathbb{R}^2$ 142
3 Building blocks 143
3.1 Construction of a multilevel system of bases 143
3.2 Compactly supported duals and local projectors 145
3.3 Two-level-split bases 146
3.4 Global duals and polynomial reproducing kernels 148
3.5 Construction of anisotropic wavelet frames 151
3.6 Discrete wavelet frames 154
3.7 Two-level-split frames 155
4 Anisotropic Besov spaces (B-spaces) 156
4.1 B-spaces induced by anisotropic covers of $\mathbb{R}^n$ 156
4.2 B-spaces induced by nested multilevel triangulations of $\mathbb{R}^2$ 158
4.3 Comparison of different B-spaces and Besov spaces 159
5 Nonlinear approximation 160
6 Measuring smoothness via anisotropic B-spaces 162
7 Application to preconditioning for elliptic boundary value problems 164
References 166

Nonlinear approximation and its applications 169
Ronald A. DeVore
1 The early years 169
2 Smoothness and interpolation spaces 171
2.1 The role of interpolation 172
3 The main types of nonlinear approximation 174
3.1 n-term approximation 174
3.2 Adaptive approximation 178
3.3 Tree approximation 178
3.4 Greedy algorithms 181
4 Image compression 185
5 Remarks on nonlinear approximation in PDE solvers 187
6 Learning theory 189
6.1 Learning with greedy algorithms 193
7 Compressed sensing 195
8 Final thoughts 199
References 199

Univariate subdivision and multi-scale transforms: The nonlinear case . . 203
Nira Dyn and Peter Oswald
1 Introduction 203
2 Nonlinear multi-scale transforms: Functional setting 210
2.1 Basic notation and further examples 210
2.2 Polynomial reproduction and derived subdivision schemes 215
2.3 Convergence and smoothness 217
2.4 Stability 223
2.5 Approximation order and decay of details 227
3 The geometric setting: Case studies 230
3.1 Geometry-based subdivision schemes 231
3.2 Geometric multi-scale transforms for planar curves 240
References 245

Rapid solution of boundary integral equations by wavelet Galerkin schemes 249
Helmut Harbrecht and Reinhold Schneider
1 Introduction 249
2 Problem formulation and preliminaries 252
2.1 Boundary integral equations 252
2.2 Parametric surface representation 253
2.3 Kernel properties 255
3 Wavelet bases on manifolds 256
3.1 Wavelets and multiresolution analyses 256
3.2 Refinement relations and stable completions 258
3.3 Biorthogonal spline multiresolution on the interval 259
3.4 Wavelets on the unit square 261

3.5 Patchwise smooth wavelet bases 266
3.6 Globally continuous wavelet bases 267
4 The wavelet Galerkin scheme 271
4.1 Historical notes 272
4.2 Discretization 273
4.3 A-priori compression 274
4.4 Setting up the compression pattern 275
4.5 Computation of matrix coefficients 277
4.6 A-posteriori compression 279
4.7 Wavelet preconditioning 279
4.8 Numerical results 281
4.9 Adaptivity 284
References 290

Learning out of leaders 295
Gérard Kerkyacharian, Mathilde Mougeot, Dominique Picard and Karine Tribouley
1 Introduction 295
2 Various learning algorithms in Wolfgang Dahmen's work 298
2.1 Greedy learning algorithms 298
2.2 Tree thresholding procedures 299
3 Learning out leaders: LOL 303
3.1 Gaussian regression model 303
3.2 LOL procedure 304
3.3 Sparsity conditions on the target function f 305
3.4 Results 306
3.5 Discussion 307
3.6 Restricted LOL 308
4 Practical performances of the LOL procedure 309
4.1 Experimental design 309
4.2 Algorithm 310
4.3 Simulation results 312
4.4 Quality reconstruction 313
4.5 Discussion 314
5 Proofs 316
5.1 Preliminaries 316
5.2 Concentration lemma 5.4 318
5.3 Proof of Theorem 3.2 319
References 323

Optimized wavelet preconditioning 325
Angela Kunoth
1 Introduction 325
2 Systems of elliptic partial differential equations (PDEs) 329
2.1 Abstract operator systems 329
2.2 A scalar elliptic boundary value problem 330

2.3 Saddle point problems involving essential boundary conditions . . . 331
2.4 PDE-constrained control problems: Distributed control . . 334
2.5 PDE-constrained control problems: Dirichlet boundary control . . . 336
3 Wavelets . . . 337
3.1 Basic properties . . . 337
3.2 Norm equivalences and Riesz maps . . . 340
3.3 Representation of operators . . . 341
3.4 Multiscale decomposition of function spaces . . . 342
4 Problems in wavelet coordinates . . . 356
4.1 Elliptic boundary value problems . . . 356
4.2 Saddle point problems . . . 358
4.3 Control problems: Distributed control . . . 361
4.4 Control problems: Dirichlet boundary control . . . 365
5 Iterative solution . . . 367
5.1 Finite systems on uniform grids . . . 368
5.2 Numerical examples . . . 372
References . . . 376

Multiresolution schemes for conservation laws . . . 379
Siegfried Müller
1 Introduction . . . 379
2 Governing equations and finite volume schemes . . . 381
3 Multiscale analysis . . . 383
4 Multiscale-based spatial grid adaptation . . . 387
5 Adaptive multiresolution finite volume schemes . . . 389
5.1 From the reference scheme to an adaptive scheme . . . 389
5.2 Approximate flux and source approximation strategies . . 390
5.3 Prediction strategies . . . 392
5.4 Multilevel time stepping . . . 393
5.5 Error analysis . . . 397
6 Numerical results . . . 397
6.1 The solver Quadflow . . . 398
6.2 Application . . . 398
7 Conclusion and trends . . . 402
References . . . 405

Theory of adaptive finite element methods: An introduction . . . 409
Ricardo H. Nochetto, Kunibert G. Siebert and Andreas Veeser
1 Introduction . . . 410
1.1 Classical vs adaptive approximation in 1d . . . 410
1.2 Outline . . . 411
2 Linear boundary value problems . . . 413
2.1 Sobolev spaces . . . 413
2.2 Variational formulation . . . 416

2.3 The inf-sup theory . . . 419
2.4 Two special problem classes . . . 423
2.5 Applications . . . 427
2.6 Problems . . . 430
3 The Petrov-Galerkin method and finite element bases . . . 432
3.1 Petrov-Galerkin solutions . . . 433
3.2 Finite element spaces . . . 438
3.3 Problems . . . 445
4 Mesh refinement by bisection . . . 447
4.1 Subdivision of a single simplex . . . 447
4.2 Mesh refinement by bisection . . . 451
4.3 Basic properties of triangulations . . . 454
4.4 Refinement algorithms . . . 457
4.5 Complexity of refinement by bisection . . . 462
4.6 Problems . . . 468
5 Piecewise polynomial approximation . . . 469
5.1 Quasi-interpolation . . . 469
5.2 A priori error analysis . . . 472
5.3 Principle of error equidistribution . . . 474
5.4 Adaptive approximation . . . 476
5.5 Problems . . . 479
6 A posteriori error analysis . . . 480
6.1 Error and residual . . . 481
6.2 Global upper bound . . . 482
6.3 Lower bounds . . . 487
6.4 Problems . . . 495
7 Adaptivity: Convergence . . . 497
7.1 The adaptive algorithm . . . 497
7.2 Density and convergence . . . 499
7.3 Properties of the problem and the modules . . . 501
7.4 Convergence . . . 503
7.5 Problems . . . 511
8 Adaptivity: Contraction property . . . 512
8.1 The modules of AFEM for the model problem . . . 513
8.2 Properties of the modules of AFEM . . . 514
8.3 Contraction property of AFEM . . . 518
8.4 Example: Discontinuous coefficients . . . 520
8.5 Problems . . . 522
9 Adaptivity: Convergence rates . . . 524
9.1 Approximation class . . . 525
9.2 Cardinality of $\mathscr{M}_k$ . . . 530
9.3 Quasi-optimal convergence rates . . . 533
9.4 Marking vs optimality . . . 534
9.5 Problems . . . 538
References . . . 539

Adaptive wavelet methods for solving operator equations: An overview 543
Rob Stevenson
1 Introduction 543
1.1 Non-adaptive methods 543
1.2 Adaptive methods 545
1.3 Best N-term approximation and approximation classes . . 546
1.4 Structure of the paper 547
1.5 Some properties of the (quasi-) norms $\|\cdot\|_{\mathscr{A}^s}$ 547
2 Well-posed linear operator equations 549
2.1 Reformulation as a bi-infinite matrix vector equation 549
2.2 Some model examples 550
3 Adaptive wavelet schemes I: Inexact Richardson iteration 552
3.1 Richardson iteration 552
3.2 Practical scheme 553
3.3 The routines **COARSE** and **APPLY** 558
3.4 Non-coercive **B** 562
3.5 Alternatives for the Richardson iteration 564
4 Adaptive wavelet schemes II: The Adaptive wavelet-Galerkin method 565
4.1 The adaptive wavelet-Galerkin method (AWGM) in a idealized setting 565
4.2 Practical scheme 568
4.3 Discussion 574
5 The approximation of operators in wavelet coordinates by computable sparse matrices 574
5.1 Near-sparsity of partial differential operators in wavelet coordinates 575
5.2 The approximate computation of the significant entries . . 580
5.3 Trees 583
6 Adaptive frame methods 585
6.1 Introduction 585
6.2 Frames 585
6.3 The adaptive solution of an operator equation in frame coordinates 586
6.4 An adaptive Schwarz method for aggregated wavelet frames 588
7 Adaptive methods based on tensor product wavelet bases 590
7.1 Tensor product wavelets 590
7.2 Non-adaptive approximation 590
7.3 Best N-term approximation and regularity 591
7.4 s^*-computability 592
7.5 Truly sparse stiffness matrices 592
7.6 Problems in space high dimension 592
7.7 Non-product domains 593

7.8 Other, non-elliptic problems . . . 594
References . . . 594

Optimal multilevel methods for H(grad), H(curl), and H(div) systems on graded and unstructured grids . . . 599
Jinchao Xu, Long Chen, and Ricardo H. Nochetto
1 Introduction . . . 599
2 The method of subspace corrections . . . 601
2.1 Iterative methods . . . 602
2.2 Space decomposition and method of subspace correction 604
2.3 Sharp convergence identities . . . 607
3 Multilevel methods on quasi-uniform grids . . . 609
3.1 Finite element methods . . . 609
3.2 Multilevel space decomposition and multigrid method . . . 611
3.3 Stable decomposition and optimality of BPX preconditioner . . . 612
3.4 Uniform convergence of V-cycle multigrid . . . 616
3.5 Systems with strongly discontinuous coefficients . . . 618
4 Multilevel methods on graded grids . . . 621
4.1 Bisection methods . . . 622
4.2 Compatible bisections . . . 624
4.3 Decomposition of bisection grids . . . 625
4.4 Generation of compatible bisections . . . 627
4.5 Node-oriented coarsening algorithm . . . 629
4.6 Space decomposition on bisection grids . . . 630
4.7 Strengthened Cauchy-Schwarz inequality . . . 634
4.8 BPX preconditioner and multigrid on graded bisection grids . . . 636
5 Multilevel methods for H(curl) and H(div) systems . . . 636
5.1 Preliminaries . . . 638
5.2 Space decomposition and multigrid methods . . . 646
5.3 Stable decomposition . . . 648
6 The auxiliary space method and HX preconditioner for unstructured grids . . . 651
6.1 The auxiliary space method . . . 651
6.2 HX preconditioner . . . 653
References . . . 655

List of Contributors

Carl de Boor
Department of Computer Sciences, University of Wisconsin-Madison,
e-mail: deboor@cs.wisc.edu, URL: http://pages.cs.wisc.edu/∼deboor/

Dietrich Braess
Mathematisches Institut, Ruhr-Universität Bochum, 44780 Bochum, Germany,
e-mail: Dietrich.Braess@rub.de,
URL: http://homepage.ruhr-uni-bochum.de/Dietrich.Braess

Long Chen
Department of Mathematics, University of California at Irvine, Irvine, CA 92697,
USA, e-mail: chenlong@math.uci.edu

Albert Cohen
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du
Chevaleret, 75013 Paris, France, e-mail: cohen@ann.jussieu.fr

Shai Dekel
GE Healthcare, 27 Hamaskit St., Herzelia 46733, Israel,
e-mail: Shai.Dekel@ge.com

Ronald A. DeVore
Department of Mathematics, Texas A&M University, College Station, Texas 77840,
USA, e-mails: ronald.a.devore@gmail.com, rdevore@math.tamu.edu,
URL: http://www.math.tamu.edu/∼rdevore

Nira Dyn
Tel Aviv University, School of Mathematical Sciences,
e-mail: niradyn@math.tau.ac.il

Wolfgang Hackbusch
Max-Planck-Institut *Mathematik in den Naturwissenschaften*, Inselstr. 22, 04103
Leipzig, Germany, e-mail: wh@mis.mpg.de,
URL: http://www.mis.mpg.de/scicomp/hackbusch_e.html

Helmut Harbrecht
Institute for Numerical Simulation, Bonn University, Wegelerstr. 6, 53115 Bonn, Germany, e-mail: harbrecht@ins.uni-bonn.de

Gérard Kerkyacharian
Université Paris-Diderot, CNRS LPMA, 175 rue du Chevaleret, 75013 Paris, France

Angela Kunoth
Institut für Mathematik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany, e-mail: kunoth@math.uni-paderborn.de,
URL: http://www2.math.uni-paderborn.de/ags/kunoth/group/angelakunoth.html

Jean-Marie Mirebeau
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France, e-mail: mirebeau@ann.jussieu.fr

Mathilde Mougeot
MODALX, Université Paris Ouest Nanterre, 200 avenue de la République, 92001 Nanterre Cedex, France

Siegfried Müller
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, D-52056 Aachen, e-mail: mueller@igpm.rwth-aachen.de

Ricardo H. Nochetto
Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742, e-mail: rhn@math.umd.edu

Peter Oswald
Jacobs University Bremen, School of Engineering and Science, D-28759 Bremen, Germany, e-mail: p.oswald@jacobs-university.de

Pencho Petrushev
Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA, e-mail: pencho@math.sc.edu

Dominique Picard
Université Paris-Diderot, CNRS LPMA, 175 rue du Chevaleret, 75013 Paris, France, e-mail: picard@math.jussieu.fr

Reinhold Schneider
Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, e-mail: schneidr@math.tu-berlin.de

Kunibert G. Siebert
Fachbereich Mathematik, Universität Duisburg-Essen, Forsthausweg 2, D-47057 Duisburg, Germany, e-mail: kg.siebert@uni-due.de

Rob Stevenson
Korteweg - de Vries (KdV) Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands,
e-mail: R.P.Stevenson@uva.nl

Karine Tribouley
MODALX, Université Paris Ouest Nanterre, 200 avenue de la République, 92001 Nanterre Cedex, France

Andreas Veeser
Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, I-20133 Milano, Italy, e-mail: andreas.veeser@unimi.it

Jinchao Xu
Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA, and LMAM, The School of Mathematical Sciences, Peking University, China, e-mail: xu@math.psu.edu.

Introduction: Wolfgang Dahmen's mathematical work

Ronald A. DeVore and Angela Kunoth

Abstract This volume is testimony to the rich and amazingly diverse mathematical life of Wolfgang Dahmen. The cornerstones of Wolfgang's research are deep theoretical analysis and extensive interdisciplinary projects in high–performance scientific computing. This article touches on some of the highlights of his work and its impact in application domains.

1 Introduction

Wolfgang Dahmen is at the peak of his mathematical career — witness his recent forays into learning theory, compressed sensing, and high dimensional problems. His accomplishments to date have unquestionable diversity and depth. Perhaps the two characteristics that best identify his mathematics are the constant exploration of new and emerging fields as well as the quest for relevancy of his work to application domains. The subsequent contributions to this volume will certainly validate this view. We therefore take the opportunity in this introductory article to give some overall feeling on how this all came about.

Ronald A. DeVore
Department of Mathematics, Texas A&M University, College Station, Texas 77840, USA,
e-mail: rdevore@math.tamu.edu,
URL: http://www.math.tamu.edu/∼rdevore

Angela Kunoth
Institut für Mathematik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany,
e-mail: kunoth@math.uni-paderborn.de,
URL: http://www2.math.uni-paderborn.de/ags/kunoth/group/angelakunoth.html

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_1,

2 The early years: Classical approximation theory

Wolfgang grew up in a small village called Linnich some forty kilometers from Aachen tucked into the northwest corner of Germany near the Dutch border. So it was natural for him to pursue his university studies at the RWTH in Aachen. He was quickly identified as a mathematical talent and offered a Research Assistantship in the Paul Leo Butzer program in Harmonic Analysis and Approximation — a group notable for its breadth and talent. He began investigating problems in classical approximation theory and set out to settle the conjecture of Golumb and Korovkin on whether convolution operators could simultaneously realize best approximation orders for the entire spectrum of smoothness classes. He disproved this conjecture in [1, 2] and these results became the core of his dissertation thesis. He went on to give general connections between the behaviour of norms of operators, asymptotically best approximation for a particular range, and spectral convergence [3, 4, 5, 6, 7]. He followed this by using these techniques to prove in [8] a slightly corrected conjecture of Stechkin. These results would already establish him as a prominent young analyst.

3 Bonn, Bielefeld, Berlin, and multivariate splines

Wolfgang took an assistant position with Karl Scherer at Universität Bonn in 1976. Bonn was at this time a hub of activity in approximation theory among the faculty and assistants as well as the steady stream of visitors who frequented the Institut für Angewandte Mathematik. It was because of some of this new exposure that Wolfgang Dahmen switched to the newly developing field of multivariate splines promoted by Carl de Boor. At that time there was not much known about such smooth piecewise polynomials in more than one spatial dimension, apart from classical finite element approaches. Although many engineering sciences built their approximations on splines, theoretical results for genuine multivariate analogues were not known.

The starting point for Wolfgang's work was a suggestion by Carl de Boor to define multivariate splines through a generalization of a geometrical interpretation of univariate B-Splines, going back to Iso Schoenberg. This eventually led to the development of a cohesive and deep theory that is well documented in the article of Carl de Boor in this volume.

There were several obvious gains that emanated from Wolfgang's entry into multivariate splines. One was the exposure to the interconnectivity of various branches of mathematics. In this case, multivariate splines intersect with commutative algebra, combinatorics, number theory, and geometry, in addition to the obvious connections to finite element methods. Another big plus for Wolfgang was his collaboration with Charles A. Micchelli. Wolfgang received an IBM PostDoctoral Fellowship for 1979/1980 which was the starting point for the very fruitful and long-term Dahmen–Micchelli collaboration.

Wolfgang returned to Germany and took his first professor position in Bielefeld in 1981. His work on multivariate splines was now at its peak. Multivariate B-splines are generated from volumes of intersections of hyperplanes with simplices. While they have many beautiful properties, they are not so easy to work with numerically. So they were soon replaced by box splines which are defined by replacing the simplex by a cube. Although box splines result in uniform grid structures, the theory for box splines needed very different techniques [28, 29, 31, 32, 38].

This box spline theory involved a number of interesting algebraic and combinatoric problems [33, 34, 37, 38]. Indeed, box-spline techniques had very surprising applications to questions of combinatorial number theory like reciprocal relations, Bell's theorem and linear systems of diophantic equations [41, 43, 65, 66, 68], culminating with far-reaching results in commutative algebra by Dahmen–Micchelli and by Rong-Qing Jia. Finally, in the long paper [67] together with Andreas Dress and Charles Micchelli, Wolfgang investigated the application of concepts from homological algebra for the treatment of the central problem of the determination of the exact dimension of the intersection of the null space of a family of endomorphisms which are characterized by a certain combinatorial structure.

3.1 Computer aided geometric design

The Dahmen–Micchelli collaboration also had lasting impact in Computer Aided Geometric Design (CAGD) [42]. This work began while in Bielefeld but continued well during his stay at the Freie Universität in Berlin (1987-1992). A number of their papers in this area treat data fitting and interpolation problems [49, 50, 51].

Univariate splines have important properties like being variation-diminishing which is strongly connected to the concept of total positivity. In its full strength, the latter is inherently one-dimensional. However, there exists a strong coupling to Polya frequency functions which can in turn be interpreted in the multivariate case as well. This viewpoint resulted in a number of partly function-theoretic-based investigations [10, 20] as well as results concerning biinfinite matrices [18, 44, 76]. Roughly speaking, a central result in this context is that totally positive matrices (which appear, for instance, for splines, Tchebycheff systems, or in the theory of small oscillations) can be completely factorized into totally positive 2-band matrices. This in turn has a number of important consequences, like corner cutting in CAGD, knot insertion by repeated convex combinations of knots, and variation-diminishing properties of B-splines.

A relevant property for data fitting is shape preservation, where the goal is to preserve properties like monotonicity or convexity [53, 56, 59, 60]. Moreover, a completely new approach for modelling surfaces with smooth interfaces based on piecewise algebraic surfaces was proposed in [54, 57, 59].

The motivation for several of Wolfgang's investigations was the premise that parametric representations of surfaces are very unsuitable for graphics of high quality. While in Berlin, Wolfgang had a close collaboration with the nearby company

Mental Images founded in 1986. During this collaboration, the intent was to develop a modelling library which was able to reduce all CAD formats used at that time to a unified format and to which after polygonalization a highly efficient ray tracer could be applied. Since algebraic representations would also support the ray tracer, one could circumvent polygonalization and the handling of different data formats. The prototype of such a library has been developed and implemented together with Mental Images and has in the aftermath triggered a number of related projects. Nowadays, their website states that "Mental Images is the recognized international leader in providing rendering and 3D modeling technology to the entertainment, computer-aided design, scientific visualization, architecture, and other industries that require sophisticated images".

3.2 Subdivision and wavelets

An integral part of CAGD is played by subdivision algorithms for quickly generating, displaying and controlling geometrical surfaces. Here Wolfgang had major contributions [37, 39, 45, 46, 48] culminating in the monograph [62], written together with Alfred Cavaretta and Charles Micchelli and published in the Memoirs of the American Mathematical Society.

CAGD and subdivision in particular were a forerunner of the development of wavelets which dominated the applied harmonic analysis and image processing communities in the 80's and 90's. So it was natural that the emphasis of Wolfgang Dahmen's work at the beginning of the 90's became the theory of multiscale analysis with a particular eye to the application of these new sophisticated tools tor the numerics of partial differential and integral equations. From 1992 on, this was fueled by the fact that his new position at RWTH Aachen brought ample research opportunities with engineers.

Another important factor in this new direction was his exposure to nonlinear approximation during the Bonn years. Nonlinear spline approximation was a favorite topic of Karl Scherer. Two of the main proponents of nonlinear methods were Dietrich Braess and Ronald A. DeVore who were frequent visitors to Bonn during the 70's. In fact, Dahmen and DeVore had several collaborations through the years but this became particularly intense with the wavelet evolution when they teamed up with the young star Albert Cohen. The impact of their work would be felt in many applied disciplines including image compression, compressed sensing, learning theory, and numerical partial differential equations (PDEs).

4 Wavelet and multiscale methods for operator equations

4.1 Multilevel preconditioning

Wolfgang's work in PDEs took on great momentum at the end of the 80's. His first major contributions centered around preconditioning elliptic operators. At that time, multigrid methods were known to provide fast numerical solvers for the system of linear equations arising from finite element discretization. A much-discussed question from the point of finite elements at the end of the 80's was under which conditions multilevel preconditioners such as the hierarchical and the BPX preconditioner (proposed by James H. Bramble, Joseph E. Pasciak and Jinchao Xu) provided uniformly bounded spectral condition numbers for the system matrix. This is an essential point to guarantee fast iterative solvers. Corresponding conditions were derived in [75] and, specifically, the uniform boundedness of the BPX preconditioner was established; a result independently obtained by Peter Oswald. From the angle of wavelets, the corresponding ingredients for optimal preconditioning became much more transparent [91, 92]. In particular, in [75] the uniform boundedness was established for the first time also for adaptively refined grids in a classical finite element framework. Two articles on preconditioning systems of PDEs by multilevel ingredients are collected in this volume: Angela Kunoth's on optimized preconditioning with wavelets, and the survey by Long Chen, Ricardo Nochetto and Jinchao Xu on BPX and multigrid preconditioning.

Wavelet-based preconditioners directly give norm equivalences for functions in Sobolev spaces and are therefore not restricted to operators of positive order. This led Wolfgang and his collaborators to a number of investigations for pseudodifferential and boundary integral operators for whose compression of operators is an even more relevant issue, see Section 4.2.

Within the last few years, discontinuous Galerkin methods for PDEs have become increasingly popular, due to the fact that one can relatively easy increase the polynomial degree and therefore ensure higher order convergence where solutions are smooth. Very recently, multilevel preconditioners for such interior discontinuous Galerkin methods have been presented for the first time in [172, 173].

These studies led to two well received extensive surveys on wavelet and multiscale methods for the numerical solution of partial differential equations in an Acta Numerica article [108] and, a few years later, [126].

4.2 Compression of operators

The goal of the multiscale approach to operator equations is to extract efficient representations for both the operator and the solution. For boundary integral formulations, conventional discretizations yield fully populated matrices and were therefore not feasible for 3D problems in reasonable time. Multipole expansions and panel

clustering methods became very popular as a way of thinning out the system matrix and making it more amenable to computation.

Multiscale wavelet methods arrived on the scene a little later spurred on by the observations of Gregory Beylkin, Ronald Coifman and Vladimir Rokhlin that certain (one dimensional) operators are almost sparse in a wavelet representation. These authors showed that for a fixed tolerance ε a matrix vector multiplication can be realized in $\mathcal{O}(N \log N)$ arithmetic operations with accuracy ε. Wolfgang and his collaborators, mainly Siegfried Prößdorf und Reinhold Schneider, were set on developing a rigorous theory that quantified the gain of multiscale methods [79, 80, 84, 86, 89, 90, 97, 101]. The results in [146] were the first to show for a large class of elliptic operators, including those of negative order, that the system matrices can be compressed up to optimal complexity $\mathcal{O}(N)$ while at the same time admitting optimal preconditioning of the system matrix and the solution of the resulting problems with asymptotically optimal convergence order. It also made the relevant point to bring preconditioning into play. (This topic was not addressed by Beylkin et al. since they had chosen an operator of order zero for which no conditioning issue arises.) Surveys about the different stages of the investigations for pseudodifferential and boundary integral operators were provided in [97, 103, 108].

So far, these results referred to discretizations on uniform grids. Another milestone for complexity reduction was achieved by the introduction of adaptively refined a–posteriori discretizations. Optimal complexity estimates for adaptive methods based on wavelets for integral equations were proved in [162].

A survey of the main results and the current state of the art of wavelet methods for integral equations is provided by the article by Helmut Harbrecht and Reinhold Schneider in this volume.

5 Adaptive solvers

Multiscale decompositions have long played an important role in image processing which leads to efficient compression of images. It is natural to try to bring these ideas into the realm of numerical PDEs. But of course, new problems arise since the solution is not known to us but can only be seen through computation.

Wolfgang had already collaborated with Albert Cohen and Ron DeVore on various aspects of image compression and multiscale systems but in the latter part of the 1990's, they begin to turn their attention to PDE solvers with the goal of providing a rigorous foundation for both theory and algorithms for adaptive methods. Of course, adaptive finite element methods had been around for some time. However, theoretical results which established their advantages over non adaptive methods were still lacking.

It seemed natural to first tackle adaptivity in the context of wavelet methods since these were already known to have the advantages mentioned above of yielding highly compressible matrix representations. The result was a penetrating theory and algorithms much beyond the state of the art known for partial differential and in-

tegral operators. It all began with the work [118] for linear elliptic partial differential equations which not only developed convergent algorithms but also algorithms with asymptotically optimal rates of convergence when compared to the wavelet-best N term approximation. This seminal work was followed by extensions to more general settings including nonlinear problems [133, 147]. Other collaborations of Wolfgang extended and substantially refined the notion of adaptive methods of optimal complexity to more general problems including saddle point problems [139, 144], adaptivity steered by goals different than the energy norm [165] and problems in optimal control constrained by an elliptic partial differential equation [150].

Interestingly, these results also applied to operators of negative order. Specifically for ensuring optimal complexity, a number of techniques known from optimal coding and compression [121] had to be interwined in a sophisticated way, yielding a novel approximate matrix-vector multiplication of optimal complexity. The derivation of adaptive methods with optimal complexity estimates for nonlinear stationary partial differential equations posed yet another difficulty which was attacked in [147] based on tree approximations. Recently, these techniques have been extended for the first time to deriving convergent adaptive schemes for elliptic eigenvalue problems in [174].

A survey of the basic principles and the main results is provided in the Encyclopedia Article [151] or the longer survey article [153]. Rob Stevenson gives a very nice overview of the state of the art of adaptive wavelet methods for operator equations in this volume.

The principles for proving convergence of adaptive methods for linear elliptic PDEs based on finite elements were known since Willy Dörfler's article in '96. However, it was a much discussed question in practical finite element codes whether a heuristically used derefinement/coarsening step was really needed for 'good complexity'. Establishing a sound theory was a difficult issue since a finite element discretization does not characterize in a natural way the underlying Sobolev space or the solution like in wavelet theory. Again tree approximations turned out to be the key for success, resulting in [145] for the first time in optimal complexity estimates for adaptive finite element methods. This seminal paper triggered numerous followers and extended into a rich theory, for which an account is given in the article of Ricardo Nochetto, Kunibert G. Siebert and Andreas Veeser in this volume.

6 Construction and implementation

The numerical implementation of wavelet algorithms is not without challenges. The main obstacle is to design multiscale systems for domains and manifolds which arise in applications. The classical development of wavelets is tied to $\mathbb{R}^n$ or, via periodization, to the torus. Wolfgang was at the heart of multiscale constructions for practical settings which took place largely in the 1990's. The initial studies [76, 78, 81] were devoted to systematically developing Fourier-free concepts and clarifying the basic relations of stability, Riesz bases, and norm equivalences. These investigations led

to Fourier-free stability criteria [81, 92, 95] which include biorthogonality as well as regularity and approximation properties. This results in the preservation of the relevant functional analytic properties of function spaces on domains and manifolds.

Flexible constructions of biorthogonal wavelets for a whole range of Sobolev spaces on domains and manifolds were provided in [88, 93, 100, 104, 105, 107, 112, 113, 115]. These tools also led to new evaluation algorithms [117, 158]. For instance, the exact computation of integrals of products of derivatives of scaling functions can be reduced to solving linear systems whose size is independent of the discretization [71]. Based on biorthogonal wavelets one can systematically construct for any spatial discretization stable pairs of approximation spaces for saddle point problems like for the Stokes problem [94] or for treating essential boundary conditions by Lagrange multipliers [122]. In addition, they admit optimal preconditioners. Starting from wavelets on the interval [104], tensorizing these and finally using parametric mappings, one can construct wavelets in particular for boundary integral equations on manifolds in 3D [100, 107, 115]. The construction in [107] is intrinsically tied to characterizations of function spaces on manifolds through manifold decompositions developed earlier by Zbigniew Ciesielski and Tadeusz Figiel using orthogonalized B-splines. This approach also leads naturally to domain decompositions for boundary integral equations. Finally, in [112] local wavelet bases with the desired stability and compression properties were constructed for standard finite element decompositions in up to 3D.

A discussion of implementation and numerical experiments for adaptive wavelet schemes for linear elliptic partial differential equations, exhibiting the theoretically predicted convergence rates in an exemplary manner, was provided in [125].

The foundation for many efficient algorithms are norm equivalences for multiscale expansions which can be used for preconditioning. So far these concepts have been derived for discretizations, initially on uniform grids. On the other hand, in view of problems in dimensions higher than three, in the past years also partition-of-unity methods which can be applied to essentially nonuniform grids have become fashionable. However, for these methods there was up to [169] no efficient and analytically proven optimal preconditioner available. Decompositions of spaces on nonuniform or anisotropic grids are the subject of an ongoing collaboration of Wolfgang Dahmen with Shai Dekel and Pencho Pestrushev, see their contribution in this volume.

7 Hyperbolic partial differential equations and conservation laws

Adaptive methods are also frequently applied to non–elliptic problems, primarily non-stationary hyperbolic problems. Wolfgang and his collaborators became interested in this area around the time that Ami Harten developed his compression approach for conservation laws. Harten's approach was improved in an essential way in [119] both with respect to practicability (2D problems on curvilinear grids) as

well as conceptually with respect to the amount of expensive flux evaluations in an adaptive grid refinement, see the contribution by Siegfried Müller in this volume. For the compressible Navier-Stokes equations, these ideas were elaborated in [141].

Another non-elliptic example centers around adaptive multigrid methods for convection dominated problems in [123]. Here adaptivity not only reduces the complexity but also stabilizes a standard Galerkin discretization so that no modification by artificial introduction of viscosity is necessary.

8 Engineering collaborations

Wolfgang was not immune to collaborating with Engineers and he had an excellent environment for such collaborations in Aachen. For example, he worked with mechanical engineers at RWTH Aachen, specifically Josef Ballmann, to efficiently simulate transport phenomena, grid generation, hypersonic flow problems, and the interaction of aerodynamics and structure. This also resulted in theoretical results on Riemann solvers for non-convex flux functions including phase transitions for hyperbolic conservation laws [154] or the well-posedness of modeling problems for nonlinear elasticity [161].

Wolfgang's long-term collaboration with the chemical engineer Wolfgang Marquardt at RWTH Aachen is one of the prominent examples of bringing novel mathematical concepts into practical applications to substantially improve simulation results. Here real-time optimization of dynamical chemical processes requires compression of systems of ordinary differential equations [102, 111, 127, 128, 129, 135]–[138].

9 The present

Wolfgang's research program continues to find interesting new avenues. In the past few years, Wolfgang Dahmen's work has primarily been driven by problems in learning theory and compressed sensing. These investigations have taken place largely in collaboration with Albert Cohen and Ron DeVore. Their goal, quite similar to the adaptive PDE program, is to understand in what sense various algorithms in these areas are optimal. This program has led to a series of results [160, 166, 167, 168, 171, 175] which clarify the gains of sparsity and nonlinearity. These remarkable accomplishments are well documented in the article of Ron DeVore in this volume.

10 Final remarks

Wolfgang Dahmen has had an illustrious mathematical career full of high points. Perhaps his most significant recognition was the awarding of the Gottfried Wilhelm Leibniz-Award in 2002, the highest scientific award in Germany given to him by the Deutsche Forschungsgemeinschaft (German Science Foundation). The authors of this introduction and his many collaborators thank him for his friendship and years of stimulating mathematics. We look forward to more in the future.

Publications by Wolfgang Dahmen (as of summer 2009)

1. (with E. Görlich) A conjecture of Golomb on optimal and nearly-optimal approximation, Bull. Amer. Math. Soc., **80** (1974), 1199-1202.
2. (with E. Görlich) Asymptotically optimal linear approximation processes and a conjecture of Golomb, In: Linear Operators and Approximation II, ed. by P.L. Butzer and B. Sz.-Nagy, ISNM 25, Birkhäuser, Basel, 1974.
3. (with E. Görlich) Best approximation with exponential error orders and intermediate spaces, Math. Z., **148** (1976), 7-21.
4. (with E. Görlich) The characterization problem for best approximation with exponential error orders and evaluation of entropy, Math. Nachr., **76** (1977), 163-179.
5. Trigonometric approximation with exponential error orders. I. Construction of asymptotically optimal processes; generalized de la Vallée Poussin sums, Math. Ann., **230** (1977), 57-74.
6. Trigonometric approximation with exponential error orders. II. Properties of asymptotically optimal processes; impossibility of arbitrarily good error estimates, Journal of Mathematical Analysis and Applications, **68** (1979), 118-129.
7. Trigonometric approximation with exponential error orders. III. Criteria for asymptotically optimal processes, Proc. of the International Conference on Approximation and Function Spaces, ed. by Z. Ciesielski, Gdansk, 1979.
8. On best approximation and de la Vallée sums, Mat. Zametkii, **5** (1978), 671-683.
9. (with K. Scherer) Best approximation with piecewise polynomials with variable knots and degrees, J. Approx. Theory, **26** (1979), 1-13.
10. Multivariate *B*-splines - recurrence relations and linear combinations of truncated powers, in: Multivariate Approximation Theory, ed. by W. Schempp, K. Zeller, Birkhäuser, Basel, (1979), 64-82.
11. Polynomials as linear combinations of multivariate *B*-splines, Math. Z., **169** (1979), 93-98.
12. Konstruktion mehrdimensionaler *B*-Splines and ihre Anwendungen auf Approximationsprobleme, in: Numerische Methoden der Approximationstheorie, vol. **5**, ed. by L. Collatz, G. Meinardus, H. Werner, Birkhäuser, Basel, (1980), 84-110.
13. On multivariate *B*-splines, SIAM J. Numer. Anal., **17** (1980), 179-191.
14. (with R. DeVore and K. Scherer) Multidimensional spline approximation, SIAM J. Numer. Anal., **17** (1980), 380-402.
15. Approximation by smooth multivariate splines on non-uniform grids, in: Quantitative Approximation, ed. by R. DeVore and K. Scherer, Academic Press, 1980, 99-114.
16. (with C.A. Micchelli) On limits of multivariate *B*-splines, J. d'Analyse Math., **39** (1981), 256-278.
17. (with A.S. Cavaretta, C.A. Micchelli and P.W. Smith) On the solvability of certain systems of linear difference equations, SIAM J. Math. Anal., **12** (1981), 833-841.
18. (with A.S. Cavaretta, C.A. Micchelli and P.W. Smith) A factorization theorem for banded matrices, Linear Algebra and its Applications, **39** (1981), 229-245.

19. (with C.A. Micchelli) On entire functions of affine lineage, Proc. Amer. Math. Soc., **84** (1982), 344-346.
20. Approximation by linear combinations of multivariate *B*-splines, J. Approx. Theory, **31** (1981), 299-324.
21. (with C.A. Micchelli) Computation of integrals and inner products of multivariate *B*-splines, Numer. Funct. Anal. and Optimiz. **3** (1981), 367-375.
22. (with C.A. Micchelli) On the linear independence of multivariate *B*-splines. I. Triangulations of simploids, SIAM J. Numer. Anal., **19** (1982), 993-1012.
23. Adaptive approximation by smooth multivariate splines, J. Approx. Theory, **36** (1982), 119-140.
24. (with C.A. Micchelli) Numerical algorithms for least squares approximation by multivariate *B*-splines, in: Numerische Methoden der Approximationstheorie, ed. by L. Collatz, G. Meinardus and H. Werner, Birkhäuser, Basel, 1981, 85-114.
25. (with C.A. Micchelli) On the linear independence of multivariate *B*-splines. II. Complete configurations, Math. Comp., **41** (1982), 143-163.
26. (with C.A. Micchelli) Some remarks on multivariate *B*-splines, in: Muiltivariate Approximation Theory, ed. by W. Schempp, and K. Zeller, ISNM 61, Birkhäuser, Basel, 1982, 81-87.
27. (with C.A. Micchelli) Multivariate splines - a new constructive approach, in: Surfaces in Computer Aided Geometric Design, ed. by R.E. Barnhill and W. Boehm, North-Holland, 1982, 191-215.
28. (with C.A. Micchelli) Translates of multivariate splines, Linear Algebra and its Applications, **52/53** (1983), 217-234.
29. (with C.A. Micchelli) On the approximation order of certain multivariate spline spaces, Journal of the Australian Mathematical Society, Ser. **B 26** (1984), 233-246.
30. (with C.A. Micchelli) Recent Progress in multivariate splines, in: Approximation Theory IV, ed. by C.K. Chui, L.L. Schumaker and J. Ward, Academic Press, 1983, 27-121.
31. (with C.A. Micchelli) On the approximation order of criss- cross finite element spaces, Journal of Computational and Applied Mathematics, **10** (1984), 255-273.
32. (with C.A. Micchelli) Some results on box splines, IBM RC 10094, July 1983, Bull. AMS, **1** (1984), 147-150.
33. (with C.A. Micchelli) On the local linear independence of translates of a box spline, Studia Mathematica **82** (1985), 243-263.
34. (with C.A. Micchelli) On the solution of certain systems of linear partial difference equations and linear dependence of translates of box splines, Trans. Amer. Math. Soc., **292** (1985), 305-320.
35. (with C.A. Micchelli) Subdivision algorithms for the generation of box spline surfaces, Computer Aided Geometric Design **1** (1984), 115-129.
36. (with C.A. Micchelli) On the multivariate Euler-Frobenius polynomials, in: Constructive Theory of Functions, ed. by B. Sendov, P. Petrushev, R. Maleev, S. Tashev, Publishing House of the Bulgarian Academy of Sciences, 1984, 237-243.
37. (with N. Dyn and D. Levin) On the convergence rates of subdivision algorithms for box spline surfaces, Constr. Approx., **1** (1985), 305-322.
38. (with C.A. Micchelli and P.W. Smith) Asymptotically optimal sampling schemes for periodic functions, Math. Proc. Camb. Phil. Soc., **99** (1986), 171-177.
39. Subdivision algorithms converge quadratically, Journal Computational and Applied Mathematics, **16** (1986), 145-158.
40. (with C.A. Micchelli) Line average algorithm: A method for the computer generation of smooth surfaces, Computer Aided Geometric Design **2** (1985), 77-85.
41. (with C.A. Micchelli) On the number of solutions to systems of linear diophantine equations and multivariate splines, Trans. Amer. Math. Soc., **308** (1988), 509-532.
42. (with C.A. Micchelli) Convexity of multivariate Bernstein polynomials and box spline surfaces, Studia Scientiarum Mathematicarum Hungarica, **23** (1988), 265-287.
43. (with C.A. Micchelli) Combinatorial aspects of multivariate splines, in: "Multivariate Approximations III", ed. by W. Schempp, K. Zeller, Birkhäuser, Basel, 1985, 120-137.

44. (with C.A. Micchelli, P.W. Smith) On factorization of bi-infinite totally positive block Toeplitz matrices, Rocky Mountain J. of Math., **16** (1986), 335-364.
45. Subdivision algorithms - recent results, some extensions and further developments, in Algorithms for Approximation, ed. by J.C. Mason, M.G. Cox, Clarendon Press, Oxford, 1987, 21-49.
46. (with C.A. Micchelli) On the piecewise structure of discrete box splines, Computer Aided Geometric Design, **3** (1986), 185-191.
47. (with C.A. Micchelli) Statistical encounters with B-splines, AMS contemporary Mathematics, **59** (1986), 17-48.
48. (with C.A. Micchelli) Algebraic properties of discrete box splines, Constr. Approx., **3** (1987), 209-221.
49. (with C.A. Micchelli) Some remarks on ridge functions, Approximation Theory and its Applications, **3** (1987), 139-143.
50. (with C.A. Micchelli) Theory and Applications of exponential splines, in: Proceedings of the International. Workshop on Multivariate Approximation, Santiago de Chile, Dec. 15-19, 86, 37-46, 1987.
51. (with C.A. Micchelli) On multivariate E-splines, Advances in Mathematics, **76** (1989), 33-93.
52. (with C.A. Micchelli, T.N.T. Goodman) Compactly supported fundamental functions for spline interpolation, Numerische Mathematik, **52** (1988), 639-664.
53. (with C.A. Micchelli, T.N.T. Goodman) Local spline interpolation schemes in one and several variables, in: Approximation and Optimization, Proceedings Havana, 1987, A. Gomez, F. Guerra, M.A. Jimenez, G. Lopez (Eds.) Springer Lecture Notes in Mathematics **1354**, 11-24, 1988.
54. (with R. Gmelig Meyling, J. Ursem) Scattered data interpolation by bivariate $C^{\mathbf{1}}$-piecewise polynomials, Approximation Theory and its Applications, **6** (1990), 6-29.
55. (with L. Elsner) Algebraic multigrid methods and the Schur complement, in: Robust Multi-Grid Methods, Proceedings of the Fourth GAMM-Seminar, Kiel, January 22-24, 1988, Notes on Numerical Fluid Mechanics, **23** (1988), 57-68.
56. (with C.A. Micchelli) Convexity and Bernstein Polynomials on k-Simploids, Acta Mathematicae Applicatae Sinica, **1** (1990), 50-66.
57. (with A.S. Cavaretta, C.A. Micchelli) The volume of restricted moment spaces, Rendiconti del Circolo Matematico di Palermo, Serie II, **38** (1989), 419-429.
58. (with C.A. Micchelli) Local dimension of piecewise polynomial spaces, syzygies and solutions of systems of partial differential equations, Mathematische Nachrichten, **148** (1990), 117-136.
59. Smooth piecewise quadric surfaces, in: Mathematical Methods in Computer Aided Geometric Design, T. Lyche, L.L. Schumaker, eds. Academic Press, 1989, 181-193.
60. A basis for certain spaces of multivariate polynomials and exponentials, Algorithms for Approximation II, J.C. Mason, M.G. Cox (Herausg.), Chapman and Hall, 1990, 80-98.
61. (with R.Q. Jia, C.A. Micchelli) On linear dependence relations for integer translates of compactly supported distributions, Mathematische Nachrichten, **151** (1991), 303-310.
62. (with A.S. Cavaretta, C.A. Michelli) Stationary subdivision, Memoirs of the Amer. Math. Soc., No. **453**, 1991.
63. (with R.Q. Jia, C.A. Micchelli) Linear dependence of cube splines revisited, in "Approximation Theory VI", C.K. Chui, L.L. Schumaker, J.D. Ward eds., Academic Press, Volume I, 1989, 161-164.
64. (with C.A. Micchelli) Stationary subdivision, fractals and wavelets, FU Preprint A-89-21, in: the Proceedings of the NATO ADVANCED STUDY INSTITUTE series, Computation of curves and Surfaces, Hersg. M. Gasca, W. Dahmen, C.A. Micchelli, Kluwer Academic Publishers, 3-26, 1990.
65. (with C.A. Micchelli) On stationary subdivision and the construction of compactly supported orthonormal wavelets, in: Multivariate Approximation and Interpolation, W. Haussmann, K. Jetter (Herausg.), Birkhäuser, 1990, 60-90.

66. (with T. M. Thamm) Cubicoids: Modeling and visualization, Computer Aided Geometric Design, **10** (1993), 89-108.
67. (with A. Dress, C.A. Micchelli) On multivariate splines, matroids and the ext-functor, Advances in Applied Mathematics, **17** (1996), 251-307.
68. Convexity and Bernstein-Bézier polynomials, in: Curves and Surfaces, P.J. Laurent, A. Le Méhauté, L.L. Schumaker eds., Academic Press, 1991, 107-134.
69. (with H.P. Seidel, C.A. Micchelli) Blossoming begets B-spline bases built better by B-patches, Mathematics of Computation, **59** (1992), 97-115.
70. (with R. DeVore, C.A. Micchelli) On monotone extensions of boundary data, Numer. Math., **60** (1992), 477-492.
71. (with C.A. Micchelli) Using the refinement equation for evaluating integrals of wavelets, SIAM J. Numer. Anal., **30** (1993), 507-537.
72. (with J.M. Carnicer) Characterization of local strict convexity preserving interpolation methods by C^1 functions, Journal of Approximation Theory, **77** (1994), 2-30.
73. (with J.M. Carnicer) Convexity preserving interpolation and Powell-Sabin elements, Computer Aided Geometric Design, **9** (1992), 279-289.
74. (with P. Oswald, X.Q. Shi) C^1-conforming hierarchical bases, Journal of Computational and Applied Mathematics, 9(1993), 263-281.
75. (with A. Kunoth) Multilevel preconditioning, Numer. Math., **63** (1992), 315-344.
76. (with C.A. Micchelli) Banded matrices with banded inverses II: Locally finite decompositions of spline spaces, Constr. Approx., **9** (1993), 263-281.
77. (with C.A. Micchelli) Continuous refinement equations and subdivision, Advances in Computational Mathematics, **1** (1993), 1-37.
78. Decomposition of refinable spaces and applications to operator equations, Numerical Algorithms, **5** (1993), 229-245.
79. (with S. Prössdorf, R. Schneider) Wavelet approximation methods for pseudodifferential equations I: Stability and convergence, Mathematische Zeitschrift, **215** (1994), 583-620.
80. (with S. Prössdorf, R. Schneider) Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution, Advances in Computational Mathematics, **1** (1993), 259-335.
81. (with J.M. Carnicer, J.M. Peña) Local decomposition of refinable spaces and wavelets, Applied and Computational Harmonic Analysis, **3** (1996), 127-153.
82. (with M Butzlaff, S. Diekmann, A. Dress, E. Schmitt, E. V. Kitzing) A hierarchical approach to force field calculations through spline approximations, Journal of Mathematical Chemistry, **15** (1994), 77-92.
83. (mit B. Raabe, T.-M. Thamm) Optimal geometry representation for rendering, in: Visualisierung in Mathematik, Technik und Kunst, Grundlagen und Anwendungen, A. Dress, G. Jäger (eds.), Vieweg-Verlag, Braunschweig, 1999, 23–49.
84. (with S. Prössdorf, R. Schneider) Multiscale methods for pseudodifferential equations, in: 'Recent Advances in Wavelet Analysis' L.L. Schumaker, G. Webb eds., Academic Press, 1993, 191-235.
85. (with S. Dahlke, V. Latour) Smooth refinable functions and wavelets obtained by convolution products, Applied and Computational Harmonic Analysis, **2** (1995), 68-84.
86. (with B. Kleemann, S. Prössdorf, R. Schneider) A multiscale method for the double layer potential equation on a polyhedron, in: Advances in Computational Mathematics, H.P. Dikshit, C.A. Micchelli, eds., World Scientific, 1994, 15-57.
87. Some remarks on multiscale transformations, stability and biorthogonality, in: Wavelets, Images and Surface Fitting, P.J. Laurent, A. Le Méhauté, L.L. Schumaker eds., Academic Press, 157-188, 1994.
88. (with A. Cohen, R. DeVore) Multiscale decompositions on bounded domains, Trans. Amer. Math. Soc., No. 8, **352** (2000), 3651–3685.
89. (with S. Prössdorf, R. Schneider) Multiscale methods for pseudo-differential equations on smooth manifolds, in: Proceedings of the International Conference on Wavelets: Theory, Algorithms, and Applications, C.K. Chui, L. Montefusco, L. Puccio (eds.), Academic Press, 385-424, 1994.

90. (with S. Prössdorf, R. Schneider) Wavelets zur schnellen Lösung von Randintegralgleichungen und angewandte harmonische Analyse, Z. Angew. Math. Mech. 74, No 6 (1994), 505-507.
91. Multiscale Techniques - Some Concepts and Perspective, in: Proceedings of the International Congress of Mathematicians, Zürich 94, Birkhäuser, Basel, 1995, 1429-1439.
92. Stability of multiscale transformations, Journal of Fourier Analysis and Applications, **2** (1996), 341-361.
93. (with S. Dahlke, E. Schmitt, I. Weinreich) Multiresolution analysis and wavelets on S^2 and S^3, Numerical Functional Analysis and Optimization, **16** (1995), 19-41.
94. (with A. Kunoth, K. Urban) Wavelet-Galerkin method for the Stokes equations, Computing, **56** (1996), 259-301.
95. (with C.A. Micchelli) Biorthogonal wavelet expansions, Constr. Approx., **13** (1997), 293–328.
96. Multiskalen-Methoden und Wavelets – Konzepte und Anwendungen, DMV–Jahresbericht, **97** (1995), 97-114.
97. (with B. Kleemann, S. Prössdorf, R. Schneider) Multiscale methods for pseudodifferential equations, *Proceedings des ICIAM /GAMM Hamburg 1995 - Kongresses*, (1996). Z. Angew. Math. Mech. 76 Suppl.1 7–10.
98. Multiscale analysis, approximation, and interpolation spaces, in: Approximation Theory VIII, Vol 2., Wavelets and Multilevel Approximation, C.K. Chui, L.L. Schumaker (eds.), World Scientific Publishing Co, 1995, 47-88.
99. (with S. Dahlke, R. Hochmuth, R. Schneider) Stable multiscale bases and local error estimation for elliptic problems, Applied Numerical Mathematics, **23** (1997), 21-48.
100. (with R. Schneider) Composite wavelet bases for operator equations, Math. Comp., **68** (1999), 1533-1567.
101. (with B. Kleemann, S. Prössdorf, R. Schneider) Multiscale methods for the solution of Helmholtz and Laplace equations, in: Boundary Element Methods, Reports from the Final Conference of the Priority Research Programme 1989-1995 of the German Research Foundation, Oct. 2-4, 1995 in Stuttgart, W. Wendland, ed., Springer-Verlag, 1996.
102. (with R. V. Watzdorf, K. Urban, W. Marquardt) A wavelet-Galerkin method applied to separation processes, in: *Scientific Computing in Chemical Engineering*, S. Keil, W. Mackens, H. Voß, J. Werther (Hersg.), Springer-Verlag, Berlin, 246-252, 1996.
103. (with A. Kunoth, R. Schneider) Operator equations, multiscale concepts and complexity, in: Proceedings, *1995 AMS-SIAM Summer Seminar, Mathematics of numerical Analysis: Real Number Algorithms* J. Renegar, M. Shub, S. Smale (eds.), Park City, 1995, Lectures in Applied Mathematics, **32** (1996), 225-261.
104. (with A. Kunoth, K. Urban) Biorthogonal spline-wavelets on the interval – Stability and moment conditions, Applied and Computational Harmonic Analysis, **6** (1999), 132-196.
105. (with A. Kunoth, K. Urban) Wavelets in numerical analysis and their quantitative properties, in: Surface Fitting and Multiresolution Methods, A. Le Méhauté, C. Rabut, L.L. Schumaker (eds.), Vanderbilt University Press, 1997, 93-130.
106. (with S. Dahlke, R. DeVore) Nonlinear approximation and adaptive techniques for solving elliptic operator equations, in: Wavelet Multiscale Methods for PDEs, Academic Press, W. Dahmen, A. Kurdila, P. Oswald eds., London, 1997, 237-283, 1997.
107. (with R. Schneider) Wavelets on manifolds I. Construction and domain decomposition, SIAM Journal on Mathematical Analysis, **31** (1999), 184-230.
108. Wavelet and Multiscale Methods for Operator Equations (invited contribution), Acta Numerica, Cambridge University Press, **6** (1997), 55–228.
109. (with D. Braess) A cascade algorithm for the Stokes equations, Numer. Math., **82** (1999), 179-191.
110. (with S. Müller, T. Schlinkmann) Multigrid and multiscale decompositions, in: Large-Scale Scientific Computations of Engineering and Environmental Problems, M. Griebel, O.P. Iliev, S.D. Margenov, P.S. Vassilevski, eds., Notes on Numerical Fluid Mechanics, Vol. 62, Vieweg, Braunschweig/Wiesbaden, 18–41, 1998.

111. (with T. Binder, L. Blank, W. Marquardt) Towards multiscale dynamic data reconciliation, in: *Nonlinear Model based Process Control* NATO ASI, (R. Berber and C. Kravaris, eds.), Kluwer Academic Publishers, 623–665, 1998.
112. (with R. Stevenson) Element-by element construction of wavelets satisfying stability and moment conditions, SIAM J. Numer. Anal., **37** (No. 1) (1999), 319–325.
113. (with B. Han, R.-Q. Jia, A. Kunoth) Biorthogonal multiwavelets on the interval: Cubic Hermite splines, Constr. Approx. **16** (2000), 221–259.
114. B-splines in analysis, algebra and applications, Traveaux Mathématiques, 1999, 15–76.
115. (with R. Schneider) Wavelets with complementary boundary conditions – Function spaces on the cube, Results in Mathematics, **34** (1998), 255–293.
116. (with D. Braess, C. Wieners) A multigrid algorithm for the mortar finite element method, SIAM J. Numer. Anal., **37** (No. 1) (1999), 48–69.
117. (with R. Schneider, Y. Xu) Nonlinear functions of wavelet expansions – Adaptive reconstruction and fast evaluation, Numer. Math., **86** (2000), 49-101.
118. (with A. Cohen, R. DeVore) Adaptive wavelet methods for elliptic operator equations – Convergence rates, Math. Comp. **70** (2001), 27–75.
119. (with B. Gottschlich-Müller, S. Müller) Multiresolution schemes for conservation laws, Numer. Math., **88** (2001), 399-443.
120. (with D. Braess) Stability estimates of the mortar finite element method for 3-dimensional problems, East West J. Numer. Math., **6** (4) (1998), 249–264.
121. (with A. Cohen, I. Daubechies, R. DeVore) Tree approximation and optimal encoding, Applied and Computational Harmonic Analysis, **11** (2001), 192–226.
122. (with A. Kunoth) Appending boundary conditions by Lagrange multipliers: Analysis of the LBB condition, Numer. Math., **88** (2001), 9–42.
123. (with S. Müller, T. Schlinkmann) On an adaptive multigrid solver for convection-dominated problems, SIAM J. Scient. Comp., **23** (No 3)(2001), 781–804.
124. (with A. Kunoth, R. Schneider) Wavelet least squares methods for boundary value problems, Siam J. Numer. Anal. **39**, No. 6, (2002), 1985–2013.
125. (with A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, K. Urban) Adaptive wavelet schemes for elliptic problems – implementation and numerical experiments, SIAM J. Sci. Comp., **23**, No. 3 (2001), 910–939.
126. Wavelet methods for PDEs – Some recent developments, J. Comp. Appl. Math., **128** (2001), 133–185.
127. (with T. Binder, L. Blank, W. Marquardt) Iterative algorithms for multiscale state estimation, Part I, Concepts, J. Opt. Theo. Appl. **111**(3)(2001), 529–551.
128. (with T. Binder, L. Blank, W. Marquardt) An adaptive multiscale method for real time moving horizon optimization, Proc. American Control Conference 2000, Chicago, Illinois June 2000, pp 4234 - 4238, Omnipress.
129. (with T. Binder, L. Blank, W. Marquardt) Grid refinement in multiscale dynamic optimization, Proc. European Symposium on Computer Aided Process Engineering 10, Florence, Italy, Mai 2000, (Ed. S. Pierucci), Elsevier, Amsterdam, pp 31-37.
130. (with I. Graham, B. Faehrmann, W. Hackbusch, S. Sauter) Inverse inequalities on non-quasiuniform meshes and application to the mortar element method, Math. Comp. **73** (2004), 1107-1138.
131. (with A. Cohen, I. Daubechies, R. DeVore) Harmonic analysis of the space BV, Revista Matematica Iberoamericana **19** (2003), 1-29.
132. (with Th. Binder, L. Blank, W. Marquardt) Regularization of dynamic data reconciliation problems by projection, Proc. IFAC Symposium on Advanced Control of Chemical Processes, ADCHEM 2000, Pisa, Italy, June 2000. Ed. L. T. Biegler, A. Brambilla, C. Scali, pp 689 - 694, Vol. 2
133. (with A. Cohen, R. DeVore) Adaptive wavelet methods II - Beyond the elliptic case, Foundations of Computational Mathematics, **2** (2002), 203–245.
134. (with H.P. Dikshit, A. Ojha) On Wachspress quadrilateral patches, Computer Aided Geometric Design, **17** (2000), 879–890.

135. (with T. Binder, L. Blank, W. Marquardt) Iterative algorithms for multiscale state estimation, Part II: Numerical investigations, J. Opt. Theo. Appl. **111**(3)(2001), 529–551.
136. (with T. Binder, L. Blank, W. Marquardt) Multiscale concepts for moving horizon optimization, in: *Online Optimization for Large Scale Systems* (M. Grötschel, S.O. Krumke, and J.Rambau, eds.), Springer, Berlin, 341–362, 2001.
137. (with T. Binder, L. Blank, W. Marquardt) On the regularization of dynamic data reconciliation problems, Journal of Process Control, **12** (2002), 557-567.
138. (with T. Binder, L. Blank, W. Marquardt) Iterative multiscale methods for process monitoring, in: Proc. of 'Fast Solutions of Discretized Optimization Problems', WIAS Berlin, May 8–12, Birkhäuser Verlag, 2000.
139. (with S. Dahlke, K. Urban) Adaptive wavelet methods for saddle point problems – Convergence rates, SIAM J. Numer. Anal., **40** (No. 4) (2002), 1230–1262.
140. (with D. Braess) The mortar element method revisited – What are the right norms ? in: Domain Decomposition Methods in Science and Engineering, N. Debit, M. Garbey, R. Hoppe, D. Keyes, Y. Kutznetzov, J. Périaux, eds., International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 2002, 27–40.
141. (with F. Bramkamp, B. Gottschlich-Müller, Ph. Lamby, M. Hesse, S. Müller, J. Ballmann, K.-H. Brakhage) H-adaptive multiscale schemes for the compressible Navier-Stokes equations – Polyhedral discretization, data compression and mesh generation, in: Notes on Numerical Fluid Mechanics, Flow Modulation and Fluid-Structure-Interaction at Airplane Wings (ed. by J. Ballmann), Springer-Verlag, Vol. 84, 125–204, 2003.
142. Wavelets als mathematisches Mikroskop, Festschrift, 100 Jahre Berliner Mathematische Gesellschaft, WB-Druck, Rieden im Allgäu, Berlin, 2001.
143. (with T. Klint, K. Urban) On fictitious domain formulations for Maxwell's equations, J. Foundation of Computational Mathematics, **3** (2)(2003), 135–160.
144. (with K. Urban, J. Vorloeper) Adaptive wavelet methods – basic concepts and applications, in: Wavelet Analysis – Twenty Years Developments, Ding–Xuan Zhou ed., World Scientific, New Jersey, 2002, 39–80.
145. (with P. Binev, R. DeVore) Adaptive Finite Element Methods with Convergence Rates, Numer. Math., **97** (2004), 219–268.
146. (with H. Harbrecht, R. Schneider) Compression techniques for boundary integral equations – Asymptotically optimal complexity estimates, SIAM J. Numer. Anal., **43**(6) (2006), 2251–2271.
147. (with A. Cohen, R. DeVore) Adaptive wavelet schemes for nonlinear variational problems, SIAM J. Numer. Anal., **41**(5) (2003), 1785–1823.
148. (with A. Cohen, R. DeVore) Sparse evaluation of compositions of functions using multiscale expansions, SIAM J. Math. Anal., 35 (2003), 279–303.
149. (with P. Binev, R. DeVore, P. Petruchev) Approximation classes for adaptive methods, Serdica Math. J., **28** (No 4) (2002), 391–416.
150. (with A. Kunoth) Adaptive wavelet methods for linear–quadratic elliptic control problems: Convergence rates, SIAM J. Contr. Optim. **43** (5) (2005), 1640–1675.
151. (with A. Cohen, R. DeVore) Adaptive wavelet techniques in Numerical Simulation, in: *Encyclopedia of Computational Mechanics*, (R. De Borste, T. Hughes, E. Stein, eds.), Wiley-Interscience, 2004, 157–197.
152. (with A. Barinka, S. Dahlke) Adaptive application of operators in standard representation, Advances in Computational Mathematics, **24** (2006), 5–34.
153. Multiscale and Wavelet Methods for Operator Equations, C.I.M.E. Lecture Notes in Mathematics, *Multiscale Problems and Methods in Numerical Simulation*, Springer Lecture Notes in Mathematics, vol 1825, Springer-Verlag, Heidelberg, 2003, 31–96.
154. (with S. Müller, A. Voß) Riemann problem for the Euler equations with non-convex equation of state including phase transitions, in: *Analysis and Numerics for Conservation Laws*, G. Warnecke ed., Springer-Verlag, Berlin-Heidelberg, 2005, 137–162.
155. (with M. Campos-Pinto, A. Cohen, R. DeVore) On the stability of nonlinear conservation laws in the Hausdorff metric, Journal of Hyperbolic Differential Equations, **2**(1) (2005), 1-14.

156. (with P. Petrushev) "Push the error" algorithm for nonlinear n-term approximation, Constructive Approximation, **23**(3)(2006), 261–304.
157. (with P. Binev, R. DeVore, N. Dyn) Adaptive approximation of curves, in: Approximation Theory: a volume dedicated to Borislav Bojanov, 43-57, Prof. M. Drinov Acad. Publ. House, Sofia, 2004.
158. (with A. Barinka, R. Schneider) Fast computation of adaptive wavelet expansions, Numer. Math., **110**(4) (2007), 549–589.
159. (with A. Kurdila, M. Nechyba, R. Prazencia, P. Binev, R. DeVore, R. Sharpley) Vision-based control of micro-air-vehicles — Progress and problems in estimation, 43rd IEEE Conference on Decision and Control, Dec. 14–17, 2004, Atlantis, Paradise Island, Bahamas.
160. (with P. Binev, A. Cohen, R. DeVore, V. Temlyakov) Universal algorithms for learning theory - Part I : piecewise constant functions, Journal of Machine Learning Research, **6** (2005), 1297-1321.
161. (with R. Massjung, J. Hurka, J. Ballmann) On well-posedness and modeling for nonlinear aeroelasticity, Notes on Numerical Fluid Mechanics, Springer Verlag, vol. 84, 2003.
162. (with H. Harbrecht, R. Schneider) Adaptive methods for boundary integral equations — Complexity estimates, Math. Comp., **76** (259) (2007), 1243–1274.
163. (with M. Jürgens) Error controlled regularization by projection, ETNA, **25** (2006), 67–100.
164. (with P. Binev, A. Cohen, R. DeVore) Universal algorithms for learning theory - Part II : piecewise polynomial functions, Constr. Approx. **26**(2) (2007), 127-152.
165. (with A. Kunoth, J. Vorloeper) Convergence of adaptive wavelet methods for goal-oriented error estimation, in: Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2005, A. Bermúdez de Castro, D. Gómez, P. Quintela, P. Salgado, eds., Springer-Verlag, Berlin-Heidelberg, 2006.
166. (with A. Barron, A. Cohen, R. DeVore) Approximation and learning by greedy algorithms, Annals of Statistics, **3** (No 1)(2008), 64–94.
167. (with A. Cohen, R. DeVore) Compressed Sensing and best k-term approximation, J. Amer. Math. Soc. 22 (2009), no. 1, 211–231.
168. (with P. Binev, A. Cohen, R. DeVore) Universal Piecewise Polynomial Estimators for Machine Learning, in: Curve and Surface Design: Avignon 2006, P. Chenin, T. Lyche, L.L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2007, 48–77.
169. (with S. Dekel, P. Petrushev) Multilevel preconditioning for partition of unity methods — Some analytic concepts, Numer. Math., **107** (2007), 503–532.
170. (with S. Dekel, P. Petrushev) Two-level split decomposition of anisotropic Besov spaces, IGPM Report # 269, RWTH Aachen, January 2007.
171. (with A. Cohen, R. DeVore) A taste of compressed sensing, SPIE Conference Proceedings: Independent Component Analyses, Wavelets, Unsupervised Nano-Biometric Sensors, and Neural Networks V, Harold H. Szu, Jack Agee, editors, Vol. 6576, 9 April, 2007, ISBN: 9780819466983.
172. (with K. Brix, M. Campos Pinto) A multilevel preconditioner for the interior penalty discontinuous Galerkin method, SIAM J. Numer. Anal., **46**(5) (2008), 2742–2768.
173. (with K. Brix, M. Campos Pinto, R. Massjung) A multilevel preconditioner for the interior penalty discontinuous Galerkin method II - Quantitative studies, Commun. Comput. Phys., **5** (2009), 296-325.
174. (with T. Rohwedder, R. Schneider, A. Zeiser) Adaptive eigenvalue computation - complexity estimates, Numer. Math. 110 (2008), no. 3, 277–312.
175. (with A. Cohen, R. DeVore) Instance optimal decoding by thresholding in compressed sensing, IGPM Report, Nov. 2008, RWTH Aachen.

The way things were in multivariate splines: A personal view

Carl de Boor

Abstract A personal account of the author's encounters with multivariate splines during their early history.

1 Tensor product spline interpolation

My first contact with multivariate splines occurred in August 1960. In my first year at the Harvard Graduate School, working as an RA for Garrett Birkhoff, I had not done too well but, nevertheless, had gotten married and so needed a better income than the RAship provided. On the (very kind and most helpful) recommendation of Birkhoff who consulted for the Mathematics Department at General Motors Research in Warren MI, I had been hired in that department in order to be of assistance to Leona Junko, the resident programmer in that department.

Birkhoff and Henry L. Garabedian, the head of that department, had developed a scheme for interpolation to data on a rectangular grid meant to mimic cubic spline interpolation; see [BG]. They would use what they called "linearized spline interpolation" and what is now called cubic spline interpolation, along the meshlines in both directions, in order to obtain values of the first derivative in both directions at each meshpoint, and then fill in each rectangle by a C^1 piecewise low-degree harmonic polynomial function that would match the given information, of value and two first-order derivatives, at each corner and, thereby, match the cubic spline interpolants along the mesh-lines.

It occurred to me that the same information could be matched by a scheme that would, say, construct the cubic spline interpolants along all the mesh-lines in the x-direction, and then use cubic spline interpolation to the resulting spline coefficients as a function of y to obtain an interpolant that was a cubic spline in x for every value of y, and C^2 rather than just C^1. Of course, one could equally well start with the

Carl de Boor
Department of Computer Sciences, University of Wisconsin-Madison,
e-mail: deboor@cs.wisc.edu, URL: http://pages.cs.wisc.edu/∼deboor/

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_2,

cubic spline interpolants along all the mesh-lines in the y-direction, and interpolate the resulting spline coefficients as a function of x and so obtain an interpolant that was a cubic spline in y for every x, and it took me some effort to convince Birkhoff that these two interpolants are the same. This is now known as bicubic spline interpolation [dB0], the tensor-product (I learned that term from Don Thomas there) of univariate cubic spline interpolation, and has become a mainstay in the construction of smooth interpolants to gridded data. I did write up a paper on n-variable tensor product interpolation, but Birkhoff thought publication of such a paper unnecessary.

Much later (see [dB4]), I realized that it is quite simple to form and use in a multivariate context tensor products of univariate programs for the approximation and evaluation of functions, provided the univariate programs can handle vector-valued functions.

Around 1960, there was related work (I learned much later) by Feodor Theilheimer of the David Taylor Model Basin, see [TS], and, in computer graphics, parametric bicubic splines were introduced around that time by J. C. Ferguson at Boeing, see [Fe], though Ferguson set the crossderivatives D_xD_yf at all mesh points to zero, thereby losing C^2 and introducing flat spots.

In this connection, I completely missed out on parametric spline work, believing (incorrectly, I now know) that it is sufficient to work with spline functions on a suitably oriented domain. Nor did I get involved in the blending approach to the construction of spline surfaces, even though I was invited by Garabedian on a visit in 1962 to Coons at M.I.T. (my first plane ride) and saw there, first-hand, an ashtray being machined as a Coons' surface [C]. The paper [BdB] (which has my name on it only grace Birkhoff's generosity) contains a summary of what was then known about multivariate splines. I had left General Motors Research by the time that Bill Gordon did his work on spline-blended surfaces there; see, e.g., [G].

2 Quasiinterpolation

My next foray into multivariate splines occurred in joint work with George Fix, though my contribution to [dBF] was the univariate part (Birkhoff objected to the publication of two separate papers). Fix had worked out the existence of a local linear map into the space of tensor-product splines of (coordinate-)degree $< k$ for a given mesh, which depended only on the value of derivatives of order $< k_0 \le k$ at all the mesh points but did not necessarily reproduce those values (hence Fix' name "quasi-interpolate" for the resulting approximation) but did reproduce all polynomials of (total) degree $< k_0$ in such a way that the approximation error can be shown to be of order k_0 in the mesh size. However, there was an unresolved argument between Fix and his thesis advisor, Garrett Birkhoff, about whether, in the univariate case, Fix' scheme was "better" than Birkhoff's "Local spline approximation by moments" [B], and Birkhoff had invited me to Cambridge MA for July 1970 to settle the matter, perhaps. ([B] started out as a joint paper but, inexplicably, did not so end up; I published the case of even-degree splines later on in [dB1].) Fortunately, once I had

derived an explicit formula for Fix' map, the two methods could easily be seen to be identical.

For $k_0 = k$, Fix's univariate scheme amounted to interpolation in the sense that it was a linear projector; nevertheless, it was called "quasi-interpolation" in the spirit of finite elements of that time since its purpose was not to match given function values but, rather, to match some suitable linear information in such a way that the process was local, stable, and reproduced all polynomials of order k, thus ensuring approximation order k. In this sense, Birkhoff's local spline approximation by moments is the first quasi-interpolation spline scheme I am aware of (with [dB2] a close and derivative-free second).

Unfortunately, it was only ten years later that I became aware of Frederickson's immediate reaction [Fr1] to [dBF] in which he constructed quasi-interpolant schemes onto smooth piecewise polynomials on what we now would call the 3-direction mesh, using bump functions obtained from the characteristic function of a triangle in the same way we now obtain a bivariate box spline from the characteristic function of a square; see [Fr2].

3 Multivariate B-splines

In 1972, I moved to Madison WI, to the Mathematics Research Center (MRC) funded since 1957 by the United States Army Research Office to carry out research in applied mathematics. It had an extensive postdoc and visitors program, the only fly in the ointment its location far from the center of the University of Wisconsin-Madison because its former housing there was bombed in August 1970, as a protest against the Vietman war, by people who took the very absence of any mention of military research in the semi-annual reports of that Army-financed institution as proof of the importance of the military research supposedly going on there. I had been hired at the time of I. J. Schoenberg's retirement from MRC.

The univariate spline theory was in good shape by that time, and, thanks to my contacts with Martin Schultz and George Fix, and to having been asked to handle the MRC symposium on the "Mathematical Aspects of Finite Elements in Partial Differential Equations" in the summer of 1973, I had begun to look at smooth piecewise polynomials in two and more variables, as they were being used in finite elements. That same summer, I participated in the Numerical Analysis conference in Dundee and heard Gil Strang's talk [St] there, in which he raised the question of the dimension of the space of bivariate $C^{(1)}$-cubics on a given triangulation. I felt like a fraud for not being able to solve that problem right then and there. As it turned out, except for "nice" triangulations, this problem is still not understood in 2009, and neither is the approximation power of such spaces known, although many have worked on it; see [LS] for what was known by 2007.

At the same time, in practice, the finite element method did not work with the space of *all* piecewise polynomials of a certain degree and smoothness on a

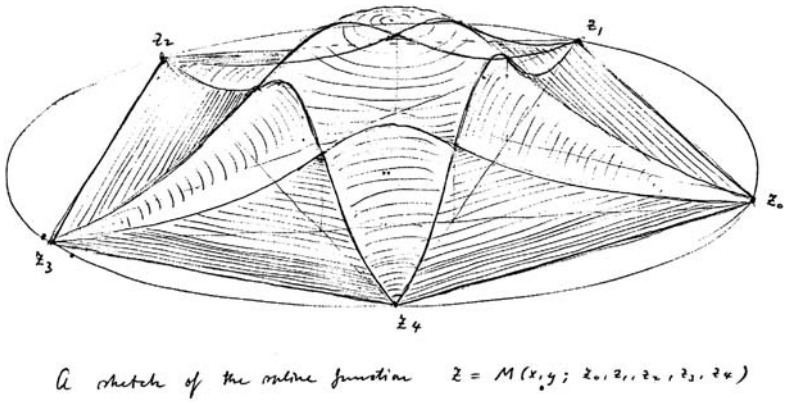


Fig. 1 Schoenberg's sketch of a bivariate quadratic B-spline

given triangulation, but with suitable subspaces, usually the linear span of suitable compactly supported more or less smooth piecewise polynomials called bump or hill functions. This, together with the essential role played by B-splines in the univariate spline theory (as summarized, e.g., in [dB3]), made me look for "B-splines", i.e., smooth compactly supported piecewise polynomials, in the multivariate setting. When discussing this issue in January 1975 with Iso Schoenberg in his home study, he went to his files and pulled out a letter [Sc] he had written to Phil Davis in 1965, with a drawing of a bivariate compactly supported piecewise quadratic function, with several planar sections drawn in as univariate quadratic B-splines; see Figure 1. The letter was in response to Davis' paper [Dav], meant to publicize the following formula, due to Motzkin and Schoenberg,

$$\frac{1}{2A}\int_T f''(z)\,\mathrm{d}x\,\mathrm{d}y = \frac{f(z_0)}{(z_0-z_1)(z_0-z_2)} + \frac{f(z_1)}{(z_1-z_0)(z_1-z_2)} + \frac{f(z_2)}{(z_2-z_0)(z_2-z_1)}, \tag{1}$$

valid for all functions f regular in the triangle T in the complex plane with vertices z_0, z_1, z_2, and with A the area of T. Schoenberg points out that, in as much as the right side of (1) is the second divided difference $\Delta(z_0, z_1, z_2)f$ of f at z_0, z_1, z_2, therefore the Genocchi-Hermite formula for the nth divided difference

$$\Delta(z_0,\ldots,z_n)f = \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{n-1}} f^{(n)}(z_0 + s_1\nabla z_1 + \cdots + s_n \nabla z_n)\,\mathrm{d}s_n \cdots \mathrm{d}s_1 \tag{2}$$

provides a ready generalization of (1) to an arbitrary finite collection of z_i in the complex plane. Moreover, it is possible to write the integral as a weighted integral over the convex hull of the z_i, i.e., in the form

$$\mathfrak{A}(z_0,\dots,z_n)f = \int_{\mathrm{conv}(z_0,\dots,z_n)} f^{(n)}(x+\mathrm{i}y)M(x,y;z_0,\dots,z_n)\,\mathrm{d}x\,\mathrm{d}y,$$

with the value at (x,y) of the weight function $M(\cdot,\cdot;z_0,\dots,z_n)$ the volume of $\sigma\cap P^{-1}\{(x,y)\}$, with P the orthogonal projector of $\mathbb{R}^n$ onto $\mathbb{C}\sim\mathbb{R}^2\subset\mathbb{R}^n$, and σ any n-simplex of unit volume whose set of vertices is mapped by P onto $\{z_0,\dots,z_n\}$. This makes $M(\cdot,\cdot;z_0,\dots,z_n)$ the two-dimensional "X-ray" or "shadow" of an n-dimensional simplex. Hence, $M(\cdot,\cdot;z_0,\dots,z_n)$ is piecewise polynomial in x,y of total degree $n-2$, nonnegative, and nonzero only in the convex hull of the z_j, and, generically, in $C^{(n-3)}$. This is strikingly illustrated in Figure 1, which shows Schoenberg's sketch of the weight function for the case $n=4$, with the z_j the five fifth-root of unity, giving a C^1 piecewise quadratic weight function.

I was much taken by this geometric construction since it immediately suggested a way to get a nonnegative partition of unity consisting of compactly supported smooth piecewise d-variate polynomials of order k: In $\mathbb{R}^k$, take a convex set C of unit k-dimensional volume (e.g., a simplex), and subdivide the cylinder $C\times\mathbb{R}^d$ into non-trivial ($k+d$-dimensional) simplices. Then their shadows on $\mathbb{R}^d$ under the orthogonal projection of $\mathbb{R}^{d+k}$ onto $\mathbb{R}^d$ provide that partition of unity. For the case $d=1$, Schoenberg was very familiar with the resulting 1-dimensional shadows of $1+k$-dimensional simplices. By the Hermite-Genocchi formula, they are univariate B-splines, a fact used by him in [CS] to prove the log-concavity of the univariate B-spline.

In a talk [dB3] at the second Texas conference in 1976, on the central role played by B-splines in the univariate spline theory, I finished with a brief discussion of what little I knew about Schoenberg's multivariate B-splines. In particular, I stressed the lack of recurrence relations to match those available for univariate B-splines, and should have pointed out that I had no idea (except when $d=1$) how to choose the partition of $C\times\mathbb{R}^d$ into simplices in order to ensure that the linear span of the resulting d-dimensional shadows has nice properties. A very alliterative solution to this difficult problem was offered in [DMS] but, to me, the most convincing solution is the one finally given by Mike Neamtu; see [N] and the references therein (although Höllig's solution [H2] is not mentioned).

Subsequently, Karl Scherer informed me that his new "Assistent", Dr. Wolfgang Dahmen, intended to provide the missing recurrence relations. It seems that Scherer had given him [dB3] to read as an introduction to splines.

4 Kergin interpolation

In January 1978, I was asked by T. Bloom of Toronto (possibly because his colleague, Peter Rosenthal, and I had been students together at Ann Arbor) my opinion of a recent result of one of his students, Paul Kergin, and, for this purpose, was sent a handwritten draft of Kergin's Ph.D. thesis [K1]. The thesis proposed a remarkable generalization of univariate Lagrange interpolation from $\Pi_{\le k}$ at a

$k+1$-set $Z = \{z_0, \dots, z_k\}$ of sites to the multivariate setting, with the interpolant chosen uniquely from $\Pi_{\le k}$ (:= the space of polynomials in d variables of total degree $\le k$) and depending continuously on the sites even when there was coalescence and, correspondingly, Hermite interpolation. To be sure, in $d > 1$ dimensions, $\dim \Pi_{\le k} = \binom{k+d}{d}$ is much larger than $k+1$, hence Kergin had to choose additional interpolation conditions in order to single out a particular element $Pf \in \Pi_{\le k}$ for given f. This he did in the following way. He required that P be linear and such that, for every $0 \le j \le k$ and every homogeneous polynomial q of degree j, and every $j+1$-subset Σ of Z, $q(D)(\mathrm{id} - P)f$ should vanish at some site in $\mathrm{conv}(\Sigma)$.

The thesis (and subsequent paper [K2]) spends much effort settling the question of how all these conditions could be satisfied simultaneously, and, in discussions with members and visitors at MRC that Spring, we looked for some simplification. Michael Golomb pointed to the "lifting" Kergin used in his proof as a possible means for simplification: If the interpoland f is a "ridge function", i.e., of the form $g \circ \lambda$ with λ a linear functional on $\mathbb{R}^d$, then Pf is of the same form; more precisely, then $Pf = (Qg) \circ \lambda$, with Qg the univariate polynomial interpolant to g at the possibly coalescent sites $\lambda(Z)$.

Fortunately, C. A. Micchelli was visiting MRC that year, from 1apr to 15sep, and readily entered these ongoing discussions on Kergin interpolation (and the missing recurrence relations for multivariate B-splines). He extended (see [M1]) the linear functional occurring in the Genocchi-Hermite formula (2) to functions of d variables by setting

$$\int_{[z_0,\dots,z_n]} h := \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{n-1}} h(z_0 + s_1 \nabla z_1 + \cdots + s_n \nabla z_n)\, \mathrm{d}s_n \cdots \mathrm{d}s_1 \qquad (3)$$

for arbitrary $z_0, \dots, z_n \in \mathbb{R}^d$, recalled the Newton form

$$Qg = \sum_{j=0}^{k} (\cdot - \lambda z_0) \cdots (\cdot - \lambda z_{j-1}) \mathbb{A}(\lambda z_0, \dots, \lambda z_j) g$$

for the univariate polynomial interpolant to g at the sites $\lambda(Z)$, and realized that, with $D_y := \sum_j y_j D_j$ the directional derivative in the direction y,

$$D_{x-z_0} \cdots D_{x-z_{j-1}} (g \circ \lambda) = \lambda(x - z_0) \cdots \lambda(x - z_{j-1}) (D^j g) \circ \lambda,$$

hence, using Genocchi-Hermite, saw that

$$\begin{aligned} (Qg) \circ \lambda &= \sum_{j=0}^{k} \lambda(\cdot - z_0) \cdots \lambda(\cdot - z_{j-1}) \int_{[\lambda z_0, \dots, \lambda z_j]} D^j g \\ &= \sum_{j=0}^{k} \int_{[z_0,\dots,z_j]} D_{\cdot - z_0} \cdots D_{\cdot - z_{j-1}} (g \circ \lambda), \end{aligned}$$

and so knew that the ansatz

$$Pf = \sum_{j=0}^{k} \int_{[z_0,\dots,z_j]} D_{\cdot - z_0} \cdots D_{\cdot - z_{j-1}} f$$

for the Kergin projector was correct for all ridge functions (given Kergin's result concerning interpolation to ridge functions), hence must be correct.

I remember the exact spot on the blackboard in the coffee room at MRC where Micchelli wrote this last formula down for me, and can still experience my astonishment and admiration. I had no inkling that this was coming, hence declined his gracious offer of making this a joint paper.

It turned out that P. Milman, who is acknowledged in [K2] for many helpful discussions, also had this formula, resulting in [MM].

5 The recurrence for multivariate B-splines

Shortly after Micchelli had left MRC that fall, I received from him the one-page letter shown in Figure 2, containing the sought-after recurrence relations for multivariate B-splines, a second occasion for me to be astonished. Micchelli had not made my mistake, of concentrating on the geometric definition of the multivariate B-spline, but had stuck with the setting in which Schoenberg first thought of these multivariate B-splines, namely as the representers of the "divided difference" functionals $f \mapsto \int_{[z_0,\dots,z_n]} f$ defined in (3).

The formula was first published in MRC TSR 1895 in November 1978, a preliminary version of [M1].

The paperclip shown in the upper left corner of Figure 2 holds a copy of an MRC memo, saying: "Diese schoene Formel schickte mir Charlie Micchelli kuerzlich. Ihr Carl de Boor". The memo accompanied a copy of Micchelli's letter which I mailed to Wolfgang Dahmen, knowing from my short visit to Bonn in August 1978 that he thought he was on the track to getting recurrence relations.

Dahmen's response was swift: in a missive dated 30oct78, he submitted to me directly for possible publication in SJNA the first version of [D5], containing a proof of the recurrence relations but based on what we now call multivariate truncated powers or cone splines since they can be thought of as shadows of high-dimensional polyhedral cones. A second version reached me 14nov78 which I promptly sent to Micchelli for refereeing, who was wondering how Dahmen could have found out so quickly about his formula. In January, Micchelli asked permission (granted, of course) to contact Dahmen directly during his visit to Germany in February, and this led to Dahmen's application (granted, of course) for a research fellowship at Micchelli's home institution, the mathematics department at IBM Watson Research Center in Yorktown Heights NY, and the rest is history. While Dahmen published various results on multivariate B-splines alone, including papers in conference proceedings [D1], [D3], [D4], the construction of spaces spanned by such B-splines and their approximation order [D6], requiring the determination of the polynomials con-

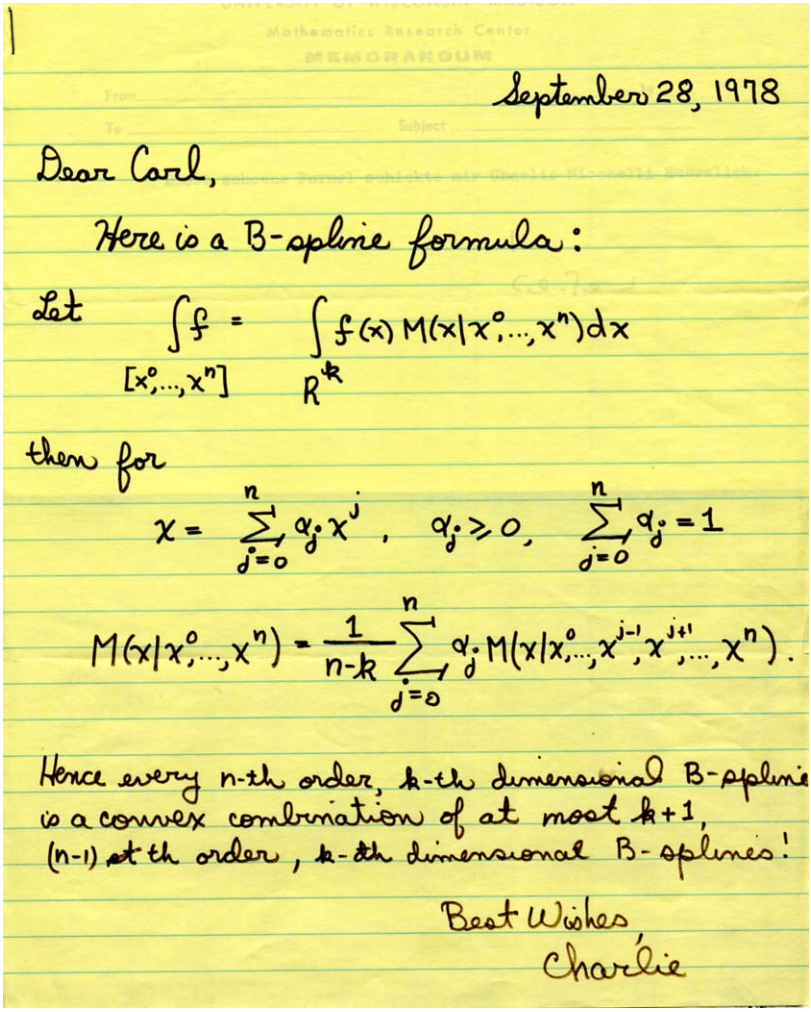


Fig. 2 Micchelli's recurrence relation for simplex splines

tained in such a span [D2], all leading up to his Habilitationsschrift [D7], his joint results with Micchelli on the mathematics of box splines were the pay-off of their joining forces in 1979. But, for that, the box splines had to make their appearance first.

6 Polyhedral splines

It must have happened during my visit with Ron DeVore at the University of South Carolina in April 1980 that he and I started a discussion on the relative merits in multivariate piecewise polynomial approximation of using total degree *vs.* coordinate degree whose outcome is [dBD]. The discussion was motivated by the fact that the approximation order achievable from a space A_h, of piecewise polynomials on a partition of mesh size h, is bounded by the maximum k for which $\Pi_{\le k}$ is contained in the approximation space A_h, and it seems that a tensor product spline space of coordinate degree k employs many more degrees of freedom (involving polynomial pieces of total degree $> k$) than seem necessary to have $\Pi_{\le k}$ contained in it.

To be sure, it is not sufficient to have $\Pi_{\le k} \subset A_h$ (see, e.g., [dBH3]); rather, $\Pi_{\le k}$ must be contained in A_h locally and stably, i.e., there must be a (local and stable) quasiinterpolant scheme with range A_h available that reproduces $\Pi_{\le k}$. It is this requirement that becomes increasingly hard and eventually, impossible if one increases the required smoothness of the approximating piecewise polynomials of order $\le k$ for a given partition or mesh. We only considered the bivariate case and considered only two partitions, a square mesh, and, in order to get some feeling for triangulations, the square mesh with all northeast diagonals drawn in (now called a 3-direction mesh or uniform type I triangulation). But how to get the smooth compactly supported piecewise polynomials needed? In the case of the 3-direction mesh, Courant's hat function offers itself for degree 1 and smoothness 0. But it was the sudden (and very pleasant) realization that this function is the 2-dimensional (skewed) shadow of a 3-cube that provided us with a recipe for the needed "bump functions" for the 3-direction mesh, as appropriate shadows of higher-dimensional cubes. We realized that other finite elements, e.g., the piecewise quadratic finite element constructed by Powell in [P], and, earlier, by Zwart in [Z], or certain elements discussed by Sablonnière, see [Sa], as well as those constructed by Sabin [S], could also be obtained as shadows of higher-dimensional cubes.

However, these new multivariate B-splines might not have been looked at carefully all that quickly but for the arrival at MRC, in the summer of 1980, of Klaus Höllig, for a 2-year postdoc. I had met Höllig the previous summer during an extended stay with Karl Scherer at the University of Bonn (during which Ron DeVore and I worked successfully in a local "Weinstube" on a problem of mixed-norm n-width that had arisen in Höllig's thesis work; see [dBDH]). Höllig produced in short order the two papers [H1], [H2], rederiving Micchelli's (and Dahmen's) results via Fourier transforms, and proposing a particular way of choosing a collection of simplices so that their shadows span a linear space of piecewise polynomials of order k with approximation order k. But, more than that, Höllig was swift to follow up on the suggestion that Micchelli's recurrence might be a simple consequence of Stokes' theorem, hence there is a version for shadows of cubes and, more generally, for shadows of convex polyhedra, as follows.

In the spirit of Micchelli's view of Schoenberg's multivariate B-spline, for a convex body B in $\mathbb{R}^n$ and a linear map P from $\mathbb{R}^n$ to $\mathbb{R}^d$, define the corresponding distribution M_B on $\mathbb{R}^d$ by

$$M_B\varphi := \int_B \varphi \circ P, \qquad \text{all test functions } \varphi,$$

with $\int_K$ the k-dimensional integral over the convex K in case the flat $\flat(K)$ spanned by K is k-dimensional. Assuming that $\flat(P(B)) = \mathbb{R}^d$, M_b is a nonnegative piecewise polynomial function, with $P(B)$ its support. Moreover, at each corner of its support, it agrees with one of Dahmen's truncated powers.

Assume that the boundary of B is the essentially disjoint union of finitely many $(n-1)$-dimensional convex bodies B_i. Then

$$D_{Pz}M_B = -\sum_i \langle z, n_i\rangle M_{B_i}, \qquad z \in \mathbb{R}^n, \tag{4}$$

$$(n-d)M_B(Pz) = \sum_i \langle b_i - z, n_i\rangle M_{B_i}(Pz), \qquad z \in \mathbb{R}^n, \tag{5}$$

with n_i the outside normal to $\flat(B_i)$, $\langle x, y\rangle$ the scalar product of x with y, and b_i a point in B_i, hence $\langle b_i - z, n_i\rangle$ is the signed distance of z from $\flat(B_i)$. The pointwise equality has to be taken in the sense of distributions. The proof of (5) in [dBH1] follows Hakopian's proof of (5) in [Ha] for the special case that B is a simplex. Under the assumption that B is a convex polytope, repeated application of (4) establishes that M_B is piecewise polynomial of degree at most $n-d$, and in $C^{n-\rho-2}$, with ρ the greatest integer with the property that a ρ-dimensional face of B is mapped by P into a $(d-1)$-dimensional set.

In [dBH2], we called M_B a polyhedral spline. Schoenberg's B-spline became a simplex spline, Dahmen's truncated power a cone spline, and the one introduced in [dBD] a box spline (though Micchelli prefers "cube spline"). These three examples seem, at present, the only ones carefully studied, probably because their polyhedra are the only ones whose facets are polyhedra of the same type.

7 Box splines

In contrast to the simplex splines, the construction of a collection of box splines spanning a useful space of piecewise polynomials is quite simple. If the box spline in question is

$$M := M_\Xi : \varphi \mapsto \int_{[0..1)^\Xi} \varphi(\Xi x)\,\mathrm{d}x$$

for some multiset or matrix Ξ of full rank of integer-valued nontrivial directions in $\mathbb{R}^d$, then

$$S(\Xi) := S_{M_\Xi} := \operatorname{span}(M_\Xi(\cdot - j) : j \in \mathbb{Z}^d)$$

is a cardinal, i.e., shift-invariant, spline space which contains all polynomials of (total) degree k where k is maximal with respect to the property that, for any subset Z of Ξ, $\Xi\backslash$Z is still of full rank. The full space of polynomials contained in $S(\Xi)$ is, in general, larger; it is denoted by $D(\Xi)$; it is the joint kernel of the differen-

tial operators $D_{\mathrm{H}} := \prod_{\eta\in \mathrm{H}} D_\eta$ where H ranges over the set $\mathscr{A}(\Xi)$ of all $\mathrm{H}\subset\Xi$ that intersect every basis in Ξ. In this connection, for any $\mathrm{Z}\subset\Xi$, $D_{\mathrm{Z}} M_\Xi = \nabla_{\mathrm{Z}} M_{\Xi\setminus \mathrm{Z}}$ and, in particular, $D_\Xi M_\Xi = \nabla_\Xi \delta$, with $\delta : \varphi \mapsto \varphi(0)$. Also, $M_\Xi * M_{\mathrm{Z}} = M_{\Xi\cup \mathrm{Z}}$. However, linear independence of $(M(\cdot - j) : j\in\mathbb{Z}^d)$ cannot hold unless Ξ is "unimodular", i.e., $|\det \mathrm{Z}| = 1$ for all bases $\mathrm{Z}\subset\Xi$. Nevertheless, even when there is no linear independence, one can construct, for $k_0 \le k$, a quasi-interpolant scheme Q into $S(\Xi)$ whose dilation $Q_h : f \mapsto Qf(\cdot/h)(\cdot h)$ provides approximation of order h^{k_0} for every smooth enough f. It is also clear that Schoenberg's theory of univariate cardinal spline interpolation (see, e.g., [Sc2]) can be extended to multivariate box spline interpolation in case of linear independence of $(M(\cdot - j) : j\in\mathbb{Z}^d)$ (a beginning is made in [dBHR]), and that the Strang-Fix theory [FS] of the approximation order of spaces spanned by the shifts of one function is applicable here.

While Höllig and I derived such basic results, eventually published in [dBH2], Dahmen and Micchelli pursued, unknown to us, vigorously much bigger game. We first learned details of their remarkable results from their survey [DM3] in the proceedings of the January 1983 Texas conference and from their summary [DM2] submitted in August 1983, with the former the only reference in the latter, and from reading [DM1], [DM4] and [DM5] for the details of some of the results announced in [DM2].

Not only did they prove that $(M_\Xi(\cdot - j) : j\in\mathbb{Z}^d)$ is (globally or locally) linearly independent iff Ξ is unimodular (something proved independently by Jia [J1], [J2]), they showed that the volume of the support of M_Ξ equals the number of $j\in\mathbb{Z}^d$ for which the support of $M_\Xi(\cdot - j)$ has a nontrivial intersection with the support of M_Ξ, and showed that support to be the essentially disjoint union of $\tau_{\mathrm{Z}} + \mathrm{Z}[0\,..\,1]^{\mathrm{Z}}$ for suitable τ_{Z} as Z runs over the set $\mathscr{B}(\Xi)$ of bases in Ξ, hence $\mathrm{vol}_d(M_\Xi[0\,..\,1]^\Xi) = \sum_{\mathrm{Z}\in\mathscr{B}(\Xi)} |\det \mathrm{Z}|$. They also completely characterized the space $E(\Xi)$ of linear dependence relations for $(M_\Xi(\cdot - j) : j\in\mathbb{Z}^d)$, i.e., the kernel of the linear map $M_\Xi^* : \mathbb{C}^{\mathbb{Z}^d} \to S(\Xi) : c \mapsto \sum_j M_\Xi(\cdot - j)c(j)$ (with the sum well-defined pointwise), and showed the space of polynomials in $S(\Xi)$, i.e., the joint kernel $D(\Xi)$ of the differential operators D_{H}, $\mathrm{H}\in\mathscr{A}(\Xi)$, to have dimension equal to $\#\mathscr{B}$. Remarkably, this last assertion holds even without the restriction that Ξ be an integer matrix.

But there is more. Recall the truncated power $T_\Xi : \varphi \mapsto \int_{\mathbb{R}_+^\Xi} \varphi(\Xi x)\,\mathrm{d}x$ introduced by Dahmen in [D5] for the case that $0\notin \mathrm{conv}(\Xi)$, i.e., the shadow of a cone. Already in [DM2], Dahmen and Micchelli define, under the assumption that $0\notin\mathrm{conv}(\Xi)$, the discrete truncated power $t(\cdot|\Xi)$ associated with Ξ as the map on $\mathbb{Z}^d$ for which

$$\sum_{\alpha\in\mathbb{Z}_+^\Xi} \varphi(\Xi\alpha) =: \sum_{j\in\mathbb{Z}^d} t(j|\Xi)\varphi(j)$$

for any finitely supported φ, hence $t(j|\Xi) = \#\{\alpha\in\mathbb{Z}^\Xi : \Xi\alpha = j\}$. In other words, $t(\cdot|\Xi)$ counts the number of nonnegative integer solutions for the linear system $\Xi ? = j$ with integer coefficients. They prove $T_\Xi = \sum_{j\in\mathbb{Z}^d} t(j|\Xi) M_\Xi(\cdot - j)$, and so obtain the remarkable formula $\nabla_\Xi T_\Xi = M_\Xi$. Their subsequent study of the discrete

truncated power enabled them, as reported in [DM6], to reprove certain conjectures concerning magic squares, thus opening up a surprising application of box spline theory.

On the other hand, box splines have had some difficulty in being accepted in areas of potential applications. A particularly striking example is Rong-Qing Jia's beautiful paper [J3] which contains a carefully crafted account of the relevant parts of the theory used in his proof of a long-outstanding conjecture of Stanley's concerning the number of symmetric magic squares. Referees from Combinatorics seemed unwilling to believe that such conjectures could be successfully tackled with spline theory.

In good part because of these (and other) results of Dahmen and Micchelli, there was a great outflow of work on box splines in the 80s, and it was hard to keep up with it. For this reason, Höllig, Riemenschneider and I decided to try to tell the whole story in a cohesive manner, resulting in [dBHR2].

I now wish we had included in the book the exponential box splines of Amos Ron [R] (followed closely by [DM7]). For, as Amos Ron has pointed out to me since (and is made clear in [BR]), the (polynomial) box splines can be understood as a limiting situation of the much simpler setup of exponential box spline. Here is an example.

Recall the Dahmen-Micchelli result that the dimension of the space $D(\Xi)$ of polynomials in the span $S(\Xi)$ of the shifts of the box spline M_Ξ equals the number $\#\mathscr{B}(\Xi)$ of bases in Ξ (provided Ξ is of full rank). This is (II.32)Theorem in the book, and its proof (a version of the Dahmen-Micchelli proof) is inductive and takes about three pages, with the main issue the claim that $\dim D(\Xi) \geq \#\mathscr{B}(\Xi)$. However, this inequality is almost immediate along the following lines suggested by Amos Ron: Choose, as we may, $\lambda : \Xi \to \mathbb{R}$ so that $(p_\xi : x \mapsto \langle x, \xi\rangle - \lambda(\xi))$ is generic, meaning that the unique common zero, v_B say, of $(p_\xi : \xi \in B)$ is different for different $B \in \mathscr{B}(\Xi)$. Consider $\mathrm{H} \in \mathscr{A}(\Xi)$. Since H intersects each $B \in \mathscr{B}(\Xi)$, the polynomial $p_\mathrm{H} := \prod_{\eta\in\mathrm{H}} p_\eta$ vanishes on $V := \{v_B : B \in \mathscr{B}(\Xi)\}$. Let $\mathrm{e}_v : x \mapsto \exp(\langle v, x\rangle)$. Then, for arbitrary $y \in \mathbb{R}^d$, $D_y \mathrm{e}_v = \langle v, y\rangle \mathrm{e}_v$, hence, for $p \in \Pi$, $p(D)\mathrm{e}_v = p(v)\mathrm{e}_v$. In particular, $p_\mathrm{H}(D)\mathrm{e}_v = 0$ for $v \in V$, hence $p_\mathrm{H}(D)f = 0$ for arbitrary $f = \sum_\alpha \widehat{f}(\alpha)()^\alpha \in \mathrm{Exp}(V) := \mathrm{span}\{\mathrm{e}_v : v \in V\}$. But $p(D)f = 0$ implies $p_\uparrow(D)f_\downarrow = 0$, with $p_\uparrow$ the "leading term" of p, i.e., the homogeneous polynomial for which $\deg(p - p_\uparrow) < \deg p$ and, correspondingly, $f_\downarrow$ the "least term" of f, i.e., the homogeneous polynomial for which $\mathrm{ord}(f - f_\downarrow) > \mathrm{ord} f := \min\{|\alpha| : \widehat{f}(\alpha) \neq 0\}$. Since $(p_\mathrm{H})_\uparrow(D) = D_\mathrm{H}$ and H was an arbitrary element of $\mathscr{A}(\Xi)$, it follows that $D(\Xi) = \cap_{\mathrm{H}\in\mathscr{A}(\Xi)} \ker D_\mathrm{H} \supset \mathrm{Exp}(V)_\downarrow := \mathrm{span}\{f_\downarrow : f \in \mathrm{Exp}(V)\}$. However, $\mathrm{Exp}(V)_\downarrow$ has dimension $\geq \#V = \#\mathscr{B}(\Xi)$, since $(\delta_v : v \in V)$ is linearly independent on $\mathrm{Exp}(V)_\downarrow$. Indeed, for any $v \in \mathbb{R}^d$ and $p \in \Pi$, $p(v) = (p(D)\mathrm{e}_v)(0)$, hence if $\sum_{v\in V} c(v)\delta_v = 0$ on $\mathrm{Exp}(V)_\downarrow$ yet $(c(v) : v \in V) \neq 0$, then $f := \sum_{v\in V} c(v)\mathrm{e}_v \neq 0$ and so $0 = f_\downarrow(D)f = \sum_{|\alpha|=\mathrm{ord} f} \widehat{f}(\alpha)^2\alpha! \neq 0$ which is nonsense.

8 Smooth multivariate piecewise polynomials and the B-net

I had given up quite early on the study of the space of all piecewise polynomials of a given order and smoothness on a given partition in more than one variable, preferring instead the finite element method approach of seeking suitable spaces of smooth piecewise polynomials spanned by bump functions. This was surely quite narrow-minded of me as, starting in the 70's, a very large, interesting and often challenging literature developed whose results are very well reported in the recent comprehensive book [LS] by Ming-Jun Lai and Larry Schumaker.

However, in the early 80's, Peter Alfeld, as a visitor at MRC, introduced me to the wonderful tool of what is now called the B-form. In this representation, the elements of the space $S_k^{(\rho)}(\Delta)$ of piecewise polynomials of degree $\le k$ on the given triangulation Δ and in $C^{(\rho)}$ are represented, on each triangle $\tau = \mathrm{conv}(V)$ in Δ, in the form

$$p = \sum_{|\alpha|=k} c(\alpha)\binom{|\alpha|}{\alpha}\ell^\alpha, \tag{6}$$

with $\alpha = (\alpha(v) : v \in V) \in \mathbb{Z}_+^V$, $\binom{|\alpha|}{\alpha} := |\alpha|!/\prod_v \alpha(v)!$, $\ell^\alpha := \prod_{v\in V}(\ell_v)^{\alpha(v)}$, and with $\ell_v := \ell_{v,\tau}$ the affine polynomial that vanishes on $V\backslash v$ and takes the value 1 at v, i.e., the ℓ_v are the Lagrange polynomials for linear interpolation to data given at V, hence $(\ell_{v,\tau}(x) : v \in V)$ are the socalled barycentric coordinates of x with respect to the vertex set V of τ. Further, it turns out to be very helpful to associate the coefficient $c(\alpha) = c(\alpha,\tau)$ with the "domain point"

$$\xi_{\alpha,\tau} := \sum_{v\in V} \alpha(v)v/k$$

(which happens to be the location of the unique maximum of ℓ_v^α (in τ)). For example, $v \in V$ is a domain point, namely $\xi_{k\delta_v,\tau}$, with δ_v the vector whose only nonzero entry is a 1 in position v, and all ℓ_w with $w \ne v$ vanish at that point, hence the corresponding coefficient, $c(k\delta_v)$, equals $p(v)$. More generally, on the edge of τ not containing v, i.e., on the zero set of ℓ_v, the only terms in (6) not obviously zero are those with $\alpha(v) = 0$, i.e., whose domain point lies on that edge. Hence continuity across that edge of a piecewise polynomial function is guaranteed by having the B-form coefficients of the two polynomial pieces abutting along that edge agree in the sense that coefficients associated with the same domain point coincide. Tbis sets up a 1-1 linear correspondence between the elements of $S_k^{(0)}(\Delta)$ and their "B-net", i.e., the scalar-valued map $\xi_{\alpha,\tau} \mapsto c(\alpha,\tau)$ on $\{\xi_{\alpha,\tau} : \alpha \in \mathbb{Z}_+^V, |\alpha| = k; \tau \in \Delta\}$.

Further, for any vector y, $D_y\ell_v = \ell_{v\uparrow}(y)$, with $\ell_{v\uparrow}$ the homogeneous linear part of the affine map ℓ_v, hence

$$D_y \sum_{|\alpha|=k} c(\alpha)\binom{|\alpha|}{\alpha}\ell^\alpha = k \sum_{|\beta|=k-1} \left(\sum_{v\in V} c(\beta+\delta_v)\ell_{v\uparrow}(y) \right) \binom{|\beta|}{\beta}\ell^\beta. \tag{7}$$

Hence, as Gerald Farin, in [Fa], was the first to stress, C^1-continuity across the edge of τ not containing v is guaranteed by the equalities

$$\sum_{w\in V} c(\beta+\delta_w)\ell_{w,\tau\uparrow}(y) = \sum_{w\in V'} c(\beta+\delta_w)\ell_{w,\sigma\uparrow}(y), \qquad |\beta| = k-1, \beta \in \mathbb{Z}_+^{V\cap V'},$$

with V' the vertex set of the triangle σ sharing that edge with τ. Note that the coefficients in these homogeneous equations are independent of the index β.

It is clear how ρ-fold iteration of this process produces the homogeneous linear equations that the B-net coefficients of an element of $S_k^{(0)}(\Delta)$ must satisfy for C^ρ continuity across the edge of τ not containing v. Each such equation involves the "quadrilateral" of coefficients $c(\beta+\gamma)$ and $c(\beta+\gamma')$, with $\beta \in \mathbb{Z}_+^{V\cap V'}$, $|\beta| = k-\rho$, and, $\gamma \in \mathbb{Z}_+^V$, $\gamma' \in \mathbb{Z}_+^{V'}$, $|\gamma| = \rho = |\gamma'|$.

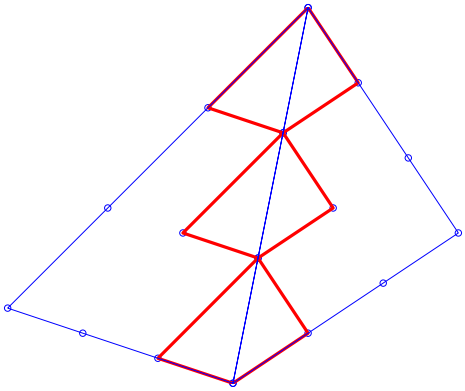


Fig. 3 C^1-conditions across an edge in the cubic case

In Figure 3, the situation is illustrated for the cubic case, $k = 3$. It shows the relevant domain points in the two triangles τ and σ sharing an edge, as well as the quadruples of domain points whose corresponding B-net coefficients must satisfy the *same* homogeneous linear equation for C^1-continuity across that edge.

This figure makes it immediate why the question of the dimension and approximation order of the space of bivariate C^1-cubics on a given triangulation might be difficult: there is only one domain point in the interior of each triangle, and its coefficient is involved in three homogeneous equations. Hence, the determination of an element of $S_3^{(1)}(\Delta)$ involves a global linear system. Correspondingly, it is not even clear whether there is an element of $S_3^{(1)}(\Delta)$ with prescribed values at the vertices of all the triangles, i.e., with the B-net coefficients corresponding to the vertices prescribed.

On the other hand, it has been known for some time that there is a local quasi-interpolant onto $S_5^{(1)}(\Delta)$ reproducing $\Pi_{\le 5}$ for any triangulation Δ (though its stability will depend on the smallest angle in the triangulation). Checking the geometry of

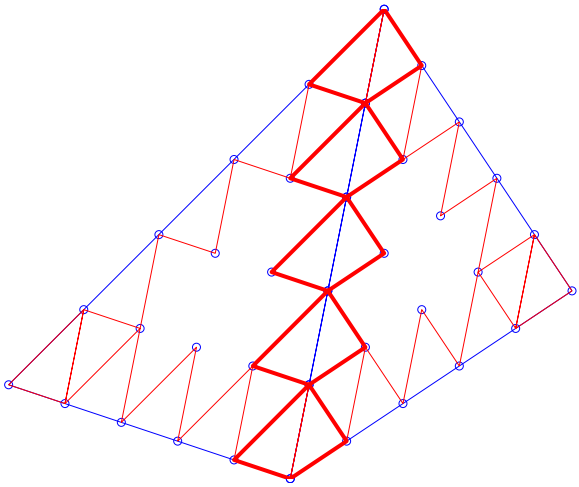


Fig. 4 C^1-conditions across an edge in the quintic case

the smoothness conditions, one realizes (see Figure 4) that 5 is the smallest value of k for which there is on each edge a "free" C^1-smoothness condition, i.e., one not touching a smoothness condition for any other edge. This led to the guess that, in the general case, $S_k^{(\rho)}(\Delta)$ has a local quasi-interpolant reproducing $\Pi_{\leq k}$ if there is a "free" C^ρ-smoothness condition on each edge, i.e., one not belonging to the "ring" of C^ρ-smoothness conditions associated with some vertex v by virtue of the fact that its edge and the edge of a smoothness condition it touches both contain v. For, one could hope to use such "free" conditions to "disentangle" or separate neighboring vertex rings. If ξ_α is the apex of such a "free" $C^{(\rho)}$-condition, it would have $\alpha(v) = \rho$ for some v, and would have $\alpha(w) > \rho$ for all $w \in V \backslash v$, hence $k = |\alpha| \geq 3\rho + 2$. For that case, [dBH6] contains a "proof" that, for a triangulation in which all the angles are bounded below by a constant, the approximation order is full, i.e., of the order h^{k+1}, where h is the mesh size. Unfortunately, the "proof" fails to take into account the possibility that the quadrangles corresponding to smoothness conditions across an edge can become nearly, but not exactly, flat which spoils a certain estimate on which the "proof" relies. This is explained in more detail in [dB6] which also contains a detailed account of the construction of a local basis for such spaces. A satisfactory proof of the main claim of [dBH6] was first given in [CHJ].

The above description of B-form and B-net readily applies to d dimensions (with the role of triangles played by d-simplices and the role of edges played by faces). However, in d dimension, existence of "free" $C^{(\rho)}$-smoothness conditions requires $k \geq (d+1)\rho + d$ for a generic partition into simplices. In particular, already for $d = 3$ one would need $k \geq 7$ for $C^{(1)}$, which discouraged me from pursuing the study of all smooth piecewise polynomials on a "triangulation" in higher dimension.

Another result using B-nets in an essential way was the discovery in [dBH3] that, even on a certain regular triangulation, namely the 3-direction mesh Δ_3, $S_3^{(1)}(\Delta_3)$

does not have full approximation order, even though the space contains $\Pi_{\leq 3}$ locally. This has been reproved in more generality and with very different methods in [dBDR].

Altogether, the appearance of the B-net revolutionized the analysis of smooth piecewise polynomials even (and particularly) in the bivariate case, as is illustrated by its prominence in [LS].

References

[BR] Ben-Artzi, A., Ron, A.: Translates of exponential box splines and their related spaces. Trans. Amer. Math. Soc. 309, 683–710 (1988)

[B] Birkhoff, G.: Local spline approximation by moments. J. Math. Mech. 16, 987–990 (1967)

[BdB] Birkhoff, G., Boor, C. R. de: Piecewise polynomial interpolation and approximation. In: L. Garabedian (ed.) Approximation of Functions, pp. 164–190. Elsevier New York (1965)

[BG] Birkhoff, G., Garabedian, H.: Smooth surface interpolation. J. Math. Phys. 39, 258–268 (1960)

[dB0] Boor, C. de: Bicubic spline interpolation. J. Math. Phys. 41, 212–218 (1962)

[dB1] Boor, C. de: On local spline approximation by moments. J. Math. Mech. 17, 729–735 (1968)

[dB2] Boor, C. de: On uniform approximation by splines. J. Approx. Theory 1, 219–235 (1968)

[dB3] Boor, C. de: Splines as linear combinations of *B*–splines, a survey. In: G.G. Lorentz, K. Chui, and L.L. Schumaker (eds.) Approximation Theory, II, pp. 1–47. Academic Press New York (1976)

[dB4] Boor, C. de: Efficient computer manipulation of tensor products. ACM Trans. Math. Software 5, 173–182. *Corrigenda:* 525 (1979)

[dB5] Boor, C. de: The polynomials in the linear span of integer translates of a compactly supported function. Constr. Approx. 3, 199–208 (1987)

[dB6] Boor, C. de: A local basis for certain smooth bivariate pp spaces. In: C. Chui, W. Schempp, and K. Zeller (eds.) Multivariate Approximation Theory IV, ISNM 90, pp. 25–30. Birkhäuser Verlag Basel (1989)

[dBD] Boor, C. de, DeVore, R.: Approximation by smooth multivariate splines. Trans. Amer. Math. Soc. 276(2), 775–788 (1983)

[dBDH] Boor, C. de, DeVore, R., Höllig, K.: Mixed norm *n*-widths. Proc. Amer. Math. Soc. 80, 577–583 (1980)

[dBDR] Boor, C. de, DeVore, R., Ron, A.: Approximation orders of FSI spaces in $L_2(\mathbb{R}^d)$. Constr. Approx. 14, 631–652 (1998)

[dBF] Boor, C. de, Fix, G.J.: Spline approximation by quasi-interpolants. J. Approx. Theory 8, 19–45 (1973)

[dBH1] Boor, C. de, Höllig, K.: Recurrence relations for multivariate B-splines. Proc. Amer. Math. Soc. 85, 397–400 (1982)

[dBH2] Boor, C. de, Höllig, K.: B-splines from parallelepipeds. J. Analyse Math. 42, 99–115 (1982/83)

[dBH3] Boor, C. de, Höllig, K.: Approximation order from bivariate C^1-cubics: a counterexample. Proc. Amer. Math. Soc. 87, 649–655 (1983)

[dBH4] Boor, C. de, Höllig, K.: Bivariate box splines and smooth pp functions on a three direction mesh. J. Comput. Appl. Math. 9, 13–28 (1983)

[dBH5] Boor, C. de, Höllig, K.: Minimal support for bivariate splines. Approx. Theory Appl. 3, 11–23 (1987)

[dBH6] Boor, C. de, Höllig, K.: Approximation power of smooth bivariate pp functions. Math. Z. 197, 343–363 (1988)

[dBHR] Boor, de C., Höllig, K., Riemenschneider, S.: Bivariate cardinal interpolation. In: C. Chui, L. Schumaker, and J. Ward (eds.) Approximation Theory IV, pp. 359–363. Academic Press New York (1983)

[dBHR2] Boor, C. de, Höllig, K., Riemenschneider, S.D.: Box Splines. Springer-Verlag, New York (1993)

[CD] Chui, C.K., Diamond, H.: A natural formulation of quasi-interpolation by multivariate splines. Proc. Amer. Math. Soc. 99, 643–646 (1987)

[CHJ] Chui, C.K., Hong, Dong, Jia, Rong-Qing: Stability of optimal-order approximation by bivariate splines over arbitrary triangulations. Trans. Amer. Math. Soc. 347(9), 3301–3318 (1995)

[CJW] Chui, Charles K., Jetter, K., Ward, J.D.: Cardinal interpolation by multivariate splines. Math. Comp. 48(178), 711–724 (1987)

[CL] Chui, C.K., Lai, M.-J.: A multivariate analog of Marsden's identity and a quasi-interpolation scheme. Constr. Approx. 3, 111–122 (1987)

[C] Coons, S.A.: Surfaces for computer-aided design of space forms. TR, Project MAC, Design Div., Mech. Engin. Dep., M.I.T. (1964)

[CS] Curry, H.B., Schoenberg, I.J.: On Pólya frequency functions IV: the fundamental spline functions and their limits. J. Analyse Math. 17, 71–107 (1966)

[Dav] Davis, Philip J.: Triangle formulas in the complex plane. Math. Comp. 18, 569–577 (1964)

[D1] Dahmen, W.: Multivariate B-splines—Recurrence relations and linear combinations of truncated powers. In: W. Schempp and K. Zeller (eds.) Multivariate Approximation Theory, pp. 64–82. Birkhäuser Basel (1979)

[D2] Dahmen, W.: Polynomials as linear combinations of multivariate B-splines. Math. Z. 169, 93–98 (1979)

[D3] Dahmen, W.: Konstruktion mehrdimensionaler B-Splines und ihre Anwendung of Approximationsprobleme. In: L. Collatz, G. Meinardus, and H. Werner (eds.) Numerical Methods of Approximation Theory Vol. 5, pp. 84–110. Birk häuser Verlag Basel (1980)

[D4] Dahmen, W.: Approximations by smooth multivariate splines on non-uniform grids. In: R. DeVore and K. Scherer (eds.) Quantitative Approximation, pp. 99–114. Academic Press New York (1980)

[D5] Dahmen, W.: On multivariate B-splines. SIAM J. Numer. Anal. 17, 179–191 (1980)

[D6] Dahmen, W.: Approximation by linear combinations of multivariate B-splines. J. Approx. Theory 31, 299–324 (1981)

[D7] Dahmen, W.: Multivariate B-splines, ein neuer Ansatz im Rahmen der konstruktiven mehrdimensionalen Approximationstheorie. Habilitation, Bonn (1981)

[DDS] Dahmen, W., DeVore, R., Scherer, K.: Multidimensional spline approximations. SIAM J. Numer. Anal. 17, 380–402 (1980)

[DM1] Dahmen, W., Micchelli, C.A.: Translates of multivariate splines. Linear Algebra Appl. 52, 217–234 (1983)

[DM2] Dahmen, W., Micchelli, C.A.: Recent progress in multivariate splines. In: C. Chui, L. Schumaker, and J. Ward (eds.) Approximation Theory IV, pp. 27–121. Academic Press New York (1983)

[DM3] Dahmen, W., Micchelli, C.A.: Some results on box splines. Bull. Amer. Math. Soc. 11, 147–150 (1984)

[DM4] Dahmen, W., Micchelli, C.A.: On the solution of certain systems of partial difference equations and linear independence of translates of box splines. Trans. Amer. Math. Soc. 292, 305–320 (1985)

[DM5] Dahmen, W., Micchelli, C.A.: On the local linear independence of translates of a box spline. Studia Math. 82, 243–263 (1985)

[DM6] Dahmen, W., Micchelli, C.A.: The number of solutions to linear Diophantine equations and multivariate splines. TAMS 308, 509–532 (1988)

[DM7] Dahmen, W., Micchelli, Charles A.: On multivariate E-splines. Advances in Math. 76, 33–93 (1989)
[DM8] Dahmen, W., Micchelli, C.: Local dimension of piecewise polynomial spaces, syzygies, and solutions of systems of partial differential equations. Mathem. Nachr. 148, 117–136 (1990)
[DMS] Dahmen, Wolfgang, Micchelli, Charles A., Seidel, Hans-Peter: Blossoming begets B-spline bases built better by B-patches. Math. Comp. 59(199), 97–115 (1992)
[Fa] Farin, G.: Subsplines über Dreiecken. PhD thesis, Braunschweig (1979)
[Fe] Ferguson, J.C.: Multivariable curve interpolation. J. Assoc. Comput. Mach. II, 221–228 (1964)
[FS] Fix, G., Strang, G.: Fourier analysis of the finite element method in Ritz-Galerkin theory. Studies in Appl. Math. 48, 265–273 (1969)
[Fr1] Frederickson, P.O.: Quasi-interpolation, extrapolation, and approximation on the plane. In: R.S.D. Thomas and H.C. Williams (eds.) Proc. Manitoba Conf. Numer. Math., pp. 159–167. Utilitas Mathematica Publishing Inc. Winnipeg (1971)
[Fr2] Frederickson, P.O.: Generalized triangular splines. Lakehead Rpt. 7 (1971)
[G] Gordon, W.J.: Spline-blended surface interpolation through curve networks. J. Math. Mech. 18, 931–952 (1969)
[Ha] Hakopian, Hakop: Multivariate spline functions, B-spline basis and polynomial interpolations. SIAM J. Numer. Anal. 19(3), 510–517 (1982)
[H1] Höllig, K.: A remark on multivariate B-splines. J. Approx. Theory 33, 119–125 (1982)
[H2] Höllig, Klaus: Multivariate splines. SIAM J. Numer. Anal. 19, 1013–1031 (1982)
[J1] Jia, R.Q.: Linear independence of translates of a box spline. J. Approx. Theory 40, 158–160 (1984)
[J2] Jia, R.Q.: Local linear independence of the translates of a box spline. Constr. Approx. 1, 175–182 (1985)
[J3] Jia, Rong-Qing: Symmetric magic squares and multivariate splines. Linear Algebra Appl. 250, 69–103 (1997)
[K1] Kergin, P.: Interpolation of C^k Functions. PhD thesis, University of Toronto, Canada (1978)
[K2] Kergin, P.: A natural interpolation of C^k functions. J. Approx. Theory 29, 278–293 (1980)
[LS] Lai, Ming-Jun, Schumaker, Larry L.: Spline functions on triangulations. Cambridge U. Press, Cambridge, UK (2007)
[M1] Micchelli, C.A.: A constructive approach to Kergin interpolation in $\mathbb{R}^k$: multivariate B-splines and Lagrange interpolation. Rocky Mountain J. Math. 10, 485–497 (1980)
[M2] Micchelli, C.A.: Mathematical Aspects of Geometric Modeling. CBMS-NSF Reg. Conf. Appl. Math. 65, SIAM, Philadelphia (1995)
[MM] Micchelli, C.A., Milman, P.: A formula for Kergin interpolation in $\mathbb{R}^k$. J. Approx. Theory 29, 294–296 (1980)
[N] Neamtu, M.: Delaunay configurations and multivariate splines: A generalization of a result of B.N. Delaunay. Trans. Amer. Math. Soc. 359, 2993–3004 (2007)
[P] Powell, M.J.D.: Piecewise quadratic surface fitting for contour plotting. In: D.J. Evans (ed.) Software for Numerical Mathematics, pp. 253–271. Academic Press London (1974)
[R] Ron, A.: Exponential box splines. Constr. Approx. 4, 357–378 (1988)
[S] Sabin, M.A.: The use of piecewise forms for the numerical representation of shape. PhD thesis, MTA Budapest (1977)
[Sa] Sablonnière, P.: unpublished notes. parts of which eventually appeared in reports, e.g., in [Sa2] (1979)
[Sa2] Sablonnière, P.: De l'existence de spline á support borné sur une triangulation équilatérale du plan. Publication ANO-39, U.E.R. d'I.E.E.A.-Informatique, Université de Lille (1981)

[Sc] Schoenberg, I.J.: Letter to Philip J. Davis. 31 May (1965).
See http://pages.cs.wisc.edu/∼deboor/HAT/isolet.pdf
or [M2, pp. 201–203].

[Sc2] Schoenberg, I.J.: Cardinal Spline Interpolation. Vol. 12, CBMS, SIAM, Philadelphia (1973)

[St] Strang, G.: The dimension of piecewise polynomials, and one-sided approximation. In: G.A. Watson (ed.) Numerical Solution of Differential Equations, pp. 144–152. Springer Berlin (1974)

[TS] Theilheimer, Feodor, Starkweather, William: The fairing of ship lines on a high-speed computer. Math. Comp. 15(76), 338–355 (1961)

[Z] Zwart, P.B.: Multivariate splines with non-degenerate partitions. SIAM J. Numer. Anal. 10, 665–673 (1973)

On the efficient computation of high-dimensional integrals and the approximation by exponential sums

Dietrich Braess and Wolfgang Hackbusch

Abstract The approximation of the functions $1/x$ and $1/\sqrt{x}$ by exponential sums enables us to evaluate some high-dimensional integrals by products of one-dimensional integrals. The degree of approximation can be estimated via the study of rational approximation of the square root function. The latter has interesting connections with the Babylonian method and Gauss' arithmetic-geometric process.

Key words: exponential sums, rational functions, Chebyshev approximation, best approximation, completely monotone functions, Heron's algorithm, complete elliptic integrals, Landen transformation.

AMS Subject Classifications: 11L07, 41A20.

1 Introduction

The approximation of transcendental or other complicated functions by polynomials, rational functions, and spline functions is at the centre of classical approximation theory. In the last decade the numerical solution of partial differential equations gave rise to quite different problems in approximation theory. In this paper we will study the approximation of $x^{-\alpha}$ ($\alpha = 1/2$ or 1) by exponential sums. Here a simple function is approximated by a more complicated one, but it enables the fast com-

Dietrich Braess
Mathematisches Institut, Ruhr-Universität Bochum, 44780 Bochum, Germany,
e-mail: Dietrich.Braess@rub.de,
URL: http://homepage.ruhr-uni-bochum.de/Dietrich.Braess

Wolfgang Hackbusch
Max-Planck-Institut *Mathematik in den Naturwissenschaften*, Inselstr. 22, 04103 Leipzig, Germany, e-mail: wh@mis.mpg.de,
URL: http://www.mis.mpg.de/scicomp/hackbusch_e.html

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_3,

putation of some high-dimensional integrals which occur in quantum physics and quantum chemistry.

A model example is an integral of the form

$$\int \frac{g_1(x_1)\cdots g_d(x_d)}{\|x-y\|_0}\mathrm{d}x \tag{1}$$

on a domain in $\mathbb{R}^d$, where $\|\cdot\|_0$ refers to the Euclidean norm. When we insert the approximation

$$\frac{1}{\sqrt{x}} \approx \sum_{j=1}^{n} \alpha_j \mathrm{e}^{-t_j x},$$

then the integral is reduced to a sum of products of one-dimensional integrals

$$\sum_{j=1}^{n} \alpha_j \prod_{i=1}^{d} \int g_i(x_i)\exp[-t_j(x_i-y_i)^2]\,\mathrm{d}x_i,$$

and a fast computation is now possible (at least in the domain, where the approximation is valid, see Section 6.2 for more details). Other examples will be discussed in Sections 5 and 6.

There are only a few problems in nonlinear approximation theory for which the degree of approximation can be estimated. Surprisingly, the problem under consideration belongs to those rare cases. The functions $x^{-\alpha}$ ($\alpha > 0$) are monosplines for the kernel e^{-tx}. For this reason, results for the rational approximation of the square root function provide good estimates for the degree of approximation by exponential sums.

In principle, rational approximation of the square root function is well known for more than a century from Zolotarov's results. Elliptic integrals play a central role in his investigations. We find it more interesting, direct, and less technical to derive approximation properties from the Babylonian method for the computation of a square root. Gauss' arithmetic-geometric process yields a fast computation of the decay rate of the approximation error.

The rest of the paper is organised as follows. Section 2 is devoted to the connection of the approximation of $x^{-\alpha}$ by exponential sums with the rational approximation of $\sqrt{x}$. The investigation of the latter with the help of the Babylonian method and Gauss' arithmetic-geometric mean is the main purpose of Section 3. The results for the approximation of $1/x$ by exponential sums on finite and infinite intervals are presented in Section 4. Numerical results show that the theory yields the correct asymptotic law, while an improvement is possible for infinite intervals. The role of the approximation problem for the computation of high-dimensional integrals is elucidated with several examples in Sections 5 and 6. The numerical computation of the best exponential sums is discussed in Section 7. Appendix A provides auxiliary results for small intervals. Properties of complete elliptic integrals that are required for the derivation of the asymptotic rates, are derived in Appendix B. Finally it is shown in Appendix C that a competing tool yields the same law for infinite intervals.

2 Approximation of completely monotone functions by exponential sums

Good estimates for the degree of approximation are available only for a few classes of nonlinear approximation problems. Fortunately, the asymptotic behaviour is known for the functions in which we are interested. The functions $1/x$ and $1/\sqrt{x}$ are *completely monotone* for $x > 0$, i.e., they are Laplace transforms of non-negative measures:

$$f(x) = \int_0^\infty \mathrm{e}^{-tx} \mathrm{d}\mu(t), \qquad \mathrm{d}\mu \geq 0.$$

In particular,

$$\frac{1}{x} = \int_0^\infty \mathrm{e}^{-tx} \mathrm{d}t, \quad \frac{1}{\sqrt{x}} = \int_0^\infty \mathrm{e}^{-tx} \frac{\mathrm{d}t}{\sqrt{\pi t}}.$$

In order to avoid degeneracies, we assume that the support of the measure is an infinite set. We will also restrict ourselves to $\Re e\, x \geq 0$.

We consider best approximations in the sense of Chebyshev, i.e., the error is to be minimised with respect to the supremum norm on a given interval. A unique best exponential sum of order n,

$$u_n(x) = \sum_{\nu=1}^{n} \alpha_\nu \mathrm{e}^{-t_\nu x} \tag{2}$$

exists for a given completely monotone function f, while this is not true for arbitrary continuous functions. Moreover, the coefficients α_ν in the best approximation are non-negative (cf. [4]).

Our error estimates require the solution of a nonlinear interpolation problem that is also solvable for completely monotone functions.

Theorem 2.1. *Let f be completely monotone for $x > 0$ and $0 < x_1 < x_2 < \ldots < x_{2n}$. Then there exists an exponential sum u_n such that*

$$u_n(x_j) = f(x_j), \qquad j = 1, 2, \ldots, 2n.$$

Moreover

$$u_n(x) < f(x) \qquad \text{for } 0 < x < x_1 \text{ and } x > x_{2n}.$$

If in addition f is continuous at $x = 0$, also $u_n(0) < f(0)$ holds.

Sketch of proof. The complete monotonicity allows us to apply deformation arguments from nonlinear analysis. The best approximation u_n on the interval $[\frac{1}{2}x_1, x_{2n} + 1]$ has an alternant of length $2n+1$ (see Definition 3.1). Therefore, $f - u_n$ has $2n$ zeros $y_1 < y_2 < \ldots, y_{2n}$. Set

$$x_j(s) := (1-s)y_j + sx_j, \qquad 0 \leq s \leq 1, \quad j = 1, 2, \ldots, 2n.$$

The set of numbers $s \in [0, 1]$ for which the interpolation at the points $x_j(s)$ is solvable, contains the point $s = 0$. The rules on the zeros of extended exponential sums

$\sum_j(\alpha_j+\beta_j x)\mathrm{e}^{-t_j x}$ and the Newton method imply that the set is open. It follows from compactness properties of exponential sums that the set is also closed. Hence, the value $s=1$ is included.

The given function f and the approximating exponential sums are analytic functions in the right half plane of $\mathbb{C}$, and the complete monotonicity provides some a priori bounds. For this reason, we can derive error bounds for our approximation problem in the interval $[a,b]$ from the knowledge of a function with small values in $[a,b]$ and symmetry properties. The latter is provided by the rational approximation of the square root function and is related to other fast computations, as we will see in the next section.

The extra assumption in the following lemma concerning the continuity of f at $x=0$ will be no drawback, since a shift $x\mapsto x+1/n$ will recover it.

Lemma 2.1. *Let f be completely monotone for $x>0$ and continuous at $x=0$. Assume that p_n and q_{n-1} are polynomials of degree n and $n-1$, respectively, and that*

$$\left|\frac{p_n(x)}{q_{n-1}(x)}-\sqrt{x}\right|\le\varepsilon\sqrt{x}\qquad \text{for } x\in[a^2,b^2] \tag{3}$$

$$\text{or}\quad \left|\frac{p_n(x)/q_{n-1}(x)-\sqrt{x}}{p_n(x)/q_{n-1}(x)+\sqrt{x}}\right|\le\varepsilon\qquad \text{for } x\in[a^2,b^2], \tag{4}$$

holds for some $\varepsilon>0$. Assume also that $p_n/q_{n-1}-\sqrt{x}$ has $2n$ zeros in the interval $[a^2,b^2]$. Then there exists an exponential sum u_n with n terms such that

$$|f(x)-u_n(x)|\le 2\varepsilon f(0)\quad \text{for } x\in[a,b].$$

Proof. Put $x=z^2$. Obviously, we can restrict ourselves to the case $\varepsilon<1$. Now (3) implies (4) and by assumption

$$\left|\frac{p_n(z^2)-zq_{n-1}(z^2)}{p_n(z^2)+zq_{n-1}(z^2)}\right|\le\varepsilon\quad \text{for } z\in[a,b].$$

Set $P_{2n}(z):=p_n(z^2)-zq_{n-1}(z^2)$ and write

$$\left|\frac{P_{2n}(z)}{P_{2n}(-z)}\right|\le\varepsilon\quad \text{for } z\in[a,b]. \tag{5}$$

Obviously,

$$\left|\frac{P_{2n}(z)}{P_{2n}(-z)}\right|=1\quad \text{for } \Re e\, z=0\quad \text{or } |z|\to\infty.$$

Let u_n be the interpolant of f at the $2n$ zeros of P_{2n}. The last inequality in

$$|f(z)|\le f(0),\quad |u_n(z)|\le u_n(\Re e\, z)\le u_n(0)<f(0)\quad \text{for } \Re e\, z\ge 0$$

follows from Theorem 2.1. Hence,

$$\left| \frac{P_{2n}(-z)}{P_{2n}(z)} \Big(f(z) - u_n(z) \Big) \right| \leq 2f(0) \tag{6}$$

holds at the boundary of the right half-plane. By the maximum principle for analytic functions (6) holds also in the interior, and

$$|f(z) - u_n(z)| \leq 2f(0) \left| \frac{P_{2n}(z)}{P_{2n}(-z)} \right|$$

completes the proof.

A similar method is sketched in Appendix 10. The maximum principle is applied to an analytic function on a sector of the complex plane and with properties different from (5). The inequality (5) is related to the capacity of a capacitor with the plates $[a,b]$ and $[-b,-a]$. We are looking for a rational function, whose absolute value is small on $[a,b]$ and large on $[-b,-a]$.

Lemma 2.1 provides only upper bounds for the degree of approximation. Surprisingly, numerical results in Section 3 lead to the conjecture that the asymptotic behaviour and the exponential decay is precisely described for finite intervals. The estimates for infinite intervals reflect the asymptotic behaviour, but are not optimal, although they are sharper than the estimate obtained via Sinc approximation methods [7] and §11.

3 Rational approximation of the square root function

3.1 Heron's algorithm and Gauss' arithmetic-geometric mean

At the beginning of the second century, Heron of Alexandria described a method to calculate the square root of a given positive number a using some initial approximation. The method was probably also known to the Babylonians. A modification – more precisely a rescaling – will help us to construct best rational approximations of the square root function in the sense of Chebyshev [22].

Let x_n be an approximation of $\sqrt{a}$. Obviously $\sqrt{a}$ is the *geometric mean* of x_n and a/x_n. Heron replaced it by the *arithmetic mean*, i.e., in modern notation:

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}).$$

Convergence follows from the recursion relation for the error

$$x_{n+1} - \sqrt{a} = \frac{(x_n - \sqrt{a})^2}{2x_n}. \tag{7}$$

Gauss considered the two means in a different context. At an early age, he became enamoured of a sequential procedure that is now known as the arithmetic-geometric

process (see, e.g., [3]). Given two numbers $0 < a_0 < b_0$, one successively takes the arithmetic mean and the geometric mean:

$$a_{j+1} := \sqrt{a_j b_j}, \qquad b_{j+1} := \frac{1}{2}(a_j + b_j). \tag{8}$$

He expressed the common limit as an elliptic integral (see Appendix 9). The distance of the two numbers is characterised by $\lambda_j := b_j/a_j$. A direct calculation yields

$$\lambda_{j+1} = \frac{1}{2}\left(\sqrt{\lambda_j} + \frac{1}{\sqrt{\lambda_j}}\right) \quad \text{or} \quad \lambda_j = \left(\lambda_{j+1} + \sqrt{\lambda_{j+1}^2 - 1}\right)^2. \tag{9}$$

The mapping $\lambda \mapsto (\lambda + \sqrt{\lambda^2 - 1})^2$ is called the *Landen transformation*. The numbers in the table below show that a few steps forwards or backwards brings us either to large numbers or to numbers very close to 1. – Finally, we mention that the numbers $(\lambda+1)/(\lambda-1)$ and λ' with $(\lambda')^{-2} + \lambda^{-2} = 1$ are moved by the same rule, but in the opposite direction.

Table 1 Arithmetic-geometric process with $\lambda_0 = 1 + \sqrt{2}$ and $\lambda_j^{-2} + (\lambda_j')^{-2} = 1$

j	λ_j	$\frac{\lambda_{j+1}}{\lambda_j - 1}$	λ_j'
-4	$6.825 \cdot 10^{14}$	$1 + 2.930 \cdot 10^{-15}$	$1 + 1.07 \cdot 10^{-30}$
-3	$1.306 \cdot 10^{7}$	$1 + 1.531 \cdot 10^{-7}$	$1 + 2.930 \cdot 10^{-15}$
-2	1807.08	1.001107	$1 + 1.531 \cdot 10^{-7}$
-1	21.26	1.099	1.001107
0	2.414	2.414	1.099
1	1.099	21.26	2.414
2	1.001107	1807.08	21.26
3	$1 + 1.531 \cdot 10^{-7}$	$1.306 \cdot 10^{7}$	1807.08
4	$1 + 2.930 \cdot 10^{-15}$	$6.825 \cdot 10^{14}$	$1.306 \cdot 10^{7}$

3.2 Heron's method and best rational approximation

In view of Lemma 2.1 we are interested in the best relative Chebyshev approximation of $\sqrt{x}$ by rational functions in $R_{n,n-1}$. Specifically, v_n is called a best approximation if it yields the solution of the minimisation problem:

$$E_{n,n-1} := E_{n,n-1,[a,b]} := \inf_{v_n \in R_{n,n-1}} \left\| \frac{v_n - \sqrt{x}}{\sqrt{x}} \right\|_{L_\infty[a,b]}.$$

Definition 3.1. An error curve $\eta(x)$ has an alternant of length ℓ, if there are ℓ points $x_1 < x_2 < \ldots < x_\ell$ such that

$$\operatorname{sign}\eta(x_{i+1}) = -\operatorname{sign}\eta(x_i) \quad \text{for } i = 1,2,\ldots\ell-1 \tag{10}$$

and

$$|\eta(x_i)| = \|\eta\|_{L_\infty} \quad \text{for } i = 1,2,\ldots\ell \tag{11}$$

holds.

The following characterisation goes back to Chebyshev. Some degeneracies that are possible with rational approximation, cannot occur here.

Theorem 3.1 (characterisation theorem). *v_n is optimal in $R_{n,n-1}$ if and only if the error curve $(v_n - \sqrt{x})/\sqrt{x}$ has an alternant of length $2n+1$.*

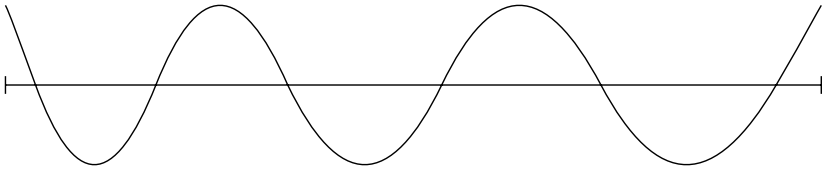


Fig. 1 Alternant of length 7

Let $p_n/q_{n-1} \in R_{n,n-1}$ be an approximation of $\sqrt{x}$. The application of one step of Heron's algorithm yields the rational function

$$\frac{1}{2}\left(\frac{p_n}{q_{n-1}} + \frac{x}{p_n/q_{n-1}}\right) = \frac{p_n^2 + xq_{n-1}^2}{2p_nq_{n-1}} \in R_{2n,2n-1}$$

From (7) we conclude that the associated error curve is non-negative and cannot be a best approximation; see Figure 2. A rescaling before and after the procedure, however, will yield a solution. This was already observed by Rutishauser [22], although he stopped at (12) and did not mention the connection with Gauss' arithmetic-geometric process.

Let v_n be the best approximation in $R_{n,n-1}$. By definition,

$$1 - E_{n,n-1} \le \frac{v_n(x)}{\sqrt{x}} \le 1 + E_{n,n-1}\,.$$

The corresponding relations for $w_n := \frac{1}{\sqrt{1-E_{n,n-1}^2}} v_n$ are

$$\sqrt{\frac{1-E_{n,n-1}}{1+E_{n,n-1}}} \le \frac{w_n(x)}{\sqrt{x}} \le \sqrt{\frac{1+E_{n,n-1}}{1-E_{n,n-1}}}\,.$$

The result of a Heron step is denoted by w_{2n} and

$$1 \le \frac{w_{2n}(x)}{\sqrt{x}} = \frac{1}{2}\left(\frac{w_n(x)}{\sqrt{x}} + \frac{\sqrt{x}}{w_n(x)}\right)$$
$$\le \frac{1}{2}\left(\sqrt{\frac{1+E_{n,n-1}}{1-E_{n,n-1}}} + \sqrt{\frac{1-E_{n,n-1}}{1+E_{n,n-1}}}\right) = \frac{1}{\sqrt{1-E_{n,n-1}^2}}.$$

We rescale the new rational function, set $v_{2n} := \frac{2\sqrt{1-E_{n,n-1}^2}}{1+\sqrt{1-E_{n,n-1}^2}} w_{2n}$, and obtain

$$\frac{2\sqrt{1-E_{n,n-1}^2}}{1+\sqrt{1-E_{n,n-1}^2}} \le \frac{v_{2n}(x)}{\sqrt{x}} \le \frac{2}{1+\sqrt{1-E_{n,n-1}^2}}.$$

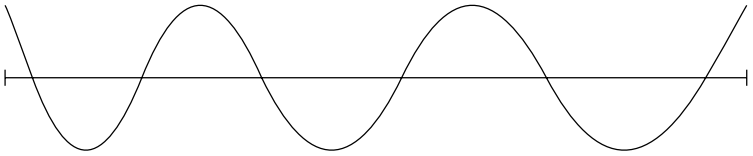


Fig. 2 Error curves and Heron's procedure

Figure 2 elucidates that the number of sign changes is doubled, and the equilibration above yields the desired alternant of length $4n+1$. Hence,

$$E_{2n,2n-1} = \frac{2}{1+\sqrt{1-E_{n,n-1}^2}} - 1 = \frac{1-\sqrt{1-E_{n,n-1}^2}}{1+\sqrt{1-E_{n,n-1}^2}} = \frac{E_{n,n-1}^2}{\left(1+\sqrt{1-E_{n,n-1}^2}\right)^2} \quad (12)$$

or

$$E_{2n,2n-1}^{-1} = \left(E_{n,n-1}^{-1} + \sqrt{E_{n,n-1}^{-2}-1}\right)^2. \quad (13)$$

Remark 3.1. The inverse $E_{2n,2n-1}^{-1}$ is obtained from $E_{n,n-1}^{-1}$ by the Landen transformation. In particular,

$$(\frac{1}{4}E_{n,n-1})^2 \le \frac{1}{4}E_{2n,2n-1}. \quad (14)$$

We will make repeated use of the following consequence: The inequality $E_{2n,2n-1} \le 4A^2$ with some $A > 0$ implies $E_{n,n-1} \le 4A$.

A start for the recursive procedure is the best constant function. The best constant for the interval $[a^2, b^2]$ and the approximation error follow from a simple optimisation of the constant:

$$v_{0,0} = \frac{2ab}{a+b}, \qquad E_{0,0,[a^2,b^2]} = \frac{b-a}{b+a} =: \frac{1}{\rho}. \quad (15)$$

Here ρ is the parameter of the ellipse on which the square root is an analytic function, if the interval $[a^2, b^2]$ is transformed into the interval $[-1, +1]$; cf. Appendix 8. Another important parameter is

$$\kappa := \frac{a}{b}.$$

Note that $v_{0,0}$ is the *harmonic mean* of the function values at the end points.

When Heron's method is applied to a constant function, a linear function with an alternant of length 3 is produced. Hence, $E_{1,0}^{-1} = \left(E_{0,0}^{-1} + \sqrt{E_{0,0}^{-2} - 1}\right)^2$, and Landen transformations provide the sequence

$$\rho = E_{0,0}^{-1} \to E_{1,0}^{-1} \to E_{2,1}^{-1} \to E_{4,3}^{-1} \to E_{8,7}^{-1} \to \dots \tag{16}$$

The asymptotic behaviour of $E_{n,n-1}$ for $n = 2^m$ can be determined already from this sequence. There are the trivial inequalities for the sequence (9)

$$4\lambda_j \le (4\lambda_{j+1})^2. \tag{17}$$

Let

$$\omega := \omega(\kappa) := \omega[a^2, b^2] := \lim_{m\to\infty} \left(\frac{1}{4} E_{2^m, 2^m-1, [a^2,b^2]}\right)^{-1/2^{m+1}}. \tag{18}$$

By (14) the sequence on the right-hand side is monotone, the limit exists, and the monotonicity also implies that

$$E_{n,n-1,[a^2,b^2]} \le 4\omega^{-2n} \tag{19}$$

holds for $n = 2^m$. We will establish the inequality for all $n \in \mathbb{N}$. Moreover, $\omega(\kappa)$ will be expressed in terms of elliptic integrals although the fast convergence of the arithmetic-geometric process is used for its fast computation, as we will see below.

Remark 3.2. We focus on upper bounds for the degree of rational approximation although lower bounds can be obtained by suitable modifications. We elucidate this for a bound corresponding to (19). Let $\lambda_j \ge \frac{1}{4}A + \frac{2}{A}$ for some $A > 1$. Hence,

$$\lambda_{j-1} \ge \left[\frac{1}{4}A + \frac{2}{A} + \sqrt{(\frac{1}{16}A^2 + 1 + \frac{4}{A^2}) - 1}\right]^2 \ge \left(\frac{1}{2}A + \frac{2}{A}\right)^2 \ge \frac{1}{4}A^2 + \frac{2}{A^2}.$$

The bound of λ_{j-1} has the same structure as the bound for λ_j. Now we obtain by induction and from (18)

$$E_{n,n-1,[a^2,b^2]} \ge \frac{4}{\omega^{2n} + 8\omega^{-2n}} \tag{20}$$

for $n = 2^m$. A comparison with (19) elucidates the fast convergence.

3.3 *Extension of the estimate* (19)

A transformation of the interval will be used for the extension of inequality (19) which was previously announced. It enables us to derive sharp estimates from the results for small intervals in Appendix 8. We encounter the arithmetic-geometric process once more.

Lemma 3.1. *Let $n \geq 1$ and (a_j, b_j) be a sequence according to the arithmetic-geometric mean process (8). Then*

$$E_{n,n-1,[a_{j+1}^2,b_{j+1}^2]} = E_{2n,2n-1,[a_j^2,b_j^2]}\,. \tag{21}$$

Proof. Set $r(x) := (x + a_j b_j)/2$. The function $r^2(x)/x$ maps the two subintervals $[a^2, ab]$ and $[ab, b^2]$ monotonously onto $[a_j b_j, (a_j + b_j)^2/4] = [a_{j+1}^2, b_{j+1}^2]$. Next, note that $\sqrt{x} = r(x)\sqrt{x/r^2(x)} = r(x)\sqrt{\xi}$ where $\xi = x/r(x)^2$.

Let $p/q \in R_{n,n-1}$ be the best approximation to $\sqrt{x}$ on $[a_{j+1}^2, b_{j+1}^2]$. Then

$$\frac{P(x)}{Q(x)} := r(x)\frac{p(x/r^2(x))}{q(x/r^2(x))} \in R_{2n,2n-1}$$

provides an approximation for the original interval with the same size of the maximal relative error as p/q on the smaller interval. The monotonicity of the mapping $r^2(x)/x$ assures that there is an alternant of length $4n+1$. Therefore, P/Q is the best approximation, and the proof is complete.

As a by-product we obtain a closed expression for the approximation by linear functions. For completeness, we also recall (15):

$$E_{1,0,[a^2,b^2]} = \Big(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\Big)^2, \qquad E_{0,0,[a^2,b^2]} = \frac{b-a}{b+a}\,. \tag{22}$$

Theorem 3.2. *Let ω be defined by (18). Then the degree of approximation is bounded by (19) for all $n \in \mathbb{N}$.*

Proof. Let $[a_0^2, b_0^2]$ be the interval for which the degree of approximation in $R_{n,n-1}$ is to be estimated and $\omega = \omega[a_0^2, b_0^2]$. Moreover, let $\ell = 2^k$. By Lemma 3.1 we know that

$$E_{0,0,[a_{k+1}^2,b_{k+1}^2]} = E_{1,0,[a_k^2,b_k^2]} = E_{\ell,\ell-1,[a_0^2,b_0^2]} \leq 4\omega^{-2\ell}.$$

From (45) it follows that the parameter of the regularity ellipse associated to the interval $[a_{k+1}^2, b_{k+1}^2]$ is the inverse, i.e.,

$$\rho = \frac{1}{4}\omega^{2\ell}.$$

By (44) we have

$$E_{n,n-1,[a_{k+1}^2,b_{k+1}^2]} \le 4(4\rho-3)^{-2n} \le 4(\omega^{2\ell}-3)^{-2n}.$$

By using the preceding lemma once more we return to the original interval,

$$E_{2\ell n,2\ell n-1,[a_0^2,b_0^2]} \le 4(\omega^{2\ell}-3)^{-2n}.$$

The degree of the numerator is $2\ell n = 2^{k+1}n$. Now we perform $k+1$ Landen transformations in the opposite direction and recall Remark 3.1 to obtain

$$E_{n,n-1,[a_0^2,b_0^2]} \le 4(\omega^{2\ell}-3)^{-2n/2\ell} = 4\omega^{-2n}(1-3\omega^{-2\ell})^{-2n/2\ell}.$$

Since we may choose an arbitrarily large ℓ, the proof is complete.

3.4 An explicit formula

The asymptotic behaviour of the degree of approximation was determined for finite intervals without the knowledge of elliptic integrals – in contrast to [32]. Explicit formulae will be useful for the treatment of the approximation with exponential sums on infinite intervals. The relevant properties of complete elliptic integrals are provided in Appendix 9.

Theorem 3.3. *Let $k = a/b$, then*

$$E_{n,n-1,[a^2,b^2]} \le 4\omega^{-2n}, \quad \textit{for } n = 1,2,3,\ldots \tag{23}$$

with

$$\omega(k) = \exp\left[\frac{\pi \mathbf{K}(k)}{\mathbf{K}'(k)}\right]. \tag{24}$$

Proof. Set $\kappa_{-j} := E_{2^j,2^j-1}$ and $\lambda_{-j} := 1/\kappa_{-j} :=$ for $j = 0,1,2,\ldots$ and extend the two sequences by the backward Landen transformation. From (16) we know that λ_{-j} obeys the rule of the arithmetic-geometric process, and $\kappa_{-j+1} = 2\sqrt{\kappa_{-j}}/(1+\kappa_{-j})$. By Lemma 9.1 and (53) we obtain

$$\begin{aligned}\lim_{j\to\infty}\left(\frac{1}{4}E_{2^j,2^j-1}\right)^{-1/2^j} &\ge \exp\left[\frac{\pi \mathbf{K}'(\kappa_0)}{2\mathbf{K}(\kappa_0)}\right] = \exp\left[\frac{2\pi \mathbf{K}'(\kappa_2)}{\mathbf{K}(\kappa_2)}\right] \\ &= \exp\left[\frac{2\pi \mathbf{K}(\kappa_2')}{\mathbf{K}'(\kappa_2)}\right] \\ &= \exp\left[2\pi \mathbf{K}\left(\frac{1-\kappa_1}{1+\kappa_1}\right)\Big/\mathbf{K}'\left(\frac{1-\kappa_1}{1+\kappa_1}\right)\right]. \end{aligned} \tag{25}$$

It follows from $\kappa_1 = E_{0,0}$ and (22) that

$$\frac{1-\kappa_1}{1+\kappa_1} = \frac{a}{b}.$$

Now the left-hand side of (25) can be identified with ω^2, and the proof is complete.

Example 3.1. We consider the approximation problem on the interval $[1,2]$, and from (22) we know that $E_{0,0} = (\sqrt{2}-1)/(\sqrt{2}+1)$. The sequence (16) and the successive calculation of square roots for modelling (18) yields the tableau

$$\begin{array}{r} \frac{\sqrt{2}-1}{\sqrt{2}+1} = 5.828427 \to 133.87475 \to 4613.84 \to 71687.79 \\ \times 4 \downarrow \\ 23.140689 \leftarrow 535.4915 \leftarrow 286751.2 \end{array}$$

Since $\mathbf{K}(1/\sqrt{2}) = \mathbf{K}'(1/\sqrt{2})$, the evaluation of ω by formula (18) is easy. We get $\omega = \exp(\pi) = 23.1406924$ in accordance with the result in the tableau above.

4 Approximation of $1/x^\alpha$ by exponential sums

4.1 Approximation of $1/x$ on finite intervals

The symbol $E_{n,[a,b]}(f)$ with only one integer index refers to the approximation by exponential sums of order n. In order to have a short notation we start with the approximation of $1/x$: First we note that

$$E_{n,[a,b]}(1/x) = \frac{1}{a} E_{n,[1,b/a]}(1/x). \tag{26}$$

Indeed, let u_n be the best approximation of $1/x$ on the interval $[1,b/a]$. The transformation $x = at$ yields

$$\frac{1}{x} - \frac{1}{a} u_n\left(\frac{x}{a}\right) = \frac{1}{a}\left[\frac{1}{t} - u_n(t)\right]. \tag{27}$$

Since the alternant is transformed into an alternant, we have (26).

Theorem 4.1. *Let $0 < a < b$ and $k = a/b$. Then*

$$E_{n,[a,b]}(1/x) \le \frac{c(k)}{a}\, n\, \omega(k)^{-2n}$$

with $\omega(k)$ given by (24) and $c(k)$ depending only on k.

Proof. By (26) it is sufficient to study approximation on the interval $[1,1/k]$. To this end, we consider the approximation of $f(x) := \frac{1}{x+1/n}$ on the interval $[1-1/n, 1/k-$

$1/n]$. Since we are interested in upper bounds, we may enlarge the interval to $[1-1/n, 1/k]$. It follows from Lemma 2.1, Theorem 3.3, and $f(0) = n$ that

$$E_{n,[1,1/k]}(1/x) \le E_{n,[1-1/n,1/k]}(1/(x+1/n)) \le 2n4\left[\omega\left(\frac{1-1/n}{1/k}\right)\right]^{-2n}.$$

Since the function $k \mapsto \omega(k)$ is differentiable, we have

$$\omega\left(\frac{1-1/n}{1/k}\right) = \omega(k[1-1/n]) \ge \omega(k)(1-\frac{c}{n})$$

with $c = c(k)$ being a bound of the derivative in a neighbourhood of k. We complete the proof by recalling $\lim_{n\to\infty}(1-c/n)^{2n} = \mathrm{e}^{-2c}$.

Theorem 4.1 provides only an upper bound. The following examples for small and large intervals, respectively, show that the order of exponential decay proved there is sharp. The numerical results give rise to the conjecture that the polynomial term is too conservative and that

$$E_{n,[a,b]}(1/x) \approx n^{1/2}\omega(k)^{-2n}.$$

Example 4.1. The parameter for the (small) interval $[1,2]$, i.e., $[a^2,b^2] = [1,4]$ is evaluated in the following tableau and is to be compared with the numbers in the third column of Table 2.

$$\begin{array}{|rrrrr|}\hline 3 & \to 33.970 & \to 4613.84 & \to & 85150133 \\ & & & & \times 4 \downarrow \\ 11.655 & \leftarrow 135.85 & \leftarrow 18445.3 & \leftarrow & 340600530 \\ \hline \end{array}$$

Example 4.2. The parameter for the large interval $[1,1000]$, i.e., $[a^2,b^2] = [1,10^6]$ is evaluated in the following tableau and is to be compared with the numbers in the third column of Table 3.

$$\begin{array}{|rrrrr|}\hline 1001/999 & \to 1.13488 & \to 2.79396 & \to 29.1906 \to & 3406.37 \\ & & & & \times 4 \downarrow \\ & 1.813 \leftarrow & 3.2869 \leftarrow & 10.804 \leftarrow & 116.728 \leftarrow 13625 \\ \hline \end{array}$$

4.2 *Approximation of* $1/x$ *on* $[1,\infty)$

If we fix n and consider the approximation problem on the interval $[1,R]$, then the bound in Theorem 4.1 increases with R. This does not reflect the right asymptotic behaviour.

Table 2 Numerical results for $1/x$ (left) and $1/\sqrt{x}$ (right) on $[1,2]$

f	$1/x$		$1/\sqrt{x}$
n	E_n	$\frac{2n}{2n-1}\frac{E_{n-1}}{E_n}$	E_n
1	$2.12794\cdot 10^{-2}$		$1.26035\cdot 10^{-2}$
2	$2.07958\cdot 10^{-4}$	136.43	$9.28688\cdot 10^{-5}$
3	$1.83414\cdot 10^{-6}$	136.06	$6.83882\cdot 10^{-7}$
4	$1.54170\cdot 10^{-8}$	135.96	$5.03516\cdot 10^{-9}$
5	$1.26034\cdot 10^{-10}$	135.92	$3.70688\cdot 10^{-11}$
6	$1.01179\cdot 10^{-12}$	135.89	$2.72889\cdot 10^{-13}$

Table 3 Numerical results for $1/x$ (left) and $1/\sqrt{x}$ (right) on $[1,1000]$

f	$1/x$		$1/\sqrt{x}$
n	E_n	$\frac{2n}{2n-1}\frac{E_{n-1}}{E_n}$	E_n
5	$6.38478\cdot 10^{-4}$		$1.21681\cdot 10^{-3}$
6	$2.17693\cdot 10^{-4}$	3.1995	$3.68730\cdot 10^{-4}$
7	$7.15300\cdot 10^{-5}$	3.2776	$1.11788\cdot 10^{-4}$
8	$2.32088\cdot 10^{-5}$	3.2875	$3.39264\cdot 10^{-5}$
9	$7.46801\cdot 10^{-6}$	3.2905	$1.03020\cdot 10^{-5}$
10	$2.38880\cdot 10^{-6}$	3.2908	$3.12940\cdot 10^{-6}$
11	$7.60494\cdot 10^{-7}$	3.2907	$9.50867\cdot 10^{-7}$
12	$2.41164\cdot 10^{-7}$	3.2905	$2.88981\cdot 10^{-7}$
13	$7.62271\cdot 10^{-8}$	3.2903	$8.78389\cdot 10^{-8}$

The error curve for the best approximation u_n has $2n$ zeros in $[1,R]$. It follows from Theorem 2.1 that $u_n(x) < 1/x$ and

$$\left|\frac{1}{x} - u_n(x)\right| < \frac{1}{x} < \frac{1}{R}$$

holds for $x > R$. Hence, for all $R > 1$,

$$E_{n,[1,\infty]}(1/x) \le \max\left\{E_{n,[1,R]}(1/x), \frac{1}{R}\right\}. \tag{28}$$

It is our aim to choose R such that the right-hand side of (28) is close to the minimal value.

Table 4 Numerical results for $1/x$ (left) and $1/\sqrt{x}$ (right) on $[1,\infty)$

f	$1/x$			$1/\sqrt{x}$
n	R_n	E_n	$E_n e^{\pi\sqrt{2n}}/\log(2+n)$	E_n
1	8.667	$8.55641\cdot 10^{-2}$	6.62	$1.399\cdot 10^{-1}$
2	41.54	$1.78498\cdot 10^{-2}$	6.89	$4.087\cdot 10^{-2}$
5	1153	$6.42813\cdot 10^{-4}$	6.82	$3.297\cdot 10^{-3}$
10	56502	$1.31219\cdot 10^{-5}$	6.67	$1.852\cdot 10^{-4}$
15	$1.175\cdot 10^{6}$	$6.31072\cdot 10^{-7}$	6.62	$2.011\cdot 10^{-5}$
20	$1.547\cdot 10^{7}$	$4.79366\cdot 10^{-8}$	6.60	$3.083\cdot 10^{-6}$
25	$1.514\cdot 10^{8}$	$4.89759\cdot 10^{-9}$	6.60	$5.898\cdot 10^{-7}$
30	$1.198\cdot 10^{9}$	$6.18824\cdot 10^{-10}$	6.61	$1.321\cdot 10^{-7}$
35	$8.064\cdot 10^{9}$	$9.19413\cdot 10^{-11}$	6.62	$3.336\cdot 10^{-8}$
40	$4.771\cdot 10^{10}$	$1.55388\cdot 10^{-11}$	6.64	$9.264\cdot 10^{-9}$
45	$2.540\cdot 10^{11}$	$2.91895\cdot 10^{-12}$	6.66	$2.780\cdot 10^{-9}$
50	$1.237\cdot 10^{12}$	$5.99210\cdot 10^{-13}$	6.68	$8.901\cdot 10^{-10}$

In order to avoid the singularity at $x=0$, we consider the approximation of $f(x) := 1/(x+1/2)$ on the interval $[\frac{1}{2}, R-\frac{1}{2}]$. The constant shift of $1/2$ is better suited for estimates on large intervals. Now it follows from Lemma 2.1, Theorem 3.3, and $f(0)=2$ that

$$E_{n,[1,R]}(1/x) \le 2\cdot 2\cdot 4\exp\left[-\frac{2n\pi\mathbf{K}(k)}{\mathbf{K}'(k)}\right]$$

with $k = 1/(2R-1)$. From (48) we know that $\mathbf{K}(k) \ge \pi/2$. This inequality and (50) imply

$$E_{n,[1,R]}(1/x) \le 16\exp\left[-\frac{\pi^2 n}{\log(\frac{4}{k}+2)}\right] \le 16\exp\left[-\frac{\pi^2 n}{\log(8R)}\right]. \tag{29}$$

The choice $R = \frac{1}{8}\exp[\pi\sqrt{n}]$ yields the final result:

$$E_{n,[1,\infty]}(1/x) \le 16\exp\left[-\pi\sqrt{n}\right]. \tag{30}$$

The results in Table 4 are based on numerically computed best approximations. They lead to the conjecture that

$$E_{n,[1,\infty]}(1/x) \approx \log n\cdot\exp\left[-\pi\sqrt{2n}\right]. \tag{31}$$

In particular, the exponents in (30) and (31) differ by a factor of $\sqrt{2}$. The same gap is found with the method discussed in Appendix 10. (The approximation by sinc functions leads even to a larger gap [7] and §11.)

Table 5 Comparison of the approximation of $\sqrt{x}$ by rational functions and $1/x$ by exponential sums

k^{-1}	$E_{4,3,[1,k^{-2}]}(\sqrt{x})$	$E_{4,[1,k^{-1}]}(1/x)$
2	$1.174\cdot 10^{-8}$	$1.542\cdot 10^{-8}$
10	$8.935\cdot 10^{-5}$	$5.577\cdot 10^{-5}$
100	$9.781\cdot 10^{-3}$	$1.066\cdot 10^{-3}$
500	$2.220\cdot 10^{-2}$	$1.700\cdot 10^{-3}$

The gap may be surprising since the numerical results in the Tables 2 and 3 show that the theory provides sharp estimates for the asymptotic behaviour for large n. It is the factor in front of the exponential term in Theorem 4.1 that is responsible. We have compared the data for $n = 4$, i.e., for a small n in Table 5. They show that the application of Lemma 2.1 provides estimates which are too conservative on large intervals, although the behaviour for large n is well modelled.

The logarithmic factor in front of (31) also shows that it will not be easy to establish sharper estimates for the infinite interval.

4.3 Approximation of $1/x^{\alpha}$, $\alpha > 0$

When more freedom in the exponent of the given function is admitted, there are no substantial changes on finite intervals. Proceeding as in the proof of Theorem 4.1 we obtain with $k = a/b$:

$$E_{n,[a,b]}(x^{-\alpha}) \le \frac{c(k)}{a^{\alpha}}\, n^{\alpha}\, \omega(k)^{-2n} \tag{32}$$

with $\omega(k)$ given in Theorem 4.1. The exponential term on the right-hand side that dominates the asymptotic behaviour for large n is unchanged.

The situation on infinite intervals is different. Given $R > 1$, we obtain with $k = 1/(2R-1)$ in analogy to (29)

$$E_{n,[1,R]}(x^{-\alpha}) \le 2^{\alpha}\, 8 \exp\left[-\frac{\pi^2 n}{\log(\frac{4}{k}+2)}\right] \le 2^{\alpha}\, 8 \exp\left[-\frac{\pi^2 n}{\log(8R)}\right] \tag{33}$$

Moreover, we have $E_{n,[1,\infty]}(x^{-\alpha}) \le \max\left\{E_{n,[1,R]}(x^{-\alpha}), R^{-\alpha}\right\}$ in analogy to (28). A suitable choice is $R = \frac{1}{8}\exp[\pi\sqrt{n/\alpha}]$. It yields

$$E_{n,[1,\infty]}(x^{-\alpha}) \le 2^{\alpha}\, 8 \exp[-\pi\sqrt{\alpha n}]. \tag{34}$$

The asymptotic decay depends heavily on α.

5 Applications of $1/x$ approximations

5.1 About the exponential sums

Let $[a,b] \subset (0,\infty]$ be a possibly semi-infinite interval, e.g. $b = \infty$ is allowed. The best approximation in $[a,b]$ is denoted by

$$\frac{1}{x} \approx u_{n,[a,b]}(x) = \sum_{\nu=1}^{n} \alpha_{\nu,[a,b]} \exp(-t_{\nu,[a,b]}\, x).$$

The rule (27) is inherited by the coefficients,

$$\alpha_{\nu,[a,b]} := \frac{1}{a}\alpha_{\nu,[1,b/a]}, \quad t_{\nu,[a,b]} := \frac{1}{a}t_{\nu,[1,b/a]},$$

and allows us to reduce the considerations to intervals of the form $[1,R]$. Due to (26) the approximation errors $E_{n,[a,b]}$ are related by $E_{n,[a,b]} = \frac{1}{a}E_{n,[1,b/a]}$. The coefficients of $v_{n,[1,R]}$ for various n and R can be found in [31].

5.2 Application in quantum chemistry

The so-called Coupled Cluster (CC) approaches are rather accurate but expensive numerical methods for solving the electronic many-body problems. The cost may be $\mathcal{O}(N^7)$, where N is the number of electrons. One of the bottlenecks is an expression of the form

$$\frac{numerator}{\varepsilon_a + \varepsilon_b + \ldots - \varepsilon_j - \varepsilon_i},$$

where $\varepsilon_i, \varepsilon_j, \ldots < 0$ are energies related to occupied orbitals $i, j, \ldots$, while $\varepsilon_a, \varepsilon_b, \ldots > 0$ are energies related to virtual orbitals $a, b, \ldots$ The denominator belongs to an interval $[E_{\min}, E_{\max}]$, where the critical lower energy bound $E_{\min}$ depends on the so-called HOMO-LUMO gap.

The denominator leads to a coupling of all orbitals $a, b, \ldots, i, j, \ldots$, whereas the numerator possesses a partial separation of variables. Therefore one tries to replace $1/(\varepsilon_a + \varepsilon_b + \ldots - \varepsilon_i - \varepsilon_j)$ by a separable expression. Such a separation saves one order in the complexity.

Any exponential sum approximation $\frac{1}{x} \approx \sum_{\nu=1}^{n} \alpha_\nu \exp(-t_\nu x)$ leads to the separable expression

$$1/(\varepsilon_a + \varepsilon_b + \ldots - \varepsilon_i - \varepsilon_j) \approx \sum_{\nu=1}^{n} \alpha_\nu \mathrm{e}^{-t_\nu \varepsilon_a} \mathrm{e}^{-t_\nu \varepsilon_b} \cdots \mathrm{e}^{t_\nu \varepsilon_j} \mathrm{e}^{t_\nu \varepsilon_i}.$$

In quantum chemistry, Almlöf [1] used the representation

$$\frac{1}{x} = \int_0^\infty \mathrm{e}^{-sx} \mathrm{d}s$$

together with a quadrature formula $\sum_{\nu=1}^{n} \alpha_\nu e^{-t_\nu x}$. This ansatz has been used in many places like the Møller-Plesset second order perturbation theory (MP2, cf. [1, 15, 16, 30]), computation of connected triples contribution in MP4 (cf. [15]), atomic orbital (AO)-MP2 (cf. [16, 2, 19]), AO-MP2 energy gradient (cf. [16, 24]), combinations with the resolution of the identity (RI)-MP2 (cf. [10, 9]), and the density-matrix-based MP2 (cf. [27, 17]).

It is hard to adapt the quadrature to the interval $[E_{\min}, E_{\max}]$ where the approximation is needed. The favourite choice among those used in quantum chemistry is the Gauss-Legendre quadrature applied to the transformed integral

$$\int_0^\infty e^{-sx} ds = \int_0^1 e^{-tx/(1-t)} \frac{dt}{(t-1)^2} \qquad (s = t/(1-t)).$$

Best approximations are only considered with respect to a weighted L^2-norm (cf. [23]). Best approximations in the supremum norm has not been considered in this community. The recent paper [28] contains a comparison between the Gauss-Legendre approach and the best approximation $u_{n,[E_{\min},E_{\max}]}$ for various applications. For instance, an error of size ≈ 0.005 of the MP2 energies for benzene with the aug-cc-pCVTZ basis set is obtained by the Gauss-Legendre quadrature with 14 terms, while the same accuracy is already obtained by the best approximation with 4 terms (best approximations with 14 terms yield an accuracy of $2 \cdot 10^{-10}$). The value $R = E_{\max}/E_{\min}$ for this example is about 278.

5.3 Inverse matrix

The previous application refers to the scalar function $1/x$. Now we consider its matrix-valued version M^{-1} for a matrix M with positive spectrum $\sigma(M) \subset [a,b] \subset (0,\infty]$. Formally, we have

$$M^{-1} \approx u_{n,[a,b]}(M) = \sum_{\nu=1}^{n} \alpha_{\nu,[a,b]} \exp(-t_{\nu,[a,b]} M).$$

Additionally, we assume that M is diagonalisable: $M = T^{-1} D T$. Then a simple calculation shows the estimate

$$\left\| M^{-1} - u_{n,[a,b]}(M) \right\|_2 \le cond_2(T)\, E_{n,[a,b]}$$

with respect to the spectral norm. We emphasize that the spectral norm estimate hinges on a *uniform* estimate of $\frac{1}{x} - u_{n,[a,b]}$ on the spectral interval $[a,b]$. Approximations of $1/x$ by exponential sums with respect to the L^2-norm would not be helpful.

The approximation of M^{-1} seems to be rather impractical since now matrix exponentials $\exp(-t_\nu M)$ have to be evaluated. The interesting applications, however, are matrices which are sums of Kronecker products.

We recall that a differential operator L is called separable in $x_1, \ldots, x_d$, if $L = \sum_{i=1}^{d} L_i$, where the operator L_i applies only to the variable x_i and the coefficients of L_i depend only on x_i. Let the domain of the boundary value problem be of product form: $\Omega = \Omega_1 \times \ldots \times \Omega_d$. Then a suitable discretisation leads to an index set I of product form: $I = \prod_{i=1}^{d} I_i$, where I_i contains the indices of the i-th coordinate direction.

The system matrix for a suitable discretisation has the form

$$\mathbf{M} = \sum_{i=1}^{d} I \otimes \cdots \otimes M^{(i)} \otimes \cdots \otimes I, \quad M^{(i)} \in \mathbb{R}^{I_i \times I_i} \tag{35}$$

(factor $M^{(i)}$ at i-th position). We assume that all $M^{(i)}$ are positive definite with smallest eigenvalue $\lambda_{\min}^{(i)}$. Since the spectrum of $\mathbf{M}$ is the sum $\sum_{i=1}^{d} \lambda^{(i)}$ of all $\lambda^{(i)} \in \sigma(M^{(i)})$, the minimal eigenvalue of $\mathbf{M}$ is $\lambda_{\min} := \sum_{i=1}^{d} \lambda_{\min}^{(i)}$. Since $\lambda_{\min}^{(i)}$ approximates the smallest eigenvalue of L_i, we have $\lambda_{\min} = \mathcal{O}(1)$.

Now we take the best approximation E_n^* with respect $[\lambda_{\min}, b]$ ($b = \sum_{i=1}^{d} \lambda_{\max}^{(i)}$ or $b = \infty$). We know that for the symmetric matrices

$$\left\| u_n(\mathbf{M}) - \mathbf{M}^{-1} \right\| \leq E_{n,[\lambda_{\min}, b]}.$$

For the evaluation of $u_n(\mathbf{M}) = \sum_{\nu=1}^{n} \alpha_\nu \exp(-t_\nu \mathbf{M})$ we make use of the identity

$$\exp(-t_\nu \mathbf{M}) = \bigotimes_{i=1}^{d} \exp\left(-t_\nu M^{(i)}\right)$$

with $M^{(i)}$ from (35) (cf. [14, §15.5.2]) and obtain

$$\mathbf{M}^{-1} \approx \sum_{\nu=1}^{n} \alpha_\nu \bigotimes_{i=1}^{d} \exp\left(-t_\nu M^{(i)}\right).$$

As described in [14, §13.3.1], the hierarchical matrix format allows us to approximate $\exp\left(-t_\nu M^{(i)}\right)$ with a cost almost linear in the size of $M^{(i)}$. The total number of arithmetical operations is $\mathcal{O}(n \sum_{i=1}^{d} \#I_i \log^* \#I_i)$. For $\#I_i = N$ $(1 \leq i \leq d)$ this expression is $\mathcal{O}(ndN \log^* N)$ and depends only linearly on d.

Therefore, it is possible to treat cases with large N and d. In [11], examples can be found with $N = 1024$ and $d \approx 1000$. Note that in this case $\mathbf{M}^{-1} \in \mathbb{R}^{M \times M}$ with $M \approx 10^{3000}$.

6 Applications of $1/\sqrt{x}$ approximations

6.1 Basic facts

Let $[a,b] \subset (0,\infty]$ be as above. We consider the best approximation of $1/\sqrt{x}$ in $[a,b]$:

$$\frac{1}{\sqrt{x}} \approx u_{n,[a,b]}(x) = \sum_{\nu=1}^{n} \alpha_{\nu,[a,b]} \exp(-t_{\nu,[a,b]}\, x).$$

In this case the relations

$$\alpha_{\nu,[a,b]} = \frac{1}{\sqrt{a}}\alpha_{\nu,[1,b/a]}, \quad t_{\nu,[a,b]} = \frac{1}{a}t_{\nu,[1,b/a]}, \quad E_{n,[a,b]} = \frac{1}{\sqrt{a}}E_{n,[1,b/a]} \tag{36}$$

hold, and again it is sufficient to determine $v_{n,[1,R]}$ with $R := b/a$. The coefficients of $v_{n,[1,R]}$ for various n and R can be obtained from [31].

The standard application uses the substitution $x = \|y\|^2 = \sum_{i=1}^{d} y_i^2$ with a vector $y \in \mathbb{R}^d$. Then we obtain the sum

$$G_{n,[a,b]}(y) := \sum_{\nu=1}^{n} \alpha_{\nu,[a^2,b^2]} \prod_{i=1}^{d} \exp(-t_{\nu,[a^2,b^2]}\, y_i^2)$$

of Gaussians which is the best approximation of $1/\|y\|$ for $\|y\| \in [a,b]$. Since, in 3D, $1/\|y\|$ is the Newton potential or Coulomb potential, this function appears in many problems.

6.2 Application to convolution

A further application refers to the convolution integral

$$\Phi(x) := \int_{\mathbb{R}^3} \frac{f(y)}{\|x-y\|}\,\mathrm{d}y.$$

We assume that f can be written as a sum of simple products. For simplicity we consider only one term:

$$f(y) = f_1(y_1)\, f_2(y_2)\, f_3(y_3). \tag{37}$$

When we replace $1/\|x-y\|$ by an approximation of the form

$$G_n(x-y) = \sum_{\nu=1}^{n} \alpha_\nu \prod_{i=1}^{3} \exp(-t_\nu\,(x_i-y_i)^2),$$

the convolution integral becomes

$$\Phi_n(x) := \int_{\mathbb{R}^3} G_n(x-y) f(y) \mathrm{d}y = \sum_{\nu=1}^{n} \alpha_\nu \prod_{i=1}^{3} \int_{\mathbb{R}} \exp(-t_\nu (x_i - y_i)^2) f_i(y_i) \mathrm{d}y_i,$$

and the 3D convolution is reduced to three 1D convolutions. This fact reduces the computational cost substantially. In the paper [13] this technique is applied for the case that the functions f_i are piecewise polynomials. However, there still remains a gap is to be closed. We have used some best approximation $G_n = G_{n,[a,b]}$ of $1/\|\cdot\|$. The value of b may be infinite or finite, if the support of f is finite and the evaluation of $\Phi(x)$ is required for x in a bounded domain. The lower bound a may be small but positive. Therefore the difference $\Phi(x) - \Phi_n(x)$ contains the term $\delta\Phi_n(x) := \int_{\|x-y\|\le A} \big(\frac{1}{\|x-y\|} - G_n(x-y)\big) f(y) \mathrm{d}y$, where the approximation fails. This contribution can be treated separately to obtain $\Phi_n + \delta\Phi_n \approx \Phi$. As shown in [13] the numerical cost arising from the extra term, is low.

A related problem is the integral $I := \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{g(x)f(y)}{\|x-y\|} \mathrm{d}x\mathrm{d}y$ which appears for example as "two-electron integral" in Quantum Chemistry. It can be considered and computed as the scalar product of g with the function Φ from above. Another approach is the replacement of $1/\|\cdot\|$ by the exponential sum G_n. Assuming again that f and g are simple products like in (37), the identity

$$\begin{aligned} I_n &:= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} G_n(x-y) g(x) f(y) \mathrm{d}x\mathrm{d}y \\ &= \sum_{\nu=1}^{n} \alpha_\nu \prod_{i=1}^{3} \int_{\mathbb{R}} \int_{\mathbb{R}} \exp(-t_\nu (x_i - y_i)^2) g_i(x_i) f_i(y_i) \mathrm{d}y_i \end{aligned}$$

shows that the six-dimensional integral is reduced to three two-dimensional ones. Concerning the error analysis, we split the integral $I = I_{\mathrm{near}} + I_{\mathrm{far}}$ into the near-field and far-field parts

$$I_{\mathrm{near}} := \int_{\|z\|\le r} \int_{\mathbb{R}^3} \frac{g(z+y)f(y)}{\|z\|} \mathrm{d}z\mathrm{d}y, \quad I_{\mathrm{far}} := \int_{\|z\|\ge r} \int_{\mathbb{R}^3} \frac{g(z+y)f(y)}{\|z\|} \mathrm{d}z\mathrm{d}y.$$

Let $I_n = I_{\mathrm{near},n} + I_{\mathrm{far},n}$ be the corresponding splitting with $1/\|\cdot\|$ replaced by G_n. We assume that $f, g \in C(\mathbb{R}^3)$ have bounded support[1]. Then for $\|z\| \le r$ the error can be bounded by $|I_{\mathrm{near},n}| + |I_{\mathrm{near}}| \lesssim \int_{\|z\|\le r} \frac{\mathrm{d}z}{\|z\|} = \mathcal{O}(r^2)$. If an error ε is desired, we need $r \sim \sqrt{\varepsilon}$. This requires the choice $G_n = G_{n,[\sqrt{\varepsilon},\infty)}$. The approximation error of G_n is $\left\| 1/\|\cdot\| - G_{n,[\sqrt{\varepsilon},\infty)} \right\|_{\infty,\|z\|\ge\sqrt{\varepsilon}} = E_{n,[\varepsilon,\infty)} = \frac{1}{\sqrt{\varepsilon}} E_{n,[1,\infty)} = \mathcal{O}(\frac{1}{\sqrt{\varepsilon}} \exp(-c\sqrt{n}))$. To equilibrate both terms, we have to choose $n = \mathcal{O}(\log^2 \varepsilon)$.

[1] In Quantum Chemistry, the functions have infinite support but decay exponentially. Therefore, similar error estimates hold.

6.3 Modification for wavelet applications

Let f be the function which is to be approximated by an exponential sum E_n. There are wavelet applications, where scalar products $\langle f, \psi \rangle$ with wavelets ψ appear. Wavelets have a certain number of vanishing moments, i.e., $\langle p, \psi \rangle = 0$ for all polynomials of degree $\le \ell$ for some $\ell \in \mathbb{N}_0$. In order to keep the moments, one can approximate f by the mixed ansatz

$$\sum_{\nu=1}^{n} \alpha_\nu \exp(-t_\nu x) + \sum_{\nu=0}^{\ell} \beta_\nu x^\nu .$$

Let $u_n^*(x) + p_\ell^*(x)$ be the best approximation of this form in the interval $[a,b] \supset \mathit{support}(\psi)$. By definition we have

$$\langle f, \psi \rangle \approx \langle u_n^* + p_\ell^*, \psi \rangle = \langle v_n^*, \psi \rangle .$$

Therefore, the polynomial part p_ℓ^* need not be stored, and the storage and quadrature costs of $\langle u_n^*, \psi \rangle$ are the same as for the usual best approximation u_n. Of course, the approximation is improved: $\|f - (u_n^* + p_\ell^*)\|_{\infty,[a,b]} \le \|f - u_n\|_{\infty,[a,b]}$.

For an illustration we give the approximation accuracy for $f(x) = 1/\sqrt{x}$ and $n = 4$, $\ell = 1$ in the interval $[1,10]$. The standard approximation is $E_{4,[1,10]} = 2.856 \cdot 10^{-5}$, while the new approach yields the better result $E^*_{4,[1,10]} = \|f - (u_4^* + p_1^*)\|_{\infty,[1,10]} = 2.157 \cdot 10^{-6}$. When these approximations are used after the substitution $x = \|y\|^2 = \sum_{i=1}^d y_i^2$, one has to take into account that $p_\ell^*(\|y\|^2)$ is a polynomial of degree 2ℓ, i.e., a corresponding number of vanishing moments is required. More details can be found in [12].

6.4 Expectation values of the H-atom

In [8, 18] the reduction similar to (1) is applied to the evaluation of expectation values of the H-atom at the ground state. The error is given in terms of the integral

$$4\alpha^2 \int_0^\infty \{v_n(r^2) - r^{-1}\} \, \mathrm{e}^{-2\alpha r} r^2 \mathrm{d}r,$$

where $v(x) = v_n(r^2)$ is an exponential sum that approximates $1/\sqrt{x}$. It is independent of α, if v_n is adapted for each α in the spirit of (36). According to [8, p.138] the asymptotic behaviour is

$$An^{1/2} \exp\left[-\pi\sqrt{\frac{4}{3}n}\right]. \tag{38}$$

We will estimate the more conservative integral

$$\varepsilon_n := 4\int_0^\infty |v_n(r^2) - r^{-1}|\,\mathrm{e}^{-2r} r^2 \mathrm{d}r. \tag{39}$$

for (almost) best approximations v_n. Without loss of generality we set $\alpha = 1$. Specifically, (39) is a weighted L_1 norm, and the treatment is typical for the estimation of weighted L_1 norms of the error [6]. The infinite interval is split into three parts $[0,a]$, $[a,b]$, and $[b,\infty)$. The points a and b are chosen such that the contributions of the first and the third interval are small. A bound for the contribution of $[a,b]$ is determined from the maximal error on this subinterval. Here the results of Section 4 are applied.

We set $a := \beta\sqrt{2n}\exp[-\frac{1}{2}\beta\sqrt{n}]$ and $b := \frac{1}{2}\beta\sqrt{n}$ with β to be fixed later with $\beta \geq 1$. Let v_n be the best approximation or, more generally, be determined by a procedure such that it interpolates $\sqrt{x}$ at $2n$ points in $[a,b]$). In these cases $|v_n - r^{-1}| < r^{-1}$ holds for $x < a$ and $x > b$. Hence,

$$4\int_0^a |v_n(r^2) - r^{-1}|\,\mathrm{e}^{-2r} r^2 \mathrm{d}r \leq 4\int_0^a r\mathrm{d}r = 2a^2 = 4\beta^2 n\exp[-\beta\sqrt{n}].$$

Similarly,

$$\begin{aligned} 4\int_b^\infty |v_n(r^2) - r^{-1}|\,\mathrm{e}^{-2r} r^2 \mathrm{d}r &\leq 4\int_b^\infty \mathrm{e}^{-2r} r\mathrm{d}r = (2b+1)\mathrm{e}^{-2b} \\ &\leq 2\beta\sqrt{n}\exp[-\beta\sqrt{n}]. \end{aligned}$$

Next, set $E := \max_{a\leq r\leq b} |v_n(r^2) - r^{-1}|$ and observe that

$$4\int_a^b |v_n(r^2) - r^{-1}|\,\mathrm{e}^{-2r} r^2 \mathrm{d}r \leq 4E\int_0^\infty \mathrm{e}^{-2r} r^2 \mathrm{d}r = E.$$

The substitution $x = r^2$ shows that we have to consider the approximation on the interval $[a^2, b^2]$. Recalling (36) we apply the guaranteed bound (33) to the best approximation for $R = (b/a)^2 = \frac{1}{8}\exp[-\beta\sqrt{n}]$:

$$\begin{aligned} E = \max_{a^2\leq x\leq b^2} |v_n(x) - \frac{1}{\sqrt{x}}| &\leq \frac{1}{a} 12\exp\left[-\frac{\pi^2 n}{\log(8R)}\right] \\ &\leq \exp[\frac{1}{2}\beta\sqrt{n}]\frac{1}{\beta\sqrt{2n}}\,12\exp\left[-\frac{\pi^2 n}{\beta\sqrt{n}}\right] \\ &\leq 12\exp\left[\frac{1}{2}\beta\sqrt{n} - \frac{\pi^2 n}{\beta}\right]. \end{aligned}$$

Finally we set $\beta = \pi\sqrt{\frac{2}{3}}$ to obtain $E \leq 12\mathrm{e}^{-\beta\sqrt{n}}$. The collection of the integrals yields

$$\varepsilon_n \leq cn\exp\left[-\pi\sqrt{\frac{2}{3}n}\right].$$

This bound is not as good as (38), while the sinc method yields bounds for norms of the error that are not as sharp as the results in Section 4; cf. [7] and §11.

7 Computation of the best approximation

Let $f(x)$ be the function $1/x$ or $1/\sqrt{x}$ to be approximated. We make the ansatz $u_n(x;\{\alpha_\nu\},\{t_\nu\}) = \sum_{\nu=1}^{n} \alpha_\nu \exp(-t_\nu x)$ and define the error

$$\eta_n(x;\{\alpha_\nu\},\{t_\nu\}) := u_n(x;\{\alpha_\nu\},\{t_\nu\}) - f(x).$$

As described in Definition 3.1, the best approximation in the interval $[1,R]$ is characterised by an *alternant* consisting of $2n+1$ points $x_0 < x_1 < \cdots < x_{2n}$ in the interval satisfying the equi-oscillation conditions (10) and (11).

Then $E_{n,[1,R]} := |\eta_n(x_i,\{\alpha_\nu\},\{t_\nu\})|$ is the optimal error $\|u_n(\cdot;\{\alpha_\nu\},\{t_\nu\}) - f\|_{\infty,[1,R]}$ over all $\{\alpha_\nu\},\{t_\nu\}$. The Remez algorithm determines the $2n$ unknown coefficients $\{\alpha_\nu\}$ and $\{t_\nu\}$ from the $2n$ equations $\eta_n(x_i) = -\eta_n(x_{i+1})$. Details of the implementation which we use will follow below.

There is a specific difference between best approximations by polynomials and exponential sums. For polynomials, the error $|\eta_n|$ approaches ∞ as $|x| \to \infty$. Since in our setting $f(x) \to 0$ and $E_n(x;\{\alpha_\nu\},\{t_\nu\}) \to 0$ as $x \to \infty$, the error satisfies $|\eta_n| \to 0$ for $x \to \infty$ (cf. §4.2). As a consequence, for each n there is a unique $R_n > 0$ such that all best approximations in intervals $[1,R]$ with $R \geq R_n$ have the same alternants. In particular, $x_{2n} = R_n$ holds. Hence, best approximations in $[1,R_n]$ are already best approximations in $[1,\infty)$. On the other hand, best approximations in $[1,R]$ with $R < R_n$ satisfy $x_{2n} = R$ and lead to larger errors $|\eta_n(x)| > E_{n,[1,R]}$ for $x > R$ beyond the end of the interval.

From the equi-oscillation property (10) we conclude that there are zeros $\xi_i \in (x_{i-1},x_i)$ of η_n for $1 \leq i \leq 2n$. Formally, we set $\xi_0 := 1$ and $\xi_{2n+1} := R$. Then $\eta_n(x_i)$ is the (local) extremum in the interval $[\xi_i,\xi_{i+1}]$ for $0 \leq i \leq 2n$. Remez-like algorithms start from *quasi-alternants*, i.e., sets of points which satisfy (10), but not yet (11). They replace x_i by the true extrema in $[\xi_i,\xi_{i+1}]$ and try to satisfy $\eta_n(x_i) = -\eta_n(x_{i+1})$ or a relaxed version with updated exponential sums (cf. Remez [21]).

Since the underlying equations are nonlinear, one must use Newton-like methods. The natural choice of parameters of u_n are the coefficients $\{\alpha_\nu\}$ and $\{t_\nu\}$. However variations in these parameters may change the sign structure of $\eta_n = u_n - f$ completely[2], but the Remez algorithms relies on the condition (10). Therefore, we use the zeros ξ_i $(1 \leq i \leq 2n)$ as parameters: $u_n(x;\{\xi_\nu\})$. Since by definition $u_n(\xi_i;\{\xi_\nu\}) = f(\xi_i)$, the function $u_n(\cdot;\{\xi_\nu\})$ can be considered as the interpolating exponential sum.

[2] Note that η_n might be very small, say 1E-12, for good approximations. Then tiny variations of t_ν may yield a new η_n which is completely positive.

In the case of polynomials we have explicit formulae (Lagrange representation) for the interpolating polynomial. Here, we need a secondary Newton process to compute the coefficients $\{\alpha_\nu\} = \{\alpha_\nu(\xi_1,\ldots,\xi_{2n})\}$ and $\{t_\nu\} = \{t_\nu(\xi_1,\ldots,\xi_{2n})\}$. This makes the algorithm more costly, but stability has priority. For the implementation leading to the results in [31] extended precision is used. Then it is possible to determine, e.g., the approximation u_7 of $1/x$ in $[1,2]$, which leads to the error $E_{7,[1,2]} = 8.020\text{E-}15$, which is rather close to machine precision.

We conclude this section with some practical remarks concerning the computation. Once a best approximation u_n is known in some interval $[1,R]$, it can be used as a good starting value for a next interval $[1,R']$ with R' sufficiently close[3] to R. In general, computations with larger R are easier than those with smaller R because the corresponding size of the error η_n. To determine the first u_n for a value n, one should proceed as follows. For rather small n, it is not so difficult to get convergence from reasonable starting values. Assume that u_{n-1} is known (preferably for a larger value of R). The structure of the coefficients $\{\alpha_\nu\},\{t_\nu\}$ and of the zeros ξ_i allow to "extrapolate" for the missing starting values α_n, t_n and ξ_{2n-1}, ξ_{2n}. The search for reasonable starting values becomes extremely simple, if one makes use of the precomputed values in [31].

Appendices

8 Rational approximation of $\sqrt{x}$ on small intervals

The rule for the transformation of the intervals allows us to extend the error bound (3.1) from all powers of 2 to all $n \in \mathbb{N}$, if we verify them for small intervals. Here we can use Newman's trick that was first applied to the approximation of e^x (see [20]). It is based on the following observation. The special product of linear polynomials is a linear and not a quadratic function if considered on the unit circle in $\mathbb{C}$:

$$(z+\beta)(\bar{z}+\beta) = 2\beta\,\Re e\, z + (r^2+\beta^2) \quad \text{if } |z| = 1.$$

Moreover, the winding number of functions on a circle provide additional information that gives rise to estimates from below.

In particular, given $\rho > 1$, we observe that

$$(\rho+z)(\rho+\bar{z}) = \rho^2+1+2\rho x = 2\rho(a+x) \qquad \text{for } |z| = 1,\ x = \Re e\, z,$$

where $a := \frac{1}{2}(\rho+\rho^{-1})$. Setting $f(z) := \sqrt{\rho+z}$, the induced function in the sense of the lemma below is $F(x) = 2\rho\sqrt{a+x}$. The quotient of the arguments at the left

[3] Let ξ_{2n} belong to $[1,R]$. Then $R' > \xi_{2n}$ is required to maintain the quasi-alternant condition (10). If one wants to get immediately the results for $R' < \xi_{2n}$, also the interpolation points ξ_ν must be diminished (e.g., by $\xi'_\nu := (\xi_\nu - 1)\,\frac{R'-1}{R-1} + 1$).

and the right boundary of the unit interval $[-1,+1]$ is

$$\frac{a-1}{a+1} = \left(\frac{\rho-1}{\rho+1}\right)^2. \tag{40}$$

We note that ρ equals the sum of the semi-axes of that ellipse in $\mathbb{C}$ with foci $+1$ and -1 in which $F(x)$ is an analytic function.

We emphasize that the symbols a and b are generic parameters in this appendix. Next, we recall a simple formula for complex numbers: $f\bar{f} - g\bar{g} = 2\Re e[\bar{f}(f-g)] - |f-g|^2$.

Lemma 8.1 (Newman's trick). *Let $r > 0$. Assume that f is a real analytic function in the disk $|z| < 1$ and that $qf - p$ with $p/q \in R_{mn}$ has $m+n+1$ zeros in the disk while q and f have none. Moreover, let $F(x) = f(z)f(\bar{z})$ where $|z| = r$, $\Re e\, z = rx$, Then*

$$2\min_{|z|=r}\left|f\left(f-\frac{p}{q}\right)\right| \le E_{m,n}(F)\ (1+o(1)) \le 2\max_{|z|=r}\left|f\left(f-\frac{p}{q}\right)\right|. \tag{41}$$

Proof. Since we are concerned with the case $|f - p/q| \ll |f|$, we write

$$\begin{aligned} \bar{f}f - \frac{\bar{p}}{\bar{q}}\frac{p}{q} &= 2\Re e[\bar{f}\left(f-\frac{p}{q}\right)] - \left|f-\frac{p}{q}\right|^2 \\ &= 2\Re e[\bar{f}\left(f-\frac{p}{q}\right)]\,(1+o(1)), \end{aligned} \tag{42}$$

and the upper bound follows from the fact that $\bar{p}p/\bar{q}q$ defines a function in $R_{m,n}$.

The lower bound will be derived by using de la Vallée–Poussin's theorem. Note that

$$\Re e\left[\bar{f}\left(f-\frac{p}{q}\right)\right] = \begin{cases} +\left|f\left(f-\frac{p}{q}\right)\right| & \text{if } \arg\left[\bar{f}\left(f-\frac{p}{q}\right)\right] \equiv 0\ (mod\ 2\pi), \\ -\left|f\left(f-\frac{p}{q}\right)\right| & \text{if } \arg\left[\bar{f}\left(f-\frac{p}{q}\right)\right] \equiv \pi\ (mod\ 2\pi). \end{cases} \tag{43}$$

By assumption $f^{-1}q^{-1}(qf-p)$ has $m+n+1$ zeros counting multiplicities but no pole in the disk $|z| < r$. The winding number of this function is $m+n+1$. The argument of $\bar{f}(f-p/q) = f^{-1}q^{-1}(qf-p)|f|^2$ is increased by $(m+n+1)\,2\pi$ when an entire circuit is performed. The argument is increased by $(m+n+1)\,\pi$ as z traverses the upper half of the circle. Since the function is real for $z = +1$ and $z = -1$, we get a set of $m+n+2$ points with sign changes as in (10). By de la Vallée–Poussin's theorem, the degree of approximation cannot be smaller than the minimum of the absolute values at those points, and the proof is complete.

The trick was invented by Newman [20] for deriving an upper bound of the error when e^x is approximated. The application to lower bounds may be traced back to [5]. The treatment of the square root function followed in [3].

A rational approximant $p_n/q_{n-1} \in R_{n,n-1}$ to $f(z) := \sqrt{\rho+z}$ is given by

$$p_n(z) = \tfrac{1}{2}\left\{\left(\sqrt{\rho}+\sqrt{\rho+z}\right)^{2n} + \left(\sqrt{\rho}-\sqrt{\rho+z}\right)^{2n}\right\},$$
$$q_{n-1}(z) = \frac{1}{2\sqrt{\rho+z}}\left\{\left(\sqrt{\rho}+\sqrt{\rho+z}\right)^{2n} - \left(\sqrt{\rho}-\sqrt{\rho+z}\right)^{2n}\right\},$$

and the error can be written in the form

$$\sqrt{\rho+z} - \frac{p_n}{q_{n-1}} = -\frac{\left(\sqrt{\rho}-\sqrt{\rho+z}\right)^{2n}}{q_{n-1}(z)}.$$

The error curve has a zero of order $2n$ at $z=0$. Therefore, p_n/q_{n-1} is a Padé approximant and Newman's trick with $x = \Re e\, z$ (and $r=1$) yields

$$2\rho\sqrt{a+x} - \frac{p_n(z)p_n(\bar{z})}{q_{n-1}(z)q_{n-1}(\bar{z})} = -2\,\Re e\left[\sqrt{\rho+\bar{z}}\,\frac{\left(\sqrt{\rho}-\sqrt{\rho+z}\right)^{2n}}{q_{n-1}(z)}\right](1+o(1))$$
$$= 8\rho\sqrt{a+x}\,\Re e\frac{z^{2n}}{\left(\sqrt{\rho}-\sqrt{\rho+z}\right)^{4n} - z^{2n}}(1+o(1)).$$

Note that $4\rho - 3 \le |(\sqrt{\rho}+\sqrt{\rho+z})^2| \le \rho+3$. Having upper and lower bounds, the winding number $2n$ yields $2n+1$ points close to an alternant. The relative error is

$$E_{n,n-1}\left(\sqrt{a+x}\right) = \frac{4}{(4\rho+\delta)^{2n}}(1+o(1)) \tag{44}$$

with some $|\delta| \le 3$. The parameters a and ρ are related as given by (40). The approximation of $\sqrt{a+x}$ on the unit interval describes the approximation of $\sqrt{x}$ on $[a-1, a+1]$. From (22) and (40) it follows that

$$E_{0,0}(\sqrt{a+x}) = \frac{1}{\rho}. \tag{45}$$

9 The arithmetic-geometric mean and elliptic integrals

Given two numbers $0 < a_0 < b_0$, the common limit $\lim_{j\to\infty} a_j = \lim_{j\to\infty} b_j$ of the double sequence (8) is called the *arithmetic-geometric mean* of a_0 and b_0 and is denoted as $m(a_0, b_0)$. It can be expressed in terms of a complete elliptic integral

$$I(a,b) = \int_0^\infty \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}}. \tag{46}$$

Gauss' crucial observation for establishing the relation between $m(a,b)$ and $I(a,b)$ is that $I(a,b)$ is invariant under the transformation $(a,b) \mapsto (a_1,b_1) = (\sqrt{ab}, \frac{a+b}{2})$.

We see this by the substitution $t = \frac{1}{2}(x - \frac{ab}{x})$. As x goes from 0 to ∞, the variable t increases from $-\infty$ to ∞. Moreover,

$$dt = \frac{x^2+ab}{2x^2}\,dx, \quad t^2 + \left(\frac{a+b}{2}\right)^2 = \frac{(x^2+a^2)(x^2+b^2)}{4x^2}, \quad t^2+ab = \frac{(x^2+ab)}{4x^2}.$$

Hence,

$$I(a_1,b_1) = \frac{1}{2}\int_{-\infty}^{\infty} \frac{dt}{\sqrt{(a_1^2+t^2)(b_1^2+t^2)}} = \int_0^{\infty} \frac{dx}{\sqrt{(a^2+x^2)(b^2+x^2)}} = I(a,b) \quad (47)$$

yields the invariance.

Let $m = m(a,b)$, and set $a_0 = a$, $b_0 = b$. By induction it follows that $I(a_0,b_0) = I(a_j,b_j)$ for all j, and by continuity $I(a_0,b_0) = I(m,m)$. Obviously, $I(m,m) = \int_0^{\infty} \frac{dt}{m^2+t^2} = \frac{\pi}{2m}$, and we conclude that

$$m(a,b) = \frac{\pi}{2I(a,b)}.$$

The *elliptic integrals* are defined by $\mathbf{K}'(k) := I(k,1)$ and $\mathbf{K}'(k) = \mathbf{K}(k')$. Here the module k and the complementary module k' are related by $k^2 + (k')^2 = 1$. A scaling argument shows that

$$I(a,b) = b^{-1}\mathbf{K}'(a/b) \qquad \text{for } 0 < a \le b. \quad (48)$$

Since the arithmetic-geometric mean of 1 and k lies between the arithmetic mean and the geometric mean, we get an estimate that is good for $k \approx 1$.

$$\frac{\pi}{1+k} \le \mathbf{K}'(k) \le \frac{\pi}{2\sqrt{k}}. \quad (49)$$

An estimate that is good for small k is more involved:

$$\begin{aligned}\mathbf{K}'(k) &= 2\int_0^{\sqrt{k}} \frac{dt}{\sqrt{(1+t^2)(k^2+t^2)}} \le 2\int_0^{\sqrt{k}} \frac{dt}{\sqrt{k^2+t^2}} = 2\int_0^{1/\sqrt{k}} \frac{dt}{\sqrt{1+t^2}} \\ &= 2\log\left(\sqrt{\frac{1}{k}} + \sqrt{\frac{1}{k}+1}\right) \le \log\left(4\left(\frac{1}{k}+\frac{1}{2}\right)\right). \end{aligned} \quad (50)$$

As a consequence, we have $(\pi/2)\mathbf{K}'(k)/\mathbf{K}(k) \le \log(\frac{4}{k}+2)$ and

$$\frac{1}{k} \ge \frac{1}{4}\exp\left[\frac{\pi\mathbf{K}'(k)}{2\mathbf{K}(k)}\right] - \frac{1}{2}. \quad (51)$$

Lemma 9.1. *Let λ_0, $\lambda_{-1},\lambda_{-2},\ldots$ be a sequence generated by the Landen transformation and $\kappa_0 := 1/\lambda_0$. Then*

$$\lambda_{-j} \geq \frac{1}{4} \exp\left[2^j \frac{\pi \mathbf{K}'(\kappa_0)}{2\mathbf{K}(\kappa_0)}\right]. \tag{52}$$

Proof. Let $0 < \kappa < 1$ and $\kappa_1 = \frac{2\sqrt{\kappa}}{1+\kappa}$. Note that

$$\kappa_1{}' = \frac{1-\kappa}{1+\kappa}. \tag{53}$$

From (47) and (48) it follows that

$$\mathbf{K}'(\kappa) = I(\kappa, 1) = I\left(\sqrt{\kappa}, \frac{1+\kappa}{2}\right) = \frac{2}{1+\kappa}\mathbf{K}'\left(\frac{2\sqrt{\kappa}}{1+\kappa}\right) = \frac{2}{1+\kappa}\mathbf{K}'(\kappa_1)$$

and with the two means of $1-\kappa$ and $1+\kappa$:

$$\begin{aligned}\mathbf{K}(\kappa_1) = I(\kappa_1', 1) = I\left(\frac{1-\kappa}{1+\kappa}, 1\right) &= (1+\kappa) I(1-\kappa, 1+\kappa) \\ &= (1+\kappa) I\left(\sqrt{1-\kappa^2}, 1\right) = (1+\kappa)\mathbf{K}(\kappa).\end{aligned}$$

Hence,

$$\frac{\mathbf{K}'(\kappa)}{\mathbf{K}(\kappa)} = 2\frac{\mathbf{K}'(\kappa_1)}{\mathbf{K}(\kappa_1)} \tag{54}$$

and by induction $\mathbf{K}'(\kappa_{-j})/\mathbf{K}(\kappa_{-j}) = 2^j \mathbf{K}'/(\kappa_0)\mathbf{K}(\kappa_0)$. Now (51) yields the preliminary estimate

$$\lambda_{-j} \geq \frac{1}{4} \exp\left[2^j \frac{\pi \mathbf{K}'(\kappa_0)}{2\mathbf{K}(\kappa_0)}\right] - \frac{1}{2}.$$

If we apply the estimate to $j+1$ instead of j, return to j noting that $\sqrt{A^2-2} \geq A + 2/A$, we see that we can drop the extra term $1/2$, and the proof is complete.

10 A direct approach to the infinite interval

There is also a one-step proof for the special function $1/x$. It is based on a result of Vjačeslavov [29] which in turn requires complicated evaluations of some special integrals; see also [25]. Since constructions on finite intervals are circumvented, it supports the argument that the non-optimal bound (30) is not induced by the limit process with large intervals.

Given $\alpha > 0$ and $n \in \mathbb{N}$, there exists a polynomial p of degree n with n zeros in $[0,1]$ such that

$$\left| x^\alpha \frac{p(x)}{p(-x)} \right| \leq c_0(\alpha) \cdot \mathrm{e}^{-\pi\sqrt{\alpha n}} \qquad \text{for } 0 \leq x \leq 1.$$

Let p be the polynomial for $\alpha = 1/4$ as stated above. Since $p(\bar{z}) = \bar{p}(z)$, it follows that $p(z)/p(-z) = 1$ for $\Re e\, z = 0$ and

$$\left| \frac{p(z^2)}{p(-z^2)} \right| = 1 \qquad \text{for } \Re e\, z = |\Im m\, z| \geq 0. \tag{55}$$

We consider $P(z) := p^2(1/z^2)$ on the sector $\mathscr{S} := \{z \in \mathbb{C} : \ |\arg z| \leq \pi/4\}$. By construction P has n double zeros in $[1,\infty)$, and from (55) it follows that

$$\left| \frac{P(z)}{P(-z)} \right| = 1 \quad \text{for } z \in \partial\mathscr{S}, \qquad \frac{P(x)}{xP(-x)} \leq \left(c_0(1/4) \cdot \mathrm{e}^{-\pi\sqrt{n/4}} \right)^2 \quad \text{for } x \geq 1.$$

Now let u_n be the exponential sum interpolating $1/x$ and its first derivative at the (double) zeros of P. Since $1/x - u_n$ has no more zeros than the specified ones, we have $u_n(x) \leq 1/x$ for $x \geq 0$. Hence,

$$|u_n(z)| \leq u_n(\Re e\, z) \leq 1/\Re e\, z \leq \sqrt{2}/|z| \qquad \text{on the boundary of } \mathscr{S}.$$

Arguing as in Section 2, we introduce the auxiliary function $g(z) := (\frac{1}{z} - u_n(z)) z \frac{P(-z)}{P(z)}$. We know that $|g(z)| \leq 1 + \sqrt{2}$ holds on the boundary of $\mathscr{S}$ and therefore in $\mathscr{S}$. Finally,

$$\left| \frac{1}{z} - u_n(z) \right| = \left| g(z) \frac{P(z)}{zP(-z)} \right| \leq (1+\sqrt{2})\, c_0^2(1/4) \mathrm{e}^{-\pi\sqrt{n}}.$$

11 Sinc quadrature derived approximations

The sinc quadrature discussed in this section approximates integrals of the form $\int_{-\infty}^{\infty} F(t)\mathrm{d}t$ under certain conditions on F. In particular, we are interested in functions that depend on a further parameter x like $F(t,x) = F_1(t)\exp[F_2(t)x]$, and the evaluation at a quadrature point $t = \tau_\nu$ yields $\alpha_\nu \mathrm{e}^{-t_\nu x}$ with $\alpha_\nu := F_1(\tau_\nu)$ and $t_\nu := F_2(\tau_\nu)$. Therefore the sinc quadrature applied to

$$f(x) := \int_{-\infty}^{\infty} F(t,x)\mathrm{d}t \tag{56}$$

is a popular method to obtain exponential sums even with guaranteed upper bounds [8, 18]. Concerning literature we refer to the monograph of Stenger [26] or [14, Anhang D]. Next, we introduce the sinc quadrature rule $T(F,h)$, its truncated form $T_N(f,h)$, and its application to $1/x$ (the application to $1/\sqrt{x}$ is quite similar).

The sinc function $\mathrm{sinc}(x) := \frac{\sin(\pi x)}{\pi x}$ is an analytic functions with the value one at $x = 0$ and zero at $x \in \mathbb{Z}\backslash\{0\}$. Given a step size $h > 0$, the family of functions

$$S_{k,h}(x) := \mathrm{sinc}(\frac{x}{h} - k) \qquad (k \in \mathbb{Z}),$$

satisfies $S_{k,h}(\nu h) = \delta_{k\nu}$ ($\delta_{k\nu}$: Kronecker symbol). Let $F \in C(\mathbb{R})$ decay sufficiently fast for $x \to \pm\infty$. Then the sum

$$F_h(x) := \sum_{k\in\mathbb{Z}} F(kh) S_{k,h}(x)$$

converges and interpolates F at all grid points $x = \nu h \in h\mathbb{Z}$. This fact suggests the interpolatory quadrature rule $\int_{-\infty}^{\infty} F(t)\mathrm{d}t \approx \int_{-\infty}^{\infty} F_h(t)\mathrm{d}t$. Since $\int_{-\infty}^{\infty} \mathrm{sinc}(t)\mathrm{d}t = 1$, the right-hand side leads to the *sinc quadrature rule*

$$T(F,h) := h \sum_{k\in\mathbb{Z}} F(kh)$$

for $\int_{-\infty}^{\infty} F(t)\mathrm{d}t$, and $T(f,h)$ can be considered as the infinite trapezoidal rule. The next step is the truncation (cut-off) of the infinite series to the finite sum

$$T_N(f,h) := h \sum_{k=-N}^{N} F(kh).$$

For convenience, we will use N as truncation parameter. It will be related to the number n of terms in (2) by $n = 2N+1$.

Before we discuss the quadrature error of $T(F,h)$, we show how to get exponential sums from this approach. As example we consider the representation of $\frac{1}{x}$ by $\int_0^{\infty} \exp(-xs)\mathrm{d}s$. Let $s = \varphi(t)$ be any differentiable transformation of $(-\infty,\infty)$ onto $(0,\infty)$. This yields the integral

$$\frac{1}{x} = \int_{-\infty}^{\infty} \exp(-x\varphi(t))\varphi'(t)\mathrm{d}t \tag{57}$$

to which the sinc quadrature $T_N(f,h)$ can be applied:

$$\frac{1}{x} \approx T_N(\exp(-x\varphi(\cdot))\varphi'(\cdot),h) = h \sum_{k=-N}^{N} \varphi'(kh)\mathrm{e}^{-x\varphi(kh)}.$$

Obviously, the right-hand side is the exponential sum (2) with $\alpha_\nu = h\varphi'((\nu-1-N)h)$ and $t_\nu = \varphi((\nu-1-N)h)$ for $1 \le \nu \le n = 2N+1$. Note that different transformations φ yield different exponential sums.

A good candidate for φ is $\varphi(t) := \exp(t)$ leading to[4]

$$\frac{1}{x} = \int_{-\infty}^{\infty} \exp(-x\mathrm{e}^t)\mathrm{e}^t\mathrm{d}t. \tag{58}$$

The exponential behaviour $t_\nu = const \cdot \mathrm{e}^{\nu h}$ of the coefficients is sometimes used as an explicit ansatz for (2). Indeed, the coefficients t_ν of the best approximations from

[4] Also $\varphi(t) = \exp(At)$ for $A > 0$ is possible. The reader may try to analyse the influence of A to the error analysis.

Section 4 lead to similar quotients $t_{\nu+1}/t_\nu$ for ν in the middle range with deviations for ν close to 1 and n.

Next we study the quadrature error of T_N. It is the sum of $\int_{-\infty}^{\infty} F(t)\mathrm{d}t - T(F,h)$ and $T(F,h) - T_N(F,h)$. The quadrature error of the sinc quadrature

$$\eta(F,h) := \left| \int_{-\infty}^{\infty} F(t)\mathrm{d}t - T(F,h) \right|$$

tends to zero as $h \to 0$. The truncation error $|T(F,h) - T_N(F,h)|$ vanishes as $N \to \infty$. Both discretisation parameters h and N will be related in such a way that both errors are (asymptotically) equal.

The analysis of $\eta(F,h)$ requires holomorphy of F in a stripe. Let

$$D_d := \{z = x + iy : x \in \mathbb{R}, |y| < d\} \subset \mathbb{C}$$

be the open stripe along the real axis with width $2d$. The function F is assumed to be holomorphically extendable to D_d such that the L^1 integral

$$\|F\|_{D_d} := \int_{\mathbb{R}} \{|F(x+id| + |F(x-id|\}\,\mathrm{d}x$$

over the boundary of D_d exists and is finite. As proved in [26, p. 144 f] the error $\eta(F,h)$ of the infinite quadrature rule $T(F,h)$ is bounded by

$$\eta(F,h) \le \|F\|_{D_d} \exp[-2\pi d/h]. \tag{59}$$

The truncation error $|T(F,h) - T_N(F,h)|$ equals $h\left|\sum_{|k|>N} F(kh)\right|$ and depends on the decay of F as $x \to \pm\infty$ (note that this concerns only the behaviour on the real axis, not in the stripe D_d).
If, for instance, $|F(t)| \le c \cdot \mathrm{e}^{-\alpha|t|}$ holds, then $|T(F,h) - T_N(F,h)| \le (2c/\alpha)\mathrm{e}^{-\alpha Nh}$ follows. In this case, the error $\eta(F,h) = \mathcal{O}(\mathrm{e}^{-2\pi d/h})$ and the truncation error $\mathcal{O}(\mathrm{e}^{-\alpha Nh})$ are asymptotically equal if $-2\pi d/h = -\alpha Nh$, i.e., $h = \sqrt{2\alpha\pi dN}$. This leads to the estimate of the total error

$$\left| \int_{-\infty}^{\infty} F(t)\mathrm{d}t - T_N(F,h) \right| \le \left(\|F\|_{D_d} + \frac{2c}{\alpha} \right) \exp[-\sqrt{2\pi d/(\alpha N)}] \quad \text{for } h = \sqrt{2\alpha\pi dN}. \tag{60}$$

So far, F is a function of t only and the integral $\int_{-\infty}^{\infty} F(t)\mathrm{d}t$ is a real number. Now we replace F by $F(t,x)$ as in (56) such that the integral defines a function $f : D \to \mathbb{R}$ on a domain D. The error estimate (60) is still correct, but it holds only pointwise for $x \in D$. We note that $\|F\|_{D_d}$ becomes a function $\|F(\cdot,x)\|_{D_d}$ of x, and even the width d of the stripe may change with x. Moreover, if decay inequality $|F(t,x)| \le c \cdot \mathrm{e}^{-\alpha|t|}$ holds with x-dependent factors $c(x)$ and $\alpha(x)$, also these quantities in (60) become variable. We have to take care that the estimate (60) (with $\|F(\cdot,x)\|_{D_d}$ replaced by an upper bound) is *uniform* in $x \in D$, and the error $|f(x) - T_N(F(\cdot,x),h)|$ is uniform too.

In the following, we will simply write $F(t)$ instead of $F(t,x)$, i.e., $F(t)$ is understood to be function-valued.

We apply this strategy to the error estimation of the integral in (58). The integrand $F(t) = \exp(-x\mathrm{e}^t)\mathrm{e}^t$ is an entire function in t, and to obtain a bounded norm $\|F\|_{D_d}$ we choose $d < \pi/2$. Then $|F(t \pm id)| = \exp(-x\mathrm{e}^t \cos(d))\mathrm{e}^t$ implies $\|F\|_{D_d} = \frac{1}{x\cos(d)}$. Inequality (59) yields

$$|\eta(F,h)| \leq \frac{\exp(-2\pi d/h)}{x\cos(d)} \qquad \text{for all } 0 < d < \pi/2.$$

Optimisation with respect to d yields $d = \arctan(2\pi/h) < \pi/2$ and

$$|\eta(F,h)| \leq \frac{\sqrt{1+(2\pi/h)^2}}{x} \exp\Big(\frac{-2\pi \arctan(2\pi/h)}{h}\Big).$$

Concerning $|T(F,h) - T_N(F,h)| = h\left|\sum_{|k|>N} F(kh)\right|$ notice the different behaviour of $F(kh)$ for $k \to \infty$ and $k \to -\infty$. As $k \to -\infty$, the factor e^{kh} describes a uniformly exponential decay, while $\exp(-x\mathrm{e}^{kh}) \to 1$. For $k \to +\infty$, the factor $\exp(-x\mathrm{e}^{kh})$ shows a doubly exponential behaviour which, however, depends on the value of x. The precise asymptotics are given by

$$\begin{aligned} h\left|\sum_{k<-N} F(kh)\right| &\leq h \sum_{k<-N} \mathrm{e}^{kh} \leq \int_{-\infty}^{-Nh} \exp(t)\mathrm{d}t = \mathrm{e}^{-Nh}, \\ h\left|\sum_{k>+N} F(kh)\right| &\leq h \sum_{k>N} \exp(-x\mathrm{e}^{kh})\mathrm{e}^{kh} \\ &\leq \int_{Nh}^{\infty} \exp(-x\mathrm{e}^t)\mathrm{e}^t \mathrm{d}t = \frac{1}{x}\exp(-x\mathrm{e}^{Nh}). \end{aligned}$$

Here we assume $x\mathrm{e}^{Nh} \geq 1$ for the second inequality, so that the function $\exp(-x\mathrm{e}^t)\mathrm{e}^t$ is monotonously decreasing in $[Nh,\infty)$. Altogether, we get the following error estimate between the integral (58) and the exponential sum $T_N(F,h)$

$$\begin{aligned} \left|\frac{1}{x} - T_N(F,h)\right| \leq{}& \frac{\sqrt{1+(2\pi/h)^2}}{x} \exp\left(\frac{-2\pi \arctan(2\pi/h)}{h}\right) \\ &+ \mathrm{e}^{-Nh} + \frac{1}{x}\exp(-x\mathrm{e}^{Nh}). \end{aligned} \tag{61}$$

To simplify the analysis, we assume $x \in [1,R]$, which implies the relation

$$\frac{1}{x}\exp(-x\mathrm{e}^{Nh}) \leq \mathrm{e}^{-Nh-1},$$

i.e., the last term in (61) is smaller than the second one. Further, we use the asymptotic behaviour $2\pi \arctan(2\pi/h) = \pi^2 - h + \mathscr{O}(h^3)$ to show that

$$\exp\frac{-2\pi\arctan(2\pi/h)}{h} = \mathcal{O}(\exp(-\pi^2/h)).$$

Therefore, the right-hand side in (61) becomes $\mathcal{O}(\sqrt{1+(2\pi/h)^2}\exp(-\pi^2/h)) + \mathcal{O}(\exp(-Nh))$. The asymptotically best choice of h is $h=\pi/\sqrt{N}$ which leads to equal exponents: $-\pi^2/h = -Nh = -\pi\sqrt{N}$. Inserting this choice of h, we get the uniform estimate

$$\left|\frac{1}{x} - T_N(F,h)\right| \le \left(\mathcal{O}(1)+2\sqrt{N}\right)\mathrm{e}^{-\pi\sqrt{N}} \le \mathcal{O}\left(\sqrt{n}\exp\left[-\frac{\pi}{\sqrt{2}}\sqrt{n}\right]\right) \quad \text{for all } x \ge 1, \tag{62}$$

where the last expression uses the number $n = 2N+1$ of terms in $T_N(F,h)$. The exponential behaviour $\exp[-\pi\sqrt{n/2}]$ is not as good as $\exp[-\pi\sqrt{n}]$ from (30).

Although we get the same behaviour $\exp[-const\sqrt{n}]$ of the error as in (30), the reason is a different one. In the case of the best approximation, we could show an error behaviour $\exp[-const\cdot n]$ for finite intervals $[1,R]$, whereas $\exp[-\pi\sqrt{n}]$ was caused by the unboundedness of $[1,\infty)$. The $\exp[-\pi\sqrt{n/2}]$ behaviour of the sinc quadrature is independent of the choice $x\in[1,R]$, R finite, or $x\in[1,\infty)$. Even if we restrict x to a single point x_0, the error is like in (62). The reason for $\exp[-const\sqrt{n}]$ in the sinc case is due to the fact that we have to equalise the exponents in $\mathcal{O}(\exp(-const/h)) + \mathcal{O}(\exp(-const\cdot Nh))$. The error $\mathcal{O}(\exp(-const/h))$ of the infinite sinc quadrature can hardly be improved (see (59)), but the truncation error $\mathcal{O}(\exp(-const\cdot Nh))$ of $|T(F,h)-T_N(F,h)|$ depends on the decay behaviour of F. If, for instance, $|F(t)| \le c\cdot\exp(-\alpha|t|^\gamma)$ holds for some $\gamma>1$, this faster decay yields the smaller truncation error $\mathcal{O}(\exp(-\alpha(Nh)^\gamma))$. Finally, $h=\mathcal{O}(N^{-\gamma/(\gamma+1)})$ leads to the total error $\exp\left[-const\cdot n^{\gamma/(\gamma+1)}\right]$. For large γ, the exponent comes close to $-const\cdot n$.

An even better decay behaviour is the doubly exponential decrease $|F(t)| \le c_1\cdot\exp(-c_2\mathrm{e}^{c_3|t|})$. In this case, the total error can be estimated by $\mathcal{O}\left(\|F\|_{D_d}\exp\left(\frac{-2\pi dc_3N}{\log(2\pi dc_3N)}\right)\right)$ (cf. [14, Satz D.4.3]). To obtain a doubly exponential decay, one can follow the following lines: Start with an integral $\int_{-\infty}^{\infty}F(t)\mathrm{d}t$, where F has the usual exponential asymptotic $|F(t)| \le c\cdot\exp(-\alpha|t|)$. Then apply the transformation $t=\sinh\tau$. The new integral is $\int_{-\infty}^{\infty}G(\tau)\mathrm{d}\tau$ with the doubly exponential integrand $G(\tau)=F(\sinh\tau)\cosh\tau$. The drawback is that one must ensure that G is still holomorphic in a stripe D_d and that $\|F\|_{D_d}$ is finite. The mentioned transformation applied to $F(t)=\exp(-x\mathrm{e}^t)\mathrm{e}^t$ from above does not succeed. For any $d>0$ the real part of $\mathrm{e}^{\sinh\tau}$ may be negative in D_d and, because of the exponentially increasing function $\exp(-x\mathrm{e}^{\sinh(\tau+id)})$, the integral with respect to $\tau\in\mathbb{R}$ does not exist, i.e. $\|F\|_{D_d}=\infty$.

A possible approach is to replace the first transformation $\varphi(t) := \exp(t) : [0,\infty) \to (-\infty,\infty)$ in (57) by $\varphi(t) := \log(1+\exp(\sinh t))$, which yields

$$\frac{1}{x} = \int_{-\infty}^{\infty} \exp\left(-x\log\left(1+\mathrm{e}^{t}\right)\right) \frac{\mathrm{d}t}{1+\mathrm{e}^{-t}}; \tag{63}$$

cf. [14, §D.4.3.2]. The integrand $F = \exp\left(-x\log\left(1+\mathrm{e}^{t}\right)\right)/\left(1+\mathrm{e}^{-t}\right)$ in (63) behaves simply exponential for $t \to \infty$ and $t \to -\infty$. Thanks to this property, the second transformation $t = \sinh\tau$ succeeds in providing an integrand G which is holomorphic in D_d for $d < \pi/2$ with finite norm $\|G\|_{D_d}$. However, pointwise finiteness $\|G(\cdot,x)\|_{D_d} < \infty$ is not enough. It turns out that in general $\|G(\cdot,x)\|_{D_d} \leq \mathcal{O}(\mathrm{e}^{x})$, which destroys the error estimates. For $x \in [1,R]$ one has to reduce the stripe D_d to the width $d = d(R) := \mathcal{O}(1/\log R)$. Then involved estimates show that the error $\left|\frac{1}{x} - T_N(G,h)\right|$ in $1 \leq x \leq R$ is bounded by

$$\mathcal{O}\left(\exp\left(-\frac{2\pi d(R)N}{\log(2\pi d(R)N}\right)\right) \qquad \text{with} \quad d(R) := \mathcal{O}(1/\log R)$$

(cf. [14, §D.4.3.2]). Since a detailed analysis shows $d(R) = \frac{\pi}{2}\frac{1}{\log(3R)} - \mathcal{O}(\log^{-2}(3R))$, this estimate is almost of the form $\exp(-Cn)$ with $C := \frac{2\pi^2}{\log(3R)\log(2\pi^2 n/\log(3R))}$ and may be compared with $\exp(-C^* n)$ from (29) with $C^* = \frac{\pi^2}{\log(8R)}$. Obviously, $C < C^*$ holds for sufficiently large n, but even for small n, $C < C^*$ holds, e.g., for $R \leq 1600$ ($n = 4$), $R \leq 24700$ ($n = 5$), $R \leq 3.7_{10}5$ ($n = 6$), $R \leq 5.5_{10}6$ ($n = 7$), and $R \leq 8.1_{10}7$ ($n = 8$). The latter bounds of R are (much) larger than the value $R = \frac{1}{8}\exp[\pi\sqrt{n}]$ introduced in the line before (30). Hence, for values of R for which the best approximation on $[1,R]$ is not already a best approximation for $[1,\infty)$, (29) gives a better result than the sinc estimate from above.

References

1. J. Almlöf: *Elimination of energy denominators in Møller-Plesset perturbation theory by a Laplace transform approach*. Chem. Phys. Lett. **176** (1991), 319–320
2. P.Y. Ayala and G. Scuseria: *Linear scaling second-order Moller-Plesset theory in the atomic orbital basis for large molecular systems*. J. Chem. Phys. **110** (1999), 3660
3. J.M. Borwein and P.B. Borwein: *Pi and the AGM*. John Wiley & Sons, 1987.
4. D. Braess: *Nonlinear Approximation Theory*. Springer-Verlag, Berlin, 1986.
5. D. Braess: *On the conjecture of Meinardus on the rational approximation of e^x*. J. Approximation Theory **36** (1982), 317–320.
6. D. Braess: *Asymptotics for the approximation of wave functions by exponential sums*. J. Approximation Theory **83** (1995), 93–103.
7. D. Braess and W. Hackbusch: *Approximation of 1/x by exponential sums in [1,∞)*. IMA J. Numer. Anal. **25** (2005), 685–697
8. E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday: *Computational quantum chemistry: a primer*. In: 'Handbook of Numerical Analysis', **X**, pp. 3–270, C. Le Bris (ed.), Elsevier, Amsterdam 2003
9. A.F. Izmaylov and G.E. Scuseria: *Resolution of the identity atomic orbital Laplace transformed second order Møller–Plesset theory for nonconducting periodic systems*. Phys. Chem. Chem. Phys. **10** (2008), 3421–3429

10. Y. Jung, R.C. Lochan, A.D. Dutoi, and M. Head-Gordon: *Scaled opposite-spin second order Møller–Plesset correlation energy: An economical electronic structure method*. J. Chem. Phys. **121** (2004), 9793
11. L. Grasedyck: *Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure*. Computing, **72** (2004), 247–265
12. W. Hackbusch: *Approximation of* $1/\|x-y\|$ *by exponentials for wavelet applications*. Computing, **76** (2006), 359–366
13. W. Hackbusch: *Efficient convolution with the Newton potential in d dimensions*. Numer. Math. **110** (2008), 449–489
14. W. Hackbusch: *Hierarchische Matrizen – Algorithmen und Analysis*. Springer-Verlag, Berlin (to appear in 2009)
15. M. Häser and J. Almlöf: *Laplace transform techniques in Møller–Plesset perturbation theory*. J. Chem. Phys. **96** (1992), 489
16. M. Häser: *Møller-Plesset (MP2) perturbation theory for large molecules*. Theor. Chim. Acta **87** (1993), 147–173
17. M. Kobayashi and H. Nakai: *Implementation of Surján's density matrix formulae for calculating second-order Møller–Plesset energy*. Chem. Phys. Lett. **420** (2006), 250–255
18. W. Kutzelnigg: *Theory of the expansion of wave functions in a Gaussian basis*. Int. J. of Quantum Chemistry **51** (1994), 447–463.
19. D.S. Lambrecht, B. Doser, and C. Ochsenfeld: *Rigorous integral screening for electron correlation methods*. J. Chem. Phys. **123** (2005), 184102
20. D.J. Newman: *Rational approximation to* e^x. J. Approximation Theory **27** (1979), 234–235
21. E.J. Remez: *Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation*. Compt. Rend. Acad. Sc. **198** (1934), 2063–2065
22. R. Rutishauser: *Betrachtungen zur Quadratwurzeliteration*. Monatshefte Math. **67** (1963), 452–464
23. D. Kats, D. Usvyat and M. Schütz: *On the use of the Laplace transform in local correlation methods*. Phys. Chem. Chem. Phys., 2008, DOI: 10.1039/b802993h
24. S. Schweizer, B. Doser, and C. Ochsenfeld: *An atomic orbital-based reformulation of energy gradients in second-order Møller–Plesset perturbation theory*. J. Chem. Phys. **128** (2008), 154101
25. H.R. Stahl: *Best uniform rational approximation of* x^α *on* $[0,1]$. Acta Math. **190** (2003), 241–306.
26. F. Stenger: *Numerical Methods Based of Sinc and Analytic Functions*. Springer-Verlag, New York 1993
27. P.R. Surján: *The MP2 energy as a functional of the Hartree–Fock density matrix*. Chem. Phys. Lett. **406** (2005), 318–320
28. A. Takatsuka, S. Ten-no, and W. Hackbusch: *Minimax approximation for the decomposition of energy denominators in Laplace-transformed Møller-Plesset perturbation theories*. J. Chem. Phys. **129** (2008), 044112
29. N.S. Vjačeslavov: *On the least deviation of the function* sign$x$ *and its primitives from the rational functions in the* L_p *metrics,* $0 < p < \infty$. Math. USSR Sbornik **32** (1977), 19–31
30. A.K. Wilson and J. Almlöf: *Møller-Plesset correlation energies in a localized orbital basis using a Laplace transform technique*. Theor. Chim. Acta **95** (1997), 49–62
31. Webpages `www.mis.mpg.de/scicomp/EXP_SUM/1_x/` and `ldots/1_sqrtx/` with explanations in `.../1_x/tabelle` and `.../1_sqrtx/tabelle`
32. E.I. Zolotarov: *Application of elliptic functions to questions of functions deviating least and most from zero* (Russian). Zap. Imp. Akad. Nauk (1877). St. Petersburg 30 no. 5; reprinted in collected works II, pp. 1–59. Izdat, Akad. Nauk SSSR, Moscow 1932.

Adaptive and anisotropic piecewise polynomial approximation

Albert Cohen and Jean-Marie Mirebeau

Abstract We survey the main results of approximation theory for adaptive piecewise polynomial functions. In such methods, the partition on which the piecewise polynomial approximation is defined is not fixed in advance, but adapted to the given function f which is approximated. We focus our discussion on (i) the properties that describe an optimal partition for f, (ii) the smoothness properties of f that govern the rate of convergence of the approximation in the L^p-norms, and (iii) fast refinement algorithms that generate near optimal partitions. While these results constitute a fairly established theory in the univariate case and in the multivariate case when dealing with elements of isotropic shape, the approximation theory for adaptive and anisotropic elements is still building up. We put a particular emphasis on some recent results obtained in this direction.

1 Introduction

1.1 Piecewise polynomial approximation

Approximation by piecewise polynomial functions is a procedure that occurs in numerous applications. In some of them such as terrain data simplification or image compression, the function f to be approximated might be fully known, while it might be only partially known or fully unknown in other applications such as denoising, statistical learning or in the finite element discretization of PDE's. In all

Albert Cohen
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France, e-mail: cohen@ann.jussieu.fr

Jean-Marie Mirebeau
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France, e-mail: mirebeau@ann.jussieu.fr

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_4,

these applications, one usually makes the distinction between *uniform* and *adaptive* approximation. In the uniform case, the domain of interest is decomposed into a partition where all elements have comparable shape and size, while these attributes are allowed to vary strongly in the adaptive case. The partition may therefore be adapted to the local properties of f, with the objective of optimizing the trade-off between accuracy and complexity of the approximation. This chapter is concerned with the following fundamental questions:

- Which mathematical properties describe an optimally adapted partition for a given function f ?
- For such optimally adapted partitions, what smoothness properties of f govern the convergence properties of the corresponding piecewise polynomial approximations ?
- Can one construct optimally adapted partitions for a given function f by a fast algorithm ?

For a given bounded domain $\Omega \subset \mathbb{R}^{\mathrm{d}}$ and a fixed integer $m > 0$, we associate to any partition $\mathscr{T}$ of Ω the space

$$V_{\mathscr{T}} := \{f \text{ s.t. } f_{|T} \in \mathbb{P}_{\mathrm{m}-1},\, \mathrm{T} \in \mathscr{T}\}$$

of piecewise polynomial functions of total degree $m-1$ over $\mathscr{T}$. The dimension of this space measures the complexity of a function $g \in V_{\mathscr{T}}$. It is proportional to the cardinality of the partition:

$$\dim(V_{\mathscr{T}}) := C_{m,d}\#(\mathscr{T}), \text{ with } C_{m,d} := \dim(\mathbb{P}_{\mathrm{m}-1}) = \binom{\mathrm{m}+\mathrm{d}-1}{\mathrm{d}}.$$

In order to describe how accurately a given function f may be described by piecewise polynomial functions of a prescribed complexity, it is therefore natural to introduce the error of best approximation in a given norm $\|\cdot\|_X$ which is defined as

$$\sigma_N(f)_X := \inf_{\#(\mathscr{T})\leq N} \min_{g\in V_{\mathscr{T}}} \|f-g\|_X.$$

This object of study is too vague if we do not make some basic assumptions that limitate the set of partitions which may be considered. We therefore restrict the definition of the above infimum to a class $\mathscr{A}_N$ of "admissible partitions" of complexity at most N. The approximation to f is therefore searched in the set

$$\Sigma_N := \cup_{\mathscr{T}\in\mathscr{A}_N} V_{\mathscr{T}},$$

and the error of best approximation is now defined as

$$\sigma_N(f)_X := \inf_{g\in\Sigma_N} \|f-g\|_X = \inf_{\mathscr{T}\in\mathscr{A}_N} \inf_{g\in V_{\mathscr{T}}} \|f-g\|_X.$$

The assumptions which define the class $\mathscr{A}_N$ are usually of the following type:

1. The elementary *geometry* of the elements of $\mathscr{T}$. The typical examples that are considered in this chapter are: intervals when $d = 1$, triangles or rectangles when $d = 2$, simplices when $d > 2$.
2. Restrictions on the *regularity* of the partition, in the sense of the relative size and shape of the elements that constitute the partition $\mathscr{T}$.
3. Restrictions on the *conformity* of the partition, which impose that each face of an element T is common to at most one adjacent element T'.

The conformity restriction is critical when imposing global continuity or higher smoothness properties in the definition of $V_{\mathscr{T}}$, and if one wants to measure the error in some smooth norm. In this survey, we limitate our interest to the approximation error measured in $X = L^p$. We therefore do not impose any global smoothness property on the space $V_{\mathscr{T}}$ and ignore the conformity requirement.

Throughout this chapter, we use the notation

$$e_{m,\mathscr{T}}(f)_p := \min_{g \in V_{\mathscr{T}}} \|f - g\|_{L^p},$$

to denote the L^p approximation error in the space $V_{\mathscr{T}}$ and

$$\sigma_N(f)_p := \sigma_N(f)_{L^p} = \inf_{g \in \Sigma_N} \|f - g\|_{L^p} = \inf_{\mathscr{T} \in \mathscr{A}_N} e_{m,\mathscr{T}}(f)_p.$$

If $T \in \mathscr{T}$ is an element and f is a function defined on Ω, we denote by

$$e_{m,T}(f)_p := \min_{\pi \in \mathbb{P}_{m-1}} \|f - \pi\|_{L^p(T)},$$

the local approximation error. We thus have

$$e_{m,\mathscr{T}}(f)_p = \Big(\sum_{T \in \mathscr{T}} e_{m,T}(f)_p^p \Big)^{1/p},$$

when $p < \infty$ and

$$e_{m,\mathscr{T}}(f)_\infty = \max_{T \in \mathscr{T}} e_{m,T}(f)_\infty.$$

The norm $\|f\|_{L^p}$ without precision on the domain stands for $\|f\|_{L^p(\Omega)}$ where Ω is the full domain where f is defined.

1.2 From uniform to adaptive approximation

Concerning the restrictions ont the regularity of the partitions, three situations should be distinguished:

1. *Quasi-uniform partitions*: all elements have approximately the same size. This may be expressed by a restriction of the type

$$C_1 N^{-1/d} \leq \rho_T \leq h_T \leq C_2 N^{-1/d}, \tag{1}$$

for all $T \in \mathscr{T}$ with $\mathscr{T} \in \mathscr{A}_N$, where $0 < C_1 \leq C_2$ are constants independent of N, and where h_T and ρ_T respectively denote the diameters of T and of it largest inscribed disc.

2. *Adaptive isotropic partitions*: elements may have arbitrarily different size but their aspect ratio is controlled by a restriction of the type

$$\frac{h_T}{\rho_T} \leq C, \tag{2}$$

for all $T \in \mathscr{T}$ with $\mathscr{T} \in \mathscr{A}_N$, where $C > 1$ is independent of N.

3. *Adaptive anisotropic partitions*: element may have arbitrarily different size and aspect ratio, i.e. no restriction is made on h_T and ρ_T.

A classical result states that if a function f belongs to the Sobolev space $W^{m,p}(\Omega)$ the L^p error of approximation by piecewise polynomial of degree m on a given partition satisfies the estimate

$$e_{m,\mathscr{T}}(f)_p \leq C h^m |f|_{W^{m,p}}, \tag{3}$$

where $h := \max_{T\in\mathscr{T}} h_T$ is the maximal mesh-size, $|f|_{W^{m,p}} := \left(\sum_{|\alpha|=m} \|\partial^\alpha f\|_{L^p}^p\right)^{1/p}$ is the standard Sobolev semi-norm, and C is a constant that only depends on (m,d,p). In the case of quasi-uniform partitions, this yields an estimate in terms of complexity:

$$\sigma_N(f)_p \leq C N^{-m/d} |f|_{W^{m,p}}, \tag{4}$$

where the constant C now also depends on C_1 and C_2 in (1).

Here and throughout the chapter, C *denotes a generic constant which may vary from one equation to the other. The dependence of this constant with respect to the relevant parameters will be mentionned when necessary.*

Note that the above estimate can be achieved by restricting the family $\mathscr{A}_N$ to a single partition: for example, we start from a coarse partition $\mathscr{T}_0$ into cubes and recursively define a nested sequence of partition $\mathscr{T}_j$ by splitting each cube of $\mathscr{T}_{j-1}$ into 2^d cubes of half side-length. We then set

$$\mathscr{A}_N := \{\mathscr{T}_j\}, \text{ if } \#(\mathscr{T}_0)2^{dj} \leq N < \#(\mathscr{T}_0)2^{d(j+1)}.$$

Similar uniform refinement rules can be proposed for more general partitions into triangles, simplices or rectangles. With such a choice for $\mathscr{A}_N$, the set Σ_N on which one picks the approximation is thus a standard linear space. Piecewise polynomials on quasi-uniform partitions may therefore be considered as an instance of *linear approximation*.

The interest of adaptive partitions is that the choice of $\mathscr{T} \in \mathscr{A}_N$ may vary depending on f, so that the set Σ_N is inherently a nonlinear space. Piecewise polynomials on adaptive partitions are therefore an instance of *nonlinear approximation*. Other instances include approximation by rational functions, or by N-term linear combinations of a basis or dictionary. We refer to [28] for a general survey on nonlinear approximation.

The use of adaptive partitions allows to improve significantly on (4). The theory that describes these improvements is rather well established for adaptive isotropic partitions: as explained further, a typical result for such partitions is of the form

$$\sigma_N(f)_p \leq CN^{-m/d}|f|_{W^{m,\tau}}, \tag{5}$$

where τ can be chosen smaller than p. Such an estimate reveals that the same rate of decay $N^{-\frac{m}{d}}$ as in (4) is achieved for f in a smoothness space which is larger than $W^{m,p}$. It also says that for a smooth function, the multiplicative constant governing this rate might be substantially smaller than when working with quasi-uniform partitions.

When allowing adaptive anisotropic partitions, one should expect for further improvements. From an intuitive point of view, such partitions are needed when the function f itself displays locally anisotropic features such as jump discontinuities or sharp transitions along smooth manifolds. The available approximation theory for such partitions is still at its infancy. Here, typical estimates are also of the form

$$\sigma_N(f)_p \leq CN^{-m/d}A(f), \tag{6}$$

but they involve quantities $A(f)$ which are not norms or semi-norms associated with standard smoothness spaces. These quantities are highly nonlinear in f in the sense that they do not satisfy $A(f+g) \leq C(A(f)+A(g))$ even with $C \geq 1$.

1.3 Outline

This chapter is organized as follows. As a starter, we study in §2 the simple case of piecewise constant approximation on an interval. This example gives a first illustration the difference between the approximation properties of uniform and adaptive partitions. It also illustrates the principle of *error equidistribution* which plays a crucial role in the construction of adaptive partitions which are optimally adapted to f. This leads us to propose and study a *multiresolution greedy refinement algorithm* as a design tool for such partitions. The distinction between isotropic and anisotropic partitions is irrelevant in this case, since we work with one-dimensional intervals.

We discuss in §3 the derivation of estimates of the form (5) for adaptive isotropic partitions. The main guiding principle for the design of the partition is again error equidistribution. Adaptive greedy refinement algorithms are discussed, similar to the one-dimensional case.

We study in §4 an elementary case of adaptive anisotropic partitions for which all elements are two-dimensional rectangles with sides that are parallel to the x and y axes. This type of anisotropic partitions suffer from an intrinsic lack of directional selectivity. We limitate our attention to piecewise constant functions, and identify the quantity $A(f)$ involved in (6) for this particular case. The main guiding principles for the design of the optimal partition are now error equidistribution combined with a local *shape optimization* of each element.

In §5, we present some recently available theory for piecewise polynomials on adaptive anisotropic partitions into triangles (and simplices in dimension $d > 2$) which offer more directional selectivity than the previous example. We give a general formula for the quantity $A(f)$ which can be turned into an explicit expression in terms of the derivatives of f in certain cases such as piecewise linear functions i.e. $m = 2$. Due to the fact that $A(f)$ is not a semi-norm, the function classes defined by the finiteness of $A(f)$ are not standard smoothness spaces. We show that these classes include piecewise smooth objects separated by discontinuities or sharp transitions along smooth edges.

We present in §6 several greedy refinement algorithms which may be used to derive anisotropic partitions. The convergence analysis of these algorithms is more delicate than for their isotropic counterpart, yet some first results indicate that they tend to generate optimally adapted partitions which satisfy convergence estimates in accordance with (6). This behaviour is illustrated by numerical tests on two-dimensional functions.

2 Piecewise constant one-dimensional approximation

We consider here the very simple problem of approximating a continuous function by piecewise constants on the unit interval $[0,1]$, when we measure the error in the uniform norm. If $f \in C([0,1])$ and $I \subset [0,1]$ is an arbitrary interval we have

$$e_{1,I}(f)_\infty := \min_{c\in\mathbf{R}} \|f-c\|_{L^\infty(I)} = \frac{1}{2}\max_{x,y\in I}|f(x)-f(y)|.$$

The constant c that achieves the minimum is the median of f on I. Remark that we multiply this estimate at most by a factor 2 if we take $c = f(z)$ for any $z \in I$. In particular, we may choose for c the average of f on I which is still defined when f is not continuous but simply integrable.

If $\mathscr{T}_N = \{I_1, \cdots, I_N\}$ is a partition of $[0,1]$ into N sub-intervals and $V_{\mathscr{T}_N}$ the corresponding space of piecewise constant functions, we thus find hat

$$e_{1,\mathscr{T}_N}(f)_\infty := \min_{g\in V_{\mathscr{T}_N}} \|f-g\|_{L^\infty} = \frac{1}{2}\max_{k=1,\cdots,N}\max_{x,y\in I_k}|f(x)-f(y)|. \tag{7}$$

2.1 Uniform partitions

We first study the error of approximation when the $\mathscr{T}_N$ are uniform partitions consisting of the intervals $I_k = [\frac{k}{N}, \frac{(k+1)}{N}]$. Assume first that f is a Lipschitz function i.e. $f' \in L^\infty$. We then have

$$\max_{x,y\in I_k} |f(x) - f(y)| \leq |I_k| \|f'\|_{L^\infty(I_k)} = N^{-1} \|f'\|_{L^\infty}.$$

Combining this estimate with (7), we find that for uniform partitions,

$$f \in \mathrm{Lip}([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-1}, \tag{8}$$

with $C = \frac{1}{2}\|f'\|_{L^\infty}$. For less smooth functions, we may obtain lower convergence rates: if f is Hölder continuous of exponent $0 < \alpha < 1$, we have by definition

$$|f(x) - f(y)| \leq |f|_{C^\alpha} |x-y|^\alpha,$$

which yields

$$\max_{x,y\in I_k} |f(x) - f(y)| \leq N^{-\alpha} |f|_{C^\alpha}.$$

We thus find that

$$f \in C^\alpha([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-\alpha}, \tag{9}$$

with $C = \frac{1}{2}|f|_{C^\alpha}$.

The estimates (8) and (9) are sharp in the sense that they admit a converse: it is easily checked that if f is a continuous function such that $\sigma_N(f)_\infty \leq CN^{-1}$ for some $C > 0$, it is necessarily Lipschitz. Indeed, for any x and y in $[0,1]$, consider an integer N such that $\frac{1}{2}N^{-1} \leq |x-y| \leq N^{-1}$. For such an integer, there exists a $f_N \in V_{\mathscr{T}_N}$ such that $\|f - f_N\|_{L^\infty} \leq CN^{-1}$. We thus have

$$|f(x) - f(y)| \leq 2CN^{-1} + |f_N(x) - f_N(y)|.$$

Since x and y are either contained in one interval or two adjacent intervals of the partition $\mathscr{T}_N$ and since f is continuous, we find that $|f_N(x) - f_N(y)|$ is either zero or less than $2CN^{-1}$. We therefore have

$$|f(x) - f(y)| \leq 4CN^{-1} \leq 8C|x-y|,$$

which shows that $f \in \mathrm{Lip}([0,1])$. In summary, we have the following result.

Theorem 2.1. *If f is a continuous function defined on $[0,1]$ and if $\sigma_N(f)_\infty$ denotes the L^∞ error of piecewise constant approximation on uniform partitions, we have*

$$f \in \mathrm{Lip}([0,1]) \Leftrightarrow \sigma_N(f)_\infty \leq CN^{-1}. \tag{10}$$

In an exactly similar way, is can be proved that

$$f \in C^{\alpha}([0,1]) \Leftrightarrow \sigma_N(f)_\infty \leq CN^{-\alpha}, \tag{11}$$

These equivalences reveal that Lipschitz and Holder smoothness are the properties that do govern the rate of approximation by piecewise constant functions in the uniform norm.

The estimate (8) is also optimal in the sense that it describes the *saturation rate* of piecewise constant approximation: a higher convergence rate cannot be obtained, even for smoother functions, and the constant $C = \frac{1}{2}\|f'\|_{L^\infty}$ cannot be improved. In order to see this, consider an arbitrary function $f \in C^1([0,1])$, so that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$|x-y| \leq \eta \Rightarrow |f'(x) - f'(y)| \leq \varepsilon.$$

Therefore if N is such that $N^{-1} \leq \eta$, we can introduce on each interval I_k an affine function $p_k(x) = f(x_k) + (x-x_k)f'(x_k)$ where x_k is an arbitrary point in I_k, and we then have

$$\|f - p_k\|_{L^\infty(I_k)} \leq N^{-1}\varepsilon.$$

It follows that

$$\begin{aligned} e_{1,I_k}(f)_\infty &\geq e_{1,I_k}(p_k)_\infty - e_{1,I_k}(f-p_k)_\infty \\ &\geq e_{1,I_k}(p_k)_\infty - \tfrac{1}{2}N^{-1}\varepsilon \\ &= \tfrac{1}{2}N^{-1}(|f'(x_k)| - \varepsilon), \end{aligned}$$

where we have used the triangle inequality

$$e_{m,T}(f+g)_p \leq e_{m,T}(f)_p + e_{m,T}(g)_p, \tag{12}$$

Choosing for x_k the point that maximize $|f'|$ on I_k and taking the supremum of the above estimate over all k, we obtain

$$e_{1,\mathscr{T}_N}(f)_\infty \geq \frac{1}{2}N^{-1}(\|f'\|_{L^\infty} - \varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, this implies the lower estimate

$$\liminf_{N\to+\infty} N\sigma_N(f)_\infty \geq \frac{1}{2}\|f'\|_{L^\infty}. \tag{13}$$

Combining with the upper estimate (8), we thus obtain the equality

$$\lim_{N\to+\infty} N\sigma_N(f)_\infty = \frac{1}{2}\|f'\|_{L^\infty}, \tag{14}$$

for any function $f \in C^1$. This identity shows that for smooth enough functions, the numerical quantity that governs the rate of convergence N^{-1} of uniform piecewise constant approximations is exactly $\frac{1}{2}\|f'\|_{L^\infty}$.

2.2 Adaptive partitions

We now consider an adaptive partition $\mathscr{T}_N$ for which the intervals I_k may depend on f. In order to understand the gain in comparison to uniform partitions, let us consider a function f such that $f' \in L^1$, i.e. $f \in W^{1,1}([0,1])$. Remarking that

$$\max_{x,y\in I} |f(x) - f(y)| \leq \int_I |f'(t)|dt,$$

we see that a natural choice fo the I_k can be done by imposing that

$$\int_{I_k} |f'(t)|dt = N^{-1} \int_0^1 |f'(t)|dt,$$

which means that the L^1 norm of f' is equidistributed over all intervals. Combining this estimate with (7), we find that for adaptive partitions,

$$f \in W^{1,1}([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-1}, \tag{15}$$

with $C := \frac{1}{2}\|f'\|_{L^1}$. This improvement upon uniform partitions in terms of approximation properties was firstly established in [35]. The above argument may be extended to the case where f belongs to the slightly larger space $BV([0,1])$ which may include discontinuous functions in contrast to $W^{1,1}([0,1])$, by asking that the I_k are such that

$$|f|_{BV(I_k)} \leq N^{-1}|f|_{BV}.$$

We thus have

$$f \in BV([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-1}, \tag{16}$$

Similar to the case of uniform partitions, the estimate (16) is sharp in the sense that a converse result holds: if f is a continuous function such that $\sigma_N(f)_\infty \leq CN^{-1}$ for some $C > 0$, then it is necessarily in $BV([0,1])$. To see this, consider $N > 0$ and any set of points $0 \leq x_1 < x_2 < \cdots < x_N \leq 1$. We know that there exists a partition $\mathscr{T}_N$ of N intervals and $f_N \in V_{\mathscr{T}_N}$ such that $\|f - f_N\|_{L^\infty} \leq CN^{-1}$. We define a set of points $0 \leq y_1 < y_2 \cdots < y_M \leq 1$ by unioning the set of the x_k with the nodes that define the partition $\mathscr{T}_N$, excluding 0 and 1, so that $M < 2N$. We can write

$$\sum_{k=0}^{N-1} |f(x_{k+1}) - f(x_k)| \leq 2C + \sum_{k=0}^{N-1} |f_N(x_{k+1}) - f_N(x_k)| \leq 2C + \sum_{k=0}^{M-1} |f_N(y_{k+1}) - f_N(y_k)|.$$

Since y_k and y_{k+1} are either contained in one interval or two adjacent intervals of the partition $\mathscr{T}_N$ and since f is continuous, we find that $|f_N(y_{k+1}) - f_N(y_k)|$ is either zero or less than $2CN^{-1}$, from which it follows that

$$\sum_{k=0}^{N-1} |f(x_{k+1}) - f(x_k)| \leq 6C,$$

which shows that f has bounded variation. We have thus proved the following result.

Theorem 2.2. *If f is a continuous function defined on $[0,1]$ and if $\sigma_N(f)_\infty$ denotes the L^∞ error of piecewise constant approximation on adaptive partitions, we have*

$$f \in BV([0,1]) \Leftrightarrow \sigma_N(f)_\infty \leq CN^{-1}. \tag{17}$$

In comparison with (8) we thus find that same rate N^{-1} is governed by a *weaker* smoothness condition since f' is not assumed to be bounded but only a finite measure. In turn, adaptive partitions may significantly outperform uniform partition for a given function f: consider for instance the function $f(x) = x^\alpha$ for some $0 < \alpha < 1$. According to (11), the convergence rate of uniform approximation for this function is $N^{-\alpha}$. On the other hand, since $f'(x) = \alpha x^{\alpha-1}$ is integrable, we find that the convergence rate of adaptive approximation is N^{-1}.

The above construction of an adaptive partition is based on equidistributing the L^1 norm of f' or the total variation of f on each interval I_k. An alternative is to build $\mathscr{T}_N$ in such a way that all local errors are equal, i.e.

$$\varepsilon_{1,I_k}(f)_\infty = \eta, \tag{18}$$

for some $\eta = \eta(N) \geq 0$ independent of k. This new construction of $\mathscr{T}_N$ does not require that f belongs to $BV([0,1])$. In the particular case where $f \in BV([0,1])$, we obtain that

$$N\eta \leq \sum_{k=1}^{N} e_{1,I_k}(f)_\infty \leq \frac{1}{2}\sum_{k=1}^{N} |f|_{BV(I_k)} \leq \frac{1}{2}|f|_{BV},$$

from which it immediately follows that

$$e_{1,\mathscr{T}_N}(f)_\infty = \eta \leq CN^{-1},$$

with $C = \frac{1}{2}|f|_{BV}$. We thus have obtained the same error estimate as with the previous construction of $\mathscr{T}_N$.

The basic principle of error equidistribution, which is expressed by (18) *in the case of piecewise constant approximation in the uniform norm, plays a central role in the derivation of adaptive partitions for piecewise polynomial approximation.*

Similar to the case of uniform partitions we can express the optimality of (15) by a lower estimate when f is smooth enough. For this purpose, we make a slight restriction on the set $\mathscr{A}_N$ of admissible partitions, assuming that the diameter of all intervals decreases as $N \to +\infty$, according to

$$\max_{I_k \in \mathscr{T}_N} |I_k| \leq AN^{-1},$$

for some $A > 0$ which may be arbitrarily large. Assume that $f \in C^1([0,1])$, so that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$|x-y| \leq \eta \Rightarrow |f'(x)-f'(y)| \leq \frac{\varepsilon}{A}. \tag{19}$$

If N is such that $AN^{-1} \leq \eta$, we can introduce on each interval I_k an affine function $p_k(x) = f(x_k) + (x-x_k)f'(x_k)$ where x_k is an arbitrary point in I_k, and we then have

$$\|f-p_k\|_{L^\infty(I_k)} \leq N^{-1}\varepsilon.$$

It follows that

$$\begin{aligned} e_{1,I_k}(f)_\infty &\geq e_{1,I_k}(p_k)_\infty - e_{1,I_k}(f-p_k)_\infty \\ &\geq e_{1,I_k}(p_k)_\infty - \tfrac{1}{2}N^{-1}\varepsilon \\ &= \tfrac{1}{2}\left(\int_{I_k}|p_k'(t)|dt - N^{-1}\varepsilon\right) \\ &\geq \tfrac{1}{2}\left(\int_{I_k}|f'(t)|dt - 2N^{-1}\varepsilon\right). \end{aligned}$$

Since there exists at least one interval I_k such that $\int_{I_k}|f'(t)|dt \geq N^{-1}\|f'\|_{L^1}$, it follows that

$$e_{1,\mathscr{T}_N}(f)_\infty \geq \frac{1}{2}N^{-1}(\|f'\|_{L^1} - 2\varepsilon).$$

This inequality becomes an equality only when all quantities $\int_{I_k}|f'(t)|dt$ are equal, which justifies the equidistribution principle for the design of an optimal partition. Since $\varepsilon > 0$ is arbitrary, we have thus obtained the lower estimate

$$\liminf_{N\to+\infty} N\sigma_N(f) \geq \frac{1}{2}\|f'\|_{L^1}. \tag{20}$$

The restriction on the family of adaptive partitions $\mathscr{A}_N$ is not so severe since A maybe chosen arbitrarily large. In particular, it is easy to prove that the upper estimate is almost preserved in the following sense: for a given $f \in C^1$ and any $\varepsilon > 0$, there exists $A > 0$ depending on ε such that

$$\limsup_{N\to+\infty} N\sigma_N(f) \leq \frac{1}{2}\|f'\|_{L^1} + \varepsilon,$$

These results show that for smooth enough functions, the numerical quantity that governs the rate of convergence N^{-1} of adaptive piecewise constant approximations is exactly $\frac{1}{2}\|f'\|_{L^1}$. Note that $\|f'\|_{L^\infty}$ may be substantially larger than $\|f'\|_{L^1}$ even for very smooth functions, in which case adaptive partitions performs at a similar rate as uniform partitions, but with a much more favorable multiplicative constant.

2.3 A greedy refinement algorithm

The principle of error distribution suggests a simple algorithm for the generation of adaptive partitions, based on a greedy refinement algorithm:

1. Initialization: $\mathscr{T}_1 = \{[0,1]\}$.

2. Given $\mathscr{T}_N$ select $I_m \in \mathscr{T}_N$ that maximizes the local error $e_{1,I_k}(f)_\infty$.
3. Split I_m into two sub-intervals of equal size to obtain $\mathscr{T}_{N+1}$ and return to step 2.

The family $\mathscr{A}_N$ of adaptive partitions that are generated by this algorithm is characterized by the restriction that all intervals are of the dyadic type $2^{-j}[n,n+1]$ for some $j \geq 0$ and $n \in \{0,\cdots,2^j-1\}$. We also note that all such partitions $\mathscr{T}_N$ may be identified to a finite subtree with N leaves, picked within an infinite dyadic *master tree* $\mathscr{M}$ in which each node represents a dyadic interval. The root of $\mathscr{M}$ corresponds to $[0,1]$ and each node I of generation j corresponds to an interval of length 2^{-j} which has two *children* nodes of generation $j+1$ corresponding to the two halves of I. This identification, which is illustrated on Figure 1, is useful for coding purposes since any such subtree can be encoded by $2N$ bits.

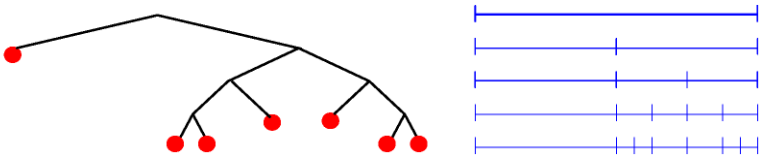


Fig. 1 Adaptive dyadic partitions identify to dyadic trees

We now want to understand how the approximations generated by adaptive refinement algorithm behave in comparison to those associated with the optimal partition. In particular, do we also have that $e_{1,\mathscr{T}_N}(f)_\infty \leq CN^{-1}$ when $f' \in L^1$? The answer to this question turns out to be negative, but it was proved in [30] that a slight strengthening of the smoothness assumption is sufficient to ensure this convergence rate: we instead assume that the *maximal function* of f' is in L^1. We recall that the maximal function of a locally integrable function g is defined by

$$M_g(x) := \sup_{r>0} |B(x,r)|^{-1} \int_{B(x,r)} |g(t)|dt,$$

It is known that $M_g \in L^p$ if and only if $g \in L^p$ for $1 < p < \infty$ and that $M_g \in L^1$ if and only if $g \in L\log L$, i.e. $\int_0^1 |g(t)|\log(1+|g(t)|)dt < \infty$, see [42]. In this sense, the assumption that $M_{f'}$ is integrable is only slightly stronger than $f \in W^{1,1}$.

If $\mathscr{T}_N := (I_1,\cdots,I_N)$, define the accuracy

$$\eta := \max_{1\leq k\leq N} e_{1,I_k}(f)_\infty.$$

For each k, we denote by J_k the interval which is the *parent* of I_k in the refinement process. From the definition of the algorithm, we necessarily have

$$\eta \leq \|f - a_{J_k}(f)\|_{L^\infty} \leq \int_{J_k} |f'(t)|dt.$$

For all $x \in I_k$, the ball $B(x, 2|I_k|)$ contains J_k and it follows therefore that

$$M_{f'}(x) \geq |B(x, 2|I_k|)|^{-1} \int_{B(x,2|I_k|)} |f'(t)|dt \geq [4|I_k|]^{-1}\eta,$$

which implies in turn

$$\int_{I_k} M_{f'}(t)dt \geq \eta/4.$$

If $M_{f'}$ is integrable, this yields the estimate

$$N\eta \leq 4 \int_0^1 M_{f'}(t)dt.$$

It follows that

$$e_{1,\mathscr{T}_N}(f)_\infty = \eta \leq CN^{-1}$$

with $C = 4\|M_{f'}\|_{L^1}$. We have thus established the following result.

Theorem 2.3. *If f is a continuous function defined on $[0,1]$ and if $\sigma_N(f)_\infty$ denotes the L^∞ error of piecewise constant approximation on adaptive partitions of dyadic type, we have*

$$M_{f'} \in L^1([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-1}, \tag{21}$$

and that this rate may be achieved by the above described greedy algorithm.

Note however that a converse to (21) does not hold and that we do not so far know of a simple smoothness property that would be exactly equivalent to the rate of approximation N^{-1} by dyadic adaptive partitions. A by-product of (21) is that

$$f \in W^{1,p}([0,1]) \Rightarrow \sigma_N(f)_\infty \leq CN^{-1}, \tag{22}$$

for any $p > 1$.

3 Adaptive and isotropic approximation

We now consider the problem of piecewise polynomial approximation on a domain $\Omega \subset \mathbb{R}^d$, using adaptive and *isotropic* partitions. We therefore consider a sequence $(\mathscr{A}_N)_{N\geq 0}$ of families of partitions that satisfies the restriction (2). We use piecewise polynomials of degree $m-1$ for some fixed but arbitrary m.

Here and in all the rest of the chapter, we restrict our attention to partitions into geometrically simple elements which are either cubes, rectangles or simplices. These simple elements satisfy a property of *affine invariance*: there exist a *reference element* R such that any $T \in \mathscr{T} \in \mathscr{A}_N$ is the image of R by an invertible affine transformation A_T. We can choose R to be the unit cube $[0,1]^d$ or the unit simplex $\{0 \leq x_1 \leq \cdots \leq x_d \leq 1\}$ in the case of partitions by cubes and rectangles or simplices, respectively.

3.1 Local estimates

If $T \in \mathscr{T}$ is an element and f is a function defined on Ω, we study the local approximation error

$$e_{m,T}(f)_p := \min_{\pi \in \mathbb{P}_{m-1}} \|f-\pi\|_{L^p(T)}. \tag{23}$$

When $p=2$ the minimizing polynomial is given by

$$\pi := P_{m,T} f,$$

where $P_{m,T}$ is the L^2-orthogonal projection, and can therefore be computed by solving a least square system. When $p \neq 2$, the minimizing polynomial is generally not easy to determine. However it is easily seen that the L^2-orthogonal projection remains an acceptable choice: indeed, it can easily be checked that the operator norm of $P_{m,T}$ in $L^p(T)$ is bounded by a constant C that only depends on (m,d) but not on the cube or simplex T. From this we infer that for all f and T one has

$$e_{m,T}(f)_p \leq \|f-P_{m,T}f\|_{L^p(T)} \leq (1+C)e_{m,T}(f)_p. \tag{24}$$

Local estimates for $e_{m,T}(f)_p$ can be obtained from local estimates on the reference element R, remarking that

$$e_{m,T}(f)_p = \Big(\frac{|T|}{|R|}\Big)^{1/p} e_{m,R}(g)_p, \tag{25}$$

where $g = f \circ A_T$. Assume that $p, \tau \geq 1$ are such that $\frac{1}{\tau} = \frac{1}{p} + \frac{m}{d}$, and let $g \in W^{m,\tau}(R)$. We know from Sobolev embedding that

$$\|g\|_{L^p(R)} \leq C\|g\|_{W^{m,\tau}(R)},$$

where the constant C depends on p, τ and R. Accordingly, we obtain

$$e_{m,R}(g)_p \leq C \min_{\pi \in \mathbb{P}_{m-1}} \|g-\pi\|_{W^{m,\tau}(R)}. \tag{26}$$

We then invoke Deny-Lions theorem which states that if R is a connected domain, there exists a constant C that only depends on m and R such that

$$\min_{\pi \in \mathbb{P}_{m-1}} \|g-\pi\|_{W^{m,\tau}(R)} \leq C|g|_{W^{m,\tau}(R)}. \tag{27}$$

If $g = f \circ A_T$, we obtain by this change of variable that

$$|g|_{W^{m,\tau}(R)} \leq C\Big(\frac{|R|}{|T|}\Big)^{1/\tau} \|B_T\|^m |f|_{W^{m,\tau}(T)}, \tag{28}$$

where B_T is the linear part of A_T and C is a constant that only depends on m and d. A well known and easy to derive bound for $\|B_T\|$ is

$$\|B_T\| \leq \frac{h_T}{\rho_R}, \tag{29}$$

Combining (25), (26), (27), (28) and (29), we thus obtain a local estimate of the form

$$e_{m,T}(f)_p \leq C|T|^{1/p-1/\tau} h_T^m |f|_{W^{m,\tau}(T)} = C|T|^{-m/d} h_T^m |f|_{W^{m,\tau}(T)}.$$

where we have used the relation $\frac{1}{\tau} = \frac{1}{p} + \frac{m}{d}$. From the isotropy restriction (2), there exists a constant $C > 0$ independent of T such that $h_T^d \leq C|T|$. We have thus established the following local error estimate.

Theorem 3.1. *If $f \in W^{m,\tau}(\Omega)$, we have for all element T*

$$e_{m,T}(f)_p \leq C|f|_{W^{m,\tau}(T)}, \tag{30}$$

where the constant C only depends on m, R and the constants in (2).

Let us mention several useful generalizations of the local estimate (30) that can be obtained by a similar approach based on a change of variable on the reference element. First, if $f \in W^{s,\tau}(\Omega)$ for some $0 < s \leq m$ and $\tau \geq 1$ such that $\frac{1}{\tau} = \frac{1}{p} + \frac{s}{d}$, we have

$$e_{m,T}(f)_p \leq C|f|_{W^{s,\tau}(T)}. \tag{31}$$

Recall that when s is not an integer, the $W^{s,\tau}$ semi-norm is defined by

$$|f|_{W^{s,\tau}(\Omega)^q} := \sum_{|\alpha|=n} \int_{\Omega\times\Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|^\tau}{|x-y|^{(s-n)\tau+d}} dxdy,$$

where n is the largest integer below s. In the more general case where $\frac{1}{\tau} \leq \frac{1}{p} + \frac{s}{d}$, we obtain an estimate that depends on the diameter of T:

$$e_{m,T}(f)_p \leq C h_T^r |f|_{W^{s,\tau}(T)}, \quad r := \frac{d}{p} - \frac{d}{\tau} + s \geq 0. \tag{32}$$

Finally, remark that for a fixed $p \geq 1$ and s, the index τ defined by $\frac{1}{\tau} = \frac{1}{p} + \frac{s}{d}$ may be smaller than 1, in which case the Sobolev space $W^{s,\tau}(\Omega)$ is not well defined. The local estimate remain valid if $W^{s,\tau}(\Omega)$ is replaced by the *Besov space* $B^s_{\tau,\tau}(\Omega)$. This space consists of all $f \in L^\tau(\Omega)$ functions such that

$$|f|_{B^s_{\tau,\tau}} := \|\omega_k(f,\cdot)_\tau\|_{L^\tau([0,\infty[,\frac{dt}{t})},$$

is finite. Here k is the smallest integer above s and $\omega_k(f,t)_\tau$ denotes the L^τ-modulus of smoothness of order k defined by

$$\omega_k(f,t)_\tau := \sup_{|h|\leq t} \|\Delta_h^k f\|_{L^\tau},$$

where $\Delta_h f := f(\cdot + h) - f(\cdot)$ is the usual difference operator. The space $B^s_{\tau,\tau}$ describes functions which have "s derivatives in L^τ" in a very similar way as $W^{s,\tau}$.

In particular it is known that these two spaces coincide when $\tau \geq 1$ and s is not an integer. We refer to [29] and [18] for more details on Besov spaces and their characterization by approximation procedures. For all $p, \tau > 0$ and $0 \leq s \leq m$ such that $\frac{1}{\tau} \leq \frac{1}{p} + \frac{s}{d}$, a local estimate generalizing (32) has the form

$$e_{m,T}(f)_p \leq C h_T^r |f|_{B^s_{\tau,\tau}(T)}, \quad r := \frac{d}{p} - \frac{d}{\tau} + s \geq 0. \tag{33}$$

3.2 Global estimates

We now turn our local estimates into global estimates, recalling that

$$e_{m,\mathscr{T}}(f)_p := \min_{g \in V_{\mathscr{T}}} \|f - g\|_{L^p} = \Big(\sum_{T \in \mathscr{T}} e_{m,T}(f)_p^p \Big)^{1/p};$$

with the usual modification when $p = \infty$. We apply the principle of error equidistribution assuming that the partition $\mathscr{T}_N$ is built in such way that

$$e_{m,T}(f)_p = \eta, \tag{34}$$

for all $T \in \mathscr{T}_N$ where $N = N(\eta)$. A first immediate estimate for the global error is therefore

$$e_{m,\mathscr{T}_N}(f)_p \leq N^{1/p} \eta. \tag{35}$$

Assume now that $f \in W^{m,\tau}(\Omega)$ with $\tau \geq 1$ such that $\frac{1}{\tau} = \frac{1}{p} + \frac{m}{d}$. It then follows from Theorem 3.1 that

$$N \eta^\tau \leq \sum_{T \in \mathscr{T}_N} e_{m,T}(f)_p^\tau \leq C \sum_{T \in \mathscr{T}_N} |f|^\tau_{W^{m,\tau}(T)} = C |f|^\tau_{W^{m,\tau}},$$

Combining with (35) and using the relation $\frac{1}{\tau} = \frac{1}{p} + \frac{m}{d}$, we have thus obtained that for adaptive partitions $\mathscr{T}_N$ built according to the error equidistribution, we have

$$e_{m,\mathscr{T}_N}(f)_p \leq C N^{-m/d} |f|_{W^{m,\tau}}. \tag{36}$$

By using (31), we obtain in a similar manner that if $0 \leq s \leq m$ and $\tau \geq 1$ are such that $\frac{1}{\tau} = \frac{1}{p} + \frac{s}{d}$, then

$$e_{m,\mathscr{T}_N}(f)_p \leq C N^{-s/d} |f|_{W^{s,\tau}}. \tag{37}$$

Similar results hold when $\tau < 1$ with $W^{s,\tau}$ replaced by $B^s_{\tau,\tau}$ but their proof requires a bit more work due to the fact that $|f|^\tau_{B^s_{\tau,\tau}}$ is not sub-additive with respect to the union of sets. We also reach similar estimate in the case $p = \infty$ by a standard modification of the argument.

The estimate (36) suggests that for piecewise polynomial approximation on adaptive and isotropic partitions, we have

$$\sigma_N(f)_p \leq CN^{-m/d}|f|_{W^{m,\tau}}, \quad \frac{1}{\tau} = \frac{1}{p} + \frac{m}{d}. \tag{38}$$

Such an estimate should be compared to (4), in a similar way as we compared (17) with (8) in the one dimensional case: the same same rate $N^{-m/d}$ is governed by a weaker smoothness condition.

In contrast to the one dimensional case, however, we cannot easily prove the validity of (38) since it is not obvious that there exists a partition $\mathscr{T}_N \in \mathscr{A}_N$ which equidistributes the error in the sense of (34). It should be remarked that the derivation of estimates such as (36) does not require a strict equidistribution of the error. It is for instance sufficient to assume that $e_{m,T}(f)_p \leq \eta$ for all $T \in \mathscr{T}_N$, and that

$$c_1\eta \leq e_{m,T}(f)_p,$$

for at least $c_2 N$ elements of $\mathscr{T}_N$, where c_1 and c_2 are fixed constants. Nevertheless, the construction of a partition $\mathscr{T}_N$ satisfying such prescriptions still appears as a difficult task both from a theoretical and algorithmical point of view.

3.3 An isotropic greedy refinement algorithm

We now discuss a simple adaptive refinement algorithm which emulates error equidistribution, similar to the algorithm which was discussed in the one dimensional case. For this purpose, we first build a hierarchy of nested quasi-uniform partitions $(\mathscr{D}_j)_{j\geq 0}$, where $\mathscr{D}_0$ is a coarse triangulation and where $\mathscr{D}_{j+1}$ is obtained from $\mathscr{D}_j$ by splitting each of its elements into a fixed number K of children. We therefore have

$$\#(\mathscr{D}_j) = K^j \#(\mathscr{D}_0),$$

and since the partitions $\mathscr{D}_j$ are assumed to be quasi-uniform, there exists two constants $0 < c_1 \leq c_2$ such that

$$c_1 K^{-j/d} \leq h_T \leq c_2 K^{-j/d}, \tag{39}$$

for all $T \in \mathscr{D}_j$ and $j \geq 0$. For example, in the case of two dimensional triangulations, we may choose $K = 4$ by splitting each triangle into 4 similar triangles by the midpoint rule, or $K = 2$ by bisecting each triangle from one vertex to the midpoint of the opposite edge according to a prescribed rule in order to preserve isotropy. Specific rules which have been extensively studied are bisection from the most recently generated vertex [8] or towards the longest edge [41]. In the case of partitions by rectangles, we may preserve isotropy by splitting each rectangle into 4 similar rectangles by the midpoint rule.

The refinement algorithm reads as follows:

1. Initialization: $\mathscr{T}_{N_0} = \mathscr{D}_0$ with $N_0 := \#(\mathscr{D}_0)$.
2. Given $\mathscr{T}_N$ select $T \in \mathscr{T}_N$ that maximizes $e_{m,T}(f)_T$.

3. Split T into its K childrens to obtain $\mathscr{T}_{N+K-1}$ and return to step 2.

Similar to the one dimensional case, the adaptive partitions that are generated by this algorithm are restricted to a particular family where each element T is picked within an infinite dyadic *master tree* $\mathscr{M} = \cup_{j\geq 0}\mathscr{D}_j$ which roots are given by the elements $\mathscr{D}_0$. The partition $\mathscr{T}_N$ may be identified to a finite subtree of $\mathscr{M}$ with N leaves. Figure 2 displays an example of adaptively refined partitions either based on longest edge bisection for triangles, or by quad-split for squares.

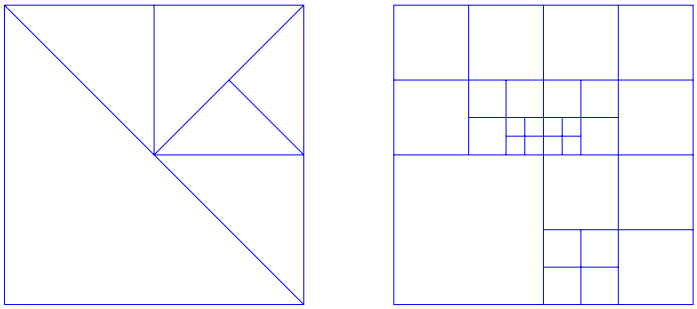


Fig. 2 Adaptively refined partitions based on longest edge bisection (left) or quad-split (right)

This algorithm cannot exactly achieve error equidistribution, but our next result reveals that it generates partitions that yield error estimates almost similar to (36).

Theorem 3.2. *If $f \in W^{m,\tau}(\Omega)$ for some $\tau \geq 1$ such that $\frac{1}{\tau} < \frac{1}{p} + \frac{m}{d}$, we then have for all $N \geq 2N_0 = 2\#(\mathscr{D}_0)$,*

$$e_{m,\mathscr{T}_N}(f)_p \leq CN^{-m/d}|f|_{W^{m,\tau}}, \tag{40}$$

where C depends on τ, m, K, R and the choice of $\mathscr{D}_0$. We therefore have for piecewise polynomial approximation on adaptively refined partitions

$$\sigma_N(f)_p \leq CN^{-m/d}|f|_{W^{m,\tau}}, \quad \frac{1}{\tau} > \frac{1}{p} + \frac{m}{d}. \tag{41}$$

Proof: The technique used for proving this result is adapted from the proof of a similar result for tree-structured wavelet approximation in [19]. We define

$$\eta := \max_{T\in\mathscr{T}_N} e_{m,T}(f)_p, \tag{42}$$

so that we obviously have when $p < \infty$,

$$e_{m,\mathscr{T}_N}(f)_p \leq N^{1/p}\eta. \tag{43}$$

For $T \in \mathscr{T}_N \setminus \mathscr{D}_0$, we denote by $P(T)$ its parent in the refinement process. From the definition of the algorithm, we necessarily have

$$\eta \leq e_{m,P(T)}(f)_p,$$

and therefore, using (32) with $s = m$, we obtain

$$\eta \leq C h_{P(T)}^r |f|_{W^{s,\tau}(P(T))}, \tag{44}$$

with $r := \frac{d}{p} - \frac{d}{\tau} + m > 0$. We next denote by $\mathscr{T}_{N,j} := \mathscr{T}_N \cap \mathscr{D}_j$ the elements of generation j in $\mathscr{T}_N$ and define $N_j := \#(\mathscr{T}_{N,j})$. We estimate N_j by taking the τ power of (44) and summing over $\mathscr{T}_{N,j}$ which gives

$$\begin{aligned} N_j \eta^\tau &\leq C^\tau \textstyle\sum_{T\in\mathscr{T}_{N,j}} h_{P(T)}^{r\tau} |f|^\tau_{W^{s,\tau}(P(T))} \\ &\leq C^\tau (\sup_{T\in\mathscr{T}_{N,j}} h_{P(T)}^{r\tau}) \textstyle\sum_{T\in\mathscr{T}_{N,j}} |f|^\tau_{W^{s,\tau}(P(T))} \\ &\leq K C^\tau (\sup_{T\in\mathscr{D}_{j-1}} h_T^{r\tau}) |f|^\tau_{W^{s,\tau}}. \end{aligned}$$

Using (39) and the fact that $\#(\mathscr{D}_j) = N_0 K^j$, we thus obtain

$$N_j \leq \min\{C\eta^{-\tau} K^{-jr\tau/d} |f|^\tau_{W^{s,\tau}}, N_0 K^j\}.$$

We now evaluate

$$N - N_0 = \sum_{j\geq 1} N_j \leq \sum_{j\geq 1} \min\{C\eta^{-\tau} K^{-jr\tau/d} |f|^\tau_{W^{s,\tau}}, N_0 K^j\}.$$

By introducing j_0 the smallest integer such that $C\eta^{-\tau} K^{-jr\tau/d} |f|^\tau_{W^{s,\tau}} \leq N_0 K^j$, we find that

$$N - N_0 \leq N_0 \sum_{j\leq j_0} K^j + C\eta^{-\tau} |f|^\tau_{W^{s,\tau}} \sum_{j>j_0} K^{-jr\tau/d},$$

which after evaluation of j_0 yields

$$N - N_0 \leq C\eta^{-\frac{d\tau}{d+r\tau}} |f|^{\frac{d\tau}{d+r\tau}}_{W^{s,\tau}} = C\eta^{-\frac{dp}{d+mp}} |f|^{\frac{dp}{d+mp}}_{W^{s,\tau}},$$

and therefore, assuming that $N \geq 2N_0$,

$$\eta \leq C N^{-1/p-m/d} |f|_{W^{s,\tau}}.$$

Combining this estimate with (43) gives the announced result. In the case $p = \infty$, a standard modification of the argument leads to a similar conclusion. □

Remark 3.1. By similar arguments, we obtain that if $f \in W^{s,\tau}(\Omega)$ for some $\tau \geq 1$ and $0 \leq s \leq m$ such that $\frac{1}{\tau} < \frac{1}{p} + \frac{s}{d}$, we have

$$e_{m,\mathscr{T}_N}(f)_p \leq C N^{-s/d} |f|_{W^{s,\tau}}.$$

The restriction $\tau \geq 1$ may be dropped if we replace $W^{s,\tau}$ by the Besov space $B^s_{\tau,\tau}$, at the price of a more technical proof.

Remark 3.2. The same approximation results can be obtained if we replace $e_{m,T}(f)_p$ in the refinement algorithm by the more computable quantity $\|f - P_{m,T} f\|_{L^p(T)}$, due to the equivalence (24).

Remark 3.3. The greedy refinement algorithm defines a particular sequence of subtrees $\mathscr{T}_N$ of the master tree $\mathscr{M}$, but $\mathscr{T}_N$ is not ensured to be the best choice in the sense of minimizing the approximation error among all subtrees of cardinality at most N. The selection of an optimal tree can be performed by an additional pruning strategy after enough refinement has been performed. This approach was developped in the context of statistical estimation under the acronyme CART (classification and regression tree), see [12, 32]. Another approach that builds a near optimal subtree only based on refinement was proposed in [7].

Remark 3.4. The partitions which are built by the greedy refinement algorithm are non-conforming. Additional refinement steps are needed when the users insists on conformity, for instance when solving PDE's. For specific refinement procedures, it is possible to bound the total number of elements that are due to additional conforming refinement by the total number of triangles T which have been refined due to the fact that $e_{m,T}(f)_T$ was the largest at some stage of the algorithm, up to a fixed multiplicative constant. In turn, the convergence rate is left unchanged compared to the original non-conforming algorithm. This fact was proved in [8] for adaptive triangulations built by the rule of newest vertex bisection. A closely related concept is the amount of additional elements which are needed in order to impose that the partition satisfies a *grading property*, in the sense that two adjacent elements may only differ by one refinement level. For specific partitions, it was proved in [23] that this amount is bounded up to a fixed multiplicative constant the number of elements contained in the non-graded partitions. Figure 3 displays the conforming and graded partitions obtained by the minimal amount of additional refinement from the partitions of Figure 2.

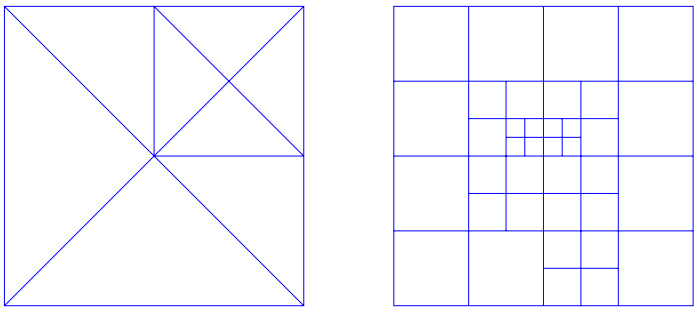


Fig. 3 Conforming refinement (left) and graded refinement (right)

The refinement algorithm may also be applied to discretized data, such as numerical images. The approximated 512×512 image is displayed on Figure 4 together

with its approximation obtained by the refinement algorithm based on newest vertex bisection and the error measured in L^2, using $N = 2000$ triangles. In this case, f has the form of a discrete array of pixels, and the $L^2(T)$-orthogonal projection is replaced by the $\ell^2(S_T)$-orthogonal projection, where S_T is the set of pixels with centers contained in T. The use of adaptive isotropic partitions has strong similarity with wavelet thresholding [28, 18]. In particular, it results in ringing artifacts near the edges.

Fig. 4 The image 'peppers" (left) and its approximation by 2000 isotropic triangles obtained by the greedy algorithm (right)

3.4 The case of smooth functions

Although the estimate (38) might not be achievable for a general $f \in W^{m,\tau}(\Omega)$, we can show that for smooth enough f, the numerical quantity that governs the rate of convergence $N^{-\frac{n}{d}}$ is exactly $|f|_{W^{m,\tau}} := \left(\sum_{|\alpha|=m} \|\partial^\alpha f\|_{L^\tau}^\tau\right)^{1/\tau}$ that we may define as so even for $\tau < 1$. For this purpose, we assume that $f \in C^m(\Omega)$. Our analysis is based on the fact that such a function can be locally approximated by a polynomial of degree m.

We first study in more detail the approximation error on a function $q \in \mathbb{P}_{\mathrm{m}}$. We denote by $\mathbb{H}_m$ the space of *homogeneous polynomials* of degree m. To $q \in \mathbb{P}_{\mathrm{m}}$, we associate its homogeneous part $\mathbf{q} \in \mathbb{H}_m$, which is such that

$$q - \mathbf{q} \in \mathbb{P}_{\mathrm{m}-1}.$$

We denote by $\mathbf{q}_\alpha$ the coefficient of $\mathbf{q}$ associated to the multi-index $\alpha = (\alpha_1, \cdots, \alpha_d)$ with $|\alpha| = m$. We thus have

$$e_{m,T}(q)_p = e_{m,T}(\mathbf{q})_p.$$

Using the affine transformation A_T which maps the reference element R onto T, and denoting by B_T its linear part, we can write

$$e_{m,T}(\mathbf{q})_p = \Big(\frac{|T|}{|R|}\Big)^{1/p} e_{R,m}(\mathbf{q}\circ A_T)_p = \Big(\frac{|T|}{|R|}\Big)^{1/p} e_{m,R}(\tilde{\mathbf{q}})_p, \;\; \tilde{\mathbf{q}} := \mathbf{q}\circ B_T \in \mathbb{H}_m$$

where we have used the fact that $\tilde{\mathbf{q}} - \mathbf{q}\circ A_T \in \mathbb{P}_{m-1}$. Introducing for any $r>0$ the quasi-norm on $\mathbb{H}_m$

$$|\mathbf{q}|_r := \Big(\sum_{|\alpha|=m} |\mathbf{q}_\alpha|^r\Big)^{1/r},$$

one easily checks that

$$C^{-1}\|B_T^{-1}\|^{-m}|\mathbf{q}|_r \le |\tilde{\mathbf{q}}|_r \le C\|B_T\|^m|\mathbf{q}|_r,$$

for some constant $C>0$ that only depends on m, r and R. We then remark that $e_{R,m}(\mathbf{q})_p$ is a norm on $\mathbb{H}_m$, which is equivalent to $|\mathbf{q}|_r$ since $\mathbb{H}_m$ is finite dimensional. It follows that there exists constants $0<C_1\le C_2$ such that for all q and T

$$C_1|T|^{1/p}\|B_T^{-1}\|^{-m}|\mathbf{q}|_r \le e_{m,T}(q)_p \le C_2|T|^{1/p}\|B_T\|^m|\mathbf{q}|_r.$$

Finally, using the bound (29) for $\|B_T\|$ and its symmetrical counterpart

$$\|B_T^{-1}\| \le \frac{h_R}{\rho_T},$$

together with the isotropy restriction (2), we obtain with $\frac{1}{\tau} := \frac{1}{p} + \frac{m}{d}$ the equivalence

$$C_1|T|^\tau|\mathbf{q}|_r \le e_{m,T}(q)_p \le C_2|T|^\tau|\mathbf{q}|_r,$$

where C_1 and C_2 only depend on m, R and the constant C in (2). Choosing $r=\tau$ this equivalence can be rewritten as

$$C_1\Big(\sum_{|\alpha|=m}\|\mathbf{q}_\alpha\|^\tau_{L^\tau(T)}\Big)^{1/\tau} \le e_{m,T}(q)_p \le C_2\Big(\sum_{|\alpha|=m}\|\mathbf{q}_\alpha\|^\tau_{L^\tau(T)}\Big)^{1/\tau}.$$

Using shorter notations, this is summarized by the following result.

Lemma 3.1. *Let $p\ge 1$ and $\frac{1}{\tau} := \frac{1}{p}+\frac{m}{d}$. There exists constant C_1 and C_2 that only depends on m, R and the constant C in (2) such that*

$$C_1|q|_{W^{m,\tau}(T)} \le e_{m,T}(q)_p \le C_2|q|_{W^{m,\tau}(T)}, \tag{45}$$

for all $q\in\mathbb{P}_m$.

In what follows, we shall frequently identify the m-th order derivatives of a function f at some point x with an homogeneous polynomial of degree m. In particular we write

$$|d^m f(x)|_r := \Big(\sum_{|\alpha|=m} |\partial^\alpha f(x)|^r \Big)^{1/r}.$$

We first establish a lower estimate on $\sigma_N(f)$, which reflects the saturation rate $N^{-m/d}$ of the method, under a slight restriction on the set $\mathscr{A}_N$ of admissible partitions, assuming that the diameter of all elements decreases as $N \to +\infty$, according to

$$\max_{T \in \mathscr{T}_N} h_T \leq A N^{-1/d}, \tag{46}$$

for some $A > 0$ which may be arbitrarily large.

Theorem 3.3. *Under the restriction (46), there exists a constant $c > 0$ that only depends on m, R and the constant C in (2) such that*

$$\liminf_{N \to +\infty} N^{m/d} \sigma_N(f)_p \geq c |f|_{W^{m,\tau}} \tag{47}$$

for all $f \in C^m(\Omega)$, where $\frac{1}{\tau} := \frac{1}{p} + \frac{m}{d}$.

Proof: If $f \in C^m(\Omega)$ and $x \in \Omega$, we denote by q_x the Taylor polynomial of order m at the point $x = (x_1, \cdots, x_d)$:

$$q_x(y) = q_x(y_1, \cdots, y_d) := \sum_{|\alpha| \leq m} \frac{1}{|\alpha|!} \partial^\alpha f(x) (y_1 - x_1)^{\alpha_1} \cdots (y_d - x_d)^{\alpha_d}. \tag{48}$$

If $\mathscr{T}_N$ is a partition in $\mathscr{A}_N$, we may write for each element $T \in \mathscr{T}_N$ and $x \in T$

$$\begin{aligned} e_{m,T}(f)_p &\geq e_{m,T}(q_x)_p - \|f - q_x\|_{L^p(T)} \\ &\geq C_1 |q_x|_{W^{m,\tau}(T)} - \|f - q_x\|_{L^p(T)} \\ &\geq c |f|_{W^{m,\tau}(T)} - C_1 |f - q_x|_{W^{m,\tau}(T)} - \|f - q_x\|_{L^p(T)}, \end{aligned}$$

with $c := C_1 \min\{1, \tau\}$, where we have used the lower bound in (45) and the quasi-triangle inequality

$$\|u + v\|_{L^\tau} \leq \max\{1, \tau^{-1}\} (\|u\|_{L^\tau} + \|v\|_{L^\tau}).$$

By the continuity of the m-th order derivative of f, we are ensured that for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|x - y| \leq \delta \Rightarrow |f(y) - q_x(y)| \leq \varepsilon |x - y|^m \text{ and } |d^m f(y) - d^m q_x|_\tau \leq \varepsilon. \tag{49}$$

Therefore if $N \geq N_0$ such that $A N_0^{-1/d} \leq \delta$, we have

$$\begin{aligned} e_{m,T}(f)_p &\geq c |f|_{W^{m,\tau}(T)} - (C_1 \varepsilon |T|^{1/\tau} + \varepsilon h_T^m |T|^{1/p}) \\ &\geq c |f|_{W^{m,\tau}(T)} - (1 + C_1) \varepsilon h_T^{m+d/p} \\ &\geq c |f|_{W^{m,\tau}(T)} - C \varepsilon N^{-1/\tau}, \end{aligned}$$

where the constant C depends on C_1 in (45) and A in (46). Using triangle inequality, it follows that

$$e_{m,\mathscr{T}_N}(f)_p = \Big(\sum_{T\in\mathscr{T}} e_{m,T}(f)_p^p\Big)^{1/p} \geq c\Big(\sum_{T\in\mathscr{T}} |f|^p_{W^{m,\tau}(T)}\Big)^{1/p} - C\varepsilon N^{-m/d}.$$

Using Hölder's inequality, we find that

$$|f|_{W^{m,\tau}} = \Big(\sum_{T\in\mathscr{T}} |f|^\tau_{W^{m,\tau}(T)}\Big)^{1/\tau} \leq N^{m/d}\Big(\sum_{T\in\mathscr{T}} |f|^p_{W^{m,\tau}(T)}\Big)^{1/p}, \tag{50}$$

which combined with the previous estimates shows that

$$N^{m/d} e_{m,\mathscr{T}_N}(f)_p \geq c|f|_{W^{m,\tau}} - C\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary this concludes the proof. □

Remark 3.5. The Hölder's inequality (50) becomes an equality if and only if all quantities in the sum are equal, which justifies the error equidistribution principle since these quantities are approximations of $e_{m,T}(f)_p$.

We next show that if $f \in C^m(\Omega)$, the adaptive approximations obtained by the greedy refinement algorithm introduced in §3.3 satisfy an upper estimate which closely matches the lower estimate (47).

Theorem 3.4. *There exists a constant C that only depends on m, R and on the choice of the hierarchy $(\mathscr{D}_j)_{j\geq 0}$ such that for all $f \in C^m(\Omega)$, the partitions $\mathscr{T}_N$ obtained by the greedy algorithm satisfy.*

$$\limsup_{N\to+\infty} N^{m/d} e_{m,\mathscr{T}_N}(f)_p \leq C|f|_{W^{m,\tau}}, \tag{51}$$

where $\frac{1}{\tau} := \frac{1}{p} + \frac{m}{d}$. In turn, for adaptively refined partitions, we have

$$\limsup_{N\to+\infty} N^{m/d} \sigma_N(f)_p \leq C|f|_{W^{m,\tau}}, \tag{52}$$

for all $f \in C^m(\Omega)$.

Proof: For any $\varepsilon > 0$, we choose $\delta > 0$ such that (49) holds. We first remark that there exists $N(\delta)$ sufficiently large such that for any $N \geq N(\delta)$ at least $N/2$ elements $T \in \mathscr{T}_N$ have parents with diameter $h_{P(T)} \leq \delta$. Indeed, the uniform isotropy of the elements ensures that

$$|T| \geq c h^d_{P(T)},$$

for some fixed constant $c > 0$. We thus have

$$\#\{T \in \mathscr{T}_N \,;\, h_{P(T)} \geq \delta\} \leq \frac{|\Omega|}{c\delta^d},$$

and the right-hand side is less than $N/2$ for large enough N. We denote by $\tilde{\mathscr{T}}_N$ the subset of $T \in \mathscr{T}_N$ such that $h_{P(T)} \leq \delta$. Defining η as previously by (42), we observe that for all $T \in \tilde{\mathscr{T}}_N \setminus \mathscr{D}_0$, we have

$$\eta \leq e_{m,P(T)}(f)_p. \tag{53}$$

If x is any point contained in T and q_x the Taylor polynomial of f at this point defined by (48), we have

$$\begin{aligned} e_{m,P(T)}(f)_p &\leq e_{m,P(T)}(q_x)_p + \|f - q_x\|_{L^p(P(T))} \\ &\leq C_2 |q_x|_{W^{m,\tau}(P(T))} + \varepsilon h_{P(T)}^m |P(T)|^{1/p} \\ &\leq C_2 \left(\tfrac{|P(T)|}{|T|}\right)^{1/\tau} |q_x|_{W^{m,\tau}(T)} + \varepsilon h_{P(T)}^m |P(T)|^{1/p} \\ &\leq C_2 \left(\tfrac{|P(T)|}{|T|}\right)^{1/\tau} |f|_{W^{m,\tau}(T)} + \varepsilon D_2 \left(\tfrac{|P(T)|}{|T|}\right)^{1/\tau} |T|^{1/\tau} + \varepsilon h_{P(T)}^m |P(T)|^{1/p}, \end{aligned}$$

where C_2 is the constant appearing in (45) and $D_2 := C_2 \max\{1, 1/\tau\}$. Combining this with (53), we obtain that for all $T \in \tilde{\mathscr{T}}_N$,

$$\eta \leq D(|f|_{W^{m,\tau}(T)} + \varepsilon |T|^{1/\tau})$$

where the constant D depends on C_2, m and on the refinement rule defining the hierarchy $(\mathscr{D}_j)_{j \geq 0}$. Elevating to the power τ and summing on all $T \in \tilde{\mathscr{T}}_N$, we thus obtain

$$(N/2 - N_0)\eta^\tau \leq \max\{1, \tau\} D^\tau (|f|^\tau_{W^{m,\tau}} + \varepsilon^\tau |\Omega|),$$

where $N_0 := \#(\mathscr{D}_0)$. Combining with (43), we therefore obtain

$$e_{m,\mathscr{T}_N}(f)_p \leq D \max\{\tau^{\frac{1}{\tau}}, 1/\tau\} N^{1/p} (N/2 - N_0)^{-1/\tau} (|f|_{W^{m,\tau}} + \varepsilon |\Omega|^{1/\tau}).$$

Taking $N > 4N_0$ and remarking that $\varepsilon > 0$ is arbitrary, we conclude that (52) holds with $C = 4^{1/\tau} D \max\{\tau^{\frac{1}{\tau}}, 1/\tau\}$. □

Theorems 3.3 and 3.4 reveal that for smooth enough functions, the numerical quantity that governs the rate of convergence $N^{-m/d}$ in the L^p norm of piecewise polynomial approximations on adaptive isotropic partitions is exactly $|f|_{W^{m,\tau}}$. In a similar way one would obtain that the same rate for quasi-uniform partitions is governed by the quantity $|f|_{W^{m,p}}$. Note however that these results are of asymptotic nature since they involve lim sup and lim inf as $N \to +\infty$, in contrast to Theorem 3.2. The results dealing with piecewise polynomial approximation on anisotropic adaptive partitions that we present in the next sections are of a similar asymptotic nature.

4 Anisotropic piecewise constant approximation on rectangles

We first explore a simple case of adaptive approximation on anisotropic partitions in two space dimensions. More precisely, we consider piecewise constant approximation in the L^p norm on adaptive partitions by rectangles with sides parallel to the x and y axes. In order to build such partitions, Ω cannot be any polygonal domain, and for the sake of simplicity we fix it to be the unit square:

$$\Omega = [0,1]^2.$$

The family $\mathscr{A}_N$ consists therefore of all partitions of Ω of at most N rectangles of the form

$$T = I \times J,$$

where I and J are intervals contained in $[0,1]$. This type of adaptive anisotropic partitions suffers from a strong coordinate bias due to the special role of the x and y direction: functions with sharp transitions on line edges are better approximated when these eges are parallel to the x and y axes. We shall remedy this defect in §5 by considering adaptive piecewise polynomial approximation on anisotropic partitions consisting of triangles, or simplices in higher dimension. Nevertheless, this first simple example is already instructive. In particular, it reveals that the numerical quantity governing the rate of approximation has an inherent non-linear structure. Throughout this section, we assume that f belongs to $C^1([0,1]^2)$.

4.1 A heuristic estimate

We first establish an error estimate which is based on the heuristic assumption that the partition is sufficiently fine so that we may consider that ∇f is constant on each T, or equivalently f coincides with an affine function $q_T \in \mathbb{P}_1$ on each T. We thus first study the local L^p approximation error on $T = I \times J$ for an affine function of the form

$$q(x,y) = q_0 + q_x x + q_y y.$$

Denoting by $\mathbf{q}(x,y) := q_x x + q_y y$ the homogeneous linear part of q, we first remark that

$$e_{1,T}(q)_p = e_{1,T}(\mathbf{q})_p, \tag{54}$$

since q and $\mathbf{q}$ differ by a constant. We thus concentrate on $e_{1,T}(\mathbf{q})_p$ and discuss the shape of T that minimizes this error when the area $|T| = 1$ is prescribed. We associate to this optimization problem a function K_p that acts on the space of linear functions according to

$$K_p(\mathbf{q}) = \inf_{|T|=1} e_{1,T}(\mathbf{q})_p. \tag{55}$$

As we shall explain further, the above infimum may or may not be attained.

We start by some observations that can be derived by elementary change of variable. If $a+T$ is a translation of T, then

$$e_{1,a+T}(\mathbf{q})_p = e_{1,T}(\mathbf{q})_p \tag{56}$$

since $\mathbf{q}$ and $\mathbf{q}(\cdot - a)$ differ by a constant. Therefore, if T is a minimizing rectangle in (55), then $a+T$ is also one. If hT is a dilation of T, then

$$e_{1,hT}(\mathbf{q})_p = h^{2/p+1} e_{1,T}(\mathbf{q})_p \tag{57}$$

Therefore, if we are interested in minimizing the error for an area $|T| = A$, we find that

$$\inf_{|T|=A} e_{1,T}(q)_p = A^{1/\tau} K_p(\mathbf{q}), \quad \frac{1}{\tau} := \frac{1}{p} + \frac{1}{2} \tag{58}$$

and the minimizing rectangles for (58) are obtained by rescaling the minimizing rectangles for (55).

In order to compute $K_p(\mathbf{q})$, we thus consider a rectangle $T = I \times J$ of unit area which barycenter is the origin. In the case $p = \infty$, using the notation $X := |q_x|\,|I|/2$ and $Y := |q_y|\,|J|/2$, we obtain

$$e_{1,T}(\mathbf{q})_\infty = X + Y.$$

We are thus interested in the minimization of the function $X+Y$ under the constraint $XY = |q_x q_y|/4$. Elementary computations show that when $q_x q_y \neq 0$, the infimum is attained when $X = Y = \frac{1}{2}\sqrt{|q_y q_x|}$ which yields

$$|I| = \sqrt{\frac{|q_y|}{|q_x|}} \text{ and } |J| = \sqrt{\frac{|q_x|}{|q_y|}}.$$

Note that the optimal aspect ratio is given by the simple relation

$$\frac{|I|}{|J|} = \frac{|q_y|}{|q_x|}, \tag{59}$$

which expresses the intuitive fact that the refinement should be more pronounced in the direction where the function varies the most. Computing $e_{1,T}(q)_\infty$ for such an optimized rectangle, we find that

$$K_\infty(\mathbf{q}) = \sqrt{|q_y q_x|}. \tag{60}$$

In the case $p = 2$, we find that

$$\begin{aligned} e_{1,T}(\mathbf{q})_2^2 &= \int_{-|I|/2}^{|I|/2}\int_{-|J|/2}^{|J|/2} |q_x x + q_y y|^2 dy\,dx \\ &= \int_{-|I|/2}^{|I|/2}\int_{-|J|/2}^{|J|/2} (q_x^2 x^2 + q_y^2 y^2 + 2q_x q_y xy) dy\,dx \\ &= 4\int_0^{|I|/2}\int_0^{|J|/2} (q_x^2 x^2 + q_y^2 y^2) dy\,dx \\ &= \tfrac{4}{3}(q_x^2(|I|/2)^3|J|/2 + q_y^2(|J|/2)^3|I|/2) \\ &= \tfrac{1}{3}(X^2+Y^2), \end{aligned}$$

where we have used the fact that $|I|\,|J| = 1$. We now want to minimize the function X^2+Y^2 under the constraint $XY = |q_x q_y|/4$. Elementary computations again show that when $q_x q_y \neq 0$, the infimum is again attained when $X = Y = \frac{1}{2}\sqrt{|q_y q_x|}$, and therefore leads to the same aspect ratio given by (59), and the value

$$K_2(\mathbf{q}) = \frac{1}{\sqrt{6}}\sqrt{|q_x q_y|}. \tag{61}$$

For other values of p the computation of $e_{1,T}(\mathbf{q})_p$ is more tedious, but leads to a same conclusion: the optimal aspect ratio is given by (59) and the function K_p has the general form

$$K_p(\mathbf{q}) = C_p\sqrt{|q_x q_y|}, \tag{62}$$

with $C_p := \left(\frac{2}{(p+1)(p+2)}\right)^{1/p}$. Note that the optimal shape of T does not depend on the L^p metric in which we measure the error.

By (54), (56) and (57), we find that for shape-optimized triangles of arbitrary area, the error is given by

$$e_{1,T}(q)_p = |T|^{1/\tau} K_p(\mathbf{q})_p = C_p\sqrt{|q_y q_x|}|T|^{1/\tau}, \tag{63}$$

Note that C_p is uniformly bounded for all $p \geq 1$.

In the case where $q \neq 0$ but $q_x q_y = 0$, the infimum in (55) is not attained, and the rectangles of a minimizing sequence tend to become infinitely long in the direction where q is constant. We ignore at the moment this degenerate case.

Since we have assumed that f coincides with an affine function on T, the estimate (63) yields

$$e_{1,T}(f)_p = C_p\left\|\sqrt{|\partial_x f \partial_y f|}\right\|_{L^\tau(T)} = \|K_p(\nabla f)\|_{L^\tau}, \quad \frac{1}{\tau} := \frac{1}{p} + \frac{1}{2}. \tag{64}$$

where we have identifed ∇f to the linear function $(x,y) \mapsto x\partial_x f + y\partial_y f$. This local estimate should be compared to those which were discussed in §3.1 for isotropic elements: in the bidimensional case, the estimate (30) of Theorem 3.1 can be restated as

$$e_{1,T}(f)_p \leq C\|\nabla f\|_{L^\tau(T)}, \quad \frac{1}{\tau} := \frac{1}{p} + \frac{1}{2}.$$

The improvement in (64) comes the fact that $\sqrt{|\partial_x f \partial_y f|}$ may be substantially smaller than $|\nabla f|$ when $|\partial_x f|$ and $|\partial_y f|$ have different order of magnitude which

reflects an anisotropic behaviour for the x and y directions. However, let us keep in mind that the validity of (64) is only when f is identified to an affine function on T.

Assume now that the partition $\mathscr{T}_N$ is built in such a way that all rectangles have optimal shape in the above described sense, and obeys in addition the error equidistribution principle, which by (64) means that

$$\|K_p(\nabla f)\|_{L^\tau(T)} = \eta, \ \ T \in \mathscr{T}_N.$$

Then, we have on the one hand that

$$e_{1,\mathscr{T}_N}(f)_p \leq \eta N^{1/p},$$

and on the other hand, that

$$N\eta^\tau \leq \|K_p(\nabla f)\|_{L^\tau}^\tau.$$

Combining the two above, and using the relation $\frac{1}{\tau} := \frac{1}{p} + \frac{1}{2}$, we thus obtain the error estimate

$$\sigma_N(f)_p \leq N^{-1/2}\|K_p(\nabla f)\|_{L^\tau}. \tag{65}$$

This estimate should be compared with those which were discussed in §3.2 for adaptive partition with isotropic elements: for piecewise constant functions on adaptive isotropic partitions in the two dimensional case, the estimate (38) can be restated as

$$\sigma_N(f)_p \leq CN^{-1/2}\|\nabla f\|_{L^\tau}, \ \ \frac{1}{\tau} = \frac{1}{p} + \frac{1}{2}.$$

As already observed for local estimates, the improvement in (64) comes from the fact that $|\nabla f|$ is replaced by the possibly much smaller $\sqrt{|\partial_x f \partial_y f|}$. It is interesting to note that the quantity

$$A_p(f) := \|K_p(\nabla f)\|_{L^\tau} = C_p \left\| \sqrt{|\partial_x f \partial_y f|} \right\|_{L^\tau},$$

is strongly nonlinear in the sense that it does not satisfy for any f and g an inequality of the type $A_p(f+g) \leq C(A_p(f) + A_p(g))$, even with $C > 1$. This reflects the fact that two functions f and g may be well approximated by piecewise constants on anisotropic rectangular partitions while their sum $f+g$ may not be.

4.2 A rigourous estimate

We have used heuristic arguments to derive the estimate (65), and a simple example shows that this estimate cannot hold as such: if f is a non-constant function that only depends on the variable x or y, the quantity $A_p(f)$ vanishes while the error $\sigma_N(f)_p$ may be non-zero. In this section, we prove a valid estimate by a rigourous

derivation. The price to pay is in the asymptotic nature of the new estimate, which has a form similar to those obtained in §3.4.

We first introduce a "tamed" variant of the function K_p, in which we restrict the search of the infimum to rectangles of limited diameter. For $M > 0$, we define

$$K_{p,M}(\mathbf{q}) = \min_{|T|=1, h_T \leq M} e_{1,T}(\mathbf{q})_p. \tag{66}$$

In contrast to the definition of K_p, the above minimum is always attained, due to the compactness in the Hausdorff distance of the set of rectangles of area 1, diameter less or equal to M, and centered at the origin. It is also not difficult to check that the functions $\mathbf{q} \mapsto e_{1,T}(\mathbf{q})_p$ are uniformly Lipschitz continuous for all T of area 1 and diameter less than M: there exists a constant C_M such that

$$|e_{1,T}(\mathbf{q})_p - e_{1,T}(\tilde{\mathbf{q}})_p| \leq C_M |\mathbf{q} - \tilde{\mathbf{q}}|, \tag{67}$$

where $|\mathbf{q}| := (q_x^2 + q_y^2)^{1/2}$. In turn $K_{p,M}$ is also Lipschitz continuous with constant C_M. Finally, it is obvious that $K_{p,M}(\mathbf{q}) \to K_p(\mathbf{q})$ as $M \to +\infty$.

If f is a C^1 function, we denote by

$$\omega(\delta) := \max_{|z-z'| \leq \delta} |\nabla f(z) - \nabla f(z')|,$$

the modulus of continuity of ∇f, which satisfies $\lim_{\delta \to 0} \omega(\delta) = 0$. We also define for all $z \in \Omega$

$$q_z(z') = f(z) + \nabla f \cdot (z' - z),$$

the Taylor polynomial of order 1 at z. We identify its linear part to the gradient of f at z:

$$\mathbf{q}_z = \nabla f(z).$$

We thus have

$$|f(z') - q_z(z')| \leq |z - z'| \omega(|z - z'|).$$

At each point z, we denote by $T_M(z)$ a rectangle of area 1 which is shape-optimized with respect to the gradient of f at z in the sense that it solves (66) with $\mathbf{q} = \mathbf{q}_z$. The following results gives an estimate of the local error for f for such optimized triangles.

Lemma 4.1. *Let $T = a + hT_M(z)$ be a rescaled and shifted version of $T_M(z)$. We then have for any $z' \in T$*

$$e_{1,T}(f)_p \leq (K_{p,M}(\mathbf{q}_{z'}) + B_M \omega(\max\{|z - z'|, h_T\}))|T|^{1/\tau},$$

with $B_M := 2C_M + M$.

Proof: For all $z, z' \in \Omega$, we have

$$\begin{aligned} e_{1,T_M}(\mathbf{q}_{z'}) &\leq e_{1,T_M}(\mathbf{q}_z) + C_M|\mathbf{q}_z - \mathbf{q}_{z'}| \\ &= K_{p,M}(\mathbf{q}_z) + C_M|\mathbf{q}_z - \mathbf{q}_{z'}| \\ &\leq K_{p,M}(\mathbf{q}_{z'}) + 2C_M|\mathbf{q}_z - \mathbf{q}_{z'}| \\ &\leq K_{p,M}(\mathbf{q}_{z'}) + 2C_M\omega(|z-z'|). \end{aligned}$$

We then observe that if $z' \in T$

$$\begin{aligned} e_{1,T}(f)_p &\leq e_{1,T}(\mathbf{q}_{z'}) + \|f - q_{z'}\|_{L^p(T)} \\ &\leq e_{1,T_M}(\mathbf{q}_{z'})|T|^{1/\tau} + \|f - q_{z'}\|_{L^\infty(T)}|T|^{1/p} \\ &\leq (K_{p,M}(\mathbf{q}_{z'}) + 2C_M\omega(|z-z'|))|T|^{1/\tau} + h_T\omega(h_T)|T|^{1/p} \\ &\leq (K_{p,M}(\mathbf{q}_{z'}) + 2C_M\omega(|z-z'|) + M\omega(h_T))|T|^{1/\tau}, \end{aligned}$$

which concludes the proof. □

We are now ready to state our main convergence theorem.

Theorem 4.1. *For piecewise constant approximation on adaptive anisotropic partitions on rectangles, we have*

$$\limsup_{N\to+\infty} N^{1/2}\sigma_N(f)_p \leq \|K_p(\nabla f)\|_{L^\tau}. \tag{68}$$

for all $f \in C^1([0,1]^2)$.

Proof: We first fix some number $\delta > 0$ and $M > 0$ that are later pushed towards 0 and $+\infty$ respectively. We define a uniform partition $\mathscr{T}_\delta$ of $[0,1]$ into squares S of diameter $h_S \leq \delta$, for example by j_0 iterations of uniform dyadic refinement, where j_0 is chosen large enough that $2^{-j_0+1/2} \leq \delta$. We then build partitions $\mathscr{T}_N$ by further decomposing the square elements of $\mathscr{T}_\delta$ in an anisotropic way. For each $S \in \mathscr{T}_\delta$, we pick an arbitrary point $z_S \in S$ (for example the barycenter of S) and consider the Taylor polynomial q_{z_S} of degree 1 of f at this point. We denote by $T_S = T_M(\mathbf{q}_{z_S})$ the rectangle of area 1 such that,

$$e_{1,T_S}(\mathbf{q}_{z_S})_p = \min_{|T|=1, h_T\leq M} e_{1,T}(\mathbf{q}_{z_S})_p = K_{p,M}(\mathbf{q}_{z_S}).$$

For $h > 0$, we rescale this rectangle according to

$$T_{h,S} = h(K_{p,M}(\mathbf{q}_{z_S}) + (B_M + C_M)\omega(\delta) + \delta)^{-\tau/2}T_S.$$

and we define $\mathscr{T}_{h,S}$ as the tiling of the plane by $T_{h,S}$ and its translates. We assume that $hC_A \leq \delta$ so that $h_T \leq \delta$ for all $T \in \mathscr{T}_{h,S}$ and all S. Finally, we define the partition

$$\mathscr{T}_N = \{T \cap S;\ T \in \mathscr{T}_{h,S} \text{ and } S \in \mathscr{T}_\delta\}.$$

We first estimate the local approximation error. By lemma (4.1), we obtain that for all $T \in \mathscr{T}_{h,S}$ and $z' \in T \cap S$

$$\begin{aligned} e_{1,T\cap S}(f)_p &\le e_{1,T}(f)_p \\ &\le (K_{p,M}(\mathbf{q}_{z'})+B_M\omega(\delta))|T|^{1/\tau} \\ &\le h^{2/\tau}(K_{p,M}(\mathbf{q}_{z_S})+(B_M+C_M)\omega(\delta))(K_{p,M}(\mathbf{q}_{z_S})+(B_M+C_M)\omega(\delta)+\delta)^{-1} \\ &\le h^{2/\tau} \end{aligned}$$

The rescaling has therefore the effect of equidistributing the error on all rectangles of $\mathscr{T}_N$, and the global approximation error is bounded by

$$e_{1,\mathscr{T}_N}(f)_p \le N^{1/p}h^{2/\tau} \tag{69}$$

We next estimate the number of rectangles $N = \#(\mathscr{T}_N)$, which behaves like

$$\begin{aligned} N &= (1+\eta(h))\textstyle\sum_{S\in\mathscr{T}_\delta}\frac{|S|}{|T_{h,S}|} \\ &= (1+\eta(h))h^{-2}\textstyle\sum_{S\in\mathscr{T}_\delta}|S|(K_{p,M}(\mathbf{q}_{z_S})+(B_M+C_M)\omega(\delta)+\delta)^\tau \\ &= (1+\eta(h))h^{-2}\textstyle\sum_{S\in\mathscr{T}_\delta}\int_S(K_{p,M}(\mathbf{q}_{z_S})+(B_M+C_M)\omega(\delta)+\delta)^\tau, \end{aligned}$$

where $\eta(h)\to 0$ as $h\to 0$. Recalling that $K_{p,M}(\mathbf{q}_{z_S})$ is Lipschitz continuous with constant C_M, it follows that

$$N \le (1+\eta(h))h^{-2}\int_\Omega (K_{p,M}(\mathbf{q}_z)+(B_M+2C_M)\omega(\delta)+\delta)^\tau. \tag{70}$$

Combining (69) and (70), we have thus obtained

$$N^{1/2}e_{1,\mathscr{T}_N}(f)_p \le (1+\eta(h))^{1/\tau}\|K_{p,M}(\mathbf{q}_z)+(B_M+2C_M)\omega(\delta)+\delta\|_{L^\tau}.$$

Observing that for all $\varepsilon>0$, we can choose M large enough and δ and h small enough so that

$$(1+\eta(h))^{1/\tau}\|K_{p,M}(\mathbf{q}_z)+(B_M+2C_M)\omega(\delta)+\delta\|_{L^\tau} \le \|K_{p,M}(\mathbf{q}_z)\|_{L^\tau}+\varepsilon,$$

this concludes the proof. □

In a similar way as in Theorem 3.3, we can establish a lower estimate on $\sigma_N(f)$, which reflects the saturation rate $N^{-1/2}$ of the method, and shows that the numerical quantity that governs this rate is exactly equal to $\|K_p(\nabla f)\|_{L^\tau}$. We again impose a slight restriction on the set $\mathscr{A}_N$ of admissible partitions, assuming that the diameter of all elements decreases as $N\to+\infty$, according to

$$\max_{T\in\mathscr{T}_N} h_T \le AN^{-1/2}, \tag{71}$$

for some $A>0$ which may be arbitrarily large.

Theorem 4.2. *Under the restriction* (71)*, we have*

$$\liminf_{N\to+\infty} N^{1/2}\sigma_N(f)_p \ge \|K_p(\nabla f)\|_{L^\tau} \tag{72}$$

for all $f \in C^1(\Omega)$, *where* $\frac{1}{\tau} := \frac{1}{p} + \frac{1}{2}$.

Proof: We assume here $p < \infty$. The case $p = \infty$ can be treated by a simple modification of the argument. Here, we need a lower estimate for the local approximation error, which is a counterpart to Lemma 4.1. We start by remarking that for all rectangle $T \in \Omega$ and $z \in T$, we have

$$|e_{1,T}(f)_p - e_{1,T}(q_z)_p| \leq \|f - q_z\|_{L^p(T)} \leq |T|^{1/p} h_T \omega(h_T),$$

and therefore

$$e_{1,T}(f)_p \geq e_{1,T}(q_z)_p - |T|^{1/p} h_T \omega(h_T) \geq K_p(\mathbf{q}_z)|T|^{1/\tau} - |T|^{1/p} h_T \omega(h_T)$$

Then, using the fact that if (a,b,c) are positive numbers such that $a \geq b - c$ one has $a^p \geq b^p - pcb^{p-1}$, we find that

$$\begin{aligned} e_{1,T}(f)_p^p &\geq K_p(\mathbf{q}_z)^p |T|^{p/\tau} - pK_p(\mathbf{q}_z)^{p-1} |T|^{(p-1)/\tau} |T|^{1/p} h_T \omega(h_T) \\ &= K_p(\mathbf{q}_z)^p |T|^{1+p/2} - pK_p(\mathbf{q}_z)^{p-1} |T|^{1+(p-1)/2} h_T \omega(h_T), \end{aligned}$$

Defining $C := p \max_{z \in \Omega} K_p(\mathbf{q}_z)^{p-1}$ and remarking that $|T|^{(p-1)/2} \leq h^{p-1}$, this leads to the estimate

$$e_{1,T}(f)_p^p \geq K_p(\mathbf{q}_z)^p |T|^{1+p/2} - Ch_T^p |T| \omega(h_T).$$

Since we work under the assumption (71), we can rewrite this estimate as

$$e_{1,T}(f)_p^p \geq K_p(\mathbf{q}_z)^p |T|^{1+p/2} - C|T| N^{-p/2} \varepsilon(N), \tag{73}$$

where $\varepsilon(N) \to 0$ as $N \to \infty$. Integrating (73) over T, gives

$$e_{1,T}(f)_p^p \geq \int_T (K_p(\mathbf{q}_z)^p |T|^{p/2} - CN^{-p/2} \varepsilon(N)) dz.$$

Summing over all rectangles $T \in \mathscr{T}_N$ and denoting by T_z the triangle that contains z, we thus obtain

$$e_{1,\mathscr{T}_N}(f)_p^p \geq \int_\Omega K_p(\nabla f(z))^p |T_z|^{p/2} dz - C|\Omega| N^{-p/2} \varepsilon(N). \tag{74}$$

Using Hölder inequality, we find that

$$\int_\Omega K_p(\nabla f(z))^\tau dz \leq \Big(\int_\Omega K_p(\nabla f(z))^p |T_z|^{p/2} dz \Big)^{\tau/p} \Big(\int_\Omega |T_z|^{-1} dz \Big)^{1-\tau/p}. \tag{75}$$

Since $\int_\Omega |T_z|^{-1} dz = \#(\mathscr{T}_N) = N$, it follows that

$$e_{1,\mathscr{T}_N}(f)_p^p \geq \|K_p(\nabla f)\|_{L^\tau}^p N^{-p/2} - C|\Omega| N^{-p/2} \varepsilon(N),$$

which concludes the proof. □

Remark 4.1. The Hölder inequality (75) which is used in the above proof becomes an equality when the quantity $K_p(\nabla f(z))^p |T_z|^{p/2}$ and $|T_z|^{-1}$ are proportional, i.e. $K_p(\nabla f(z))|T|^{1/\tau}$ is constant, which again reflects the principle of error equidistribution. In summary, the optimal partitions should combine this principe with locally optimized shapes for each element.

5 Anisotropic piecewise polynomial approximation

We turn to adaptive piecewise polynomial approximation on anisotropic partitions consisting of triangles, or simplices in higher dimension. Here $\Omega \subset \mathbb{R}^d$ is a domain that can be decomposed into such partitions, therefore a polygon when $d = 2$, a polyhedron when $d = 3$, etc. The family $\mathscr{A}_N$ consists therefore of all partitions of Ω of at most N simplices. The first estimates of the form (6) were rigorously established in [17] and [5] in the case of piecewise linear element for bidimensional triangulations. Generalization to higher polynomial degree as well as higher dimensions were recently proposed in [14, 15, 16] as well as in [39]. Here we follow the general approach of [39] to the characterization of optimal partitions.

5.1 The shape function

If f belongs to $C^m(\Omega)$, where $m-1$ is the degree of the piecewise polynomials that we use for approximation, we mimic the heuristic approach proposed for piecewise constants on rectangles in §4.1 by assuming that on each triangle T the relative variation of $d^m f$ is small so that it can be considered as a constant over T. This means that f is locally identified with its Taylor polynomial of degree m at z, which is defined as

$$q_z(z') := f(z) + \nabla f(z) \cdot (z'-z) + \sum_{k=2}^{m} \frac{1}{k!} d^k f(z)[z'-z, \cdots, z'-z].$$

If $q \in \mathbb{P}_{\mathrm{m}}$ is a polynomial of degree m, we denote by $\mathbf{q} \in \mathbb{H}^{\mathrm{m}}$ its homogeneous part of degree m. For $q = q_z$ we can identify $\mathbf{q}_z \in \mathbb{H}_{\mathrm{m}}$ with $\frac{1}{m!} d^m f(z)$. Since $\mathbf{q} - q \in \mathbb{P}_{\mathrm{m}-1}$ we have

$$e_{m,T}(q)_p = e_{m,T}(\mathbf{q})_p.$$

We optimize the shape of the simplex T with respect to $\mathbf{q}$ by introducing the function $K_{m,p}$ defined on the space $\mathbb{H}_{\mathrm{m}}$

$$K_{m,p}(\mathbf{q}) := \inf_{|T|=1} e_{m,T}(\mathbf{q})_p, \tag{76}$$

where the infimum is taken among all triangles of area 1. This infimum may or may not be attained. We refer to $K_{m,p}$ as the *shape function*. It is obviously a generalization of the function K_p introduced for piecewise constant on rectangles in §4.1.

As in the case of rectangles, some elementary properties of $K_{m,p}$ are obtained by change of variable: if $a+T$ is a shifted version of T, then

$$e_{m,a+T}(\mathbf{q})_p = e_{m,T}(\mathbf{q})_p \tag{77}$$

since $\mathbf{q}$ and $\mathbf{q}(\cdot - a)$ differ by a polynomial of degree $m-1$, and that if hT is a dilation of T, then

$$e_{m,hT}(\mathbf{q})_p = h^{d/p+m} e_{m,T}(\mathbf{q})_p \tag{78}$$

Therefore, if T is a minimizing simplex in (76), then $a+T$ is also one, and if we are interested in minimizing the error for a given area $|T| = A$, we find that

$$\inf_{|T|=A} e_{m,T}(q)_p = A^{1/\tau} K_{m,p}(\mathbf{q}), \quad \frac{1}{\tau} := \frac{1}{p} + \frac{m}{d} \tag{79}$$

and the minimizing simplex for (58) are obtained by rescaling the minimizing simplex for (55).

Remarking in addition that if φ is an invertible linear transform, we then have for all f

$$|\det(\varphi)|^{1/p} e_{m,T}(f \circ \varphi)_p = e_{m,\varphi(T)}(f)_p,$$

and using (79), we also obtain that

$$K_{m,p}(\mathbf{q} \circ \varphi) = |\det(\varphi)|^m K_{m,p}(\mathbf{q}) \tag{80}$$

The minimizing simplex of area 1 for $\mathbf{q} \circ \varphi$ is obtained by application of φ^{-1} followed by a rescaling by $|\det(\varphi)|^{1/d}$ to the minimizing simplex of area 1 for $\mathbf{q}$ if it exists.

5.2 Algebraic expressions of the shape function

The identity (80) can be used to derive the explicit expression of $K_{m,p}$ for particular values of (m,p,d), as well as the exact shape of the minimizing triangle T in (76).

We first consider the case of piecewise affine elements on two dimensional triangulations, which corresponds to $d = m = 2$. Here $\mathbf{q}$ is a quadratic form and we denote by $\det(\mathbf{q})$ its determinant. We also denote by $|\mathbf{q}|$ the positive quadratic form associated with the absolute value of the symmetric matrix associated to $\mathbf{q}$.

If $\det(\mathbf{q}) \neq 0$, there exists a φ such that $\mathbf{q} \circ \varphi$ is either x^2+y^2 or x^2-y^2, up to a sign change, and we have $|\det(\mathbf{q})| = |\det(\varphi)|^{-2}$. It follows from (80) that $K_{2,p}(\mathbf{q})$ has the simple form

$$K_{2,p}(\mathbf{q}) = \kappa_p |\det(\mathbf{q})|^{1/2}, \tag{81}$$

where $\kappa_p := K_{2,p}(x^2+y^2)$ if $\det(\mathbf{q}) > 0$ and $\kappa_p = K_{2,p}(x^2-y^2)$ if $\det(\mathbf{q}) < 0$.

The triangle of area 1 that minimizes the L^p error when $\mathbf{q} = x^2 + y^2$ is the equilateral triangle, which is unique up to rotations. For $\mathbf{q} = x^2 - y^2$, the triangle that minimizes the L^p error is unique up to an hyperbolic transformation with eigenvalues t and $1/t$ and eigenvectors $(1,1)$ and $(1,-1)$ for any $t \neq 0$. Therefore, such triangles may be highly anisotropic, but at least one of them is isotropic. For example, it can be checked that a triangle of area 1 that minimizes the L^∞ error is given by the half square with vertices $((0,0),(\sqrt{2},0),(0,\sqrt{2}))$. It can also be checked that an equilateral triangle T of area 1 is a "near-minimizer" in the sense that

$$e_{2,T}(\mathbf{q})_p \leq C K_{2,p}(\mathbf{q}),$$

where C is a constant independent of p. It follows that when $\det(\mathbf{q}) \neq 0$, the triangles which are isotropic with respect to the distorted metric induced by $|\mathbf{q}|$ are "optimally adapted" to $\mathbf{q}$ in the sense that they nearly minimize the L^p error among all triangles of similar area.

In the case when $\det(\mathbf{q}) = 0$, which corresponds to one-dimensional quadratic forms $\mathbf{q} = (ax+by)^2$, the minimum in (76) is not attained and the minimizing triangles become infinitely long along the null cone of $\mathbf{q}$. In that case one has $K_{2,p}(\mathbf{q}) = 0$ and the equality (81) remains therefore valid.

These results easily generalize to piecewise affine functions on simplicial partitions in higher dimension $d > 1$: one obtains

$$K_{2,p}(\mathbf{q}) = \kappa_p |\det(\mathbf{q})|^{1/d}, \tag{82}$$

where κ_p only takes a finite number of possible values. When $\det(\mathbf{q}) \neq 0$, the simplices which are isotropic with respect to the distorted metric induced by $|\mathbf{q}|$ are "optimally adapted" to $\mathbf{q}$ in the sense that they nearly minimize the L^p error among all simplices of similar volume.

The analysis becomes more delicate for higher polynomial degree $m \geq 3$. For piecewise quadratic elements in dimension two, which corresponds to $m = 3$ and $d = 2$, it is proved in [39] that

$$K_{3,p}(\mathbf{q}) = \kappa_p |\mathrm{disc}(\mathbf{q})|^{1/4}.$$

for any *homogeneous* polynomial $\mathbf{q} \in \mathbb{H}_3$, where

$$\mathrm{disc}(ax^3 + bx^2y + cxy^2 + dy^3) := b^2c^2 - 4ac^3 - 4b^3d + 18abcd - 27a^2d^2,$$

is the usual discriminant and κ_p only takes two values depending on the sign of $\mathrm{disc}(\mathbf{q})$. The analysis that leads to this result also describes the shape of the triangles which are optimally adapted to $\mathbf{q}$.

For other values of m and d, the exact expression of $K_{m,p}(\mathbf{q})$ is unknown, but it is possible to give equivalent versions in terms of polynomials $Q_{m,d}$ in the coefficients of $\mathbf{q}$, in the following sense: for all $\mathbf{q} \in \mathbb{H}_{\mathrm{m}}$

$$c_1(Q_{m,d}(\mathbf{q}))^{\frac{1}{r}} \leq K_{3,p}(\mathbf{q}) \leq c_2(Q_{m,d}(\mathbf{q}))^{\frac{1}{r}},$$

where $r := \deg(Q_{m,d})$, see [39].

Remark 5.1. It is easily checked that the shape functions $\mathbf{q} \mapsto K_{m,p}(\mathbf{q})$ are equivalent for all p p in the sense that there exist constant $0 < C_1 \leq C_2$ that only depend on the dimension d such that

$$C_1 K_{m,\infty}(\mathbf{q}) \leq K_{m,p}(\mathbf{q}) \leq C_2 K_{m,\infty}(\mathbf{q}),$$

for all $\mathbf{q} \in \mathbf{H}_{\mathrm{m}}$ and $p \geq 1$. In particular a minimizing triangle for $K_{m,\infty}$ is a near-minimizing triangle for $K_{m,p}$. In that sense, the optimal shape of the element does not strongly depend on p.

5.3 Error estimates

Following at first a similar heuristics as in §4.1 for piecewise constants on rectangles, we assume that the triangulation $\mathscr{T}_N$ is such that all its triangles T have optimized shape with respect to the polynomial q that coincides with f on T.

According to (79), we thus have for any triangle $T \in \mathscr{T}$,

$$e_{m,T}(f)_p = |T|^{\frac{1}{\tau}} K_{m,p}(\mathbf{q}) = \left\| K_{m,p}\left(\frac{d^m f}{m!}\right) \right\|_{L^\tau(T)}.$$

We then apply the principle of *error equidistribution*, assuming that

$$e_{m,T}(f)_p = \eta,$$

From which it follows that $e_{m,\mathscr{T}_N}(f)_p \leq N^{1/p}\eta$ and

$$N\eta^\tau \leq \left\| K_{m,p}\left(\frac{d^m f}{m!}\right) \right\|_{L^\tau}^\tau,$$

and therefore

$$\sigma_N(f)_p \leq N^{-m/d} \left\| K_{m,p}\left(\frac{d^m f}{m!}\right) \right\|_{L^\tau}. \tag{83}$$

This estimate should be compared to (38) which was obtained for adaptive partitions with elements of isotropic shape. The essential difference is in the quantity $K_{m,p}\left(\frac{d^m f}{m!}\right)$ which replaces $d^m f$ in the L^τ norm, and which may be significantly smaller. Consider for example the case of piecewise affine elements, for which we can combine (83) with (82) to obtain

$$\sigma_N(f)_p \leq CN^{-2/d} \left\| |\det(d^2 f)|^{1/d} \right\|_{L^\tau}. \tag{84}$$

In comparison to (38), the norm of the hessian $|d^2 f|$ is replaced by the quantity $|\det(d^2 f)|^{1/d}$ which is geometric mean of its eigenvalues, a quantity which is sig-

nificantly smaller when two eigenvalues have different orders of magnitude which reflects an anisotropic behaviour in f.

As in the case of piecewise constants on rectangles, the example of a function f depending on only one variable shows that the estimate (84) cannot hold as such. We may obtain some valid estimates by following the same approach as in Theorem 4.1. This leads to the following result which is established in [39].

Theorem 5.1. *For piecewise polynomial approximation on adaptive anisotropic partitions into simplices, we have*

$$\limsup_{N\to+\infty} N^{m/d}\sigma_N(f)_p \leq C \left\| K_{m,p}\left(\frac{d^m f}{m!}\right)\right\|_{L^\tau}, \quad \frac{1}{\tau} := \frac{1}{p}+\frac{m}{d}, \tag{85}$$

for all $f \in C^m(\Omega)$. The constant C can be chosen equal to 1 in the case of two-dimensional triangulations $d = 2$.

The proof of this theorem follows exactly the same line as the one of Theorem 4.1: we build a sequence of partitions $\mathscr{T}_N$ by refining the triangles S of a sufficiently fine quasi-uniform partition $\mathscr{T}_\delta$, intersecting each S with a partition $\mathscr{T}_{h,S}$ by elements with shape optimally adapted to the local value of $d^m f$ on each S. The constant C can be chosen equal to 1 in the two-dimensional case, due to the fact that it is then possible to build $\mathscr{T}_{h,S}$ as a tiling of triangles which are all optimally adapted. This is no longer possible in higher dimension, which explains the presence of a constant $C = C(m,d)$ larger than 1.

We may also obtain lower estimates, following the same approach as in Theorem 4.2: we first impose a slight restriction on the set $\mathscr{A}_N$ of admissible partitions, assuming that the diameter of the elements decreases as $N \to +\infty$, according to

$$\max_{T\in\mathscr{T}_N} h_T \leq AN^{-1/d}, \tag{86}$$

for some $A > 0$ which may be arbitrarily large. We then obtain the following result, which proof is similar to the one of Theorem 4.2.

Theorem 5.2. *Under the restriction (86), we have*

$$\liminf_{N\to+\infty} N^{m/d}\sigma_N(f)_p \geq \left\| K_{m,p}\left(\frac{d^m f}{m!}\right)\right\|_{L^\tau} \tag{87}$$

for all $f \in C^m(\Omega)$, where $\frac{1}{\tau} := \frac{1}{p}+\frac{m}{d}$.

5.4 Anisotropic smoothness and cartoon functions

Theorem 5.1 reveals an improvement over the approximation results based on adaptive isotropic partitions in the sense that $\|K_{m,p}\left(\frac{d^m f}{m!}\right)\|_{L^\tau}$ may be significantly

smaller than $\|d^m f\|_{L^\tau}$, for functions which have an anisotropic behaviour. However, this result suffers from two major defects:

1. The estimate (85) is asymptotic: it says that for all $\varepsilon > 0$, there exists N_0 depending on f and ε such that

$$\sigma_N(f)_p \leq CN^{-m/d}\left(\left\|K_{m,p}\left(\frac{d^m f}{m!}\right)\right\|_{L^\tau} + \varepsilon\right),$$

   for all $N \geq N_0$. However, it does not ensure a uniform bound on N_0 which may be very large for certain f.

2. Theorem 5.1 is based on the assumption $f \in C^m(\Omega)$, and therefore the estimate (85) only seems to apply to sufficiently smooth functions. This is in contrast to the estimates that we have obtained for adaptive isotropic partitions, which are based on the assumption that $f \in W^{m,\tau}(\Omega)$ or $f \in B^m_{\tau,\tau}(\Omega)$.

The first defect is due to the fact that a certain amount of refinement should be performed before the relative variation of $d^m f$ is sufficiently small so that there is no ambiguity in defining the optimal shape of the simplices. It is in that sense unavoidable.

The second defect raises a legitimate question concerning the validity of the convergence estimate (85) for functions which are not in $C^m(\Omega)$. It suggests in particular to introduce a class of distributions such that

$$\left\|K_{m,p}\left(\frac{d^m f}{m!}\right)\right\|_{L^\tau} < +\infty,$$

and to try to understand if the estimate remains valid inside this class which describe in some sense functions which have a certain amount anisotropic smoothness. The main difficulty is that that this class is not well defined due to the nonlinear nature of $K_{m,p}\left(\frac{d^m f}{m!}\right)$. As an example consider the case of piecewise linear elements on two dimensional triangulation, that corresponds to $m = d = 2$. In this case, we have seen that $K_{2,p}(\mathbf{q}) = \kappa_p \sqrt{|\det(\mathbf{q})|}$. The numerical quantity that governs the approximation rate N^{-1} is thus

$$A_p(f) := \left\|\sqrt{|\det(d^2 f)|}\right\|_{L^\tau}, \quad \frac{1}{\tau} = \frac{1}{p} + 1.$$

However, this quantity cannot be defined in the distribution sense since the product of two distributions is generally ill-defined. On the other hand, it is known that the rate N^{-1} can be achieved for functions which do not have C^2 smoothness, and which may even be discontinuous along curved edges. Specifically, we say that f is a *cartoon function* on Ω if it is almost everywhere of the form

$$f = \sum_{1 \leq i \leq k} f_i \chi_{\Omega_i},$$

where the Ω_i are disjoint open sets with piecewise C^2 boundary, no cusps (i.e. satisfying an interior and exterior cone condition), and such that $\overline{\Omega} = \cup_{i=1}^k \overline{\Omega}_i$, and where

for each $1 \leq i \leq k$, the function f_i is C^2 on a neighbourhood of $\overline{\Omega}_i$. Such functions are a natural candidates to represent images with sharp edges or solutions of PDE's with shock profiles.

Let us consider a fixed cartoon function f on a polygonal domain Ω associated with a partition $(\Omega_i)_{1\leq i\leq k}$. We define

$$\Gamma := \bigcup_{1\leq i\leq k} \partial\Omega_i,$$

the union of the boundaries of the Ω_i. The above definition implies that Γ is the disjoint union of a finite set of points $\mathscr{P}$ and a finite number of open curves $(\Gamma_i)_{1\leq i\leq l}$.

$$\Gamma = \Big(\bigcup_{1\leq i\leq l} \Gamma_i \Big) \cup \mathscr{P}.$$

If we consider the approximation of f by piecewise affine function on a triangulation $\mathscr{T}_N$ of cardinality N, we may distinguish two types of elements of $\mathscr{T}_N$. A triangle $T \in \mathscr{T}_N$ is called "regular" if $T \cap \Gamma = \emptyset$, and we denote the set of such triangles by $\mathscr{T}_N^r$. Other triangles are called "edgy" and their set is denoted by $\mathscr{T}_N^e$. We can thus split Ω according to

$$\Omega := (\cup_{T\in\mathscr{T}_N^r} T) \cup (\cup_{T\in\mathscr{T}_N^e} T) = \Omega_N^r \cup \Omega_N^e.$$

We split accordingly the L^p approximation error into

$$e_{2,\mathscr{T}_N}(f)_p^p = \sum_{T\in\mathscr{T}_N^r} e_{2,T}(f)_p^p + \sum_{T\in\mathscr{T}_N^e} e_{2,T}(f)_p^p.$$

We may use $\mathscr{O}(N)$ triangles in $\mathscr{T}_N^e$ and $\mathscr{T}_N^r$ (for example $N/2$ in each set). Since f has discontinuities along Γ, the approximation error on the edgy triangles does not tend to zero in L^∞ and $\mathscr{T}_N^e$ should be chosen so that Ω_N^e has the aspect of a thin layer around Γ. Since Γ is a finite union of C^2 curves, we can build this layer of width $\mathscr{O}(N^{-2})$ and therefore of global area $|\Omega_N^e| \leq CN^{-2}$, by choosing long and thin triangles in $\mathscr{T}_N^e$. On the other hand, since f is uniformly C^2 on Ω_N^r, we may choose all triangles in $\mathscr{T}_N^r$ of regular shape and diameter $h_T \leq CN^{-1/2}$. Hence we obtain the following heuristic error estimate, for a well designed anisotropic triangulation:

$$\begin{aligned} e_{2,\mathscr{T}_N}(f)_p &\leq \textstyle\sum_{T\in\mathscr{T}_N^r} |T| e_{2,T}(f)_\infty^p + \sum_{T\in\mathscr{T}_N^e} |T| e_{2,T}(f)_\infty^p \\ &\leq C|\Omega_N^r| (\sup_{T\in\mathscr{T}_N^r} h_T^2) \|d^2 f\|_{L^\infty(\Omega_N^r)}^p + C|\Omega_N^e| \|f\|_{L^\infty(\Omega_N^e)}^p, \end{aligned}$$

and therefore

$$e_{2,\mathscr{T}_N}(f)_p \leq CN^{-\min\{1,2/p\}}, \tag{88}$$

where the constant C depends on $\|d^2 f\|_{L^\infty(\Omega\setminus\Gamma)}$, $\|f\|_{L^\infty(\Omega)}$ and on the number, length and maximal curvature of the C^2 curves which constitute Γ.

These heuristic estimates have been discussed in [38] and rigorously proved in [25]. Observe in particular that the error is dominated by the edge contribution when

$p > 2$ and by the smooth contribution when $p < 2$. For the critical value $p = 2$ the two contributions have the same order.

For $p \geq 2$, we obtain the approximation rate N^{-1} which suggests that approximation results such as Theorem 5.1 should also apply to cartoon functions and that the quantity $A_p(f)$ should be finite for such functions. In some sense, we want to "bridge the gap" between results of anisotropic piecewise polynomial approximation for cartoon functions and for smooth functions. For this purpose, we first need to give a proper meaning to $A_p(f)$ when f is a cartoon function. As already explained, this is not straightforward, due to the fact that the product of two distributions has no meaning in general. Therefore, we cannot define $\det(d^2 f)$ in the distribution sense, when the coefficients of $d^2 f$ are distributions without sufficient smoothness.

We describe a solution to this problem proposed in [22] which is based on a regularization process. In the following, we consider a fixed radial nonnegative function φ of unit integral and supported in the unit ball, and define for all $\delta > 0$ and f defined on Ω,

$$\varphi_\delta(z) := \frac{1}{\delta^2}\varphi\left(\frac{z}{\delta}\right) \text{ and } f_\delta = f * \varphi_\delta. \tag{89}$$

It is then possible to gives a meaning to $A_p(f)$ based on this regularization. This approach is additionally justified by the fact that sharp curves of discontinuity are a mathematical idealisation. In real world applications, such as photography, several physical limitations (depth of field, optical blurring) impose a certain level of blur on the edges.

If f is a cartoon function on a set Ω, and if $x \in \Gamma \setminus \mathscr{P}$, we denote by $[f](x)$ the jump of f at this point. We also denote by $|\kappa(x)|$ the absolute value of the curvature at x. For $p \in [1,\infty]$ and τ defined by $\frac{1}{\tau} := 1 + \frac{1}{p}$, we introduce the two quantities

$$\begin{aligned} S_p(f) &:= \left\| \sqrt{|\det(d^2 f)|} \right\|_{L^\tau(\Omega\setminus\Gamma)} = A_p(f_{|\Omega\setminus\Gamma}), \\ E_p(f) &:= \|\sqrt{|\kappa|}[f]\|_{L^\tau(\Gamma)}, \end{aligned}$$

which respectively measure the "smooth part" and the "edge part" of f. We also introduce the constant

$$C_{p,\varphi} := \|\sqrt{|\Phi\Phi'|}\|_{L^\tau(\mathbb{R})}, \quad \Phi(x) := \int_{y\in\mathbb{R}} \varphi(x,y)dy. \tag{90}$$

Note that f_δ is only properly defined on the set

$$\Omega^\delta := \{z \in \Omega \,;\, B(z,\delta) \subset \Omega\},$$

and therefore, we define $A_p(f_\delta)$ as the L^τ norm of $\sqrt{|\det(d^2 f_\delta)|}$ on this set. The following result is proved in [22].

Theorem 5.3. *For all cartoon functions f, the quantity $A_p(f_\delta)$ behaves as follows:*

- *If $p < 2$, then*

$$\lim_{\delta\to 0} A_p(f_\delta) = S_p(f).$$

- *If $p = 2$, then $\tau = \frac{2}{3}$ and*

$$\lim_{\delta\to 0} A_2(f_\delta) = (S_2(f)^\tau + E_2(f)^\tau C_{2,\varphi}^\tau)^{1/\tau}.$$

- *If $p > 2$, then $A_p(f_\delta) \to \infty$ according to*

$$\lim_{\delta\to 0} \delta^{\frac{1}{2}-\frac{1}{p}} A_p(f_\delta) = E_p(f) C_{p,\varphi}.$$

Remark 5.2. This theorem reveals that as $\delta \to 0$, the contribution of the neighbourhood of Γ to $A_p(f_\delta)$ is neglectible when $p < 2$ and dominant when $p > 2$, which was already remarked in the heuristic computation leading to (88).

Remark 5.3. In the case $p = 2$, it is interesting to compare the limit expression $(S_2(f)^\tau + E_2(f)^\tau C_{2,\varphi}^\tau)^{1/\tau}$ with the *total variation* $TV(f) = |f|_{BV}$. For a cartoon function, the total variation also can be split into a contribution of the smooth part and a contribution of the edge, according to

$$TV(f) := \int_{\Omega\setminus\Gamma} |\nabla f| + \int_\Gamma |[f]|.$$

Functions of bounded variation are thus allowed to have jump discontinuities along edges of finite length. For this reason, *BV* is frequently used as a natural smoothness space to describe the mathematical properties of images. It is also well known that *BV* is a regularity space for certain hyperbolic conservation law, in the sense that the total variation of their solutions remains finite for all time $t > 0$. In recent years, it has been observed that the space *BV* (and more generally classical smoothness spaces) do not provide a fully satisfactory description of piecewise smooth functions arising in the above mentionned applications, in the sense that the total variation only takes into account the size of the sets of discontinuities and not their geometric smoothness. In contrast, we observe that the term $E_2(f)$ incorporates an information on the smoothness of Γ through the presence of the curvature $|\kappa|$. The quantity $A_2(f)$ appears therefore as a potential substitute to $TV(f)$ in order to take into account the geometric smoothness of the edges in cartoon function and images.

6 Anisotropic greedy refinement algorithms

In the two previous sections, we have established error estimates in L^p norms for the approximation of a function f by piecewise polynomials on optimally adapted anisotropic partitions. Our analysis reveals that the optimal partition needs to satisfy two intuitively desirable features:

1. Equidistribution of the local error.

2. Optimal shape adaptation of each element based on the local properties of f.

For instance, in the case of piecewise affine approximation on triangulations, these items mean that each triangle T should be close to equilateral with respect to a distorted metric induced by the local value of the hessian $d^2 f$.

From the computational viewpoint, a commonly used strategy for designing an optimal triangulation consists therefore in evaluating the hessian $d^2 f$ and imposing that each triangle is isotropic with respect to a metric which is properly related to its local value. We refer in particular to [10] and to [9] where this program is executed by different approaches, both based on Delaunay mesh generation techniques (see also the software package [45] which includes this type of mesh generator). While these algorithms produce anisotropic meshes which are naturally adapted to the approximated function, they suffer from two intrinsic limitations:

1. They are based on the data of $d^2 f$, and therefore do not apply well to non-smooth or noisy functions.
2. They are non-hierarchical: for $N > M$, the triangulation $\mathscr{T}_N$ is not a refinement of $\mathscr{T}_M$.

Similar remark apply to anisotropic mesh generation techniques in higher dimensions or for finite elements of higher degree.

The need for hierarchical partitions is critical in the construction of wavelet bases, which play an important role in applications to image and terrain data processing, in particular data compression [19]. In such applications, the multilevel structure is also of key use for the fast encoding of the information. Hierarchy is also useful in the design of optimally converging adaptive methods for PDE's [8, 40, 43]. However, all these developments are so far mostly limited to isotropic refinement methods, in the spirit of the refinement procedures discussed in §3. Let us mention that hierarchical *and* anisotropic triangulations have been investigated in [36], yet in this work the triangulations are *fixed in advance* and therefore generally not adapted to the approximated function.

A natural objective is therefore to design adaptive algorithmic techniques that combine hierarchy and anisotropy, that apply to any function $f \in L^p(\Omega)$, and that lead to optimally adapted partitions.

In this section, we discuss anisotropic refinement algorithms which fullfill this objective. These algorithms have been introduced and studied in [20] for piecewise polynomial approximation on two-dimensional triangulations. In the particular case of piecewise affine elements, it was proved in [21] that they lead to optimal error estimates. The main idea is again to refine the element T that maximizes the local error $e_{m,T}(f)_p$, but to allow several scenarios of refinement for this element. Here are two typical instances in two dimensions:

1. For rectangular partitions, we allow to split each rectangle into two rectangles of equal size by either a vertical or horizontal cut. There are therefore two splitting scenarios.

2. For triangular partitions, we allow to bisect each triangle from one of its vertex towards the mid-point of the opposite edge. There are therefore three splitting scenarios.

We display on Figure 5 two examples of anisotropic partitions respectively obtained by such splitting techniques. The choice between the different splitting scenarios is

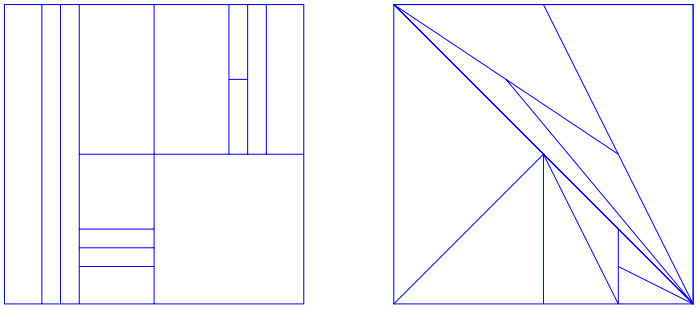


Fig. 5 Anisotropic partitions obtained by rectangle split (left) and triangle bisection (right)

done by a *decision rule* which depends on the function f. A typical decision rule is to select the split which best decreases the local error. The greedy refinement algorithm therefore reads as follows:

1. Initialization: $\mathscr{T}_{N_0} = \mathscr{D}_0$ with $N_0 := \#(\mathscr{D}_0)$.
2. Given $\mathscr{T}_N$ select $T \in \mathscr{T}_N$ that maximizes $e_{m,T}(f)_T$.
3. Use the decision rule in order to select the type of split to be performed on T.
4. Split T into K elements to obtain $\mathscr{T}_{N+K-1}$ and return to step 2.

Intuitively, the error equidistribution is ensured by selecting the element that maximizes the local error, while the role of the decision rule is to optimize the shape of the generated elements.

The problem is now to understand if the piecewise polynomial approximations generated by such refinement algorithms satisfy similar convergence properties as those which were established in §4 and §5 when using optimally adapted partitions. We first study the anisotropic refinement algorithm for the simple case of piecewise constant on rectangles, and we give a complete proof of its optimal convergence properties. We then present the anisotropic refinement algorithm for piecewise polynomials on triangulations, and give without proof the available results on its optimal convergence properties.

Remark 6.1. Let us remark that in contrast to the refinement algorithm discussed in §2.3 and 3.3, the partition $\mathscr{T}_N$ may not anymore be identified to a finite subtree within a fixed infinite master tree $\mathscr{M}$. Instead, for each f, the decision rule defines an infinite master tree $\mathscr{M}(f)$ that *depends* on f. The refinement algorithm corresponds

to selecting a finite subtree within $\mathscr{M}(f)$. Due to the finite number of splitting possibilities for each element, this finite subtree may again be encoded by a number of bits proportional to N. Similar to the isotropic refinement algorithm, one may use more sophisticated techniques such as CART in order to select an optimal partition of N elements within $\mathscr{M}(f)$. On the other hand the selection of the optimal partition within *all* possible splitting scenarios is generally of high combinatorial complexity.

Remark 6.2. A closely related algorithm was introduced in [26] and studied in [24]. In this algorithm every element is a convex polygon which may be split into two convex polygons by an arbitrary line cut, allowing therefore an infinite number of splitting scenarios. The selected split is again typically the one that decreases most the local error. Although this approach gives access to more possibilities of anisotropic partitions, the analysis of its convergence rate is still an open problem.

6.1 The refinement algorithm for piecewise constants on rectangles

As in §4, we work on the square domain $\Omega = [0,1]^2$ and we consider piecewise constant approximation on anisotropic rectangles. At a given stage of the refinement algorithm, the rectangle $T = I \times J$ that maximizes $e_{1,T}(f)_p$ is split either vertically or horizontally, which respectively corresponds to split one interval among I and J into two intervals of equal size and leaving the other interval unchanged. As already mentionned in the case of the refinement algorithm discussed in §3.3, we may replace $e_{1,T}(f)_p$ by the more computable quantity $\|f - P_{1,T} f\|_p$ for selecting the rectangle T of largest local error. Note that the $L^2(T)$-projection onto constant functions is simply the average of f on T:

$$P_{1,T} f = \frac{1}{|T|} \int_T f.$$

If T is the rectangle that is selected for being split, we denote by (T_d, T_u) the down and up rectangles which are obtained by a horizontal split of T and by (T_l, T_r) the left and right rectangles which are obtained by a vertical split of T. The most natural decision rule for selecting the type of split to be performed on T is based on comparing the two quantities

$$e_{T,h}(f)_p := \left(e_{1,T_d}(f)_p^p + e_{1,T_u}(f)_p^p \right)^{1/p} \text{ and } e_{T,v}(f)_p := \left(e_{1,T_l}(f)_p^p + e_{1,T_r}(f)_p^p \right)^{1/p},$$

which represent the local approximation error after splitting T horizontally or vertically, with the standard modification when $p = \infty$. The decision rule based on the L^p error is therefore:

If $e_{T,h}(f)_p \le e_{T,v}(f)_p$, then T is split horizontally, otherwise T is split vertically.

As already explained, the role of the decision rule is to optimize the shape of the

generated elements. We have seen in §4.1 that in the case where f is an affine function

$$q(x,y) = q_0 + q_x x + q_y y,$$

the shape of a rectangle $T = I \times J$ which is optimally adapted to q is given by the relation (59). This relation cannot be exactly fullfilled by the rectangles generated by the refinement algorithm since they are by construction dyadic type, and in particular

$$\frac{|I|}{|J|} = 2^j,$$

for some $j \in \mathbb{Z}$. We can measure the adaptation of T with respect to q by the quantity

$$a_q(T) := \left| \log_2 \left(\frac{|I|\,|q_x|}{|J|\,|q_y|} \right) \right|, \tag{91}$$

which is equal to 0 for optimally adapted rectangles and is small for "well adapted" rectangles. Inspection of the arguments leading the heuristic error estimate (65) in §4.1 or to the more rigourous estimate (68) in Theorem 4.1 reveals that these estimates also hold up to a fixed multiplicative constant if we use rectangles which have well adapted shape in the sense that $a_{q_T}(T)$ is uniformly bounded where q_T is the approximate value of f on T.

We notice that for all q such that $q_x q_y \neq 0$, there exists at least a dyadic rectangle T such that $a_T(q) \leq \frac{1}{2}$. We may therefore hope that the refinement algorithm leads to optimal error estimate of a similar form as (68), provided that the decision rule tends to generate well adapted rectangles. The following result shows that this is indeed the case when f is exactly an affine function, and when using the decision rule either based on the L^2 or L^∞ error.

Proposition 6.1. *Let $q \in \mathbb{P}_1$ be an affine function and let T be a rectangle. If T is split according to the decision rule either based on the L^2 or L^∞ error for this function and if T' a child of T obtained from this splitting, one then has*

$$a_q(T') \leq |a_q(T) - 1|. \tag{92}$$

As a consequence, all rectangles obtained after sufficiently many refinements satisfy $a_q(T) \leq 1$.

Proof: We first observe that if $T = I \times J$, the local L^∞ error is given by

$$e_{1,T}(q)_\infty := \frac{1}{2} \max\{|q_x|\,|I|, |q_y|\,|J|\},$$

and the local L^2 error is given by

$$e_{1,T}(q)_2 := \frac{1}{4\sqrt{3}} (q_x^2 |I|^2 + q_y^2 |J|^2)^{1/2}.$$

Assume that T is such that $|I|\,|q_x| \geq |J|\,|q_y|$. In such a case, we find that

$$e_{T,v}(q)_\infty = \frac{1}{2}\max\{|q_x|\,|I|, |q_y|\,|J|/2\} = |q_x|\,|I|/2,$$

and

$$e_{T,h}(q)_\infty = \frac{1}{2}\max\{|q_x|\,|I|/2, |q_y|\,|J|\} \le |q_x|\,|I|/2.$$

Therefore $e_{T,h}(q)_\infty \le e_{T,v}(q)_\infty$ which shows that the horizontal cut is selected by the decision rule based on the L^∞ error. We also find that

$$e_{T,v}(q)_2 := \frac{1}{\sqrt{6}}(q_x^2|I|^2 + q_y^2|J|^2/4)^{1/2},$$

and

$$e_{T,h}(q)_2 := \frac{1}{\sqrt{6}}(q_x^2|I|^2/4 + q_y^2|J|^2)^{1/2},$$

and therefore $e_{T,h}(q)_2 \le e_{T,v}(q)_2$ which shows that the horizontal cut is selected by the decision rule based on the L^2 error. Using the fact that

$$\log_2\Big(\frac{|I|\,|q_x|}{|J|\,|q_y|}\Big) \ge 0,$$

we find that if T' is any of the two rectangle generated by both decision rules, we have $a_q(T') = a_q(T) - 1$ if $a_q(T) \ge 1$ and $a_q(T') = 1 - a_q(T)$ if $a_q(T) \le 1$. In the case where $|I|\,|q_x| < |J|\,|q_y|$, we reach a similar conclusion observing that the vertical cut is selected by both decision rules. This proves (92) □

Remark 6.3. We expect that the above result also holds for the decision rules based on the L^p error for $p \notin \{2, \infty\}$ which therefore also lead to well adapted rectangles when f is an affine. In this sense all decision rules are equivalent, and it is reasonable to use the simplest rules based on the L^2 or L^∞ error in the refinement algorithm that selects the rectangle which maximizes $e_{1,T}(f)_p$, even when p differs from 2 or ∞.

6.2 Convergence of the algorithm

From an intuitive point of view, we expect that when we apply the refinement algorithm to an arbitrary function $f \in C^1(\Omega)$, the rectangles tend to adopt a locally well adapted shape, provided that the algorithm reaches a stage where f is sufficiently close to an affine function on each rectangle. However this may not necessarily happen due to the fact that we are not ensured that the diameter of all the elements tend to 0 as $N \to \infty$. Note that this is not ensured either for greedy refinement algorithms based on isotropic elements. However, we have used in the proof of Theorem 3.4 the fact that for N large enough, a fixed portion - say $N/2$ - of the elements have arbitrarily small diameter, which is not anymore guaranteed in the anisotropic setting.

We can actually give a very simple example of a smooth function f for which the approximation produced by the anisotropic greedy refinement algorithm *fails to converge* towards f due to this problem. Let φ be a smooth function of one variable which is compactly supported on $]0,1[$ and positive. We then define f on $[0,1]^2$ by

$$f(x,y) := \varphi(4x) - \varphi(4x-1).$$

This function is supported in $[0,1/2]\times[0,1]$. Due to its particular structure, we find that if $T=[0,1]^2$, the best approximation in $L^p(T)$ is achieved by the constant $c=0$ and one has

$$e_{1,T}(f)_p = 2^{1/p}\|\varphi\|_{L^p}.$$

We also find that $c=0$ is the best approximation on the four subrectangles T_d, T_u, T_l and T_r and that $e_{T,h}(f)_p = e_{T_v}(f)_p = e_{1,T}(f)_p$ which means both horizontal and vertical split do not reduce the error. According to the decision rule, the horizontal split is selected. We are then facing a similar situation on T_d and T_u which are again both split horizontally. Therefore, after $N-1$ greedy refinement steps, the partition $\mathscr{T}_N$ consists of rectangles all of the form $[0,1]\times J$ where J are dyadic intervals, and the best approximation remains $c=0$ on each of these rectangles. This shows that the approximation produced by the algorithm fails to converge towards f, and the global error remains

$$e_{1,\mathscr{T}_N}(f)_p = 2^{1/p}\|\varphi\|_{L^p},$$

for all $N>0$.

The above example illustrates the fact that the anisotropic greedy refinement algorithm may be defeated by simple functions that exhibit an oscillatory behaviour. One way to correct this defect is to impose that the refinement of $T=I\times J$ reduces its largest side-length the case where the refinement suggested by the original decision rule does not sufficiently reduce the local error. This means that we modify as follow the decision rule:

Case 1: if $\min\{e_{T,h}(f)_p, e_{T,v}(f)_p\} \le \rho e_{1,T}(f)_p$, then T is split horizontally if $e_{T,h}(f)_p \le e_{T,v}(f)_p$ or vertically if $e_{T,h}(f)_p > e_{T,v}(f)_p$. We call this a *greedy split*.

Case 2: if $\min\{e_{T,h}(f)_p, e_{T,v}(f)_p\} > \rho e_{1,T}(f)_p$, then T is split horizontally if $|I| \le |J|$ or vertically if $|I| > |J|$. We call this a *safety split*.

Here ρ is a parameter chosen in $]0,1[$. It should not be chosen too small in order to avoid that all splits are of safety type which would then lead to isotropic partitions. Our next result shows that the approximation produced by the modified algorithm does converge towards f.

Theorem 6.1. *For any $f\in L^p(\Omega)$ or in $C(\Omega)$ in the case $p=\infty$, the partitions $\mathscr{T}_N$ produced by the modified greedy refinement algorithm with parameter $\rho\in]0,1[$ satisfy*

$$\lim_{N\to+\infty} e_{1,\mathscr{T}_N}(f)_p = 0. \tag{93}$$

Proof: Similar to the original refinement procedure, the modified one defines a infinite master tree $\mathscr{M} := \mathscr{M}(f)$ with root Ω which contains all elements that can be generated at some stage of the algorithm applied to f. This tree depends on f, and the partition $\mathscr{T}_N$ produced by the modified greedy refinement algorithm may be identified to a finite subtree within $\mathscr{M}(f)$. We denote by $\mathscr{D}_j := \mathscr{D}_j(f)$ the partition consisting of the rectangles of area 2^{-j} in $\mathscr{M}$, which are thus obtained by j refinements of Ω. This partition also depends on f.

We first prove that $e_{1,\mathscr{D}_j}(f)_p \to 0$ as $j \to \infty$. For this purpose we split $\mathscr{D}_j$ into two sets $\mathscr{D}_j^g$ and $\mathscr{D}_j^s$. The first set $\mathscr{D}_j^g$ consists of the element T for which more than half of the splits that led from Ω to T were of greedy type. Due to the fact that such splits reduce the local approximation error by a factor ρ and that this error is not increased by a safety split, it is easily cheched by an induction argument that

$$e_{1,\mathscr{D}_j^g}(f)_p = \Big(\sum_{T \in \mathscr{D}_j^g} e_{1,T}(f)_p^p \Big)^{1/p} \leq \rho^{j/2} e_{1,\Omega}(f)_p \leq \rho^{j/2} \|f\|_{L^p},$$

which goes to 0 as $j \to +\infty$. This result also holds when $p = \infty$. The second set $\mathscr{D}_j^s$ consists of the elements T for which at least half of the splits that led from Ω to T were safety split. Since two safety splits reduce at least by 2 the diameter of T, we thus have

$$\max_{T \in \mathscr{D}_j^s} h_T \leq 2^{1-j/4},$$

which goes to 0 as $j \to +\infty$. From classical properties of density of piecewise constant functions in L^p spaces and in the space of continuous functions, it follows that

$$e_{1,\mathscr{D}_j^s}(f)_p \to 0 \ \text{ as } \ j \to +\infty.$$

This proves that

$$e_{1,\mathscr{D}_j}(f)_p = \Big(e_{1,\mathscr{D}_j^g}(f)_p^p + e_{1,\mathscr{D}_j^s}(f)_p^p \Big)^{1/p} \to 0 \ \text{ as } \ j \to +\infty,$$

with the standard modification if $p = \infty$.

In order to prove that $e_{1,\mathscr{T}_N}(f)_p$ also converges to 0, we first observe that since $e_{1,\mathscr{D}_j}(f)_p \to 0$, it follows that for all $\varepsilon > 0$, there exists only a finite number of $T \in \mathscr{M}(f)$ such that $e_{1,T}(f)_p \geq \varepsilon$. In turn, we find that

$$\varepsilon(N) := \max_{T \in \mathscr{T}_N} e_{1,T}(f)_p \to 0 \ \text{ as } \ N \to +\infty.$$

For some $j > 0$, we split $\mathscr{T}_N$ into two sets $\mathscr{T}_N^{j+}$ and $\mathscr{T}_N^{j-}$ which consist of those $T \in \mathscr{T}_N$ which are in $\mathscr{D}_l$ for $l \geq j$ and $l < j$ respectively. We thus have

$$e_{1,\mathscr{T}_N}(f)_p = \Big(e_{1,\mathscr{T}_N^{j+}}(f)_p^p + e_{1,\mathscr{T}_N^{j+}}(f)_p^p \Big)^{1/p} \leq \Big(e_{1,\mathscr{D}_j}(f)_p^p + 2^j \varepsilon(N)^p \Big)^{1/p}.$$

Since $e_{1,\mathscr{D}_j}(f)_p \to 0$ as $j \to +\infty$ and $\varepsilon(N) \to 0$ as $N \to \infty$, and since j is arbitrary, this concludes the proof, with the standard modification if $p = \infty$. □

6.3 Optimal convergence

We now prove that using the specific value $\rho = \frac{1}{\sqrt{2}}$ the modified greedy refinement algorithm has optimal convergence properties similar to (68) in the case where we measure the error in the L^∞ norm. Similar results can be obtained when the error is measured in L^p with $p < \infty$, at the price of more technicalities.

Theorem 6.2. *There exists a constant $C > 0$ such that for any $f \in C^1(\Omega)$, the partition $\mathscr{T}_N$ produced by the modified greedy refinement algorithm with parameter $\rho = \frac{1}{\sqrt{2}}$ satisfy the asymptotic convergence estimate*

$$\limsup_{N\to+\infty} N^{1/2} e_{1,\mathscr{T}_N}(f)_\infty \leq C \left\| \sqrt{|\partial_x f \ \partial_y f|} \right\|_{L^2} \tag{94}$$

The proof of this theorem requires a preliminary result. Here and after, we use the ℓ^∞ norm on $\mathbb{R}^2$ for measuring the gradient: for $z = (x,y) \in \Omega$

$$|\nabla f(z)| := \max\{|\partial_x f(z)|, |\partial_y f(z)|\},$$

and

$$\|\nabla f\|_{L^\infty(T)} := \sup_{z\in T} |\nabla f(z)| = \max\{\|\partial_x f\|_{L^\infty(T)}, \|\partial_y f\|_{L^\infty(T)}\}.$$

We recall that the local L^∞-error on T is given by

$$e_{1,T}(f)_\infty = \frac{1}{2}\left(\max_{z\in T} f(z) - \min_{z\in T} f(z)\right).$$

We also recall from the proof of Theorem 6.1 that

$$\varepsilon(N) := \max_{T\in\mathscr{T}_N} e_{1,T}(f)_\infty \to 0 \ \text{ as } \ N \to +\infty.$$

Finally we sometimes use the notation $x(z)$ and $y(z)$ to denote the coordinates of a point $z \in \mathbb{R}^2$.

Lemma 6.1. *Let $T_0 = I_0 \times J_0 \in \mathscr{T}_M$ be a dyadic rectangle obtained at some stage M of the refinement algorithm, and let $T = I \times J \in \mathscr{T}_N$ be a dyadic rectangle obtained at some later stage $N > M$ and such that $T \subset T_0$. We then have*

$$|I| \geq \min\left\{|I_0|, \frac{\varepsilon(N)}{4\|\nabla f\|_{L^\infty(T_0)}}\right\} \ \text{and} \ |J| \geq \min\left\{|J_0|, \frac{\varepsilon(N)}{4\|\nabla f\|_{L^\infty(T_0)}}\right\}.$$

Proof: Since the coordinates x and y play symmetrical roles, it suffices to prove the first inequality. We reason by contradiction. If the inequality does not hold, there exists a rectangle $T' = I' \times J'$ in the chain that led from T_0 to T_1 which is such that

$$|I'| < \frac{\varepsilon(N)}{4\|\nabla f\|_{L^\infty(T_0)}},$$

and such that T' is split vertically by the algorithm. If this was a safety split, we would have that $|J'| \leq |I'|$ and therefore

$$e_{1,T'}(f)_\infty \leq (|I'| + |J'|)\|\nabla f\|_{L^\infty(T)} \leq 2|I'|\|\nabla f\|_{L^\infty(T)} < \varepsilon(N),$$

which is a contradiction, since all ancestors of T should satisfy $e_{1,T'}(f)_\infty \geq \varepsilon(N)$. Hence this split was necessarily a greedy split.

Let $z_m := \mathrm{Argmin}_{z\in T'} f(z)$ and $z_M := \mathrm{Argmax}_{z\in T'} f(z)$, and let T'' be the child of T' (after the vertical split) containing z_M. Then T'' also contains a point z'_m such that $|x(z'_m) - x(z_m)| \leq |I'|/2$ and $y(z'_m) = y(z_m)$. It follows that

$$\begin{aligned} e_{T',v}(f)_\infty &= e_{1,T''}(f)_\infty \\ &\geq \tfrac{f(z_M) - f(z'_m)}{2} \\ &\geq \tfrac{f(z_M) - f(z_m) + \|\partial_x f\|_{L^\infty(T')}|I'|/2}{2} \\ &\geq \tfrac{3}{4} e_{1,T'}(f)_\infty \\ &> \rho e_{1,T'}(f)_\infty. \end{aligned}$$

The error was therefore insufficiently reduced which contradicts a greedy split. □

Proof of Theorem 6.2: We consider a small but fixed $\delta > 0$, we define $h(\delta)$ as the maximal $h > 0$ such that

$$\forall z, z' \in \Omega,\ |z - z'| \leq 2h(\delta) \Rightarrow |\nabla f(z) - \nabla f(z')| \leq \delta.$$

For any rectangle $T = I \times J \subset \Omega$, we thus have

$$\begin{aligned} e_{1,T}(f)_\infty &\geq (\|\partial_x f\|_{L^\infty(T)} - \delta)\min\{h(\delta), |I|\}, \\ e_{1,T}(f)_\infty &\geq (\|\partial_y f\|_{L^\infty(T)} - \delta)\min\{h(\delta), |J|\}. \end{aligned} \tag{95}$$

Let $\delta > 0$ and $M = M(f, \delta)$ be the smallest value of N such that $\varepsilon(N) \leq 9\delta h(\delta)$. For all $N \geq M$, and therefore $\varepsilon(N) \leq 9\delta h(\delta)$, we consider the partition $\mathscr{T}_N$ which is a refinement of $\mathscr{T}_M$. For any rectangle $T_0 = I_0 \times J_0 \in \mathscr{T}_M$, we denote by $\mathscr{T}_N(T_0)$ the set of rectangles of $\mathscr{T}_N$ that are contained T_0. We thus have

$$\mathscr{T}_N := \cup_{T_0 \in \mathscr{T}_M} \mathscr{T}_N(T_0),$$

and $\mathscr{T}_N(T_0)$ is a partition of T_0. We shall next bound by below the side length of $T = I \times J$ contained in $\mathscr{T}_N(T_0)$, distinguishing different cases depending on the behaviour of f on T_0.

Case 1. If $T_0 \in \mathscr{T}_M$ is such that $\|\nabla f\|_{L^\infty(T_0)} \leq 10\delta$, then a direct application of Lemma 6.1 shows that for all $T = I \times J \in \mathscr{T}_N(T_0)$ we have

$$|I| \geq \min\left\{|I_0|, \frac{\varepsilon(N)}{40\delta}\right\} \text{ and } |J| \geq \min\left\{|J_0|, \frac{\varepsilon(N)}{40\delta}\right\} \tag{96}$$

Case 2. If $T_0 \in \mathscr{T}_M$ is such that $\|\partial_x f\|_{L^\infty(T_0)} \geq 10\delta$ and $\|\partial_y f\|_{L^\infty(T_0)} \geq 10\delta$, we then claim that for all $T = I \times J \in \mathscr{T}_N(T_0)$ we have

$$|I| \geq \min\{|I_0|, \frac{\varepsilon(N)}{20\|\partial_x f\|_{L^\infty(T_0)}}\} \text{ and } |J| \geq \min\{|J_0|, \frac{\varepsilon(N)}{20\|\partial_x f\|_{L^\infty(T_0)}}\}, \tag{97}$$

and that furthermore

$$|T_0|\ \|\partial_x f\|_{L^\infty(T_0)}\|\partial_y f\|_{L^\infty(T_0)} \leq \left(\frac{10}{9}\right)^2 \int_{R^*} |\partial_x f\ \partial_y f| dxdy. \tag{98}$$

This last statement easily follows by the following observation: combining (95) with the fact that $\|\partial_x f\|_{L^\infty(T_0)} \geq 10\delta$ and $\|\partial_y f\|_{L^\infty(T_0)} \geq 10\delta$ and that $e_{1,T}(f)_\infty \leq \varepsilon(N) \leq 9\delta h(\delta)$, we find that for all $z \in T_0$

$$|\partial_x f(z)| \geq \|\partial_x f\|_{L^\infty(T_0)} - \delta \geq \frac{9}{10}\|\partial_x f\|_{L^\infty(T_0)},$$

and

$$|\partial_y f(z)| \geq \|\partial_y f\|_{L^\infty(T_0)} - \delta \geq \frac{9}{10}\|\partial_y f\|_{L^\infty(T_0)},$$

Integrating over T_0 yields (98). Moreover for any rectangle $T \subset T_0$, we have

$$\frac{9}{10} \leq \frac{e_{1,T}(f)_\infty}{\|\partial_x f\|_{L^\infty(T_0)}|I| + \|\partial_y f\|_{L^\infty(T_0)}|J|} \leq 1. \tag{99}$$

Clearly the two inequalities in (97) are symmetrical, and it suffices to prove the first one. Similar to the proof of Lemma 6.1, we reason by contradiction, assuming that a rectangle $T' = I' \times J'$ with $|I'|\|\partial_x f\|_{L^\infty(T_0)} < \frac{\varepsilon(N)}{10}$ was split vertically by the algorithm in the chain leading from T_0 to T. A simple computation using inequality (99) shows that

$$\frac{e_{T',h}(f)_\infty}{e_{1,T'}(f)_\infty} \leq \frac{e_{T',h}(f)_\infty}{e_{T',v}(f)_\infty} \leq \frac{5}{9} \times \frac{1+2\sigma}{1+\sigma/2} \text{ with } \sigma := \frac{\|\partial_x f\|_{L^\infty(T_0)}|I'|}{\|\partial_y f\|_{L^\infty(T_0)}|J'|}.$$

In particular if $\sigma < 0.2$ the algorithm performs a horizontal greedy split on T', which contradicts our assumption. Hence $\sigma \geq 0.2$, but this also leads to a contradiction since

$$\varepsilon(N) \leq e_{1,T'}(f)_\infty \leq \|\partial_x f\|_{L^\infty(T_0)}|I'| + \|\partial_y f\|_{L^\infty(T_0)}|J'| \leq (1+\sigma^{-1})\frac{\varepsilon(N)}{10} < \varepsilon(N)$$

Case 3. If $T_0 \in \mathscr{T}_M$ be such that $\|\partial_x f\|_{L^\infty(T_0)} \leq 10\delta$ and $\|\partial_y f\|_{L^\infty(T_0)} \geq 10\delta$, we then claim that for all $T = I \times J \in \mathscr{T}_N(T_0)$ we have

$$|I| \geq \min\left\{|I_0|, \frac{\varepsilon(N)}{C\delta}\right\} \text{ and } |J| \geq \min\left\{|J_0|, \frac{\varepsilon(N)}{4\|\nabla f\|_{L^\infty}}\right\}, \text{ with } C = 200, \quad (100)$$

with symmetrical result if T_0 is such that $\|\partial_x f\|_{L^\infty(T_0)} \geq 10\delta$ and $\|\partial_y f\|_{L^\infty(T_0)} \leq 10\delta$. The second part of (100) is a direct consequence of Lemma 6.1, hence we focus on the first part. Applyting the second inequality of (95) to $T = T_0$, we obtain

$$9\delta h(\delta) \geq e_{1,T_0}(f)_\infty \geq (\|\partial_y f\|_{L^\infty}(T_0) - \delta)\min\{h(\delta), |J_0|\} \geq 9\delta \min\{h(\delta), |J_0|\},$$

from which we infer that $|J_0| \leq h(\delta)$. If $z_1, z_2 \in T_0$ and $x(z_1) = x(z_2)$ we therefore have $|\partial_y f(z_1)| \geq |\partial_y f(z_2)| - \delta$. It follows that for any rectangle $T = I \times J \subset T_0$ we have

$$(\|\partial_y f\|_{L^\infty(T)} - \delta)|J| \leq e_{1,T}(f)_\infty \leq \|\partial_y f\|_{L^\infty(T)}|J| + 10\delta|I|. \quad (101)$$

We then again reason by contradiction, assuming that a rectangle $T' = I' \times J'$ with $|I'| \leq \frac{2\varepsilon(N)}{C\delta}$ was split vertically by the algorithm in the chain leading from T_0 to T. If $\|\partial_y f\|_{L^\infty(T')} \leq 10\delta$, then $\|\nabla f\|_{L^\infty(T')} \leq 10\delta$ and Lemma 6.1 shows that T' should not have been split vertically, which is a contradiction. Otherwise $\|\partial_y f\|_{L^\infty(T')} - \delta \geq \frac{9}{10}\|\partial_y f\|_{L^\infty(T')}$, and we obtain

$$(1 - 20/C)e_{1,T'}(f)_\infty \leq \|\partial_y f\|_{L^\infty(T')}|J'| \leq \frac{10}{9}e_{1,T'}(f)_\infty. \quad (102)$$

We now consider the children T'_v and T'_h of T' of maximal error after a horizontal and vertical split respectively, and we inject (102) in (101). It follows that

$$\begin{aligned} e_{T',h}(f)_\infty &= e_{1,T'_h}(f)_\infty \\ &\leq \|\partial_y f\|_{L^\infty(T')}|J'|/2 + 10\delta|I'| \\ &\leq \tfrac{5}{9}e_{1,T'}(f)_\infty + 20\varepsilon(N)/C \\ &\leq (\tfrac{5}{9} + 20/C)e_{1,T'}(f)_\infty = \tfrac{59}{90}e_{1,T'}(f)_\infty, \end{aligned}$$

and

$$\begin{aligned} e_{T',v}(f)_\infty &= e_{1,T'_h}(f)_\infty \\ &\geq (\|\partial_y f\|_{L^\infty(T')} - \delta)|J| \\ &\geq \tfrac{9}{10}\|\partial_y f\|_{L^\infty(T')}|J'| \\ &\geq \tfrac{9}{10}(1 - 20/C)e_{1,T'}(f)_\infty = \tfrac{81}{100}e_{1,T'}(f)_\infty. \end{aligned}$$

Therefore $e_{T',v}(f)_\infty > e_{T',h}(f)_\infty$ which is a contradiction, since our decision rule would then select a horizontal split.

We now choose N large enough so that the minimum in (96), (97) and (100) is are always equal to the second term. For all $T \in \mathscr{T}_N(T_0)$, we respectively find that

$$\frac{\varepsilon(N)^2}{|T|} \leq C \begin{cases} \delta^2 & \text{if } \|\nabla f\|_{L^\infty(T_0)} \leq 10\delta \\ \frac{1}{|T_0|}\int_{T_0} |\partial_x f\ \partial_y f| & \text{if } \|\partial_x f\|_{L^\infty(T_0)} \geq 10\delta \text{ and } \|\partial_y f\|_{L^\infty(T_0)} \geq 10\delta \\ \delta \|\nabla f\|_{L^\infty} & \text{if } \|\partial_x f\|_{L^\infty(T_0)} \leq 10\delta \text{ and } \|\partial_y f\|_{L^\infty(T_0)} \geq 10\delta \\ & \text{(or reversed).} \end{cases}$$

with $C = \max\{40^2,\ 20^2(10/9)^2,\ 800\} = 1600$. For $z \in \Omega$, we set $\psi(z) := \frac{1}{|T|}$ where $T \in \mathscr{T}_N$ such $z \in T$, and obtain

$$N = \#(\mathscr{T}_N) = \int_\Omega \psi \leq C\varepsilon(N)^{-2} \left(\int_\Omega |\partial_x f\ \partial_y f| dxdy + \delta \|\nabla f\|_{L^\infty} + \delta^2 \right).$$

Taking the limit as $\delta \to 0$, we obtain

$$\limsup_{N\to\infty} N^{\frac{1}{2}} \|f - f_N\| \leq 40 \left\| \sqrt{|\partial_x f \partial_y f|} \right\|_{L^2},$$

which concludes the proof. □

Remark 6.4. The proof of the Theorem can be adapted to any choice of parameter $\rho \in]\frac{1}{2}, 1[$.

6.4 Refinement algorithms for piecewise polynomials on triangles

As in §5, we work on a polygonal domain $\Omega \subset \mathbb{R}^2$ and we consider piecewise polynomial approximation on anisotropic triangles. At a given stage of the refinement algorithm, the triangle T that maximizes $e_{m,T}(f)_p$ is split from one of its vertices $a_i \subset \{a_1, a_2, a_3\}$ towards the mid-point b_i of the opposite edge e_i. Here again, we may replace $e_{m,T}(f)_p$ by the more computable quantity $\|f - P_{m,T} f\|_p$ for selecting the triangle T of largest local error.

If T is the triangle that is selected for being split, we denote by (T_i', T_i'') the two children which are obtained when T is split from a_i towards b_i. The most natural decision rule is based on comparing the three quantities

$$e_{T,i}(f)_p := \left(e_{m,T_i'}(f)_p^p + e_{m,T_i''}(f)_p^p \right)^{1/p}, \quad i = 1,2,3.$$

which represent the local approximation error on T after the three splitting options, with the standard modification when $p = \infty$. The decision rule based on the L^p error is therefore:

T is split from a_i towards b_i for an i that minimizes $e_{T,i}(f)_p$.

A convergence analysis of this anisotropic greedy algorithm is proposed in [21] in the case of piecewise affine functions corresponding to $m = 2$. Since it is by far more involved than the convergence analysis presented in §6.1, §6.2 and §6.3 for piecewise constants on rectangles, but possess several similar features, we discuss without proofs the main available results and we also illustrate their significance through numerical tests.

No convergence analysis is so far available for the case of higher order piecewise polynomial $m > 2$, beside a general convergence theorem similar to Theorem 6.1. The algorithm can be generalized to simplices in dimension $d > 2$. For instance, a 3-d simplex can be split into two simplices by a plane connecting one of its edges to the midpoint of the opposite edge, allowing therefore between 6 possibilities.

As remarked in the end of §6.1, we may use a decision rule based on a local error measured in another norm than the L^p norm for which we select the element T of largest local error. In [21], we considered the "L^2-projection" decision rule based on minimizing the quantity

$$e_{T,i}(f)_2 := \left(\|f - P_{2,T_i'}(f)\|^2_{L^2(T_i')} + \|f - P_{2,T_i''}(f)\|^2_{L^2(T_i'')} \right)^{1/2},$$

as well as the "L^∞-interpolation" decision rule based on minimizing the quantity

$$d_{T,i}(f)_2 := \|f - I_{2,T_i'}(f)\|_{L^\infty(T_i')} + \|f - I_{2,T_i''}(f)\|_{L^\infty(T_i'')},$$

where $I_{2,T}$ denotes the local interpolation operator: $I_{2,T}(f)$ is the affine function that is equal to f at the vertices of T. Using either of these two decision rules, it is possible to prove that the generated triangles tend to adopt a well adapted shape.

In a similar way to the algorithm for piecewise constant approximation on rectangles, we first discuss the behaviour of the algorithm when f is exactly a quadratic function q. Denoting by $\mathbf{q}$ its the homogeneous part of degree 2, we have seen in §5.1 that when $\det(\mathbf{q}) \neq 0$, the approximation error on an optimally adapted triangle T is given by

$$e_{2,T}(q)_p = e_{2,T}(q)_p = |T|^{1/\tau} K_{2,p}(\mathbf{q}), \quad \frac{1}{\tau} := \frac{1}{p} + 1.$$

We can measure the adaptation of T with respect to $\mathbf{q}$ by the quantity

$$\sigma_{\mathbf{q}}(T)_p = \frac{e_{2,T}(\mathbf{q})_p}{|T|^{1/\tau} K_{2,p}(\mathbf{q})},$$

which is equal to 1 for optimally adapted triangles and small for "well adapted" triangles. It is easy to check that the functions $(\mathbf{q}, T) \mapsto \sigma_T(\mathbf{q})_p$ are equivalent for all p, similar to the shape functions $K_{2,p}$ as observed in §5.2.

The following theorem, which is a direct consequence of the results in [21], shows that the decision rule tends to make "most triangles" well adapted to $\mathbf{q}$.

Theorem 6.3. *There exists constants $0 < \theta, \mu < 1$ and a constant C_p that only depends on p such that the following holds. For any $\mathbf{q} \in \mathbb{H}_2$ such that $\det(\mathbf{q}) \neq 0$ and any triangle T, after j refinement levels of T according to the decision rule, a proportion $1 - \theta^j$ of the 2^j generated triangles T' satisfies*

$$\sigma_{\mathbf{q}}(T')_p \leq \min\{\mu^j \sigma_{\mathbf{q}}(T)_p, C_p\}. \tag{103}$$

As a consequence, for $j > j(\mathbf{q}, T) = -\frac{\log C_p - \log(\sigma_{\mathbf{q}}(T)_p)}{\log \mu}$ one has

$$\sigma_{\mathbf{q}}(T')_p \leq C_p, \tag{104}$$

for a proportion $1 - \theta^j$ of the 2^j generated triangles T'.

This result should be compared to Proposition 6.1 in the case of rectangles. Here it is not possible to show that *all* triangles become well adapted to q, but a proportion that tends to 1 does. It is quite remarkable that with only three splitting options, the greedy algorithm manages to drive most of the triangles to a near optimal shape. We illustrate this fact on Figure 6, in the case of the quadratic form $\mathbf{q}(x,y) := x^2 + 100y^2$, and an initial triangle T which is equilateral for the euclidean metric and therefore not well adapted to $\mathbf{q}$. Triangles such that $\sigma_{\mathbf{q}}(T')_2 \leq C_2$ are displayed in white, others in grey. We observe the growth of the proportion of well adapted triangles as the refinement level increases.

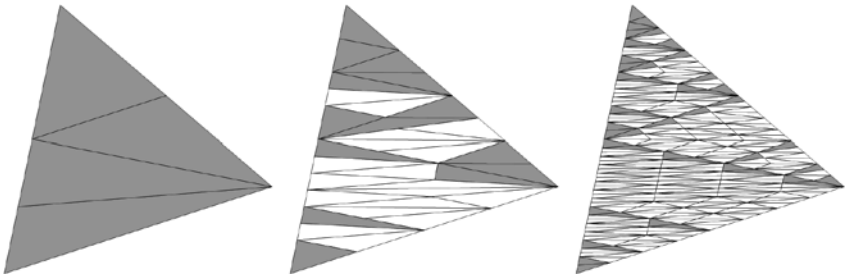


Fig. 6 Greedy refinement for $\mathbf{q}(x,y) := x^2 + 100y^2$: $j = 2$ (left), $j = 5$ (center), $j = 8$ (right)

From an intuitive point of view, we expect that when we apply the anisotropic greedy refinement algorithm to an arbitrary function $f \in C^2(\Omega)$, the triangles tend to adopt a locally well adapted shape, provided that the algorithm reaches a stage where f is sufficiently close to an quadratic function on each triangle. As in the case of the greedy refinement algorithm for rectangles, this may not always be the case. It is however possible to prove that this property holds in the case of strictly convex or concave functions, using the "L^∞-interpolation" decision rule. This allows to prove in such a case that the approximation produced by the anisotropic greedy algorithm satisfies an optimal convergence estimate in accordance with Theorem 5.1. These results from [21] can be summarized as follows.

Theorem 6.4. *If f is a C^2 function such that $d^2 f(x) \geq \alpha I$ or $d^2 f(x) \leq -\alpha I$, for all $x \in \Omega$ and some $\alpha > 0$, then the triangulation generated by the anisotropic greedy refinement algorithm (with the L^∞-interpolation decision rule) satisfies*

$$\lim_{N \to +\infty} \max_{T \in \mathcal{T}_N} h_T = 0. \tag{105}$$

Moreover, there exists a constant $C > 0$ such that for any such f, the approximation produced by the anisotropic greedy refinement algorithm satisfies the asymptotic convergence estimate

$$\limsup_{N \to +\infty} N e_{2,\mathcal{T}_N}(f)_p \leq C \left\| \sqrt{|\det(d^2 f)|} \right\|_{L^\tau}, \quad \frac{1}{\tau} := \frac{1}{p} + 1. \tag{106}$$

For a non-convex function, we are not ensured that the diameter of the elements tends to 0 as $N \to \infty$, and similar to the greedy algorithm for rectangles, it is possible to produce examples of smooth functions f for which the approximation produced by the anisotropic greedy refinement algorithm fails to converge towards f. A natural way to modify the algorithm in order to circumvent this problem is to impose a type of splitting that tend to diminish the diameter, such as longest edge or newest vertex bisection, in the case where the refinement suggested by the original decision rule does not sufficiently reduce the local error. This means that we modify as follow the decision rule:

Case 1: if $\min\{e_{T,1}(f)_p, e_{T,2}(f)_p, e_{T,3}(f)_p\} \leq \rho e_{2,T}(f)_p$, then split T from a_i towards b_i for an i that minimizes $e_{T,i}(f)_p$. We call this a *greedy split*.

Case 2: if $\min\{e_{T,1}(f)_p, e_{T,2}(f)_p, e_{T,3}(f)_p\} > \rho e_{2,T}(f)_p$, then split T from the most recently generated vertex or towards its longest edge in the euclidean metric. We call this a *safety split*.

As in modified greedy algorithm for rectangles, ρ is a parameter chosen in $]0,1[$ that should not be chosen too small in order to avoid that all splits are of safety type which would then lead to isotropic triangulations. It was proved in [20] that the approximation produced by this modified algorithm does converge towards f for any $f \in L^p(\Omega)$. The following result also holds for the generalization of this algorithm to higher degree piecewise polynomials.

Theorem 6.5. *For any $f \in L^p(\Omega)$ or in $C(\Omega)$ in the case $p = \infty$, the approximations produced by the modified anisotropic greedy refinement algorithm with parameter $\rho \in]0,1[$ satisfies*

$$\limsup_{N \to +\infty} e_{2,\mathcal{T}_N}(f)_p = 0. \tag{107}$$

Similar to Theorem 6.2, we may expect that the modified anisotropic greedy refinement algorithm satisfies optimal convergence estimates for all C^2 function, but this is an open question at the present stage.

Conjecture. *There exists a constant $C > 0$ and $\rho^* \in]0,1[$ such that for any $f \in C^2$, the approximation produced by the modified anisotropic greedy refinement algorithm with parameter $\rho \in]\rho^*, 1[$ satisfies the asymptotic convergence estimate* (106).

We illustrate the performance of the anisotropic greedy refinement algorithm algorithm for a function f which has a sharp transition along a curved edge. Specifically we consider

$$f(x,y) = f_\delta(x,y) := g_\delta(\sqrt{x^2+y^2}),$$

where g_δ is defined by $g_\delta(r) = \frac{5-r^2}{4}$ for $0 \leq r \leq 1$, $g_\delta(1+\delta+r) = -\frac{5-(1-r)^2}{4}$ for $r \geq 0$, g_δ is a polynomial of degree 5 on $[1,1+\delta]$ which is determined by imposing that g_δ is globally C^2. The parameter δ therefore measures the sharpness of the transition. We apply the anisotropic refinement algorithm based on splitting the triangle that maximizes the local L^2-error and we therefore measure the global error in L^2.

Figure 7 displays the triangulation $\mathscr{T}_{10000}$ obtained after 10000 steps of the algorithm for $\delta = 0.2$. In particular, triangles T such that $\sigma_{\mathbf{q}}(T)_2 \leq C_2$ - where $\mathbf{q}$ is the quadratic form associated with $d^2 f$ measured at the barycenter of T - are displayed in white, others in grey. As expected, most triangles are of the first type therefore well adapted to f. We also display on this figure the adaptive isotropic triangulation produced by the greedy tree algorithm based on newest vertex bisection for the same number of triangles.

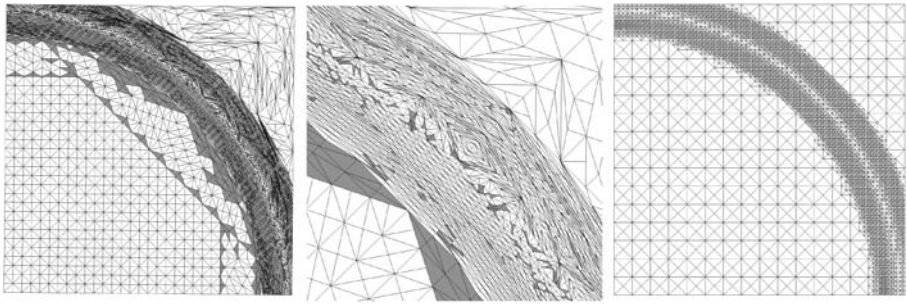


Fig. 7 The anisotropic triangulation $\mathscr{T}_{10000}$ (left), detail (center), isotropic triangulation (right)

Since f is a C^2 function, approximations by uniform, adaptive isotropic and adaptive anisotropic triangulations all yield the convergence rate $\mathscr{O}(N^{-1})$. However the constant

$$C := \limsup_{N \to +\infty} N e_{2,\mathscr{T}_N}(f)_2,$$

strongly differs depending on the algorithm and on the sharpness of the transition. We denote by C_U, C_I and C_A the empirical constants (estimated by $N\|f - f_N\|_2$ for $N = 8192$) in the uniform, adaptive isotropic and adaptive anisotropic case respectively, and by $U(f) := \|d^2 f\|_{L^2}$, $I(f) := \|d^2 f\|_{L^{2/3}}$ and $A(f) := \|\sqrt{|\det(d^2 f)|}\|_{L^{2/3}}$

the theoretical constants suggested by the convergence estimates. We observe on Figure 8. that C_U and C_I grow in a similar way as $U(f)$ and $I(f)$ as $\delta \to 0$ (a detailed computation shows that $U(f) \approx 10.37\delta^{-3/2}$ and $I(f) \approx 14.01\delta^{-1/2}$). In contrast C_A and $A(f)$ remain uniformly bounded, a fact which is in accordance with Theorem 5.3 and reflects the superiority of anisotropic triangulations as the layer becomes thinner and f_δ tends to a cartoon function.

δ	$U(f)$	$I(f)$	$A(f)$	C_U	C_I	C_A
0.2	103	27	6.75	7.87	1.78	0.74
0.1	602	60	8.50	23.7	2.98	0.92
0.05	1705	82	8.48	65.5	4.13	0.92
0.02	3670	105	8.47	200	6.60	0.92

Fig. 8 Comparison between theoretical and empirical convergence constants for uniform, adaptive isotropic and anisotropic refinements, and for different values of δ

We finally apply the anisotropic refinement algorithm to the numerical image of Figure 4 based on the discretized L^2 error and using $N = 2000$ triangles. We observe on Figure 9 that the ringing artefacts produced by the isotropic greedy refinement algorithm near the edges are strongly reduced. This is due to the fact that the anisotropic greedy refinement algorithm generates long and thin triangles aligned with the edges. We also observe that the quality is slightly improved when using the modified algorithm. Let us mention that a different approach to the approximation of image by adaptive anisotropic triangulations was proposed in [27]. This approach is based on a *thinning* algorithm, which starts from a fine triangulation and iteratively coarsens it by point removal. The use of adaptive adaptive anisotropic partitions has also strong similarities with thresholding methods based on representations which have more directional selectivity than wavelet decompositions [4, 13, 31, 37]. It is not known so far if these methods satisfy asymptotic error estimates of the same form as (106).

Fig. 9 Approximation by 2000 anisotropic triangles obtained by the greedy (left) and modified (right) algorithm

References

1. F. Alauzet and P.J. Frey, *Anisotropic mesh adaptation for CFD computations*, Comput. Methods Appl. Mech. Engrg. 194, 5068–5082, 2005.
2. B. Alpert, *A class of bases in L^2 for the sparse representation of integral operators*, SIAM J. Math. Anal. 24, 246–262, 1993.
3. T. Apel, *Anisotropic finite elements: Local estimates and applications*, Advances in Numerical Mathematics, Teubner, Stuttgart, 1999.
4. F. Arandiga, A. Cohen, R. Donat, N. Dyn and B. Matei, *Approximation of piecewise smooth images by edge-adapted techniques*, ACHA 24, 225–250, 2008.
5. V. Babenko, Y. Babenko, A. Ligun and A. Shumeiko, *On Asymptotical Behavior of the Optimal Linear Spline Interpolation Error of C^2 Functions*, East J. Approx. 12(1), 71–71, 101.
6. Yuliya Babenko, *Asymptotically Optimal Triangulations and Exact Asymptotics for the Optimal L^2-Error for Linear Spline Interpolation of C^2 Functions*, submitted.
7. P. Binev and R. DeVore, *Fast Computation in Adaptive Tree Approximation*, Numerische Mathematik 97, 193–217, 2004.
8. P. Binev, W. Dahmen and R. DeVore, *Adaptive Finite Element Methods with Convergence Rates*, Numerische Mathematik 97, 219–268, 2004.
9. J.-D. Boissonnat, C. Wormser and M. Yvinec. *Locally uniform anisotropic meshing.* To appear at the next Symposium on Computational Geometry, june 2008 (SOCG 2008)
10. H. Borouchaki, P.J. Frey, P.L. George, P. Laug and E. Saltel, *Mesh generation and mesh adaptivity: theory, techniques*, in Encyclopedia of computational mechanics, E. Stein, R. de Borst and T.J.R. Hughes ed., John Wiley & Sons Ltd., 2004.
11. Sebastien Bougleux and Gabriel Peyré and Laurent D. Cohen. *Anisotropic Geodesics for Perceptual Grouping and Domain Meshing.* Proc. tenth European Conference on Computer Vision (ECCV'08), Marseille, France, October 12-18, 2008.
12. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees*, Wadsworth international, Belmont, CA, 1984.
13. E. Candes and D.L. Donoho, *Curvelets and curvilinear integrals*, J. Approx. Theory. 113, 59–90, 2000.
14. W. Cao, *An interpolation error estimate on anisotropic meshes in $\mathbb{R}^n$ and optimal metrics for mesh refinement*. SIAM J. Numer. Anal. 45 no. 6, 2368–2391, 2007.
15. W. Cao, *Anisotropic measure of third order derivatives and the quadratic interpolation error on triangular elements*, to appear in SIAM J. Sci. Comput., 2007.
16. W. Cao, *An interpolation error estimate in $\mathbb{R}^2$ based on the anisotropic measures of higher order derivatives*. Math. Comp. 77, 265–286, 2008.
17. L. Chen, P. Sun and J. Xu, *Optimal anisotropic meshes for minimizing interpolation error in L^p-norm*, Math. of Comp. 76, 179–204, 2007.
18. A. Cohen, *Numerical analysis of wavelet methods*, Elsevier, 2003.
19. A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree-structured approximation and optimal encoding*, App. Comp. Harm. Anal. 11, 192–226, 2001.
20. A. Cohen, N. Dyn, F. Hecht and J.-M. Mirebeau, *Adaptive multiresolution analysis based on anisotropic triangulations*, preprint, Laboratoire J.-L. Lions, submitted 2008.
21. A. Cohen, J.-M. Mirebeau, *Greedy bisection generates optimally adapted triangulations*, preprint, Laboratoire J.-L. Lions, submitted 2008.
22. A. Cohen, J.-M. Mirebeau, *Anisotropic smoothness classes: from finite element approximation to image processing*, preprint, Laboratoire J.-L. Lions, submitted 2009.
23. W. Dahmen, *Adaptive approximation by multivariate smooth splines*, J. Approx. Theory 36, 119–140, 1982.
24. S. Dahmen, S. Dekel and P. Petrushev, *Two-level splits of anisotropic Besov spaces*, to appear in Constructive Approximation, 2009.
25. S. Dekel, D. Leviatan and M. Sharir, *On Bivariate Smoothness Spaces Associated with Nonlinear Approximation*, Constructive Approximation 20, 625–646, 2004

26. S. Dekel and D. Leviathan, *Adaptive multivariate approximation using binary space partitions and geometric wavelets*, SIAM Journal on Numerical Analysis 43, 707–732, 2005.
27. L. Demaret, N. Dyn, M. Floater and A. Iske, *Adaptive thinning for terrain modelling and image compression*, in Advances in Multiresolution for Geometric Modelling, N.A. Dodgson, M.S. Floater, and M.A. Sabin (eds.), Springer-Verlag, Heidelberg, 321–340, 2005.
28. R. DeVore, *Nonlinear approximation*, Acta Numerica 51-150, 1998
29. R. DeVore and G. Lorentz, *Constructive Approximation*, Springer, 1993.
30. R. DeVore and X.M. Yu, *Degree of adaptive approximation*, Math. of Comp. 55, 625–635.
31. D. Donoho, *Wedgelets: nearly minimax estimation of edges*, Ann. Statist. 27(3), 859–859, 897.
32. D. Donoho, *CART and best basis: a connexion*, Ann. Statist. 25(5), 1870–1870, 1911.
33. W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. 33, 1106–1124, 1996.
34. P.J. Frey and P.L. George, *Mesh generation. Application to finite elements*, Second edition. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2008.
35. J.-P. Kahane, *Teoria constructiva de functiones*, Course notes, University of Buenos Aires, 1961.
36. B. Karaivanov and P. Petrushev, *Nonlinear piecewise polynomial approximation beyond Besov spaces*, Appl. Comput. Harmon. Anal. 15(3), 177–177, 223.
37. E. Le Pennec and S. Mallat, *Bandelet image approximation and compression*, SIAM Journal of Multiscale Modeling. and Simulation, 4(3), 992–992, 1039.
38. S. Mallat *A Wavelet Tour of Signal Processing - The sparse way*, 3rd Revised edition, Academic Press, 2008.
39. J.-M. Mirebeau, *Optimally adapted finite element meshes*, preprint, Laboratoire J.-L. Lions, submitted 2009.
40. P. Morin, R. Nochetto and K. Siebert, *Convergence of adaptive finite element methods*, SIAM Review 44, 631–658, 2002.
41. M.C. Rivara, *New longest-edge algorithms for the renement and/or improvement of unstructured triangulations*, Int. J. Num. Methods 40, 3313–3324, 1997.
42. E.M. Stein, *Singular integral and differentiability properties of functions*, Princeton University Press, 1970.
43. R. Stevenson, *An optimal adaptive finite element method*, SIAM J. Numer. Anal., 42(5), 2188–2188, 2217.
44. R. Verfurth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, 1996.
45. The 2-d anisotropic mesh generator BAMG: http://www.freefem.org/ff++/ (included in the FreeFem++ software)

Anisotropic function spaces with applications

Shai Dekel and Pencho Petrushev

Abstract In this survey we review the recently developed theory of anisotropic spaces and representations of functions based on anisotropic multilevel ellipsoid covers (dilations) of $\mathbb{R}^n$. We also exhibit the relations of the ellipsoid cover approach to earlier concepts of anisotropic structures as well as to the framework of general spaces of homogeneous type. A number of open problems are presented and discussed.

1 Introduction

Anisotropic phenomena appear in various contexts in mathematical analysis and its applications. For instance, functions are frequently very smooth on subdomains of $\mathbb{R}^n$ separated by smooth curves or manofolds, where they have jump discontinuities. This sort of singularities reduce significantly the classical smoothness of the functions and create problems when attempting to find sparse representations of them.

One perhaps useful approach to resolving the singularities of functions along smooth curves and manifolds (and more general singular behaviors) is the utilization of the framework of anisotropic multiscale ellipsoid covers (dilations) of $\mathbb{R}^n$ which may change rapidly from point to point at any level and in depth. The second important element of this concept is to use anisotropic ellipsoid covers adaptively by allowing them to adjust to the singularities of the function under question. Other critical issues are related, in particular, to the anisotropic representation of functions and definition and characterization of the respective anisotropic smoothness spaces.

Shai Dekel
GE Healthcare, 27 Hamaskit St., Herzelia 46733, Israel, e-mail: Shai.Dekel@ge.com

Pencho Petrushev
Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA,
e-mail: pencho@math.sc.edu

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_5,

The purpose of this survey paper is to review the main concepts and problems of this relatively new undertaking, documented so far in [11, 12, 14]. Although we will have some answers to reveal to some of the important questions, there will be plenty of open problems presented as well.

This theory has three main components with the first being the structure of the underlying **ellipsoid covers** of $\mathbb{R}^n$. The main role here is played by discrete multilevel ellipsoid covers of $\mathbb{R}^n$ of the form: $\Theta = \cup_{m\in\mathbb{Z}}\Theta_m$, where each Θ_m consists of ellipsoids of volume $\sim 2^{-a_0 j}$ which cover $\mathbb{R}^n$ and any ellipsoids $\theta_1, \theta_2 \in \Theta_m$ with $\theta_1 \cap \theta_2 \neq \emptyset$ have similar shapes and orientations. In depth the behavior of the ellipsoids is similar, namely, if $\theta_1 \in \Theta_m$, $\theta_2 \in \Theta_{m+1}$ and $\theta_1 \cap \theta_2 \neq \emptyset$, then θ_1 and θ_2 are similar in shape and orientation. An important feature of the set of all ellipsoid covers of $\mathbb{R}^n$ is that it is invariant under affine transforms. Another important issue is that any ellipsoid cover of $\mathbb{R}^n$ generates a quasi-distance, which coupled with the Lebesgue measure transforms $\mathbb{R}^n$ into a homogeneous type space. The properties of anisotropic covers are explored in [12]. A short description of them is given in §2, where we also compare ellipsoid covers of $\mathbb{R}^2$ with the so called multilavel strong local regular (SLR) triangulations of $\mathbb{R}^2$, introduced in [20].

The **anisotropic elements (building blocks)** introduced in [12] and the related representations of functions is the second component of our theory. A sequence of bases $\{\Phi_m\}_{m\in\mathbb{Z}}$ is naturally associated to each discrete ellipsoid cover $\Theta = \cup_{m\in\mathbb{Z}}\Theta_m$. Here each Φ_m consists of C^∞ functions which are supported on the ellipsoids in Θ_m, reproduce polynomials of degree $< k$ and are locally linear independent. The key property of these bases is that each Φ_m is a stable basis in L_p for $0 < p \leq \infty$. This allows to define local projectors into the spaces $S_m = \operatorname{span}(\Phi_m)$ which preserve polynomials of degree $< k$. In turn, these maps induce a sequence of two-level-split bases which provide representation of functions and are aligned with the underlying anisotropic structure in $\mathbb{R}^n$. As is shown in [12] these representations also allow to characterize the anisotropic Besov spaces of positive smoothness. The next step is to define smooth (global) duals to $\{\Phi_m\}$ and thereby to construct kernels $\{S_m\}$ which reproduce polynomials of a certain degree in both variables. This enabled us to deploy the machinery of homogeneous spaces to the construction of continuous and discrete anisotropic wavelet frames. All these constructions and results are presented in §3.

The third component of the theory we review here consists of **anisotropic spaces** associated with anisotropic ellipsoid covers of $\mathbb{R}^n$. The anisotropic homogeneous ($\dot{B}^\alpha_{pq}(\Theta)$) and inhomogeneous ($B^\alpha_{pq}(\Theta)$) Besov spaces (B-spaces) of positive smoothness are developed in [12] and briefly introduced in §4. In the same section we compare them with the anisotropic B-spaces induced by multilevel SLR-triangulations of $\mathbb{R}^2$ and with classical Besov spaces. In §5 we show that, in analogy to the classical case, certain B-spaces naturally occur in nonlinear N-term approximation from the two-level-split bases. In §6 we advance the idea of using adaptively anisotropic B-space for measuring the smoothness of the functions, which is closely related to the problem for sparse representation of functions. The development of anisotropic Triebel-Lizorkin of an arbitrary smoothness is the grand open problem in this theory. The key is to construct anisotropic frames with well local-

ized elements and prescribed vanishing moments which are faithfully aligned with the underlying anisotropic ellipsoid cover.

Candès and Donoho (e.g. [5, 6]) have developed the so called *curvlets*, which provide an alternative scheme for resolving singularities of functions along smooth curves in $\mathbb{R}^2$. The advantage of curvlets over our approach is that the curvlets form a frame, while our scheme is adaptive, and hence curvlets are easier to implement. On the other hand, the curvlet frame is overly redundant. More precisely at every location and scale there are numerous directional elements with various orientations (the number of orientations increases with the scale). Curvlets are purely L_2-creatures which rely on fine cancelations and are unusable for decomposition of functions in L_p, $p \neq 2$.

Yet another approach to resolving singularities of functions along smooth curves is developed in [1, 15] and is based on the so called *Adaptive Geometric Wavelets*. This method is closely related to the schemes employing ellipsoid covers and nested triangulations considered here; it proved to be very effective in image compression.

In the final Section 7 the two-level-split bases and the machinery of Besov spaces are applied in a regular set-up to the development of meshless multilevel Schwarz preconditioners for elliptic boundary value problems. The details of this development are given in [11], which was the starting point of this work.

Throughout we will use $|E|$ to denote the Lebesgue measure of $E \subset \mathbb{R}^n$; we will denote by c, c_1, c_2, etc. positive constants which may vary at every appearance. The equivalence $a \sim b$ means $c_1 a \leq b \leq c_2 a$.

2 Anisotropic multiscale structures on $\mathbb{R}^n$

In this article we are mainly concerned with anisotropic structures on $\mathbb{R}^n$ induced by anisotropic ellipsoid covers (dilations) of $\mathbb{R}^n$ and the related function spaces . For comparison we will also briefly review the anisotropic multilevel nested triangulations of $\mathbb{R}^2$.

2.1 Anisotropic multilevel ellipsoid covers (dilations) of $\mathbb{R}^n$

We denote by $B(x,r)$ the Euclidean ball in $\mathbb{R}^n$ of radius r centered at x. The image of the unit ball $B^* := B(0,1)$ in $\mathbb{R}^n$ via an affine transform will be called an *ellipsoid*.

Definition 2.1. We call

$$\Theta = \bigcup_{m\in\mathbb{Z}} \Theta_m$$

a *discrete multilevel ellipsoid cover* of $\mathbb{R}^n$ if the following conditions are obeyed, where $a_0, \ldots, a_8$, and N_1 are positive constants:

(a) Every level Θ_m $(m \in \mathbb{Z})$ consists of ellipsoids θ such that

$$a_1 2^{-a_0 m} \le |\theta| \le a_2 2^{-a_0 m} \tag{1}$$

and Θ_m is a cover of $\mathbb{R}^n$, i.e. $\mathbb{R}^n = \bigcup_{\theta\in\Theta_m} \theta$.

(b) For $\theta \in \Theta$ let A_θ be an affine transform of the form

$$A_\theta(x) = M_\theta x + v_\theta, \quad M_\theta \in \mathbb{R}^{n\times n},$$

such that $\theta = A_\theta(B^*)$ and $v_\theta := A(0)$ is the center of θ. We postulate that for any $\theta \in \Theta_m$ and $\theta' \in \Theta_{m+\nu}$ $(m \in \mathbb{Z}, \nu \ge 0)$ with $\theta \cap \theta' \neq \emptyset$, we have

$$a_3 2^{-a_4 \nu} \le 1/\|M_{\theta'}^{-1} M_\theta\|_{\ell_2\to\ell_2} \le \|M_\theta^{-1} M_{\theta'}\|_{\ell_2\to\ell_2} \le a_5 2^{-a_6 \nu}. \tag{2}$$

(c) Each ellipsoid $\theta \in \Theta_m$ can be intersected by at most N_1 ellipsoids from Θ_m.

(d) For every $x \in \mathbb{R}^n$ and $m \in \mathbb{Z}$ there exists $\theta \in \Theta_m$ such that $x \in \theta^\diamond$, where $\theta^\diamond$ is the dilated version of θ by a factor of $a_7 < 1$, i.e. $\theta^\diamond = A_\theta(B(0,a_7))$.

(e) If $\theta \cap \eta \neq \emptyset$ with $\theta \in \Theta_m$ and $\eta \in \Theta_m \cup \Theta_{m+1}$, then $|\theta \cap \eta| > a_8 |\eta|$.

We will denote by $\mathbf{p}(\Theta) := \{a_0, a_1, \dots, a_8, N_1\}$ the set of all parameters in the above definition.

Several clarifying remarks are in order.

1. It is crucial that the set of all discrete ellipsoid covers of $\mathbb{R}^n$ is invariant under affine transforms. More precisely, the images $A(\theta)$ of all ellipsoids $\theta \in \Theta$ of a given cover Θ of $\mathbb{R}^n$ via an affine transform A of the form $A(x) = Mx + v$ with $|\det M| = 1$ form an ellipsoid cover of $\mathbb{R}^n$ with the same parameters as Θ.
2. Condition (b) above indicates that if $\theta \cap \theta' \neq \emptyset$, then the ellipsoids θ and θ' are similar in shape and orientation when they are from close levels. In particular, if $M := M_\theta^{-1} M_{\theta'}$ and $M = UDV$ is the singular value decomposition of M, where U and V are orthogonal matrices, and $D = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ is diagonal with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$, then

$$\|M\|_{\ell_2\to\ell_2} = \sigma_1 \quad \text{and} \quad \|M_{\theta'}^{-1} M_\theta\|_{\ell_2\to\ell_2} = \|M^{-1}\|_{\ell_2\to\ell_2} = 1/\sigma_n.$$

Therefore, condition (b) is equivalently expressed as

$$a_3 2^{-a_4 \nu} \le \sigma_n \le \cdots \le \sigma_1 \le a_5 2^{-a_6 \nu}. \tag{3}$$

This condition has a clear geometric interpretation: The affine transform A_θ^{-1}, which maps the ellipsoid θ onto the unit ball B^*, maps the ellipsoid θ' onto an ellipsoid with semi-axes $\sigma_1, \sigma_2, \dots, \sigma_n$ satisfying (3).

3. Condition (e) may seem restrictive, but this is not the case. As is shown in [12] if Θ is a discrete ellipsoid cover satisfying conditions $(a)-(d)$ above, then there exists a discrete ellipsoid cover $\widetilde{\Theta}$ of $\mathbb{R}^n$ which obeys conditions $(a)-(e)$ (with possibly different constants a_1 and a_7) obtained by dilating every ellipsoid $\theta \in \Theta$ by a factor r_θ satisfying $(a_7+1)/2 \le r_\theta \le 1$.

Continuous and semi-continuous ellipsoid covers. Discrete ellipsoid covers of $\mathbb{R}^n$ are easy to derive from semi-continuous or continuous covers, which are in general easier to construct.

In the case of a *semi-continuous ellipsoid cover* $\Theta = \cup_{m\in\mathbb{Z}}\Theta_m$, an ellipsoid $\theta(v,m) \in \Theta_m$ is associated to every $v \in \mathbb{R}^n$ and $m \in \mathbb{Z}$ such that

$$a_1 2^{-a_0 m} \le |\theta(v,m)| \le a_2 2^{-a_0 m},$$

which replaces (1) and the respective affine transforms satisfy a condition similar to (2); conditions (c)-(e) are void.

In the case of a *continuous ellipsoid cover* $\Theta := \cup_{t\in\mathbb{R}}\Theta_t$, an ellipsoid $\theta(v,t) \in \Theta_t$ is associated to every $v \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that

$$a_1 2^{-a_0 t} \le |\theta(v,t)| \le a_2 2^{-a_0 t},$$

i.e. the scale is continuous as well. For more detail and the exact definitions of ellipsoid covers, see §2.2 in [12]

Examples. (i) The one parameter family of diagonal dilation matrices

$$D_t = \operatorname{diag}(2^{-tb_1}, 2^{-tb_2}, \dots, 2^{-tb_n}), \quad b_j > 0,\ j = 1,\dots,n,$$

apparently induces a continuous ellipsoid cover of $\mathbb{R}^n$.

(ii) Suppose A is an $n \times n$ real matrix with eigenvalues λ satisfying $|\lambda| > 1$. Then it is easy to see that the affine transforms $A_{v,m}(x) := A^{-m}x + v$, $v \in \mathbb{R}^n$, $m \in \mathbb{Z}$, define a semi-continuous ellipsoid cover (dilations) of $\mathbb{R}^n$. This particular kind of dilations are used in [2, 3, 4] for the development of anisotropic Hardy, Besov, and Triebel-Lizorkin spaces.

(iii) The continuous covers used in Section 6 (see also §7 in [12]) are nontrivial examples of anisotropic ellipsoid covers of $\mathbb{R}^2$, where the ellipsoids change rapidly from point to point and in depth.

The point is that, on the one hand, continuous and semi-continuous covers are easier to construct and, on the other, given a semi-continuous or continuous cover one can construct a discrete ellipsoid cover with essentially the same (equivalent) metric (see [12]).

Quasi-distance. A quasi-distance is naturally associated with any discrete, semi-continuous or continuous ellipsoid covers of $\mathbb{R}^n$. Recall that a *quasi-distance* on a set $X \neq \emptyset$ is a map $\rho : X \times X \to [0,\infty)$ satisfying the conditions:

(a) $\rho(x,y) = 0 \Longleftrightarrow x = y$,
(b) $\rho(y,x) = \rho(x,y)$,
(c) $\rho(x,z) \le \kappa(\rho(x,y) + \rho(y,z))$.

Here $\kappa \ge 1$ is a constant.

Definition 2.2. Assuming that Θ is a continuous, semi-continuous or discrete ellipsoid cover of $\mathbb{R}^n$, we define $\rho : \mathbb{R}^n \times \mathbb{R}^n \to [0,\infty)$ by

$$\rho(x,y) := \inf\{|\theta| : \theta \in \Theta \text{ and } x,y \in \theta\}. \tag{4}$$

Proposition 2.1. [12] *For any ellipsoid cover Θ of $\mathbb{R}^n$ the map $\rho : \mathbb{R}^n \times \mathbb{R}^n \to [0,\infty)$ defined above is a quasi-distance on $\mathbb{R}^n$.*

Spaces of homogeneous type were first introduced in [8] (see also [9, 16]) as a means for extending the Calderón-Zygmund theory of singular integral operators to more general settings. Let X be a topological space endowed with a Borel measure μ and a quasi-distance $\rho(\cdot,\cdot)$. Assume that the balls $B_\rho(x,r) := \{y \in X : \rho(x,y) < r\}$, $x \in X$, $r > 0$, form a basis for the topology in X. The space (X,ρ,μ) is said to be of *homogenous type* if there exists a constant $A > 0$ such that for all $x \in X$ and $r > 0$,

$$\mu(B_\rho(x,2r)) \le A\mu(B_\rho(x,r)). \tag{5}$$

If (5) holds then μ is said to be a *doubling measure* [25, Chapter 1, 1.1]. A space of homogeneous type is said to be *normal*, if uniformly $\mu(B(x,r)) \sim r$.

Suppose Θ is an ellipsoid cover of $\mathbb{R}^n$ and let $\rho(\cdot,\cdot)$ be the associated quasi-distance, defined in (4). Denote $B_\rho(x,r) := \{y \in \mathbb{R}^n : \rho(x,y) < r\}$ for $x \in \mathbb{R}^n$, $r > 0$. As is shown in [12] there exist ellipsoids $\theta', \theta'' \in \Theta$ such that $\theta' \subset B_\rho(x,r) \subset \theta''$ and $|\theta'| \sim |B_\rho(x,r)| \sim |\theta''| \sim r$. Consequently, $\mathbb{R}^n$ equipped with the distance $\rho(\cdot,\cdot)$ and the Lebesgue measure, i.e. $(\mathbb{R}^n,\rho,dx)$ is a homogeneous type space. Therefore, the machinery of spaces of homogeneous type can be employed to our purposes here.

2.2 Comparison of ellipsoid covers with nested triangulations in $\mathbb{R}^2$

An alternative way of introducing anisotropic structures in $\mathbb{R}^2$ is through multilevel nested triangulations. The strong locally regular (SLR) triangulations, introduced in [20], provide a structure compatible with ellipsoid covers. We next recall briefly the definition of SLR-triangulations.

We call $\mathcal{T} = \bigcup_{m\in\mathbb{Z}} \mathcal{T}_m$ an SLR-triangulation of $\mathbb{R}^2$ with levels $\{\mathcal{T}_m\}$ if the following conditions are obeyed:

(a) Every level $\mathcal{T}_m$ consists of closed triangles with disjoint interiors which cover $\mathbb{R}^2$ and there are no hanging vertices.

(b) $\mathcal{T}_{m+1}$ is a refinement of $\mathcal{T}_m$ ($m \in \mathbb{Z}$) and each triangle $\triangle \in \mathcal{T}_m$ is subdivided and has uniformly bounded number of children in $\mathcal{T}_{m+1}$.

(c) For each $\triangle \in \mathcal{T}$ let $A_\triangle$ be an affine transform of the form

$$A_\triangle(x) = M_\triangle x + v_\triangle, \quad M_\triangle \in \mathbb{R}^{n\times n},$$

such that $\triangle = A_\triangle(\triangle^*)$, where $\triangle^*$ is an equilateral reference triangle. Now the condition is that for any $\triangle \in \mathcal{T}_m$ and $\triangle' \in \mathcal{T}_m \cup \mathcal{T}_{m+1}$ such that $\triangle' \cap \triangle \neq \emptyset$ one has

$$c_1 \le 1/\|M_{\triangle'}^{-1} M_\triangle\|_{\ell_2\to\ell_2} \le \|M_\triangle^{-1} M_{\triangle'}\|_{\ell_2\to\ell_2} \le c_2. \tag{6}$$

In [20] condition (c) is formulated in an equivalent form via a minimum angle condition.

Note that the multilevel SLR-triangulations provide a means for constructing discrete ellipsoid covers of $\mathbb{R}^2$. Given an SLR-triangulation $\mathscr{T}$ one considers for each triangle $\triangle \in \mathscr{T}$ the minimum area circumscribed ellipse. Then one dilates the resulting ellipses by a sufficiently large factor > 1 to obtain a discrete ellipse cover of $\mathbb{R}^2$.

The main advantage of ellipse covers over SLR-triangulations is that the latter are nested which makes them less flexible and harder to construct. On the other hand, as shown in [13] in presence of an SLR-triangulation it is easier to construct building blocks consisting of piecewise polynomials. Also the respective generalized Besov spaces and nonlinear approximation theory are easier to develop. We will be more specific about these issues later on.

3 Building blocks

The construction of simple elements (building blocks) which allow to represent the functions and characterize the norms of the spaces of interest is imperative for our theory. Here we first define a sequence of of bases consisting of C^∞ functions supported on the ellipsoids of the underlying anisotropic ellipsoid cover. Secondly, we construct compactly supported duals which generate local projectrors and two-level-split bases. Thirdly, we develop smooth global duals which provide polynomial reproducing kernels that we utilize to the construction of anisotropic frames.

3.1 Construction of a multilevel system of bases

Given a discrete ellipsoid cover Θ of $\mathbb{R}^n$, we first construct for each level $m \in \mathbb{Z}$ a stable basis Φ_m whose elements are smooth functions supported on the ellipsoids of Θ_m. The procedure begins by first *coloring the ellipsoids in* Θ. It is easy to see that Θ can be split into no more than $2N_1$ disjoint subsets (colors) $\{\Theta^\ell\}_{\ell=1}^{2N_1}$ so that for any $m \in \mathbb{Z}$ neither two ellipsoids $\theta', \theta'' \in \Theta_m \cup \Theta_{m+1}$ with $\theta' \cap \theta'' \neq \emptyset$ are of the same color.

Our second step is to construct locally independent piecewise polynomial bumps. For fixed positive integers M and k ($M \geq k$) we define

$$\tilde{\phi}_\ell(x) := (1-|x|^2)_+^{M+\ell k}, \quad \ell = 1,2,\ldots,2N_1.$$

Notice that $\tilde{\phi}_\ell$, $\ell = 1,\ldots,2N_1$, being of different degrees are linearly independent on any ball contained in $B^* = B(0,1)$.

The next step is to smooth out each $\tilde{\phi}_\ell$ by convolving it with a compactly supported C^∞ function. Namely, let $h \in C^\infty(\mathbb{R}^n)$ be such that $\operatorname{supp} h = \overline{B^*}$, $h \geq 0$, and

$\int_{\mathbb{R}^n} h = 1$. Denote $h_\delta(x) := \delta^{-n} h(\delta^{-1}x)$. Then for $0 < \delta < 1$ (we choose δ sufficiently small) the bump

$$\phi_\ell^* := \tilde{\phi}_\ell * h_\delta$$

belongs to C^∞, ϕ_ℓ^* is a polynomial of degree exactly $2(M+\ell k)$ on $B(0,1-\delta)$ and $\operatorname{supp}\phi_\ell^* = \overline{B(0,1+\delta)}$. Now we define $\phi_\ell(x) := \phi_\ell^*((1+\delta)x)$.

For any $\theta \in \Theta$ we let A_θ denote the affine transform from Definition 2.1 such that $A_\theta(B^*) = \theta$ and set

$$\phi_\theta := \phi_\ell \circ A_\theta^{-1} \quad \text{for} \quad \theta \in \Theta^\ell,\ 1 \le \ell \le 2N_1.$$

We introduce an mth level partition of unity by defining for each $\theta \in \Theta_m$

$$\varphi_\theta := \frac{\phi_\theta}{\sum_{\theta' \in \Theta_m} \phi_{\theta'}}. \tag{7}$$

By property (d) of ellipsoids covers it follows that $\sum_{\theta\in\Theta_m} \varphi_\theta(x) = 1$ for $x \in \mathbb{R}^n$.

Let

$$\{P_\beta : |\beta| \le k-1\}, \quad \text{where } \deg P_\beta = |\beta|, \tag{8}$$

be an orthonormal basis in $L_2(B^*)$ for the space $\mathscr{P}_k$ of all polynomials in n variables of total degree $k-1$. For each $\theta \in \Theta$ and $|\beta| < k$ we define

$$P_{\theta,\beta} := |\theta|^{-1/2} P_\beta \circ A_\theta^{-1} \quad \text{and} \quad g_{\theta,\beta} := \varphi_\theta P_{\theta,\beta}. \tag{9}$$

To simplify our notation, we denote

$$\Lambda_m := \{\lambda := (\theta,\beta) : \theta \in \Theta_m, |\beta| < k\} \quad \text{and} \quad g_\lambda := g_{\theta,\beta}, \quad \lambda = (\theta,\beta). \tag{10}$$

Also, for $\lambda = (\theta,\beta)$ we will denote by θ_λ and β_λ the components of λ.

Now we define the mth level basis Φ_m by

$$\Phi_m := \{g_\lambda : \lambda \in \Lambda_m\} \quad \text{and set} \quad S_m := \operatorname{span}(\Phi_m), \tag{11}$$

where "span" means "closed span".

By the definition of $\{g_\lambda\}$ it readily follows that $\mathscr{P}_k \subset S_m$. More importantly, Φ_m is locally linearly independent and L_p-stable, which will be recorded in the next theorem.

Theorem 3.1. *Any function $f \in S_m$ has a unique representation*

$$f(x) = \sum_{\lambda\in\Lambda_m} \langle f, \tilde{g}_\lambda\rangle g_\lambda(x), \tag{12}$$

where for every $x \in \mathbb{R}^n$ the sum is finite and the functions $\{\tilde{g}_\lambda\}$ have the following properties: $\operatorname{supp}(\tilde{g}_\lambda) \subset \theta_\lambda$, $\|\tilde{g}_{\theta,\beta}\|_p \sim |\theta|^{1/p-1/2}$ *and the biorthogonal relation $\langle g_{\lambda'}, \tilde{g}_\lambda\rangle = \delta_{\lambda',\lambda}$ holds. Moreover, for any $f \in S_m \cap L_p$, $0 < p \le \infty$, such that $f = \sum_{\lambda\in\Lambda_m} c_\lambda g_\lambda$ we have*

$$\|f\|_p \sim \Big(\sum_{\lambda \in \Lambda_m} \|c_\lambda g_\lambda\|_p^p \Big)^{1/p} \tag{13}$$

with the obvious modification when $p = \infty$.

The proof of this theorem is based on the local linear independence of the functions $\{g_\lambda : \lambda \in \Lambda_m\}$ and uses a compactness argument, see [12] for the details.

We will denote $\widetilde{\Phi}_m := \{\tilde{g}_\lambda : \lambda \in \Lambda_m\}$.

3.2 Compactly supported duals and local projectors

Our next step is to introduce simple operators which map L_p^{loc} into S_m and locally preserve the polynomials $P \in \mathscr{P}_k$ with $\mathscr{P}_k$ being the set of all polynomials of degree $< k$. These operators will give us a vehicle for developing a decomposition scheme which allows to characterize the anisotropic Besov norms induced by ellipsoid covers of $\mathbb{R}^n$.

Using the bases $\{\Phi_m\}$ and their duals $\{\widetilde{\Phi}_m\}$ from Theorem 3.1 we introduce projectors Q_m mapping L_p^{loc} $(1 \le p \le \infty)$ onto the spaces S_m defined by

$$Q_m f := \int_{\mathbb{R}^n} Q_m(x,y) f(y) dy \quad \text{with} \quad Q_m(x,y) := \sum_{\lambda \in \Lambda_m} \tilde{g}_\lambda(y) g_\lambda(x). \tag{14}$$

Evidently, Q_m is a linear operator which maps L_p^{loc} into S_m and preserves locally all polynomials from $\mathscr{P}_k$. To be more specific, setting

$$\theta^* := \cup \{\theta' \in \Theta_m : \theta \cap \theta' \neq \emptyset\} \quad \text{for } \theta \in \Theta_m, \tag{15}$$

it is easy to see that if $f|_{\theta^*} = P|_{\theta^*}$ with $P \in \mathscr{P}_k$, then $Q_m f|_\theta = P|_\theta$.

Another simple operator with similar properties is given in [12].

Evidently, the operators $\{Q_m\}$ from above are no longer usable, when working in L_p with $p < 1$. In this case, for a given ellipsoid $\theta \in \Theta$, we let $T_{\theta,p} : L_p(\theta) \to \mathscr{P}_k|_\theta$ be a projector such that

$$\|f - T_{\theta,p} f\|_{L_p(\theta)} \le c E_k(f,\theta)_p, \quad f \in L_p(\theta), \tag{16}$$

where $E_k(f,\theta)_p := \inf_{P \in \mathscr{P}_k} \|f - P\|_{L_p(\theta)}$. Thus $T_{\theta,p} f$ is simply a near best approximation to f from $\mathscr{P}_k$ in $L_p(\theta)$, and hence $T_{\theta,p}$ can be realized as a linear projector onto $\mathscr{P}_k|_\theta$ if $p \ge 1$ by using, say, the Averaged Taylor polynomials, see e.g. [13]. Of course, $T_{\theta,p}$ is a nonlinear operator if $p < 1$.

We now define the operator $T_{m,p} : L_p^{\mathrm{loc}} \to S_m$ by

$$T_{m,p} f := \sum_{\theta \in \Theta_m} \varphi_\theta T_{\theta,p} f. \tag{17}$$

Evidently, the operator $T_{m,p}$ $(0 < p \le \infty)$ is a local projector onto $\mathscr{P}_k$ (nonlinear if $p < 1$) just like Q_m. Since $T_{m,p}f \in S_m$, it can be represented in terms of the basis functions g_λ as

$$T_{m,p}f = \sum_{\theta\in\Theta_m}\sum_{|\beta|<k} b_{\theta,\beta}(f)g_{\theta,\beta} = \sum_{\lambda\in\Lambda_m} b_\lambda(f)g_\lambda, \tag{18}$$

where $b_\lambda(f) := \langle T_{m,p}f, \tilde{g}_\lambda\rangle$ depends nonlinearly on f if $p < 1$.

In summary, if $\hat{T}_m := Q_m$ or $\hat{T}_m := T_{m,p}$, then

$$\hat{T}_m f = \sum_{\lambda\in\Lambda_m} b_\lambda(f)g_\lambda, \ \text{ where } \ b_\lambda(f) = \begin{cases} \langle f, \tilde{g}_\lambda\rangle & \text{if } \hat{T}_m = Q_m, \\ \langle T_{m,p}f, \tilde{g}_\lambda\rangle & \text{if } \hat{T}_m = T_{m,p}. \end{cases} \tag{19}$$

We now recall briefly the definition of local and global moduli of smoothness that are standard means for describing the quality of approximation. The forward differences of a function f on a set $E \subset \mathbb{R}^n$ in direction $h \in \mathbb{R}^n$ are defined by

$$\Delta_h^k f(x) := \sum_{j=0}^{k} (-1)^{k+j}\binom{k}{j} f(x+jh) \quad \text{if } [x, x+kh] \subset E$$

and $\Delta_h^k f(x) := 0$ otherwise. Then the kth L_p-moduli of smoothness on E and $\mathbb{R}^n$ are defined by

$$\omega_k(f,E)_p := \sup_{h\in\mathbb{R}^n} \|\Delta_h^k f\|_{L_p(E)} \quad \text{and} \quad \omega_k(f,t)_p := \sup_{|h|\le t} \|\Delta_h^k f\|_p, \quad t > 0. \tag{20}$$

We next give the most important properties of the operators Q_m and $T_{m,p}$ from above.

Proposition 3.1. *Suppose $\hat{T}_m$ is any of the operators Q_m or $T_{m,p}$ if $1 \le p \le \infty$, and $\hat{T}_m := T_{m,p}$ if $0 < p < 1$. Then for $f \in L_p^{\mathrm{loc}}$ and $\theta \in \Theta_m$ $(m \in \mathbb{Z})$*

$$\|f - \hat{T}_m f\|_{L_p(\theta)} \le c \sum_{\theta'\in\Theta_m:\, \theta'\cap\theta\neq\emptyset} \omega_k(f,\theta')_p,$$

and $\|f - \hat{T}_m f\|_{L_p(K)} \to 0$ as $m \to \infty$ for any compact $K \subset \mathbb{R}^n$.

Furthermore, if $f \in L_p$ $(L_\infty := C_0)$, then $\|f - \hat{T}_m f\|_p \to 0$ as $m \to \infty$.

3.3 Two-level-split bases

Assume that T_m $(m \in \mathbb{Z})$ is one of the operators Q_m or $T_{m,p}$ if $p \ge 1$, and $T_m := T_{m,p}$ if $p < 1$, defined in §3.2. These operators and the bases $\{\Phi_m\}_{m\in\mathbb{Z}}$ from (11) will be used to define two-level-split bases which will play an important role in what follows.

We will make use of the following representation of consecutive level polynomial bases, defined in (9):

$$P_{\theta,\alpha} =: \sum_{|\beta|<k} C^{\theta,\eta}_{\alpha,\beta} P_{\eta,\beta}, \quad \theta \in \Theta_m, \quad \eta \in \Theta_{m+1}. \tag{21}$$

Then since $\sum_{\eta\in\Theta_{m+1}} \varphi_\eta = 1$, we have

$$P_{\theta,\alpha} = \sum_{\eta\in\Theta_{m+1}:\theta\cap\eta\neq\emptyset} \sum_{|\beta|<k} C^{\theta,\eta}_{\alpha,\beta} P_{\eta,\beta}\varphi_\eta \quad \text{on} \quad \theta.$$

This yields

$$\begin{aligned} T_{m+1}f - T_m f &= \sum_{\eta\in\Theta_{m+1}} \sum_{|\beta|<k} b_{\eta,\beta}(f) P_{\eta,\beta}\varphi_\eta - \sum_{\theta\in\Theta_m} \sum_{|\alpha|<k} b_{\theta,\alpha}(f) P_{\theta,\alpha}\varphi_\theta \\ &= \sum_{\theta\in\Theta_m} \varphi_\theta \sum_{\eta\in\Theta_{m+1}} \sum_{|\beta|<k} b_{\eta,\beta}(f) P_{\eta,\beta}\varphi_\eta \\ &\quad - \sum_{\theta\in\Theta_m} \sum_{|\alpha|<k} b_{\theta,\alpha}(f) \sum_{\theta\cap\eta\neq\emptyset} \sum_{|\beta|<k} m^{\theta,\eta}_{\alpha,\beta} P_{\eta,\beta}\varphi_\theta\varphi_\eta \\ &= \sum_{\eta\in\Theta_{m+1}} \sum_{\theta\in\Theta_m:\theta\cap\eta\neq\emptyset} \sum_{|\beta|<k} \Big\{ b_{\eta,\beta}(f) - \sum_{|\alpha|<k} m^{\theta,\eta}_{\alpha,\beta} b_{\theta,\alpha}(f) \Big\} P_{\eta,\beta}\varphi_\eta\varphi_\theta, \end{aligned} \tag{22}$$

where $b_\lambda(f)$ are given by (19) and depends on the choice of T_m. Thus, setting

$$\mathscr{V}_m := \{\nu = (\eta,\theta,\beta) : \eta \in \Theta_{m+1}, \theta \in \Theta_m, \theta\cap\eta \neq 0, |\beta| < k\}, \quad m \in \mathbb{Z}, \tag{23}$$

the building blocks in (22) have the form

$$F_\nu := P_{\eta,\beta}\varphi_\eta\varphi_\theta, \quad \nu = (\eta,\theta,\beta) \in \mathscr{V}_m, \tag{24}$$

where $P_{\eta,\beta}$ are defined in (9) and φ_η, φ_θ are from (7). We define

$$\mathscr{F}_m := \{F_\nu : \nu \in \mathscr{V}_m\} \quad \text{and} \quad W_m := \operatorname{span}(\mathscr{F}_m), \quad m \in \mathbb{Z}. \tag{25}$$

Note that $F_\nu \in C^\infty$, $\operatorname{supp} F_\nu = \overline{\theta\cap\eta}$ if $\nu = (\eta,\theta,\beta)$ and $\|F_\nu\|_2 \sim 1$.

One uses the argument of the proof of Theorem 3.1 (see [12]) to establish the stability of the two-level-split bases:

Theorem 3.2. *Any $f \in W_m$ has a unique representation*

$$f = \sum_{\nu\in\mathscr{V}_m} c_\nu(f) F_\nu, \tag{26}$$

where the dual functionals $c_\nu(\cdot)$ are of the following form: For each $\nu \in \mathscr{V}_m$, $\nu = (\eta,\theta,\beta)$, there is an ellipsoid $B_\nu \subset \theta\cap\eta$ with $|B_\nu| \sim |\eta|$ and $B_\nu = A_\eta(B^*_\nu)$ for some ball $B^*_\nu \subset B^*$ such that $c_\nu(f) = \langle f, \tilde{F}_\nu\rangle$, where $\operatorname{supp}\tilde{F}_\nu \subset \overline{B}_\nu$, $\|\tilde{F}_\nu\|_p \sim |\eta|^{1/p-1/2}$.

Moreover, if $f \in W_m$ and $f = \sum_{\nu\in\mathscr{V}_m} a_\nu F_\nu$, then

$$\|f\|_p \sim \Big(\sum_{\nu \in \mathscr{V}_m} \|a_\nu F_\nu\|_p^p \Big)^{1/p}, \quad 0 < p \le \infty, \tag{27}$$

with the obvious modification when $p = \infty$.

Using the results from §3.2 one easily derives multilevel decompositions of functions using the two-level-split bases from above.

Theorem 3.3. *For any* $f \in L_p^{\mathrm{loc}}(\mathbb{R}^n)$, $0 < p \le \infty$,

$$f = T_0 f + \sum_{m \ge 0} (T_{m+1} f - T_m f) = \sum_{m \ge -1} \sum_{\nu \in \mathscr{V}_m} d_\nu(f) F_\nu, \tag{28}$$

where the convergence is in $L_p(K)$ *for all compacta* $K \subset \mathbb{R}^n$. *Here for* $m \ge 0$

$$d_\nu(f) = b_{\eta,\beta}(f) - \sum_{|\alpha|<k} C_{\alpha,\beta}^{\theta,\eta} b_{\theta,\beta}(f), \quad \nu := (\eta, \theta, \beta) \tag{29}$$

with $C_{\alpha,\beta}^{\theta,\eta}$ *from* (21), *while* $\mathscr{V}_{-1} := \Lambda_0$, $F_\lambda := g_\lambda$ *and* $d_\lambda(f) := b_\lambda(f)$ *if* $\lambda \in \mathscr{V}_{-1}$.

Moreover, if $f \in L_p$ ($L_\infty := C_0$), *then* (28) *as well as*

$$f = \sum_{m \in \mathbb{Z}} (T_{m+1} - T_m) f \tag{30}$$

hold in L_p.

3.4 Global duals and polynomial reproducing kernels

A substantial drawback of the operators Q_m and $T_{m,p}$ considered in §§3.2-3.3 is that their transposed operators do not reproduce polynomials. For instance, it is easy to see that for the operator Q_m from (14) we have $Q_m P(x) = \int_{\mathbb{R}^n} Q_m(x,y) P(y) dy$ $\forall P \in \mathscr{P}_k$, however, $Q_m P(y) = \int_{\mathbb{R}^n} Q_m(x,y) P(x) dx$ is no longer true for $P \in \mathscr{P}_k$. Consequently, these operators do not fit in the general framework of approximation to the identity operators in homogeneous spaces, which allows to construct anisotropic wavelet frames (see e.g. [16]). This problem is fixed in [14] by introducing smooth duals to the bases $\{g_\lambda\}_{\lambda \in \Lambda_m}$, which we describe next.

As in [14] to simplify our set-up we will assume for the rest of this section that in the definition of ellipsoid covers of $\mathbb{R}^n$ we have $a_0 = 0$ (see Definition 2.1). Also to make our presentation more compatible with [14] we will slightly change our notation assuming that all operators of interest reproduce polynomials of degree $< r$ instead of degree $< k$.

The next step is to introduce an appropriate generalization to higher orders of the *approximation to the identity* definition given in [16]. To this end we first have to establish some convenient notation. Let $K(x,y)$ be a smooth kernel. For $x, y \in \mathbb{R}^n$ the Taylor representation of $K(x,y)$ centered at x with y fixed can be written in the

form

$$K(z,y) = T_{r-1,x}(K(\cdot,y))(z) + R_{r,x}(K(\cdot,y))(z), \quad z \in \mathbb{R}^n, \tag{31}$$

where T_{r-1} is the Taylor polynomial of degree $r-1$ and $R_{r,x}$ is the rth order Taylor remainder.

In the particular case of spaces of homogeneous type generated by an anisotropic ellipsoid cover of $\mathbb{R}^n$ with a qusi-distance $\rho(\cdot,\cdot)$ we will need the notation

$$\mu(x,y,d) := \begin{cases} \mu_0 & \text{if} \quad \rho(x,y) < d, \\ \mu_1 & \text{if} \quad \rho(x,y) \geq d. \end{cases} \tag{32}$$

Definition 3.1. Let $(\mathbb{R}^n, \rho, dx)$ be a normal space of homogeneous type. A sequence of kernel operators $\{S_m\}_{m\in\mathbb{Z}}$, formally defined by $S_m(f)(x) := \int_{\mathbb{R}^n} S_m(x,y)f(y)dy$, is an *approximation to the identity of order* (μ,δ,r), where $\mu = (\mu_0,\mu_1)$, $0 < \mu_0 \leq \mu_1 \leq 1$, $\delta > 0$, $r \in \mathbb{N}$, with respect to $\rho(\cdot,\cdot)$, if for some constant $c > 0$ the following conditions are satisfied:

(i) $|S_m(x,y)| \leq c\frac{2^{-m\delta}}{(2^{-m}+\rho(x,y))^{1+\delta}}, \quad \forall x,y \in \mathbb{R}^n.$

(ii) For $1 \leq k \leq r$ and all $x,y,z \in \mathbb{R}^n$,

$$|R_{k,x}(S_m(\cdot,y))(z)| \leq c\rho(x,z)^{\mu(x,z,2^{-m})k} \times \left(\frac{2^{-m\delta}}{(2^{-m}+\rho(x,y))^{1+\delta+\mu(x,z,2^{-m})k}} + \frac{2^{-m\delta}}{(2^{-m}+\rho(y,z))^{1+\delta+\mu(x,z,2^{-m})k}} \right),$$

$$|R_{k,y}(S_m(x,\cdot))(z)| \leq c\rho(y,z)^{\mu(y,z,2^{-m})k} \times \left(\frac{2^{-m\delta}}{(2^{-m}+\rho(x,y))^{1+\delta+\mu(y,z,2^{-m})k}} + \frac{2^{-m\delta}}{(2^{-m}+\rho(x,z))^{1+\delta+\mu(y,z,2^{-m})k}} \right),$$

(iii) For $1 \leq k \leq r$ and all $x,x',y,y' \in \mathbb{R}^n$

$$\begin{aligned} &|R_{k,y}(R_{k,x}(S_m(\cdot,\cdot))(x'))(y')|, |R_{k,x}(R_{k,y}(S_m(\cdot,\cdot))(y'))(x')| \\ &\leq c\rho(x,x')^{\mu(x,x',2^{-m})k}\rho(y,y')^{\mu(y,y',2^{-m})k} \\ &\times \left(\frac{2^{-m\delta}}{(2^{-m}+\rho(x,y))^{1+\delta+\mu(x,x',2^{-m})k+\mu(y,y',2^{-m})k}} \right. \\ &\qquad + \frac{2^{-m\delta}}{(2^{-m}+\rho(x,y'))^{1+\delta+\mu(x,x',2^{-m})k+\mu(y,y',2^{-m})k}} \\ &+ \frac{2^{-m\delta}}{(2^{-m}+\rho(x',y))^{1+\delta+\mu(x,x',2^{-m})k+\mu(y,y',2^{-m})k}} \\ &\qquad \left. + \frac{2^{-m\delta}}{(2^{-m}+\rho(x',y'))^{1+\delta+\mu(x,x',2^{-m})k+\mu(y,y',2^{-m})k}} \right) \end{aligned}$$

[To clarify our notation, denote $g_m(x,x',y) := R_{k,x}\left(S_m(\cdot,y)\right)(x')$, then for fixed $x,x' \in \mathbb{R}^n$, $R_{k,y}\left(R_{k,x}\left(S_m(\cdot,\cdot)\right)(x')\right)(y') := R_{k,y}\left(g_m(x,x',\cdot)\right)(y')$].

(iv) $\quad P(x) = \int_{\mathbb{R}^n} S_m(x,y)P(y)dy$ and $P(y) = \int_{\mathbb{R}^n} S_m(x,y)P(x)dx$ for all $P \in \mathscr{P}_r$.

Note that the definition of an approximation of the identity given in [16] corresponds to the case $0 < \delta < r = 1$.

To construct well localized kernels $S_m(x,y)$ which reproduce polynomials we need to construct an appropriate dual basis to Φ_m. Let G_m be the Gram matrix

$$G_m := \left[A_{\lambda,\lambda'}\right]_{\lambda,\lambda' \in \Lambda_m}, \quad A_{\lambda,\lambda'} := \langle g_\lambda, g_{\lambda'} \rangle := \int_{\mathbb{R}^n} g_\lambda g_{\lambda'}.$$

By Theorem 3.1, for any sequence $t = (t_\lambda)_{\lambda \in \Lambda_m}$ in $l_2(\Lambda_m)$ we have

$$c_1 \|t\|_{l_2} \le \langle G_m t, t \rangle = \Big\| \sum_{\lambda \in \Lambda_m} t_\lambda g_\lambda \Big\|_2 \le c_2 \le \|t\|_{l_2},$$

where the constants $c_1, c_2 > 0$ are independent on t and m. Therefore, the operator $G_m : l_2 \to l_2$ with matrix G_m is symmetric, positive and $c_1 I \le G_m \le c_2 I$. Hence, G_m^{-1} exists and $c_2^{-1} I \le G_m^{-1} \le c_1^{-1} I$. Denote by $G_m^{-1} =: \left[B_{\lambda,\lambda'}\right]_{\lambda,\lambda' \in \Lambda_m}$ the matrix of the operator G_m^{-1}.

The next lemma shows that the entries of G_m^{-1} decay away from its main diagonal at sub-exponential rate.

Lemma 3.1. [14] *There exist constants $0 < q_*, \gamma < 1$ and $c > 0$ depending only on $\mathbf{p}(\Theta)$ and r such that for any entry $B_{\lambda,\lambda'}$ of G_m^{-1} $(\lambda, \lambda' \in \Lambda_m)$ and points $x \in \theta_\lambda$, $y \in \theta_{\lambda'}$*

$$|B_{\lambda,\lambda'}| \le c q_*^{(2^m \rho(x,y))^\gamma}. \tag{33}$$

Definition of smooth duals. We define new duals by

$$\tilde{\tilde{g}}_\lambda := \sum_{\lambda' \in \Lambda_m} B_{\lambda,\lambda'} g_{\lambda'}, \quad \lambda \in \Lambda_m, \tag{34}$$

and set $\tilde{\tilde{\Phi}}_m := \{\tilde{\tilde{g}}_\lambda\}_{\lambda \in \Lambda_m}$. For $\lambda \in \Lambda_m$, let x_0 be any point in θ_λ. Combining (33) and (34) it follows that

$$|\tilde{\tilde{g}}_\lambda(x)| \le c 2^{-m/2} \sum_{x \in \theta_{\lambda'}} |B_{\lambda,\lambda'}| \le c 2^{-m/2} q_*^{(2^m \rho(x,x_0))^\gamma}. \tag{35}$$

Therefore, each $\tilde{\tilde{g}}_\lambda$ has sub-exponential decay with respect to the quasi-distance induced by Θ. Also, it is easy to verify the biorthogonality relation, namely,

$$\langle g_\lambda, \tilde{\tilde{g}}_{\lambda'} \rangle = \sum_{\lambda'' \in \Lambda_m} B_{\lambda',\lambda''} \langle g_\lambda, g_{\lambda''} \rangle = \left(G_m^{-1} G_m\right)_{\lambda',\lambda} = \delta_{\lambda,\lambda'}.$$

We use the bases Φ_m and $\tilde{\tilde{\Phi}}_m$ to introduce an approximation to the identity determined by the operators $\{S_m\}_{m\in\mathbb{Z}}$ with kernels

$$S_m(x,y) := \sum_{\lambda\in\Lambda_m} g_\lambda(x)\tilde{\tilde{g}}_\lambda(y). \tag{36}$$

In the next theorem we record the fact that these kernels define the desired approximation to the identity.

Theorem 3.4. [14] *For a discrete ellipsoid cover Θ, the kernels from (36) define an approximation to the identity with respect to the quasi-distance $\rho(\cdot,\cdot)$ induced by Θ. Here the vector μ can be defined as $\mu := (a_6, a_4)$, the parameter δ can be selected arbitrarily large and the parameter r is the degree of the polynomials used in the construction of the local ellipsoid "bumps" in §3.1.*

3.5 Construction of anisotropic wavelet frames

Wavelet operators. Let $\{S_m\}_{m\in\mathbb{Z}}$ be an approximation to the identity of order (μ,δ,r). Then evidently the kernels of the *wavelet operators* $D_m := S_{m+1} - S_m$ satisfy conditins (i)-(iii) in Definition 3.1, while the polynomial reproduction condition (iv) is replaced by the following *zero moment* condition

$$\int_{\mathbb{R}^n} D_m(x,y)P(y)dy = 0, \quad \int_{\mathbb{R}^n} D_m(x,y)P(x)dx = 0 \ \ \forall P\in\mathscr{P}_r. \tag{37}$$

The next lemma shows that any two wavelet operators (kernels) from different scales are "almost orthogonal".

Lemma 3.2. [14] *Suppose two kernel operators $\{D_m^1\}_{m\in\mathbb{Z}}$ and $\{D_m^2\}_{m\in\mathbb{Z}}$ satisfy (37) for some $r\geq 1$ and conditions (i)-(ii) of an approximation to the identity of order (μ,δ,r) for some $\delta\geq\mu_1 r$. Then*

$$|D_k^1 D_l^2(x,y)| \leq c2^{-|k-l|\mu_0 r}\frac{2^{-\min\{k,l\}\delta}}{\left(2^{-\min\{k,l\}}+\rho(x,y)\right)^{1+\delta}}, \quad k,l\in\mathbb{Z}. \tag{38}$$

Dual wavelet operators. In this section we leverage significantly on the results of Han and Sawyer [19] (see also [16]) concerning the Calderón reproducing formula in spaces of homogeneous type and adapt them to our specific setting. We begin with the definitions for anisotropic test functions and molecules.

Definition 3.2. Let $\rho(\cdot,\cdot)$ be a quasi-distance on $\mathbb{R}^n$. A function $f\in C(\mathbb{R}^n)$ is said to be in the *anisotropic test function space* $\mathscr{M}(\varepsilon,\delta,x_0,t)$, $0<\varepsilon,\delta\leq 1$, $x_0\in\mathbb{R}^n$, $t\in\mathbb{R}$, if there exists a constant $C>0$ such that

(i) $|f(x)|\leq C\frac{2^{-t\delta}}{(2^{-t}+\rho(x,x_0))^{1+\delta}}$ $\forall x\in\mathbb{R}^n$.

(ii) $|f(x)-f(y)|\leq C\rho(x,y)^\varepsilon\frac{2^{-t\delta}}{(2^{-t}+\rho(x,x_0))^{1+\delta+\varepsilon}}$ for all $x,y\in\mathbb{R}^n$,

where $\rho(x,y) \le \frac{1}{2\kappa}(2^{-t}+\rho(x,x_0))$ with κ the constant of the quasi-distance (see §2.1).

One can easily show that $\mathscr{M}(\varepsilon,\delta,x_0,t)$ is a Banach space with norm $\|f\|_{\mathscr{M}}$ defined as the infimum of all constants C such that (i)-(ii) are valid. We also denote $\mathscr{M}(\varepsilon,\delta) := \mathscr{M}(\varepsilon,\delta,0,0)$.

Definition 3.3. The set of *molecules* $\mathscr{M}_0(\varepsilon,\delta,x_0,t)$ is defined as the set of all anisotropic test functions $f \in \mathscr{M}(\varepsilon,\delta,x_0,t)$ such that $\int_{\mathbb{R}^n} f(y)\,dy = 0$.

We denote by $\mathscr{M}_0(\varepsilon,\delta)$ the subspace of all molecules in $\mathscr{M}(\varepsilon,\delta)$.

For some $\gamma > \varepsilon$, let $\overset{\circ}{\mathscr{M}}(\varepsilon,\delta)$ be the closure of $\mathscr{M}(\gamma,\delta)$ in the norm of $\mathscr{M}(\varepsilon,\delta)$. Then, we define $\overset{\circ}{\mathscr{M}}{}'(\varepsilon,\delta)$ as the dual of $\overset{\circ}{\mathscr{M}}(\varepsilon,\delta)$.

We are now prepared to state the Calderón reproducing formula which implies the existence of dual wavelet operators.

Theorem 3.5. [Continuous Calderón reproducing formula] *Suppose $(\mathbb{R}^n,\rho,dx)$ is a normal space of homogeneous type and let $\{S_m\}_{m\in\mathbb{Z}}$ be an approximation to the identity of order (μ,δ,r) with respect to $\rho(\cdot,\cdot)$. Set $D_m := S_{m+1} - S_m$ for $m \in \mathbb{Z}$. Then there exist linear operators $\{\tilde{D}_m\}_{m\in\mathbb{Z}}$ and $\{\hat{D}_m\}_{m\in\mathbb{Z}}$ such that for any $f \in \mathscr{M}_0(\varepsilon,\gamma)$, $0 < \varepsilon,\gamma < \mu_0$,*

$$f = \sum_{m\in\mathbb{Z}} \tilde{D}_m D_m(f) = \sum_{m\in\mathbb{Z}} D_m \hat{D}_m(f), \tag{39}$$

where the series converge in the norm of $\mathscr{M}(\varepsilon',\gamma')$, $\varepsilon' < \varepsilon$, $\gamma' < \gamma$, and in $L_p(\mathbb{R}^n)$, $1 < p < \infty$. Furthermore, for any $\varepsilon < \mu_0$, the kernels of $\{\tilde{D}_m\}$ and $\{\hat{D}_m\}$ satisfy conditions (i)-(iii) of an approximation to the identity of order $(\mu,\varepsilon,1)$ (with constants depending on ε) and the r-th zero moments condition (37).

By a duality argument we obtain

Corollary 3.1. *Under the hypothesis of Theorem 3.5 for any $f \in \overset{\circ}{\mathscr{M}}{}'(\varepsilon,\delta)$ the series in (39) converges in $\overset{\circ}{\mathscr{M}}{}'(\varepsilon_*,\delta_*)$ with $\varepsilon < \varepsilon_* < \mu_0$, $\gamma < \gamma_* < \mu_0$.*

We next sketch the proof of Theorem 3.5. The method of proof is essentially similar to the method used in [19]. We use Coifman's idea to write the identity operator I as

$$I = \sum_k D_k = \sum_k D_k \sum_l D_l = \sum_{k,l} D_k D_l.$$

For an integer $N > 0$ we introduce the operator $D_m^N := \sum_{|j|\le N} D_{m+j}$ and define the operators T_N and R_N by

$$I = \sum_{k,l} D_k D_l = \sum_{k\in\mathbb{Z}} D_k^N D_k + \sum_{|j|>N} \sum_{k\in\mathbb{Z}} D_{k+j} D_k =: T_N + R_N.$$

Let $0 < \varepsilon,\gamma < \mu_0$. We claim that R_N is bounded on $\mathscr{M}_0(\varepsilon,\gamma,x_0,t)$ for any $x_0 \in \mathbb{R}^n$ and $t \in \mathbb{R}$. Moreover, there exist constants $\tau > 0$ and $c > 0$ such that

$$\|R_N f\|_{\mathscr{M}_0(\varepsilon,\gamma,x_0,t)} \le c2^{-N\tau}\|f\|_{\mathscr{M}_0(\varepsilon,\gamma,x_0,t)} \quad \text{for} \quad f \in \mathscr{M}_0(\varepsilon,\gamma,x_0,t). \tag{40}$$

Assume the claim for a moment. Choosing N so that $c2^{-N\tau} < 1$, then (40) implies that the operator T_N^{-1} exists and is bounded on $\mathscr{M}_0(\varepsilon,\gamma,x_0,t)$. Thus, we obtain

$$I = T_N^{-1}T_N = \sum_m \left(T_N^{-1}D_m^N\right) D_m = \sum_m \tilde{D}_m D_m,$$

where $\tilde{D}_m := T_N^{-1}D_m^N$. The regularity conditions on the kernels $\{D_m\}$ and (37) imply that for any fixed N and $y \in \mathbb{R}^n$ the function $D_m^N(\cdot,y)$ is in $\mathscr{M}_0(\mu_0,\delta)$. This immediately implies that $\tilde{D}_m(\cdot,y) = T_N^{-1}D_m^N(\cdot,y)$ is in $\mathscr{M}_0(\varepsilon,\gamma)$ for any $0 < \varepsilon,\gamma < \mu_0$. Similarly, we can write

$$I = T_N T_N^{-1} = (\sum_m D_m^N D_m)T_N^{-1} = \sum_m D_m D_m^N T_N^{-1} = \sum_m D_m \hat{D}_m,$$

where $\hat{D}_m := D_m^N T_N^{-1}$. By the same token, for any fixed N and $x \in \mathbb{R}^n$, the function $\hat{D}_m(x,\cdot)$ is in $\mathscr{M}_0(\varepsilon,\gamma)$ for any $0 < \varepsilon,\gamma < \mu_0$ and the proof is complete.

Discussion. In the proof of Theorem 3.5 we applied tools from the general theory of spaces of homogeneous type to construct dual wavelet operators. Although the kernels of the dual operators $\{\tilde{D}_m\}$ and $\{\hat{D}_m\}$ have the same vanishing moments as $\{D_m\}$, we only claim very "modest" regularity and decay on them. For example, in Theorem 3.5 we claim that for any $0 < \gamma < \mu_0$, there exists a constant $c > 0$ such that

$$|\tilde{D}_m(x,y)|, |\hat{D}_m(x,y)| \le \frac{c2^{-m\gamma}}{(2^{-m}+\rho(x,y))^{1+\gamma}}.$$

At the same time, the construction of the anisotropic approximation of the identity over an ellipsoid cover in §3.4 (see Theorem 3.4) produces wavelet kernels $\{D_m\}$ such that for any $\delta > 0$

$$|D_m(x,y)| \le \frac{c2^{-m\delta}}{(2^{-m}+\rho(x,y))^{1+\delta}}, \quad c = c(\delta).$$

It is an *open problem* to define higher order anisotropic test function spaces and prove that the operators $R_N := \sum_{|j|>N}\sum_{k\in\mathbb{Z}} D_{k+j}D_k$ are bounded on these higher order spaces as in (40).

Applying the Calderón reproducing formula we obtain the following Littlewood-Paley type result (see [16]).

Proposition 3.2. *Suppose $\{S_m\}_{m\in\mathbb{Z}}$ is an anisotropic approximation of the identity and let $D_m = S_{m+1} - S_m$, $m \in \mathbb{Z}$. Then for any $f \in L_p(\mathbb{R}^n)$, $1 < p < \infty$, we have*

$$\|f\|_p \sim \left\| \left(\sum_m |D_m(f)(\cdot)|^2 \right)^{1/2} \right\|_p.$$

3.6 Discrete wavelet frames

Here we describe briefly the construction of wavelet frames using the discrete Calderón reproducing formula, which in turn is obtained by "sampling" the continuous Calderón reproducing formula (see e.g. [16, 14]). We first introduce the following sampling process.

Definition 3.4. Let $\rho(\cdot,\cdot)$ be a quasi-distance on $\mathbb{R}^n$. We call a set of closed domains $\Omega_{m,k} \subset \mathbb{R}^n$, $m \in \mathbb{Z}$, $k \in I_m$, and points $y_{m,k} \in \Omega_{m,k}$, a *sampling set* if the following conditions are satisfied:

(a) For each $m \in \mathbb{Z}$, the sets $\Omega_{m,k}$, $k \in I_m$, have disjoint interiors.
(b) $\mathbb{R}^n = \cup_{k\in I_m} \Omega_{m,k}$ for $m \in \mathbb{Z}$.
(c) Each set $\Omega_{m,k}$ satisfies $\Omega_{m,k} \subset B_\rho(x_{m,k}, c2^{-m})$ for some point $x_{m,k} \in \mathbb{R}^n$ ($c > 0$ is a constant).
(d) There exists a constant $c' > 0$ such that for any $m \in \mathbb{Z}$ and $k \in I_m$, we have $\rho(y_{m,k}, y_{m,k'}) > c'2^{-m}$ for all $k' \in I_m$, $k' \neq k$, except perhaps for a set of uniformly bounded number of points.

In the next theorem we present the discrete Calderón reproducing formula.

Theorem 3.6. [14] *Let $\{S_m\}_{m\in\mathbb{Z}}$ be an anisotropic approximation to the identity of order (μ,δ,r) with respect to the quasi-distance induced by an ellipsoid cover Θ of $\mathbb{R}^n$. Denote $D_m := S_{m+1} - S_m$ and let $\{\Omega_{m,k}\}$ and $\{y_{m,k}\}$ with $y_{m,k} \in \Omega_{m,k}$ be a sampling set for Θ. Then there exist $N > 0$ and linear operators $\{\hat{E}_m\}$ such that for any $f \in \mathcal{M}_0(\varepsilon,\gamma)$, $0 < \varepsilon, \gamma < \mu_0$,*

$$f = \sum_{m\in\mathbb{Z}} \sum_{k\in I_{m+N}} |\Omega_{m+N,k}| \hat{E}_m(f)(y_{m+N,k}) D_m(\cdot, y_{m+N,k}), \tag{41}$$

where the convergence is in $\mathcal{M}(\varepsilon',\gamma')$, $\varepsilon' < \varepsilon$, $\gamma' < \gamma$, and in $L_p(\mathbb{R}^n)$, $1 < p < \infty$. Furthermore, the kernels of $\{\hat{E}_m\}$ satisfy conditions (i)-(iii) of anisotropic approximations to the identity of order $(\mu,\varepsilon,1)$ for any $\varepsilon < \mu_0$ (with constants depending on ε) and the rth degree zero moments condition (37).

The proof of this theorem follows in the footsteps of the proof in the general case of homogeneous spaces (see e.g. [16]).

Definition of anisotropic wavelet frames. We denote briefly $K_m := I_{m+N}$ and define the functions $\{\psi_{m,k}\}$ by

$$\psi_{m,k}(x) := |\Omega_{m+N,k}|^{1/2} D_m(x, y_{m+N,k})$$

and the functionals $\{\tilde{\psi}_{m,k}\}$ by

$$\tilde{\psi}_{m,k}(x) := |\Omega_{m+N,k}|^{1/2} \hat{E}_m(y_{m+N,k}, x), \quad m \in \mathbb{Z},\ k \in K_m.$$

Then (41) takes the form

$$f = \sum_m \sum_{k \in K_m} \langle f, \tilde{\psi}_{m,k} \rangle \psi_{m,k}. \tag{42}$$

The next theorem shows that $\{\psi_{m,k}\}$, $\{\tilde{\psi}_{m,k}\}$ is a pair of dual frames.

Theorem 3.7. [14] *Let $\{S_m\}_{m \in \mathbb{Z}}$ be an anisotropic approximation to the identity of order (μ, δ, r). Denote $D_m := S_{m+1} - S_m$ and let $\{\Omega_{m,k}\}$ and $\{y_{m,k}\}$, $y_{m,k} \in \Omega_{m,k}$ be a sampling set for Θ. Then there exist constants $0 < A \le B < \infty$ such that for any $f \in L_2(\mathbb{R}^n)$*

$$A\|f\|_2^2 \le \sum_m \sum_{k \in K_m} |\langle f, \tilde{\psi}_{m,k} \rangle|^2 \le B\|f\|_2^2. \tag{43}$$

3.7 Two-level-split frames

We now use the two-level-split construction from §3.3 and the smooth duals $\{\tilde{\tilde{g}}_\lambda\}$ from §3.4 to derive a useful representation for the wavelet kernels $D_m(x,y)$.

For $\lambda = (\theta, \beta)$ we denote $\tilde{\tilde{g}}_{\theta,\beta} := \tilde{\tilde{g}}_\lambda$, where $\tilde{\tilde{g}}_\lambda$ is defined in (34). Then the kernel $S_m(x,y)$, defined in (36), has the representation

$$S_m(x,y) = \sum_{\theta \in \Theta_m} \sum_{|\beta| < r} \tilde{\tilde{g}}_{\theta,\beta}(y) P_{\theta,\beta} \varphi_\theta(x).$$

Now precisely as in §3.3 we get

$$\begin{aligned} D_m(x,y) &:= S_{m+1}(x,y) - S_m(x,y) \\ &= \sum_{\eta \in \Theta_{m+1}} \sum_{\theta \in \Theta_m : \theta \cap \eta \neq \emptyset} \sum_{|\beta| < r} \Big\{ \tilde{\tilde{g}}_{\eta,\beta}(y) - \sum_{|\alpha| < r} C_{\alpha,\beta}^{\theta,\eta} \tilde{\tilde{g}}_{\theta,\alpha}(y) \Big\} P_{\eta,\beta}(x) \varphi_\eta(x) \varphi_\theta(x), \end{aligned}$$

The new dual functions $\tilde{\tilde{F}}_\nu$, $\nu = (\eta, \theta, \beta) \in \mathscr{V}_m$, are defined by

$$\tilde{\tilde{F}}_\nu = \tilde{\tilde{F}}_{\eta,\theta,\beta} := \tilde{\tilde{g}}_{\eta,\beta} - \sum_{|\alpha| < r} C_{\alpha,\beta}^{\theta,\eta} \tilde{\tilde{g}}_{\theta,\alpha}. \tag{44}$$

Thus we arrive at the following representation

$$D_m(x,y) = \sum_{\nu \in \mathscr{V}_m} \tilde{\tilde{F}}_\nu(y) F_\nu(x).$$

Observe that since each $\theta \in \Theta_m$ is intersected by finitely many ellipsoids from Θ_{m+1} it follows by (35) that the duals $\{\tilde{\tilde{F}}_\nu\}$ have sub-exponential localization as the duals $\{\tilde{\tilde{g}}_\lambda\}$. Also, Theorem 3.2 and Proposition 3.2 imply that $\{F_\nu\}$, $\{\tilde{\tilde{F}}_\nu\}$ is a pair of dual frames.

Proposition 3.3. *For any $f \in L_2(\mathbb{R}^n)$*

$$\|f\|_2 \sim \Big(\sum_m \|D_m(f)\|_2^2\Big)^{1/2} \sim \Big(\sum_\nu \langle f, \tilde{\tilde{F}}_\nu\rangle^2\Big)^{1/2}.$$

4 Anisotropic Besov spaces (B-spaces)

In this section we review the anisotropic Besov spaces of positive smoothness induced by discrete ellipsoid covers of $\mathbb{R}^n$, introduced in [12], and compare them with the B-spaces based on anisotropic nested triangulations of $\mathbb{R}^2$ from [13, 20]. We will be mainly interested in the homogeneous versions of these spaces.

4.1 B-spaces induced by anisotropic covers of $\mathbb{R}^n$

Assuming that Θ is discrete ellipsoid cover of $\mathbb{R}^n$ (see Definition 2.1) we will define the homogeneous B-spaces $\dot{B}^\alpha_{pq}(\Theta)$ of positive smoothness $\alpha > 0$. In this definition there is a hidden parameter k which we choose to be the smallest integer satisfying the condition

$$k > \frac{a_0}{a_6} \cdot \frac{\alpha}{n}. \tag{45}$$

This will guarantee the equivalence of the norms in $\dot{B}^\alpha_{pq}(\Theta)$ introduced below. Here a_0 and a_6 are the constants from Definition 2.1, §2.1.

Definition of $\dot{B}^\alpha_{pq}(\Theta)$ via local moduli of smoothness. For $\alpha > 0$ and $0 < p, q \le \infty$ the space $\dot{B}^\alpha_{pq}(\Theta)$ is defined as the set of all functions $f \in L^{\rm loc}_p$ such that

$$\|f\|_{\dot{B}^\alpha_{pq}(\Theta)} := \Big(\sum_{m\in\mathbb{Z}} \Big(\sum_{\theta\in\Theta_m} |\theta|^{-\alpha p/n}\, \omega_k(f,\theta)_p^p\Big)^{q/p}\Big)^{1/q} < \infty, \tag{46}$$

where $\omega_k(f,\theta)_p$ is the kth local modulus of smoothness of f (see (20)).

This definition needs some additional clarification. Observe that $\|P\|_{\dot{B}^\alpha_{pq}(\Theta)} = 0$ for $P \in \mathscr{P}_k$ and hence the norm in $\dot{B}^\alpha_{pq}(\Theta)$ is a semi-(quasi-)norm and $\dot{B}^\alpha_{pq}(\Theta)$ is a quotient space modulo $\mathscr{P}_k$. We will use the operators Q_m and $T_{m,p}$ from §3.2 to construct a meaningful representation of each $f \in \dot{B}^\alpha_{pq}(\Theta)$. Let T_m ($m \in \mathbb{Z}$) be one of the operators Q_m or $T_{m,p}$ if $p \ge 1$, and $T_m := T_{m,p}$ if $p < 1$. We define

$$\|f\|^T_{\dot{B}^\alpha_{pq}(\Theta)} := \Big(\sum_{m\in\mathbb{Z}} \Big(2^{a_0 m\alpha/n}\|(T_{m+1} - T_m)f\|_p\Big)^q\Big)^{1/q}. \tag{47}$$

Proposition 3.1 and property (c) of ellipsoid covers imply

$$\|f - T_m f\|_p \le c\Big(\sum_{\theta\in\Theta_m} \omega_k(f,\theta)_p^p\Big)^{1/p}$$

and since $\|(T_{m+1}-T_m)f\|_p \le c\|f-T_{m+1}f\|_p + c\|f-T_mf\|_p$, we get

$$\|f\|^T_{\dot B^\alpha_{pq}(\Theta)} \le c\|f\|_{\dot B^\alpha_{pq}(\Theta)}. \tag{48}$$

For more precise description of $\dot B^\alpha_{pq}(\Theta)$ we have to distinguish between two basic cases.

Case 1: $0<\alpha<n/p$ or $\alpha=n/p$ and $q\le 1$. Then as is shown in [12] for any $f\in \dot B^\alpha_{pq}(\Theta)$ there exists a polynomial $P\in\mathscr{P}_k$ such that

$$f=\sum_{m\in\mathbb{Z}}(T_{m+1}-T_m)f+P \quad \text{in } L_p(K) \tag{49}$$

for all compact sets $K\subset\mathbb{R}^n$.

Case 2: $\alpha>n/p$ or $\alpha=n/p$ and $q>1$. Now the space $\dot B^\alpha_{pq}(\Theta)$ can be viewed as the set of all regular tempered distributions f such that $\|f\|_{\dot B^\alpha_{pq}(\Theta)}<\infty$ and

$$f=\sum_{m\in\mathbb{Z}}(T_{m+1}-T_m)f,$$

where the convergence is in $\mathscr{S}'/\mathscr{P}_k$. This means that there exist polynomials $P\in\mathscr{P}_k$ and $P_m\in\mathscr{P}_k$, $m\in\mathbb{Z}$, such that

$$f=P+\lim_{j\to-\infty}\sum_{m=j}^{\infty}(T_{m+1}-T_m)f+P_m \quad \text{in} \quad \mathscr{S}'.$$

In addition, $\dot B^\alpha_{pq}(\Theta)$ is continuously embedded in $\mathscr{S}'$.

Other norms in $\dot B^\alpha_{pq}(\Theta)$. The good understanding of the B-spaces depends on having several equivalent norms in $\dot B^\alpha_{pq}(\Theta)$. Note that if $\{d_\nu(f)\}$ are defined from $(T_{m+1}-T_m)f=\sum_{\nu\in\mathscr{V}_m}d_\nu(f)F_\nu$, then using Theorem 3.2

$$\|f\|^T_{\dot B^\alpha_{pq}(\Theta)}\sim\Big(\sum_{m\in\mathbb{Z}}\Big(\sum_{\nu\in\mathscr{V}_m}(|\eta_\nu|^{-\alpha/n}\|d_\nu(f)F_\nu\|_p)^p\Big)^{q/p}\Big)^{1/q}. \tag{50}$$

Observe that the above equivalence holds if $d_\nu(f)$ are replaced by $\langle f,\tilde{\tilde F}_\nu\rangle$ due to the sub-exponential localization of the duals $\{\tilde{\tilde F}_\nu\}$.

Also, we define

$$\|f\|^A_{\dot B^\alpha_{pq}(\Theta)}:=\inf_{f=\sum_{\nu\in\mathscr{V}}a_\nu F_\nu}\Big(\sum_{m\in\mathbb{Z}}\Big(\sum_{\nu\in\mathscr{V}_m}(|\eta_\nu|^{-\alpha/n}\|a_\nu F_\nu\|_p)^p\Big)^{q/p}\Big)^{1/q}. \tag{51}$$

Here the infimum is taken over all representations $f=\sum_{\nu\in\mathscr{V}}a_\nu F_\nu$, where the convergence is to be understood as described in Cases 1-2 above.

In the next theorem we record the equivalence of the above norms.

Theorem 4.1. [12] *If $\alpha > 0$, $0 < p,q \le \infty$, and condition (45) is satisfied, then the norms $\|\cdot\|_{\dot{B}^\alpha_{pq}(\Theta)}$, $\|\cdot\|^T_{\dot{B}^\alpha_{pq}(\Theta)}$, and $\|\cdot\|^A_{\dot{B}^\alpha_{pq}(\Theta)}$ are equivalent.*

The embedding of $\dot{B}^\alpha_{pq}$ in $\mathscr{S}'$ or (49) readily imply the completeness of $\dot{B}^\alpha_{pq}(\Theta)$.

Inhomogeneous B-spaces. Sometimes it is more convenient to use the inhomogeneous versions $B^\alpha_{pq}(\Theta^+)$ of the B-spaces induced by anisotropic ellipsoid covers of $\mathbb{R}^n$, which are simpler than the homogeneous counterparts $\dot{B}^\alpha_{pq}(\Theta)$.

For the definition of the inhomogeneous spaces $B^\alpha_{pq}(\Theta^+)$ one only needs ellipsoid covers with levels $m = 0, 1, \dots$, i.e. covers of the form

$$\Theta^+ := \bigcup_{m=0}^{\infty} \Theta_m.$$

The space $B^\alpha_{pq}(\Theta^+)$, $\alpha > 0$, $0 < p,q \le \infty$, is defined as the set of all functions $f \in L_p(\mathbb{R}^n)$ such that

$$|f|_{B^\alpha_{pq}(\Theta^+)} := \Big(\sum_{m\ge 0} \Big(\sum_{\theta\in\Theta_m} (|\theta|^{-\alpha p/n}\, \omega_k(f,\theta)_p^p \Big)^{q/p} \Big)^{1/q} < \infty, \tag{52}$$

where $\omega_k(f,\theta)_p$ is the kth local modulus of smoothness of f in $L_p(\theta)$.

The (quasi-)norm in $B^\alpha_{pq}(\Theta^+)$ is defined by

$$\|f\|_{B^\alpha_{pq}(\Theta^+)} := \|f\|_p + |f|_{B^\alpha_{pq}(\Theta^+)}.$$

Other equivalent norms in $B^\alpha_{pq}(\Theta^+)$ can be defined similarly as for the homogeneous B-spaces from above. In particular, using the notation from from Theorem 3.3 one has

$$\|f\|_{B^\alpha_{pq}(\Theta^+)} \sim \Big(\sum_{m\ge -1} \Big(\sum_{\nu\in\mathscr{V}_m} (|\eta_\nu|^{-\alpha/n} \|d_\nu(f) F_\nu\|_p)^p \Big)^{q/p} \Big)^{1/q}. \tag{53}$$

For more details about anisotropic B-spaces induced by ellipsoid covers and proofs we refer the reader to [12].

4.2 B-spaces induced by nested multilevel triangulations of $\mathbb{R}^2$

We first recall briefly some basic definitions and facts from [20, 13].

Spline multiresolution analysis (MRA). Let $\mathscr{T} = \bigcup_{m\in\mathbb{Z}} \mathscr{T}_m$ be an SLR-triangulation of $\mathbb{R}^2$ (see §2.2). Denote by V_m the set of all vertices of triangles from $\mathscr{T}_m$.

For $r \ge 0$ and $k \ge 2$, we denote by $S^{k,r}_m = S^{k,r}(\mathscr{T}_m)$ the set of all r times differentiable piecewise polynomial functions of degree $< k$ over $\mathscr{T}_m$, i.e. $s \in S^{k,r}_m$ if $s \in C^r(\mathbb{R}^2)$ and $s = \sum_{\triangle\in\mathscr{T}_m} \mathbb{1}_\triangle \cdot P_\triangle$ with $P_\triangle \in \mathscr{P}_k$.

It will be convenient to denote, for any vertex $v \in V_m$, by $\mathrm{Star}^1(v)$ the union of all triangles $\triangle \in \mathscr{T}_m$ attached to v. Inductively for $\ell \geq 2$, we define $\mathrm{Star}^\ell(v)$ as the union of $\mathrm{Star}^{\ell-1}(v)$ and the stars of all vertices of $\mathrm{Star}^{\ell-1}(v)$.

We assume that for each $m \in \mathbb{Z}$ there exists a subspace S_m of $S_m^{k,r}$ and a family $\Phi_m = \{\varphi_\theta : \theta \in \Theta_m\} \subset S_m$ satisfying the following conditions:

(a) $S_m \subset S_{m+1}$ and $\mathscr{P}_{\tilde{k}} \subset S_m$, for some $1 \leq \tilde{k} \leq k$,

(b) Φ_m is a stable basis for S_m in L_p $(1 \leq p \leq \infty)$,

(c) For every $\theta \in \Theta_m$ there is a vertex $v_\theta \in V_m$ such that φ_θ and its dual are supported on $\mathrm{Star}^\ell(v_\theta)$, where $\ell \geq 1$ is a constant independent of θ and m.

We denote $\Phi := \bigcup_{m\in\mathbb{Z}} \Phi_m$ and $\Theta := \bigcup_{m\in\mathbb{Z}} \Theta_m$.

A simple example of spline MRA is the sequence $\{S_m\}_{m\in\mathbb{Z}}$ of all continuous piecewise linear functions ($r = 0$, $k = 2$) on the levels $\{\mathscr{T}_m\}_{m\in\mathbb{Z}}$ of a given SLR-triangulation $\mathscr{T}$ of $\mathbb{R}^2$. A basis for each space S_m is given by the set Φ_m of the Courant elements φ_θ, supported on the cells θ of $\mathscr{T}_m$ (θ is the union of all triangles of $\mathscr{T}_m$ attached to a vertex, say, v_θ). The function φ_θ takes value 1 at v_θ and 0 at all other vertices.

A concrete construction of a spline MRA for an arbitrary SLR-triangulation $\mathscr{T}$ is given in [13], where $S_m = S_m^{k,r} = S^{k,r}(\mathscr{T}_m)$ for given $r \geq 1$ and $k > 4r+1$.

Local spline approximation. For $\triangle \in \mathscr{T}_m$ we set

$$\Omega_\triangle^\ell := \cup\{\mathrm{Star}^\ell(v) : v \in V_m,\ \triangle \subset \mathrm{Star}^\ell(v)\}.$$

We now let $\mathbb{S}_\triangle(f)_p$ denote the error of L_p-approximation from S_m on $\Omega_\triangle^\ell$, i.e.

$$\mathbb{S}_\triangle(f)_p := \inf_{s\in S_m} \|f - s\|_{L_p(\Omega_\triangle^\ell)}. \tag{54}$$

Definition of $\dot{\mathscr{B}}_{pq}^\alpha(\Phi)$. Given a spline MRA $\{S_m\}_{m\in\mathbb{Z}}$ over an SLR-tiangulation $\mathscr{T}$ of $\mathbb{R}^2$ and an associated family of basis functions Φ, as described above, we define the B-space $\dot{\mathscr{B}}_{pq}^\alpha(\Phi)$, $\alpha > 0$, $0 < p, q \leq \infty$, as the set of all $f \in L_p^{\mathrm{loc}}(\mathbb{R}^2)$ such that

$$\|f\|_{\dot{\mathscr{B}}_{pq}^\alpha(\Phi)} := \left(\sum_{m\in\mathbb{Z}} \left[2^{m\alpha} \left(\sum_{\triangle\in\mathscr{T},\, 2^{-m}\leq|\triangle|<2^{-m+1}} \mathbb{S}_\triangle(f)_p^p \right)^{1/p} \right]^q \right)^{1/q} < \infty \tag{55}$$

with the ℓ_q-norm replaced by the sup-norm if $q = \infty$.

4.3 Comparison of different B-spaces and Besov spaces

The most substantial distinction between $\dot{B}_{pq}^\alpha(\Theta)$ and $\dot{\mathscr{B}}_{pq}^\alpha(\Phi)$ is that the spaces $\dot{B}_{pq}^\alpha(\Theta)$ are defined via *local polynomial* approximation $\sim \omega_k(f,\theta)_p$, while $\dot{\mathscr{B}}_{pq}^\alpha(\Phi)$ are defined via *local spline* approximation: $\mathbb{S}_\triangle(f)_p$. As a result, loosely speaking the spaces $\dot{B}_{pq}^\alpha(\Theta)$ have larger norms that the spaces $\dot{\mathscr{B}}_{pq}^\alpha(\Phi)$. However, if $\mathbb{S}_\triangle(f)_p$ in (55) is replaced by $\omega_k(f,\Omega_\triangle^1)_p$ then the resulting quantity would be equivalent to

$\|f\|_{\dot{B}^{\alpha}_{pq}(\Theta)}$, where Θ is the ellipse cover of $\mathbb{R}^2$ obtained by dilating the minimum area circumscribed ellipses for all triangle $\triangle \in \mathscr{T}$ as mentioned in §2.2.

Another important distinction between $\dot{B}^{\alpha}_{pq}(\Theta)$ and $\dot{\mathscr{B}}^{\alpha}_{pq}(\Phi)$ is that the underlying multilevel triangulation for the later space is nested, while the ellipsoid cover generating the former is not nested. Therefore, in constructing ellipsoid cover and dealing with B-spaces $\dot{B}^{\alpha}_{pq}(\Theta)$ one has much more freedom.

It is quite easy to show that (see [11]) if Θ is an ellipsoid cover of $\mathbb{R}^n$ consisting of Euclidean balls, then the B-spaces $\dot{B}^{\alpha}_{pq}(\Theta)$ are the same as the respective classical Besov spaces $\dot{B}^{\alpha}_{q}(L_p)$ (with equivalent norms). We maintain that local moduli of smoothness rather than global ones are more natural for the definition of anisotropic (and even classical) Besov spaces of positive smoothness since they more adequately reflect the nature of the spaces. For the theory of (classical) Besov spaces we refer the reader to [23, 26].

As already mentioned the powers A^j of a real $n \times n$ matrix A with eigenvalues λ obeying $|\lambda| > 1$ generate a semi-continuous and hence discrete ellipsoid cover of $\mathbb{R}^n$. It can be shown that for $\alpha > n(1/p - 1)_+$ the associated B-spaces $\dot{B}^{\alpha}_{pq}$ are exactly the same (with equivalent norms) as the anisotropic Besov spaces (with weight 1) developed in [3].

As indicated in §2.1, $\mathbb{R}^n$ equipped with the distance $\rho(\cdot,\cdot)$ introduced in Definition 2.2 and the Lebesgue measure is a space of homogeneous type and hence the general theory of Besov spaces on homogeneous spaces applies (see e.g. [19]). In fact, in the specific setting of this paper the anisotropic Besov spaces given by the general theory are the same as the B-spaces from here for sufficiently small $\alpha > 0$. The main distinction between the two theories is that we can handle B-spaces of an arbitrary smoothness $\alpha > 0$, while the general theory of Besov spaces on homogeneous spaces is only feasible for smoothness α with $|\alpha| < \varepsilon$ for some sufficiently small ε.

5 Nonlinear approximation

One of the main applications of the anisotropic B-spaces is to nonlinear N-term approximation from the two-level-split bases introduced in §3.3, which is the purpose of this section. We will also compare here the two-level-split bases with anisotropic hierarchical spline bases as tools for nonlinear approximation.

The B-spaces of nonlinear approximation. A particular type of B-spaces plays an important role in nonlinear N-term approximation in L_p. Given $0 < p < \infty$ and $\alpha > 0$ let τ be defined by

$$1/\tau = \alpha/n + 1/p, \tag{56}$$

which in the case of classical Besov spaces signifies the critical embedding in L_p. For nonlinear approximation in $L_\infty := C_0$ τ is determined by $1/\tau = \alpha/n$ and necessarily $\alpha \geq 1$ (otherwise the embedding (60) below is not valid).

For a given discrete ellipsoid cover Θ of $\mathbb{R}^n$, the homogeneous B-spaces $\dot{B}^\alpha_\tau(\Theta) := \dot{B}^\alpha_{\tau\tau}(\Theta)$ are of a particular importance in nonlinear approximation from the two-level-split bases. From (46) we have

$$\|f\|_{\dot{B}^\alpha_\tau(\Theta)} := \Big(\sum_{\theta\in\Theta} |\theta|^{-\alpha\tau/n} \omega_k(f,\theta)_\tau^\tau \Big)^{1/\tau}. \tag{57}$$

Observe that in general $\tau < 1$, however, just as in [20] it can be showen that for any $0 < q < p$

$$\|f\|_{\dot{B}^\alpha_\tau(\Theta)} \sim \Big(\sum_{\theta\in\Theta} |\theta|^{(1/p-1/q)\tau} \omega_k(f,\theta)_q^\tau \Big)^{1/\tau}. \tag{58}$$

This allows to work in L_q with $q \geq 1$ if $p > 1$ instead of L_τ.

The key point here is that the norm in $\dot{B}^\alpha_\tau(\Theta)$ has the representation

$$\|f\|_{\dot{B}^\alpha_\tau(\Theta)} \sim \Big(\sum_{\nu\in\mathscr{V}} \|d_\nu(f) F_\nu\|_p^\tau \Big)^{1/\tau}, \quad \mathscr{V} := \cup_{m\in\mathbb{Z}} \mathscr{V}_m, \tag{59}$$

which implies the embedding of $\dot{B}^\alpha_\tau(\Theta)$ in L_p: Every $f \in \dot{B}^\alpha_\tau(\Theta)$ can be identified modulo $\mathscr{P}_k$ as a function in $L_p(\mathbb{R}^n)$ such that

$$\|f\|_p \leq c \|f\|_{\dot{B}^\alpha_\tau(\Theta)}. \tag{60}$$

This identification will always be assumed in what follows. In fact, the above shows that $\dot{B}^\alpha_\tau(\Theta)$ lies on the Sobolev embedding line.

The situation is quite the same for the inhomogeneous B-spaces $B^\alpha_\tau := B^\alpha_{\tau\tau}(\Theta^+)$ associated with a discrete ellipsoid cover $\Theta^+ = \cup_{m\geq 0}\Theta_m$ of $\mathbb{R}^n$.

Nonlinear N-term approximation from $\mathscr{F} := \cup_{m\in\mathbb{Z}}\mathscr{F}_m = \{F_\nu : \nu \in \mathscr{V}\}$. We let $\mathscr{E}_N$ denote the nonlinear set of all functions g of the form

$$g = \sum_{\nu\in\Gamma_N} a_\nu F_\nu,$$

where $\Gamma_N \subset \mathscr{V}$, $\#\Gamma \leq N$, and Γ is allowed to vary with g. Then the error $\sigma_N(f)_p$ of best L_p-approximation of $f \in L_p(\mathbb{R}^n)$ from $\mathscr{E}_N$ is defined by

$$\sigma_N(f)_p := \inf_{g\in\mathscr{E}_N} \|f - g\|_p.$$

Theorem 5.1 (Jackson estimate). *If $f \in \dot{B}^\alpha_\tau(\Theta)$, $\alpha > 0$, $0 < p \leq \infty$, then*

$$\sigma_N(f)_p \leq c N^{-\alpha/n} \|f\|_{\dot{B}^\alpha_\tau(\Theta)}, \tag{61}$$

where c depends only on α, p, and the parameters of Θ.

When $0 < p < \infty$, estimate (61) follows by the general Theorem 3.4 in [20] and in the case $p = \infty$ its proof can be carried out as the proof of Theorem 3.1 in [21].

In a standard way the Jackson estimate (61) leads to a direct estimate for nonlinear N-term approximation from $\mathscr{F}$ which involves the K-functional between L_p and $\dot{B}^\alpha_\tau(\Theta)$. It is a challenging *open problem* to prove a companion inverse estimate due to the fact that $\mathscr{F}$ is possibly redundant and nonnested.

Comparison with nonlinear N-term approximation from nested spline bases. Nonlinear N-term approximation in L_p $(0 < p \le \infty)$ from the spline basis elements in $\Phi = \cup_{m\in\mathbb{Z}}\Phi_m$ (see §4.2) has been developed in [20, 13, 21, 10]. In [20, 13] Jackson and Bernstein estimates are established involving the B-spaces $\dot{\mathscr{B}}^\alpha_\tau(\Phi) := \dot{\mathscr{B}}^\alpha_{\tau\tau}(\Phi)$ with norm

$$\|f\|_{\dot{\mathscr{B}}^\alpha_\tau(\Phi)} := \Big(\sum_{\triangle\in\mathscr{T}} (|\triangle|^{-\alpha}\mathbb{S}_\triangle(f)_\tau)^\tau\Big)^{1/\tau}, \tag{62}$$

where $1/\tau := \alpha + 1/p$ for $\alpha > 0$ if $0 < p < \infty$ and $\alpha \ge 1$ if $p = \infty$. Then the standard machinery of Approximation theory is used to characterize the respective approximation spaces as real interpolation spaces between L_p and $\dot{\mathscr{B}}^\alpha_\tau(\Phi)$.

The most important difference between the nonlinear N-term approximation from $\mathscr{F}$ and Φ is that the spaces $\dot{\mathscr{B}}^\alpha_\tau(\Phi)$ (defined by local spline approximation) are specifically designed for the purposes of nonlinear spline approximation and allow to characterize the rates of approximation $O(N^{-\beta})$ for all $\beta > 0$, while in the former case β is limited. On the other hand, the spaces $\dot{B}^\alpha_\tau(\Theta)$ are of more general nature and are direct generalization of Besov spaces. They are much less sensitive to changes in the underlying ellipsoid cover Θ compared to changes in $\dot{\mathscr{B}}^\alpha_\tau(\Phi)$ when changing the respective triangulation $\mathscr{T}$. In general, the spaces $\dot{B}^\alpha_\tau(\Theta)$ are better than $\dot{\mathscr{B}}^\alpha_\tau(\Phi)$ as a tool for measuring the anisotropic features of functions (see below).

6 Measuring smoothness via anisotropic B-spaces

It has always been a question in analysis how to measure the smoothness of a given function, and as a consequence, there is a variety of smoothness space. We next show how the anisotropic B-spaces $\dot{B}^\alpha_\tau(\Theta)$ can be deployed to measuring the smoothness of functions and how this is related to nonlinear N-term approximation from the two-level-split bases.

We focus on two "simple" examples of discontinuous functions on $\mathbb{R}^2$, namely, $\mathbb{1}_{B(0,1)}$ the characteristic function of the unit disk $B(0,1)$ and $\mathbb{1}_Q$ the characteristic function of a square $Q \subset \mathbb{R}^2$. As shown in [12] each of these functions has higher order smoothness α in $\dot{B}^\alpha_\tau(\Theta)$ for an appropriately selected ellipse cover Θ compared with its (classical) Besov space smoothness. Moreover, their smoothness via suitable covers will be seen to differ substantially.

As in the previous section, for given $0 < p < \infty$ and $\alpha > 0$, let τ be defined by $1/\tau = \alpha/2 + 1/p$.

Theorem 6.1. [12] (i) *There exists an anisotropic ellipsoid cover* Θ *of* $\mathbb{R}^2$ *such that* $\mathbb{1}_{B(0,1)} \in \dot{B}^{\alpha}_{\tau}(\Theta)$ *for any* $\alpha < 4/p$. *In comparison, in the scale of Besov spaces* $\dot{B}^{\alpha}_{\tau\tau}$ *one has* $\mathbb{1}_{B(0,1)} \in \dot{B}^{\alpha}_{\tau\tau}$ *for* $\alpha < 2/p$. *Here the bounds for* α *are sharp.*

(ii) *For any square* Q *in* $\mathbb{R}^2$ *and any* $\alpha > 0$ *there exists an anisotropic ellipsoid cover* Θ *of* $\mathbb{R}^2$ *such that* $\mathbb{1}_Q \in \dot{B}^{\alpha}_{\tau}(\Theta)$, *while in the scale of Besov spaces* $\dot{B}^{\alpha}_{\tau\tau}$ *one has only* $\mathbb{1}_Q \in \dot{B}^{\alpha}_{\tau\tau}$ *for* $\alpha < 2/p$ *and this bound for* α *is sharp.*

This theorem coupled with the Jackson estimate (61) leads to the following approximation result.

Corollary 6.1. [12] (i) *There exists a discrete ellipse cover* Θ *of* $\mathbb{R}^2$ *such that for any* $0 < p < \infty$ *the nonlinear N-term approximation from* $\mathscr{F}_\Theta$ *satisfies*

$$\sigma_N(\mathbb{1}_{B(0,1)})_p \le cN^{-\gamma} \quad \text{for all} \quad \gamma < 2/p.$$

(ii) *For any* $\alpha > 0$ *there exists a discrete ellipse cover* Θ *of* $\mathbb{R}^2$ *such that for any* $0 < p < \infty$ *the nonlinear N-term approximation from* $\mathscr{F}_\Theta$ *satisfies*

$$\sigma_N(\mathbb{1}_Q)_p \le cN^{-\alpha}.$$

For comparison, if $\sigma^W_m(f)_p$ *denotes the best N-term approximation of* f *in* L_p $(p \ge 1)$ *from any reasonable wavelet basis, then for* $E = B(0,1)$ *or* $E = Q$

$$\sigma^W_N(\mathbb{1}_E)_p \le cN^{-\gamma} \quad \text{for all} \quad \gamma < 1/p.$$

All estimates above are sharp.

Discussion. As indicated above for appropriate ellipse covers, the B-space smoothness of the characteristic functions of the unit ball and any square in $\mathbb{R}^2$ is higher than their Besov space smoothness. Thus by using adaptive dilations the anisotropic B-spaces are better able to resolve the singularities along smooth or piecewise smooth curves. Consequently, the two-level-split decompositions of these functions are substantially sparser than their wavelet decompositions, which leads to better rates of nonlinear N-term approximation. It might surprise that characteristic functions of polygonal domains have, in a sense, infinite smoothness while those of domains with smooth boundaries have limited regularity. However, the covers that yield higher and higher smoothness in the polygonal case have to become less and less constrained, which means that the parameters in $\mathbf{p}(\Theta)$ are subjected to more and more generous bounds. Keeping these parameters within a compact set would limit the regularity that could be described in this way.

The above two examples illustrate clearly the concept of measuring the smoothness of functions via anisotropic B-space and in particular by the B-spaces of nonlinear approximation $\dot{B}^{\alpha}_{\tau}(\Theta)$. The key idea is to allow the underlying ellipsoid cover to adapt to the given function.

It is a challenging *open problem* to devise a scheme which for a given function f finds an optimal (or near optimal) ellipsoid cover Θ such that f exhibits the highest order α of smoothness in $\dot{B}^{\alpha}_{\tau}(\Theta)$ in the above sense.

7 Application to preconditioning for elliptic boundary value problems

In this section we apply the two-level-split bases from §3.3 in a regular set-up to the development of multilevel Schwarz preconditioners for elliptic boundary value problems. We consider the following model problem. Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a symmetric bilinear form on a Hilbert space V with norm $\|\cdot\|_V = \langle\cdot,\cdot\rangle^{1/2}$ that is V-elliptic, i.e. there exist positive constants c_a, C_a such that

$$a(v,v) \geq c_a\|v\|_V^2, \quad |a(v,w)| \leq C_a\|v\|_V\|w\|_V, \quad v,w \in V. \tag{63}$$

The problem is, for a given $f \in V'$ to find $u \in V$ such that

$$a(u,v) = \langle f,v\rangle, \quad \forall\ v \in V. \tag{64}$$

For simplicity we only consider the model case $V = H_0^1(\Omega)$ corresponding to Dirichlet boundary conditions. Higher order problems could be treated in an analogous way. We assume that Ω is a bounded *extension* domain, which means that Ω has a sufficiently regular boundary to permit any element v of any Sobolev or Besov space $X(\Omega)$ over Ω to be extended to $\tilde{v} \in X(\mathbb{R}^n)$, $\tilde{v}|_\Omega = v$, so that $\|v\|_{X(\mathbb{R}^n)} \leq C_X\|v\|_{X(\Omega)}$. This is e.g. the case when the boundary of Ω is piecewise smooth and Ω obeys a uniform cone condition. The homogeneous boundary conditions are supposed to be realized in the trial spaces by suitable polynomial factors in the atoms.

We assume that $\Theta = \cup_{m\geq -1}\Theta_m$ is a regular multilevel cover of $\mathbb{R}^n$ consisting of balls. We will utilize the atoms $\{F_\gamma\}$ defined in §3.3 for $\gamma \in \mathscr{V} = \bigcup_{m=-1}^\infty \mathscr{V}_m$, see Theorem 3.3. For better notation we will index the elements F_γ of the two-level-split bases $\mathscr{F}_m$ by γ instead of ν as before.

We will put this in the context of *stable splittings* in the theory of *multilevel Schwarz preconditioners*, see e.g. [22, 27].

Let $V_\gamma := \operatorname{span}(F_\gamma)$, so that $H_0^1(\Omega) := V = \sum_\gamma V_\gamma$. The key fact is that $\{V_\gamma\}_{\gamma\in\mathscr{V}}$ form a *stable splitting* for V:

Theorem 7.1. *There exist constants $c_V, C_V > 0$ such that for any $v \in V$*

$$c_V\|v\|_V \leq \inf_{v=\sum_\gamma v_\gamma}\left(\sum_{\gamma\in\mathscr{V}}|\eta_\gamma|^{-2/d}\|v_\gamma\|_2^2\right)^{1/2} \leq C_V\|v\|_V. \tag{65}$$

Moreover, $\{V_\gamma\}_{\gamma\in\mathscr{V}^\ell}$ with $\mathscr{V}^\ell := \bigcup_{m=-1}^\ell \mathscr{V}_m$ form a uniformly stable splitting for the spaces $S_m := \operatorname{span}(\Phi_m)$ in the sense of (65) with the same constants c_V, C_V.

Using that the norms $a(\cdot,\cdot)^{1/2}$ and $\|\cdot\|_{H^1(\Omega)}$ are equivalent and the well known fact that $\|\cdot\|_{H^1(\Omega)} \sim \|\cdot\|_{B_2^1(L_2(\Omega))}$, estimates (65) are immediate from Theorem 4.1 taking into account that Besov and B-norms are equivalent in the regular setting. The second part of Theorem 7.1 follows from the fact that the telescoping expansions

underlying the inhomogeneous version of $\|\cdot\|^T_{B^\alpha(\Theta)}$ (see (47) and (53)) terminate without affecting this norm. For more details, see [11].

This allows us to apply the theory of Schwarz methods along the following lines. For $V_0 := S_0 = \operatorname{span}(\Phi_0)$ define $P_{V_0} : V \to V_0$ and $r_{V_0} \in S_0$ by

$$a(P_{V_0}v, F_\gamma) = a(v, F_\gamma), \quad (r_{V_0}, F_\gamma)_{L_2} = \langle f, F_\gamma\rangle, \quad \gamma \in \mathscr{V}_0 = \Theta_0.$$

Furthermore, we introduce the auxiliary bilinear forms:

$$b_\gamma(v,w) := |\eta_\gamma|^{-2/d}(v,w)_{L_2}, \quad v,w \in V_\gamma, \quad \gamma \in \mathscr{V} \setminus \mathscr{V}_0. \tag{66}$$

We now consider the spaces V_γ with norms $\|v\|_{V_\gamma} := (b_\gamma(v,v))^{1/2}$ and define the linear operators $P_{V_\gamma} : V \to V_\gamma$ and $f_\gamma \in V_\gamma$ by

$$\begin{aligned} |\eta_\gamma|^{-2/d}(P_{V_\gamma}v, F_\gamma)_{L_2} &= a(v, F_\gamma), \\ |\eta_\gamma|^{-2/d}(f_\gamma, F_\gamma)_{L_2} &= \langle f, F_\gamma\rangle. \end{aligned} \tag{67}$$

Thus, as usual,

$$P_{V_\gamma}v = a_\gamma(v)F_\gamma, \qquad f_\gamma = r_\gamma(f)F_\gamma, \tag{68}$$

where

$$a_\gamma(v) = \frac{|\eta_\gamma|^{2/d}a(v,F_\gamma)}{\langle F_\gamma, F_\gamma\rangle}, \quad r_\gamma(f) = \frac{|\eta_\gamma|^{2/d}\langle f, F_\gamma\rangle}{\langle F_\gamma, F_\gamma\rangle}. \tag{69}$$

The following theorem now is an immediate consequence of the results in [18, 22].

Theorem 7.2. *Problem (64) is equivalent to the operator equation*

$$P_V u = \bar{f}, \quad \textit{where} \tag{70}$$

$$P_V := P_{V_0} + \sum_{\gamma\in\mathscr{V}\setminus\mathscr{V}_0} P_{V_\gamma}, \quad \bar{f} := r_{V_0} + \sum_{\gamma\in\mathscr{V}\setminus\mathscr{V}_0} f_\gamma.$$

Moreover, the spectral condition number $\kappa(P_V)$ of the additive Schwarz operator P_V satisfies

$$\kappa(P_V) \le \frac{C_a C_V}{c_a c_V}, \tag{71}$$

where c_a, C_a, c_V, C_V are the constants from (63) and (65).

Estimate (71) yields that simple iterative schemes, such as Richardson iterations,

$$u^{n+1} = u^n + \alpha(\bar{f} - P_V u^n), \quad n = 0,1,2,\ldots, \tag{72}$$

converge with a fixed error reduction rate per step.

We conclude with a few remarks. First, the operator equation (70) is formulated in the full infinite dimensional space. Alternatively, restricting the summation to a finite subset $\bar{\mathscr{V}}$ of $\mathscr{V}$ (e.g. $\bar{\mathscr{V}} = \mathscr{V}^\ell$), we obtain a finite dimensional discrete problem whose condition fulfills (on account of Theorem 7.1) the same bound uniformly in the size and choice of $\bar{\mathscr{V}}$. In this sense our preconditioner is asymptotically optimal.

On the other hand, it is conceptually useful to consider the full infinite dimensional problem (70). Then (72) has to be understood as an *idealized* scheme whose numerical implementation requires appropriate *approximate* applications of the (infinite dimensional) operator P_V quite in the spirit of [7]. This can be done by computing in addition to solving the coarse scale problem on $S_0 = V_0$ only finitely many but properly selected components P_{V_γ} each requiring only the solution of a one-dimensional problem. This hints at the adaptive potential of such an approach similar to the developments in [7]. This, in turn, raises the question what accuracy can be achieved at best when using linear combinations of at most N of the atoms. Thus we arrive at the problem for nonlinear N-term approximation from $\{F_\gamma\}$ in H^1.

For more details we refer the reader to [11].

Acknowledgements The second author has been supported by NSF Grant DMS-0709046.

References

1. Alani, A., Averbuch, A., Dekel, S.: Image coding using geometric wavelets, IEEE Trans. Image Process. **16**, 69–77 (2007)
2. Bownik, M.: Anisotrpopic Hardy spaces and wavelets. Mem. Amer. Math. Soc. **164**, No 781 (2003)
3. Bownik, M.: Atomic and molecular decompositions of anisotrpopic Besov spaces. Math. Z. **250**, 539–571 (2005)
4. Bownik, M., Ho, K.-P.: Atomic and molecular decompositions of anisotrpopic Triebel-Lizorkin spaces. Trans. Amer. Math. Soc. **358**, 1469–1510 (2006)
5. Candès, E., Demanet, L., Donoho, D., Ying, L.: Fast Discrete Curvelet Transforms, Multiscale Model. Simul. **5**, 861–899 (2006)
6. Candès, E., Donoho, D.: Continuous Curvelet Transform: I. Resolution of the Wavefront Set, Appl. Comput. Harmon. Anal. **19**, 162–197 (2005)
7. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods II - Beyond the elliptic case. Found. Comput. Math. **2**, 203–245 (2002)
8. Coifman, R., Weiss, G.: Analyse harmonique non-comutative sur certains espaces homogenes. Lecture Notes in Math. **242**, Springer-Verlag (1971)
9. Coifman, R., Weiss, G.: Extensions of Hardy spaces and their use in analysis. Bull. Amer. Math. Soc. **83**, 569–645 (1977)
10. Dahmen, D., Petrushev, P.: "Push-the-Error" algorithm for nonlinear n-term approximation. Constr. Approx. **23**, 261–304 (2006)
11. Dahmen, W., Dekel, S., Petrushev, P.: Multilevel Preconditioning for Partition of Unity Methods - Some Analytic Concepts. Numer. Math. **107**, 503–532 (2007)
12. Dahmen, W., Dekel, S., Petrushev, P.: Two-level-split Decomposition of Anisotropic Besov Spaces. . Constr. Approx. (to appear)
13. Davydov, O., Petrushev, P.: Nonlinear approximation from differentiable piecewise polynomials. SIAM J. Math. Anal. **35**, 708–758 (2003)
14. Dekel, D., Han, Y., Petrushev, P.: Anisotropic meshless frames on $\mathbb{R}^n$. J. Fourier Anal. and Appl. (to appear)
15. Dekel, D., Leviatan, D.: Adaptive multivariate approximation using binary space partitions and geometric wavelets, SIAM J. Numer. Anal. **43**, 707–732 (2005)
16. Deng, D., Han, Y.: Harmonic analysis on spaces of homogeneous type. Lecture Notes in Math. **1966** (2009)

17. Folland, G., Stein, E.: Hardy spaces on homogeneous groups. Princeten University Press, N. J. (1982)
18. Griebel, M., Oswald, P.: Remarks on the abstract theory of additive and multiplicative Schwarz methods. Numer. Math. **70**, 163–180 (1995)
19. Han, H., Sawyer, E.: Littlewood-Paley theory on spaces of homogeneous type and classical function spaces. Mem. Amer. Math. Soc. **530** (1994)
20. Karaivanov, K., Petrushev, P.: Nonlinear piecewise polynomial approximation beyond Besov spaces. Appl. Comput. Harmon. Anal. **15**, 177–223 (2003)
21. Karaivanov, K., Petrushev, P., Sharpley, R.C.: Algorithms for nonlinear piecewise polynomial approximation. Trans. Amer. Math. Soc. **355**, 2585–2631 (2003)
22. Oswald, O.: Multilevel Finite Element Approximation. Teubner Skripten zur Numerik, Teubner (1994)
23. Peetre, J.: New thoughts on Besov spaces. Duke Univ. Math. Series, Duke Univ. Durham (1976)
24. Petrushev, P.: Anisotropic spaces and nonlinear n-term spline approximation. Approximation Theory XI: Gatlinburg 2004, 363–394, Mod. Methods Math., Nashboro Press, Brentwood, TN (2005)
25. Stein, E.: Harmonic Analysis: Real-variable methods, Orthogonality and oscillatory integrals. Princeton University Press (1993)
26. Triebel, H.: Theory of Function Spaces. Monographs in Math. vol. 78, Birkhäuser (1983)
27. Xu, J.: Iterative methods by space decomposition and subspace correction, SIAM Review **34**, 581–613 (1992)

Nonlinear approximation and its applications

Ronald A. DeVore

Abstract I first met Wolfgang Dahmen in 1974 in Oberwolfach. He looked like a high school student to me but he impressed everyone with his talk on whether polynomial operators could produce both polynomial and spectral orders of approximation. We became the best of friends and frequent collaborators. While Wolfgang's mathematical contributions spread across many disciplines, a major thread in his work has been the exploitation of nonlinear approximation. This article will reflect on Wolfgang's pervasive contributions to the development of nonlinear approximation and its application. Since many of the contributions in this volume will address specific application areas in some details, my thoughts on these will be to a large extent anecdotal.

1 The early years

I was first exposed to approximation theory in a class taught by Ranko Bojanic in the Fall of 1964 at Ohio State University. Each student was allowed one optional class (outside of the required algebra and analysis). I do not know why I chose this from among the other options - perhaps another student had recommended it to me as a well structured interesting class - but I was immediately hooked. It just seemed like a natural subject answering natural questions. If we cannot explicitly solve most real world problems then we better learn how to approximate them.

The course was more on the theory than on the computational side since the demand for fast computational algorithms did not yet seem as urgent. There were no wavelets and splines were in their infancy. But there was plenty to intrigue the student including the Jackson-Bernstein theory of polynomial approximation which remains to this day as the prototype for understanding the quantitative side of ap-

Ronald A. DeVore
Texas A& M University, College Station, TX, USA, e-mail: ronald.a.devore@gmail.com

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_6,

proximation. Let us describe the modern form of this theory since it will be useful as we continue this exposition.

Suppose that we are interested in approximating the elements from a space X equipped with a norm $\|\cdot\| := \|\cdot\|_X$ by using the elements of the spaces X_n, $n = 1,2,\ldots$. Typical examples are $X = L_p$ or a Sobolev space while the usual suspects for X_n are spaces of polynomials, splines, or rational functions. We assume that for all $n,m \geq 1$, we have

$$X_n + X_m \subset X_{c(n+m)}, \text{ for some fixed } c \geq 1, \tag{1}$$

which is certainly the case for the above examples. Given $f \in X$, we define

$$E_n(f) := \inf_{g \in X_n} \|f - g\|. \tag{2}$$

The main challenge in the quantitative arena of approximation is to describe precisely the elements of X which have a prescribed order of approximation. Special attention is given to the approximation orders which are of the form n^{-r} since these occur most often in numerical computation. This gives the *primary approximation spaces* $\mathscr{A}^r := \mathscr{A}^r(X,(X_n))$, $r > 0$, consisting of all $f \in X$ for which

$$|f|_{\mathscr{A}^r} := \sup_{n \geq 1} n^r E_n(f) \tag{3}$$

is finite. The left side of (3) serves to define a semi-norm on $\mathscr{A}^r$. We obtain the norm for this space by adding $\|f\|_X$ to the semi-norm.

While the spaces $\mathscr{A}^r$ are sufficient to understand most approximation methods, it is sometimes necessary to go to a finer scale of spaces when dealing with nonlinear approximation. Accordingly, if $q > 0$, we define $\mathscr{A}^r_q$ via the quasi-norm

$$|f|_{\mathscr{A}^r_q(X)} := \|(2^{kr} E_{2^k}(f))\|_{\ell_q}. \tag{4}$$

Again, we obtain the norm for this space by adding $\|f\|_X$ to the semi-norm. When $q = \infty$, we obtain the spaces $\mathscr{A}^r$ because of (1).

The problem of characterizing $\mathscr{A}^r$ was treated in the following way for the case when $X = C[-\pi,\pi]$ is the space of continuous 2π periodic functions and X_n is the space of trigonometric polynomials of degree $\leq n$. One proves two fundamental inequalities for trigonometric approximation. The first of these is the following inequality proved by D. Jackson:

$$E_n(f) \leq C_k \|f^{(k)}\|_{C[-\pi,\pi]} n^{-k}, \quad n,k = 1,2,\ldots. \tag{5}$$

A companion to this is the famous Bernstein inequality which says

$$\|T^{(k)}\|_{C[-\pi,\pi]} \leq n^k \|T\|_{C[-\pi,\pi]}, \quad n,k = 1,2,\ldots \tag{6}$$

From these two fundamental inequalities, one can show that $\mathscr{A}^r$ is the generalized Lipschitz space Lip r space (defined later in §2) and more generally the $\mathscr{A}^r_q$ are

the same as the Besov spaces $B_q^r(L_\infty)$ which are also discussed in §2. It is interesting to note that the modern way of deriving such characterizations is not much different than the classical approach for trigonometric polynomials except that everything is now encasted in the general framework of interpolations spaces. This leads to the following theory.

Suppose for our approximation setting, we can find a space Y_k such that the following generalized Jackson and Bernstein inequalities hold

$$E_n(f)_X \le C_k \|f\|_{Y_k} n^{-k}, \quad n = 1,2,\dots. \tag{7}$$

and

$$\|S\|_{Y_k} \le C_k n^k \|S\|_X, \quad S \in X_n, \quad n = 1,2,\dots. \tag{8}$$

Then for any $0 < r < k$ and $0 < q \le \infty$, we have

$$\mathscr{A}_q^r(X,(X_n)) = (X,Y_k)_{\theta,q}, \quad \theta := r/k, \tag{9}$$

where the spaces on the right are the interpolation spaces given by the real method of interpolation (K-functionals) as described in the next section. In our case of trigonometric polynomial approximation the space Y_k is C^k with its usual semi-norm. It is well known that the interpolation spaces between C and C^k are the Besov spaces and in particular the generalized Lipschitz spaces when $q = \infty$.

The beauty of the above theory is that it boils down the problem of characterizing the approximation spaces for a given method of approximation to one of proving two inequalities: the generalized Jackson and Bernstein inequalities for the given approximation process. This recipe has been followed many times. An interesting question is whether the characterization (9) provides essential new information. That this is indeed the case rests on the fact that these interpolation spaces can be given a concrete description for most pairs (X,Y_k) of interest. This fact will be discussed next.

2 Smoothness and interpolation spaces

We all learn early on that the more derivatives a function has then the smoother it is. This is the coarse idea of smoothness spaces. Modern analysis carries this idea extensively forward by introducing a myriad of spaces to delineate properties of functions. We will touch on this with a very broad stroke only to communicate the heuerestic idea behind the smoothness spaces we shall need for describing rates of approximation.

For an integer $s > 0$, the Sobolev space $W^s(L_p(\Omega))$, on a domain $\Omega \subset \mathbb{R}^d$ consists of all $f \in L_p(\Omega)$ for which all of the distributional derivatives $D^\nu f$ of order s are also in $L_p(\Omega)$. This space is equipped with the semi-norm

$$|f|_{W^s(L_p(\Omega))} := \max_{|\nu|=s} \|D^\nu f\|_{L_p(\Omega)}. \tag{10}$$

We obtain the norm on $W^s(L_p(\Omega))$ by adding $\|\cdot\|_{L_p(\Omega)}$ to this semi-norm.

It is of great interest to extend this definition to all $s > 0$. One can initiate such an extension from many viewpoints. But the most robust of these approaches is to replace derivatives by differences. Suppose that we wish to define fractional order smoothness spaces on $\mathbb{R}^d$. The translation operator T_h for $h \in \mathbb{R}^d$ is defined on a function f by $T_h(f) := f(\cdot + h)$ and leads to the difference operators

$$\Delta_h^r := \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} T_{kh}. \tag{11}$$

If we apply Δ_h^r to a smooth function f then $h^{-r}\Delta_h^r(f)(x) \to r! f^{(r)}(x)$ as $h \to 0$. We can obtain smoothness spaces in L_p by placing conditions on how fast $\|\Delta_h^r(f)\|_{L_p}$ tends to zero as $h \to 0$. To measure this we introduce the *moduli of smoothness*

$$\omega_r(f,t)_p := \sup_{|h| \le t} \|\Delta_h^r(f)\|_{L_p(\Omega_{rh})}, \tag{12}$$

where Ω_t consists of all $x \in \Omega$ for which the line segment $[x, x+t]$ is contained in Ω.

We get a variety of spaces by placing decay conditions on $\omega_r(f,t)_p$ as $t \to 0$. The most classical of these are the generalized Lipschitz spaces $\mathrm{Lip}\,\alpha := \mathrm{Lip}(\alpha, L_p)$ in L_p which consist of all f for which

$$|f|_{\mathrm{Lip}(\alpha,L_p)} := \sup_{t>0} t^{-\alpha} \omega_r(f,t)_p, \quad \alpha < r, \tag{13}$$

is finite. We obtain the norm on this space by adding $\|f\|_{L_p}$ to (13). The above definition holds for all $0 < p \le \infty$. We usually make the convention that L_∞ is replaced by the space of continuous functions. Note that the above definition apparently depends on r but it is easy to show that one obtains exactly the same spaces no matter which r one choses (as long as $r > \alpha$) and the (quasi-)seminorms (13) are equivalent.

The generalized Lipschitz spaces are fine for a good understanding of approximation. However, certain subtle questions require a finer scaling of spaces provided by the Besov scale. Now, in addition to α we introduce a second fine scale parameter $q \in [0, \infty)$. Then the Besov space $B_q^\alpha(L_p)$ is defined by its semi-norm

$$|f|_{B_q^\alpha(L_p)} := \{ \int_{t>0} [t^{-\alpha} \omega_r(f,t)_p]^q \frac{dt}{t} \}^{1/q}, \quad \alpha < r. \tag{14}$$

2.1 The role of interpolation

We have already noted that approximation spaces can be characterized as interpolation spaces provided the fundamental Bernstein and Jackson type inequalities have been proven. For this characterization to be of use, we need to be able to describe

these interpolation spaces. Although this is not always simple, it has been carried out for all pairs of spaces that arise in linear and nonlinear approximation. To describe these results we will make a very brief incursion into interpolation.

The subject of operator interpolation grew out of harmonic analysis in the quest to have a unified approach to characterizing the mapping properties of its primary operators such as Fourier transforms, conjugate operators, maximal functions and singular integrals. Of primary interest in approximation theory are the real interpolation spaces. Given a pair of normed linear spaces X, Y which are both embedded in a common topological space, we can define the K-functional

$$K(f,t) := K(f,t;X,Y) := \inf_{f=f_0+f_1} \{\|f_0\|_X + t\|f_1\|_Y\}. \tag{15}$$

Often, the norm on Y is replaced by a semi-norm as is the case below when considering Y as a Sobolev space. The real interpolation spaces $(X,Y)_{\theta,q}$ are now defined for any $\theta \in (0,1)$ and $q > 0$ by the quasi-norm

$$\|f\|_{(X,Y)_{\theta,q}} := \|t^{-\theta} K(f,t)\|_{L_q(\mu)}, \tag{16}$$

where $\mu(t) := \frac{dt}{t}$ is Haar measure. By this time the reader is sure to observe the common flavor of all these norms (approximation spaces, Besov spaces, and interpolation spaces).

We have already mentioned that these interpolation spaces are identical to the approximation spaces whenever we have the Jackson and Bernstein inequalities in fold. What is ever more enlightening is that for classical pairs of spaces the K-functional and the interpolation spaces are always familiar quantities which have been walking the streets of analysis for decades. Let us give a couple of examples which will certainly convince even the most skeptical reader of the beautiful way in which the whole story pieces together.

The L_p spaces are interpolation spaces for the pair (L_1, L_∞) as is encapsulated in the Riesz-Thorin interpolation theorem (usually proved by means of complex interpolation). This theorem also follows from the real method of interpolation since the K-functional for this pair is easy to describe

$$K(f,t,L_1,L_\infty) = \int_0^t f^*(s)\,ds, \tag{17}$$

where f^* is the nondecreasing rearrangement of f as introduced by Hardy and Littlewood. From this characterization, one easily deduces that the interpolation spaces $(L_1,L_\infty)_{\theta,q}$ are identical to the Lorentz spaces $L_{p,q}$ with the identification $\theta = 1 - 1/p$. When $q = p$, we obtain $L_p = L_{p,p}$.

As a second example, consider the K-functional for the pair $(L_p(\Omega), W^k(L_p(\Omega)))$ on a Lipschitz domain $\Omega \subset \mathbb{R}^d$. Johnen and Scherer [37] showed that

$$K(f,t,L_p(\Omega),W^k(L_p(\Omega)) \approx \omega_r(f,t)_p \tag{18}$$

our old friend the modulus of smoothness. From this, one immediately deduces that $(L_p(\Omega), W^k(L_p(\Omega))_{\theta,q} = B^s_q(L_p(\Omega))$ for $\theta = s/k$.

There are numerous other examples of this sort beautifully reported on in the book by Bennett and Sharpley [8] that unquestionably convince us that the K-functional is indeed a natural object. These results make our job of characterizing the approximation spaces quite clear. We need only establish corresponding Jackson and Bernstein inequalities for the given approximation process and then finish the characterization via interpolation theory. This will be our modus operandi in the sequel.

3 The main types of nonlinear approximation

In application domains, there are four types of nonlinear approximation that are dominant. We want to see what form the general theory takes for these cases. We suppose that we are interested in approximating the elements $f \in X$ where X is a (quasi-) Banach space equipped with a norm $\|\cdot\|_X$.

3.1 n-term approximation

A set $\mathscr{D} \subset X$ of elements from X is called a dictionary if each element $g \in \mathscr{D}$ has norm one and the finite linear combinations of the elements in $\mathscr{D}$ are dense in X. The simplest example of a dictionary is when $\mathscr{D}$ is a basis for X. However, redundant systems $\mathscr{D}$ are also important. An issue is how much redundancy is possible while retaining reasonable computation.

Given a positive integer n, we define Σ_n as the set of all linear combinations of at most n elements from $\mathscr{D}$. Thus, the general element in Σ_n takes the form

$$S = \sum_{g\in\Lambda} c_g g, \quad \#(\Lambda) = n. \tag{19}$$

If we use the elements of Σ_n to approximate a function $f \in X$, then it induces an error

$$\sigma_n(f)_X := \inf_{S\in\Sigma_n} \|f - S\|_X. \tag{20}$$

Here we are following tradition to denote the error of nonlinear approximation by σ_n rather than using the generic E_n introduced earlier. The approximation spaces $\mathscr{A}^r_q(X)$ are defined as in the general setting. The approximation problem before us is whether we can characterize these spaces.

Let us consider the simplest case of the above setting where $X = \mathscr{H}$ is a real Hilbert space and $\mathscr{D} = \{\phi_j\}_{j=1}^\infty$ is an orthonormal basis for $\mathscr{H}$. Then, each $f \in \mathscr{H}$ has an orthogonal expansion

$$f = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j, \quad \|f\|_{\mathscr{H}}^2 = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2. \tag{21}$$

Because of the $\mathscr{H} \longrightarrow \ell_2$ isometry, a best n term approximation to a given $f \in \mathscr{H}$ is obtained by retaining its n largest terms (the possibility of ties in the size of the coefficients shows that this best approximation is not necessarily unique). Thus, if we let $c_j = \langle f, \phi_j \rangle$ and (c_j^*) be the rearrangement of their absolute values into nonincreasing order, then the approximation error of n-term approximation to f is

$$\sigma_n^2(f) = \sum_{j>n} [c_j^*]^2, \quad n = 1, 2, \ldots. \tag{22}$$

There is a simple characterization of the approximation spaces in this setting of n-term approximation. For example, for the primary spaces, we have that $f \in \mathscr{A}^r$ if and only if the coefficients (c_j) are in weak ℓ_τ with $1/\tau = s + 1/2$ and

$$\|f\|_{\mathscr{A}^r} \approx \|(c_j)\|_{w\ell_\tau}, \tag{23}$$

where we recall that weak ℓ_τ is the space of all sequences (a_j) which satisfy

$$\|(a_j)\|_{w\ell_\tau} := \sup_{n \geq 1} n^{1/\tau} a_n^* < \infty. \tag{24}$$

Similar results hold for the secondary spaces $\mathscr{A}_q^r$ characterizing them by the membership of the coefficient sequences in the Lorentz spaces $\ell_{\tau,q}$, $1/\tau = s + 1/2$. Indeed, this can be proved by establishing generalized Jackson-Bernstein inequalities for the pair $\mathscr{H}$ and Y_k as the set of $f \in \mathscr{H}$ whose coefficient are in weak ℓ_p with $1/p = k + 1/2$. We refer the reader to [25] for details.

In the case where we are interested in approximation in other spaces than $\mathscr{H}$, for example in L_p, $p \neq 2$, things are more subtle and depend very much on the particular basis $\{\phi_j\}$. Let us restrict our attention to to the wavelet basis which will play a special role in our discussion.

Suppose that φ is a compactly supported univariate scaling function (i.e. φ satisfies a two scale relationship) whose shifts form an orthonormal system. Let ψ be the compactly supported mother wavelet associated to φ normalized in $L_2(\mathbb{R}^d)$: $\|\psi\|_{L_2} = 1$. There are two ways to form an orthonormal wavelet system from this pair. The standard construction is to define $\psi^0 := \varphi$ and $\psi^1 := \psi$. If E' is the set of vertices of the unit cube and E the set of nonzero vertices, we define

$$\psi^e(x_1, \ldots, x_d) := \psi^{e_1}(x_1) \cdots \psi^{e_d}(x_d), \quad e \in E'. \tag{25}$$

The shifted dilates $\psi_{j,k}^e(x) := 2^{jd/2} \psi^e(2^j(x-k))$, $j \in \mathbb{Z}$, $k \in \mathbb{Z}^d$, $e \in E$, form an orthonormal system for $L_2(\mathbb{R}^d)$.

It is convenient to index these wavelets according to their spacial scaling. Let $\mathscr{D}(\mathbb{R}^d)$ denote the set of all dyadic cubes in $\mathbb{R}^d$. Each $I \in \mathscr{D}(\mathbb{R}^d)$ has the form $I = 2^{-jd}[k, k + \underline{1}]$ with $\underline{1} := (1, \ldots, 1)$. We identify the wavelets with the dyadic cubes via

$$\psi_I^e := \psi_{j,k}^e, \quad I \in \mathscr{D}(\mathbb{R}^d), e \in E. \tag{26}$$

This gives the wavelet decomposition

$$f = \sum_{I \in \mathscr{D}} \sum_{e \in E} f_{I,e} \psi_I^e, \quad f_{I,e} := \langle f, \psi_I^e \rangle, \tag{27}$$

which is valid for each $f \in L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$.

There is a second wavelet basis built directly from tensor products of univariate wavelets. If $R = I_1 \times \cdots \times I_d$, $I_j \in \mathscr{D}(\mathbb{R}^d)$, $j = 1, \ldots, d$, is a d dimensional dyadic rectangle, then we define

$$\psi_R(x) := \psi_{I_1}(x_1) \cdots \psi_{I_d}(x_d), \tag{28}$$

where each ψ_{I_j} is a univariate wavelet. This basis is sometimes called the hyperbolic wavelet basis or sparse grid basis in PDEs. The support of ψ_R is now associated to the rectangle R and in the case that ψ is the univariate Haar wavelet it is precisely this rectangle.

To continue the discussion, let us consider the first of these bases. Some of the results for L_2 approximation carry over to other approximation norms. The vehicle for doing this is the Littlewood-Paley theory for wavelets which allows one to compute other norms such as the L_p norms by simple expressions (the square function) of the wavelet coefficients. Rather than go too far down this road, which is well reported on in [25], we mention only some of the consequences of this. The first of which is the fact that it is possible to characterize the approximation spaces $\mathscr{A}_q^r(L_p)$ for certain special values of q even when the approximation takes place in an L_p space, $p > 1$. This even extends to $p \leq 1$ if we replace the L_p space by the Hardy space H_p. Namely, $\mathscr{A}_q^r(L_p(\mathbb{R}^d)) = B_q^{rd}(L_q(\mathbb{R}^d))$, provided $1/q = r + 1/p$. These results carry over to approximation on domains $\Omega \subset \mathbb{R}^d$ but now more care must be taken to define appropriate wavelet bases. The only case that is completely straightforward is to use the Haar wavelets for a cube such as $[0,1]^d$ in $\mathbb{R}^d$.

From the Besov characterizations of the approximation spaces given in the previous paragraph, we can see the power of nonlinear approximation. If we use the elements from linear spaces of dimension n (such as polynomials or splines on uniform partitions) to approximate a function $f \in L_p(\Omega)$, $\Omega \subset \mathbb{R}^d$, then we will obtain approximation of order $O(n^{-r})$ if and only if $f \in B_\infty^{rd}(L_p(\Omega)$, i.e. roughly speaking we need f to have rd derivatives in L_p. However, when using nonlinear methods such as n-term wavelet approximation it is sufficient to have $f \in B_q^{rd}(L_q)$, $1/q = r + 1/p$, i.e. rd derivatives in L_q. The gain here is not in the number of derivatives (rd) but in the space where these derivatives must lie. Since $q < p$ this requirement is much weaker in the case of nonlinear approximation. Indeed, functions with singularities may be in $f \in B_q^{rd}(L_q)$ but not in $f \in B_\infty^{rd}(L_p)$.

Here is a useful way to think about this comparison between linear and nonlinear for approximation in L_p. If we use linear methods, there will be a largest value s_L such that $f \in B_\infty^s(L_p)$ for all $s < s_L$. Similarly, there will be a largest s_{NL} such that $f \in B_q^s(L_q)$, $1/q = s/d + 1/p$ for all $s < s_{NL}$. We always have $s_{NL} \geq s_L$. However,

in many cases s_{NL} is much larger than s_L. This translates into being able to approximate such f with accuracy $O(n^{-s_{NL}/d})$ for nonlinear methods with n parameters but only accuracy $O(n^{-s_L/d})$ for linear methods with the same number of parameters. Consider the case $d = 1$ and a function f which is piecewise analytic with a finite number of jump discontinuities. If we approximate this function in $L_2[0,1]$ using linear spaces of dimension n, we will never get approximation orders better than $O(n^{-1/2})$ because $s_L = 1/2$, but using nonlinear methods we obtain order $O(n^{-r})$ for all $r > 0$ because $s_{NL} = \infty$.

Let us turn to the question of how we build a good n-term approximation to a function $f \in L_p$ where there is an important story to tell. It is very simple to describe how to choose a near best n-term approximation to a given f by simply choosing the n-terms in the wavelet expansion for which $\|f_{I,e}\psi_I^e\|_{L_p}$ is largest. Let $\tilde{\Lambda}_n(f) := \{(I,e)\}$ be the indices of these n largest terms (with ties in the size of the coefficients handled in an arbitrary way) and $S_n(f) := \sum_{(I,e)\in\tilde{\Lambda}_n(f)} f_{I,e}\psi_I^e$. Then we have the beautiful result of Temlyakov[50]

$$\|f - S_n(f)\|_{L_p(\mathbb{R}^d)} \le C\sigma_n(f)_{L_p(\mathbb{R}^d)}, \tag{29}$$

with the constant C depending only on d and p.

Sometimes it is notationally beneficial to renormalize the wavelets in L_p. Let us denote by $\psi_{I,p}^e$ these renormalized wavelets and by $f_{I,e,p}$ the coefficients of f with respect to this renormalized bases. Then a consequence of (29) is that a simple thresholding of the wavelet coefficients yields near best approximants. Namely, given any threshold $\delta > 0$, we denote by $\Lambda_\delta(f) := \Lambda_{\delta,p}(f) := \{(I,e) : |f_{I,e,p}| > \delta\}$, and the approximation

$$T_\delta(f) := \sum_{(I,e)\in\Lambda_\delta(f)} f_{I,e,p}\psi_{I,p}^e. \tag{30}$$

Then, $T_\delta(f)$ is a near best n-term approximation to f in $L_p(\mathbb{R}^d)$ for $n = \#(\Lambda_\delta(f))$. Notice that there is a slight distinction here between $T_\delta(f)$ and $S_n(f)$ because for some values of n, $S_n(f)$ cannot be obtained by thresholding because of possible ties in the size of coefficients.

Let us conclude this discussion of n-term approximation by remarking that it cannot be implemented directly in a numerical application because it requires a search over all wavelet coefficients which is an infinite task. In numerical practice this search is limited by fixing a maximal dyadic level J to limit the search. Other numerically friendly nonlinear algorithms are adaptive and tree based algorithm which we discuss next.

3.2 Adaptive approximation

This type of approximation has a long history and owes a lot of its interest to its usefulness in describing certain numerical methods for PDEs. To drive home the main ideas behind adaptive approximation, let us consider the simple setting of approximating a function f on the unit cube $\Omega := [0,1]^d$ in $\mathbb{R}^d$ using piecewise polynomials on partitions consisting of dyadic cubes from $\mathscr{D}(\Omega) := \{I \in \mathscr{D}(\mathbb{R}^d) : I \subset \Omega\}$. Given an integer $r > 0$ and an $f \in L_p(\Omega)$, we denote by

$$E_r(f,I)_p := \inf_{Q \in \mathscr{P}_{r-1}} \|f - Q\|_{L_p(I)}, \tag{31}$$

the L_p error in approximating f on I by polynomials of order r (total degree $r-1$). The simplest adaptive algorithms are built on an estimator $E(I)$ for $E_r(f,I)_p$:

$$E_r(f,I)_p \le E(I), \quad I \in \mathscr{D}(\Omega). \tag{32}$$

To build an adaptive approximation to f, we let $\Lambda_0 := \{\Omega)\}$ and given that $\Lambda_n = \Lambda_n(f)$ has been defined, we generate Λ_{n+1} by choosing the dyadic cube $I = I_n$ from Λ_n for which the estimator $E(I_n)$ is largest (with again ties handled arbitrarily) and then removing I and replacing it by its 2^d children. Thus, the idea is to only subdivide where the error is largest. There have been several papers discussing the approximation properties of such adaptive algorithms starting with the pioneering work of Birman and Solomjak [13] which established convergence rates (in the case $E(I) = E(f,I)_p$) very similar to the estimates of the previous section for n-term wavelet approximation. A typical result is that if a function f is in a Besov space $B^s_q(L_\tau)$ which compactly embeds into L_p then a suitable adaptive algorithm will provide an approximation to f with accuracy $O(n^{-s/d})$ where n is the number of parameters (proportional to the number of cells in the adaptive partition). One can easily argue that one cannot do away with the assumption of compact embedding. Such results on adaptive approximation are only slightly weaker than those for n-term approximation. In the latter one does not assume compactness of the embedding into L_p.

One can even guarantee a certain near optimal performance of adaptive algorithms although now the rule for subdividing is more subtle. These will be described in the next section in the more general setting of tree approximation.

3.3 Tree approximation

We have already noted that trees arise in a natural way in nonliner approximation. The wavelet decomposition organizes itself on trees whose nodes are dyadic cubes in $\mathbb{R}^d$. We have also seen that adaptive partitioning is described by a tree whose nodes are the cells created during the adaptive algorithm. It is useful to formalize

tree approximation and extract its main features since we shall see that it plays a significant role in applications of nonlinear approximation.

We assume that we have a (generally infinite) master tree $\mathscr{T}^*$ with one root node. In the case of adaptive partitioning this root node would be the domain Ω. We also assume that each node has exactly K children. This matches both the wavelet tree and the usual refinement rules in adaptive partitioning. Note that in the case the master tree arises from adaptive partitioning, it fixes the way a cell must be subdivided when it arises in an adaptive algorithm. So this setting does not necessarily cover all possible adaptive strategies.

We shall be interested in finite subtrees $\mathscr{T} \subset \mathscr{T}^*$. Such a tree $\mathscr{T}$ has the property that for any node in $\mathscr{T}$ its parent is also in $\mathscr{T}$. We define $\mathscr{L}(T)$ to be the leaves of $\mathscr{T}$. This is the set of all terminal nodes in $\mathscr{T}$, i.e. such a node has none of its children in $\mathscr{T}$. We say that the tree is complete if whenever a node is in $\mathscr{T}$ all of its siblings are also in $\mathscr{T}$. We shall restrict our discussion to complete trees. Any adaptively generated partition is associated to a complete tree $\mathscr{T}$. We define $\mathscr{N}(\mathscr{T})$ to be the set of the internal nodes of $\mathscr{T}$, i.e. the ones which are not leaves. Then $\mathscr{T} = \mathscr{N}(\mathscr{T}) \cup \mathscr{L}(\mathscr{T})$, if considered as sets.

As the measure of complexity of a tree $\mathscr{T} \subset \mathscr{T}^*$ we consider the number of subdivisions $\mathbf{n}(\mathscr{T})$ needed to create $\mathscr{T}$ from its root. We shall often use the fact that

$$\mathbf{n}(\mathscr{T}) = \#(\mathscr{N}(\mathscr{T})). \tag{33}$$

It follows that

$$\#(\mathscr{T}) = K\mathbf{n}(\mathscr{T}) + 1 \tag{34}$$

Also, for a complete tree, $\mathscr{L}(\mathscr{T}) = 1 + (K-1)\mathbf{n}(\mathscr{T})$. So, $\mathbf{n}(\mathscr{T})$ is a fair measure of the complexity of $\mathscr{T}$.

In tree approximation, we assume that to every node $I \in \mathscr{T}^*$, we have an error or *energy* $e(I)$. We measure the performance of a finite tree $\mathscr{T}$ by

$$E(\mathscr{T}) := \sum_{I \in \mathscr{L}(\mathscr{T})} e(I). \tag{35}$$

If we are considering trees corresponding to adaptive partitioning then we would take $e(I) = E(f,I)_p^p$ where $E(f,I)_p$ is the local $L_p(I)$ error on the cell I. Similarly, if we are doing wavelet approximation in L_2 then we would take $e(I) := \sum_{J \subset I} \sum_{e \in E} |f_I^e|^2$ which would be the energy in the wavelet coefficients on all nodes of the tree below I (this corresponds to the error contributed by not including these coefficients). We are interested in the best performance of trees of size $\mathbf{n}(\mathscr{T}) \leq n$ which is given by

$$\sigma_n := \inf_{\mathbf{n}(\mathscr{T}) \leq n} E(\mathscr{T}). \tag{36}$$

Using this definition of σ_n gives the approximation classes $\mathscr{A}_q^r(L_p)$ for tree approximation in L_p.

What is the cost of tree approximation versus n-term approximation? The main point of our work with Wolfgang on wavelet tree approximation given in [20] is

that the cost is almost negligible. Recall that for n-term wavelet approximation in $L_p(\Omega)$, $\Omega \subset \mathbb{R}^d$, we achieve error $O(n^{-r/d})$ for a function f if it is in the Besov space $B^r_q(L_q(\mathbb{R}^d))$ with $1/q = r/d + 1/p$. These latter spaces are barely embedded in L_p and are not compactly embedded. We prove in [20] that whenever a Besov space $B^r_q(L_\tau)$ is compactly embedded into $L_p(\Omega)$ then wavelet tree approximation gives the same approximation rate $O(n^{-r/d})$. Said in another way, this Besov space is embedded into $\mathcal{A}_\infty^{r/d}(L_p)$. Of course, we get such a compact embedding whenever $\tau > (r/d + 1/p)^{-1}$ because of the Sobolev embedding theorem. Thus, from this point of view, tree approximation performs almost as well as n-term approximation.

The proof of the above result on the performance of wavelet tree approximation requires the counting of the new nodes added in order to guarantee the tree structure. However, the number of these new nodes can be controlled by grouping the nodes according to the size of the wavelet coefficients and counting each grouping. Finally, let us remark that in [11] we prove similar theorems on tree approximation for trees generated by adaptive partitioning. This plays an important role in understanding which solutions to elliptic partial differential equations can be well approximated by adaptive finite element methods.

Let us turn to the discussion of finding near best trees. Finding the best tree that matches σ_k in (36) is practically infeasible since it would require searching over all trees $\mathcal{T} \subset \mathcal{T}^*$ with $\mathbf{n}(\mathcal{T}) = k$ and the number of such trees is exponential in k. Remarkably, however, it is possible to design practical algorithms that do almost as well while involving only $O(n)$ computations. The first algorithms of this type were given in [12]. We shall describe a modification of this approach that gives slightly better constants in the estimation of performance.

The tree algorithm we shall consider can be implemented in the general setting of [12]. However, here, we shall limit ourselves to the following setting. We assume the error functionals are *subadditive* in the sense that

$$e(\mathbf{I}) \geq \sum_{\mathbf{I}' \in \mathcal{C}(\mathbf{I})} e(\mathbf{I}'), \tag{37}$$

where $\mathcal{C}(\mathbf{I})$ is the set of children of $\mathbf{I}$. This property holds for the examples we have described above.

A naive strategy to generate a good tree for adaptive approximation would be to mark for subdivision the cells which have largest local errors. However, such a strategy would not generate near optimal trees because it could happen that subdividing a cell and its successive generations would not reduce at all the global error and so a better strategy would have been to subdivide some other cell. To obtain near optimal algorithms, one has to be more clever and penalize successive subdivisions which do not markedly reduce the error. This is done through certain *modified error functionals* $\tilde{e}(\mathbf{I})$ whose precise definition we postpone for a moment. The tree algorithm we propose will grow a given tree $\mathbb{T}$ by including the children of $\mathbf{I}$ as new nodes when $\tilde{e}(\mathbf{I})$ is the largest among all $\tilde{e}(\mathbf{I}') \in \Lambda(\mathbb{T})$.

In our formulation and analysis of the tree algorithm, the local error functional e can be any functional defined on the nodes $\mathbf{I}$ in $\mathcal{T}$ which is subadditive.

Tree-Algorithm:

- Let $\mathscr{T}^0 := \{X\}$ be the root tree.
- If $\mathscr{T}^k$ has been defined for some $k \geq 0$, then define

$$\mathbf{I}^* = \operatorname{argmax}\{\tilde{e}(\mathbf{I}) : \mathbf{I} \in \mathscr{L}(\mathscr{T}_k)\}$$

and $\mathscr{T}^{k+1} := \mathscr{T}^k \cup \{\mathscr{C}(\mathbf{I}^*)\}$.

As the modified error functional, we employ

$$\tilde{e}(\mathbf{I}) := e(\mathbf{I}) \quad \text{for} \quad \mathbf{I} = X \quad \text{and} \quad \tilde{e}(\mathbf{I}) := \left(\frac{1}{e(\mathbf{I})} + \frac{1}{\tilde{e}(\mathbf{I}')}\right)^{-1} \quad \text{for} \quad \mathbf{I} \in \mathscr{C}(\mathbf{I}'). \tag{38}$$

The purpose of the modified error is to penalize children of cells which are chosen for subdivision but the resulting refinement does not significantly decrease the total error. Notice that in such a case the modified error $\tilde{e}$ decreases for the children and therefore makes them less apt to be chosen in later subdivisions.

The following theorem describes the performance of the tree algorithm.

Theorem 3.1. *At each step n of the above* tree algorithm *the output tree $\mathbb{I} = \mathbb{I}_n$ satisfies*

$$E(\mathscr{T}) \leq \left(\frac{n}{n-k}\right) \sigma_k, \tag{39}$$

whenever $k < n$.

The main distinction of the above results from previous ones in [12] is that the constant on the right hand side of (39) is now completely specified and, in particular, does not involve the total number of children of a node. Note that the computational complexity of implementing the tree algorithm with a resulting tree $\mathscr{T}$ depends only on $\mathbf{n}(\mathscr{T})$. Therefore, when applying this algorithm to adaptive partitioning, it is independent of the spatial dimension d. The proof of the above theorem will be given in a forthcoming paper with Peter Binev, Wolfgang, and Phillipp Lamby.

3.4 Greedy algorithms

In application domains, there is a desire to have as much approximation power as possible. This is accomplished by choosing a large dictionary $\mathscr{D}$ to increase approximation power. However, their sheer size can cause a stress on computation. Greedy algorithms are a common approach to keeping computational tasks reasonable when dealing with large dictionaries. They have a long history in statistics and signal processing. A recent survey of the approximation properties of such algorithms is given in [51] where one can find the main results of this subject.

We shall consider only the problem of approximating a function f from a Hilbert space $\mathscr{H}$ by a finite linear combination $\hat{f}$ of elements of a given dictionary $\mathscr{D} =$

$(g)_{g\in\mathscr{D}}$. We have already discussed the case where $\mathscr{D}$ is an orthonormal basis. One of the motivations for utilizing general dictionaries rather than orthonormal systems is that in many applications, such as signal processing or statistical estimation, it is not clear which orthonormal system, if any, is best for representing or approximating f. Thus, dictionaries which are a union of several bases or collections of general waveforms are preferred. Some well known examples are the use of Gabor sytems, curvelets, and wavepackets in signal processing and neural networks in learning theory.

When working with dictionaries $\mathscr{D}$ which are not orthonormal bases, the realization of a best n-term approximation is usually out of reach from a computational point of view since it would require minimizing $\|f-\hat{f}\|$ over all $\hat{f}$ in an infinite or huge number of n dimensional subspaces. *Greedy algorithms* or matching pursuit aim to build "sub-optimal yet good" n-term approximations through a greedy selection of elements g_k, $k=1,2,\cdots$, within the dictionary $\mathscr{D}$, and to do so with a more manageable number of computations.

There exist several versions of these algorithms. The four most commonly used are the *pure greedy*, the *orthogonal greedy*, the *relaxed greedy* and the *stepwise projection* algorithms, which we respectively denote by the acronyms PGA, OGA, RGA and SPA. All four of these algorithms begin by setting $f_0 := 0$. We then define recursively the approximant f_k based on f_{k-1} and its residual $r_{k-1} := f - f_{k-1}$.

In the PGA and the OGA, we select a member of the dictionary as

$$g_k := \operatorname{argmax}_{g\in\mathscr{D}} |\langle r_{k-1}, g\rangle|. \tag{40}$$

The new approximation is then defined as

$$f_k := f_{k-1} + \langle r_{k-1}, g_k\rangle g_k, \tag{41}$$

in the PGA, and as

$$f_k = P_k f, \tag{42}$$

in the OGA, where P_k is the orthogonal projection onto $V_k := \operatorname{Span}\{g_1,\cdots,g_k\}$. It should be noted that when $\mathscr{D}$ is an orthonormal basis both algorithms coincide with the computation of the best k-term approximation.

In the RGA, the new approximation is defined as

$$f_k = \alpha_k f_{k-1} + \beta_k g_k, \tag{43}$$

where (α_k,β_k) are real numbers and g_k is a member of the dictionary. There exist many possibilities for the choice of (α_k,β_k,g_k), the most greedy being to select them according to

$$(\alpha_k,\beta_k,g_k) := \operatorname*{argmin}_{(\alpha,\beta,g)\in\mathbb{R}^2\times\mathscr{D}} \|f-\alpha f_{k-1}-\beta g\|. \tag{44}$$

Other choices specify one or several of these parameters, for example by taking g_k as in (40) or by setting in advance the value of α_k and β_k, see e.g. [38] and [4]. Note that the RGA coincides with the PGA when the parameter α_k is set to 1.

In the SPA, the approximation f_k is defined by (42) as in the OGA, but the choice of g_k is made so as to minimize over all $g \in \mathscr{D}$ the error between f and its orthogonal projection onto $\mathrm{Span}\{g_1, \cdots, g_{k-1}, g\}$.

Note that, from a computational point of view, the OGA and SPA are more expensive to implement since at each step they require the evaluation of the orthogonal projection $P_k f$ (and in the case of SPA a renormalization). Such projection updates are computed preferrably using Gram-Schmidt orthogonalization (e.g. via the QR algorithm) or by solving the normal equations

$$G_k a_k = b_k, \tag{45}$$

where $G_k := (\langle g_i, g_j \rangle)_{i,j=1,\cdots,k}$ is the Gramian matrix, $b_k := (\langle f, g_i \rangle)_{i=1,\cdots,k}$, and $a_k := (\alpha_j)_{j=1,\cdots,k}$ is the vector such that $f_k = \sum_{j=1}^{k} \alpha_j g_j$.

In order to describe the known results concerning the approximation properties of these algorithms, we introduce the class $\mathscr{L}_1 := \mathscr{L}_1(\mathscr{D})$ consisting of those functions f which admit an expansion $f = \sum_{g \in \mathscr{D}} c_g g$ where the coefficient sequence (c_g) is absolutely summable. We define the norm

$$\|f\|_{\mathscr{L}_1} := \inf\{ \sum_{g \in \mathscr{D}} |c_g| \ : \ f = \sum_{g \in \mathscr{D}} c_g g \} \tag{46}$$

for this space. This norm may be thought of as an ℓ_1 norm on the coefficients in representation of the function f by elements of the dictionary; it is emphasized that it is not to be confused with the L_1 norm of f. An alternate and closely related way of defining the $\mathscr{L}_1$ norm is by the infimum of numbers V for which f/V is in the closure of the convex hull of $\mathscr{D} \cup (-\mathscr{D})$. This is known as the "variation" of f as introduced in [3].

In the case where $\mathscr{D}$ is an orthonormal basis, we find that if $f \in \mathscr{L}_1$,

$$\sigma_N(f) = (\sum_{g \notin \Lambda_n(f)} |c_g|^2)^{1/2} \leq (\|f\|_{\mathscr{L}_1} \min_{g \in \Lambda_n(f)} |c_g|)^{1/2} \leq \|f\|_{\mathscr{L}_1} N^{-1/2}, \tag{47}$$

which is contained in (23).

For the PGA, it was proved in [29] that $f \in \mathscr{L}_1$ implies that

$$\|f - f_N\| \lesssim N^{-1/6}. \tag{48}$$

This rate was improved to $N^{-\frac{11}{62}}$ in [40], but on the other hand it was shown [43] that for a particular dictionary there exists $f \in \mathscr{L}_1$ such that

$$\|f - f_N\| \gtrsim N^{-0.27}. \tag{49}$$

When compared with (47), we see that the PGA is far from being optimal.

The RGA, OGA and SPA behave somewhat better: it was proved respectively in [38] for the RGA and SPA, and in [29] for the OGA, that one has

$$\|f - f_N\| \lesssim \|f\|_{\mathscr{L}_1} N^{-1/2}, \tag{50}$$

for all $f \in \mathscr{L}_1$.

For each of these algorithms, it is known that the convergence rate $N^{-1/2}$ cannot in general be improved even for functions which admit a very sparse expansion in the dictionary $\mathscr{D}$ (see [29] for such a result with a function being the sum of two elements of $\mathscr{D}$).

At this point, some remarks are in order regarding the meaning of the condition $f \in \mathscr{L}_1$ for some concrete dictionaries. A commonly made statement is that greedy algorithms break the *curse of dimensionality* in that the rate $N^{-1/2}$ is independent of the dimension d of the variable space for f, and only relies on the assumption that $f \in \mathscr{L}_1$. This is not exactly true since in practice the condition that $f \in \mathscr{L}_1$ becomes more and more stringent as d grows. For instance, in the case where we work in the Hilbert space $\mathscr{H} := L_2([0,1]^d)$ and when $\mathscr{D}$ is a *wavelet basis* (ψ_λ), it follows from our earlier observations in §3.1 that the smoothness property which ensures that $f \in \mathscr{L}_1$ is that f should belong to the Besov space $B_1^s(L_1)$ with $s = d/2$, which roughly means that f has all its derivatives of order less or equal to $d/2$ in L_1 (see [25] for the characterization of Besov spaces by the properties of wavelet coefficients). Another instance is the case where $\mathscr{D}$ consists of sigmoidal functions of the type $\sigma(v \cdot x - w)$ where σ is a fixed function and v and w are arbitrary vectors in $\mathbb{R}^d$, respectively real numbers. For such dictionaries, it was proved in [4] that a sufficient condition to have $f \in \mathscr{L}_1$ is the convergence of $\int |\omega||\mathscr{F} f(\omega)|d\omega$ where $\mathscr{F}$ is the Fourier operator. This integrability condition requires a larger amount of decay on the Fourier transform $\mathscr{F} f$ as d grows. Assuming that $f \in \mathscr{L}_1$ is therefore more and more restrictive as d grows. Similar remarks also hold for other dictionaries (hyperbolic wavelets, Gabor functions etc.).

The above discussion points to a significant weakness in the theory of greedy algorithms in that there are no viable bounds for the performance of greedy algorithms for general functions $f \in \mathscr{H}$. This is a severe impediment in some application domains (such as learning theory) where there is no a priori knowledge that would indicate that the target function is in $\mathscr{L}_1$. One of the main contributions of the work with Wolfgang [7] was to provide error bounds for the performance of greedy algorithms for general functions $f \in \mathscr{H}$. This was accomplished by developing a technique based on interpolation of operators that provides convergence rates N^{-s}, $0 < s < 1/2$, whenever f belongs to a certain intermediate space between $\mathscr{L}_1$ and the Hilbert space $\mathscr{H}$. Namely, we used the spaces

$$\mathscr{B}_p := [\mathscr{H}, \mathscr{L}_1]_{\theta,\infty}, \quad \theta := 2/p - 1, \quad 1 < p < 2, \tag{51}$$

which are the real interpolation spaces between $\mathscr{H}$ and $\mathscr{L}_1$. We showed that if $f \in \mathscr{B}_p$, then the OGA and RGA, when applied to f, provide approximation rates CN^{-s} with $s := \theta/2 = 1/p - 1/2$. Thus, if we set $\mathscr{B}_1 = \mathscr{L}_1$, then these spaces provide a full range of approximation rates for greedy algorithms. Recall, as discussed previously, for general dictionaries, greedy algorithms will not provide convergence rates better than $N^{-1/2}$ for even the simplest of functions. The results we obtained were optimal in the sense that they recovered the best possible convergence rate in

the case where the dictionary is an orthonormal basis. For an arbitrary target function $f \in \mathscr{H}$, convergence of the OGA and RGA holds without rate.

4 Image compression

The emergence of wavelets as a good representation system took place in the late 1980's. One of the most impressive applications of the wavelet system occured in image processing, especially compression and denoising. There are a lot of stories to be told here including the method of thresholding wavelet coefficients for denoising, first suggested by Donoho and Johnstone [31], as a simple methodology for effectively solving imaging problems. But we shall restrict our attention to the problem of understanding the best implementation of wavelets in compression (image encoding).

What is an image? Too often the view is a digitized image. While this matches what we treat in application, it is not the correct launching point for a theory. Engineers usually view images and signals as realizations of a stochastic process. One can debate the efficacy of this viewpoint versus the deterministic viewpoint I am going to now advocate.

In [26], we proposed to view images as functions f defined on a continuum which we shall normalize as the unit square $[0,1]^2$. The digitized images we observe are then simply samples of f given as averages over small squares (pixels). Thus, any representation system for functions on $[0,1]^2$ can be used to for images and computations are made from the samples. We advocated the use of wavelets because of its multiscale structure and the remainder of our discussion of image processing will be limited to wavelet decompositions.

Suppose we we wish to compress functions using wavelet decompositions. The first step is to chose the norm or metric in which we wish to measure distortion. This is traditionally done using the L_2 norm which corresponds to what Engineers use in their measure of Peak Signal to Noise Ratio (PSNR). However, for the purposes of this discussion any L_p norm would work equally well. We have already seen that a near best n term approximation (actually best when $p = 2$) is gotten by simply keeping the n largest terms (measured in L_p) of the wavelet decomposition. So this must be how to do compression. However to convert everything to a binary bitstream one has to further quantize the coefficients since in general the wavelet coefficients are real numbers.

Understanding how to quantize is quite easy if one recalls the connection between n-term approximation and thresholding. Namely, as explained earlier, except for possible ties in the sizes of wavelet coefficients, choosing the biggest n terms corresponds to setting a threshold and retaining the wavelet coefficients above this threshold. Since thresholding takes the view that coefficients below the threshold size $\eta > 0$ should not be retained, it makes perfect sense that quantizing a wavelet coefficient a should be made by taking the smallest number of binary bits of a so that the recovery $\hat{a}$ from these bits satisfies $|a - \hat{a}| \leq \eta$. This makes a perfectly reason-

able compression scheme except that in addition one has to send bits to identify the index of the wavelet coefficient. Here the matter becomes a little more interesting.

Before embarking on the index identification problem, let us remark that the characterization (given in §3.1) of the approximation classes $\mathscr{A}_\tau^r(L_p)$ as Besov spaces $B_\tau^s(L_\tau)$ when $1/\tau = r + 1/p$ and $r = s/2$ (because we are in two space dimensions) gives a very satisfying characterization of which images can be compressed with a given distortion rate if we measure complexity of the encoding by the number of terms retained in the wavelet decomposition. This was the story told in [26]. However, there was rightfully considerable objection to this theory since it was based on the number of terms n retained and not on the number of bits needed to encode this information.

A major step in the direction of giving a theory based on the number of bits was taken in the paper of Cohen, Daubechies, Gulyeruz, and Orchard [21]. It was however limited to measuring distortion in the L_2 norm. With Wolfgang, we wanted to give a complete theory that would include measuring distortion in any L_p space. The key step in developing such a theory was to consider the notion of tree approximation and in fact this is where the theory of tree approximation characterizing the spaces $\mathscr{A}_q^r(L_p, tree)$ for the wavelet basis (described earlier) was developed. Let us see how this solves our encoding problem.

To build a compression for functions, we first choose our compression metric L_p. We then agree on a minimal smoothness ε that we shall assume of the functions in L_p. This step is necessary so that the encoder is applied to a compact set of functions. Next, we find the wavelet coefficients of the wavelet decomposition of the image with respect to the wavelet basis normalized in L_p. We then build a sequence of trees $\mathscr{T}_k$ associated to the image as follows. We consider the set Λ_k of all wavelet indices for which the coefficient of the image is in absolute value $\geq 2^{-k}$. The nodes in Λ_k will not form a tree so we complete them to the smallest tree $\mathscr{T}_k$ which contains Λ_k. An important point here is that the sets Λ_k and the tree $\mathscr{T}_k$ can be found without computing and searching over an infinite set of wavelet coefficients because of our assumption on minimal smoothness in L_p.

Notice that the tree $\mathscr{T}_k$ is contained in $\mathscr{T}_{k+1}$. Therefore $\Delta_k := \mathscr{T}_k \setminus \mathscr{T}_{k-1}$ will tell us how to obtain $\mathscr{T}_k$ once $\mathscr{T}_{k-1}$ is known. This process is called *growing the tree*.

We shall send a progressive bitstream to the receiver. After receiving any portion of this bitstream the receiver will be able to construct an approximation of the image with higher and higher resolution (in our chosen L_p metric) as more and more bits are received. The first bits will identify the smallest value of k_0 for which Λ_{k_0} is nonempty. Then come the bits to identify $\mathscr{T}_{k_0}$ followed by bits to identify the sign of the coefficients in $\mathscr{T}_{k_0}$ and one bit of the binary expansion of each of the coefficients. Later bits come in packets. Each packet tells us how to go from $\mathscr{T}_{k-1}$ to $\mathscr{T}_k$ and how to increase the resolution of each of the coefficients in hand.

Precisely, in the k-th packet we first send bits that tell how to grow $\mathscr{T}_{k-1}$ to $\mathscr{T}_k$. Next, we send a bit for each new coefficient (i.e. those in Δ_k) to identify its sign, next comes one bit (the lead bit) of the binary expansion for each new coefficient. Finally, we send one additional bit for each of the old coefficients that had been previously sent.

For the resulting encoder one can prove the following result of [20]:
Performance of image encoder: *If the image $f \in B^s_q(L_\tau)$ for some $s > 0$ and $\tau > (s/2+1/p)^{-1}$, then after receiving n bits, these bits can be decoded to give an image $\hat{f}$ such that $\|f - \hat{f}\|_{L_p} \leq Cn^{-s/2}$.*

There were two key ingredients in proving the above result on the performance of the encoder. The first of these is to show that tree approximation is as effective as n-term approximation when approximating functions in Besov classes that compactly embed into L_p. We have already discussed this issue in our section on tree approximation. The second new ingredient is to show that any quad tree with m nodes can be encoded using at most $4m$ bits. Here, we borrowed the ideas from [21].

5 Remarks on nonlinear approximation in PDE solvers

Certainly, the construction of numerical algorithms based on nonlinear approximation for solving PDEs has been one of Wolfgang's major accomplishments. An extensive description of this development for elliptic PDEs will be presented in the contribution of Morin, Nochetto and Siebert in this volume. We will restrict our remarks to some historical comments.

We shall discuss only the model Laplace problem

$$-\Delta(u) = f \text{ on } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{52}$$

where $f \in H^{-1}$ and the solution u is to be captured in the energy norm which in this case is the $H^1_0(\Omega)$ norm. The solution to such equations is well known to generate singularities of two types. The first is due to singularities in f itself while the other come from the boundary of the domain, for example corner singularities. So it is natural to envision nonlinear approximation methods as the basis for effective numerical solvers. Indeed, it was already shown in [23], that the solutions to (52) on Lipschitz domains always have higher smoothness in the scale of Besov spaces corresponding to nonlinear approximation than they do in the scale for linear approximation. So the theoretical underpinnings were there to advocate nonlinear methods and they were certainly in vogue beginning with the work of Ivo Babuska and his collaborators (starting with [1]). Surprisingly, there was no algorithm based on nonlinear methods which was proven to outperform linear methods save for some univariate results.

Wolfgang brought Albert and I this problem and explained the bulk chasing technique of Doerfler [32] which can be used to show convergence (but no rates) for adaptive finite element methods (with some massaging as provided by Morin, Nochetto, and Siebert [44]). We thought that the easiest type algorithm to analyze would be based on wavelet decompositions. One advantage of chosing wavelets is that (52) can be converted to an infinite matrix operator equation

$$\mathcal{A}\bar{u} = \bar{f} \tag{53}$$

where $\mathscr{A}$ is bounded and boundedly invertible on ℓ_2. Here one employs the wavelet preconditioning (diagonal rescaling) utilized in the analysis of preconditioning in [24]. The key property inherited by this matrix is off diagonal decay which can also be described as a compressibility in that $\mathscr{A}$ can be well approximated by finite rank matrices.

In analogy with the results on image encoding, we wanted to create a Galerkin algorithm for numerically solving (52) based on wavelet tree approximation such that whenever u is in one of the approximation classes $\mathscr{A}^s$ then the algorithm produces an approximant to u (in the energy norm) with near optimal rate distortion. Namely, if N is the cardinality of the tree $\mathscr{T}$ associated to the numerical approximation $u_{\mathscr{T}}$, then

$$\|u - u_{\mathscr{T}}\|_{H^1_0} \leq C_0 \|u\|_{\mathscr{A}^s} N^{-s}. \tag{54}$$

In the end we actually did much better since we showed the operational count needed to compute $u_{\mathscr{T}}$ could also be kept proportional to N.

We were quickly able to build the framework for the wavelet numerical algorithm. However, we wrestled for quite some time to derive optimal bounds for the number of terms in the wavelet decomposition of the approximant. This of course is necessary for any rate distortion theory. In the end, we went back to our analogy with image compression where one discards small coefficients in such decompositions when seeking optimal compression and noise reduction. This led to our coarsening algorithm and a subsequent proof of optimal performance of the numerical algorithm. It was an important contribution of Stevenson [48] that it is actually possible to build adaptively wavelet algorithms without coarsening with the same optimal rate distortion theory. Heuristically, if one is not too aggressive with the bulk chasing then the majority of the nodes chosen will in the end survive coarsening.

Our first paper [16] on adaptive wavelet methods was built on solving finite discrete problems formed by taking appropriate subsections of the matrix $\mathscr{A}$. This actually turned out to be the wrong view. Wolfgang proposed the idea that we should retain as long as possible the infinite matrix form (53) and algorithms should be viewed as solving this infinite dimensional problem. This turned out to be not only the right conceptual view but also very powerful in algorithm development. This allowed us to solve non-coercive problems and provide a very robust and elegant theory in [17].

With Peter Binev, Wolfgang and I wondered why we could not carry our wavelet theory over to finite element methods based on adaptive triangulations . We quickly found out that these algorithms had major differences from wavelet algorithms. First of all, in contrast to having one matrix (53) governing the algorithm, the matrices changed at each iteration. This made the effect of refining triangles much more subtle than the growing wavelet trees. Fortunately, we were able to borrow the theory of local error estimators for finite elements developed by Morin, Nochetto, and Siebert [44]. Another major difficulty was the fact the problem of hanging nodes (or non-conforming elements). This required us to develop a way to count the additional refinements necessary to guarantee conforming elements. This was eventually given

by a nice maximal function type algorithm. Our algorithm for adaptive finite element methods again had a coarsening step based on the tree algorithm of [12]. Again, Rob Stevenson was able to show that one can proceed without coarsening. Now there is a much finer understanding of adaptive finite element algorithms which will be well presented in the contribution of Morin, Nochetto, and Siebert in this volume.

6 Learning theory

Learning theory is a problem in data fitting. The data is assumed to be generated by an unknown measure ρ defined on a product space $Z := X \times Y$. We shall assume that X is a bounded domain of $\mathbb{R}^d$ and $Y = \mathbb{R}$. The article of Gerard Kerkyacharian, Mathilde Mougeot, Dominique Picard, and Karine Tribouley in this volume will give a general exposition of this subject. Here we want to touch on some aspects of this subject that relate to nonlinear approximation.

We assume that we are given m independent random observations $z_i = (x_i, y_i)$, $i = 1, \dots, m$, identically distributed according to ρ. We are interested in finding the function f_ρ which best describes the relation between the y_i and the x_i. This is the *regression function* $f_\rho(x)$ defined as the conditional expectation of the random variable y at x:

$$f_\rho(x) := \int_Y y d\rho(y|x) \tag{55}$$

with $\rho(y|x)$ the conditional probability measure on Y with respect to x. We shall use $\mathbf{z} = \{z_1, \dots, z_m\} \subset Z^m$ to denote the set of observations.

One of the goals of learning is to provide estimates under minimal restrictions on the measure ρ since this measure is unknown to us. We shall work under the mild assumption that this probability measure is supported on an interval $[-M, M]$

$$|y| \le M, \tag{56}$$

almost surely. It follows in particular that $|f_\rho| \le M$. This property of ρ can usually be inferred in practical applications.

We denote by ρ_X the marginal probability measure on X defined by

$$\rho_X(S) := \rho(S \times Y). \tag{57}$$

We shall assume that ρ_X is a Borel measure on X. We have

$$d\rho(x,y) = d\rho(y|x) d\rho_X(x). \tag{58}$$

It is easy to check that f_ρ is the minimizer of the risk functional

$$\mathscr{E}(f) := \int_Z (y - f(x))^2 d\rho, \tag{59}$$

over $f \in L_2(X,\rho_X)$ where this space consists of all functions from X to Y which are square integrable with respect to ρ_X. In fact one has

$$\mathscr{E}(f) = \mathscr{E}(f_\rho) + \|f - f_\rho\|^2, \tag{60}$$

where

$$\|\cdot\| := \|\cdot\|_{L_2(X,\rho_X)}. \tag{61}$$

The goal in learning is to find an *estimator* $f_{\mathbf{z}}$ for f_ρ from the given data $\mathbf{z}$. The usual way of evaluating the performance of such an estimator is by studying its convergence either in probability or in expectation, i.e. the rate of decay of the quantities

$$\text{Prob}\{\|f_\rho - f_{\mathbf{z}}\| \geq \eta\}, \quad \eta > 0 \quad \text{or} \quad E(\|f_\rho - f_{\mathbf{z}}\|^2) \tag{62}$$

as the sample size m increases. Here both the expectation and the probability are taken with respect to the product measure ρ^m defined on Z^m. Estimations in probability are to be preferred since they give more information about the success of a particular algorithm and they automatically yield an estimate in expectation by integrating with respect to η. Much more is known about the performance of algorithms in expectation. This type of regression problem is referred to as *random design* or *distribution-free* because there are no a priori assumption on ρ_X. An excellent survey on distribution free regression theory is provided in the book [35], which includes most existing approaches as well as the analysis of their rate of convergence in the expectation sense.

A common approach to regression estimation is to choose an hypothesis (or *model*) class $\mathscr{H}$ and then to define $f_{\mathbf{z}}$, in analogy to (59), as the minimizer of the empirical risk

$$f_{\mathbf{z}} := \operatorname*{argmin}_{f\in\mathscr{H}} \mathscr{E}_{\mathbf{z}}(f), \quad \text{with} \quad \mathscr{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{j=1}^{m}(y_j - f(x_j))^2. \tag{63}$$

In other words, $f_{\mathbf{z}}$ is the best approximation to $(y_j)_{j=1}^m$ from $\mathscr{H}$ in the the empirical norm

$$\|g\|_m^2 := \frac{1}{m}\sum_{j=1}^{m}|g(x_j)|^2. \tag{64}$$

Typically, $\mathscr{H} = \mathscr{H}_m$ depends on a finite number $n = n(m)$ of parameters. Of course, we advocate the use of nonlinear families $\mathscr{H}_m$ for the reasons already made abundantly clear in this exposition. In some algorithms, the number n is chosen using an a priori assumption on f_ρ. Better algorithms avoid such prior assumptions and the number n is adapted to the data in the algorithm. This is usually done by what is called model selection in statistics but this can be sometimes be an expensive numerical procedure in practical implementations.

Estimates for the decay of the quantities in (62) are usually obtained under certain assumptions (called *priors*) on f_ρ. We emphasize that the algorithms should not depend on prior assumptions on f_ρ. Only in the analysis of the algorithms do we impose such prior assumptions in order to see how well the algorithm performs.

Priors on f_ρ are typically expressed by a condition of the type $f_\rho \in \Theta$ where Θ is a class of functions that necessarily must be contained in $L_2(X, \rho_X)$. If we wish the error, as measured in (62), to tend to zero as the number m of samples tends to infinity then we necessarily need that Θ is a compact subset of $L_2(X, \rho_X)$. There are three common ways to measure the compactness of a set Θ: (i) minimal coverings, (ii) smoothness conditions on the elements of Θ, (iii) the rate of approximation of the elements of Θ by a specific approximation process.

In studying the estimation of the regression function, the question arises at the outset as to what are the best approximation methods to use in deriving algorithms for approximating f_ρ and therefore indirectly in defining prior classes? With no additional knowledge of ρ (and thereby f_ρ) there is no general answer to this question. This is in contrast to numerical methods for PDEs where regularity theorems for the PDEs can lead to the optimal recovery schemes.

However, it is still possible in learning to draw some distinctions between certain strategies. Suppose that we seek to approximate f_ρ by the elements from a hypothesis class $\mathscr{H} = \Sigma_n$. Here the parameter n measures the complexity associated to the process. In the case of approximation by elements from linear spaces we will take the space Σ_n to be of dimension n. For nonlinear methods, the space Σ_n is not linear and now n represents the number of parameters used in the approximation.

If we have two approximation methods corresponding to sequences of approximation spaces (Σ_n) and (Σ_n'), then the second process would be superior to the first in terms of rates of approximation if $E_n'(g) \le CE_n(g)$ for all g and an absolute constant $C > 0$. For example, approximation using piecewise linear functions would in this sense be superior to using approximation by piecewise constants. In our learning context however, there are other considerations since: (i) the rate of approximation need not translate directly into results about estimating f_ρ because of the uncertainty in our observations, (ii) it may be that the superior approximation method is in fact much more difficult (or impossible) to implement in practice. For example, a typical nonlinear method may consist of finding an approximation to g from a family of linear spaces each of dimension N. The larger the family the more powerful the approximation method. However, too large of a family will generally make the numerical implementation of this method of approximation impossible.

Suppose that we have chosen the space Σ_n to be used as our hypothesis class $\mathscr{H}$ in the approximation of f_ρ from our given data $\mathbf{z}$. How should we define our approximation? As we have already noted, the most common approach is empirical risk minimization which gives the function $\hat{f}_\mathbf{z} := \hat{f}_{\mathbf{z},\Sigma_n}$ defined by (63). However, since we know $|f_\rho| \le M$, the approximation will be improved if we post-truncate $\hat{f}_\mathbf{z}$ by M. For this, we define the truncation operator

$$T_M(x) := \min(|x|, M)\operatorname{sign}(x) \tag{65}$$

for any real number x and define

$$f_\mathbf{z} := f_{\mathbf{z},\mathscr{H}} := T_M(\hat{f}_{\mathbf{z},\mathscr{H}}). \tag{66}$$

There are general results that provide estimates for how well $f_{\mathbf{z}}$ approximates f_ρ. One such estimate given in [35] (see Theorem 11.3) applies when $\mathscr{H}$ is a linear space of dimension n and gives

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim \frac{n\log(m)}{m} + \inf_{g\in\mathscr{H}} \|f_\rho - g\|^2. \tag{67}$$

The second term is the bias and equals our approximation error $E_n(f_\rho)$ for approximation using the elements of $\mathscr{H}$. The first term is the variance which bounds the error due to uncertainty. One can derive rates of convergence in expectation by balancing both terms (see [35] and [27]) for specific applications.

The deficiency of this approach is that one needs to know the behavior of $E_n(f_\rho)$ in order to choose the best value of n and this requires a priori knowledge of f_ρ. There is a general procedure known as model selection which circumvents this difficulty and tries to automatically choose a good value of n (depending on f_ρ) by introducing a penalty term. Suppose that $(\Sigma_n)_{n=1}^m$ is a family on linear spaces each of dimension n. For each $n = 1,2,\dots,m$, we have the corresponding function $f_{\mathbf{z},\Sigma_n}$ defined by (66) and the empirical error

$$\hat{E}_{n,\mathbf{z}} := \frac{1}{m}\sum_{j=1}^m (y_j - f_{\mathbf{z},\Sigma_n}(x_j))^2. \tag{68}$$

Notice that $E_{n,\mathbf{z}}$ is a computable quantity which we can view as an estimate for $E_n(f_\rho)$. In complexity regularization, one chooses a value of n by

$$n^* := n^*(\mathbf{z}) := \operatorname{argmin}\{E_{n,\mathbf{z}} + \frac{n\log m}{m}\}. \tag{69}$$

We now define

$$\hat{f}_{\mathbf{z}} := f_{\mathbf{z},\Sigma_{n^*}} \tag{70}$$

as our estimator to f_ρ. One can then prove (see Chapter 12 of [35]) that whenever f_ρ can be approximated to accuracy $E_n(f_\rho) \le Mn^{-s}$ for some $s > 0$, then

$$E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(X,\rho_X)}) \le C[\frac{(\log m)^2}{m}]^{\frac{2s}{2s+1}} \tag{71}$$

which save for the logarithm is an optimal rate estimation in expectation. For a certain range of s, one can also prove similar estimates in probability (see [27]). Notice that the estimator did not need to have knowledge of s and nevertheless obtains the optimal performance.

Model selection can also be applied in the setting of nonlinear approximation, i.e. when the spaces Σ_n are nonlinear but in this case, one needs to invoke conditions on the compatibility of the penalty with the complexity of the approximation process as measured by an entropy restriction. We refer the reader to Chapter 12 of [35] for a more detailed discussion of this topic

Let us also note that the penalty approach is not always compatible with the practical requirement of *on-line* computations. By on-line computation, we mean that the estimator for the sample size m can be derived by a simple update of the estimator for the sample size $m-1$. In penalty methods, the optimization problem needs to be globally re-solved when adding a new sample. However, when there is additional structure in the approximation process such as the adaptive partitioning, then there are algorithms that circumvent this difficulty.

With Wolfgang, we wanted to develop algorithms based on nonlinear piecewise polynomials which are universally optimal and in addition are numerically easy to implement. Our first paper [9] built such an algorithm based on piecewise constant approximation. Its implementation is very simple (wavelet like) and can be done on line with streaming data. We proved theorems which showed the optimality of this algorithm in terms of the desirable probability estimates.

While proving the results in [9], we were puzzled by the fact that these results did not carry over nontrivially to general piecewise polynomials. Through a family of counterexamples, we found that if we wanted estimators which perform well in probability then either we must assume something more about the underlying probability measure ρ or we must find an alternative to empirical risk minimization. The simplest way out of this dilemma was to use post truncation as described in (66). Using this type of truncation, we developed in [7] optimal adaptive partitioning learning algorithms for arbitrary polynomial degrees and proved their universal optimality.

6.1 Learning with greedy algorithms

We have already emphasized that keeping the computational task reasonable in learning algorithms is a significant issue. For this reason, with Wolfgang we studied the application of greedy algorithms for learning. The main goal of our extension of the theory of greedy algorithms, as discussed in §3.4 was to apply these to the learning problem. Indeed, we built an estimator based on the application of the OGA or RGA to the noisy data (y_i) in the Hilbert space defined by the empirical norm

$$\|f\|_n := \frac{1}{n}\sum_{i=1}^{n} |f(x_i)|^2, \tag{72}$$

and its associated inner product. At each step k, the algorithm generates an approximation $\hat{f}_k$ to the data. Our estimator was then defined by

$$\hat{f} := T\hat{f}_{k^*} \tag{73}$$

where T is the truncation operator (65) and the value of k^* is selected by a complexity regularization procedure. Our main result for this estimator was (roughly) that when the regression function f_ρ is in $\mathscr{B}_p$ (where this space is defined with respect

to the norm $\|u\|^2 := E(|u(x)|^2)$ as in §3.4), the estimator has convergence rate

$$E(\|\hat{f} - f_\rho\|^2) \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2s}{1+2s}}, \tag{74}$$

again with $s := 1/p - 1/2$. In the case where $f_\rho \in \mathscr{L}_1$, we obtain the same result with $p = 1$ and $s = 1/2$. We also show that this estimator is universally consistent.

In order to place these results into the current state of the art of statistical learning theory, let us first remark that similar convergence rate for the denoising and the learning problem could be obtained by a more "brute force" approach which would consist in selecting a proper subset of $\mathscr{D}$ by complexity regularization with techniques such as those in [2] or Chapter 12 of [35]. Following for instance the general approach of [35], this would typically first require restricting the size of the dictionary $\mathscr{D}$ (usually to be of size $O(n^a)$ for some $a > 1$) and then considering all possible subsets $\Lambda \subset \mathscr{D}$ and spaces $\mathscr{G}_\Lambda := \mathrm{Span}\{g \in \Lambda\}$, each of them defining an estimator

$$\hat{f}_\Lambda := T\left(\mathrm{Argmin}_{f \in \mathscr{G}_\Lambda} \|y - f\|_n^2\right) \tag{75}$$

The estimator $\hat{f}$ is then defined as the $\hat{f}_\Lambda$ which minimizes

$$\min_{\Lambda \subset \mathscr{D}} \{\|y - \hat{f}_\Lambda\|_n^2 + \mathrm{Pen}(\Lambda, n)\} \tag{76}$$

with $\mathrm{Pen}(\Lambda, n)$ a complexity penalty term. The penalty term usually restricts the size of Λ to be at most $\mathscr{O}(n)$ but even then the search is over $O(n^{an})$ subsets. In some other approaches, the sets $\mathscr{G}_\Lambda$ might also be discretized, transforming the subproblem of selecting $\hat{f}_\Lambda$ into a discrete optimization problem.

The main advantage of using the greedy algorithm in place of (76) for constructing the estimator is a dramatic reduction of the computational cost. Indeed, instead of considering all possible subsets $\Lambda \subset \mathscr{D}$ the algorithm only considers the sets $\Lambda_k := \{g_1, \cdots, g_k\}$, $k = 1, \ldots, n$, generated by the empirical greedy algorithm. This approach was proposed and analyzed in [41] using a version of the RGA in which

$$\alpha_k + \beta_k = 1 \tag{77}$$

which implies that the approximation f_k at each iteration stays in the convex hull $\mathscr{C}_1$ of $\mathscr{D}$. The authors established that if f does not belong to $\mathscr{C}_1$, the RGA converges to its projection onto $\mathscr{C}_1$, In turn, the estimator was proved to converge in the sense of (74) to f_ρ, with rate $(n/\log n)^{-1/2}$, if f_ρ lies in $\mathscr{C}_1$, and otherwise to its projection onto $\mathscr{C}_1$. In that sense, this procedure is not universally consistent.

Our main contribution in the work with Wolfgang was to remove requirements of the type $f_\rho \in \mathscr{L}_1$ when obtaining convergence rates. In the learning context, there is indeed typically no advanced information that would guarantee such restrictions on f_ρ. The estimators that we construct for learning are now universally consistent and have provable convergence rates for more general regression functions described by means of interpolation spaces. One of the main ingredient in our analysis of the performance of our greedy algorithms in learning is a powerful exponential con-

centration inequality which was introduced in [41]. Let us mention that a closely related analysis, which however does not involve interpolation spaces, was developed in [5, 6].

Let us finally mention that there exist some natural connections between the greedy algorithms which we have discussed and other numerical techniques for building a sparse approximation in the dictionnary based on the minimization of an ℓ_1 criterion. In the statistical context, these are the celebrated LASSO [52, 36] and LARS [33] algorithms. The relation between ℓ_1 minimization and greedy selection is particularly transparent in the context of deterministic approximation of a function f in an orthonormal basis: if we consider the problem of minimizing

$$\|f - \sum_{g\in\mathscr{D}} d_g g\|^2 + t \sum_{g\in\mathscr{D}} |d_g| \tag{78}$$

over all choices of sequences (d_g), we see that it amounts in minimizing $|c_g - d_g|^2 + t|d_g|$ for each individual g, where $c_g := \langle f, g\rangle$. The solution to this problem is given by the *soft thresholding* operator

$$d_g := c_g - \frac{t}{2}\mathrm{sign}(c_g) \text{ if } |c_g| > \frac{t}{2}, \ 0 \text{ else}, \tag{79}$$

and is therefore very similar to picking the largest coefficients of f.

7 Compressed sensing

Compressed sensing came into vogue during the last few years but its origins lie in results from approximation and functional analysis dating back to the 1970's. The primary early developers were Kashin [39] and Gluskin [34]. Donoho [30] and Candés and Tao [14] showed the importance of this theory in signal processing and added substantially to the theory and its numerical implementation, especially how to do decoding in a practical way.

In discrete compressed sensing, we want to capture a vector (signal) $x \in \mathbb{R}^N$ with N large. Of course if we make N measurements we will know x exactly. The problem is to make comparably fewer measurements and still have enough information to accurately recover x. Since the subject is intimately intertwined with sparsity and nonlinear approximation, the problems of compressed sensing immediately peaked our interest.

The m measurements we are allowed to make about x are of the form of an inner product of x with prescribed vectors. These measurements are represented by a vector

$$y = \Phi x, \tag{80}$$

of dimension $m < N$, where Φ is an $m \times N$ measurement matrix (called a CS matrix). To extract the information that the measurement vector y holds about x, one uses a decoder Δ which is a mapping from $\mathbb{R}^m$ into $\mathbb{R}^N$. The vector $x^* := \Delta(y) = \Delta(\Phi x)$

is our approximation to x extracted from the information y. In contrast to Φ, the operator Δ is allowed to be non-linear.

In recent years, considerable progress has been made in understanding the performance of various choices of the measurement matrices Φ and decoders Δ. Although not exclusively, by far most contributions focus on the ability of such an encoder-decoder pair (Φ,Δ) to recover a *sparse* signal. For example, a typical theorem says that there are pairs (Φ,Δ) such that whenever $x \in \Sigma_k$, with $k \le am/\log(N/k)$, then $x^* = x$.

Our view was that from both a theoretical and a practical perspective, it is highly desirable to have pairs (Φ,Δ) that are robust in the sense that they are effective even when the vector x is not assumed to be sparse. The question arises as to how we should measure the effectiveness of such an encoder-decoder pair (Φ,Δ) for non-sparse vectors. In [18] we have proposed to measure such performance in a metric $\|\cdot\|_X$ by the largest value of k for which

$$\|x-\Delta(\Phi x)\|_X \le C_0\sigma_k(x)_X, \quad \forall x \in I\!R^N, \tag{81}$$

with C_0 a constant independent of k,n,N. We say that a pair (Φ,Δ) which satisfies property (81) is *instance-optimal* of order k with constant C_0. It was shown that this measure of performance heavily depends on the norm employed to measure error. Let us illustrate this by two contrasting results from [18]:

(i) If $\|\cdot\|_X$ is the ℓ_1-norm, it is possible to build encoding-decoding pairs (Φ,Δ) which are instance-optimal of order k with a suitable constant C_0 whenever $m \ge ck\log(N/k)$ provided c and C_0 are sufficiently large. Moreover, the decoder Δ can be taken as

$$\Delta(y) := \operatorname*{argmin}_{\Phi z=y} \|z\|_{\ell_1}. \tag{82}$$

Therefore, in order to obtain the accuracy of k-term approximation, the number m of non-adaptive measurements need only exceed the amount k of adaptive measurements by the small factor $c\log(N/k)$. We shall speak of the range of k which satisfy $k \le am/\log(N/k)$ as the *large range* since it is the largest range of k for which instance-optimality can hold.

(ii) In the case $\|\cdot\|_X$ is the ℓ_2-norm, if (Φ,Δ) is any encoding-decoding pair which is instance-optimal of order $k=1$ with a fixed constant C_0, then the number of measurement m is always larger than aN, where $a>0$ depends only on C_0. Therefore, the number of non-adaptive measurements has to be very large in order to compete with even one single adaptive measurement.

The matrices Φ which have the largest range of instance-optimality for ℓ_1 are all given by stochastic constructions. Namely, one creates an appropriate random family $\Phi(\omega)$ of $m\times N$ matrices on a probability space (Ω,ρ) and then shows that with high probability on the draw, the resulting matrix $\Phi=\Phi(\omega)$ will satisfy instance-optimality for the large range of k. There are no known deterministic constructions. The situation is even worse in the sense that given an $m\times N$ matrix Φ there is no simple method for checking its range of instance-optimality.

While the above results show that instance-optimality is not a viable concept in ℓ_2, it turns out that the situation is not as bleak as it seems. For example, a more optimistic result was established by Candes, Romberg and Tao in [15]. They show that if $m \geq ck\log(N/k)$, it is possible to build pairs (Φ, Δ) such that for all $x \in \mathbb{R}^N$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \frac{\sigma_k(x)_{\ell_1}}{\sqrt{k}}, \tag{83}$$

with the decoder again defined by (82). This implies, in particular, that k-sparse signals are exactly reconstructed and that signals x in the space weak ℓ_p with $\|x\|_{w\ell_p} \leq M$ for some $p < 1$ are reconstructed with accuracy $C_0 M k^{-s}$ with $s = 1/p - 1/2$. This bound is of the same order as the best estimate available on $\max\{\sigma_k(x)_{\ell_2} : \|x\|_{w\ell_p} \leq M\}$. Of course, this result still falls short of instance-optimality in ℓ_2 as it must.

What intrigued us was that instance-optimality can be attained in ℓ_2 if one accepts a probabilistic statement. A first result in this direction, obtained by Cormode and Mutukrishnan in [22], shows how to construct random $m \times N$ matrices $\Phi(\omega)$ and a decoder $\Delta = \Delta(\omega)$, $\omega \in \Omega$, such that for any $x \in \mathbb{R}^N$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2} \tag{84}$$

holds with overwhelming probability (larger than $1 - \varepsilon(m)$ where $\varepsilon(m)$ tends rapidly to 0 as $m \to +\infty$) as long as $k \leq am/(\log N)^{5/2}$ with a suitably small. Note that this result says that given x, the set of $\omega \in \Omega$ for which (84) fails to hold has small measure. This set of failure will depend on x.

From our viewpoint, *instance-optimality in probability* is the proper formulation in ℓ_2. Indeed, even in the more favorable setting of ℓ_1, we can never put our hands on matrices Φ which have the large range of instance-optimality. We only know with high probability on the draw, in certain random constructions, that we can attain instance-optimality. So the situation in ℓ_2 is not that much different from that in ℓ_1.

The results in [18] pertaining to instance-optimality in probability asked two fundamental questions: (i) can we attain instance-optimality for the largest range of k, i.e. $k \leq an/\log(N/k)$, and (ii) what are the properties of random families that are needed to attain this performance. We showed that instance-optimality can be obtained in the probabilistic setting for the largest range of k, i.e. $k \leq an/\log(N/k)$ using quite general constructions of random matrices. Namely, we introduced two properties for a random matrix Φ which ensure instance-optimality in the above sense and then showed that these two properties hold for rather general constructions of random matrices (such as Gaussian and Bernoulli). However, one shortcoming of the results in [18] is that the decoder used in establishing instance-optimality was defined by minimizing $\|y - \Phi x\|_{\ell_2}$ over all k-sparse vectors, a task which cannot be achieved in any reasonable computational time.

This led us to consider other possible decoders which are numerically friendly and can be coupled with standard constructions of random matrices to obtain an encoding/decoding pair which is instance-optimal for the largest range of k. There are two natural classes of decoders.

The first is based on ℓ_1 minimization as described in (82). It was a nontrivial argument given by Przemek Wojtaczszyk [54] that this decoder gives ℓ_2 instance optimality in probability when coupled with random Gaussian matrices. The key feature of his proof was the fact that such an $m \times N$ Gaussian matrix maps the unit ball in ℓ_1^N onto a set that contains the ball of radius $\frac{\log(N/m)}{m}$ in ℓ_2^m.

The above mapping property fails to hold for general random matrices. For example for the Bernouli family, any point that maps into the vector $e_1 = (1,0,\dots,0)$ must have ℓ_1^N norm $\geq \sqrt{n}$. So some new ideas were needed to prove instance optimality in probability for general random families. This is provided by new mapping properties which state that the image of the unit ℓ_1^N ball covers a certain clipped ℓ_2^N ball. These remarkable mapping properties were first proved in [42] and rediscovered in [28] where the instance optimality is proved.

The other natural decoders for compressed sensing are greedy algorithms. The idea to apply greedy algorithms for compressed sensing originated with Gilbert and Tropp [53] who proposed to use the orthogonal greedy algorithm or orthogonal matching pursuit (OMP) in order to decode y. Namely, the greedy algorithm is applied to the dictionary of column vectors of Φ and the input vector y. After k iterations, it identifies a set of Λ of k column indices (those corresponding to the vectors used to approximate y by the greedy algorithm. Once the set Λ is found, we decode y by taking the minimizer of $\|y - \Phi(z)\|_{\ell_2}$ among all z supported on Λ. The latter step is least squares fitting of the residual and is very fast.

These authors proved the following result for a probabilistic setting for general random matrices which include the Bernouli and Gaussian families: if $m \geq ck\log N$ with c sufficiently large, then for any k sparse vector x, the OMP algorithm returns exactly $x^k = x$ after k iterations, with probability greater than $1 - N^{-b}$ where b can be made arbitrarily large by taking c large enough.

Decoders like OMP are of high interest because of their efficiency. The above result of Gilbert and Tropp remains as the only general statement about OMP in the probabilistic setting. A significant breakthrough on decoding using greedy pursuit was given in the paper of Needel and Vershynin [46] (see also their followup [47]) where they showed the advantage of adjoining a batch of coordinates at each iteration rather than just one coordinate as in OMP. They show that such algorithms can deterministically capture sparse vectors for a slightly smaller range than the largest range of k.

With Wolfgang, we were interested in whether decoders based on thresholding could be used as decoders to yield ℓ_2 instance-optimality in probability for general families of random matrices for the large range of k. In [19] we give an algorithm which does exactly that. This algorithm adds a batch of coordinates at each iteration and then uses a thinning procedure to possibly remove some of them at later iterations. Conceptually, one thinks in terms of a bucket holding all of the coordinates to be used in the construction of x. In the analysis of such algorithms it is important to not allow more than a multiple of k coordinates to gather in the bucket. The thinning is used for this purpose. Thinning is much like the coarsening used in PDE solvers which we described earlier Our algorithm is similar in nature to the COSAMP algorithm of Needel and Tropp [45].

8 Final thoughts

As has been made abundantly clear in this brief survery, Wolfgang Dahmen's contributions to both the theory of nonlinear approximation and to its application in a wide range of domains has been pervasive. Fortunately, the story is still going strong and I am happy to be going along for the ride.

References

1. I. Babuska and M. Vogelius, *Feedack and adaptive finite element solution of one dimensional boundary value problems*, Num. Math., **44**(1984), 75–102.
2. A. Barron, *Complexity regularization with application to artificial neural network*, in *Nonparametric functional estimation and related topics*, G. Roussas (ed.), 1990, 561–576, Kluwer Academic Publishers.
3. A. Barron, *Neural net approximation*, Proc 7th Yale Workshop on Adaptive and Learning Systems, K.S. Narendra Ed, New Haven, CT, 1992, pp. 69–72.
4. A. Barron, *Universal approximation bounds for superposition of n sigmoidal functions*, IEEE Trans. Inf. Theory, **39**(1993), 930–945.
5. A. Barron and G. Cheang *Penalized least squares, model selection, convex hull classes, and neural nets*, in Verleysen, M. (editor). Proceedings of the 9th ESANN, Brugge, Belgium, De-Facto press, 2001. pp. 371-376.
6. A. Barron, and G.H.L. Cheang *Risk bounds and greedy computations for penalized least squares, model selection, and neural networks* , Preprint, Department of Statistics, Yale University.
7. A. Barron, A. Cohen, W. Dahmen and R. DeVore, *Approximation and learning by greedy algorithms*, Annals of Statistics, **36**(2008), 64–94.
8. C. Bennett and R. Sharpley, *Interpolation of Operators*, in Pure and Applied Mathematics, 1988, Academic Press, N.Y.
9. P. Binev, A. Cohen, W. Dahmen, R. DeVore and V. Temlyakov, *Universal Algorithms for Learning Theory Part I: Piecewise Constant Functions*, J. Machine Learning, **6**(2005), 1297–1321.
10. P. Binev, W. Dahmen, and R. DeVore, *Adaptive Finite Element Methods with Convergence Rates*, Numerische Mathematik, **97**(2004), 219–268.
11. P. Binev, W. Dahmen, R. DeVore, and P. Petrushev, *Approximation Classes for Adaptive Methods*, Serdica Math. J., **28**(2002), 391–416.
12. P. Binev and R. DeVore, *Fast Computation in Adaptive Tree Approximation*, Num. Math., **97**(2004), 193–217.
13. M. Birman and M. Solomjak, *Piecewise polynomial approximations of functions of the classes* W_p^α, Mat. Sbornik, **73**(1967), 331–355.
14. E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inf. Theory, **51**(2005), 4203–4215.
15. E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure and Appl. Math., **59**(2006), 1207–1223.
16. A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: convergence rates*, Math. Comp., **70**(2000), 27–75.
17. A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for operator equations: beyond the elliptic case*, J. FoCM, **2**(2002), 203–245.
18. A. Cohen, W. Dahmen and R. DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc., **22**(2009), 211–231.

19. A. Cohen, W. Dahmen and R. DeVore, *Instance Optimal Decoding by Thresholding in Compressed Sensing*, Contemporary Math., to appear.
20. A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, *Tree Approximation and Encoding*, ACHA, **11**(2001), 192–226.
21. A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard, *On the importance of combining wavelet based non-linear approximation in coding strategies*, IEEE Trans. Inf. Th., **48**(2002), 1895–1921.
22. G. Cormode and S. Muthukrishnan, *Towards an algorithmic theory of compressed sensing*, Technical Report 2005-25, DIMACS, 2005. Graham Cormode, S. Muthukrishnan: Combinatorial Algorithms for Compressed Sensing. SIROCCO 2006: 280–294.
23. S. Dahlke and R. DeVore, *Besov regularity for elliptic boundary value problems*, Communication in PDE's, **22**(1997), 1–16.
24. W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Num. Math., **63**(1992), 315–344.
25. R. DeVore, *Nonlinear approximation*, Acta Numerica, **7**(1998), 51–150.
26. R. DeVore, B. Jawerth and B. Lucier, *Image compression through transform coding*, IEEE Proceedings on Information Theory, **38**(1992), 719–746.
27. R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov, *On Mathematical Methods for Supervised Learning*, J. of FOCM, **6**(2006), 3–58.
28. R. DeVore, G. Petrova and P. Wojtaszczyk, *Instance-Optimality in Probability with an ℓ_1-minimization decoder*, ACHA, to appear.
29. R. DeVore and V. Temlyakov, *Some remarks on greedy algorithms*, Advances in Computational Mathematics, **5**(1996), 173–187.
30. D. Donoho, *Compressed Sensing*, IEEE Trans. Information Theory, **52**(2006), 1289–1306.
31. Donoho, D.L. and I.M. Johnstone, *Minimax Estimation via Wavelet shrinkage*, Annals of Statistics, **26**(1998), 879–921.
32. W. Dörfler, *A convergent adaptive scheme for Poisson's equation*, SIAM J. Num. Analysis, **33**(1996), 1106–1124.
33. B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Least angle regression*, Ann. Statist., **32**(2004), 407–499
34. A. Garnaev, E. Gluskin, *The widths of a Euclidean ball*, Doklady AN SSSR, **277**(1984), 1048–1052.
35. L. Györfy, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*, 2002, Springer Verlag, Berlin.
36. T. Hastie, R. Tibshirani and J. Friedman *The Elements of Statistical Learning*, 2001, Springer.
37. H. Johnen and K. Scherer, *On the equivalence of the K-functional and moduli of continuity and some applications*, in *Constr. Theory of functions of several variables* , Proc. Conf. Oberwolfach 1976, Springer Lecture Notes 571), 119–140.
38. L. Jones, *A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training*, Annals of Statistics, **20**(1992), 608–613.
39. B. Kashin, *The widths of certain finite dimensional sets and classes of smooth functions*, Izvestia, **41**(1977), 334–351.
40. S. Konyagin and V. Temlyakov, *Rate of convergence of Pure greedy Algorithm*, East J. Approx., **5**(1999), 493–499.
41. W. Lee, P. Bartlett and R. Williamson, *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Trans. Inf. Theory, **42**(1996), 2118–2132.
42. A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Advances in Math., **195**(2005), 491–523.
43. Livshitz, E.D. and V.N. Temlyakov, *Two lower estimates in greedy approximation*, Constr. Approx., **19**(2003), 509–524.
44. P. Morin, R. Nochetto, K. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. of Num. Anal., **38**(2000), 466–488.
45. D. Needell and J. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, ACHA, to appear.

46. D. Needell and R. Vershynin, *Uniform Uncertainty Principle and signal recovery via Regularized Orthogonal Matching Pursuit*, J. of FOCM, **9**(2009), 317–334.
47. D. Needell and R. Vershynin, *Signal Recovery from Inaccurate and Incomplete Measurements via Regularized Orthogonal Matching Pursuit*, preprint, 2007,
48. R. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Num. Anal., **41**(2003), 1074–1100.
49. R. Stevenson, *An optimal adaptive finite element method*, SIAM J. Numer. Anal., **42**(2005), 2188–2217.
50. V. Temlyakov, *The best m-term approximation and greedy algorithms*, Adv. Comput. Math., **8**(1998), 249–265.
51. V. Temlyakov, *Nonlinear methods of approximation*, J. of FOCM, **3**(2003), 33–107.
52. R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society, Series B, **58**(1995), 267–288.
53. J. Tropp and A. Gilbert, *Signal recovery from random measurements via Orthogonal Matching Pursuit*, IEEE Trans. Info. Theory, **53**(2007), 4655–4666.
54. P. Wojtaszczyk, *Stability and instance optimality for Gaussian measurements in compressed sensing*, J. of FOCM, to appear

Univariate subdivision and multi-scale transforms: The nonlinear case

Nira Dyn and Peter Oswald

1 Introduction

Over the past 25 years, fast multi-scale algorithms such as wavelet-type pyramid transforms for hierarchical data representation, multi-grid solvers for the numerical solution of operator equations, and subdivision methods in computer-aided geometric design lead to tremendous successes in data and geometry processing, and in scientific computing in general. While linear multi-scale analysis is in a mature state [10, 18, 26, 15, 12, 23], not so much is known in the nonlinear case. Nonlinearity arises naturally, e.g. in data-adaptive algorithms, in image and geometry processing, robust de-noising, or due to nonlinear constraints on the analyzed objects themselves that need to be preserved on all scales.

For illustration, and to guide our further discussions, let us introduce three univariate examples of nonlinear multi-scale transforms that have played a central role in the development of the emerging theory.

Example 1: (W)ENO multi-scale transforms for piecewise smooth functions. Adaption of data representations to jump discontinuities is motivated by applications to hyperbolic PDEs, and serves as simplified model for developing edge-adaptive algorithms in image analysis. Motivated by his work on essentially non-oscillatory (ENO) schemes for numerically solving hyperbolic conservation laws, A. Harten [37, 38, 6] introduced ENO schemes for adaptive multi-scale data representation. For simplicity, assume that a piecewise smooth function $f \in L_\infty(\mathbb{R})$ is sampled at dyadic points, and represented by data vectors $v^j \in \ell_\infty(\mathbb{Z})$ with entries $v_i^j = f(i2^{-j})$ corresponding to uniform grids Γ^j of sampling step-size 2^{-j}.

Nira Dyn
Tel Aviv University, School of Mathematical Sciences, Tel Aviv, Israel,
e-mail: niradyn@math.tau.ac.il

Peter Oswald
Jacobs University Bremen, School of Engineering and Science, D-28759 Bremen, Germany,
e-mail: p.oswald@jacobs-university.de

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_7,

For smooth f, a very popular way to encode the whole sequence $\{v^j\}_{j\ge 0}$ is the use of the cubic Deslauriers-Dubuc wavelet transform (sometimes called the standard linear interpolating 4-point scheme, see [20, 21]), where v^j is split into v^{j-1} and the j-th detail sequence d^j given by $d^j = v^j - Sv^{j-1}$, where the linear operator $S:\ \ell_\infty(\mathbb{Z}) \to \ell_\infty(\mathbb{Z})$ (called prediction or *subdivision operator*) is given by

$$\begin{aligned} (Sv)_{2i} &= v_i, \\ (Sv)_{2i+1} &= -\tfrac{1}{16}v_{i-1} + \tfrac{9}{16}v_i + \tfrac{9}{16}v_{i+1} - \tfrac{1}{16}v_{i+2}, \end{aligned} \qquad i \in \mathbb{Z}. \tag{1}$$

Indeed, knowing $\{v^0, d^j\}_{j\ge 1}$ allows us to recursively reconstruct the original sequence $\{v^j\}_{j\ge 0}$. Obviously, $d^j_{2i} = 0$ implies that only odd-indexed entries of d^j need to be stored, and storing v^{j-1} and d^j as floating-point numbers is as expensive as storing v^j. Thus, the transform and its finite realizations

$$\{v^j\}_{j=0}^J \qquad \longleftrightarrow \qquad \{v^0, d^j\}_{j=1}^J, \qquad J \ge 1,$$

belong to the class of non-expansive $1-1$ *multi-scale transforms*.

The above S has some properties that are characteristic for most of the multi-scale transforms and are key to their analysis: S is *local* (i.e., data associated with a grid point of Γ^j are predicted from data associated with finitely many grid points of Γ^{j-1} close to it), and *r-shift invariant* with dilation factor $r = 2$. The latter property can be formalized by the operator identity $ST_k = T_{rk}S$, where T_k is the shift-operator given by $(T_k v)_i = v_{i+k}$, $i \in \mathbb{Z}$. Another property that is central to the subject is the *polynomial reproduction* and, closely related, *approximation order* of S. Detailed definitions will be given later. For the above S it is well-known that it reproduces cubic polynomials because the formula for $(Sv^{j-1})_{2i+1}$ comes from interpolating the four data $\{v^{j-1}_s\}_{s=i-1,\dots,i+2}$ at the corresponding sub-grid of Γ^{j-1} by a cubic polynomial p_i, and evaluating its value at the point $(i+\frac{1}{2})2^{-j+1}$ of Γ^j central to them. As a result, for smooth f the ℓ_p norms of the detail sequences d^j decay at a rate 2^{-4j}. Thus, if representation is required up to a certain accuracy only then fewer bits are necessary to encode the detail information.

This savings effect is to some extent lost when jump discontinuities are present. A remedy is to detect potential jump discontinuities from the data v^{j-1}, and use a smarter, data-dependent and thus nonlinear, prediction rule. For ENO schemes, one chooses the "least oscillating" among the interpolating cubic polynomials p_{i-1}, p_i, p_{i+1} for assigning an appropriate value corresponding to the point $(i+\frac{1}{2})2^{-j+1}$. The effect of this modification is illustrated in Fig. 1, the nonlinear scheme obviously suppresses the spurious oscillations associated with the Gibbs phenomenon for linear wavelet-type transforms, and reduces the number of large detail entries d^j_i near the jump discontinuity.

Weighted essentially non-oscillatory (WENO) transforms use a convex combination of the three predictions, with weights smoothly depending on the measured oscillations. Instead of interpreting the entries v^j_i as values of f at dyadic points $i2^{-j}$, one could equally well interpret them as averages on dyadic intervals $(i2^{-j}, (i+1)2^{-j})$. In this case other subdivision operators S would be preferable for symmetry reasons, and the restriction would be more naturally defined by averaging

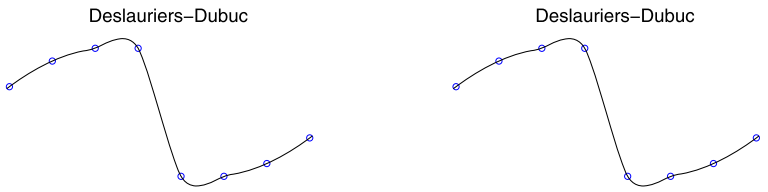


Fig. 1 While the linear Deslauriers-Dubuc subdivision scheme produces overshoots near a jump (left), the nonlinear ENO scheme avoids artifacts (right) but produces a less smooth limit near the jump. The coarse grid data points are indicated by circles

$v_i^{j-1} = (v_{2i}^j + v_{2i+1}^j)/2$ rather than by simple down-sampling. The convergence, limit smoothness, and stability properties of these schemes have been systematically studied in [13] within a framework of *quasi-linear, data-dependent subdivision* which will be reviewed in below. Functional limits of the sequence $\{v^j\}_{j\geq 0}$ are understood in the usual way as limits of an associated sequence $\{f^j\}_{j\geq 0}$, where the function $f^j \in C(\mathbb{R})$ is typically defined as the linear spline interpolant to the data (Γ^j, v^j). We note that there exist other families of nonlinear schemes of a similar flavor, such as monotonicity and convexity preserving schemes, PPH, and power-p schemes, see e.g. [43, 46, 5, 2].

Example 2: Median-interpolating schemes for robust de-noising. In [22], motivated by applications to heavy-tail, non-Gaussian noise removal, a nonlinear multi-scale transform with dilation factor $r = 3$ was introduced where point evaluations and linear averaging operations are systematically replaced by median calculations. More precisely, let us assume that the entries v_i^J of the fine-scale data vector v^J represent noisy measurements of average values on triadic intervals $I_i^J = (i3^{-J}, (i+1)3^{-J})$ of a smooth function f. Then, [22] uses the rule

$$v_i^{j-1} = \mathrm{med}(v_{3i}^j, v_{3i+1}^j, v_{3i+2}^j), \qquad i \in \mathbb{Z}, \qquad j = 1, \ldots, J, \tag{2}$$

to define coarse-scale representations of the measured data (the median of three numerical values is defined in the obvious way). Detail sequences d^j are formally defined by $d^j = v^j - Sv^{j-1}$, $j = 1, \ldots, J$, as before but using a new type of nonlinear *median-interpolating subdivision operator* S. To define it, recall that the values $\{v_s^{j-1}\}_{s=i-1,i,i+1}$ can be interpreted as approximate coarse-scale medians of f on the three consecutive intervals I_s^{j-1}, $s = i-1, i, i+1$. It turns out that there is a unique quadratic polynomial p_i whose median with respect to these three intervals concides with the given v_s^{j-1}, $s = i-1, i, i+1$, i.e.,

$$p_i(t) = At^2 + Bt + C \ : \qquad v_s^{j-1} = \mathrm{med}(p_i; I_s^{j-1}), \quad s = i-1, i, i+1. \tag{3}$$

Then, for the three subintervals of I_i^{j-1} with respect to the next triadic grid, set

$$(Sv^{j-1})_s := \mathrm{med}(p_i; I_s^j), \qquad s = 3i, 3i+1, 3i+2. \tag{4}$$

Note that this subdivision scheme is closely related to a linear scheme which one obtains if one replaces the median conditions in both the interpolation step (3) and the imputation step (4) by evaluation at the corresponding interval midpoints (for a monotone continuous function f on an interval I, the median indeed coincides with the value of f at the midpoint of I). The idea of studying nonlinear subdivision processes and multi-scale transforms by relating them to close-by linear schemes, and using perturbation arguments is very fruitful, and has been followed by various authors [25, 65, 62, 61, 19].

Fig. 2 The rules for the nonlinear median-interpolating subdivision scheme (left), and the linear midpoint-interpolating scheme (right) are close but not identical

The convergence and smoothness properties of the limits of the median-interpolating subdivision process have been studied in a series of papers [22, 52, 64], for the stability of the associated multi-scale transform, see [36]. The remarkable paper [64] solves the smoothness problem in the Hölder scale, and is based on a detailed analysis of associated nonlinear dynamical systems. Various extensions of the median-interpolating multi-scale transform have been proposed as well: one can consider higher-order median-interpolation [30], other robust estimators [53] or nonlinear interpolation conditions [51, 65].

Example 2 is an *expansive multi-scale transform*. Indeed, an easy calculation shows that the 3^J data per unit interval to be stored for v^J are replaced by $1+3+\ldots+3^J \approx 3^{J+1}/2$ data to be stored for $\{v^0, d^j\}_{j=1}^J$ resulting in an increase of storage requirements by a factor $3/2$. Expansive multi-scale transforms occur also if linear frame representations are explored, and offer sometimes even some advantages (e.g., robustness with respect to erasures in the case of frame decompositions).

Example 3: Normal multi-resolution for efficient geometry compression. While the previous examples serve scalar data associated with the (appropriately interpreted) samples of a function with respect to a uniform grid (a situation which we call *functional setting*), in geometry processing there is no fixed or natural parametrization of a geometric object by a function, and finding an appropriate parametrization is often part of the processing task. Normal multi-resolution is a remarkable example of a nonlinear multi-scale transform that has originally been developed for surface compression [41, 34] and image analysis [39], works directly on vector data, and cannot be reduced to the scalar case. It serves as an example for what we call *geometric nonlinear multi-scale transforms*. To reveal the main

idea, we describe the scheme for representing a closed smooth curve $\mathscr{C}$ embedded into $\mathbb{R}^2$ for $r = 2$ following [19, 55]. The scalar-valued data vectors $v^j \in \ell_\infty(\mathbb{Z})$ from the previous examples will be replaced by $\mathbb{R}^2$-valued periodic sequences $\mathbf{v}^j$ of length $2^j n_0$ with $n_0 \geq 3$, which consist of points $\mathbf{v}_i^j$ on $\mathscr{C}$. Periodicity means that $\mathbf{v}_{i+2^j n_0}^j \equiv \mathbf{v}_i^j$ for all $i \in \mathbb{Z}$. The analysis step of the normal multi-resolution scheme starts with a sufficiently dense sampling $\mathbf{v}^0$ of $n_0 \geq 3$ curve points ordered such that the polygonal line obtained by connecting consecutive $\mathbf{v}_i^0$ by straight line segments is a faithful approximation to $\mathscr{C}$. To construct $\mathbf{v}^j$ from $\mathbf{v}^{j-1}$, we keep the curve points from $\mathbf{v}^{j-1}$ by setting $\mathbf{v}_{2i}^j = \mathbf{v}_i^{j-1}$ for all $i \in \mathbb{Z}$, and insert new points $\mathbf{v}_{2i+1}^j \in \mathscr{C}$ by first predicting "base points" $\hat{\mathbf{v}}_{2i+1}^j$ using any reasonable interpolating subdivision operator S, i.e., $\hat{\mathbf{v}}^j = S\mathbf{v}^{j-1}$. The point $\mathbf{v}_{2i+1}^j$ is obtained by intersecting the normal to the edge vector $\mathbf{e}_i^{j-1} := \mathbf{v}_{i+1}^{j-1} - \mathbf{v}_i^{j-1}$ through the base point $\hat{\mathbf{v}}_{2i+1}^j$ with $\mathscr{C}$. We do not dwell on implementation aspects such as the subtle issue of which curve point to select if there are many intersection points. What needs to be stored as entry d_i^j of the detail sequence d^j is the signed distance of $\mathbf{v}_{2i+1}^j$ from the base point $\hat{\mathbf{v}}_{2i+1}^j$. The reconstruction (or synthesis) step $\{\mathbf{v}^0, d^j\}_{j=1}^J \longmapsto \mathbf{v}^J$ is recursively given by

$$\mathbf{v}_{2i}^j = \mathbf{v}_i^{j-1}, \qquad \mathbf{v}_{2i+1}^j = (S\mathbf{v}^{j-1})_{2i+1} + d_i^j \mathbf{n}_i^{j-1}, \qquad i = 0, \ldots, 2^{j-1} n_0,$$

for $j = 1, \ldots, J$, where $\mathbf{n}_i^{j-1}$ denotes the unit normal vector to the edge vector $\mathbf{e}_i^{j-1}$. This time, the nonlinearity is hidden in the normal computation for the detail update, and not in the subdivision part as in the previous examples. Fig. 3 illustrates the construction.

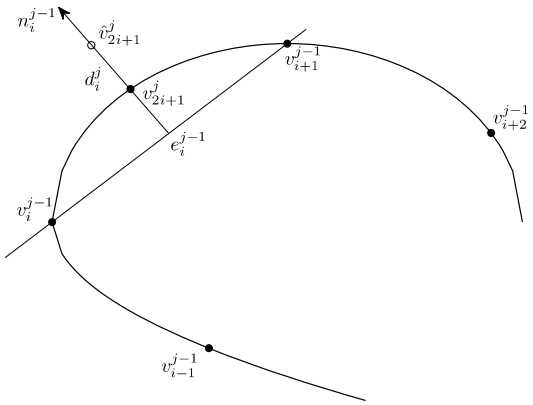


Fig. 3 Illustration of the normal multi-resolution scheme

Normal multi-resolution offers two obvious advantages. First of all, for smooth $\mathscr{C}$ and appropriate S, choosing the locally geometry-adapted frame consisting of the edge/normal vector pairs $(\mathbf{e}_i^{j-1}, \mathbf{n}_i^{j-1})$ results in much smaller detail magnitudes than using fixed coordinate axes for all $i \in \mathbb{Z}$, and, more importantly, the detail sequences contain scalar data. Indeed, the representation of the curve by its fine-scale sampling $\mathbf{v}^J$ requires $2^{J+1}n_0$ reals while its multi-scale $\{\mathbf{v}^0, d^j\}_{j=1}^J$ is given by $2n_0 + (1 + \ldots + 2^{J-1})n_0 \approx 2^J n_0$ reals. The savings are even more impressive when the idea is applied to surfaces and piecewise smooth multivariate functions, and combined with compression by thresholding detail entries d_i^j, see the performance reports in [41, 34, 39]. The papers [19, 55] give the analysis of normal multi-scale transforms with a linear interpolating subdivision operator S for the case of smooth curves based on perturbation arguments. The surface case still awaits its theoretical analysis.

In general, nonlinear multi-scale transforms operate on grid functions $\Gamma^j \to X$, where X typically coincides with $\mathbb{R}^n$ for some $n \geq 1$ (or with a manifold embedded into $\mathbb{R}^n$). The grids Γ^j are generated in a systematic way by a certain *topology refinement* (data-adapted topology refinement is an area of future research). In the above examples, $\Gamma^j = r^{-j}\mathbb{Z}$ is created by uniform r-adic refinement for $r = 2$, resp. $r = 3$. In the multivariate case, much more general grid topologies and refinement rules are possible. The values of the grid functions are collected into vectors v^j with entries indexed by the elements of Γ^j. They are related to each other by down-sampling operations using restriction operators R_j,

$$v^{j-1} = R_j v^j, \qquad j \geq 1, \tag{5}$$

detail computations

$$d^j = D_j(v^j, v^{j-1}, S_j v^{j-1}), \qquad j \geq 1, \tag{6}$$

involving prediction or subdivision operators S_j, and up-sampling operations

$$v^j = P_j(v^{j-1}, d^j) \qquad j \geq 1, \tag{7}$$

where at least one of these components involves nonlinear maps. In our univariate examples, we can use the fact that all Γ^j are isomorphic to $\mathbb{Z}$, and interpret the data vectors v^j as elements of $\ell_\infty(\mathbb{Z} \to X)$, where $X = \mathbb{R}$ in Example 1 and 2, and $X = \mathbb{R}^2$ for Example 3. This allows us to work with operators R, S, D, and P that act on this sequence space, and do not depend on j. For Example 1, the operator R is given by $(Rv)_i = v_{2i}$, is linear, and corresponds to trivial down-sampling for $r = 2$, while S is the ENO-modified nonlinear Deslauriers-Dubuc subdivision operator, $D(\tilde{v}, v, \hat{v}) := \tilde{v} - \hat{v}$, and $P(v, d) = Sv + d$. For Example 3, $X = \mathbb{R}^2$, R is as above, S is the linear Deslauriers-Dubuc subdivision operator, while the nonlinearity is induced through the normal map $v \to \mathbf{n}(v)$ that enters the detail computation, and up-sampling operation:

$$D(\tilde{v},v,\hat{v})_i := (\tilde{v}-\hat{v})_i \cdot \mathbf{n}(v)_i, \qquad i \in \mathbb{Z},$$

$$P(v,d)_{2i} = (Sv)_{2i} = v_i, \qquad P(v,d)_{2i+1} = (Sv)_{2i+1} + d_i \mathbf{n}_i, \qquad i \in \mathbb{Z}.$$

Finally, Example 2 is characterized by nonlinear R given by (2), and nonlinear S given by (3-4), while D and P remain the same as in Example 1.

The theoretical investigation of the properties of a nonlinear transform

$$v^J \quad \longleftrightarrow \quad \{v^0, d^1, \ldots, d^J\}, \qquad J \geq 1, \tag{8}$$

requires the study of the limit behavior for $J \to \infty$, and concentrates on answering the following natural questions:

- **Convergence**. Given reasonable v^0 and d^j, $j \geq 1$, do the reconstructed v^j converge to a reasonable limit object? This can be cast in terms of convergence of the associated sequence of functions f^j to a limit f^∞ in some function space.
- **Smoothness**. To judge the visual appearance of the results of reconstruction (for instance after a compression step), or in applications to numerical discretization schemes for elliptic boundary value problems, the guaranteed smoothness of these limits f^∞ in the Hölder or Sobolev scale is of essential interest.
- **Approximation and detail decay**. If the limit object f^∞ is sufficiently smooth, can we guarantee that the functions f^j that represent the grid data v^j converge to f^∞ at a certain prescribed rate typical for this smoothness class and comparable linear approximation processes? A related question is whether the smoothness properties of f^∞ can be characterized in terms of the detail sequences d^j, $j \geq 1$, as is well-known for many linear wavelet transforms.
- **Stability**. While the stability of the decomposition step

  $$v^J \quad \longrightarrow \quad \{v^0, d^1, \ldots, d^J\}, \qquad J \geq 1,$$

  (small perturbations in the fine-grid data v^J for J large will not lead to big perturbations of the details d^j, $j \leq J$, or the coarse grid data v^0) is often easy to understand (e.g., for interpolatory transforms when trivial down-sampling is used), the stability of the multi-scale reconstruction step

  $$\{v^0, d^1, \ldots, d^J\} \quad \longrightarrow \quad v^J, \qquad J \geq 1,$$

  is a difficult question of extreme importance for introducing compression strategies based on thresholding of detail sequences in a multi-scale decomposition.

Our aim is to survey the existing case studies for nonlinear multi-scale transforms and the emerging approaches to the development of a theory that tries to give first answers to the above questions. The exposition concentrates on the univariate case, namely multi-scale processing of data sampled from univariate functions or from curves. It is split into presenting the basic theory of nonlinear transforms in the functional setting in section 2 (this covers Example 1 and 2), and an exemplary discussion of what we call geometric subdivision schemes and multi-scale transforms, for which Example 3 is prototypical, in section 3. Extensions to multivariate

schemes, schemes that process manifold- or set-valued data, and some other recent developments will not be reviewed. This way, we hope to be able to expose the main ideas more clearly, and still provide enough guidance for future research in this exciting research field.

2 Nonlinear multi-scale transforms: Functional setting

2.1 Basic notation and further examples

Throughout Section 2, we consider local, r-shift invariant, stationary multi-scale transforms (8), recursively acting on data sequences from $\ell_p(\mathbb{Z})$ ($1 \le p \le \infty$) according to a simplified version of (5), (6), (7), where

$$v^{j-1} = Rv^j, \qquad d^j = D(v^j - Sv^{j-1}); \qquad v^j = Sv^{j-1} + Pd^j, \qquad j \ge 1, \tag{9}$$

with bounded but generally nonlinear operators $P, D, R, S : \ell_p(\mathbb{Z}) \to \ell_p(\mathbb{Z})$. Abusing a bit conventions in the nonlinear case, we call an operator $T : X \to Y$ between two Banach space X and Y bounded if there is a constant C_0 such that $\|Tx\|_Y \le C_0\|x\|_X$ for all $x \in X$, and Lipschitz continuity of such a T always means that there exists a constant C_1 such that $\|Tx - Ty\|_Y \le C_1\|x - y\|_X$ for all $x, y \in X$). For consistency in (9), the relation $(\mathrm{Id} - PD)(\mathrm{Id} - SR) = 0$ needs to hold. Here Id is the identity operator. That the operators in (9) are independent of the scale index $j \ge 1$ makes the scheme stationary. Example 1 and 2 from Section 1 fit this definition (for them $P = D = \mathrm{Id}$).

Example 4. Second generation linear and nonlinear wavelet transforms. Here is the construction of a 2-shift invariant univariate $1-1$ multi-scale transform from [8] based on the lifting scheme [59, 60]. Let v^j be split into "even" and "odd" parts

$$(R_e v^j)_i := v^j_{2i}, \qquad (R_o v^j)_i = v^j_{2i+1}, \qquad i \in \mathbb{Z}.$$

Then set

$$d^j := R_o v^j - T_{c,1} R_e v^j, \qquad v^{j-1} := R_e v^j + T_{c,2} d^j \tag{10}$$

for decomposition and

$$R_e v^j = v^{j-1} - T_{c,2} d^j, \qquad R_o v^j := d^j + T_{c,1} R_e v^j \tag{11}$$

for reconstruction. Here,

$$(T_{c,s} v)_i = \sum_{k=-K_1}^{K_2} b_{s;k}(v_{i-K_1}, \ldots, v_{i+K_2}) v_{i-k}, \qquad i \in \mathbb{Z},$$

are linear or nonlinear convolution operators generated by finitely many coefficient functions $b_{s;k} : \mathbb{R}^{K_1+K_2+1} \to \mathbb{R}$, $k = -K_1, \ldots, K_2$, $s = 1, 2$. This is called "predict

first" transform in [8], another "update first" transform is obtained by switching the order of execution of the sub-steps in (10), (11):

$$v^{j-1} := R_e v^j + T_{c,2} R_o v^j, \qquad d^j := R_o v^j - T_{c,1} v^{j-1}, \tag{12}$$

$$R_o v^j := d^j + T_{c,1} v^{j-1}, \qquad R_e v^j = v^{j-1} - T_{c,2} d^j, \tag{13}$$

Concrete examples, e.g., multi-scale transforms, where the nonlinearity is induced by quantization, or ENO-type schemes working with variable-order interpolating polynomials near a suspected jump discontinuity, and references can be found in [8].

Both transforms can be rewritten in our standard form (9), e.g., for the "predict first" version, one would set $R = (\mathrm{Id} - T_{c,2}T_{c,1})R_e + T_{c,2}R_o$, $D = -T_{c,1}R_e + R_o$, and define S and P by

$$R_e S = \mathrm{Id}, \quad R_o S = T_{c,1}, \qquad R_e P = -T_{c,2}, \quad R_o P = \mathrm{Id} - T_{c,1}T_{c,2}.$$

General linear bi-orthogonal wavelet transforms [18, 15] with finitely supported masks have similar representations, with all involved operators being linear. Transforms of the above type are obviously non-expansive $1-1$ transforms, i.e., do not formally change storage requirements. Note that there is also growing interest in expansive transforms related to tight affine frames. Lack of space prevents us from giving further details.

Example 5. Power-p schemes. This multi-scale transform is similar to Example 1 but uses a different prediction rule. I.e., again $r = 2$, $P = D = I$, the restriction operator is given by $(Rv)_i = v_{2i}$, $i \in \mathbb{Z}$, and the interpolating subdivision operator S is given by

$$(Sv)_{2i} = v_i, \qquad (Sv)_{2i+1} = \frac{v_i + v_{i+1}}{2} - \frac{1}{8} H_p(\Delta^2 v_{i-1}, \Delta^2 v_i), \qquad i \in \mathbb{Z}, \tag{14}$$

where the so-called limiter H_p is defined by

$$H_p(x,y) = \begin{cases} \frac{x+y}{2}\left(1 - \left|\frac{x-y}{x+y}\right|^p\right), & xy > 0, \\ 0, & xy \le 0. \end{cases} \tag{15}$$

The parameter $p \in [1, +\infty)$ is fixed, $\Delta^k := (\Delta)^k$ denotes the k-th order forward difference operator acting on sequence spaces, where $(\Delta v)_i = v_{i+1} - v_i$, $i \in \mathbb{Z}$.

Power-p schemes have been introduced in the context of generalized ENO-methods for hyperbolic problems [58], and can be useful for compressing piecewise smooth data and functions. Earlier, the case $p = 2$ appeared in [29, 43]. A straightforward calculation shows that if $v|_{\{i_0,\dots,i_1\}}$ is a convex (concave, linear) segment of v, then $(Sv)|_{\{2i_0,\dots,2i_1\}}$ preserves this property if $p \in [1,2]$. The formula is constructed such that for $\Delta^2 v_{i-1} = \Delta^2 v_i$ the obtained value $(Sv)_{2i+1}$ is the same as for

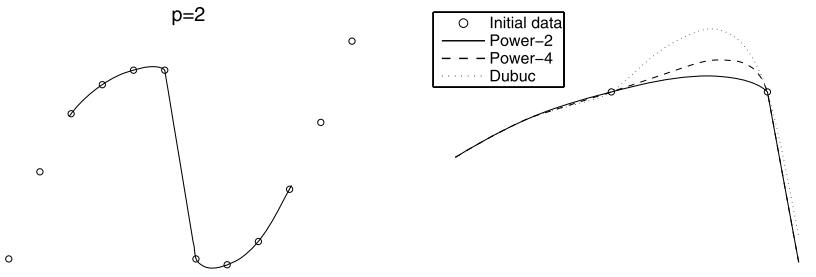


Fig. 4 Limits of power-2 subdivision (on the left), and a comparison with power-4 subdivision and the cubic Deslauriers-Dubuc scheme near a jump (on the right)

the linear interpolating cubic Deslauriers-Dubuc 4-point scheme discussed in Example 1 while at intervals, where the sign of the second differences changes, the newly inserted value is obtained by linear interpolation from its endpoint values.

Coming back to the notation for the r-shift invariant univariate case, the grids $\Gamma^j = r^{-j}\mathbb{Z}$ are systematically identified with $\mathbb{Z}$, the subdivision operator $S : \ell_p(\mathbb{Z}) \to \ell_p(\mathbb{Z})$ satisfies $ST_k = T_{rk}S$, and the restriction operator $R : \ell_p(\mathbb{Z}) \to \ell_p(\mathbb{Z})$ satisfies $RT_{kr} = T_kR$. The locality of an r-shift invariant transform is assured by assuming that the action of S is given by r multivariate functions ϕ_s according to

$$(Sv)_{ri+s} = \phi_s(v_{i-L_1}, \dots, v_{i+L_2}), \qquad s = 0, \dots, r-1, \tag{16}$$

where the integers L_1, L_2 are fixed and independent of $i \in \mathbb{Z}$, and $L = L_1 + L_2 + 1$ is the support length of the subdivision part of the transform. Similarly,

$$(Rv)_i = \phi(v_{ri-L_3}, \dots, v_{ri+L_4}) \tag{17}$$

for some function ϕ and fixed integers L_3, L_4. It is easy to see that due to locality and r-shift invariance, boundedness (Lipschitz continuity, C^1 property, ...) of S on $\ell_p(\mathbb{Z})$ spaces is equivalent to the boundedness (Lipschitz continuity, C^1 property, ...) of the coordinate functions $\phi_s : \mathbb{R}^L \to \mathbb{R}$ representing S, similarly for R. We will always silently assume that $S\mathbf{0} = R\mathbf{0} = \mathbf{0}$, where $\mathbf{0}$ is the zero sequence given by $\mathbf{0}_i = 0$, $i \in \mathbb{Z}$.

Sometimes, especially if nonlinear schemes are considered as perturbations of associated linear schemes, the alternative representation

$$(Sv)_{ri+s} = \sum_{l=-L_2}^{L_1} a_{rl+s}(v_{i-L_1}, \dots, v_{i+L_2})v_{i-l}, \qquad s = 0, \dots, r-1,$$

or, equivalently,

$$(Sv)_j = \sum_{i\in\mathbf{Z}} a_{j-ri}(v|_{I_{[j/r]}})v_i, \qquad j \in \mathbb{Z}, \tag{18}$$

is chosen. To shorten the notation, by $v|_{I_i}$ we have denoted the restriction of v to the finite index set $I_i := \{i-L_1,\dots,i+L_2\}$, $i \in \mathbb{Z}$. Coefficient functions with index $s \notin \{-rL_1,\dots,r(L_2+1)-1\}$ vanish for all arguments: $a_s(\cdot) \equiv 0$. Based on (18) we can now formally write the action of S as an infinite matrix-vector product

$$Sv = S_v v, \tag{19}$$

where S_v is a bi-infinite, data-dependent matrix operator with entries identified from (18):

$$(S_v)_{j,i} := a_{j-ri}(v|_{I_{[j/r]}}), \qquad j,i \in \mathbb{Z}.$$

Note that for a linear S, the matrix operator S_v does not depend on v (in which case we can drop the subscript v, and identify S with its matrix representation), and is given by the finitely supported sequence $a := \{a_l\}_{l\in\mathbf{Z}}$ called *mask* of the subdivision operator. The representation (18)-(19) was introduced in [13], and was the departure point for a systematic theory of data-dependent, so-called quasi-linear, subdivision schemes and multi-scale transforms which will be reviewed below. The transition from (16) to (18) and (19) is not unique, and needs to be done carefully.

Another often used approach is to write $S = S_0 + T'$, where S_0 is an appropriate linear subdivision operator and the remaining nonlinear part T' is "small" in a certain sense, see [19, 51, 2] for this perturbation approach. E.g., the power-p subdivision scheme from Example 5 naturally splits into a linear part S_0 (point insertion by linear interpolation) and nonlinear perturbation T' given by the limiter, and depending only on the 2nd order differences $\Delta^2 v$. For the median-interpolating scheme of Example 2, the natural choice for S_0 is the linear midpoint interpolation scheme given by

$$(S_0 v)_{3i+s} = \begin{cases} \frac{2v_{i-1}+8v_i-v_{i+1}}{9}, & s=0, \\ v_i, & s=1, \\ \frac{-v_{i-1}+8v_i+2v_{i+1}}{9}, & s=2, \end{cases} \qquad i \in \mathbb{Z},$$

and the resulting perturbation operator $T' = S - S_0$ given by

$$(T'v)_{3i+s} = \alpha_s(\Delta v|_{\{i-1,i\}})\Delta^2 v_i, \qquad s=0,1,2, \quad i \in \mathbb{Z},$$

depends in a specific way on Δv and $\Delta^2 v$ (e.g., the functions α_s are uniformly bounded, see [52, Section 2.2] for details). To identify the representation (16) from these formulas, set $L_1 = L_2 = 1$.

For the latter scheme, a natural choice for the representation (18), (19) is to set

$$a_{3l} = \begin{cases} \frac{2}{9}+\alpha_0(\cdot), \\ \frac{8}{9}-2\alpha_0(\cdot), \\ -\frac{1}{9}+\alpha_0(\cdot), \end{cases} \quad a_{3l+1} = \begin{cases} \alpha_1(\cdot), \\ 1-2\alpha_1(\cdot), \\ \alpha_1(\cdot), \end{cases} \quad a_{3l+2} = \begin{cases} -\frac{1}{9}+\alpha_2(\cdot), & l=1, \\ \frac{8}{9}-2\alpha_2(\cdot), & l=0, \\ \frac{2}{9}+\alpha_2(\cdot), & l=-1. \end{cases}$$

The non-zero entries of the matrix representation of S_v are contained in the 3×3 sub-blocks

$$(S_v)|_{\{3i,3i+1,3i+2\}\times\{i-1,i,i+1\}} = \begin{pmatrix} a_3(\Delta v|_{\{i-1,i\}}) & a_0(\Delta v|_{\{i-1,i\}}) & a_{-3}(\Delta v|_{\{i-1,i\}}) \\ a_4(\Delta v|_{\{i-1,i\}}) & a_1(\Delta v|_{\{i-1,i\}}) & a_{-2}(\Delta v|_{\{i-1,i\}}) \\ a_5(\Delta v|_{\{i-1,i\}}) & a_2(\Delta v|_{\{i-1,i\}}) & a_{-1}(\Delta v|_{\{i-1,i\}}) \end{pmatrix},$$

$i \in \mathbb{Z}$. In the general case, (18) results in a similar block-structured matrix operator S_v with $r \times (L_1 + L_2 + 1)$ sub-blocks.

Explicit representations (18), (19) have also been used in the study of (W)ENO schemes in [13] (for WENO, see also [2, Section 4.2]). Within the framework of Example 1, formulas for $a_s(\cdot)$ and S_v follow from

$$(Sv)_{2i+1} := \sum_{l=-1}^{1} c_{-l}(\Delta v|_{\{i-2,\dots,i+2\}})(S_l v)_{2i+l} \tag{20}$$

where S_0 is the standard 4-point scheme (1) while the formulas

$$(S_l v)_{2i+1} = \begin{cases} \frac{v_{i-2}-5v_{i-1}+10v_i+4v_{i+1}}{16}, & l = 1 \\ \frac{-v_{i-2}+9v_{i-1}+9v_i-v_{i+1}}{16}, & l = -1 \end{cases} \qquad i \in \mathbb{Z}.$$

define the linear schemes obtained from predictions using shifted cubic interpolation polynomials $p_{i\pm1}$ (all schemes are considered interpolating, i.e., $(Sv)_{2i} = (S_l v)_{2i} = v_i$, $i \in \mathbb{Z}$, $l = -1, 0, 1$). For WENO, the coefficient functions

$$c_l(\cdot) := \frac{\alpha_l(\cdot)}{\alpha_{-1}(\cdot) + \alpha_0(\cdot) + \alpha_1(\cdot)}, \qquad \alpha_l(\cdot) := \left(\frac{\gamma_l}{\varepsilon + \beta_l(\cdot)}\right)^2,$$

depend on non-negative, smooth functions $\beta_l(\cdot)$ measuring for argument $\Delta v|_{\{i-2,\dots,i+2\}}$ the degree of oscillation of the prediction polynomial p_{i+l}, and $\varepsilon, \gamma_l > 0$ are fixed constants, $l = -1, 0, 1$. The small parameter $\varepsilon > 0$ acts as regularization parameter, and avoids division by zero. Obviously, $c_l(\cdot) \in (0,1)$ and $c_1(\cdot) + c_0(\cdot) + c_1(\cdot) = 1$, i.e., the WENO subdivision operator (20) is a convex combination of the three linear operators S_l, with coefficients smoothly depending on Δv. The ENO subdivision operator has the same principal structure (20) but a generally discontinuous dependence of the coefficients $c_l(\cdot)$ on $\Delta(v)$: For ENO, we set $c_l(\Delta v|_{\{i-2,\dots,i+2\}}) = 1$ if p_{i+l} is the least oscillating of the three predictor polynomials, i.e., if $\beta_l(\Delta v|_{\{i-2,\dots,i+2\}})$ is minimal, and assign zeros to the other two coefficients. For these so-called 6-point (W)ENO schemes, set $L_1 = 2$, $L_2 = 3$ in (18).

The following subsections represent a summary of the currently available theoretical results for the family (9) of nonlinear multi-scale transforms and associated subdivision processes $v^j = Sv^{j-1}$, $j \geq 1$, which is obtained from (9) by formally setting $d^j = \mathbf{0}$ for $j \geq 1$. We survey mainly results from [13, 49, 19, 64, 52, 36, 2], where proofs and further material can be found. A few results, especially on offset

invariant S, and extensions to $L_p(\mathbb{R})$ for $1 \le p < \infty$, are new and elaborated on in more detail in a forthcoming paper.

2.2 Polynomial reproduction and derived subdivision schemes

The concept of polynomial reproduction for subdivision operators is fundamental in the study of multi-scale transforms, therefore we start the exposition with it. For nonlinear S, there are two slightly different extensions of the familiar definition for linear subdivision operators. The first definition follows [13], the second [64, 36]. Throughout the section, we denote by $\mathbb{P}_k$ the set of algebraic polynomials of degree $< k$ or, equivalently, of order $\le k$, and by $\mathbf{1}$ the constant sequence given by $\mathbf{1}_i := 1$, $i \in \mathbb{Z}$.

Definition 2.1. Let the r-shift invariant subdivision operator S be represented in the form (19). Then S has polynomial reproduction of order $k \ge 1$ if for each $v \in \ell_p(\mathbb{Z})$ the associated linear subdivision operator S_v has the following property: For any polynomial p of degree m, $0 \le m < k$, there exists a polynomial q of degree $< m$ such that

$$S_v(p|_{\mathbf{Z}}) = (p+q)|_{r^{-1}\mathbf{Z}}.$$

In particular, S reproduces constants (i.e., has polynomial reproduction of order $k = 1$) if

$$S_v(\mathbf{1}) = \mathbf{1}, \qquad \forall\, v \in \ell_p(\mathbb{Z}).$$

Definition 2.2. An r-shift invariant subdivision operator S is offset invariant for $\mathbb{P}_k$, $k \ge 1$, if for each $v \in \ell_p(\mathbb{Z})$, and any polynomial p of degree m, $0 \le m < k$, there exists a polynomial q of degree $< m$ such that

$$S(v + p|_{\mathbf{Z}}) = Sv + (p+q)|_{r^{-1}\mathbf{Z}}.$$

In particular, S is offset invariant for constants (i.e., the set $\mathbb{P}_1$) if

$$S(v + \alpha\mathbf{1}) = Sv + \alpha\mathbf{1}, \qquad \forall\, \alpha \in \mathbb{R}, \quad \forall\, v \in \ell_p(\mathbb{Z}).$$

Note that the formulation of these definitions automatically ensures that polynomial reproduction of order k implies polynomial reproduction of order m (similarly offset invariance for $\mathbb{P}_k$ implies offset invariance for $\mathbb{P}_m$) for all $1 \le m < k$. For linear S, both definitions coincide. As Example 1 and 2 demonstrate, for nonlinear schemes the two conditions are different. E.g., since the (W)ENO subdivision operator (20) is a convex combination of three linear subdivision operators, each being exact for cubic polynomials, it possesses polynomial reproduction of order $k = 4$. On the other hand, although it is obviously offset invariant for $\mathbb{P}_1$ (because the coefficient functions depend on Δv, and not on v directly), it is not offset invariant for any $\mathbb{P}_k$ with $k > 1$ (this assumes that the dependence of the oscillation indicators $\beta_l(\cdot)$ on v cannot be reduced to a direct dependence on higher order differences $\Delta^k v$). For the median-interpolating scheme of Example 2, a close look at

the coefficient representations reveals that offset invariance can hold only for $\mathbb{P}_1$ but polynomial reproduction of order at least $k=2$ holds. For the power-p scheme, both definitions apply for $k=2$.

As all our examples indicate, offset invariance for $\mathbb{P}_k$ of a nonlinear scheme is to be expected to hold for $k=1$, sometimes holds with $k=2$ but no nonlinear examples of relevance are known for larger k. However, it is the right concept for the extension of the notion of *derived subdivision operators* to the nonlinear setting.

Theorem 2.1. *Let S be a local, r-shift invariant subdivision operator. If S is offset invariant for $\mathbb{P}_k$ for some integer $k \ge 1$ then there exist local, r-shift invariant derived subdivision operators $S^{[m]} : \ell_p(\mathbb{Z}) \to \ell_p(\mathbb{Z})$ such that*

$$\Delta^m S v = S^{[m]} \Delta^m v, \qquad \forall\, v \in \ell_p(\mathbb{Z}) \tag{21}$$

for $m = 1,2,\dots,k$. Moreover, if S is written in the form (16) then its derived subdivision operators $S^{[m]}$, $m=1,\dots,k$, inherit such a representation with the same (or smaller) L_1, L_2, and with functions $\phi_s^{[k]}$ that are obtained from the ϕ_s by superpositions involving only linear transformations. In particular, if S (and thus the functions $\phi_s(\cdot)$) is bounded (continuous, Lipschitz continuous, C^1, ...) then so is $S^{[k]}$ (and the functions $\phi_s^{[k]}$). In particular, if S is bounded then

$$\|S^{[m]} w\|_{\ell_p(\mathbf{Z})} \le r^{-m+1/p} \|w\|_{\ell_p(\mathbf{Z})} + C \|\Delta w\|_{\ell_p(\mathbf{Z})}, \tag{22}$$

and if S is Lipschitz continuous then

$$\|S^{[m]} w - S^{[m]} w'\|_{\ell_p(\mathbf{Z})} \le r^{-m+1/p} \|w - w'\|_{\ell_p(\mathbf{Z})} + C \|\Delta (w - w')\|_{\ell_p(\mathbf{Z})}, \tag{23}$$

$m = 0,1,\dots,k-1$, with constants C independent of $w, w' \in \ell_p(\mathbb{Z})$.

The proof extends the standard argument for linear S, see [10, 26]. For $k=1$ it was first given in [64, Theorem 2.5], see also [36, Lemma 2.1-2]. The case $k>1$ was suggested in [36, Section 2.1] and can be obtained by induction from $k=1$.

Definition 2.2 can be replaced by a recursive one: S is offset invariant for $k \ge 2$ if it is offset invariant for $\mathbb{P}_{k-1}$, and the scaled version of the associated $(k-1)$-st derived operator $\tilde{S}^{[k-1]} = r^{k-1} S^{[k-1]}$ is offset invariant for constants. If S has polynomial reproduction of order k, then all linear subdivision operators S_v are offset invariant for $\mathbb{P}_k$, and thus derived subdivision operators $(S_v)^{[m]}$ exist for all $m = 1,\dots,k$ and v. Thus, Theorem 2.1 covers this case as well. To give a concrete example, let us consider Example 5. We already mentioned that the power-p scheme is offset invariant for $\mathbb{P}_2$ which follows from observing that $H_p((\Delta^2(v+q|_{\mathbf{Z}}))_{i-1}, (\Delta^2(v+q|_{\mathbf{Z}}))_i) = H_p((\Delta^2 v)_{i-1}, (\Delta^2 v)_i)$ for all v, $i \in \mathbb{Z}$, and $q \in \mathbb{P}_2$. From (14) we find

$$(\Delta S v)_{2i} = \frac{\Delta v_i}{2} - \frac{1}{8} H_p(\Delta^2 v_{i-1}, \Delta^2 v_i), \qquad (\Delta S v)_{2i+1} = \frac{\Delta v_i}{2} + \frac{1}{8} H_p(\Delta^2 v_{i-1}, \Delta^2 v_i),$$

and

$$\begin{aligned}(\Delta^2 Sv)_{2i} &= \tfrac{1}{4}H_p(\Delta^2 v_{i-1}, \Delta^2 v_i),\\ (\Delta^2 Sv)_{2i+1} &= \tfrac{\Delta^2 v_i}{2} - \tfrac{1}{8}(H_p(\Delta^2 v_{i-1}, \Delta^2 v_i) + H_p(\Delta^2 v_i, \Delta^2 v_{i+1})),\end{aligned} \qquad i \in \mathbb{Z}.$$

Thus, the derived subdivision operators $S^{[1]}$ and $S^{[2]}$ are given by

$$\begin{aligned}(S^{[1]}w)_{2i} &= \tfrac{w_i}{2} - \tfrac{1}{8}H_p(\Delta w_{i-1}, \Delta w_i),\\ (S^{[1]}w)_{2i+1} &= \tfrac{w_i}{2} + \tfrac{1}{8}H_p(\Delta w_{i-1}, \Delta w_i),\\ (S^{[2]}w)_{2i} &= \tfrac{1}{4}H_p(w_{i-1}, w_i),\\ (S^{[2]}w)_{2i+1} &= \tfrac{w_i}{2} - \tfrac{1}{8}(H_p(w_{i-1}, w_i) + H_p(w_i, w_{i+1})),\end{aligned} \qquad i \in \mathbb{Z}. \tag{24}$$

2.3 Convergence and smoothness

In the univariate case, L_p-convergence of the reconstruction part

$$v^j = Sv^{j-1} + Pd^j, \qquad j \geq 1, \tag{25}$$

of a multi-scale transform resp. the subdivision scheme

$$v^j = Sv^{j-1}, \qquad j \geq 1, \tag{26}$$

associated with it to a limit function, and the smoothness of the latter, can be studied by associating with v^j its linear spline interpolants f^j on the grid $\Gamma^j = r^{-j}\mathbb{Z}$:

$$f^j(x) = (i+1-r^j x)v_i^j + (r^j x - i)v_{i+1}^j), \qquad x \in [r^{-j}i, r^{-j}(i+1)), \qquad i \in \mathbb{Z}. \tag{27}$$

Alternatively, we can write $f^j = \sum_i v_i^j B_2(r^j \cdot - i)$ using linear B-splines (with $B_2(x) = 1 - |x|$ for $|x| \leq 1$, and $B_2(x) = 0$ otherwise), and think of f^j as the limit of a linear subdivision process for B-splines of order 2.

Definition 2.3. The multi-scale reconstruction algorithm (25) is called L_p convergent if, for any $v^0 \in \ell_p(\mathbb{Z})$ and detail sequences $d^j \in \ell_p(\mathbb{Z})$ satisfying

$$\sum_{j \geq 1} r^{-j/p} \|d^j\|_{\ell_p(\mathbf{Z})} < \infty, \tag{28}$$

the corresponding sequence of linear interpolants f^j converges in $L^p(\mathbb{R})$ to a limit function $f^\infty \in L_p(\mathbb{R})$.
Similarly, if the subdivision scheme (26) associated with S is called L_p convergent if $f^j \to f^\infty \not\equiv 0$ in $L_p(\mathbb{R})$ for any $v^0 \neq \mathbf{0}$.

In applications to multi-scale solvers for operator equations [15, 12] and geometric modeling [26], the smoothness characteristics and sometimes also shape properties of the limits f^∞ matter. Smoothness of functions that are limits of approximation processes (in our case the recursively constructed sequences $\{f^j\}$ of linear splines) is conveniently measured in the scale of Besov spaces (see [49] for various equiv-

alent definitions including the standard one based on moduli of smoothness). We give a definition for a subclass of Besov spaces using an approximation-theoretic characterization which is convenient for our setup. Let $1 \le p \le \infty$, $k = 1, 2, \ldots$, and $0 < s < k - 1 + 1/p$. A function $f \in L_p(\mathbb{R})$ belongs to the Besov space $B_p^s(\mathbb{R})$ if and only if there exists at least one L_p convergent series representation

$$f = \sum_{j=0}^{\infty} h^j,$$

where the functions h^j are splines of order k with knots at the grid points $\Gamma^j = r^{-j}\mathbb{Z}$, satisfying the constraint

$$\sum_{j=0}^{\infty} r^{spj} \|h^j\|_{L_p(\mathbb{R})}^p < \infty,$$

if $1 \le p < \infty$, and

$$\sup_{j \ge 0} r^{sj} \|h^j\|_{L_\infty(\mathbb{R})} < \infty,$$

if $p = \infty$. Moreover, we can define a norm in $B_p^s(\mathbb{R})$ by setting

$$\|f\|_{B_p^s(\mathbb{R})} := \begin{cases} \inf \left(\sum_{j=0}^{\infty} r^{spj} \|h^j\|_{L_p(\mathbb{R})}^p \right)^{1/p}, & 1 \le p < \infty, \\ \inf \sup_{j \ge 0} r^{sj} \|h^j\|_{L_\infty(\mathbb{R})}, & p = \infty, \end{cases} \tag{29}$$

where the infimum is taken with respect to all such representations. For given s, the choice of k is secondary: Norms for different k are equivalent (for this reason, we did not show the dependence of the Besov space norm on k). Proofs based on Jackson-Bernstein inequalities for splines and references can be found in [11, 50, 15]. Note that for the two important subcases $p = \infty$ and $p = 2$, the scale $B_p^s(\mathbb{R})$, $s > 0$, coincides with the scale of Hölder-Zygmund classes $\mathscr{C}^s$ resp. Sobolev spaces $H^s(\mathbb{R}) = W_2^s(\mathbb{R})$.

We are now ready to discuss the smoothness of the algorithms (25) and (26).

Definition 2.4. The subdivision scheme (26) associated with S possesses L_p smoothness $s > 0$ if it is L_p convergent, with limit functions satisfying

$$f^\infty \in B_p^{s-}(\mathbb{R}) := \bigcup_{0<t<s} B_p^t(\mathbb{R}), \qquad \forall v^0 \in \ell_p(\mathbb{Z}).$$

The maximal such $s > 0$ is called the L_p smoothness exponent of S, and denoted by $s_p(S)$.

The following theorem is proved in [13, 49] for S of the form (19).

Theorem 2.2. *Let S be a local, r-shift invariant, bounded subdivision operator in $\ell_p(\mathbb{Z})$, represented by (19) via a family of linear subdivision operators $\{S_v,\ v \in \ell_p(\mathbb{Z})\}$ which are uniformly bounded,*

$$\|S_v w\|_{\ell_p(\mathbf{Z})} \le C \|w\|_{\ell_p(\mathbf{Z})}, \qquad \forall w, v \in \ell_p(\mathbb{Z}),$$

and have polynomial reproduction order k for some integer $k \geq 1$. Let P be a bounded operator on $\ell_p(\mathbf{Z})$.
i) If

$$\rho_{p,k}(\{S_v\}) := \limsup_{j\to\infty} \sup_{v^0\in\ell_p(\mathbf{Z})} \|(S_{v^{j-1}})^{[k]} \dots (S_{v^1})^{[k]}(S_{v^0})^{[k]}\|^{1/j}_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})} < r^{1/p}, \quad (30)$$

then S is L_p convergent. In this case, a lower bound for the L_p smoothness exponent of S is given by

$$s_p(S) \geq \min(k, -\log_r(r^{-1/p}\rho_{p,k}(\{S_v\}))) > 0. \quad (31)$$

ii) If

$$\tilde{\rho}_{p,k}(\{S_v\}) := \limsup_{j\to\infty} \sup_{w^j\in\ell_p(\mathbf{Z}),\, l=0,\dots,j-1} \|(S_{w^{j-1}})^{[k]} \dots (S_{w^1})^{[k]}(S_{w^0})^{[k]}\|^{1/j}_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})}, \quad (32)$$

satisfies $\tilde{\rho}_{p,k}(\{S_v\}) < r^{1/p}$, then the multi-scale reconstruction algorithm (25) is L_p convergent.
Moreover, if for some s satisfying

$$0 < s < \min(k, -\log_r(r^{-1/p}\rho_{p,k}(\{S_v\})))$$

the norm

$$\|\{v^0, d^j\}_{j\geq 1}\|_{p,s;r} := \begin{cases} \left(\|v^0\|^p_{\ell_p(\mathbf{Z})} + \sum_{j\geq 1} r^{j(sp-1)}\|d^j\|^p_{\ell_p(\mathbf{Z})}\right)^{1/p}, & 1 \leq p < \infty, \\ \sup\{\|v^0\|_{\ell_p(\mathbf{Z})}, r^{j(s-1/p)}\|d^j\|_{\ell_p(\mathbf{Z})}\}_{j\geq 1}, & p = \infty. \end{cases}$$

is finite, then the limit function f of the multi-scale reconstruction (25) belongs to $B^s_p(\mathbf{R})$, and

$$\|f\|_{B^s_p(\mathbf{R})} \leq C\|\{v^0, d^j\}_{j\geq 1}\|_{p,s,r}. \quad (33)$$

The counterpart of this theorem for S that are offset invariant for $\mathbb{P}_k$ and thus possess derived subdivision operators $S^{[l]}$, $l = 1, \dots, k$, is formulated in the next theorem. In this generality it is new, although partial cases have appeared before, see, e.g., [2] for $p = \infty$.

Theorem 2.3. *Let S be a local, r-shift invariant, bounded subdivision operator operator on $\ell_p(\mathbf{Z})$. Assume that S is offset invariant for $\mathbb{P}_k$ for some integer $k \geq 1$.*
i) If

$$\rho_{p,k}(S) = \rho_p(S^{[k]}) := \limsup_{j\to\infty} \|(S^{[k]})^j\|^{1/j}_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})} < r^{1/p} \quad (34)$$

then S is L_p convergent, and

$$s_p(S) \geq \min(k, -\log_r(r^{-1/p}\rho_{p,k}(S))) > 0. \quad (35)$$

ii) If, in addition, S is Lipschitz continuous, and P is bounded, then (34) also implies the L_p convergence of the multi-scale reconstruction (25). Moreover, if for some s satisfying $0 < s < \min(k, -\log_r(r^{-1/p}\rho_{p,k}(S)))$ we have

$$\|\{v^0, d^j\}_{j\ge 1}\|_{p,s;r} < \infty,$$

then the limit function f belongs to $B_p^s(\mathbf{R})$ and (33) holds.

A couple of comments on the introduced spectral radii, and the range of applicability of the two theorems are in order. First of all, instead of lim sup one can write lim in all three cases. Also, by definition $\tilde\rho_{p,k}(\{S_v\}) \ge \rho_{p,k}(\{S_v\})$. Secondly, by the definition of derived subdivision operators, both $\rho_{p,k}(\{S_v\})$ and $\rho_{p,k}(S) = \rho_p(S^{[k]})$ are tied to geometric decay estimates for the norms of the sequences $\Delta^k v^j$, where $v^j = Sv^{j-1} = S^j v^0$. E.g., by repeatedly applying

$$\Delta^k v^j = \Delta^k S_{v^{j-1}} v^{j-1} = (S_{v^{j-1}})^{[k]} \Delta^k v^{j-1},$$

for the subdivision algorithm (19), we have

$$\begin{aligned}\|\Delta^k v^j\|_{\ell_p(\mathbf{Z})} &\le \|(S_{v^{j-1}})^{[k]} \dots (S_{v^1})^{[k]} (S_{v^0})^{[k]}\|_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})} \|\Delta^k v^0\|_{\ell_p(\mathbf{Z})} \\ &\le C\rho^j \|\Delta^k v^0\|_{\ell_p(\mathbf{Z})}\end{aligned}$$

for $j \ge 1$, whenever $\rho > \rho_{p,k}(\{S_v\})$. The constant C only depends on k and the chosen ρ.

The same argument goes through for $\rho_{p,k}(S)$. However, in this case the infimum of the set of all ρ for which such a geometric decay holds equals $\rho_{p,k}(S)$. To see this, observe that $\Delta^k v^j = (S^{[k]})^j \Delta^k v^0$ for all v_0, and thus

$$\|(S^{[k]})^j\|_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})} = \rho_{p,k;j}(S) := \sup_{\|\Delta^k v\|_{\ell_p(\mathbf{Z})}=1} \|\Delta^k S^j v\|_{\ell_p(\mathbf{Z})}, \tag{36}$$

and

$$\rho_{p,k}(S) = \limsup_{j\to\infty} \rho_{p,k;j}(S)^{1/j} = \inf\{\rho : \|\Delta^k S^j v\|_{\ell_p(\mathbf{Z})} \le C_\rho \rho^j \|\Delta^k v\|_{\ell_p(\mathbf{Z})}\}. \tag{37}$$

By definition of derived subdivision operators of S and of the linear operators $\{S_v\}$ we have $(S_v)^{[k]}\Delta^k v = \Delta^k S_v v = \Delta^k S v = S^{[k]}\Delta^k v$, consequently always

$$\tilde\rho_{p,k}(\{S_v\}) \ge \rho_{p,k}(\{S_v\}) \ge \rho_{p,k}(S), \tag{38}$$

if the conditions for the existence of these spectral radii are met. (38) holds for all admissible choices of the family of linear subdivision operators $\{S_v\}$ representing S. Since in practice offset invariance for $\mathbb{P}_k$ holds often with $k \le 2$ only, part i) of Theorem 2.2 offers sometimes greater flexibility because it may even apply for larger k. A concrete example is given by the dyadic median-interpolating subdivision scheme [52] for which offset invariance for $\mathbb{P}_k$ holds for $k = 1$ only, and

$\rho_{p,1}(S) = 1/2$, while $\{S_v\}$ in the representation (19) have polynomial reproduction of order $k = 2$ and $\rho_{p,2}(\{S_v\}) < 1/2$. The dyadic median-interpolating subdivision scheme also provides an instance when the first inequality in (38) is strict, see [52]. A comparison of $\tilde{\rho}_{p,k}(\{S_v\})$ and $\rho_{p,k}(S)$ is more subtle since S does not define $\{S_v\}$ uniquely. An example of an operator S showing that $\tilde{\rho}_{p,k}(\{S_v\}) > \rho_{p,k}(S)$ for *any* admissible choice of the family $\{S_v\}$ is, to the best of our knowledge, not known. The additional assumption of Lipschitz stability for part ii) of Theorem 2.2 represents a mild restriction since most schemes satisfy it (the exception is the ENO scheme). Moreover, Lipschitz stability of S is necessary for the stability of subdivision and multi-scale reconstruction algorithms associated with S, a desirable property which is discussed in the next subsection.

Another useful property of the spectral radii is that

$$\rho_{p,m}(S) \le \max(r^{-m+1/p}, \rho_{p,k}(S)), \qquad m = 1, \dots, k-1, \tag{39}$$

and similar inequalities hold for $\rho_{p,m}(\{S_v\})$ and $\tilde{\rho}_{p,m}(\{S_v\})$. For the proof it is enough to consider $m = k-1$, the rest follows by recursion. Set $\hat{\rho} = r^{-k+1+1/p}$, and $w = (S^{[k-1]})^{n-1} v$ in (22). Then

$$\begin{aligned} \|(S^{[k-1]})^n v\|_{\ell_p(\mathbf{Z})} &\le \hat{\rho} \|(S^{[k-1]})^{n-1} v\|_{\ell_p(\mathbf{Z})} + C \|\Delta (S^{[k-1]})^{n-1} v\|_{\ell_p(\mathbf{Z})} \\ &= \hat{\rho} \|(S^{[k-1]})^{n-1} v\|_{\ell_p(\mathbf{Z})} + C \|(S^{[k]})^{n-1} \Delta v\|_{\ell_p(\mathbf{Z})} \\ &\le \hat{\rho} \|(S^{[k-1]})^{n-1} v\|_{\ell_p(\mathbf{Z})} + C \rho^n \|v\|_{\ell_p(\mathbf{Z})}, \end{aligned}$$

where $\rho = \rho_{p,k}(S)) + \varepsilon$ is fixed with arbitrary $\varepsilon > 0$. By recursion,

$$\|(S^{[k-1]})^n v\|_{\ell_p(\mathbf{Z})} \le C \|v\|_{\ell_p(\mathbf{Z})} \sum_{i=0}^{n} \hat{\rho}^{n-i} \rho^i \le C n (\max(\hat{\rho}, \rho))^n \|v\|_{\ell_p(\mathbf{Z})}.$$

This shows that $\rho_{p,k-1}(S) \le \max(\hat{\rho}, \rho)$, and (39) follows if $\varepsilon \to 0$.

In applications, to get upper bounds for the above spectral radii, estimates for the quantities $\rho_{p,k;j}(S)$ defined in (36) are used for small values of j. Unfortunately, as experimental evidence shows, the convergence of $\rho_{p,k;j}(S)^{1/j}$ towards $\rho_{p,k}(S)$ is generally very slow. Alternatively, due to the locality of the subdivision operators involved, $\rho_{p,k}(S)$ can also be characterized as the *ℓ_p-joint spectral radius* of a certain family of nonlinear maps acting on a certain $\mathbf{R}^M$, where M depends on the dilation factor r and the support length L of S. For linear subdivision operators S, there is an extensive literature on this subject, especially for the cases $p = 2$ and $p = \infty$. In the nonlinear case there is much room for further research.

As an illustration, let us consider the power p-scheme (Example 5). As was mentioned in subsection 2.1, the associated subdivision operator is offset invariant for $\mathbf{P}_k$, $k = 1, 2$, with explicit formulas for $S^{[k]}$ given in (24). The limiter $H_p(x, y)$ from (15) vanishes whenever $xy \le 0$ and otherwise satisfies

$$0 < \alpha(x, y) := \frac{2H_p(x, y)}{x + y} \le 1, \qquad xy > 0.$$

Thus, setting $\alpha(x,y)=0$ for $xy\le 0$, and denoting $\alpha_i:=\alpha(w_{i-1},w_i)$, we easily get from (24) that

$$(S^{[2]}w)_{2i} = \frac{\alpha_i}{8}(w_{i-1}+w_i),$$
$$(S^{[2]}w)_{2i+1} = \frac{1}{16}((8-\alpha_i-\alpha_{i+1})w_i-\alpha_i w_{i-1}-\alpha_{i+1}w_{i+1}).$$

Taking absolute values, we immediately get

$$|(S^{[2]}w)_{2i}|\le\frac{1}{4}\max\{|w_{i-1}|,|w_i|\},\qquad |(S^{[2]}w)_{2i+1}|\le\frac{1}{2}\max\{|w_{i-1}|,|w_i|,|w_{i+1}|\},$$

which, according to (36) for $j=1$, gives

$$\rho_{\infty,2}(S)\le\|S^{[2]}\|_{\ell_\infty(\mathbf{Z})\to\ell_\infty(\mathbf{Z})}\le\frac{1}{2}<1.$$

Thus, this crude estimate implies uniform convergence, and gives $s_\infty(S)\ge 1$ for any power-p subdivision scheme. In this particular case, this estimate for the Hölder exponent is sharp: For the initial sequence $v^0=(\dots,0,1,0,1,\dots)$, the limit f^∞ is the linear spline interpolant to these data on $\mathbf{Z}$, and does not belong to $C^1(\mathbf{R})$, and thus to any $B^s_\infty(\mathbf{R})$ with $s>1$. The final result for the Hölder exponent of these schemes is $s_\infty(S)=1$ which is known for a long time (for $p=2$, see [29], for other p, see, e.g., [2]). We conjecture that

$$s_q(S)=-\log_2(\rho_{q,2}(S))+\frac{1}{q}=1+\frac{1}{q},\qquad 1\le q<\infty, \tag{40}$$

holds for all power-p schemes but have verified this only in partial situations such as for the convexity-preserving case $p\le 2$, where

$$\|S^{[2]}w\|_{\ell_q(\mathbf{Z})\to\ell_q(\mathbf{Z})}=\frac{1}{2},\qquad 1\le q\le\infty,$$

can be deduced from the already established result for $q=\infty$ and from the case $q=1$ by complex interpolation (the upper bound $s_q(S)\le 1+1/q$ follows from the same linear spline example as used for $q=\infty$).

However, in most examples of nonlinear S, the trivial upper estimates

$$\rho_{q,k}(\{S_v\})\le\sup_v\|(S_v)^{[k]}\|_{\ell_q(\mathbf{Z})\to\ell_q(\mathbf{Z})}$$

resp.

$$\rho_{q,k}(S)\le\|S^{[k]}\|_{\ell_q(\mathbf{Z})\to\ell_q(\mathbf{Z})}$$

for the spectral radii are just too weak (in this regard, the power-p schemes represent an exception), and one needs to resort to (36) for $j>1$ to obtain more rigorous bounds.

The computation of exact values for these spectral radii and for the smoothness exponents $s_q(S)$ becomes a subtle issue. In this respect the nonlinear case is much harder than the case of linear subdivision operators S, where it can be reduced to the q-joint spectral radius problem for finite families of matrices, or in the special case of L_2-smoothness exponents, to a finite dimensional eigenvalue problem, see, e.g., [40, 68]. Finding the smoothness exponents of nonlinear schemes usually means to enter a detailed study of the nonlinear dynamics hidden in the subdivision scheme.

The only nontrivial case, where such an investigation has led to success is the paper [64] by Xie and Yu, where

$$s_\infty(S) = 1$$

has been established for the triadic median-interpolating S (Example 2). The same authors [64, 63] have also propagated a conjecture on *smoothness equivalence*: For many nonlinear S, the smoothness exponent $s_q(S)$ coincides with the smoothness exponent of a near-by linear S_0. Currently this conjecture is established only in a few cases, in particular for manifold-valued subdivision schemes [65]. For the power-p subdivision operator S, the appropriate S_0 is equivalent to the linear B-spline subdivision, and is obtained if the limiter term $H_p(\cdot)$ is dropped from the definition. For median-interpolating schemes, S_0 is given by systematically replacing all conditions of median interpolation by interpolation conditions at the interval midpoints, more examples can be found in [65].

2.4 Stability

Stability of multi-scale transforms, i.e., the robustness with respect to small perturbations, is not a major issue for linear schemes since convergence of a linear subdivision scheme implies stability. However, for nonlinear schemes it is by no means obvious, and deserves consideration. In this subsection, we consider only the case of Lipschitz stability in $L_p(\mathbb{R})$. We will again deal with the simplified version (9) of a nonlinear multi-scale transform, and its parts: The reconstruction part (25), the associated subdivision scheme (26), and the decomposition part

$$v^{j-1} = Rv^j, \qquad d^j = D(v^j - Sv^{j-1}), \qquad j = J, J-1, \dots, 1. \tag{41}$$

Definition 2.5. The decomposition algorithm (41) is called L_p stable if there is a constant C_D such that

$$\max\{\|v^0 - \tilde{v}^0\|_{\ell_p(\mathbf{Z})}, r^{-j/p}\|d^j - \tilde{d}^j\|_{\ell_p(\mathbf{Z})}\}_{j=1,\dots,J} \le C_D r^{-J/p}\|v^J - \tilde{v}^J\|_{\ell_p(\mathbf{Z})}$$

holds for all $v^J, \tilde{v}^J \in \ell_p(\mathbb{Z})$, and $J \ge 1$.
The reconstruction algorithm (25) is called L_p stable if there is a constant C_U such that

$$r^{-J/p}\|v^J-\tilde{v}^J\|_{\ell_p(\mathbf{Z})}\le C_U(\|v^0-\tilde{v}^0\|_{\ell_p(\mathbf{Z})}+\sum_{j=1}^{J} r^{-j/p}\|d^j-\tilde{d}^j\|_{\ell_p(\mathbf{Z})})$$

holds for all $v^0,\tilde{v}^0\in\ell_p(\mathbf{Z})$, $d^j,\tilde{d}^j\in\ell_p(\mathbf{Z})$, $j=1,\dots J$, and $J\ge 1$.

The subdivision algorithm (26) is called L_p stable if there is a constant C_S such that

$$r^{-J/p}\|v^J-\tilde{v}^J\|_{\ell_p(\mathbf{Z})}\le C_S\|v^0-\tilde{v}^0\|_{\ell_p(\mathbf{Z})}$$

holds for all $v^0,\tilde{v}^0\in\ell_p(\mathbf{Z})$, and $J\ge 1$.
For all these definitions it is assumed that the associations

$$v^J \longleftrightarrow \{v^0,d^1,\dots,d^J\}$$
$$\tilde{v}^J \longleftrightarrow \{\tilde{v}^0,\tilde{d}^1,\dots,\tilde{d}^J\}$$

are given by the corresponding recursions in (9), where in the subdivision case detail sequences are set to $\mathbf{0}$.

Defining L_p stability in this form is valuable for realistic algorithms, e.g., for compression based on detail thresholding. The inclusion of the fore-factors $r^{-j/p}$ is dictated by the interpretation of the sequences v^j as representations of an L_p limit function on the grids Γ^j. Indeed, assuming L_p convergence of the reconstruction algorithm studied in the previous subsection, the stability of (25) implies

$$\|f^\infty-\tilde{f}^\infty\|_{L_p(\mathbf{Z})}\le C_U(\|v^0-\tilde{v}^0\|_{\ell_p(\mathbf{Z})}+\sum_{j=1}^{\infty} r^{-j/p}\|d^j-\tilde{d}^j\|_{\ell_p(\mathbf{Z})})$$

for the L_p limits of the associated sequences $\{f^j\}_{j\ge 0}$ and $\{\tilde{f}^j\}_{j\ge 0}$. Finally, we note that L_p stability of the decomposition part in a stronger form (e.g., symmetric with the stability condition for (25)) is probably too much to ask for.

We will briefly deal with the decomposition part, where L_p stability can often be determined easily. If R is linear then the condition

$$\|R^n\|_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})}\le Cr^{-n/p},\qquad n\ge 1, \tag{42}$$

together with the Lipschitz continuity of D and S, is a necessary and sufficient condition for the L_p stability of (41). For the nonlinear case the corresponding sufficient condition on R reads

$$\|R^n v-R^n\tilde{v}\|_{\ell_p(\mathbf{Z})}\le Cr^{-n/p}\|v-\tilde{v}\|_{\ell_p(\mathbf{Z})},\qquad n\ge 1,$$

uniformly in $v,\tilde{v}$, and $n\ge 1$. Indeed, by rephrasing this inequality we get

$$r^{-j/p}\|v^j-\tilde{v}^j\|_{\ell_p(\mathbf{Z})}=r^{-j/p}\|R^{J-j}v^J-R^{J-j}\tilde{v}^J\|_{\ell_p(\mathbf{Z})}\le Cr^{-J/p}\|v^J-\tilde{v}^J\|_{\ell_p(\mathbf{Z})}.$$

The Lipschitz continuity of D and S yields the stability inequalities for d^j, $j=1,\dots,J$, as well.

With these definitions, trivial down-sampling given by $(Rv)_i = v_{ri}$ which is typical for interpolatory multi-scale transforms leads to L_p stability in (41) only if $p = \infty$, which is logical since point evaluations on L_p functions are not well-defined. On the other hand, if R is a linear averaging restriction operator given by

$$(Rv)_i = \sum_{l \in \mathbf{Z}} b_l v_{ri+l},$$

where the sequence $\{b_l\}$ is finitely supported, non-negative, and satisfying

$$\sum_{i \in \mathbf{Z}} b_{ri+s} = r^{-1}, \qquad s = 0, \ldots, r-1,$$

then (41) is L_p stable for all $1 \le p \le \infty$. Indeed, since $(Rv)_i$ is a convex combination of entries of v, we get

$$\begin{aligned} \|Rv\|^p_{\ell_p(\mathbf{Z})} &\le \sum_{i \in \mathbf{Z}} |\sum_{l \in \mathbf{Z}} b_l v_{ri+l}|^p \le \sum_{l \in \mathbf{Z}} b_l \sum_{i \in \mathbf{Z}} |v_{ri+l}|^p \\ &= \sum_{s=0}^{r-1} (\sum_{i \in \mathbf{Z}} b_{ri+s})(\sum_{i \in \mathbf{Z}} |v_{ri+s}|^p) = r^{-1} \|v\|^p_{\ell_p(\mathbf{Z})}, \end{aligned}$$

and (42) holds with $C = 1$.

The only example of a decomposition algorithm (41) with a nonlinear R comes from Example 2 (median-interpolating schemes), where $r = 3$, and R is defined via (2). The obvious inequality

$$|\mathrm{med}(a,b,c) - \mathrm{med}(a',b',c')| \le \max(|a-a'|, |b-b'|, |c-c'|)$$

implies L_∞ stability of this R. The example of the two sequences

$$v^J_i = \begin{cases} 0, & i < (3^J-1)/2 \\ 1, & i \ge (3^J-1)/2, \end{cases} \qquad \tilde{v}^J_i = \begin{cases} 0, & i \le (3^J-1)/2 \\ 1, & i > (3^J-1)/2, \end{cases} \qquad i \in \mathbb{Z},$$

shows that $\|v^j - \tilde{v}^j\|_{\ell_p(\mathbf{Z})} = 1$ for all $j = 0, \ldots, J$ and $1 \le p \le \infty$. Thus, L_p stability cannot hold if $1 \le p < \infty$.

General results on the L_p stability of the multi-scale reconstruction (25) and of the subdivision scheme (26) for S with the representation (19) are developed in [13, 49] for $1 \le p \le \infty$, and more recently for S which is offset invariant and for $p = \infty$ in [36] and [2]. We start with formulating the main result of [13, 49] for our definition of L_p stability (note that these papers deal with the limit case $J \to \infty$, and consider both the L_p and Besov space settings).

Theorem 2.4. *In addition to the assumptions of Theorem 2.2, assume that the family $\{S_v\}$ is Lipschitz continuous as function of $v \in \ell_p(\mathbb{Z})$:*

$$\|S_v - S_w\|_{\ell_p(\mathbf{Z}) \to \ell_p(\mathbf{Z})} \le C \|v - w\|_{\ell_p(\mathbf{Z})} \qquad \forall w, v \in \ell_p(\mathbb{Z}). \tag{43}$$

If $\tilde{\rho}_{p,k}(\{S_v\}) < 1$ then (25) is L_p stable.

Whether L_p stability holds under the weaker and more natural condition $\tilde{\rho}_{p,k}(\{S_v\}) < r^{1/p}$ is an open question. In [13], L_p stability of point-interpolation and cell-average based WENO subdivision (Example 1) is established using Theorem 2.4.

However, condition (43) limits the applicability of Theorem 2.4, as the natural assumption of Lipschitz continuity of the original S does not automatically carry over to the family $\{S_v\}$. Indeed, (43) fails to hold for many concrete multi-scale transforms. Examples 2 and 5 fall into this category. A stability criterion which circumvents this difficulty and is directly based on S has recently been formulated in [36, 2] for $p = \infty$, and has its roots in earlier case studies for some convexity-preserving schemes such as the power-2 subdivision from Example 5, [45, 5]. We formulate the result from [36], and extend it to the whole range $1 \le p \le \infty$. For simplicity, we first state it for $k = 1$.

Theorem 2.5. *Let S be an r-shift invariant, local, offset invariant for $\mathbb{P}_1$, and Lipschitz continuous subdivision operator, and let P be bounded and Lipschitz continuous. Then the existence of a ρ, $0 < \rho < 1$, and an integer $n \ge 1$ such that for any two sets $\{v^0, d^j\}$, $\{\tilde{v}^0, \tilde{d}^j\}$ of multi-scale data we have the inequality*

$$r^{-n/p}\|\Delta(v^n - \tilde{v}^n)\|_{\ell_p(\mathbf{Z})} \le \rho\|\Delta(v^0 - \tilde{v}^0)\|_{\ell_p(\mathbf{Z})} + C\sum_{l=1}^{n} r^{-l/p}\|d^l - \tilde{d}^l\|_{\ell_p(\mathbf{Z})}, \quad (44)$$

implies the L_p stability of the multi-scale reconstruction (25).
If (44) holds in the special case when $v^0, \tilde{v}^0 \in \ell_p(\mathbb{Z})$ are arbitrary but $d^j = \tilde{d}^j = \mathbf{0}$, $j = 1, \dots, n$, then at least the subdivision scheme (26) is L_p stable.

The statement of this theorem carries over to $k > 1$ if an estimate of the form

$$\|Sv - Sw\|_{\ell_p(\mathbf{Z})} \le r^{1/p}\|v - w\|_{\ell_p(\mathbf{Z})} + C\|\Delta^k(v - w)\|_{\ell_p(\mathbf{Z})} \quad (45)$$

can be established, and if (44) holds with Δ replaced by Δ^k, see [36]. Moreover, in [36] condition (44) is replaced by a spectral radius estimate on the derivatives of the derived subdivision operators. In the next theorem we formulate this result under the simplifying condition, that all functions ϕ_s in the definition (16) of S possess uniformly bounded, continuous partial derivatives, and refer to [36] for the exact conditions of piecewise continuous differentiability under which the statement can be proved. Denote by $D_v S^{[k]}$ the Frechet derivative of $S^{[k]}$ at $v \in \ell_p(\mathbb{Z})$. Due to our simplifying assumption, the linear operator family $D_v S^{[k]} : \ell_p(\mathbb{Z}) \to \ell_p(\mathbb{Z})$ depends continuously on v. Now define the spectral radii $\rho_{p,k}^{stab}(S) = \rho_p^{stab}(S^{[k]})$ and $\tilde{\rho}_{p,k}^{stab}(S) = \tilde{\rho}_p^{stab}(S^{[k]})$ as follows:

$$\rho_p^{stab}(S^{[k]}) := \limsup_{j\to\infty} \sup_{w\in\ell_p(\mathbf{Z})} \|DS^{[k]}_{(S^{[k]})^{j-1}w} DS^{[k]}_{(S^{[k]})^{j-2}w} \cdots DS^{[k]}_{w}\|^{1/j}_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})}, \quad (46)$$

and

$$\tilde{\rho}_p^{stab}(S^{[k]}) := \limsup_{j\to\infty} \sup_{w^0, w^1, \dots, w^{j-1}\in\ell_p(\mathbf{Z})} \|DS^{[k]}_{w^{j-1}} DS^{[k]}_{w^{j-2}} \cdots DS^{[k]}_{w^0}\|^{1/j}_{\ell_p(\mathbf{Z})\to\ell_p(\mathbf{Z})}. \quad (47)$$

Note that in (47) the supremum is taken with respect to an arbitrary collection of w^l, $l = 0, \dots, j-1$, while in (46) it is taken with respect to a single w. Set $w^l = (S^{[k]})^l w$, $l = 0, \dots, j-1$, to see that

$$\rho_{p,k}^{stab}(S) \le \tilde{\rho}_{p,k}^{stab}(S).$$

As demonstrated in [36] for the dyadic median interpolating scheme, this inequality can be strict. On the other hand, in [33, Lemma 4.2] it was observed that

$$\lim_{n\to\infty} \tilde{\rho}_p^{stab}((S^{[k]})^n)^{1/n} = \tilde{\rho}_{p,k}^{stab}(S),$$

a property that is used there for establishing approximation results.

Theorem 2.6. *Let S be an r-shift invariant, local, Lipschitz continuous subdivision operator, and let P be bounded and Lipschitz continuous. In addition, assume that S is offset invariant for $\mathbb{P}_k$, and that (45) holds.*
i) The multi-scale reconstruction (25) is L_p stable if $\tilde{\rho}_{p,k}^{stab}(S) < r^{1/p}$, and L_p stability fails to hold when $\tilde{\rho}_{p,k}^{stab}(S) > r^{1/p}$.
ii) The subdivision algorithm (26) is L_p stable if $\rho_{p,k}^{stab}(S) < r^{1/p}$ while in the case $\rho_{p,k}^{stab}(S) > r^{1/p}$ it is not.

For the proof and the generalization to certain classes of piecewise-differentiable S, we refer to [36, Section 2.3] in the case $p = \infty$. The extension to $1 \le p < \infty$ is straightforward. The L_∞ stability of median-interpolating and power-p multi-scale transforms and subdivision schemes is partially resolved in [36] using Theorem 2.6. For the power-2 scheme (also called PPH scheme) see [45, 5]. In particular, (25) is L_∞ stable for the power-p case if $1 \le p < \frac{8}{3}$, and fails to be L_∞ stable if $p > 4$. For the stability analysis of certain monotonicity- and convexity-preserving interpolating subdivision schemes introduced in [44, 46], see [35]. More examples of univariate nonlinear multi-scale transforms, e.g., a triadic version of the power-p scheme or a non-interpolatory PPH scheme with smooth limit functions, can be found in [14, 3].

2.5 Approximation order and decay of details

In this subsection, we study the L_p convergence of nonlinear multi-scale transforms from a different angle. In contrast to subsection 2.3, where for given multi-scale data $\{v^0, d^j\}_{j\ge1}$ issues such as L_p convergence and Besov regularity of limit functions are the main concern, we here infer properties of $\{v^0, d^j\}_{j\ge1}$ from properties of f.

In more practical terms, let us interpret the decomposition part (41) of the multi-scale transform as a process of taking sampling values of a smooth function f with respect to Γ^j, and collecting them into the "sampling" vectors v^j. This is, we assume $v^j = R_j f$ for a certain sequence of sampling operators $R_j : L_p(\mathbb{R}) \to \ell_p(\mathbb{Z})$ which also implicitly define R via $R_{j-1} = RR_j$, $j \ge 1$. E.g., for $p = \infty$ and $f \in C(\mathbb{R})$ sampling by function evaluation at Γ^j given by $v_i^j := f(r^{-j}i)$, $i \in \mathbb{Z}$, $j \ge 0$, is often used, and compatible with trivial down-sampling R given by $(Rv^j)_i = v_{ri}^j$. For

sampling L_p functions ($1 \le p \le \infty$), often

$$v_i^j = (R_j f)_i := \int_{\mathbf{R}} f(t) r^j \tilde{\phi}(r^j t - i)\,\mathrm{d}t = \int_{\mathbf{R}} f(r^{-j}(x+i))\tilde{\phi}(x)\,\mathrm{d}x, \quad i \in \mathbb{Z}, \tag{48}$$

where $\tilde{\phi}(t) \in L_{p/(p-1),loc}(\mathbf{R})$ has compact support and satisfies $\int_{\mathbf{R}} \tilde{\phi}(x)\,\mathrm{d}x = 1$. If we insist on compatibility with (41) with a local R then $\tilde{\phi}$ needs to be refinable,

$$\tilde{\phi}(x) = r \sum_{l \in \mathbf{Z}} b_l \tilde{\phi}(rx - l), \tag{49}$$

with finitely supported coefficient sequence $\{b_l\}$. The restriction operator R has then the form discussed in the beginning of subsection 2.4. E.g., taking B-splines of order m as $\tilde{\phi}$ is a common choice, if $m = 1$ then the sampling is equivalent with taking averages on dyadic intervals. Note that trivial down-sampling formally results if $\tilde{\phi} = \delta_0$ is the delta-function which satisfies (49) with the coefficients $b_0 = 1$, $b_i = 0$, $i \neq 0$. In the statement below we will silently include locally supported refinable Radon measures for the generation of sampling operators if $p = \infty$.

All sampling procedures discussed in the literature are linear. The question of how to deal with multi-scale transforms with nonlinear R is open. For example, no natural sampling compatible with the R defined in Example 2 comes to mind. Median sampling on dyadic intervals as R_j is not suitable since $R_{j-1} = RR_j$ is violated, as simple examples show. In the remainder of this subsection, we therefore discuss only linear R_j. We also assume that the sequence f^j of linear splines associated with $v^j = R_j f$ converges to f for any $f \in L_p(\mathbf{R})$. For the R_j of (48) this assumption is easy to check.

We discuss the following two questions on the multi-scale transform (9). The first question is to give sharp estimates for the resulting sequence $\{v^0, d^j\}_{j \ge 1}$ in terms of the smoothness class f belongs to. These are usually called *direct or Jackson-type estimates*. For f from Besov-Hölder classes this amounts to proving inverse inequalities to (33). Implicitly, this means to characterize the decay of the detail sequences d^j. The second, related question is the *approximation order* of the multi-scale transform which addresses the simpler question of how to relate the sampling information $v^j = R_j f$ directly to f, applying only the subdivision scheme, without having access to the detail sequences d^l with $l > j$. Roughly speaking, we speak of approximation order s if any f of smoothness s can be reconstructed using S within error $\mathrm{O}(h^s)$ from its samples with respect to a grid of step-size h.

Definition 2.6. We say that the multi-scale transform (9) has L_p approximation order $s > 0$, if for any $f \in B_p^s(\mathbf{R})$ and any $j \ge 0$, the reconstruction $f_j^\infty \in L_p(\mathbf{R})$ from the sampling sequence v^j exists, i.e., the linear spline interpolants for the sequence $v^j, Sv^j, S^2v^j, \ldots$ with respect to the grids $\Gamma^j, \Gamma^{j+1}, \Gamma^{j+2}, \ldots$ converge to f_j^∞ in $L_p(\mathbf{R})$, and that $f_j^\infty \to f$ in $L_p(\mathbf{R})$ at rate s:

$$\|f - f_j^\infty\|_{L_p(\mathbf{R})} = \mathrm{O}(r^{-js}), \qquad j \to \infty. \tag{50}$$

Checking approximation order is closely related to direct estimates via L_p stability of the multi-scale reconstruction (25). To this end, it is enough to realize that f_j^∞ is the L_p limit of the reconstruction (25) with the "perturbed" data $\{v^0, d^1, \dots, d^j, \mathbf{0}, \mathbf{0}, \dots\}$. Suppose a direct estimate of the form

$$\|\{v^0, d^j\}_{j\geq 1}\|_{p,s;r} \leq C\|f\|_{B_p^s(\mathbf{R})}, \qquad \forall\, f \in B_p^s(\mathbf{R}), \tag{51}$$

is established. Then $f_j^\infty \in L_p(\mathbf{R})$ is well-defined, and using L_p stability we have

$$\begin{aligned}\|f - f_j^\infty\|_{L_p(\mathbf{R})} &\leq C\sum_{l>j} r^{-l/p}\|d^l\|_{\ell_p(\mathbf{Z})} \\ &\leq Cr^{-js}\left(\sum_{l>j} r^{(sp-1)l}\|d^l\|_{\ell_p(\mathbf{Z})}^p\right)^{1/p} \leq Cr^{-js}\|f\|_{B_p^s(\mathbf{R})}.\end{aligned}$$

In [33], this reduction is observed for $p = \infty$. There it is also shown that, under some minor additional conditions, even stability of the subdivision scheme (26) is sufficient to prove the reduction. The papers [17, 66] address related issues for manifold-valued subdivision. Note that the question of approximation order could also be discussed without reference to the full multi-scale transform (9), just as a property of sampling operators $\{R_j\}$, $j \geq 0$, on the one hand, and the subdivision operator S, on the other. Such an approach could be of interest, when a subdivision scheme is used for reconstruction, and multi-scale decomposition and detail decay are not important.

Thus, for the rest of this subsection we concentrate on direct estimates. Although this is a well-studied subject in the linear case [10, 15], very little is known for nonlinear multi-scale transforms (in particular, no results are known if R is also nonlinear). Some results in this direction can be found in [49, Section 2.1]. The following statement covers most of them, and its proof is a straightforward extension of their proofs.

Theorem 2.7. *Let the sampling operators R_j be given by (48), where $\tilde{\phi} \in L_{p/(p-1),loc}(\mathbf{R})$ is refinable. If the operator SR is bounded and offset exact for $\mathbf{P}_k$, i.e.,*

$$SR(v + q|_{\mathbf{Z}}) = SRv + q|_{\mathbf{Z}}, \qquad \forall\, q \in \mathbf{P}_k, \quad \forall\, v \in \ell_p(\mathbf{Z}),$$

and if

$$\|d\|_{\ell_p(\mathbf{Z})} \leq C\|Pd\|_{\ell_p(\mathbf{Z})}, \qquad \forall\, d \in \ell_p(\mathbf{Z}),$$

then the multi-scale transform satisfies the Jackson-type estimate (51) for all $0 < s < k$.

Note that this result extends to $s = k$ if in Definition 2.6 $B_p^k(\mathbf{R})$ is replaced by the Besov space $B_{p,\infty}^k(\mathbf{R})$, and also applies to point-evaluation sampling if $p = \infty$. Offset exactness of SR is more restrictive than offset invariance, and does not hold if $\{R_j\}$ and S are arbitrarily paired. If S is given then special efforts could go into constructing R and $\tilde{\phi}$ such that $R_j = RR_{j+1}$ holds, and SR becomes offset exact. If R_j is given

by point evaluation then offset exactness of SR for $\mathbb{P}_k$ can be reduced to polynomial exactness of S for $\mathbb{P}_k$ which yields the more familiar bounding statements in [49].

To illustrate the use of Theorem 2.7, consider again Example 5. The power-p scheme uses trivial down-sampling for R and dilation factor $r = 2$. This is compatible with defining $(R_j f)_i := f(2^{-j}i)$, and we are restricted to applying the L_∞ case of the theorem. Since $\Delta^2 q|_{\mathbf{Z}} = \mathbf{0}$ for any linear polynomial q, it is easy to check that SR is offset exact for $\mathbb{P}_2$. Thus, $f \in B^s_\infty(\mathbb{R})$ implies the estimate

$$\|d^j\|_{\ell_\infty(\mathbf{Z})} \le C 2^{-js} \|f\|_{B^s_\infty(\mathbf{R})}$$

for the decay rate of the detail coefficients resp. the estimate

$$\|f - f_j^\infty\|_{L_\infty(\mathbf{R})} \le C 2^{-js} \|f\|_{B^s_\infty(\mathbf{R})}$$

for the approximation rate for s in the range $0 < s \le 2$. These estimates cannot be expected to hold for larger s, since power-p schemes reduce to linear interpolation near points of inflection.

3 The geometric setting: Case studies

In this section we consider geometric multi-scale transforms and geometry-based subdivision schemes, again in the univariate case. The prediction operator S appearing in these schemes is also termed as a refinement step, and we present several new such refinement steps. In contrast to the functional setting, geometric schemes operate on vector data (vertex points, edges, and normal vectors of polygonal lines) in a way that prevents us from analyzing them componentwise. So far such nonlinear vector subdivision schemes and multi-scale transforms have been investigated in case studies only, and tools for their systematic analysis have yet to be developed.

In subsection 3.1, we deal with the issue of convergence of few examples of curve subdivision schemes, defined by geometry-based refinement steps, and discuss properties of the limits generated by the schemes. Subsection 3.2 is devoted to geometric multi-scale transforms for planar curves based on the idea of normal multiresolution [34, 19, 55] (discussed briefly as the third example in the Introduction). We suggest several new multi-scale transforms, mimicking the original normal multiresolution scheme, but with the linear S there replaced by a nonlinear geometry-based refinement rule. These multi-scale transforms have two sources of nonlinearity, the one is the nonlinearity in the prediction, and the other is the inherent nonlinearity in the definition of the details.

3.1 Geometry-based subdivision schemes

A geometry-based refinement step depends on all the components of the points involved simultaneously, in contrast to the linear case, where the refinement step is applied separately to each component. Therefore linear subdivision schemes in the geometric setting are analyzed by the tools for linear subdivision schemes in the functional setting. In contrast, the geometry-based subdivision schemes are the nonlinear analogue of linear vector subdivision schemes, with refinement steps defined by operations of matrices on vectors.

3.1.1 Three types of geometry-based nonlinear 4-point schemes

In this subsection we present three geometry-based 4-point schemes, all related to the linear 4-point scheme in different ways. The refinement rule of the linear 4-point scheme is

$$(S_w\mathscr{P})_{2i} = P_i, \quad (S_w\mathscr{P})_{2i+1} = -w(P_{i-1}+P_{i+2})+(\frac{1}{2}+w)(P_i+P_{i+1}). \tag{52}$$

In the above, $\mathscr{P}$ denotes a control polygon, namely a polygonal line through a sequence of points (control points), denoted by $\{P_i\}$, and w is a fixed tension parameter (to avoid discussions about boundary treatment, assume a closed or bi-infinite control polygon). This refinement rule includes the Deslauriers-Dubuc scheme (1) for $w = \frac{1}{16}$, and the linear B-spline scheme for $w = 0$ as special cases.

It is well known that the scheme given by (52) has the following attributes:

- It generates "good" curves when applied to control polygons with edges of comparable length.
- It generates curves which become smoother (have greater Hölder exponent of the first derivative), the closer the tension parameter is to $\frac{1}{16}$.
- Starting from an initial control polygon with edges of significantly different length, S_w with a tension parameter around $\frac{1}{16}$, may generate curves with artifacts.
  An artifact is a geometric feature of the generated curve which does not exist in the initial control polygon, such as an inflection point or a self-intersection point.
- S_w generates a curve which preserves the shape of an initial control polygon with edges of significantly different length, only for very small values of w. (Recall that the control polygon itself corresponds to the generated curve with zero tension parameter.)

Displacement-safe 4-point schemes. This geometry-based version of the 4-point scheme is introduced in [48], and adapts the tension parameter w in (52) to the geometry of the four control points taking part in the definition of an inserted point (a refined point with an odd index). The failure of the 4-point scheme with a fixed

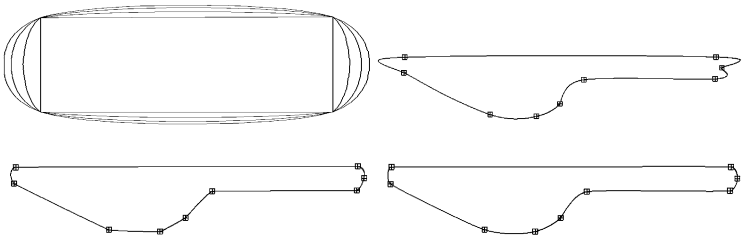


Fig. 5 Curves generated by the linear 4-point scheme: (Upper left) the effect of different tension parameters, (upper right) artifacts in the curve generated with $w = \frac{1}{16}$, (lower left) artifact-free but visually non-smooth curve generated with $w = 0.01$. Artifact-free and visually smooth curve generated in a nonlinear way with adaptive tension parameters (lower right)

tension parameter to generate smooth looking artifact free curves, when the edges of the initial control polygon are of significantly different length, is demonstrated in Figure 5. Also shown there is a high quality curve generated by a scheme with adaptive tension parameter.

To derive the refinement step with adaptive tension parameter, we write the insertion rule in (52) in terms of the edges $\{e_j = P_{j+1} - P_j\}$ of the control polygon, and relate the inserted point to the edge e_j. The insertion rule can thus be written in the form,

$$(S\mathscr{P})_{2j+1} = P_{e_j} = M_{e_j} + w_{e_j}(e_{j-1} - e_{j+1}) \tag{53}$$

with M_{e_j} the midpoint of e_j, w_{e_j} the adaptive tension parameter, and P_{e_j} the inserted point relative to the edge e_j. Defining $d_{e_j} = w_{e_j}(e_{j-1} - e_{j+1})$ as the displacement from M_{e_j}, we control its size by choosing w_{e_j} according to a geometrical criterion.

In [48] there are various geometrical criteria, all of them guaranteeing that the inserted control point P_{e_j} is different from the boundary points of the edge e_j, and that the length of each of the two edges replacing e_j is bounded by the length of e_j. One way to achieve these goals is to choose w_{e_j} so that

$$\|d_{e_j}\| \le \frac{1}{2}\|e_j\| \,. \tag{54}$$

The resulting schemes are termed *displacement-safe*. In all these schemes the value of the tension parameter w_{e_j} is restricted to the interval $(0, \frac{1}{16}]$, such that a tension parameter close to $\frac{1}{16}$ is assigned to *regular stencils*, namely to stencils of four points with three edges of almost equal length, while the *less regular* the stencil is, the closer to zero is the tension parameter assigned to it.

A *natural* choice of an adaptive tension parameter in $(0, \frac{1}{16}]$ obeying (54) is

$$w_{e_j} = \min\left\{\frac{1}{16}, c\,\frac{\|e_j\|}{\|e_{j-1} - e_{j+1}\|}\right\}, \quad \text{with a fixed } c \in [\frac{1}{8}, \frac{1}{2}) \,. \tag{55}$$

In (55), the value of c is restricted to the interval $[\frac{1}{8}, \frac{1}{2})$ to guarantee that $w_{e_j} = \frac{1}{16}$ for stencils with $\|e_{j-1}\| = \|e_j\| = \|e_{j+1}\|$. To see this, observe that in this case, $\|e_{j-1} - e_{j+1}\| = 2\sin\frac{\theta}{2}\|e_j\|$, with $\theta \in [0,\pi]$ the angle between the two vectors e_{j-1}, e_{j+1}. Thus we have $\|e_j\|/\|e_{j-1} - e_{j+1}\| = (2\sin\frac{\theta}{2})^{-1} \geq \frac{1}{2}$, and if $c \geq \frac{1}{8}$ then the minimum in (55) is $\frac{1}{16}$. The choice (55) defines irregular stencils (corresponding to small w_{e_j}) as those with $\|e_j\|$ much smaller than at least one of $\|e_{j-1}\|, \|e_{j+1}\|$, and such that when these two edges are of comparable length, the angle between them is not close to zero.

The convergence of this geometric 4-point scheme, and the continuity of the limits generated, follow from a result in [47]. There it is proved that the 4-point scheme with a varying tension parameter is convergent, and that the limits generated are continuous, whenever the tension parameters are restricted to the interval $[0, \tilde{w}]$, with $\tilde{w} < \frac{1}{8}$.

Yet, the result in [47] about C^1 limits of the 4-point scheme with a varying tension parameter does not apply to the geometric 4-point scheme defined by (53) and (55), since the tension parameters used during this subdivision process are not bounded away from zero.

Nevertheless, many simulations indicate that the curves generated by this scheme are C^1 (see [48]).

Parametrization-based 4-point schemes. This type of schemes is introduced and investigated in [24]. The idea for the geometric insertion rule comes from the insertion rule of the linear scheme with $w = \frac{1}{16}$, corresponding to the Deslaurier-Dubuc 4-point scheme. The point $(S_{1/16}\mathscr{P})_{2i+1}$ is obtained by the evaluation of a cubic polynomial interpolating the data $\{(i-k, P_{i-k}) : k = -1, 0, 1, 2\}$ at the point $i + \frac{1}{2}$ (see [20]). From this point of view, the linear scheme corresponds to a uniform parametrization of the control polygon at each refinement level. This approach fails when the initial control polygon has edges of significantly different length. Yet the use of the centripetal parametrization, instead of the uniform parametrization, leads to a geometric 4-point scheme with artifact-free limit curves, as can be seen in Figure 6.

The centripetal parametrization, which is known to be effective for interpolation of control points by a cubic spline curve (see [28]), has the form $\mathbf{t}_{\mathbf{cen}}(\mathscr{P}) = \{\tau_i\}$, with

$$\tau_i = \tau_{i-1} + \|P_i - P_{i-1}\|_2^{1/2}, \tag{56}$$

where $\|\cdot\|_2$ is the Euclidean norm, and $\mathscr{P} = \{P_i\}$.

Let $\mathscr{P}^j$ be the control polygon at refinement level j, and let $\{\tau_i^j\} = \mathbf{t}_{\mathbf{cen}}(\mathscr{P}^j)$. The refinement rule for the geometric 4-point scheme, based on the centripetal parametrization is:

$$P_{2i}^{j+1} = P_i^j, \quad P_{2i+1}^{j+1} = \pi_{j,i}\Big(\frac{\tau_i^j + \tau_{i+1}^j}{2}\Big),$$

with $\pi_{j,i}$ the vector of cubic polynomials, satisfying the interpolation conditions

$$\pi_{j,i}(\tau_{i+k}^j) = P_{i+k}^j, \quad k = -1, 0, 1, 2.$$

Note that this construction can be done with any parametrization. In fact in [24] the chordal parametrization ($\tau_{i+1} - \tau_i = \|P_{i+1} - P_i\|_2$) is also investigated, but found to be inferior to the centripetal parametrization (see Figure 6).

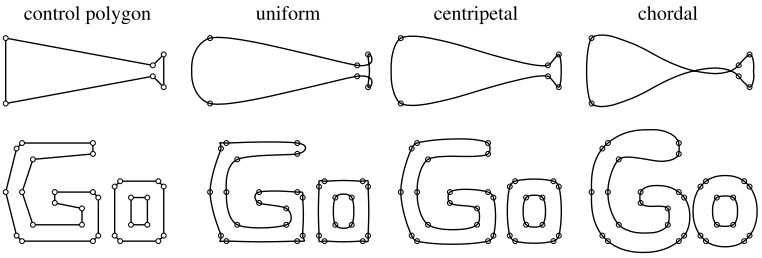


Fig. 6 Comparisons between 4-point schemes based on different parameterizations

The analysis of the schemes in [24] is rather ad-hoc. It is shown there that the centripetal and chordal schemes are well defined, in the sense that any inserted point is different from the end points of the edge to which it corresponds, and that both schemes are convergent to continuous limit curves. Although numerical simulations indicate that both schemes generate C^1 curves, as does the linear 4-point scheme, there is no proof in [24] of such a property for the geometric schemes.

Another type of information on the limit curves, which is relevant to the absence or presence of artifacts, is available in [24]. Bounds on the Hausdorff distance, $d_{\mathscr{H}}$ (see (63)), from sections of a limit curve to their corresponding edges in the initial control polygon are derived. These bounds give a partial qualitative understanding of the empirical observation that the limit curves corresponding to the centripetal parametrization are artifact free.

Let $\mathscr{C}$ denote a curve generated by the scheme based on the centripetal parametrization from an initial control polygon $\mathscr{P}^0$. Since the scheme is interpolatory, $\mathscr{C}$ passes through the initial control points. Denote by $\mathscr{C}|_{e_i^0}$ the section of $\mathscr{C}$ with boundary points as those of e_i^0. Then

$$d_{\mathscr{H}}(\mathscr{C}|_{e_i^0}, e_i^0) \le \frac{5}{7}\|e_i^0\|_2.$$

Thus the section of the curve corresponding to a short edge cannot be too far from its edge. On the other hand the corresponding bound in the linear case has the form

$$d_{\mathscr{H}}(\mathscr{C}|_{e_i^0}, e_i^0) \le \frac{3}{13}\max\{\|e_j^0\|_2\ ,\ \ |j-i| \le 2\},$$

and a section of the curve can be rather far from its corresponding short edge, if this edge has a long neighboring edge. In the case of the chordal parametrization the bound is even worse

$$d_{\mathscr{H}}(\mathscr{C}|_{e_i^0}, e_i^0) \leq \frac{11}{5} \max\{\|e_j^0\|_2\ ,\ \ |j-i| \leq 2\}.$$

Comparisons of the performance of the three 4-point schemes, based on uniform, chordal and centripetal parametrization, are given in Figure 6.

Circle preserving 4-point scheme. While the first two types of geometric 4-point schemes were designed to alleviate artifacts in the geometry (position) of the limit curves, this geometric version of the 4-point scheme was designed to overcome artifacts in the numerical curvature generated by the linear scheme [57]. The scheme is *circle preserving* in the sense that if the initial control points are ordered points on a circle, then the limit curve is that part of the circle between the first and the last initial control points.

The insertion rule requires geometric computations, as the inserted point is an intersection point between a circle and a sphere. The details of the computation of an inserted point are given in [57] as an algorithm.

Fig. 7 A control polygon and the limit curve generated by the circle preserving variant of the 4-point scheme

Figure 7 from the above paper, demonstrates the limit curve obtained from a control polygon with slowly varying numerical curvature. The numerical curvature of the limit curve is varying smoothly.

It is shown in [57] that the scheme is asymptotically equivalent to the linear 4-point scheme, and therefore according to [25] the scheme is convergent and generates continuous limit curves. Numerical simulations indicate that the scheme generates C^1 limit curves.

The available analysis of the above three types of geometric 4-point schemes, is limited to showing convergence and continuity of the limit curves. The proof of C^1 seems to be much harder due to to the lack of an appropriate parametrization for the geometrically defined curves. Perhaps the proof should be in terms of geometric arguments, such as the continuity of the tangents of the curves.

3.1.2 Convexity-preserving schemes in the plane

A shape property of planar control polygons, which is important to preserve in the curves generated by subdivision, is convexity. Here we present three different subdivision schemes which are convexity preserving in the geometric setting, namely they refine convex polygons into convex polygons.

We first introduce some geometrical notions related to polygonal lines and to convexity.

- An edge in a polygonal line such that its two neighboring edges are in the same half-plane, determined by the line through the edge, is termed a "convex edge".
- An edge in a polygonal line which is on the same line as one of its neighboring edges is termed a "straight edge".
- A line through a vertex of a polygonal line, such that the two edges meeting at the vertex are on the same side of the line, is termed a "convex tangent". A "straight tangent" at a vertex is a line through one of the edges emanating from the vertex.
- A polygonal line consisting of convex and straight edges is termed a "convex polygon". It is a "strictly convex polygon" if all its edges are convex. In Figure 8 three examples of strictly convex polygons are given.

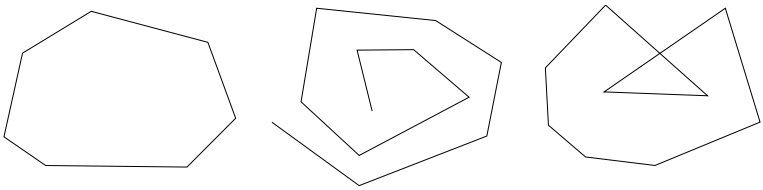


Fig. 8 Convex polygons: (left) closed, (middle) open, (right) self- intersecting

Among the three convexity preserving schemes presented here, two are nonlinear and geometry-based, while one is linear but of Hermite type. It refines control points and normals at the control points, and is inherently related to nonlinear schemes generating surfaces by refining control points and normals.

Convexity preserving 4-point scheme with adaptive tension parameter This scheme is a geometric variant of the 4-point scheme, similar to the displacement-safe schemes, but with the adaptive tension parameter chosen to preserve convexity. It is designed and analyzed in [48]. The scheme refines convex (strictly convex) control polygons into convex (strictly convex) control polygons. We describe the geometrical construction of the inserted points in terms of notation introduced in the first part of subsection 3.1.1.

As a first step in the construction, at each control point from which at least one convex edge emanates, a convex tangent is constructed. At the other control points a straight tangent is constructed, coinciding with one of the straight edges meeting at the control point. We denote the tangent at P_i by t_i.

In case of a straight edge e_i, $P_{e_i} = M_{e_i}$.

In case of a convex edge e_i, the tangents t_i and t_{i+1} together with e_i determine a triangle, T_{e_i}. By construction, the line through e_i separates T_{e_i} from the edges e_{i-1}, e_{i+1}. Thus the half-line starting from M_{e_i} along the direction $e_{i-1} - e_{i+1}$ intersects T_{e_i}. The point P_{e_i} is chosen so that $\|P_{e_i} - M_{e_i}\| / \|e_{i-1} - e_{i+1}\| \le \frac{1}{16}$ and that $P_{e_i} \in T_{e_i}$.

These two conditions for point insertion guarantee that $0 \leq w_{e_i} \leq \frac{1}{16}$ and that the refined control polygon $S\mathscr{P}$ obtained from $\mathscr{P}$ by the refinement rule

$$(S\mathscr{P})_{2i} = P_i \ , \ (S\mathscr{P})_{2i+1} = P_{e_i} \ ,$$

is convex (strictly convex) if the control polygon $\mathscr{P}$ is (see [48]).

This construction of refined control polygons, when repeated, generates a sequence of convex (strictly convex) polygons from an initial convex (strictly convex) polygon. It is proved by arguments similar to those cited in subsection 3.1.1 for the convergence of the displacement-safe schemes, that this sequence converges, and that the limit is a continuous convex (strictly convex) curve. Moreover, it is shown that the curve between two consecutive initial control points is either a line segment when the edge connecting the two points in the initial control polygon is straight, or otherwise a strictly convex curve.

Note that the subdivision scheme is interpolatory and that the inserted point between P_i and P_{i+1} depends on the points $P_{i-1}, P_i, P_{i+1}, P_{i+2}$ as in the linear 4-point scheme.

The convex tangents in this construction can be chosen in different ways. A natural choice of such a tangent is

$$t_i = P_{i+1} - P_{i-1} = e_i + e_{i-1}. \tag{57}$$

This choice was tested in many numerical experiments, and was found superior to other choices.

In Figure 9, the performance of this convexity-preserving scheme is compared on several examples with that of the displacement-safe scheme of subsection 3.1.1 and with that of the linear 4-point scheme.

The convexity-preserving 4-point scheme is extended in [48] to a co-convexity preserving scheme for general planar polygons.

Convexity preserving 2-point Hermite-type scheme. This convexity preserving scheme is a two-point interpolatory Hermite-type scheme. It operates on data in $\mathbb{R}^2$, consisting of control points and unit normals at the control points. It generates a convex limit curve from an initial convex data, namely a strictly convex control polygon, with compatible normals (compatible with the convexity).

The scheme is presented briefly in [16], as a first step in the construction of a nonlinear Hermite-type scheme for the generation of surfaces, interpolating the initial control points and the unit normals attached to them.

The insertion rule for the point between two consecutive points P_i, P_{i+1} is derived from the quadratic Bézier curve interpolating these two points and the normals at these points. First the mid control point, Q_i, of the quadratic Bézier curve is constructed, as the intersection point of the lines through P_i and P_{i+1}, which are orthogonal to the corresponding normals. The parametric midpoint of the Bézier curve determined by the control points P_i, Q_i, P_{i+1}, is the inserted point. It is given by

$$(S\mathscr{P})_{2i+1} = \frac{1}{4}(P_i + 2Q_i + P_{i+1}). \tag{58}$$

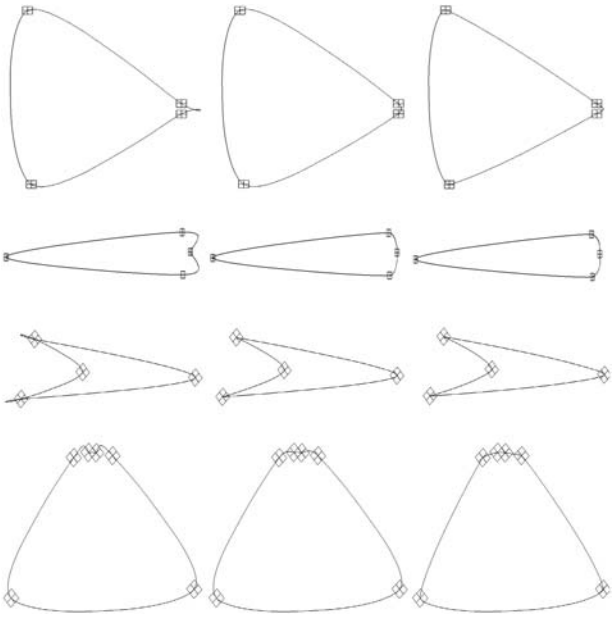


Fig. 9 Examples: (left column) the linear 4-point scheme with $w = 1/16$, (middle column) the displacement-safe scheme of subsection 3.1.1, (right column) the convexity preserving 4-point scheme

The normal at the refined point is the normal of the Bézier curve at this point, which is orthogonal to the direction $P_{i+1} - P_i$.

By construction the limit curve is a C^1 piecewise quadratic Bézier curve.

Convexity preserving scheme refining lines. This scheme is an extension of the 'dual' Chaikin scheme for lines, proposed by Sabin in [56]. It is an interpolatory scheme, which is used and analyzed in [27], as a first step towards the construction of a convexity preserving interpolatory scheme, operating on convex polyhedra and generating in the limit smooth convex surfaces. The scheme, although an interpolatory scheme refining control points, can be regarded as refining the support lines of the convex control polygon determined by the control points at each refinement level.

Given a strictly convex, closed control polygon $\mathscr{P}$, the first step in the construction of the inserted points, is the assignment of convex tangents $\{t_j\}$ (e.g. as in (57)) to the control points. Now, t_j, e_j, t_{j+1} determine a triangle T_{e_j}. The refined polygon $S\mathscr{P}$ is strictly convex if the inserted point between P_j and P_{j+1} is any point inside T_{e_j}. The rule for assigning convex tangents to the control points of $S\mathscr{P}$ is to keep the convex tangents at the control points of $\mathscr{P}$ and to choose convex tangents at the inserted points.

Denoting by $\langle\mathscr{P}\rangle$ the closed planar set enclosed by $\mathscr{P}$, and by $\mathscr{Q}$ the convex polygon generated by the convex tangents to the points of $\mathscr{P}$, it is easy to verify

that

$$\langle \mathscr{P} \rangle \subset \langle S\mathscr{P} \rangle \subset \langle \mathscr{Q} \rangle,$$

Thus repeated refinements, starting from an initial strictly convex, closed polygon, generate a sequence of "increasing" closed convex sets (in the sense of inclusion of sets), which are all contained in the closed convex set $\langle \mathscr{Q} \rangle$. This is sufficient to guarantee the convergence of the sequence of strictly convex, closed polygons to a continuous, closed convex curve. Moreover, it is proved in [27] by geometrical arguments, that the limit curve is C^1.

3.1.3 Ideas for designing new geometry-based schemes

Here we give three geometric constructions of refinement rules. The corresponding schemes have not been analyzed yet. All the schemes are interpolatory.

Interpolatory 4-point scheme based on circular arc approximation. For the construction of the inserted point between P_i and P_{i+1}, first two auxiliary points are constructed. These points are the mid-points of two circular arcs between P_i and P_{i+1}, one on the unique circular arc through P_i, P_{i+1}, P_{i+2} and the other on the unique circular arc through P_{i-1}, P_i, P_{i+1}. The inserted point is the midpoint of the two auxiliary points. It is a point on the line through M_{e_i} orthogonal to e_i. The resulting scheme is circle preserving by construction.

Interpolatory $2n$-point scheme based on the centripetal parametrization. This is an extension of the second geometric version of the 4-point scheme, presented in subsection 3.1.1. To determine the inserted point between P_i and P_{i+1}, one first parameterize the $2n$ points $P_{i+j},\ j = -n+1,\dots,n$ according to the centripetal parametrization (see (56)) to obtain the parameter differences $t_{i+j+1} - t_{i+j},\ j = -n+1,\dots,n-1$. Then the interpolating polynomial vector of degree $2n-1$ to the data $(t_{i+j}, P_{i+j}),\ j = -n+1,\dots,n$, is constructed and evaluated at the point $(t_i + t_{i+1})/2$ to yield the inserted point. This family of schemes is a geometric analogue of the Deslauriers-Dubuc family.

Convexity preserving interpolatory scheme based on quadratic Bézier curves. Given a strictly convex, closed control polygon, a refined strictly convex, closed control polygon is generated, by first assigning a convex tangent to each control point, and then computing the intersection points of consecutive convex tangents. The inserted point between P_i and P_{i+1} is the parametric midpoint (58) of the Bézier quadratic curve, determined by the three control points P_i, Q_i, P_{i+1}, where Q_i is the intersection point of the tangents at P_i and at P_{i+1}.

The rule for assigning convex tangents to the control points of the refined polygon, is to keep those at the 'old' control points, and to choose convex tangents at the inserted control points.

Note that an inserted point depends on four control points; two on each side.

3.2 Geometric multi-scale transforms for planar curves

We present here results on geometric multi-scale transforms for continuous curves in the plane, all based on the idea of normal multiresolution, which is discussed briefly in the Introduction. We also suggest new geometric constructions for normal multiresolution, which have still to be investigated.

3.2.1 The general structure of normal multiresolutions

Here we present again the main features of the normal multiresolution (NM), which aims at a multi-scale representation of curves in the plane, which can be encoded efficiently. The presentation is somewhat more general than that in Example 3 in section 1, to allow geometric prediction operators. In the following we assume that the curves are continuous.

Given a planar curve, $\mathscr{C}$, it is approximated by a sequence of polygonal lines $\{\mathscr{P}^j\}_{j\ge 0}$, where $\mathscr{P}^j = \{P_i^j\}$ with points P_i^j on the curve at refinement level j connected by edges $e_i^j = P_{i+1}^j - P_i^j$. As before, let us assume for simplicity that $\mathscr{C}$ is closed, that $\mathscr{P}^j$ contains $n_j = 2^j n_0$ points periodically enumerated by the index $i \in \mathbb{Z}$.

To obtain $\mathscr{P}^{j+1}$ from $\mathscr{P}^j$, the points $\{P_i^j\}$ are retained and denoted by $\{P_{2i}^{j+1}\}$, i.e., their indices are doubled, and between any two consecutive points $P_{2i}^{j+1}, P_{2(i+1)}^{j+1}$ a point from the curve segment $\mathscr{C}_i^j$ between the points P_i^j and P_{i+1}^j, is inserted with the index $2i+1$. The inserted point is obtained by a two-step procedure. First a prediction step $S^j\mathscr{P}^j$, with S^j an interpolatory prediction/subdivision operator, is performed which results in the point $\hat{P}_{2i+1}^{j+1} = (S^j\mathscr{P}^j)_{2i+1}$ near the curve segment $\mathscr{C}_i^j$. The prediction step is followed by a projection step, which determines a point on $\mathscr{C}$ as an intersection of $\mathscr{C}$ with the line orthogonal to the edge e_i^j through the predicted point $(S^j\mathscr{P}^j)_{2i+1}$ (see Figure 3). We denote this projection operator, acting on the predicted point and mapping it to the curve, by R^j. Note that for general S^j, the resulting point $P_{2i+1}^{j+1} := (R^jS^j\mathscr{P}^j)_{2i+1}$ is not necessarily on $\mathscr{C}_i^j$. Since the projection operator R^j is determined by the geometry of $\mathscr{P}^j$ and by the points $S^j\mathscr{P}^j$, it is a property of the prediction operator S^j and of $\mathscr{P}^j$ which guarantees that the inserted points are on the correct curve segments. We term a pair $\{S^j, \mathscr{P}^j\}$ *admissible for NM at level* j, if for all i

$$(R^jS^j\mathscr{P}^j)_{2i+1} \in \mathscr{C}_i^j.$$

Thus, if the pair $\{S^j, \mathscr{P}^j\}$ is admissible for NM at level j, then the polygonal line $\mathscr{P}^{j+1}$ consists of the vertices

$$P_{2i}^{j+1} = P_i^j, \quad P_{2i+1}^{j+1} = (R^jS^j\mathscr{P}^j)_{2i+1} \tag{59}$$

in a natural ordering along $\mathscr{C}$.

In the following we assume that the operators S^j are chosen in advance, e.g., as an insertion rule of an interpolatory linear subdivision scheme, or as an insertion rule of an interpolatory geometric subdivision scheme determined by the geometry of $\mathscr{P}^j$, so that these operators do not have to be encoded. Moreover we assume that for each $j \geq 0$ the pair $\{S^j, \mathscr{P}^j\}$ is admissible for NM at level j.

Remark: The set of pairs $\{S^j, \mathscr{P}^j\}$ admissible for NM at level j is nonempty. To see this consider the linear mid-point interpolatory subdivision scheme, S_0 in the notation of (52), with the refinement rule

$$(S_0\mathscr{P}^j)_{2i} = P_i^j, \;\; (S_0\mathscr{P}^j)_{2i+1} = \frac{1}{2}(P_i^j + P_{i+1}^j), \tag{60}$$

It is easy to verify that any pair $\{S_0, \mathscr{P}^j\}$, with $\mathscr{P}^j$ a polygonal line consisting of vertices sampled from the curve, is admissible for NM at any level. We term a scheme with this property *unconditionally admissible for NM*.

Defining the signed distances from the predicted points $S^j\mathscr{P}^j$ to their corresponding projected points $R^jS^j\mathscr{P}^j$ as *details at level* j, and denoting them by $d^j = \{d_i^j\}$, we observe that these details and the lines connecting pairs of corresponding predicted and projected points, are sufficient for computing the points of $\mathscr{P}^{j+1}$ from $\mathscr{P}^j$. Since these lines are determined by the information in $\mathscr{P}^j$, it follows that for the construction of $\mathscr{P}^{j+1}$ from $\mathscr{P}^j$ only the details at level j have to be encoded. Thus, the sequence of polygonal lines $\{\mathscr{P}^j\}_{j=1}^J$ can be reconstructed from the information

$$\mathscr{P}^0, d^j, \ldots, d^{J-1}. \tag{61}$$

The gain in the NM is that instead of encoding differences between points, which are vectors, we have to encode signed scalars, and to use the geometric information available at each level for the reconstruction of the next level.

In the following we discuss the case of linear prediction operators.

3.2.2 Normal multiresolutions with linear prediction operators

Denoting by S a linear prediction operator, and considering the stationary case $S^j = S$, equation (59) becomes

$$P_{2i}^{j+1} = P_i^j, \;\; P_{2i+1}^{j+1} = (R^jS\mathscr{P}^j)_{2i+1}. \tag{62}$$

It is clear that in this NM the nonlinearity/geometry is introduced by the projection operators R^j.

Note that in the stationary case $S^j = S$, with a linear S, the notion *admissible for NM at level* j can be replaced by *admissible for NM*. Moreover for a pair $\{S, \mathscr{P}^0\}$ to start a NM, we need the stronger notion *strongly admissible for NM*, namely that this pair and all the pairs $\{S, \mathscr{P}^j\}$ with $j > 0$ are admissible for NM, where $\mathscr{P}^j$ is the polygonal line generated by j refinement steps of the NM with S, starting from $\mathscr{P}^0$.

In [19] NMs with linear subdivision schemes as prediction operators are analyzed. The case of the mid-point prediction operator, given by (60), is easier to analyze and the results are better in some respects. The main issues addressed in [19] are the convergence of the reconstruction and the regularity of the limit, the rate of decay of the details, and the stability of the reconstruction. The rate of decay of the details is strongly related to the quality of the approximation of the curve by the polygonal lines $\{S\mathscr{P}^j\}$. The stability issue is concerned with the effect of small changes in the information (61), on the reconstructed polygonal lines $\{\mathscr{P}^j\}_{j=1}^J$. The NM is termed *stable* if the changes in the reconstructed polygonal lines due to small changes in the information are controlled. In a stable and converging NM, the polygonal lines (and hence the curve) can be well approximated without the small details, allowing a further reduction in the amount of encoded information.

Here we cite several results from [19] on the family of linear 4-point schemes $\{S_w\}$ with $w \in [0, \frac{1}{16}]$, where S_w is given by (52). These schemes constitute the main example in [19]. As is noted before, the scheme S_w with $w = 0$ corresponds to the mid-point scheme (60), and the scheme with $w = \frac{1}{16}$ corresponds to the Deslauriers-Dubuc 4-point scheme (1).

Although the results in [19] are derived in great generality, we limit our presentation to the above family of schemes. This alleviates the need to introduce the rather technical terminology, with which the general results are formulated.

To present the results, we first introduce the notion of the *regularity exponent* of a continuous curve $\mathscr{C}$ with a finite length $\ell(\mathscr{C})$. Let $(x(s), y(s)), \quad s \in [0, \ell(\mathscr{C})]$, be a representation of the curve in terms of the arc-length parametrization. The curve has Hölder regularity exponent $\nu = m + \mu$ with m a nonnegative integer and $\mu \in (0, 1]$, if both functions $x(s)$ and $y(s)$ have a continuous mth derivative which is Hölder continuous with exponent greater or equal to μ.

As is observed in subsection 3.2.1, the mid-point scheme S_0, is unconditionally admissible for NM, and hence can be used as the prediction operator for NM. It is shown in [19] that any member of the family of 4-point schemes $\{S_w\}$ with $w \in (0, \frac{1}{16}]$, can also serve as the prediction operator in NMs of smooth curves. More specifically,

Result 1: Let $\mathscr{C}$ have regularity exponent $\beta > 1$, and let $\mathscr{P}$ be a polygonal line with vertices sampled from $\mathscr{C}$. Then there exist $w \in (0, \frac{1}{16}]$, and a positive integer J, such that the pair $\{S_w, \mathscr{P}^J\}$ is strongly admissible for NM, where the polygonal line $\mathscr{P}^J$ is generated by J refinement steps of the NM with the mid-point rule, starting from $\mathscr{P}$,

Moreover, for any $w^* \in (0, \frac{1}{16}]$ there exists a positive integer J^* such that the pair $\{S_{w^*}, P^{J^*}\}$ is strongly admissible for NM, where P^{J^*} is the polygonal line generated by J^* refinement steps of the NM with S_w, starting from $\mathscr{P}^J$.

The advantage of using S_w with $w \neq 0$ in NMs of smooth curves is indicated by the next two results.

Result 2: Let $\mathscr{C}$ have regularity exponent $\beta > 1$, let $\mathscr{P}^0$ be a polygonal line consisting of sampled points from $\mathscr{C}$, and let $w \in (0, \frac{1}{16}]$. If the pair $\{S_w, \mathscr{P}^0\}$, is strongly admissible for NM, then the NM with S_w as a prediction operator, starting from $\mathscr{P}^0$ is stable and convergent.

Moreover the details of this NM $\{d^j\}_{j\geq 0}$ decay according to

$$\|d^j\|_\infty = \max_i |d_i^j| = O(j2^{-\min(2,\beta)j}).$$

In case $w = \frac{1}{16}$ and $\beta > 3$ the details decay according to

$$\|d^j\|_\infty = O(j2^{-3j}).$$

Result 3: Let $\mathscr{C}$ have regularity exponent $\beta > 0$. Then the NM with the mid-point prediction operator is stable and convergent.

Moreover the details of this NM decay according to

$$\|d^j\|_\infty = O(2^{-\min(2,\beta)j}).$$

It is easy to conclude from the last two results that if the smoothness of the curve is not known then the mid-point prediction operator should be used. Otherwise, the smoothness of the curve indicates which prediction operator to use, when aiming at small details and at a good approximation of the curve by the NM. Below we formulate these conclusions.

1. For a curve with regularity exponent $\beta > 3$ the details decay much faster with the predictor $S_{1/16}$ than with any other S_w with $w \in [0, \frac{1}{16})$.
2. For a curve with regularity exponent $\beta \in (2,3]$ all prediction operators S_w with $w \in (0, \frac{1}{16}]$ are superior to the mid-point prediction, with respect to the rate of decay of the details.
3. For a curve with regularity exponent $\beta \leq 2$, the details decay much faster with the mid-point prediction operator than with any S_w with $w \in (0, \frac{1}{16}]$.

3.2.3 Normal multiresolutions with nonlinear prediction operators

Here we suggest ideas for improving the NM by using nonlinear prediction operators. We can take S^j in (59) as one of the geometry-based schemes of subsections 3.1.1 and 3.1.3.

Among these geometry-based schemes several have the advantage of being unconditionally admissible for NM, the *displacement safe* 4*-point* schemes due to condition (54), and the *interpolatory* 4*-point scheme based on circular arc approximation* by the construction of the inserted point. In fact, with the latter prediction operator the NM generates the same polygonal lines $\{\mathscr{P}^j\}$ as those generated by the NM with the mid-point prediction, but the details are different.

We conjecture that for a curve with regularity exponent $\beta > 3$, the details in the NM with the *interpolatory* 4*-point scheme based on circular arc approximation* as prediction operator, decay as $O(2^{-3j})$. The conjecture is based on the following observation.

Observation: Let $\mathscr{C}$ be a curve with regularity exponent $\beta > 3$, and let the three points $P_i,\ i = 1,2,3$ be on $\mathscr{C}$, such that $h = \max_{i=1,2} \|P_{i+1} - P_i\|_2$ is small enough.

Then the circular arc through the three points P_i, $i = 1,2,3$ approximates the section of the curve between these three points, with error of order $O(h^3)$. (Here the error is measured by the Hausdorff metric, which is defined in the next subsection).

3.2.4 Adaptive approximation based on the NM with mid-point prediction

Here we discuss an algorithm for the adaptive approximation of planar curves, based on the NM with the mid-point prediction operator. This algorithm is presented and analyzed in [9]. We cite here quantitative results about the quality of the approximation, expressed in terms of the number of segments in the approximating polygonal lines. The results are stated with less details and not in their full generality.

For that we introduce some notation. Let I be a segment in a polygonal line with vertices sampled from $\mathscr{C}$. The curve segment between the boundary vertices of I is denoted by $\mathscr{C}_I$. The distance between two segments of curves γ, δ of finite length, is measured by the Hausdorff metric

$$d_{\mathscr{H}}(\gamma,\delta) = \max\{\mathscr{H}(\gamma,\delta), \mathscr{H}(\delta,\gamma)\}, \tag{63}$$

with the one-sided Hausdorff distance

$$\mathscr{H}(\gamma,\delta) = \max_{P\in\gamma}\min_{Q\in\delta}\|P-Q\|_2.$$

It is easy to see that $d_{\mathscr{H}}(\mathscr{C}_I,I) = \mathscr{H}((\mathscr{C}_I,I) \geq \mathscr{H}((I,\mathscr{C}_I)$.

Given an error tolerance ε, the adaptive algorithm refines a polygonal line $\mathscr{P}$ with vertices sampled from $\mathscr{C}$, by inserting a point according to the mid-point prediction between any two vertices corresponding to a segment I of $\mathscr{P}$ for which $\mathscr{H}((\mathscr{C}_I,I) > \varepsilon$. The point insertion and the computation of the corresponding detail, is according to the NM with the mid-point prediction. The algorithm terminates with a polygonal line $\mathscr{P}$ for which $\mathscr{H}(\mathscr{C}_I,I) \leq \varepsilon$ for all $I \in \mathscr{P}$. It is easy to note that the binary tree defined by this algorithm is a subtree of the binary tree generated by the NM with the mid-point prediction.

Here we cite an important result relating the number of segments in the final polygonal line obtained by the algorithm, to the error tolerance.

Result: Let $\mathscr{C}$ be a curve with finite length, and let $\varepsilon > 0$. Denote by $\mathscr{P}(\varepsilon)$ the polygonal line generated by the algorithm with the given error tolerance, and by $|\mathscr{P}(\varepsilon)|$ the number of segments in $\mathscr{P}(\varepsilon)$. Then there exists a constant $C(\mathscr{C})$, depending on the curve, such that

$$|\mathscr{P}(\varepsilon)| \leq \frac{C(\mathscr{C})}{\varepsilon}.$$

Moreover if the curve has finite curvature, then there is a constant, $\tilde{C}(\mathscr{C})$, depending on $\mathscr{C}$, such that

$$|\mathscr{P}(\varepsilon)| \leq \frac{\tilde{C}(\mathscr{C})}{\varepsilon^{\frac{1}{2}}}.$$

The above result indicates that for a curve of finite length the algorithm generates polygonal lines with error decreasing linearly with the inverse of the number of segments. The error decreases as the inverse of the square of the number of segments, for curves with finite curvature.

References

1. Amat, S., Arandiga, F., Cohen, A., Donat, R.: Tensor product multiresolution analysis with error control for compact image representation. Signal Processing **82**, 587–608 (2002)
2. Amat, S., Dadourian, K., Liandrat, J.: Analysis of a class of subdivision schemes and associated non-linear multiresolution transforms. (submitted 2008), arXiv:0810.1146[math.NA]
3. Amat, S., Dadourian, K., Liandrat, J.: On a C2-nonlinear subdivision scheme avoiding Gibbs oscillations. (submitted 2008), arXiv:0812.2562[math.NA]
4. Amat, S., Donat, R., Liandrat, J., Trillo, J.C.: Analysis of a new non-linear subdivision scheme: Applications to image processing. Found. Comput. Math. **6**, 193–226 (2006)
5. Amat, S., Liandrat, J.: On the stability of the PPH nonlinear multiresolution. Appl. Comput. Harmon. Anal. **18**, 198–206 (2005)
6. Arandiga, F., Donat, R.: Nonlinear multi-scale decompositions: The approach of A. Harten. Numerical Algorithms **23**, 175–216 (2000)
7. Arandiga, F., Cohen, A., Donat, R., Dyn, N., Matei, B.: Approximation of piecewise smooth functions and images by edge-adapted (ENO-EA) nonlinear multiresolution techniques. J. Comput. Harmon. Anal. **24**, 225–250 (2008)
8. Baraniuk, R.G., Claypoole Jr., R.L., Davis, G.M., Sweldens, W.: Nonlinear wavelet transforms for image coding via lifting. IEEE Trans. Image Process. **12**, 1449–1459 (2003)
9. Binev, P., Dahmen, W., DeVore, R., Dyn, N.: Adaptive approximation of curves. In: Dimitrov, D.K., Nikolov, G., Ulichov, R. (eds.), Approximation Theory: a volume dedicated to Borislav Boyanov, pp. 43–57, Martin Drinov Academic Publishing House, Sofia (2004)
10. Cavaretta, A.S., Dahmen, W., Micchelli, C.A.: Stationary subdivision. Memoirs AMS **93**, AMS, Providence (1991)
11. Ciesielski, Z.: Constructive function theory and spline systems. Studia Math. **58**, 277–302 (1975)
12. Cohen, A.: Numerical Analysis of Wavelet Methods. Elsevier (2003)
13. Cohen, A., Dyn, N., Matei, B.: Quasilinear subdivision schemes with application to ENO interpolation. Appl. Comput. Harmon. Anal. **15**, 89–116 (2003)
14. Dadourian, K.: Schemas de subdivision, analyses multiresolutions non-lineaires, applications. PhD thesis, Univerisite de Provence (2008)
15. Dahmen, W.: Wavelet and multiscale methods for operator equations. Acta Numerica **6**, 55–228 (1997)
16. Damme, van R.: Bivariate Hermite subdivision. Computer Aided Geometric Design **14**, 847–875 (1997)
17. Dyn, N., Grohs, P., Wallner, J.: Approximation order of interpolatory nonlinear subdivision schemes. J. Comput. Appl. Math. (2008)
18. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia PA (1992)
19. Daubechies, I., Runborg, O., Sweldens, W.: Normal multiresolution approximation of curves. Constr. Approx. **20**, 399–463 (2004)
20. Deslauriers, G., Dubuc, S.: Symmetric iterative interpolation processes. Constr. Approx. **5**, 49–68 (1989)
21. Donoho, D.L.: Interpolating wavelet transforms. Tech. Rep., Dep. Statistics, Stanford Univ. (2002)
22. Donoho, D.L., Yu, T.P.-Y.: Nonlinear pyramid transforms based on median interpolation. SIAM J. Math. Anal. **31**, 1030–1061 (2000)

23. Dyn, N.: Subdivision schemes in CAGD. In: Light, W.A. (ed.) Advances in Numerical Analysis II, pp. 36–104. Oxford Univ. Press Oxford (1992)
24. Dyn, N., Floater, M.S., Hormann K.: Four-point curve subdivision based on iterated chordal and centripetal parametrization, to appear in Computer Aided Geometric Design, also a Technical Report no. IfI-07-06, Department of Informatics, Clausthal University of Technology (2007), Germany
25. Dyn, N., Levin, D.: Analysis of asymptotically equivalent binary subdivision schemes. J. Comput. Appl. Math. **193**, 594–621 (1995)
26. Dyn, N., Levin, D.: Subdivision schemes in geometric modelling. Acta Numerica **11**, 73–144 (2002)
27. Dyn, N., Levin, D., Liu, D.: Interpolatory convexity preserving subdivision schemes for curves and surfaces. Computer Aided Design **24**, 211–216 (1992)
28. Floater, M.S.: On the deviation of a parametric cubic spline interpolant from its data polygon. Computer Aided Geometric Design **25**, 148–156 (2008)
29. Floater, M.S., Micchelli, C.A.: Nonlinear stationary subdivision. In: Govil, N.K., Mohapatra, R.N., Nashed, Z., Sharma, A., Szabados, J. (eds.) Approximation Theory: in memory of A.K. Varma, pp. 209–224, Marcel Dekker (1998)
30. Goodman, T., Yu, T.P.-Y.: Interpolation of medians. Adv. Comput. Math. **11**, 1–10 (1999)
31. Grohs, P.: Smoothness equivalence properties of univariate subdivision schemes and their projection analogues. Geometry Preprint 07/03, TU Graz (2007)
32. Grohs, P.: Smoothness analysis of subdivision schemes on regular grids by proximity. SIAM J. Numer. Anal. **46**, 2169–2182 (2008)
33. Grohs, P.: Approximation order from stability of nonlinear subdivision schemes. J. Approx. Th. (submitted), also a Geometry preprint 2008/06, TU Graz (2008), Austria
34. Guskov, I., Vidimce, K., Sweldens, W., Schröder, P.: Normal meshes. In: Akeley, K. (ed.) Computer Graphics (SIGGRAPH 2000: Proceedings), pp. 95–102, ACM Press/Addison Wesley Longman (2000)
35. Harizanov, S.: Stability of nonlinear multiresolution analysis. Proc. Appl. Math. Mech. **8**, 10933–10934 (2008)
36. Harizanov, S., Oswald, P.: Stability of nonlinear subdivision and multiscale transforms. Constr. Approx. (2009)
37. Harten, A.: Discrete multiresolution analysis and generalized wavelets. J. Appl. Numer. Anal. **12**, 153–192 (1993)
38. Harten, A.: Multiresolution representation of data: a general framework. SIAM J. Numer. Anal. **33**, 1205–1256 (1996)
39. Jansen, M., Baraniuk, R., Lavu, S.: Multiscale approximation of piecewise smooth two-dimensional functions using normal triangulated meshes. Appl. Comput. Harmon. Anal. **19**, 92–130 (2005)
40. Jia, R.-Q., Jiang, Q.: Spectral analysis of the transition operator and its application to smoothness analysis of wavelets. SIAM J. Matrix Anal. Appl. **24**, 1071–1109 (2003)
41. Khodakovsky A., Guskov, I.: Compression of normal meshes. In: (eds.) Geometric Modeling for Scientific Visualization, pp. 189–207, Springer (2003)
42. Khodakovsky A., Schröder, P., Sweldens, W.: Progressive geometry compression. In: Akeley, K. (ed.) Computer Graphics (SIGGRAPH 2000: Proceedings), pp. 271–278, ACM Press/Addison Wesley Longman (2000)
43. Kuijt, F., Damme, van R.: Smooth interpolation by a convexity preserving non-linear subdivision algorithm. In: Le Mehaute, A., Rabut, C., Schumaker, L.L. (eds.) Surface Fitting and Multiresolution Methods, pp. 219–224, Vanderbilt Univ. Press, Nashboro TN (1997)
44. Kuijt, F., Damme, van R.: Convexity preserving interpolatory subdivision schemes. Constr. Approx. **14**, 609–630 (1998)
45. Kuijt, F., van Damme, R.: Stability of subdivision schemes. TW Memorandum 1469, Fac. Appl. Math., Univ. Twente (1998)
46. Kuijt, F., Damme, van R.: Monotonicity preserving interpolatory subdivision schemes. J. Comput. Appl. Math. **101**, 203–229 (1999)

47. Levin, D.: Using Laurent polynomial representation for the analysis of nonuniform subdivision schemes. Advances of Computational Mathematics **11**, 41–54 (1999)
48. Marinov, M., Dyn, N., Levin, D.: Geometrically controlled 4-Point interpolatory schemes. In: Advances in Multiresolution for Geometric Modelling, Dodgson, N.A., Floater, M.S., Sabin, M.A., (eds.), pp. 301–315, Springer-Verlag (2005)
49. Matei, B.: Smoothness characterization and stability in nonlinear multiscale framework: theoretical results. Asymptotic Analysis **41**, 277–309 (2005)
50. Oswald, P.: Multilevel Finite Element Approximation: Theory & Applications. Teubner, Leipzig (1994)
51. Oswald, P.: Smoothness of a nonlinear subdivision scheme. In: Cohen, A., Merrien, J.-L., Schumaker, L.L. (eds.) Curve and Surface Fitting (Saint Malo 2002: Proceedings), pp. 323–332, Nashboro Press, Brentwood TM (2003)
52. Oswald, P.: Smoothness of nonlinear median-interpolation subdivision. Adv. Comput. Math. **20**, 401–423 (2004)
53. Pang, J.S., Yu, T.P.-Y.: Continuous M-estimators and their interpolation by polynomials. SIAM J. Numer. Anal. **42**, 997–1017 (2004)
54. Ur Rahman, I., Drori, I., Stodden, V.C., Donoho, D.L., Schröder, P.: Multiscale representation of manifold-valued data. Multiscale Model. Simul. **4**, 1201–1232 (2005)
55. Runborg, O.: Introduction to normal multiresolution analysis. In: Engquist, B., Løtstedt, P., Runborg, O. (eds.) Multiscale Methods in Science and Engineering, LNCSE vol. 44, pp. 205–224, Springer (2005)
56. Sabin, M.: The dual quadratic B-spline. Personal communication.
57. Sabin, M., Dodgson, N.: A circle-preserving variant of the four-point subdivision scheme. In: Dahlen, M., Morken, K., Schumaker, L.L. (eds.) Mathematical Methods for Curves and Surfaces: Tromso 2004, pp. 275-286, Nashboro Press (2005)
58. Serna, S., Marquina, A.: Power ENO methods: a fifth-order accurate weighted Power ENO method. J. Comput. Physics **194**, 632–658 (2004)
59. Sweldens, W.: The lifting scheme: A custom-design construction biorthogonal wavelets. Appl. Comput. Harmon. Anal. **3**, 186–200 (1996)
60. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. SIAM J. Math.Anal. **29**, 511–546 (1997)
61. Wallner, J.: Smoothness analysis of subdivision schemes by proximity. Constr. Approx. **24**, 289–318 (2006)
62. Wallner, J., Dyn, N.: Convergence and C^1 analysis of subdivision schemes on manifolds by proximity. Comput. Aided Geom. Design **22**, 593–622 (2005)
63. Xie, G., Yu, T.P.-Y.: On a linearization principle for nonlinear p-mean subdivision schemes. In: Neamtu, M., Saff, E.B. (eds.) Advances in Constructive Approximation, pp. 519–533, Nashboro Press (2004)
64. Xie, G., Yu, T.P.-Y.: Smoothness analysis of nonlinear subdivision schemes of homogeneous and affine invariant type. Constr. Approx. **22**, 219–254 (2005)
65. Xie, G., Yu, T.P.-Y.: Smoothness equivalence properties of manifold-valued data subdivision schemes based on the projection approach. SIAM J. Numer. Anal. **45**, 1200–1225 (2007)
66. Xie, G., Yu, T.P.-Y.: Approximation order equivalence properties of manifold-valued data subdivision schemes. (2008), to appear
67. Yang, X.: Normal based subdivision scheme for curve design. Comput. Aided Geom. Design **23**, 243–260 (2006)
68. Zhou, D.X.: The p-norm joint spectral radius for even integers. Methods Appl. Anal. **5**, 39–54 (1998)

Rapid solution of boundary integral equations by wavelet Galerkin schemes

Helmut Harbrecht and Reinhold Schneider

Abstract The present paper aims at reviewing the research on the wavelet-based rapid solution of boundary integral equations. When discretizing boundary integral equations by appropriate wavelet bases the system matrices are quasi-sparse. Discarding the non-relevant matrix entries is called wavelet matrix compression. The compressed system matrix can be assembled within linear complexity if an exponentially convergent *hp*-quadrature algorithm is used. Therefore, in combination with wavelet preconditioning, one arrives at an algorithm that solves a given boundary integral equation within discretization error accuracy, offered by the underlying Galerkin method, at a computational expense that stays proportional to the number of unknowns. By numerical results we illustrate and quantify the theoretical findings.

1 Introduction

Many mathematical models concerning for example field calculations, flow simulation, elasticity or visualization are based on operator equations with *nonlocal operators*, especially *boundary integral operators*. Discretizing such problems will then lead in general to possibly very large linear systems with *densely populated* matrices. Moreover, the involved operator may have an order different from zero which means that it acts on different length scales in a different way. This is well known to entail the linear systems to become more and more ill-conditioned when the level of resolution increases. Both features pose serious obstructions to the efficient nu-

Helmut Harbrecht
Institute for Numerical Simulation, Bonn University, Wegelerstr. 6, 53115 Bonn, Germany, e-mail: harbrecht@ins.uni-bonn.de

Reinhold Schneider
Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, e-mail: schneidr@math.tu-berlin.de

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_8,

merical treatment of such problems to an extent that desirable realistic simulations are still beyond current computing capacities.

This fact has stimulated enormous efforts to overcome these obstructions. The resulting significant progress made in recent years resulted in several methods for the rapid solution of boundary integral equations. These methods reduce the complexity to a nearly optimal rate or even an optimal rate. Denoting the number of unknowns by N_J, this means the complexity $\mathcal{O}(N_J \log^\alpha N_J)$ for some $\alpha \geq 0$. Prominent examples for such methods are the *fast multipole method* [42, 91], the *panel clustering* [48], *adaptive cross approximation* [2, 3], or *hierarchical matrices* [47, 104]. As observed in the pioneering paper [4] and investigated in [22, 27, 28, 29, 30, 79, 80, 81, 95], wavelet bases offer a further tool for the rapid solution of boundary integral equations. In fact, a Galerkin discretization with wavelet bases yields quasi-sparse matrices, i.e., the most matrix entries are negligible and can be treated as zero. Discarding these non-relevant matrix entries is called matrix compression. It has been firstly proven in [95] that only $\mathcal{O}(N_J)$ significant matrix entries remain. A precise outline of the historical development of the wavelet matrix compression can be found in Subsection 4.1.

Concerning boundary integral equations, a strong effort has been spent on the construction of appropriate wavelet bases on surfaces [24, 31, 32, 64, 59, 63, 69, 79, 95, 100]. In order to achieve the optimal complexity of the *wavelet Galerkin scheme*, wavelet bases are required that, depending on the order of the underlying operator, provide sufficiently many vanishing moments. This often rules out orthonormal wavelets. We report here on the realization of *biorthogonal spline wavelets*, derived from the multiresolution developed in [16]. These wavelets are advantageous since the regularity of the duals is known [105]. Moreover, the duals are compactly supported which preserves the linear complexity of the fast wavelet transform also for its inverse. This is an important task in applications, for instance for the coupling of FEM and BEM, cf. [41, 54, 55]. Additionally, in view of the discretization of operators of positive order, for instance, the hypersingular operator, globally continuous wavelets are available [6, 17, 31, 63].

The efficient computation of the relevant matrix coefficients turned out to be an extremely hard but important task for the successful application of the wavelet Galerkin scheme [61, 70, 81, 95]. We present a fully discrete Galerkin scheme based on numerical quadrature. Supposing that the given manifold is piecewise analytic we can use an hp-quadrature scheme [61, 92, 97] in combination with exponentially convergent quadrature rules. By combining all these ingredients we gain an algorithm which solves a given boundary integral equation in asymptotically linear complexity without compromising the accuracy of the underlying Galerkin scheme, see [22, 61]. This algorithm allows to solve boundary integral equations in reasonable time on serial computers with up to 10 million unknowns.

Wavelet matrix compression can be viewed as a non-uniform approximation of the Schwartz kernel with respect to the typical singularity on the diagonal. If the domain has corners and edges, the solution itself admits singularities. In this case an adaptive refinement will reduce the number of unknowns drastically without compromising the overall accuracy.

Adaptive methods for boundary integral equations have been considered by several authors, see e.g. [8, 9, 39, 71, 73, 96] and the references therein. However, convergence can in general only be proven under the so-called *saturation assumption*. Particularly, we are not aware of any paper concerning the combination of adaptive BEM with recent fast methods for integral equations like e.g. the fast multipole method. We emphasize further that, in difference to finite elements methods for local operators, the residuum to boundary integral equations cannot be computed exactly which makes the error estimation a quite difficult task.

Adaptive wavelet Galerkin methods for boundary integral equations have been considered first in [58, 60]. Although this method performs quite well in numerical experiments, convergence can only proven with the help of the saturation assumption. In [23], based on the adaptive algorithms from [13, 14, 15], a fully discrete adaptive wavelet Galerkin method has been presented that realizes asymptotically optimal complexity in the present context of global operators. *Asymptotically optimal* means here that any target accuracy can be achieved at a computational expense that stays proportional to the number of degrees of freedom (within the setting determined by the underlying wavelet basis) that would ideally be necessary for realizing that target accuracy if full knowledge about the unknown solution were given.

Meanwhile the wavelet Galerkin scheme has successfully been applied to a wide range of problems. For example, to the coupling of FEM and BEM (see [41, 54, 55]), to shape optimization (see e.g. [37, 38, 50], to inverse obstacle problems (see [51, 52]), and to the solution of boundary value problems with stochastic input parameters (see [62, 82]). The approach can be extended to surfaces which are represented by panels [53, 67, 94, 102, 103], particularly the illumination in virtual scenes is addressed in [67]. Moreover, recent progress has been made to employ wavelet matrix compression for high dimensional boundary integral equations, arising in particular from finance [88, 89, 90], see also [45, 68] for earlier results in this direction.

The outline of this survey is as follows. First, we specify the class of problems under consideration. Then, in Section 3 we provide suitable wavelet bases on manifolds. We review the historical development until the automatic construction of wavelets on surfaces that are described by Computer Aided Design. With the wavelet bases at hand we are able to introduce the fully discrete wavelet Galerkin scheme in Section 4. We survey on practical issues like setting up the compression pattern, assembling the system matrix and preconditioning. Especially, we present numerical results with respect to a nontrivial domain geometry in order to demonstrate the scheme. Finally, in Subsection 4.9 we present the developments concerning adaptivity, supported again by numerical results.

In the following, in order to avoid the repeated use of generic but unspecified constants, by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \sim D$ as $C \lesssim D$ and $C \gtrsim D$.

2 Problem formulation and preliminaries

2.1 Boundary integral equations

We consider a boundary integral equation on the closed boundary surface Γ of a three dimensional domain Ω

$$(\mathscr{A}u)(\mathbf{x}) = \int_\Gamma k(\mathbf{x},\mathbf{y})u(\mathbf{y})d\sigma_\mathbf{y} = f(\mathbf{x}), \quad \mathbf{x}\in\Gamma. \tag{1}$$

Here, the boundary integral operator $\mathscr{A} : H^q(\Gamma) \to H^{-q}(\Gamma)$ is assumed to be a bijective operator of order $2q$. The properties of the class of kernel functions $k(\mathbf{x},\mathbf{y})$ which are under consideration will be precisely outlined in Subsection 2.3.

Example 2.1 (Capacity computation). We suppose that the domain $\Omega \subset \mathbb{R}^3$ describes a charged condenser. The electric field generated by the condenser's charge is described by the exterior Dirichlet problem

$$\Delta u = 0 \text{ in } \Omega^c, \quad u = 1 \text{ on } \Gamma, \quad u = \mathscr{O}(\|\mathbf{x}\|^{-1}) \text{ as } \|\mathbf{x}\| \to \infty.$$

The capacity $C(\Omega)$ of the condenser is determined by the Dirichlet energy

$$C(\Omega) = \int_{\Omega^c} \|\nabla u\|^2 d\mathbf{x} = -\int_\Gamma \frac{\partial u}{\partial \mathbf{n}} d\sigma.$$

Here, the Neumann data can be computed by the so-called Dirichlet-to-Neumann map

$$\mathscr{V}\frac{\partial u}{\partial \mathbf{n}} = \left(\mathscr{K} - \frac{1}{2}\right)u \text{ on } \Gamma \tag{2}$$

where the single layer operator $\mathscr{V} : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ and the double layer operator $\mathscr{K} : L^2(\Gamma) \to L^2(\Gamma)$ are respectively defined by

$$(\mathscr{V}u)(\mathbf{x}) = \int_\Gamma \frac{u(\mathbf{y})}{4\pi\|\mathbf{x}-\mathbf{y}\|} d\sigma_\mathbf{y}, \quad (\mathscr{K}u)(\mathbf{x}) = \int_\Gamma \frac{\langle \mathbf{n}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle}{4\pi\|\mathbf{x}-\mathbf{y}\|^3} u(\mathbf{y}) d\sigma_\mathbf{y}, \quad \mathbf{x}\in\Gamma.$$

Since $\mathscr{K}1 = -1/2$ the computation of the Neumann data reduces to the solution of a Fredholm integral equation of the first kind

$$\mathscr{V}\frac{\partial u}{\partial \mathbf{n}} = -1 \text{ on } \Gamma.$$

Example 2.2 (Computation of free surfaces of liquid metals). Exterior electromagnetic shaping is the determination of the free surface of a droplet Ω of liquid metal of volume $|\Omega|$ that levitates in an electromagnetic field. The magnetic field is generated by conductors outside the droplet, i.e., the density current vector $\mathbf{j}$ is compactly supported in Ω^c and satisfies $\operatorname{div}\mathbf{j} = 0$.

The free surface $\Gamma = \partial\Omega$ is the minimizer of the shape optimization problem

$$J(\Omega) = -\int_{\Omega^c} \|\mathbf{B}\|^2 d\mathbf{x} + C\int_{\Gamma} 1 d\sigma + D\int_{\Omega} x_3 d\mathbf{x} \to \min \quad \text{subject to} \quad |\Omega| = V_0,$$

where the magnetic field $\mathbf{B} = \mathbf{B}(\Omega)$ satisfies the magneto-static Maxwell equations

$$\operatorname{curl}\mathbf{B} = \mu\mathbf{j} \text{ in } \Omega^c, \quad \operatorname{div}\mathbf{B} = 0 \text{ in } \Omega^c, \quad \langle \mathbf{B}, \mathbf{n}\rangle = 0 \text{ on } \Gamma \tag{3}$$

together with the decay condition $\mathbf{B} = \mathcal{O}(\|\mathbf{x}\|^{-2})$ as $\|\mathbf{x}\| \to \infty$. The constants C, D refer to the surface tension and the gravitational acceleration.

During an optimization procedure the magnetic field in the exterior domain Ω^c has to be computed very often on different geometries. Here, it became rather popular to exploit boundary integral equations, see e.g. [38, 75, 76, 85]. By the standard decomposition

$$\mathbf{B} = \operatorname{curl}\mathbf{A} + \nabla u, \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{\mathbf{j}(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y}$$

the problem reduces to seeking $u \in H^1_{loc}(\Omega^c)$, satisfying the following exterior Neumann problem for the Laplacian

$$\Delta u = 0 \text{ in } \Omega^c, \quad \frac{\partial u}{\partial \mathbf{n}} = -\langle \operatorname{curl}\mathbf{A}, \mathbf{n}\rangle \text{ on } \Gamma, \quad u = \mathcal{O}(\|\mathbf{x}\|^{-1}) \text{ as } \|\mathbf{x}\| \to \infty.$$

This boundary value problem can be efficiently solved by invoking the Neumann-to-Dirichlet map which follows from the relation (2).

A simulated droplet is depicted in Fig. 15.

2.2 Parametric surface representation

We assume that the boundary $\Gamma \subset \mathbb{R}^3$ is represented by piecewise parametric mappings. Let $\square := [0,1]^2$ denote the unit square. We subdivide the given manifold into several *patches*

$$\Gamma = \bigcup_{i=1}^{M} \Gamma_i, \qquad \Gamma_i = \gamma_i(\square), \qquad i = 1, 2, \ldots, M,$$

such that each $\gamma_i : \square \to \Gamma_i$ defines a diffeomorphism of $\square$ onto Γ_i. The intersection $\Gamma_i \cap \Gamma_{i'}$, $i \neq i'$, of the patches Γ_i and $\Gamma_{i'}$ is supposed to be either $\emptyset$, a common edge, or a common vertex.

A mesh of level j on Γ is induced by dyadic subdivisions of depth j of the unit square into 4^j squares $C_{j,\mathbf{k}} \subseteq \square$, where $\mathbf{k} = (k_1, k_2)$ with $0 \leq k_1, k_2 < 2^j$. This generates $4^j M$ *elements* (or elementary domains) $\Gamma_{i,j,\mathbf{k}} := \gamma_i(C_{j,\mathbf{k}}) \subseteq \Gamma_i$, $i = 1, \ldots, M$. In order to get a regular mesh of Γ, the parametric representation is subjected to the following matching condition. A bijective, affine mapping $\Xi : \square \to \square$ exists such that for all $\mathbf{x} = \gamma_i(\mathbf{s})$ on the common interface $\Gamma_i \cap \Gamma_{i'}$ it holds that $\gamma_i(\mathbf{s}) = (\gamma_{i'} \circ \Xi)(\mathbf{s})$.

In other words, the diffeomorphisms γ_i and $\gamma_{i'}$ coincide along the edges except for orientation.

The canonical inner product in $L^2(\Gamma)$ is given by

$$(u,v)_{L^2(\Gamma)} = \int_\Gamma u(\mathbf{x})v(\mathbf{x})d\sigma_{\mathbf{x}} = \sum_{i=1}^M \int_\square u\big(\gamma_i(\mathbf{s})\big)v\big(\gamma_i(\mathbf{s})\big)\,\kappa_i(\mathbf{s})d\mathbf{s}$$

with $\kappa_i \sim 1$ being the surface measure

$$\kappa_i := \left\| \frac{\partial \gamma_i}{\partial s_1} \times \frac{\partial \gamma_i}{\partial s_2} \right\|.$$

The corresponding Sobolev spaces are indicated by $H^s(\Gamma)$. Of course, depending on the global smoothness of the surface, the range of permitted $s \in \mathbb{R}$ is limited to $s \in (-s_\Gamma, s_\Gamma)$. In case of general Lipschitz domains we have at least $s_\Gamma = 1$ since for all $0 \le s \le 1$ the spaces $H^s(\Gamma)$ consist of the traces of functions $\in H^{s+1/2}(\Omega)$, cf. [20].

For the construction of wavelets on manifolds the following *modified* inner product is playing a crucial role

$$\langle u,v\rangle = \sum_{i=1}^M (u\circ\gamma_i, v\circ\gamma_i)_{L^2(\square)} = \sum_{i=1}^M \int_\square u\big(\gamma_i(\mathbf{s})\big)v\big(\gamma_i(\mathbf{s})\big)d\mathbf{s}. \tag{4}$$

We shall introduce some function spaces associated with this inner product. For arbitrary $s \ge 0$ we define the Sobolev spaces $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$ as the closure of all *patchwise* C^∞*-functions* on Γ with respect to the norm

$$\|v\|_{H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)} := \sum_{i=1}^M \|v\circ\gamma_i\|_{H^s(\square)}. \tag{5}$$

The Sobolev spaces of negative order, that is $H^{-s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$, are defined as the duals of $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$ with respect to the modified inner product (4), equipped by the norm

$$\|v\|_{H^{-s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)} := \sup_{w\in H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)} \frac{\langle v,w\rangle}{\|w\|_{H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)}}. \tag{6}$$

Since the surface measure is in general discontinuous across the interface of two neighbouring patches, the Sobolev spaces $H^s(\Gamma)$ and $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$ are only isomorphic in the range $s \in \big(-\min\{\frac{1}{2}, s_\Gamma\}, \min\{\frac{1}{2}, s_\Gamma\}\big)$, see [31] for the details.

In complete analogy, based on (5), (6), we define also the spaces $H^{s,1}_{\langle\cdot,\cdot\rangle}(\Gamma)$, $s \in \mathbb{R}$, which stem from the completion of all *globally continuous patchwise* C^∞*-functions* on Γ. According to [31], the Sobolev spaces $H^s(\Gamma)$ and $H^{s,1}_{\langle\cdot,\cdot\rangle}(\Gamma)$ are isomorphic in the range $s \in \big(-\min\{\frac{1}{2}, s_\Gamma\}, \min\{\frac{3}{2}, s_\Gamma\}\big)$.

The Sobolev spaces $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$ and $H^{s,1}_{\langle\cdot,\cdot\rangle}(\Gamma)$ will be needed later on in the analysis of patchwise smooth wavelet bases in Subsection 3.5 and globally continuous wavelet functions in Subsection 3.6, respectively.

The surface representation is in contrast to the common approximation of surfaces by panels. It has the advantage that the rate of convergence is not limited by approximation. Technical surfaces generated by tools from Computer Aided Design (CAD) are often represented in the present form.

The most common geometry representation in CAD is defined by the IGES (Initial Graphics Exchange Specification) standard. Here, the initial CAD object is a solid, bounded by a closed surface that is given as a collection of parametric surfaces which can be trimmed or untrimmed. An untrimmed surface is already a four-sided patch, parameterized over a rectangle. Whereas, a trimmed surface is just a piece of a supporting untrimmed surface, described by boundary curves. There are several representations of the parameterizations including B-splines, NURBS (nonuniform rational B-Splines), surfaces of revolution, and tabulated cylinders [65].

In [56] an algorithm has been developed to decompose a technical surface, described in the IGES format, into a collection of parameterized four-sided patches, fulfilling all the above requirements. We refer the reader to [56] for the details. Fig. 16 presents two examples of geometries produced by this algorithm.

2.3 Kernel properties

We can now specify the kernel functions of the boundary integral operator $\mathscr{A}$ under consideration. To this end, we denote by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$ multi-indices and define $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2$. Moreover, we denote by $k_{i,i'}(\mathbf{s},\mathbf{t})$ the transported kernel functions, that is

$$k_{i,i'}(\mathbf{s},\mathbf{t}) := k\big(\gamma_i(\mathbf{s}), \gamma_{i'}(\mathbf{t})\big)\,\kappa_i(\mathbf{s})\,\kappa_{i'}(\mathbf{t}), \quad 1 \le i, i' \le M. \tag{7}$$

Definition 2.1. A kernel $k(\mathbf{x},\mathbf{y})$ is called standard kernel of the order $2q$ if the partial derivatives of the transported kernel functions $k_{i,i'}(\mathbf{s},\mathbf{t})$, $1 \le i, i' \le M$, are bounded by

$$\big|\partial_{\mathbf{s}}^{\boldsymbol{\alpha}} \partial_{\mathbf{t}}^{\boldsymbol{\beta}} k_{i,i'}(\mathbf{s},\mathbf{t})\big| \le c_{\boldsymbol{\alpha},\boldsymbol{\beta}} \big\|\gamma_i(\mathbf{s}) - \gamma_{i'}(\mathbf{t})\big\|^{-(2+2q+|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|)}$$

provided that $2 + 2q + |\boldsymbol{\alpha}| + |\boldsymbol{\beta}| > 0$.

We emphasize that this definition requires patchwise smoothness but *not* global smoothness of the geometry. The surface itself needs to be only Lipschitz continuous. Generally, under this assumption, the kernel of a boundary integral operator $\mathscr{A}$ of order $2q$ is standard of order $2q$. Hence, we may assume this property in what follows.

3 Wavelet bases on manifolds

3.1 Wavelets and multiresolution analyses

We shall first focus on those aspects of biorthogonal multiresolution analyses which are useful for our purpose. Let Ω be a domain $\subset \mathbb{R}^n$ or manifold $\subset \mathbb{R}^{n+1}$. Then, in general, a biorthogonal multiresolution analysis consists of two nested families of finite dimensional subspaces

$$\begin{aligned} V_{j_0} \subset V_{j_0+1} \subset \cdots \subset V_j \subset V_{j+1} \cdots \subset \cdots \subset L^2(\Omega), \\ \widetilde{V}_{j_0} \subset \widetilde{V}_{j_0+1} \subset \cdots \subset \widetilde{V}_j \subset \widetilde{V}_{j+1} \cdots \subset \cdots \subset L^2(\Omega), \end{aligned} \tag{8}$$

such that $\dim V_j \sim \dim \widetilde{V}_j \sim 2^{nj}$ and

$$\overline{\bigcup_{j \geq j_0} V_j} = \overline{\bigcup_{j \geq j_0} \widetilde{V}_j} = L^2(\Omega). \tag{9}$$

The spaces $V_j = \operatorname{span} \boldsymbol{\Phi}_j$, $\widetilde{V}_j = \operatorname{span} \widetilde{\boldsymbol{\Phi}}_j$ are generated by biorthogonal single-scale bases

$$\boldsymbol{\Phi}_j = [\phi_{j,k}]_{k \in \Delta_j}, \qquad \widetilde{\boldsymbol{\Phi}}_j = [\widetilde{\phi}_{j,k}]_{k \in \Delta_j}, \qquad (\boldsymbol{\Phi}_j, \widetilde{\boldsymbol{\Phi}}_j)_{L^2(\Omega)} = \mathbf{I},$$

where Δ_j denotes a suitable index set with cardinality $|\Delta_j| \sim 2^{nj}$. Note that here and in the sequel the basis $\boldsymbol{\Phi}_j = [\phi_{j,k}]_{k \in \Delta_j}$ has to be understood as a row vector.

A final requirement is that these bases are uniformly stable, i.e., for any vector $\mathbf{c} \in \ell^2(\Delta_j)$ holds

$$\|\boldsymbol{\Phi}_j \mathbf{c}\|_{L^2(\Omega)} \sim \|\widetilde{\boldsymbol{\Phi}}_j \mathbf{c}\|_{L^2(\Omega)} \sim \|\mathbf{c}\|_{\ell^2(\Delta_j)} \tag{10}$$

uniformly in j.

If one is going to use the spaces V_j as trial spaces in a Galerkin scheme, then additional properties are required. At least the primal single-scale bases are supposed to satisfy the locality condition

$$\operatorname{diam} \operatorname{supp} \phi_{j,k} \sim 2^{-j}.$$

Furthermore, it is assumed that the following Jackson and Bernstein type estimates hold uniformly in j for $s < \gamma$, $s \leq t \leq d$

$$\inf_{v_j \in V_j} \|u - v_j\|_{H^s(\Omega)} \lesssim 2^{j(s-t)} \|u\|_{H^t(\Omega)}, \qquad u \in H^t(\Omega), \tag{11}$$

and for $s \leq t < \gamma$

$$\|v_j\|_{H^t(\Omega)} \lesssim 2^{j(t-s)} \|v_j\|_{H^s(\Omega)}, \qquad v_j \in V_j, \tag{12}$$

where $d, \gamma > 0$ are fixed constants given by

$$d = \sup\{s \in \mathbb{R} : \inf_{v_j \in V_j} \|u - v_j\|_{L^2(\Omega)} \le 2^{-js} \|u\|_{H^s(\Omega)}\},$$
$$\gamma = \sup\{s \in \mathbb{R} : V_j \subset H^s(\Omega)\}.$$

Usually, d is the maximal degree of polynomials which are locally contained in V_j and is referred to as the order of exactness of the multiresolution analysis $\{V_j\}$. The parameter γ denotes the regularity or smoothness of the functions in the spaces V_j. Analogous estimates are valid for the dual multiresolution analysis $\{\widetilde{V}_j\}$ with constants $\widetilde{d}$ and $\widetilde{\gamma}$.

Instead of using only a single scale j the idea of wavelet concepts is to keep track to the increment of information between two adjacent scales j and $j+1$. The biorthogonal wavelets

$$\Psi_j = [\psi_{j,k}]_{k \in \nabla_j}, \qquad \widetilde{\Psi}_j = [\widetilde{\psi}_{j,k}]_{k \in \nabla_j}, \qquad (\Psi_j, \widetilde{\Psi}_j)_{L^2(\Omega)} = \mathbf{I},$$

where $\nabla_j = \Delta_{j+1} \setminus \Delta_j$, are the bases of *uniquely* determined complement spaces $W_j = \operatorname{span} \Psi_j$, $\widetilde{W}_j = \operatorname{span} \widetilde{\Psi}_j$, satisfying

$$\begin{aligned} V_{j+1} &= V_j \oplus W_j, & V_j \cap W_j &= \{0\}, & W_j &\perp \widetilde{V}_j, \\ \widetilde{V}_{j+1} &= \widetilde{V}_j \oplus \widetilde{W}_j, & \widetilde{V}_j \cap \widetilde{W}_j &= \{0\}, & \widetilde{W}_j &\perp V_j. \end{aligned} \tag{13}$$

We claim that the primal wavelets $\psi_{j,k}$ are also local with respect to the corresponding scale j, that is

$$\operatorname{diam} \operatorname{supp} \psi_{j,k} \sim 2^{-j}, \tag{14}$$

and we will normalize the wavelets such that $\|\psi_{j,k}\|_{L^2(\Omega)} \sim \|\widetilde{\psi}_{j,k}\|_{L^2(\Omega)} \sim 1$. Furthermore, we suppose that the wavelet bases

$$\Psi = [\Psi_j]_{j \ge j_0 - 1}, \qquad \widetilde{\Psi} = [\widetilde{\Psi}_j]_{j \ge j_0 - 1}, \tag{15}$$

($\Psi_{j_0-1} := \Phi_{j_0}$, $\widetilde{\Psi}_{j_0-1} := \widetilde{\Phi}_{j_0}$), are Riesz bases of $L^2(\Omega)$.

The assumptions that (11) and (12) hold with some constants γ and $\widetilde{\gamma}$ relative to $\{V_j\}$, $\{\widetilde{V}_j\}$ provide a convenient device for switching between the norms $\|\cdot\|_{H^t(\Omega)}$ and corresponding sums of weighted wavelet coefficients. Namely, the following norm equivalences hold

$$\begin{aligned} \|v\|^2_{H^s(\Omega)} &\sim \sum_{j \ge j_0 - 1} \sum_{k \in \nabla_j} 2^{js} \left| (v, \widetilde{\psi}_{j,k})_{L^2(\Omega)} \right|^2, \quad s \in (-\widetilde{\gamma}, \gamma), \\ \|v\|^2_{H^s(\Omega)} &\sim \sum_{j \ge j_0 - 1} \sum_{k \in \nabla_j} 2^{js} \left| (v, \psi_{j,k})_{L^2(\Omega)} \right|^2, \quad s \in (-\gamma, \widetilde{\gamma}), \end{aligned} \tag{16}$$

see e.g. [21, 66, 95] for the details. Note that for $s = 0$ the norm equivalences imply the Riesz property of the wavelet bases.

From (13) we deduce that the primal wavelets provide *vanishing moments* or the *cancellation property* of order $\widetilde{d}$, that is

$$\left|(v,\psi_{j,k})_{L^2(\Gamma)}\right| \lesssim 2^{-j(\widetilde{d}+n/2)}|v|_{W^{\widetilde{d},\infty}(\operatorname{supp}\psi_{j,\mathbf{k}})}. \tag{17}$$

Here, $|v|_{W^{\widetilde{d},\infty}(\Omega)} := \sup_{|\boldsymbol{\alpha}|=\widetilde{d},\,\mathbf{x}\in\Omega}|\partial^{\boldsymbol{\alpha}}v(\mathbf{x})|$ denotes the semi-norm in $W^{\widetilde{d},\infty}(\Omega)$. Notice that the corresponding cancellation property with parameter d holds with respect to the dual wavelets.

3.2 Refinement relations and stable completions

For the construction of multiresolution bases we are interested in the *filter* or *mask coefficients* associated with the scaling functions and the wavelets. Since boundary functions have to be introduced, these filter coefficients are not fixed like in the stationary case. Therefore, we are going to compute the full *two-scale relations*

$$\begin{aligned} \Phi_j &= \Phi_{j+1}\mathbf{M}_{j,0}, & \Psi_j &= \Phi_{j+1}\mathbf{M}_{j,1},\\ \widetilde{\Phi}_j &= \widetilde{\Phi}_{j+1}\widetilde{\mathbf{M}}_{j,0}, & \widetilde{\Psi}_j &= \widetilde{\Phi}_{j+1}\widetilde{\mathbf{M}}_{j,1}, \end{aligned} \tag{18}$$

where $\mathbf{M}_{j,0},\widetilde{\mathbf{M}}_{j,0}\in\mathbb{R}^{|\Delta_{j+1}|\times|\Delta_j|}$ and $\mathbf{M}_{j,1},\widetilde{\mathbf{M}}_{j,1}\in\mathbb{R}^{|\Delta_{j+1}|\times|\nabla_j|}$. Notice that these matrices will be banded and only the filter coefficients for some specific scaling functions and wavelets have to be modified. That way, the advantages of the stationary and shift-invariant case are preserved as far as possible.

We proceed as follows. We first construct biorthogonal single-scale bases in refinable spaces V_j and $\widetilde{V}_j$. The parameters $d,\widetilde{d},\gamma$ and $\widetilde{\gamma}$ are constituted by these single-scale bases. According to (13) the complementary spaces W_j and $\widetilde{W}_j$ are determined uniquely. But as we will see there is some freedom in choosing the biorthogonal wavelet bases that generate these complementary spaces. Each pair of matrices $\mathbf{M}_{j,1},\widetilde{\mathbf{M}}_{j,1}$ satisfying

$$[\mathbf{M}_{j,0},\mathbf{M}_{j,1}]^T[\widetilde{\mathbf{M}}_{j,0},\widetilde{\mathbf{M}}_{j,1}] = \mathbf{I}$$

defines wavelets (especially Riesz bases in $L^2(\Omega)$) via their two-scale relations (18). But, for instance, this straightforward construction does not imply fixed and finite masks of the wavelets. Hence, in order to define suitable wavelet bases, we utilize the concept of the *stable completion* [7]. This concept is universal and often employed in the sequel.

Definition 3.1. Let $\breve{\Psi}_j = [\breve{\psi}_{j,k}]_{k\in\nabla_j}\subset V_{j+1}$ be a given collection of functions, satisfying

$$\breve{\Psi}_j = \Phi_{j+1}\breve{\mathbf{M}}_{j,1}, \qquad \breve{\mathbf{M}}_{j,1}\in\mathbb{R}^{|\Delta_{j+1}|\times|\nabla_j|},$$

such that $[\mathbf{M}_{j,0},\breve{\mathbf{M}}_{j,1}]$ is invertible. We define the matrix $[\mathbf{G}_{j,0},\mathbf{G}_{j,1}]$ with $\mathbf{G}_{j,0}\in\mathbb{R}^{|\Delta_{j+1}|\times|\Delta_j|}$ and $\mathbf{G}_{j,1}\in\mathbb{R}^{|\Delta_{j+1}|\times|\nabla_j|}$ as the inverse of $[\mathbf{M}_{j,0},\breve{\mathbf{M}}_{j,1}]^T$, i.e.

$$[\mathbf{M}_{j,0},\breve{\mathbf{M}}_{j,1}]^T[\mathbf{G}_{j,0},\mathbf{G}_{j,1}] = \mathbf{I}. \tag{19}$$

The collection $\breve{\Psi}_j$ is called a stable completion of Φ_j if

$$\left\|[\mathbf{M}_{j,0},\breve{\mathbf{M}}_{j,1}]\right\|_{\ell^2(\Delta_{j+1})} \sim \left\|[\mathbf{G}_{j,0},\mathbf{G}_{j,1}]\right\|_{\ell^2(\Delta_{j+1})} \sim 1. \tag{20}$$

We derive the desired wavelet basis by projecting the stable completion onto W_j, cf. [26]. In terms of the refinement matrices, the related matrix $\mathbf{M}_{j,1}$ is defined by

$$\mathbf{M}_{j,1} = \left[\mathbf{I} - \mathbf{M}_{j,0}\widetilde{\mathbf{M}}_{j,0}^T\right]\breve{\mathbf{M}}_{j,1} =: \breve{\mathbf{M}}_{j,1} - \mathbf{M}_{j,0}\mathbf{L}_j. \tag{21}$$

One readily verifies that the matrix $\mathbf{L}_j \in \mathbb{R}^{|\Delta_j|\times|\nabla_j|}$ satisfies

$$\mathbf{L}_j = \widetilde{\mathbf{M}}_{j,0}^T\breve{\mathbf{M}}_{j,1} = \left(\widetilde{\Phi}_j, \breve{\Psi}_j\right)_{L^2(\Omega)}. \tag{22}$$

Moreover, one concludes from the identity

$$\mathbf{I} = [\mathbf{M}_{j,0},\mathbf{M}_{j,1}]^T[\widetilde{\mathbf{M}}_{j,0},\widetilde{\mathbf{M}}_{j,1}] = \begin{bmatrix}\mathbf{I} & -\mathbf{L}_j\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^T [\mathbf{M}_{j,0},\breve{\mathbf{M}}_{j,1}]^T[\mathbf{G}_{j,0},\mathbf{G}_{j,1}]\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{L}_j^T & \mathbf{I}\end{bmatrix}$$

the equality

$$[\widetilde{\mathbf{M}}_{j,0},\widetilde{\mathbf{M}}_{j,1}] = [\mathbf{G}_{j,0} + \mathbf{G}_{j,1}\mathbf{L}_j^T, \mathbf{G}_{j,1}],$$

i.e. $\widetilde{\mathbf{M}}_{j,1} = \mathbf{G}_{j,1}$. Note that a compactly supported stable completion implies compactly supported wavelet bases.

Remark 3.1. The definition of $\mathbf{M}_{j,1}$ implies

$$\Psi_j = \Phi_{j+1}\mathbf{M}_{j,1} = \Phi_{j+1}\breve{\mathbf{M}}_{j,1} - \Phi_{j+1}\breve{\mathbf{M}}_{j,0}\mathbf{L}_j = \breve{\Psi}_j - \Phi_j\mathbf{L}_j.$$

Consequently, similarly to [101], the wavelets Ψ_j are obtained by updating $\breve{\Psi}_j$ by linear combinations of the coarse level generators Φ_j.

3.3 Biorthogonal spline multiresolution on the interval

Our approach is based on the biorthogonal spline multiresolution on $\mathbb{R}$ that has been developed by A. Cohen, I. Daubechies and J.-C. Feauveau [16]. These functions have several properties which make them favourite candidates for the wavelet Galerkin scheme.

– The primal multiresolution consists of cardinal B-splines of the order d as scaling functions. Therefore, we have to deal only with piecewise polynomials which simplifies the computation of the coefficients in Galerkin matrices. We like to point out that the primal multiresolution realizes the order of approximation d. The regularity of these ansatz functions is $\gamma^{\mathbb{R}} = d - 1/2$.

– The dual multiresolution is generated by compactly supported scaling functions, realizing the order of approximation $\widetilde{d} \in \mathbb{N}$ ($d + \widetilde{d}$ even). According to [105] their regularity $\widetilde{\gamma}^{\mathbb{R}}$ is known.

In accordance with [26], based on these scaling functions, we can construct refinable spaces $V_j^{[0,1]}$, $\widetilde{V}_j^{[0,1]}$ which contain all polynomials of degree less than d, $\widetilde{d}$, respectively. The goal is thus to construct a wavelet basis such that only a few boundary wavelets do not coincide with translates and dilates of the Cohen-Daubechies-Feauveau wavelets [16].

For the treatment of boundary integral equations we focus on piecewise constant and linear wavelets, i.e., $d = 1$ and $d = 2$. On the level j, we consider the interval $[0,1]$ subdivided into 2^j equidistant subintervals. Then, of course, $V_j^{[0,1]}$ is the space generated by 2^j and $2^j + 1$ piecewise constant and linear scaling functions, respectively. We prefer the Haar basis and the hierarchical basis on the given partition to define suitable stable completions. In fact, by utilizing these stable completions the interior wavelets coincide with the Cohen-Daubechies-Feauveau wavelets, cf. [26, 49].

According to [26] the following statements hold.

– The collections $\Psi^{[0,1]}$ and $\widetilde{\Psi}^{[0,1]}$, given by (15), define biorthogonal Riesz bases in $L^2([0,1])$.

– The functions of $\Psi^{[0,1]}$ and $\widetilde{\Psi}^{[0,1]}$ have $\widetilde{d}$ and d vanishing moments, respectively.

– The functions of the collections $\Psi^{[0,1]}$ and $\widetilde{\Psi}^{[0,1]}$ have the same regularity as the biorthogonal spline wavelets in $L^2(\mathbb{R})$ [105]. Therefore, the norm equivalences (16) are valid for $\gamma = \gamma^{\mathbb{R}} = d - 1/2$ and $\widetilde{\gamma} = \widetilde{\gamma}^{\mathbb{R}}$.

In view of operators of positive order, e.g. the hypersingular operator, we need globally continuous wavelet bases. According to [31, 49], for their construction, the primal and dual scaling functions as well as the stable completion are required to satisfy the following boundary conditions.

– Only one function each of the collections $\Phi_j^{[0,1]}$ and $\widetilde{\Phi}_j^{[0,1]}$, respectively, is nonvanishing at the interval endpoints $x = 0$ and $x = 1$. that is

$$\phi_{j,k}^{[0,1]}(0) = \begin{cases} 2^{j/2}, & k = 0, \\ 0, & k \neq 0, \end{cases} \qquad \widetilde{\phi}_{j,k}^{[0,1]}(0) = \begin{cases} 2^{j/2} c, & k = 0, \\ 0, & k \neq 0, \end{cases} \qquad c \neq 0, \tag{23}$$

and likewise for $x = 1$ and $k = |\triangle_j^{[0,1]}|$.

– The stable completion $\breve{\Psi}_j^{[0,1]}$ fulfills zero boundary conditions

$$\breve{\psi}_{j,k}^{[0,1]}(0) = \breve{\psi}_{j,k}^{[0,1]}(1) = 0, \qquad k \in \nabla_j^{[0,1]}. \tag{24}$$

Moreover, there holds the symmetry condition

$$\breve{\psi}_{j,k}^{[0,1]}(x) = \breve{\psi}_{j,|\nabla_j^{[0,1]}|-k}^{[0,1]}(1-x), \qquad k \in \nabla_j^{[0,1]}. \tag{25}$$

Notice that condition (23) can be realized by a suitable change of bases, cf. [31, 49]. The construction of stable completions that satisfy (24), (25), is addressed in [26, 31, 49]. Notice that in the case of the piecewise linears, the hierarchical basis satisfies (24) and (25).

3.4 Wavelets on the unit square

In general it suffices to consider two dimensional wavelets for the treatment of boundary integral equations. Hence, to keep the presentation simple, we confine ourselves to the two dimensional case. For the higher dimensional case we refer the reader to [31, 49].

3.4.1 Biorthogonal scaling functions

The canonical definition of (isotropic) biorthogonal multiresolutions on the unit square $\square := [0,1]^2$ is to take tensor products of the univariate constructions. That is, the collections of scaling functions are given by

$$\Phi_j^{\square} = \Phi_j^{[0,1]} \otimes \Phi_j^{[0,1]}, \qquad \widetilde{\Phi}_j^{\square} = \widetilde{\Phi}_j^{[0,1]} \otimes \widetilde{\Phi}_j^{[0,1]}, \tag{26}$$

with the index set $\Delta_j^{\square} = \Delta_j^{[0,1]} \times \Delta_j^{[0,1]}$. Consequently, the associated refinement matrices are

$$\mathbf{M}_{j,0}^{\square} = \mathbf{M}_{j,0}^{[0,1]} \otimes \mathbf{M}_{j,0}^{[0,1]}, \qquad \widetilde{\mathbf{M}}_{j,0}^{\square} = \widetilde{\mathbf{M}}_{j,0}^{[0,1]} \otimes \widetilde{\mathbf{M}}_{j,0}^{[0,1]}. \tag{27}$$

As an immediate consequence of the univariate case, the spaces $V_j^{\square} := \operatorname{span} \Phi_j^{\square}$ and $\widetilde{V}_j^{\square} := \operatorname{span} \widetilde{\Phi}_j^{\square}$ are nested and dense in $L^2(\square)$. Clearly, these spaces are also exact of the order d and $\widetilde{d}$, respectively. We emphasize that the complement spaces $W_j^{\square}$ and $\widetilde{W}_j^{\square}$ are uniquely determined by (13). With this in mind, the remainder of this subsection is dedicated to the construction of biorthogonal wavelet bases $\Psi_j^{\square}$ and $\widetilde{\Psi}_j^{\square}$ with $W_j^{\square} := \operatorname{span} \Psi_j^{\square}$ and $\widetilde{W}_j^{\square} := \operatorname{span} \widetilde{\Psi}_j^{\square}$.

3.4.2 Tensor product wavelets

Firstly, we introduce the simplest construction, namely tensor product wavelets

$$\Psi_j^{\square} = \left[\Phi_j^{[0,1]} \otimes \Psi_j^{[0,1]}, \Psi_j^{[0,1]} \otimes \Phi_j^{[0,1]}, \Psi_j^{[0,1]} \otimes \Psi_j^{[0,1]} \right].$$

The associated refinement matrices are defined via

$$\mathbf{M}_{j,1}^{\square} = \begin{bmatrix} \mathbf{M}_{j,0}^{[0,1]} \otimes \mathbf{M}_{j,1}^{[0,1]} \\ \mathbf{M}_{j,1}^{[0,1]} \otimes \mathbf{M}_{j,0}^{[0,1]} \\ \mathbf{M}_{j,1}^{[0,1]} \otimes \mathbf{M}_{j,1}^{[0,1]} \end{bmatrix}, \qquad \widetilde{\mathbf{M}}_{j,1}^{\square} = \begin{bmatrix} \widetilde{\mathbf{M}}_{j,0}^{[0,1]} \otimes \widetilde{\mathbf{M}}_{j,1}^{[0,1]} \\ \widetilde{\mathbf{M}}_{j,1}^{[0,1]} \otimes \widetilde{\mathbf{M}}_{j,0}^{[0,1]} \\ \widetilde{\mathbf{M}}_{j,1}^{[0,1]} \otimes \widetilde{\mathbf{M}}_{j,1}^{[0,1]} \end{bmatrix}.$$

Hence, we differ three types of wavelets on $\square$, see Figs. 1 and 2. The first type is the tensor product $\phi_{j,k}^{[0,1]} \otimes \psi_{j,\ell}^{[0,1]}$. The second type is the tensor product of $\psi_{j,k}^{[0,1]} \otimes \phi_{j,\ell}^{[0,1]}$. The third type consists of the tensor product of two wavelets $\psi_{j,k}^{[0,1]} \otimes \psi_{j,\ell}^{[0,1]}$. We mention that $|\Delta_j^{[0,1]}| \approx |\nabla_j^{[0,1]}|$ implies nearly identical cardinalities of the three types of wavelets.

3.4.3 Simplified tensor product wavelets

We consider an extension of the tensor product construction. As we will see it replaces the third type wavelets by smoother ones. Particularly this simplifies numerical integration, for instance, in the Galerkin scheme.

The idea is to involve a suitable stable completion on the unit square. Based on the univariate case it can be defined by the collection

$$\breve{\Psi}_j^{\square} = \big[\Phi_j^{[0,1]} \otimes \breve{\Psi}_j^{[0,1]}, \breve{\Psi}_j^{[0,1]} \otimes \Phi_j^{[0,1]}, \breve{\Psi}_j^{[0,1]} \otimes \breve{\Psi}_j^{[0,1]}\big].$$

The refinement matrices $\breve{\mathbf{M}}_{j,1}^{\square}$, $\mathbf{G}_{j,0}^{\square}$ and $\mathbf{G}_{j,1}^{\square}$ are computed by

$$\breve{\mathbf{M}}_{j,1}^{\square} = \begin{bmatrix} \mathbf{M}_{j,0}^{[0,1]} \otimes \breve{\mathbf{M}}_{j,1}^{[0,1]} \\ \breve{\mathbf{M}}_{j,1}^{[0,1]} \otimes \mathbf{M}_{j,0}^{[0,1]} \\ \breve{\mathbf{M}}_{j,1}^{[0,1]} \otimes \breve{\mathbf{M}}_{j,1}^{[0,1]} \end{bmatrix}, \quad \mathbf{G}_{j,0}^{\square} = \mathbf{G}_{j,0}^{[0,1]} \otimes \mathbf{G}_{j,0}^{[0,1]}, \quad \mathbf{G}_{j,1}^{\square} = \begin{bmatrix} \mathbf{G}_{j,0}^{[0,1]} \otimes \mathbf{G}_{j,1}^{[0,1]} \\ \mathbf{G}_{j,1}^{[0,1]} \otimes \mathbf{G}_{j,0}^{[0,1]} \\ \mathbf{G}_{j,1}^{[0,1]} \otimes \mathbf{G}_{j,1}^{[0,1]} \end{bmatrix}.$$

As one readily verifies, the matrix $\mathbf{L}_j^{\square}$ is given by

$$\mathbf{L}_j^{\square} = \begin{bmatrix} \mathbf{I}^{(|\Delta_j^{[0,1]}|)} \otimes \mathbf{L}_j^{[0,1]} \\ \mathbf{L}_j^{[0,1]} \otimes \mathbf{I}^{(|\Delta_j^{[0,1]}|)} \\ \mathbf{L}_j^{[0,1]} \otimes \mathbf{L}_j^{[0,1]} \end{bmatrix}.$$

This implies

$$\mathbf{M}_{j,1}^{\square} = \breve{\mathbf{M}}_{j,1}^{\square} - \mathbf{M}_{j,0}^{\square}\mathbf{L}_j^{\square} = \begin{bmatrix} \mathbf{M}_{j,0}^{[0,1]} \otimes \mathbf{M}_{j,1}^{[0,1]} \\ \mathbf{M}_{j,1}^{[0,1]} \otimes \mathbf{M}_{j,0}^{[0,1]} \\ \breve{\mathbf{M}}_{j,1}^{[0,1]} \otimes \breve{\mathbf{M}}_{j,1}^{[0,1]} - \big(\mathbf{M}_{j,0}^{[0,1]} \otimes \mathbf{M}_{j,0}^{[0,1]}\big)\big(\mathbf{L}_j^{[0,1]} \otimes \mathbf{L}_j^{[0,1]}\big) \end{bmatrix}.$$

Hence, we differ again three types of wavelets on $\square$. The first and the second type coincide with the tensor product wavelets, see Figs. 1 and 2. But now the third

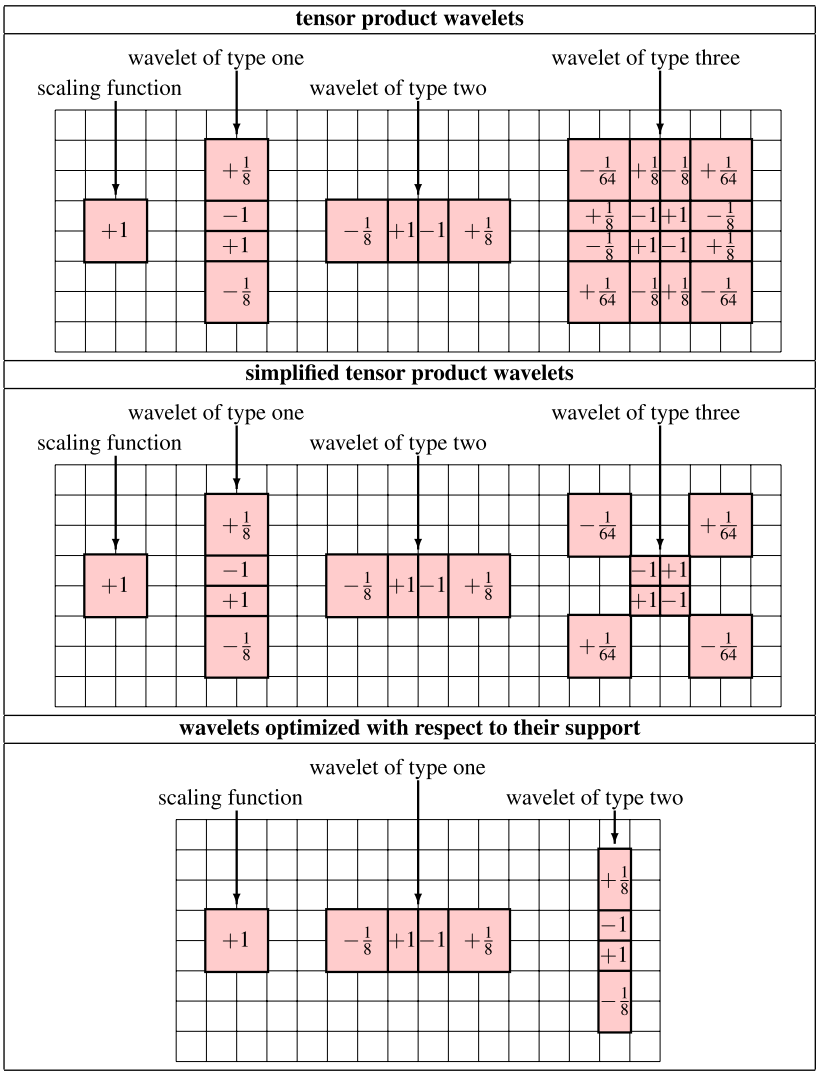


Fig. 1 Interior piecewise constant wavelets with three vanishing moments. The boundary wavelets are not depicted

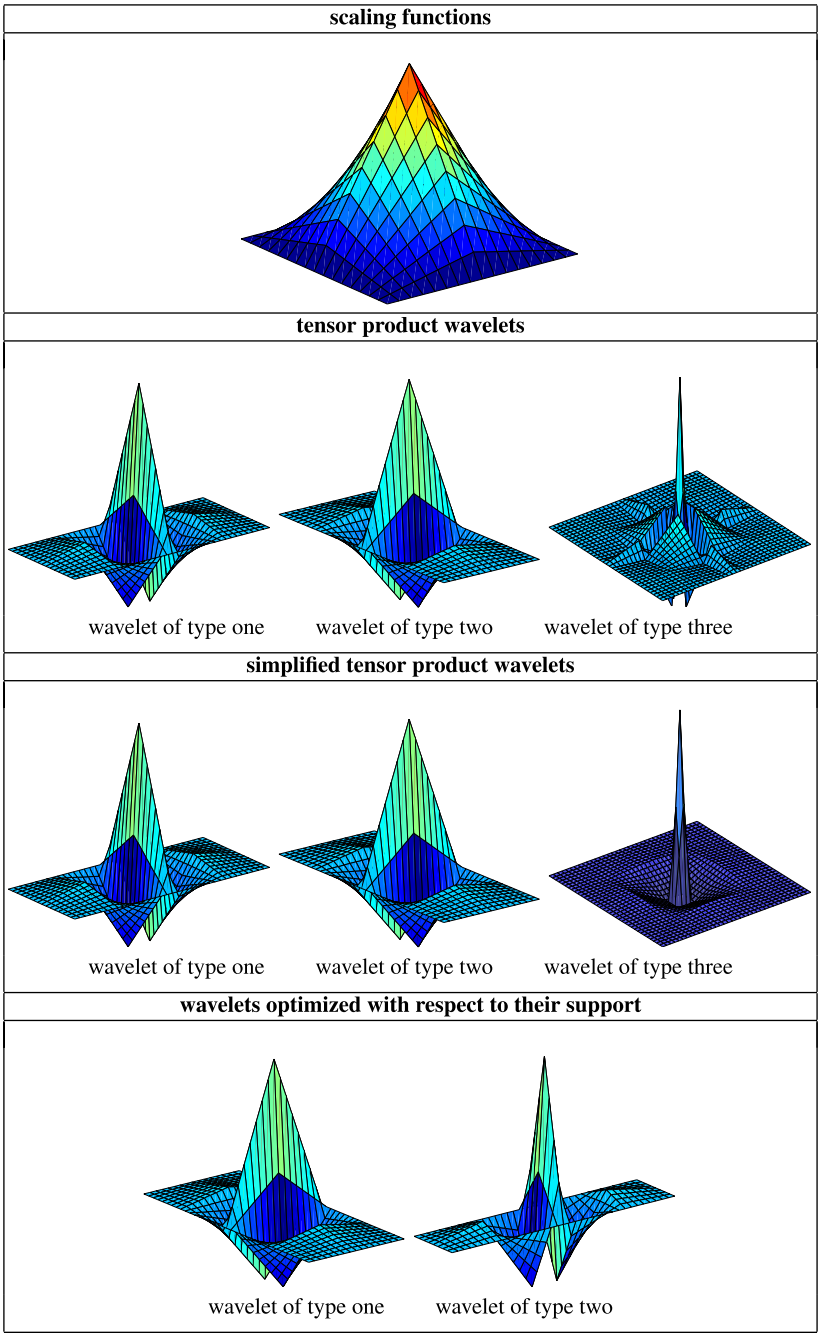


Fig. 2 Interior piecewise linear wavelets with four vanishing moments. The boundary wavelets are not depicted

type consists of the tensor product of the stable completion $\breve{\psi}^{\square}_{j,\mathbf{k}} = \breve{\psi}^{[0,1]}_{j,k} \otimes \breve{\psi}^{[0,1]}_{j,\ell}$, updated by certain scaling functions $\phi^{\square}_{j,\mathbf{k}'} = \phi^{[0,1]}_{j,k'} \otimes \phi^{[0,1]}_{j,\ell'}$ of the coarse grid j. In general, the support of this wavelet does not depend on the choice of the stable completion (except for piecewise constant wavelets, see Fig. 1). But choosing a stable completion on $[0,1]$ with small supports, the product $\breve{\psi}^{[0,1]}_{j,k} \otimes \breve{\psi}^{[0,1]}_{j,\ell} \in V^{\square}_{j+1}$ has also small support. Since the additional scaling functions belong to $V^{\square}_{j}$, the wavelet is smoother than the corresponding tensor product wavelet.

3.4.4 Wavelets optimized with respect to their Supports

Finally, we consider a more advanced construction which yields wavelets with very small supports. We define the wavelets according to

$$\begin{aligned} \Psi^{\square}_{j} &= \left[\Psi^{[0,1]}_{j} \otimes \Phi^{[0,1]}_{j}, \Phi^{[0,1]}_{j+1} \otimes \Psi^{[0,1]}_{j}\right], \\ \widetilde{\Psi}^{\square}_{j} &= \left[\widetilde{\Psi}^{[0,1]}_{j} \otimes \widetilde{\Phi}^{[0,1]}_{j}, \widetilde{\Phi}^{[0,1]}_{j+1} \otimes \widetilde{\Psi}^{[0,1]}_{j}\right]. \end{aligned} \tag{28}$$

Lemma 3.1. *The collections of wavelets $\Psi^{\square}_{j}$ and $\widetilde{\Psi}^{\square}_{j}$ introduced by (28) define biorthogonal wavelet bases with respect to the multiresolution given by $\Phi^{\square}_{j}$ and $\widetilde{\Phi}^{\square}_{j}$.*

Proof. One readily verifies the equations

$$\left(\psi^{\square}_{j,\mathbf{k}}, \widetilde{\phi}^{\square}_{j,\mathbf{k}'}\right)_{L^2(\square)} = \left(\widetilde{\psi}^{\square}_{j,\mathbf{k}}, \phi^{\square}_{j,\mathbf{k}'}\right)_{L^2(\square)} = 0, \qquad \mathbf{k} \in \nabla^{\square}_{j}, \quad \mathbf{k}' \in \Delta^{\square}_{j},$$

by inserting the definition (28) of the wavelet functions and employing the biorthogonality on the interval. Consequently, in order to show the biorthogonality of the wavelets, we only have to prove that

$$\left(\psi^{\square}_{j,\mathbf{k}}, \widetilde{\psi}^{\square}_{j,\mathbf{k}'}\right)_{L^2(\square)} = \delta_{\mathbf{k},\mathbf{k}'}, \qquad \mathbf{k}, \mathbf{k}' \in \Delta^{\square}_{j}.$$

But similar to above this is again an immediate consequence of the biorthogonality on the interval. In view of the cardinality of the sets $\Psi^{\square}_{j}$, $\widetilde{\Psi}^{\square}_{j}$ the biorthogonality implies the assertion.

The refinement matrices $\mathbf{M}^{\square}_{j,1}$ and $\widetilde{\mathbf{M}}^{\square}_{j,1}$ are computed by

$$\mathbf{M}^{\square}_{j,1} = \begin{bmatrix} \mathbf{M}^{[0,1]}_{j,1} \otimes \mathbf{M}^{[0,1]}_{j,0} \\ \mathbf{I}^{(|\Delta^{[0,1]}_{j+1}|)} \otimes \mathbf{M}^{[0,1]}_{j,1} \end{bmatrix}, \qquad \widetilde{\mathbf{M}}^{\square}_{j,1} = \begin{bmatrix} \widetilde{\mathbf{M}}^{[0,1]}_{j,1} \otimes \widetilde{\mathbf{M}}^{[0,1]}_{j,0} \\ \mathbf{I}^{(|\Delta^{[0,1]}_{j+1}|)} \otimes \widetilde{\mathbf{M}}^{[0,1]}_{j,1} \end{bmatrix}.$$

Thus, we distinguish two types of wavelets on $\square$, cf. Figs. 1 and 2. The first type is the tensor product $\psi^{[0,1]}_{j,k} \otimes \phi^{[0,1]}_{j,\ell}$. The second type is the tensor product $\phi^{[0,1]}_{j+1,k} \otimes \psi^{[0,1]}_{j,\ell}$. This wavelet type owns a very small support in comparison with

the previously introduced wavelets, since a scaling function of the fine grid $j+1$ appears in the first coordinate. Notice that the number of wavelets of type two is nearly twice as much as the number of wavelets of type one.

3.5 Patchwise smooth wavelet bases

If the wavelets are not required to be globally continuous, one may employ wavelet bases that are defined on each patch individually. This strategy reflects the canonical choice for the piecewise constants. But in the case piecewise bilinear ansatz functions we obtain double nodes along the edges of neighbouring patches which leads to somewhat more degrees of freedom than in the case of global continuity.

The primal scaling functions and wavelets are given by

$$\phi_{j,\mathbf{k}}^{\Gamma_i}(\mathbf{x}) := \begin{cases} \phi_{j,\mathbf{k}}^{\square}\big(\gamma_i^{-1}(\mathbf{x})\big), & \mathbf{x} \in \Gamma_i, \\ 0, & \text{else}, \end{cases} \qquad \psi_{j,\mathbf{k}}^{\Gamma_i}(\mathbf{x}) := \begin{cases} \psi_{j,\mathbf{k}}^{\square}\big(\gamma_i^{-1}(\mathbf{x})\big), & \mathbf{x} \in \Gamma_i, \\ 0, & \text{else}. \end{cases}$$

Setting $\boldsymbol{\Phi}_j^{\Gamma_i} = \big[\phi_{j,\mathbf{k}}^{\Gamma_i}\big]_{\mathbf{k}\in\Delta_j^{\square}}$ and $\boldsymbol{\Psi}_j^{\Gamma_i} = \big[\psi_{j,\mathbf{k}}^{\Gamma_i}\big]_{\mathbf{k}\in\nabla_j^{\square}}$, the collections of scaling functions and wavelets on Γ are defined by $\boldsymbol{\Phi}_j^{\Gamma} := \big[\boldsymbol{\Phi}_j^{\Gamma_i}\big]_{i=1}^{M}$ and $\boldsymbol{\Psi}_j^{\Gamma} := \big[\boldsymbol{\Psi}_j^{\Gamma_i}\big]_{i=1}^{M}$. Obviously, the refinement matrices with $\boldsymbol{\Phi}_j^{\Gamma} = \boldsymbol{\Phi}_{j+1}^{\Gamma}\mathbf{M}_{j,0}^{\Gamma}$ and $\boldsymbol{\Psi}_j^{\Gamma} = \boldsymbol{\Phi}_{j+1}^{\Gamma}\mathbf{M}_{j,1}^{\Gamma}$ are obtained by

$$\mathbf{M}_{j,0}^{\Gamma} = \operatorname{diag}\big(\underbrace{\mathbf{M}_{j,0}^{\square},\ldots,\mathbf{M}_{j,0}^{\square}}_{M \text{ times}}\big), \qquad \mathbf{M}_{j,1}^{\Gamma} = \operatorname{diag}\big(\underbrace{\mathbf{M}_{j,1}^{\square},\ldots,\mathbf{M}_{j,1}^{\square}}_{M \text{ times}}\big).$$

Clearly, the spaces $V_j^{\Gamma} := \operatorname{span}\boldsymbol{\Phi}_j^{\Gamma}$ are nested. In addition, we find $V_{j+1}^{\Gamma} = V_j^{\Gamma} \oplus W_j^{\Gamma}$, where $W_j^{\Gamma} := \operatorname{span}\boldsymbol{\Psi}_j^{\Gamma}$. Proceeding analogously on the dual side yields a multiresolution on Γ which is biorthogonal with respect to the modified inner product (4).

The subsequent proposition, proven in [31], states that we obtain all important properties of the univariate case with respect to the modified inner product.

Proposition 3.1. *The collection of wavelets $\boldsymbol{\Psi}^{\Gamma}$ and $\widetilde{\boldsymbol{\Psi}}^{\Gamma}$ form biorthogonal Riesz bases in $L^2_{\langle\cdot,\cdot\rangle}(\Gamma)$. The primal wavelets satisfy the cancellation property (17) with parameter $\widetilde{d}$ with respect to the modified inner product (4). Moreover, the norm equivalences (16) hold with $\gamma = \gamma^{\mathbb{R}}$ and $\widetilde{\gamma} = \widetilde{\gamma}^{\mathbb{R}}$ with respect to the spaces $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$.*

Remark 3.2. The cancellation property (17) with parameter $\widetilde{d}$ holds also with respect to the canonical inner product, since each wavelet is supported on a single patch. Due to the isomorphy of the Sobolev spaces $H^s(\Gamma)$ and $H^{s,0}_{\langle\cdot,\cdot\rangle}(\Gamma)$ for $s \in \big(-\min\{\frac{1}{2}, s_\Gamma\}, \min\{\frac{1}{2}, s_\Gamma\}\big)$, cf. Subsection 2.2, the norm equivalences with respect to the canonical Sobolev spaces $H^s(\Gamma)$ are valid with $\gamma = \min\{1/2, s_\Gamma, \gamma^{\mathbb{R}}\}$ and $\widetilde{\gamma} = \min\{1/2, s_\Gamma, \widetilde{\gamma}^{\mathbb{R}}\}$. In particular, the wavelets $\boldsymbol{\Psi}^{\Gamma}$ and $\widetilde{\boldsymbol{\Psi}}^{\Gamma}$ form biorthogonal Riesz bases in $L^2(\Gamma)$.

3.6 Globally continuous wavelet bases

Similar constructions of globally continuous wavelet bases have been presented in three different papers published at nearly the same time [6, 17, 31]. We summarize here the construction of *composite wavelet bases*, as introduced by W. Dahmen and R. Schneider in [31], which is based on the simplified tensor product wavelets. To this end, both the underlying scaling functions and the stable completion are required to satisfy the conditions specified in Subsection 3.3. In what follows we are going to glue scaling functions and wavelets along the interfaces of neighbouring patches to gain global continuity.

We introduce first some notation since we need to deal with local indices and functions defined on the parameter domain $\square$ as well as global indices and functions on the surface Γ. To this end, it is convenient to identify the basis functions with physical grid points of the mesh on the unit square, i.e., we employ a bijective mapping $q_j : \Delta_j^{\square} \to \square$ in order to redefine our index sets on the unit square. This mapping should identify the boundary functions with points on $\partial\square$. Moreover, it should preserve the symmetry, that is, in view of (25), given any affine mapping $\Xi : \square \to \square$, there holds

$$\Phi_j^{\square} = \Phi_j^{\square} \circ \Xi, \qquad \breve{\Psi}_j^{\square} = \breve{\Psi}_j^{\square} \circ \Xi, \qquad \widetilde{\Phi}_j^{\square} = \widetilde{\Phi}_j^{\square} \circ \Xi, \qquad \widetilde{\Psi}_j^{\square} = \widetilde{\Psi}_j^{\square} \circ \Xi. \quad (29)$$

Then, the boundary conditions (23) and (24) imply

$$\phi_{j,\mathbf{k}}^{\square}\Big|_{\partial\square} \equiv \widetilde{\phi}_{j,\mathbf{k}}^{\square}\Big|_{\partial\square} \equiv 0, \quad \mathbf{k} \in \Delta_j^{\square} \cap \square^{\circ},$$
$$\breve{\psi}_{j,\mathbf{k}}^{\square}\Big|_{\partial\square} \equiv 0, \quad \mathbf{k} \in \nabla_j^{\square} \cap \square^{\circ}.$$

Hence, all functions corresponding to the indices $\mathbf{k}$ located in the interior of $\square$ satisfy zero boundary conditions. In the case of piecewise bilinears the mapping q_j is simply defined by $q_j(\mathbf{k}) = 2^{-j}\mathbf{k}$. For the general case we refer to [31].

A given point $\mathbf{x} \in \Gamma$ might have several representations

$$\mathbf{x} = \gamma_{i_1}(\mathbf{s}_1) = \ldots = \gamma_{i_{r(\mathbf{x})}}(\mathbf{s}_{r(\mathbf{x})})$$

if $\mathbf{x}$ belongs to different patches $\Gamma_{i_1}, \ldots, \Gamma_{i_{r(\mathbf{x})}}$. Of course, this occurs only if $\mathbf{x}$ lies on an edge or vertex of a patch. We count the number of preimages of a given point $\mathbf{x} \in \Gamma$ by the function

$$r : \Gamma \to \mathbb{N}, \qquad r(\mathbf{x}) := \big|\{i \in \{1,2,\ldots,M\} : \mathbf{x} \in \Gamma_i\}\big|. \quad (30)$$

Clearly, one has $r(\mathbf{x}) \geq 1$, where $r(\mathbf{x}) = 1$ holds for all $\mathbf{x}$ located in the interior of the patches Γ_i. Moreover, $r(\mathbf{x}) = 2$ for all $\mathbf{x}$ which belong to an edge and are different from a vertex.

Next, given two points $\mathbf{x}, \mathbf{y} \in \Gamma$, the function

$$c : \Gamma \times \Gamma \to \mathbb{N}, \quad c(\mathbf{x},\mathbf{y}) := \big|\{i \in \{1,2,\ldots,M\} : \mathbf{x} \in \Gamma_i \wedge \mathbf{y} \in \Gamma_i\}\big| \quad (31)$$

counts the number of patches Γ_i which contain both points simultaneously.

The index sets on Γ are just given by physical grid points on the surface

$$\Delta_j^\Gamma := \{\gamma_i(\mathbf{k}) : \mathbf{k} \in \Delta_j^\square,\ i \in \{1,2,\dots,M\}\}, \quad \nabla_j^\Gamma := \Delta_{j+1}^\Gamma \setminus \Delta_j^\Gamma. \tag{32}$$

The gluing of functions at the intersections of the patches is performed as follows. According to [31], the scaling functions $\Phi_j^\Gamma := \left[\phi_{j,\boldsymbol{\xi}}^\Gamma\right]_{\boldsymbol{\xi}\in\Delta_j^\Gamma}$ and $\widetilde{\Phi}_j^\Gamma := \left[\widetilde{\phi}_{j,\boldsymbol{\xi}}^\Gamma\right]_{\boldsymbol{\xi}\in\Delta_j^\Gamma}$ are defined by

$$\phi_{j,\boldsymbol{\xi}}^\Gamma(\mathbf{x}) = \begin{cases} \phi_{j,\mathbf{k}}^\square\big(\gamma_i^{-1}(\mathbf{x})\big), & \exists (i,\mathbf{k}) : \gamma_i(\mathbf{k}) = \boldsymbol{\xi} \wedge \mathbf{x} \in \Gamma_i, \\ 0, & \text{elsewhere}, \end{cases}$$

$$\widetilde{\phi}_{j,\boldsymbol{\xi}}^\Gamma(\mathbf{x}) = \begin{cases} \frac{1}{r(\boldsymbol{\xi})}\widetilde{\phi}_{j,\mathbf{k}}^\square\big(\gamma_i^{-1}(\mathbf{x})\big), & \exists (i,\mathbf{k}) : \gamma_i(\mathbf{k}) = \boldsymbol{\xi} \wedge \mathbf{x} \in \Gamma_i, \\ 0, & \text{elsewhere}. \end{cases}$$

On the primal side, this definition reflects the canonical strategy. On the dual side, the strategy is the same except for normalization, for a visualization see also Fig. 3. The normalization factor ensures biorthogonality with respect to the modified inner product (4), i.e. $\langle \Phi_j^\Gamma, \widetilde{\Phi}_j^\Gamma \rangle = \mathbf{I}$.

The scaling functions are refinable Riesz bases of the spaces $V_j^\Gamma := \operatorname{span} \Phi_j^\Gamma$ and $\widetilde{V}_j^\Gamma := \operatorname{span} \widetilde{\Phi}_j^\Gamma$. The two-scale relations (18), associated with these scaling functions, are given by (cf. [31, 59])

$$\left[\mathbf{M}_{j,0}^\Gamma\right]_{\boldsymbol{\xi}',\boldsymbol{\xi}} = \begin{cases} \left[\mathbf{M}_{j,0}^\square\right]_{\mathbf{k}',\mathbf{k}}, & \exists (i,\mathbf{k},\mathbf{k}') : \boldsymbol{\xi} = \gamma_i(\mathbf{k}) \wedge \boldsymbol{\xi}' = \gamma_i(\mathbf{k}'), \\ 0, & \text{elsewhere}, \end{cases} \tag{33}$$

$$\left[\widetilde{\mathbf{M}}_{j,0}^\Gamma\right]_{\boldsymbol{\xi}',\boldsymbol{\xi}} = \begin{cases} \frac{c(\boldsymbol{\xi},\boldsymbol{\xi}')}{r(\boldsymbol{\xi})}\left[\widetilde{\mathbf{M}}_{j,0}^\square\right]_{\mathbf{k}',\mathbf{k}}, & \exists (i,\mathbf{k},\mathbf{k}') : \boldsymbol{\xi} = \gamma_i(\mathbf{k}) \wedge \boldsymbol{\xi}' = \gamma_i(\mathbf{k}'), \\ 0, & \text{elsewhere}. \end{cases} \tag{34}$$

In accordance with [31], the stable completion $\breve{\Psi}_j = \left[\breve{\psi}_{j,\boldsymbol{\xi}}^\Gamma\right]_{\boldsymbol{\xi}\in\nabla_j^\Gamma}$ can be defined like the primal scaling functions, namely

$$\breve{\psi}_{j,\boldsymbol{\xi}}^\Gamma(\mathbf{x}) = \begin{cases} \breve{\psi}_{j,\mathbf{k}}^\square\big(\gamma_i^{-1}(\mathbf{x})\big), & \exists (i,\mathbf{k}) : \gamma_i(\mathbf{k}) = \boldsymbol{\xi} \wedge \mathbf{x} \in \Gamma_i, \\ 0, & \text{elsewhere}. \end{cases}$$

The associated refinement matrix is thus determined analogously to $\mathbf{M}_{j,0}^\Gamma$, that is

$$\left[\breve{\mathbf{M}}_{j,1}^\Gamma\right]_{\boldsymbol{\xi}',\boldsymbol{\xi}} = \begin{cases} \left[\breve{\mathbf{M}}_{j,1}^\square\right]_{\mathbf{k}',\mathbf{k}}, & \exists (i,\mathbf{k},\mathbf{k}') : \boldsymbol{\xi} = \gamma_i(\mathbf{k}) \wedge \boldsymbol{\xi}' = \gamma_i(\mathbf{k}'), \\ 0, & \text{elsewhere}. \end{cases} \tag{35}$$

The dual wavelets $\widetilde{\Psi}_j^\Gamma := \left[\widetilde{\psi}_{j,\boldsymbol{\xi}}^\Gamma\right]_{\boldsymbol{\xi}\in\nabla_j^\Gamma}$ are derived from their refinement relation (18), where

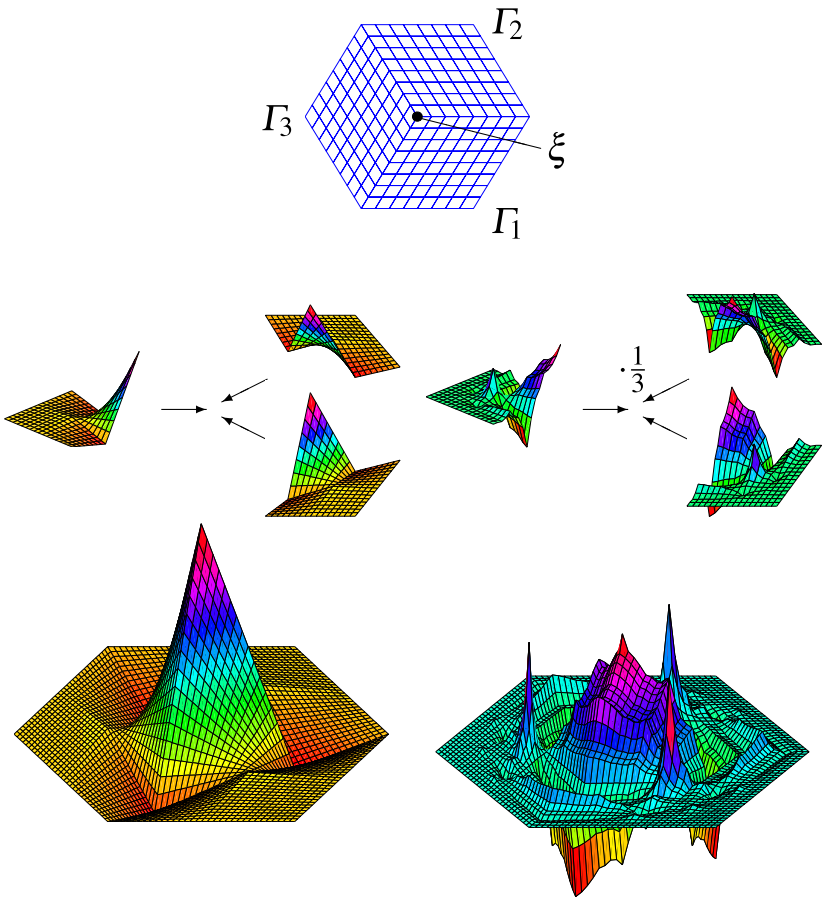


Fig. 3 The primal (left) and the dual (right) generator on a degenerated vertex in the case $d = \widetilde{d} = 2$

$$\left[\widetilde{\mathbf{M}}^{\Gamma}_{j,1}\right]_{\boldsymbol{\xi}',\boldsymbol{\xi}} = \begin{cases} \frac{c(\boldsymbol{\xi},\boldsymbol{\xi}')}{r(\boldsymbol{\xi})}\left[\widetilde{\mathbf{M}}^{\square}_{j,1}\right]_{\mathbf{k}',\mathbf{k}}, & \exists (i,\mathbf{k},\mathbf{k}') : \boldsymbol{\xi} = \gamma_i(\mathbf{k}) \wedge \boldsymbol{\xi}' = \gamma_i(\mathbf{k}'), \\ 0, & \text{elsewhere}, \end{cases} \tag{36}$$

cf. [31, 49]. Consequently, the matrix $\mathbf{L}^{\Gamma}_j := \left(\widetilde{\mathbf{M}}^{\Gamma}_{j,0}\right)^T \breve{\mathbf{M}}^{\Gamma}_{j,1}$ reads as (see [49, 59] for the proof)

$$\left[\mathbf{L}^{\Gamma}_j\right]_{\boldsymbol{\xi}',\boldsymbol{\xi}} = \begin{cases} \frac{c(\boldsymbol{\xi},\boldsymbol{\xi}')}{r(\boldsymbol{\xi}')}\left[\mathbf{L}^{\square}_j\right]_{\mathbf{k}',\mathbf{k}}, & \exists (i,\mathbf{k},\mathbf{k}') : \boldsymbol{\xi} = \gamma_i(\mathbf{k}) \wedge \boldsymbol{\xi}' = \gamma_i(\mathbf{k}'), \\ 0, & \text{elsewhere}. \end{cases} \tag{37}$$

From this the primal wavelets $\Psi^{\Gamma}_j := \left[\psi^{\Gamma}_{j,\boldsymbol{\xi}}\right]_{\boldsymbol{\xi}\in\nabla^{\Gamma}_j}$ are finally given by

$$\mathbf{M}_{j,1}^{\Gamma} = \check{\mathbf{M}}_{j,1}^{\Gamma} - \mathbf{L}_j^{\Gamma}\mathbf{M}_{j,0}^{\Gamma}. \tag{38}$$

In all we end up with a black box algorithm for the application of the discrete wavelet transform. Although the definitions of the refinement matrices seem to be very technical, the implementation of the discrete wavelet transform is rather canonical as the Algorithms 3.1 and 3.2 confirm. The dual wavelet transform (Algorithm 3.1) employs Eqs. (34) and (36), whereas the wavelet transform (Algorithm 3.2) is based on the Eqs. (33), (35), and (37) to define the wavelets according to (38). Notice that the factor $c(\boldsymbol{\xi},\boldsymbol{\xi}')$ is implicitly realized by the loop over all patches. A globally continuous wavelet with two vanishing moments, located on an edge, and its corresponding dual are depicted in Fig. 17.

Algorithm 3.1 This algorithm computes the two-scale decomposition $\widetilde{\boldsymbol{\Phi}}_{j+1}^{\Gamma}\mathbf{a}^{(j+1)} = \widetilde{\boldsymbol{\Phi}}_j^{\Gamma}\mathbf{a}^{(j)} + \widetilde{\boldsymbol{\Psi}}_j^{\Gamma}\mathbf{b}^{(j)}$, where $\mathbf{a}^{(j)} = \big[a_{\boldsymbol{\xi}}^{(j)}\big]_{\boldsymbol{\xi}\in\Delta_j^{\Gamma}}$ and $\mathbf{b}^{(j)} = \big[b_{\boldsymbol{\xi}}^{(j)}\big]_{\boldsymbol{\xi}\in\nabla_j^{\Gamma}}$.

initialization: $\mathbf{a}^{(j)} := \mathbf{b}^{(j)} := \mathbf{0}$
for $i = 1$ **to** M **do begin**
  for all $\mathbf{k} \in \Delta_j^{\square}$ **do begin** C: compute coefficients of $\widetilde{\boldsymbol{\Phi}}_j^{\Gamma}$
    for all $\mathbf{k}' \in \Delta_{j+1}^{\square}$ **do begin**
      $a_{\gamma_i(\mathbf{k})}^{(j)} = a_{\gamma_i(\mathbf{k})}^{(j)} + \big[\widetilde{\mathbf{M}}_{j,0}^{\square}\big]_{\mathbf{k}',\mathbf{k}} a_{\gamma_i(\mathbf{k}')}^{(j+1)} / r\big(\gamma_i(\mathbf{k})\big)$
  end, end
  for all $\mathbf{k} \in \nabla_j^{\square}$ **do begin** C: compute coefficients of $\widetilde{\boldsymbol{\Psi}}_j^{\Gamma}$
    for all $\mathbf{k}' \in \Delta_{j+1}^{\square}$ **do begin**
      $b_{\gamma_i(\mathbf{k})}^{(j)} = b_{\gamma_i(\mathbf{k})}^{(j)} + \big[\widetilde{\mathbf{M}}_{j,1}^{\square}\big]_{\mathbf{k}',\mathbf{k}} a_{\gamma_i(\mathbf{k}')}^{(j+1)} / r\big(\gamma_i(\mathbf{k})\big)$
end, end, end.

Proposition 3.2. *The collection of wavelets $\boldsymbol{\Psi}^{\Gamma}$ and $\widetilde{\boldsymbol{\Psi}}^{\Gamma}$ form biorthogonal Riesz bases in $L^2_{\langle\cdot,\cdot\rangle}(\Gamma)$. The primal wavelets satisfy the cancellation property (17) with parameter $\widetilde{d}$ with respect to the modified inner product (4). Moreover, the norm equivalences (16) hold for $\gamma = \gamma^{\mathbb{R}}$ and $\widetilde{\gamma} = \widetilde{\gamma}^{\mathbb{R}}$ with respect to the spaces $H^{s,1}_{\langle\cdot,\cdot\rangle}(\Gamma)$.*

Remark 3.3. The norm equivalences with respect to the canonical Sobolev spaces $H^s(\Gamma)$ are valid for $\gamma = \min\{3/2, s_\Gamma, \gamma^{\mathbb{R}}\}$ and $\widetilde{\gamma} = \min\{1/2, s_\Gamma, \widetilde{\gamma}^{\mathbb{R}}\}$ since, according to Subsection 2.2, the Sobolev spaces $H^s(\Gamma)$ and $H^{s,1}_{\langle\cdot,\cdot\rangle}(\Gamma)$ are isomorphic for $s \in \big(-\min\{\frac{1}{2}, s_\Gamma\}, \min\{\frac{3}{2}, s_\Gamma\}\big)$. Considering the canonical inner product, the cancellation property is in general not satisfied if the wavelet is supported on several patches. For such wavelets the cancellation property is true only if the surface measure is continuous across the interfaces of intersecting patches. However, a slight

Algorithm 3.2 This algorithm computes the two-scale decomposition $\Phi_{j+1}^{\Gamma}\mathbf{a}^{(j+1)} = \Phi_j^{\Gamma}\mathbf{a}^{(j)} + \Psi_j^{\Gamma}\mathbf{b}^{(j)}$, where $\mathbf{a}^{(j)} = \big[a_{\boldsymbol{\xi}}^{(j)}\big]_{\boldsymbol{\xi}\in\Delta_j^{\Gamma}}$ and $\mathbf{b}^{(j)} = \big[b_{\boldsymbol{\xi}}^{(j)}\big]_{\boldsymbol{\xi}\in\nabla_j^{\Gamma}}$.

initialization: $\mathbf{a}^{(j)} := \mathbf{b}^{(j)} := \mathbf{0}$
for $i = 1$ **to** M **do begin**
  for all $\mathbf{k} \in \Delta_j^{\square}$ **do begin** C: compute coefficients of Φ_j^{Γ}
    for all $\mathbf{k}' \in \Delta_{j+1}^{\square}$ **do begin**
      $a_{\gamma_i(\mathbf{k})}^{(j)} = a_{\gamma_i(\mathbf{k})}^{(j)} + \big[\mathbf{M}_{j,0}^{\square}\big]_{\mathbf{k}',\mathbf{k}} a_{\gamma_i(\mathbf{k}')}^{(j+1)} / r\big(\gamma_i(\mathbf{k}')\big)$
  end, end
  for all $\mathbf{k} \in \nabla_j^{\square}$ **do begin** C: compute coefficients of $\breve{\Psi}_j^{\Gamma}$
    for all $\mathbf{k}' \in \Delta_{j+1}^{\square}$ **do begin**
      $b_{\gamma_i(\mathbf{k})}^{(j)} = b_{\gamma_i(\mathbf{k})}^{(j)} + \big[\tilde{\mathbf{M}}_{j,1}^{\square}\big]_{\mathbf{k}',\mathbf{k}} a_{\gamma_i(\mathbf{k}')}^{(j+1)} / r\big(\gamma_i(\mathbf{k}')\big)$
end, end, end
for $i = 1$ **to** M **do begin**
  for all $\mathbf{k} \in \nabla_j^{\square}$ **do begin** C: add scaling functions to $\breve{\Psi}_j^{\Gamma}$
    for all $\mathbf{k}' \in \Delta_j^{\square}$ **do begin**
      $b_{\gamma_i(\mathbf{k})}^{(j)} = b_{\gamma_i(\mathbf{k})}^{(j)} - \big[\mathbf{L}_j^{\square}\big]_{\mathbf{k}',\mathbf{k}} a_{\gamma_i(\mathbf{k}')}^{(j)} / r\big(\gamma_i(\mathbf{k}')\big)$
end, end, end.

modification of the present construction will lead to globally continuous wavelets with patchwise vanishing moments, see [63] for the details.

4 The wavelet Galerkin scheme

This section is devoted to a fully discrete wavelet Galerkin scheme for boundary integral equations. After a historical overview on wavelet matrix compression (Subsection 4.1) we discretize the given boundary integral equation by the wavelets constructed in the previous section (Subsection 4.2). Then, in Subsection 4.3 we introduce the a-priori matrix compression which reduces the relevant matrix coefficients to an asymptotically linear number. In Subsections 4.4 and 4.5 we point out the computation of the compressed matrix. Next, in Subsection 4.6 we present an a-posteriori compression to reduce again the number of matrix coefficients. Subsection 4.7 is dedicated to the preconditioning of system matrices which arise from boundary integral operators of nonzero order. Subsection 4.8 is devoted to numerical results with respect to a nontrivial geometry. Finally, in Subsection 4.9 we consider the adaptive wavelet Galerkin scheme, also supported by numerical results.

4.1 Historical notes

The basic observation in the pioneering paper [4] was that the wavelet approximation of the kernel function of a singular integral operator contains a vast of small coefficients which can be neglected. G. Beylkin, R. Coifman, and V. Rokhlin demonstrated that the compression error can be controlled in terms of an operator norm. In principle, there are two different possibilities for defining bivariate wavelets, namely either wavelets with isotropic support on $\Gamma \times \Gamma$, given by a fixed level $\ell \in \mathbb{N}$, or tensor products of wavelets on Γ with various levels in the coordinate directions $\mathbf{j} = (j_1, j_2) \in \mathbb{N}^2$. The second approach reveals exactly the Galerkin matrix with respect to a univariate wavelet basis. It is called the *standard representation*, whereas the first one, called *nonstandard representation*, can be used only in conjunction with matrix-vector multiplication. The nonstandard representation has many in common with cluster methods like the fast multipole method and was favoured by the authors. For fixed accuracy in each level, the authors demonstrated that the compressed nonstandard representation scales log-linear with the size of the matrix. For linear complexity they referred to well established but rather nontrivial techniques from harmonic analysis, see e.g. [72].

The paper [4] triggered the collaboration of W. Dahmen, S. Prössdorf and R. Schneider ([24, 27, 28, 29, 30]). Their first paper [29] develops a framework of general Petrov-Galerkin methods for pseudo-differential operator equations using functions in shift-invariant spaces. In the second paper [28] they went through the arguments from harmonic analysis [18, 19] and proofed the linear complexity of matrix compression in the standard and nonstandard form under the assumptions of [4].

In [27, 28], they moreover treated a more complicated question, communicated by W. Hackbusch, namely how to adopt the compression error to the actually required accuracy. For a periodic integral equation one can define a compression of the standard form, taking a log-linear number of relevant entries into account, such that the solution of the compressed scheme differs from the exact solution at most by an error bounded by the actual discretization accuracy. Linear complexity has been concluded for a nearly optimal convergence rate. The analysis is based on the investigation of consistency and exploits the smoothness of the underlying solution. However, it does not apply to the nonstandard form. The conclusion drawn from this result is that, under the perspective of asymptotic accuracy, the standard form is superior compared to the nonstandard form, contradicting the suggestion from [4].

The estimate for the matrix coefficients which correspond to distant basis functions is based on the Taylor expansion. Compared to the standard scaling functions, there is an additional decay induced by the vanishing moments of the wavelets. The decay estimate employed in the early papers was essentially improved in [30, 79, 80, 95] by using a twofold Taylor expansion, in a similar way as in a sparse grid approximation. The resulting compression is referred to as the *first compression*.

A more detailed investigation is required in case of interactions of wavelets with overlapping supports. A pseudo-differential operator does not spoil the local

smoothness of a piecewise polynomial function. In fact, it does not change the singular support, a well known fact in harmonic analysis. If a fine scale wavelet $\psi_{j,\mathbf{k}}$ is supported on a region where a coarse scale wavelet $\psi_{j',\mathbf{k}'}$ is smooth, then one can exploit again vanishing moments when testing $\mathscr{A}\psi_{j,\mathbf{k}}$ with the locally smooth function $\psi_{j',\mathbf{k}'}$. The related matrix coefficient becomes small, depending on the distance of the support of the small wavelet to the singular support of the coarse wavelet. The resulting additional compression is called the *second compression* [22, 95]. By this advanced technique it is possible to achieve optimal convergence rates in linear complexity.

Engineers often prefer the collocation instead of the Galerkin scheme. The collocation scheme fits into the general framework of Petrov-Galerkin methods. It has been shown in [95] how the biorthogonal basis, namely the hierarchical basis, together with a biorthogonal multiresolution analysis can be used for wavelet matrix compression. Therein, log-linear complexity for the collocation scheme has been derived. This result has been improved to linear complexity in [86, 87]. Multiwavelet bases dual to Dirac distributions have been developed in [10] and applied to wavelet matrix compression for collocating a second kind Fredholm integral equation in [11].

4.2 Discretization

In what follows, the collection Ψ_J with a capital J denotes the finite wavelet basis in the space V_J, i.e., $\Psi_J := \bigcup_{j=j_0-1}^{J-1} \Psi_j$. Further, $N_J := \dim V_J \sim 4^J$ indicates the number of unknowns.

The variational formulation of the given boundary integral equation (1) reads:

$$\text{seek } u \in H^q(\Gamma): \quad (\mathscr{A}u, v)_{L^2(\Gamma)} = (f, v)_{L^2(\Gamma)} \quad \text{for all } v \in H^q(\Gamma). \tag{39}$$

It is well known, that the variational formulation (39) is equivalent to the boundary integral equation (1), see e.g. [46, 93] for details.

To gain the Galerkin method we replace the energy space $H^q(\Gamma)$ in the variational formulation (39) by the finite dimensional space V_J of piecewise constant or bilinear functions, as introduced in the previous section. Then, we arrive at the problem

$$\text{seek } u_J \in V_J: \quad (\mathscr{A}u_J, v_J)_{L^2(\Gamma)} = (f, v_J)_{L^2(\Gamma)} \quad \text{for all } v_J \in V_J.$$

Traditionally this equation is discretized by the single-scale basis of V_J which yields a densely populated system matrix (see the left plot of Fig. 18). Whereas, since the kernel function is smooth apart from the diagonal, the discretization by wavelets with a sufficiently strong cancellation property (17) leads to a quasi-sparse system matrix (see the right plot of Fig. 18). Most matrix coefficients are negligible without compromising the order of convergence of the Galerkin scheme. Thus, we shall employ the wavelet basis in V_J for discretization, making the ansatz $u_J = \Psi_J \mathbf{u}_J$,

and obtain the *wavelet Galerkin scheme*

$$\mathbf{A}_J\mathbf{u}_J = \mathbf{f}_J, \qquad \mathbf{A}_J = \big(\mathscr{A}\Psi_J, \Psi_J\big)_{L^2(\Gamma)}, \qquad \mathbf{f}_J = \big(f, \Psi_J\big)_{L^2(\Gamma)}. \tag{40}$$

Remark 4.1. Replacing in (40) the wavelet basis Ψ_J by the single-scale basis Φ_J yields the traditional single-scale Galerkin scheme $\mathbf{A}_J^\phi \mathbf{u}_J^\phi = \mathbf{f}_J^\phi$, where we have set $\mathbf{A}_J^\phi := \big(\mathscr{A}\Phi_J, \Phi_J\big)_{L^2(\Gamma)}$, $\mathbf{f}_J^\phi := \big(f, \Phi_J\big)_{L^2(\Gamma)}$ and $u_J = \Phi_J \mathbf{u}_J^\phi$. This scheme is related to the wavelet Galerkin scheme by

$$\mathbf{A}_J^\psi = \mathbf{T}_J \mathbf{A}_J^\phi \mathbf{T}_J^T, \qquad \mathbf{u}_J^\psi = \mathbf{T}_J^{-T} \mathbf{u}_J^\phi, \qquad \mathbf{f}_J^\psi = \mathbf{T}_J \mathbf{f}_J^\phi,$$

where $\mathbf{T}_J$ denotes the wavelet transform. Since the system matrix $\mathbf{A}_J^\phi$ is densely populated, the naive solution of a given boundary integral equation in the single-scale basis costs at least $\mathscr{O}(N_J^2)$.

4.3 A-priori compression

The system matrix in wavelet coordinates is quasi-sparse and can be compressed without compromising the discretization error. In a first compression step, all matrix entries, for which the distance of the supports between the associated trial and test functions is larger than a level depending cut-off parameter $\mathscr{B}_{j,j'}$, are set to zero. The second compression, reflected by the cut-off parameter $\mathscr{B}'_{j,j'}$, affects those remaining matrix entries, for which the corresponding trial and test functions have overlapping supports. Both situations are illustrated in Fig. 4. The resulting compression pattern of the matrix is descriptively called *finger structure*, see Fig. 19.

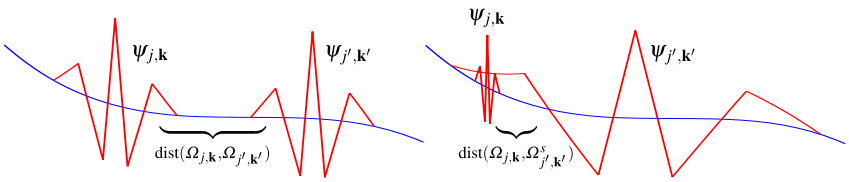


Fig. 4 The situations affected by the first (left) and the second compression (right)

To formulate the compression rules, we introduce the abbreviation

$$\Omega_{j,\mathbf{k}} := \mathrm{conv\,hull}(\mathrm{supp}\,\psi_{j,\mathbf{k}}), \qquad \Omega^s_{j,\mathbf{k}} := \mathrm{sing\,supp}\,\psi_{j,\mathbf{k}}. \tag{41}$$

Notice that the first expression denotes the convex hull of the support of a wavelet with respect to the Euclidean space $\mathbb{R}^3$. The second expression indicates the *singular support*, i.e., that subset of Γ where the wavelet is not smooth.

Theorem 4.1 (A-priori compression [22]). *Let $\Omega_{j,\mathbf{k}}$ and $\Omega'_{j,\mathbf{k}}$ be given as in (41) and define the compressed system matrix $\mathbf{A}_J$, corresponding to the boundary integral operator $\mathscr{A}$, by*

$$[\mathbf{A}_J]_{(j,\mathbf{k}),(j',\mathbf{k}')} := \tag{42}$$
$$\begin{cases} 0, & \operatorname{dist}(\Omega_{j,\mathbf{k}},\Omega_{j',\mathbf{k}'}) > \mathscr{B}_{j,j'} \text{ and } j,j' \geq j_0, \\ 0, & \operatorname{dist}(\Omega_{j,\mathbf{k}},\Omega_{j',\mathbf{k}'}) \lesssim 2^{-\min\{j,j'\}} \text{ and} \\ & \operatorname{dist}(\Omega^s_{j,\mathbf{k}},\Omega_{j',\mathbf{k}'}) > \mathscr{B}^s_{j,j'} \text{ if } j' > j \geq j_0 - 1, \\ & \operatorname{dist}(\Omega_{j,\mathbf{k}},\Omega^s_{j',\mathbf{k}'}) > \mathscr{B}^s_{j,j'} \text{ if } j > j' \geq j_0 - 1, \\ (\mathscr{A}\psi_{j',\mathbf{k}'},\psi_{j,\mathbf{k}})_{L^2(\Gamma)}, & \text{otherwise.} \end{cases}$$

Fixing

$$a > 1, \qquad d < \delta < \widetilde{d} + 2q, \tag{43}$$

the cut-off parameters $\mathscr{B}_{j,j'}$ and $\mathscr{B}^s_{j,j'}$ are set as follows

$$\begin{aligned} \mathscr{B}_{j,j'} &= a \max\left\{2^{-\min\{j,j'\}}, 2^{\frac{2J(\delta-q)-(j+j')(\delta+\widetilde{d})}{2(\widetilde{d}+q)}}\right\}, \\ \mathscr{B}^s_{j,j'} &= a \max\left\{2^{-\max\{j,j'\}}, 2^{\frac{2J(\delta-q)-(j+j')\delta-\max\{j,j'\}\widetilde{d}}{\widetilde{d}+2q}}\right\}. \end{aligned} \tag{44}$$

Then, the system matrix $\mathbf{A}_J$ has only $\mathscr{O}(N_J)$ nonzero coefficients. Moreover, the error estimate

$$\|u - u_J\|_{H^{2q-d}(\Gamma)} \lesssim 2^{2J(q-d)} \|u\|_{H^d(\Gamma)} \tag{45}$$

holds for the solution u_J of the compressed Galerkin system provided that u and Γ are sufficiently regular.

The next theorem shows that the over-all complexity of assembling the compressed system matrix with sufficient accuracy can be kept of the order $\mathscr{O}(N_J)$, even when a computational cost of logarithmic order is allowed for each entry. This theorem will be used later as the essential ingredient for proving that the quadrature strategy proposed in Subsection 4.5 scales linearly.

Theorem 4.2 (Complexity [22, 49]). *The complexity of computing the compressed system matrix $\mathbf{A}_J$ is $\mathscr{O}(N_J)$ if the calculation of its relevant entries $(\mathscr{A}\psi_{j',\mathbf{k}'},\psi_{j,\mathbf{k}})_{L^2(\Gamma)}$ is performed in $\mathscr{O}\big(\big[J - \frac{j+j'}{2}\big]^\alpha\big)$ operations with some $\alpha \geq 0$.*

4.4 Setting up the compression pattern

Checking the distance criterion (42) for each matrix coefficient, in order to assemble the compressed matrix, would require $\mathscr{O}(N_J^2)$ function calls. To realize linear com-

plexity, we exploit the underlying tree structure with respect to the supports of the wavelets, to predict negligible matrix coefficients. We will call a wavelet $\psi_{j+1,\mathrm{son}}$ a son of $\psi_{j,\mathrm{father}}$ if $\Omega_{j+1,\mathrm{son}} \subseteq \Omega_{j,\mathrm{father}}$. The following observation is an immediate consequence of the relations $\mathscr{B}_{j,j'} \geq \mathscr{B}_{j+1,j'} \geq \mathscr{B}_{j+1,j+1'}$, and $\mathscr{B}^s_{j,j'} \geq \mathscr{B}^s_{j+1,j'}$ if $j > j'$.

Lemma 4.1. *We consider $\Omega_{j+1,\mathrm{son}} \subseteq \Omega_{j,\mathrm{father}}$ and $\Omega_{j'+1,\mathrm{son}} \subseteq \Omega_{j',\mathrm{father}}$.*

1. *If*

$$\operatorname{dist}\left(\Omega_{j,\mathrm{father}}, \Omega_{j',\mathrm{father}'}\right) > \mathscr{B}_{j,j'}$$

then there holds

$$\operatorname{dist}\left(\Omega_{j+1,\mathrm{son}}, \Omega_{j',\mathrm{father}'}\right) > \mathscr{B}_{j+1,j'},$$
$$\operatorname{dist}\left(\Omega_{j+1,\mathrm{son}}, \Omega_{j'+1,\mathrm{son}'}\right) > \mathscr{B}_{j+1,j+1'}.$$

2. *For $j > j'$ suppose*

$$\operatorname{dist}\left(\Omega_{j,\mathrm{father}}, \Omega^s_{j',\mathrm{father}'}\right) > \mathscr{B}^s_{j,j'}$$

then we can conclude that

$$\operatorname{dist}\left(\Omega_{j+1,\mathrm{son}}, \Omega^s_{j',\mathrm{father}'}\right) > \mathscr{B}^s_{j+1,j'}$$

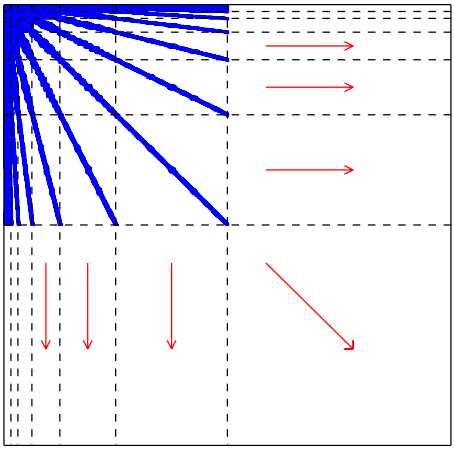


Fig. 5 The compression pattern are computed successively by starting from the coarse grids

With the aid of this lemma we have to check the distance criteria only for coefficients which stem from subdivisions of calculated coefficients on a coarser level, cf. Fig. 5. Therefore, the resulting procedure of checking the distance criterion is still of linear complexity.

4.5 Computation of matrix coefficients

The significant matrix coefficients $(\mathscr{A}\psi_{j',\mathbf{k}'}, \psi_{j,\mathbf{k}})_{L^2(\Gamma)}$ retained by the compression strategy can generally neither be determined analytically nor be computed exactly. Therefore we have to approximate the matrix coefficients by quadrature rules. This causes an additional error which has to be controlled with regard to our overall objective of realizing asymptotically optimal accuracy while preserving efficiency. Thm. 4.2 describes the maximal allowed computational expenses for the computation of the individual matrix coefficients so as to realize still overall linear complexity.

The following theorem tells us that sufficient accuracy requires only a level dependent precision of quadrature for computing the retained matrix coefficients.

Theorem 4.3 (Accuracy [22, 49]). *Let the error of quadrature for computing the relevant matrix coefficients $(\mathscr{A}\psi_{j',\mathbf{k}'}, \psi_{j,\mathbf{k}})_{L^2(\Gamma)}$ be bounded by the level dependent threshold*

$$\varepsilon_{j,j'} \sim \min\left\{2^{-|j-j'|}, 2^{-4(J-\frac{j+j'}{2})\frac{\delta-q}{d+q}}\right\} 2^{2Jq} 2^{-2\delta(J-\frac{j+j'}{2})} \tag{46}$$

with $\delta \in (d, \widetilde{d}+r)$ from (43). Then, the Galerkin scheme is stable and converges with the optimal order (45).

From (46) we conclude that the entries on the coarse grids have to be computed with the full accuracy while the entries on the finer grids are allowed to have less accuracy. Unfortunately, the domains of integration are very large on coarser scales.

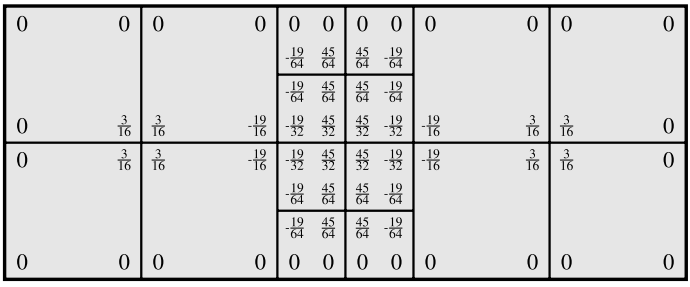


Fig. 6 Element-based representation of a piecewise bilinear wavelet with four vanishing moments

According to the fact that a wavelet is a linear combination of scaling functions, the numerical integration can be reduced to interactions of polynomial shape functions on certain elements. This suggests to employ an element-based representation of the wavelets like illustrated in Fig. 6 in the case of a piecewise bilinear wavelet. Consequently, we have only to deal with integrals of the form

$$I_{\ell,\ell'}(\Gamma_{i,j,\mathbf{k}}, \Gamma_{i',j',\mathbf{k}'}) := \int_{C_{j,\mathbf{k}}} \int_{C_{j',\mathbf{k}'}} k_{i,i'}(\mathbf{s},\mathbf{t})\, p_\ell(\mathbf{s})\, p_{\ell'}(\mathbf{t})\, d\mathbf{t}\, d\mathbf{s} \tag{47}$$

with p_ℓ denoting the polynomial shape functions. This is quite similar to the traditional Galerkin discretization. The main difference is that in the wavelet approach the elements may appear on different levels due to the multilevel nature of wavelet bases.

Difficulties arise if the domains of integration are very close together relatively to their size. We have to apply numerical integration with some care in order to keep the number of evaluations of the kernel function at the quadrature nodes moderate and to fulfill the requirements of Thm. 4.2. The necessary accuracy can be achieved within the allowed expenses if we employ an exponentially convergent quadrature method.

In [49, 61, 95] a geometrically graded subdivision of meshes is proposed in combination with varying polynomial degrees of approximation in the integration rules, cf. Fig. 7. Exponential convergence can be achieved for boundary integral operators which are *analytically standard*.

Definition 4.1. We call the kernel $k(\mathbf{x},\mathbf{y})$ an analytically standard kernel of the order $2q$ if the partial derivatives of the transported kernel functions $k_{i,i'}(\mathbf{s},\mathbf{t})$, $1 \le i,i' \le M$, satisfy

$$\left|\partial_{\mathbf{s}}^{\boldsymbol{\alpha}}\partial_{\mathbf{t}}^{\boldsymbol{\beta}} k_{i,i'}(\mathbf{s},\mathbf{t})\right| \le \frac{(|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|)!}{(r\left\|\gamma_i(\mathbf{s})-\gamma_{i'}(\mathbf{t})\right\|)^{2+2q+|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|}}$$

for some $r>0$ provided that $2+2q+|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|>0$.

Generally, the kernels of boundary integral operators are analytically standard under the assumption that the underlying manifolds are patchwise analytic. It is shown in [49, 61] that an *hp*-quadrature scheme, based on tensor product Gauß-Legendre quadrature rules, leads to a number of quadrature points that satisfies the assumptions of Thm. 4.2 with $\alpha = 4$. Since the proofs are rather technical we refer the reader to [49, 61, 81, 95, 97] for further details.

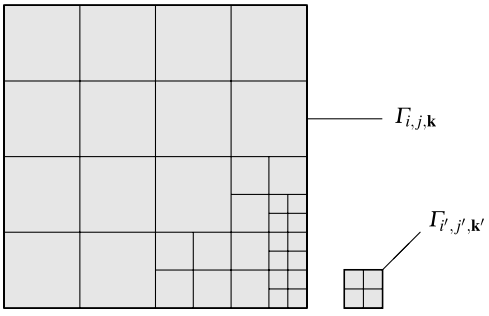


Fig. 7 Adaptive subdivision of the domains of integration

Since the kernel function has a singularity on the diagonal, we are still confronted with singular integrals if the domains of integration live on the same level and have some points in common. This happens if the underlying elements are identical or

share a common edge or vertex. When we do not deal with weakly singular integral operators, the operators can be regularized, for instance by integration by parts [74]. So we end up with weakly singular integrals. Such weakly singular integrals can be treated by the so-called *Duffy-trick* [36, 92] in order to transform the singular integrands into analytical ones.

4.6 A-posteriori compression

If the entries of the compressed system matrix $\mathbf{A}_J$ have been computed, we may apply an a-posteriori compression by setting all entries to zero, which are smaller than a level dependent threshold. That way, a matrix $\widehat{\mathbf{A}}_J$ is obtained which has less nonzero entries than the matrix $\mathbf{A}_J$. Clearly, this does not accelerate the calculation of the matrix coefficients. But the requirement to the memory is reduced if the system matrix needs to be stored. Especially if the linear system of equations has to be solved for several right hand sides, like for instance in shape optimization (cf. [37, 50]) or inverse obstacle problems (cf. [51, 52]), the faster matrix-vector multiplication pays off. To our experience the a-posteriori compression reduces the number of nonzero coefficients by a factor 2–5.

Theorem 4.4 (A-posteriori compression [22, 49]). *We define the a-posteriori compression by*

$$\left[\widehat{\mathbf{A}}_J\right]_{(j,\mathbf{k}),(j',\mathbf{k}')} = \begin{cases} 0, & \text{if } \left|[\mathbf{A}_J]_{(j,\mathbf{k}),(j',\mathbf{k}')}\right| \le \varepsilon_{j,j'}, \\ [\mathbf{A}_J]_{(j,\mathbf{k}),(j',\mathbf{k}')}, & \text{if } \left|[\mathbf{A}_J]_{(j,\mathbf{k}),(j',\mathbf{k}')}\right| > \varepsilon_{j,j'}, \end{cases}$$

where the level dependent threshold $\varepsilon_{j,j'}$ is chosen as in (46) with $\delta \in (d, \widetilde{d}+r)$ from (43). Then, the optimal order of convergence (45) of the Galerkin scheme is not compromised.

4.7 Wavelet preconditioning

The system matrices arising from operators of nonzero order are ill conditioned since there holds $\operatorname{cond}_{\ell^2} \mathbf{A}_J \sim 2^{2J|q|}$. According to [21, 25, 95], the wavelet approach offers a simple diagonal preconditioner based on the norm equivalences (16).

The norm equivalences have been stated first in [72] for orthogonal wavelet bases. At the same time a new multilevel preconditioner, nowadays called the BPX scheme, has been discussed in the literature [5]. In [77, 78] it has been shown that this preconditioner leads to uniformly bounded condition numbers. The arguments used there are based on the moduli of smoothness and Besov norms. This approach was extended in [25] by additional approximation theoretic arguments like K-functionals to general multilevel bases, including wavelets. These papers inspired

a new effort in the development of norm equivalences, see e.g. the monograph [12] and the literature cited therein, and generating systems [43, 44].

In [107] a strengthened Cauchy-Schwarz inequality has been used to show bounded condition numbers of the BPX. This technique was employed in [95] to precondition operators of negative order by wavelets. The main result is that biorthogonal wavelets satisfy the norm equivalence to Sobolev norms in H^s if and only if $-\widetilde{\gamma} < s < \gamma$. As an immediate consequence one has to look for wavelets with $H^{1/2}$-regular duals to gain optimal preconditioners for operators of the order -1, for instance the single layer operator. Wavelets on manifolds with arbitrarily prescribed smoothness have been developed for this purpose in [32]. However, these wavelets have never been used since first numerical experiments were completely discouraging. On the other hand, the composite wavelet bases from [31] lead to moderate condition numbers, even for extremely large systems.

Theorem 4.5 (Preconditioning [25, 95]). *Let the diagonal matrix $\mathbf{D}_J^r$ be defined by*

$$\left[\mathbf{D}_J^r\right]_{(j,\mathbf{k}),(j',\mathbf{k}')} = 2^{rj}\delta_{j,j'}\delta_{\mathbf{k},\mathbf{k}'}, \quad \mathbf{k} \in \nabla_j, \quad \mathbf{k}' \in \nabla_{j'}, \quad j_0 - 1 \le j, j' < J. \tag{48}$$

Then, if $\widetilde{\gamma} > -q$, the diagonal matrix $\mathbf{D}_J^{2q}$ defines an asymptotically optimal preconditioner to $\mathbf{A}_J$, i.e., $\operatorname{cond}_{\ell^2}(\mathbf{D}_J^{-q}\mathbf{A}_J\mathbf{D}_J^{-q}) \sim 1$.

It should be stressed that while the above scaling is asymptotically optimal, the quantitative performance may vary significantly among different scalings with the same asymptotic behavior. In particular, since Ψ is, on account of the mapping properties of $\mathscr{A}$ and the norm equivalences (16), also a Riesz basis with respect to the energy norm, it would be natural to normalize the wavelets in this energy norm which would suggest the specific scaling $\left(\mathscr{A}\psi_{j,\mathbf{k}}, \psi_{j,\mathbf{k}}\right)_{L^2(\Gamma)} \sim 2^{2qj}$. In fact, this latter diagonal scaling improves and even simplifies the wavelet preconditioning.

As the numerical results in e.g. [57] confirm, this preconditioning works well in the two dimensional case. However, in three dimensions, the results are not satisfactory. Fig. 8 refers to the ℓ^2-condition numbers of the stiffness matrices arising from the single layer operator on the unit square, discretized by piecewise bilinears. Even though the condition numbers with respect to the wavelet bases are bounded, they are not significantly lower than with respect to the single-scale basis. We mention that the situation becomes even worse for operators which are defined on more complicated geometries.

A slight modification of the wavelet preconditioner offers a significant improvement. The simple trick is to combine the above preconditioner with the mass matrix which yields an appropriate *operator based preconditioning*, cf. [49].

Theorem 4.6. *Let $\mathbf{D}_J^r$ be defined as in (48) and $\mathbf{B}_J := (\Psi_J, \Psi_J)_{L^2(\Gamma)}$ denote the mass matrix. Then, if $\widetilde{\gamma} > -q$, the matrix $\mathbf{C}_J^{2q} = \mathbf{D}_J^q\mathbf{B}_J\mathbf{D}_J^q$ defines an asymptotically optimal preconditioner to $\mathbf{A}_J$, i.e.,*

$$\operatorname{cond}_{\ell^2}\left(\left(\mathbf{C}_J^{2q}\right)^{-1/2}\mathbf{A}_J\left(\mathbf{C}_J^{2q}\right)^{-1/2}\right) \sim 1.$$

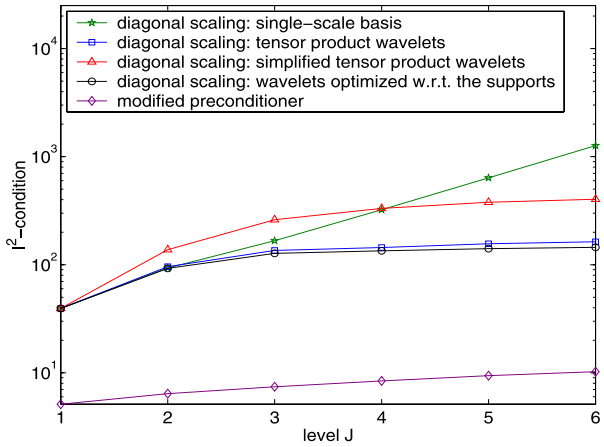


Fig. 8 ℓ^2-condition numbers for the single layer operator on the unit square

This preconditioner decreases the condition numbers impressively, cf. Fig. 8. Moreover, the condition depends now only on the underlying spaces but not on the particular wavelet basis. To our experiences the condition reduces about the factor 10 compared to the preconditioner (48). We like to mention that, by employing the fast wavelet transform, the application of this preconditioner requires only the inversion of a single-scale mass matrix, which is diagonal in case of piecewise constant and very sparse in case of piecewise bilinear ansatz functions.

4.8 Numerical results

To complement the theoretical results by quantitative numerical studies we consider as a first example the case of a smooth surface, namely the unit sphere. Here we expect to encounter the highest obtainable convergence rate.

We solve an interior Dirichlet problem for the Laplacian by the indirect approach using the single layer potential operator. This gives rise to a Fredholm integral equation of the first kind for an unknown density $\rho \in H^{-1/2}(\Gamma)$. Hence, in particular, preconditioning is an issue. The surface of the sphere is parameterized with the aid of six patches. As Dirichlet data we choose the restriction of the harmonic function

$$U(\mathbf{x}) = \frac{\langle \mathbf{a}, \mathbf{x}-\mathbf{b}\rangle}{\|\mathbf{x}-\mathbf{b}\|^3}, \quad \mathbf{a} = [1,2,4]^T, \quad \mathbf{b} = [1.5,0,0]^T \notin \Omega \tag{49}$$

to Γ. Then, U is the unique solution of the Dirichlet problem. We discretize the given boundary integral equation by piecewise constant wavelets with three vanishing moments. For the computation of the potential U we expect the pointwise convergence rate

$$|U(\mathbf{x}) - U_J(\mathbf{x})| \lesssim \|\rho - \rho_J\|_{H^{-2}(\Gamma)} \lesssim 2^{-3J}\|\rho\|_{H^1(\Gamma)}, \quad \mathbf{x} \in \Omega,$$

see e.g. [46, 93, 98, 106].

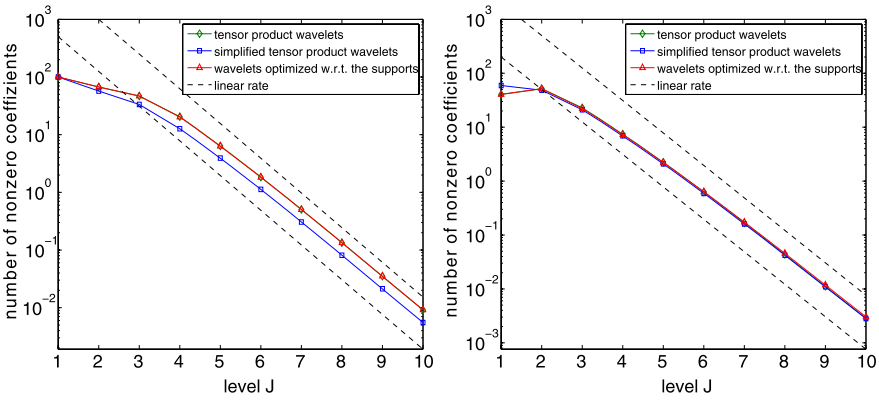


Fig. 9 A-priori (left) and a-posteriori (right) compression rates in case of the unit sphere

In Fig. 9 the a-priori (left) and a-posteriori (right) compression rates are reported via the ratio of the number of nonzero entries of the compressed matrix and N_J^2. The dashed lines indicate linear behaviour. We computed the compression rates with respect to all the three wavelet constructions from Section 3. The best a-priori compression is produced by the wavelets with optimized supports which issues from their much smaller supports. Whereas, with respect to the a-posteriori compression, all constructions yield comparable results.

Regarding the computing times, we observe once more that the wavelets with optimized supports perform best. The computing time is about half as much as in case of the tensor product wavelets. The speed-up is still about 10 percent in comparison with the simplified tensor product wavelets.

Fig. 10 displays the behavior of the approximation error of our scheme (here, in case of the optimized wavelet basis). We evaluate the approximate potential at the points $[0,0,0]^T$, $[0.25,0.25,0.25]^T$, and $[0.5,0.5,0.5]^T$ and measure its absolute error when compared with the exact solution. In addition, we evaluate 1681 points, distributed uniformly in the interior of the sphere, and measure the ℓ^∞-norm of the pointwise absolute errors. We see that the compression does not spoil the optimal order of convergence which is indicated by the dashed lines.

As a second example we consider a more complicated geometry, namely a crankshaft of a parallel twin motor (as used in old British motorcycles), cf. Fig. 11. The surface of this crankshaft is parameterized with the aid of 142 patches. The problem under consideration is the same as above, in particular, we choose the same function U (49).

In order to measure the error produced by the method, we calculate the approximate solution $U_J = \mathcal{V}\rho_J$ in several points $\mathbf{x}_i$ located inside the domain, depicted in

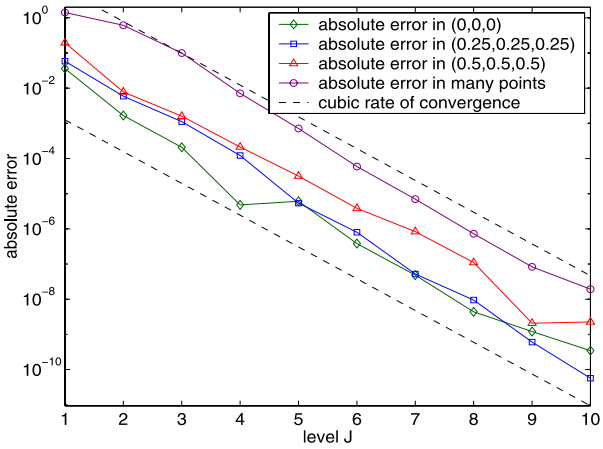


Fig. 10 Absolute errors of the approximate solution in case of the sphere

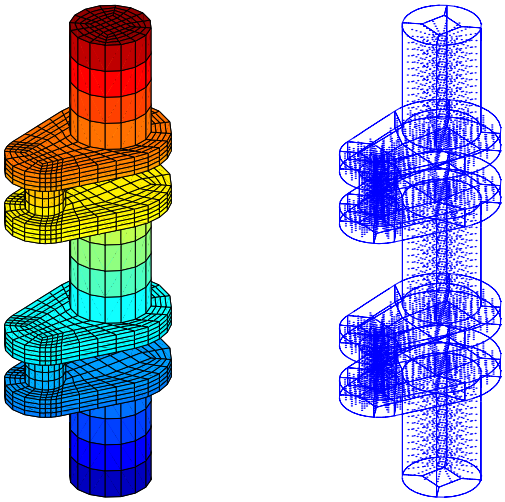


Fig. 11 The surface mesh and the evaluation points x_i of the potential

Fig. 11. The discrete potentials are denoted by

$$\mathbf{U} := [U(\mathbf{x}_i)], \qquad \mathbf{U}_J := [(\mathscr{Y}\rho_J)(\mathbf{x}_i)].$$

We list in Table 1 the results produced by the wavelet Galerkin scheme. Due to present edge singularities the solution is not in $H^1(\Gamma)$ and we thus cannot expect the full convergence rate. However, the presence of the singularities does not require any

change of parameters (see (44)). For $N_J \leq 9088$ we have also computed the solution of the uncompressed scheme. The corresponding absolute errors for the traditional Galerkin method are 1.0 if $N_J = 2272$ and $2.4 \cdot 10^{-1}$ if $N_J = 9088$. This shows again that the present compression does not degrade the accuracy of the Galerkin scheme. Since there is no difference in accuracy at low level compared to the unperturbed scheme and since we can inspect the convergence history we can expect to have a reliably accurate solution.

unknowns		piecewise constant wavelets $\psi^{(1,3)}$					
J	N_J	$\|\mathbf{U}-\mathbf{U}_J\|_\infty$	cpu-time (in sec.)	a-priori compression (nnz in %)	a-posteriori compression (nnz in %)	cg-iterations	memory requirements
1	568	13.7	0	27	20	24	3.2 MB
2	2272	1.0 (14)	0	8.7	6.8	36	11 MB
3	9088	$2.4 \cdot 10^{-1}$ (4.3)	7	3.2	1.9	54	32 MB
4	36352	$1.6 \cdot 10^{-2}$ (15)	52	0.93	0.42	77	128 MB
5	145408	$5.4 \cdot 10^{-3}$ (3.0)	280	0.25	0.097	86	524 MB
6	581632	$2.1 \cdot 10^{-3}$ (2.5)	1773	0.064	0.024	92	2.1 GB
7	2.3 Mio.	$1.9 \cdot 10^{-4}$ (9.0)	9588	0.016	0.0059	101	8.3 GB
8	9.3 Mio.	$2.7 \cdot 10^{-5}$ (7.0)	49189	0.0040	0.0015	110	29 GB

Table 1 Numerical results with respect to the crankshaft

For 9.3 million unknowns, only 0.0040% of the entries have to be computed. After the a-posteriori compression even only 0.0015% nonzero entries are used for the computation of the solution ρ_J. In the mean one has thus 375 (a-priori), respectively 138 (a-posteriori) coefficients per unknown. In our wavelet Galerkin scheme we have allocated about 29 Gigabyte storage for the solution of 9.3 million unknowns.

In general more than 95% of the computing time is consumed by the precomputational steps, namely setting up the matrix pattern and assembling the compressed Galerkin matrix. Whereas, the iterative solution is quite fast. A precise breakdown of the computing times can be found in Fig. 12 in terms of a bar diagramme. All computations were carried out on a single processor of SUN Fire X4600 M2 Server, equipped with eight 3.0 GHz AMD Opteron DualCore processors and 32 GB RAM per processor.

4.9 Adaptivity

A simple adaptive refinement strategy for the wavelet Galerkin scheme has been proposed in [58, 60]. Here, a mesh refinement takes place if the hierarchical increment is large. However, convergence of such an empirical strategy can be shown only under the assumption of the so-called *saturation assumption*. A strictly prov-

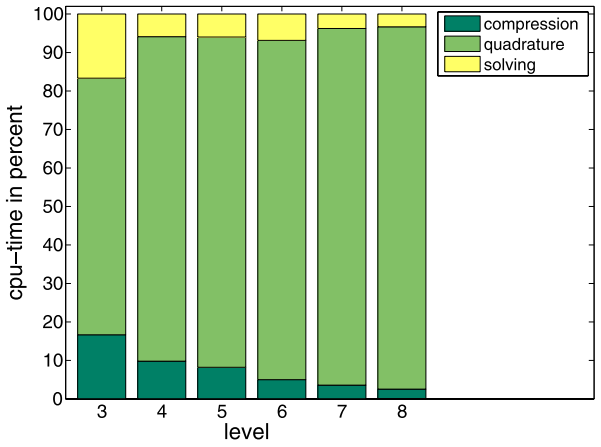


Fig. 12 Distribution of computing times in percent

able and efficient adaptive algorithm, developed in [23], is based on the techniques of A. Cohen, W. Dahmen, and R. DeVore from [13, 14, 15].

A core ingredient of the adaptive strategy is the approximate application of (infinite dimensional) operators that ensures asymptotically optimal complexity in the following sense. If (in the given discretization framework) the unknown solution u can be approximated in the energy norm with an optimal choice of N degrees of freedom at a rate N^{-s}, then the adaptive scheme matches this rate by producing for any target accuracy ε an approximate solution u_ε such that $\|u-u_\varepsilon\|_{H^q(\Gamma)} \leq \varepsilon$ at a computational expense that stays proportionally to $\varepsilon^{-1/s}$ as ε tends to zero, see [13, 14]. Notice that $N^{-\bar{s}}$, where $\bar{s} := (d-q)/2$, is the best possible rate of convergence, gained in case of uniform refinement if $u \in H^d(\Gamma)$. Since the computation of the relevant matrix coefficients is by far the most expensive step in our algorithm, we cannot use the approach of [13, 14]. Thus, in [23] the strategy of the *best N-term approximation* has been adopted by the notion of *tree approximation*, as considered in [1, 15, 33].

The algorithm is based on an iterative method for the *continuous equation* (1), expanded with respect to the wavelet basis. To this end we assume the wavelet basis Ψ to be normalized in the energy space. Then, (1) is equivalent to the well posed problem of finding $u = \Psi\mathbf{u}$ such that the *infinite dimensional* linear system of equations

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \quad \mathbf{A} = (\mathscr{A}\Psi, \Psi)_{L^2(\Gamma)}, \quad \mathbf{f} = (f, \Psi)_{L^2(\Gamma)} \tag{50}$$

holds.

The application of the operator $\mathscr{A}$ to a function is then approximated by an appropriate (finite dimensional) matrix-vector multiplication. Given a finitely supported vector $\mathbf{v}$ and a target accuracy ε, we choose *wavelet trees* τ_j according to

$$\left\|\mathbf{v}-\mathbf{v}|_{\tau_j}\right\|_{\ell^2} \le 2^{j\bar{s}}\varepsilon, \quad j=0,1,\ldots,J := \left\lceil \frac{\log_2(\|\mathbf{v}\|_{\ell^2}/\varepsilon)}{\bar{s}} \right\rceil$$

and define the *layers* $\Delta_j := \tau_{j+1} \setminus \tau_j$. These layers play now the role of the levels in case of the non-adaptive scheme, i.e., the accuracy will be balanced *layer-dependent*.

By adopting the compression rules from [99] we can define operators $\mathbf{A}_j$, having only $\mathcal{O}\big(2^j(1+j)^{-6}\big)$ relevant coefficients per row and column while satisfying

$$\|\mathbf{A}-\mathbf{A}_j\|_{\ell^2} \le \frac{2^{-j\bar{s}}}{(1+j)^6}.$$

Then, the approximate matrix-vector multiplication

$$\mathbf{w} := \sum_{j=0}^{J-1} \mathbf{A}_j \mathbf{v}|_{\Delta_j}$$

gives raise to the estimate

$$\|\mathbf{A}\mathbf{v}-\mathbf{w}\|_{\ell^2} \le \sum_{j=0}^{J-1} \left\|(\mathbf{A}-\mathbf{A}_j)\mathbf{v}|_{\Delta_j}\right\|_{\ell^2} \le \sum_{j=0}^{J-1} \frac{2^{-j\bar{s}}}{(1+j)^6} 2^{j\bar{s}}\varepsilon \le \varepsilon.$$

For this error estimate a logarithmic factor $(1+j)^{-2}$ would be sufficient but the larger exponent is required to avoid logarithmic terms in the complexity estimates, stemming from the quadrature in accordance with Subsection 4.5. By combing this approximate matrix-vector product with a suitable iterative solver for (50) (cf. [13]) or the adaptive Galerkin type algorithm from [40] we achieve the desired goal of optimal complexity. We skip further details here and refer the reader to [23].

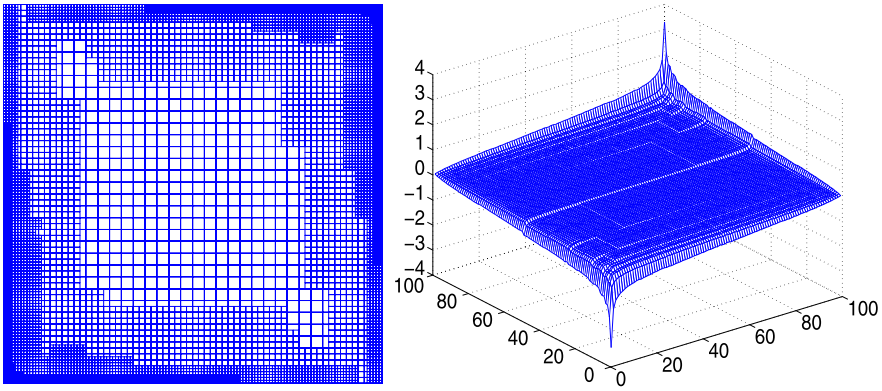


Fig. 13 The adaptive mesh on the domain Ω and the associated approximate solution σ_ε

For the numerical illustration we consider a magneto-static problem arising in thin-film micromagnetics. For a weak external field the magnetic charge σ fulfills the variational formulation of the Dirichlet screen problem. Its solution amounts to the inversion of the single layer potential on the sample's surface Γ, that is

$$\frac{1}{4\pi}\int_{\Gamma}\frac{\sigma(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|}d\mathbf{y} = -\langle \mathbf{H}_{\text{ext}}, \mathbf{x}\rangle \quad \text{on } \Gamma \tag{51}$$

with $\mathbf{H}_{\text{ext}}$ being the applied external field, see [34, 35] for the details concerning the modeling.

We discretize Eq. (51) by piecewise constant wavelets with three vanishing moments. We consider the sample's surface Γ to be the unit square and apply the external field $\mathbf{H}_{\text{ext}} := [0.2, 0.2]^T$. The left plot of Fig. 13 shows the mesh that is produced by the adaptive algorithm. The approximate magnetic charge σ_ε, seen in the left plot of Fig. 14, exhibits the well known characteristic singularities near the edges and corners [83, 84].

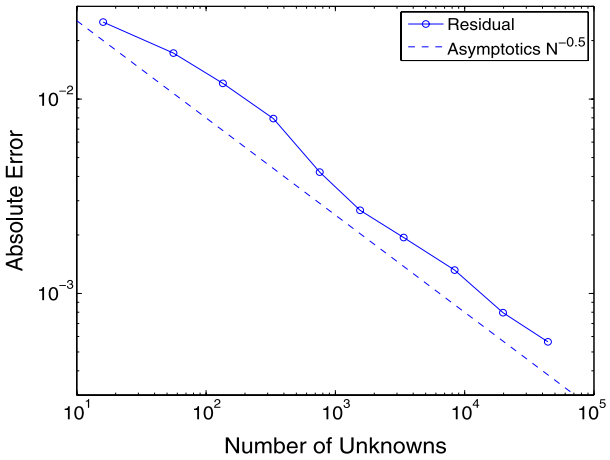


Fig. 14 Accuracy versus the degrees of freedom

In Fig. 14 we plotted the energy norm of the residuum versus the degrees of freedom. The curve validates that the discretization error behaves like $N^{-0.5}$. This is in fact the best possible rate in the presence of the edge singularities which are of *anisotropic* nature. Notice that the best possible rate of convergence, offered by the present set-up, would be $N^{-0.75}$ in case of isotropic singularities.

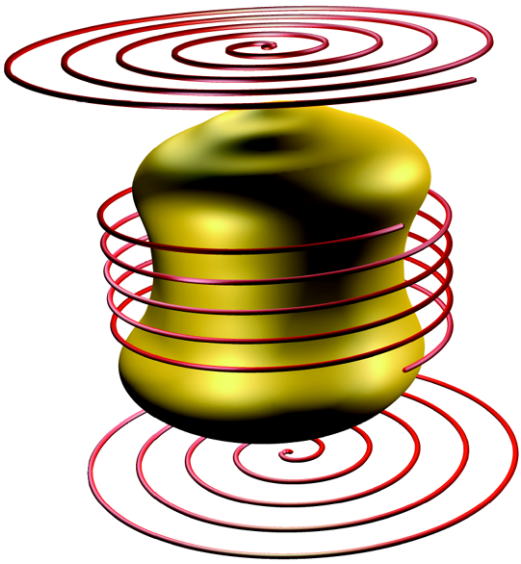


Fig. 15 A droplet of liquid metal levitating in an electric field

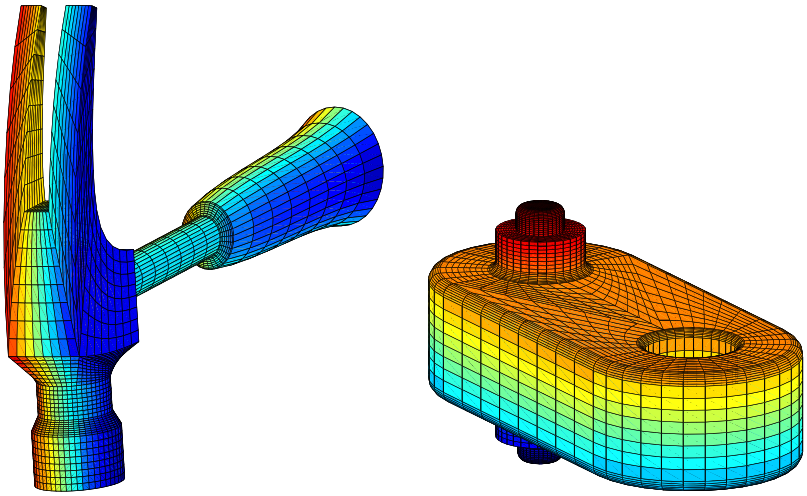


Fig. 16 Automatically produced parameterizations of a hammer and a link

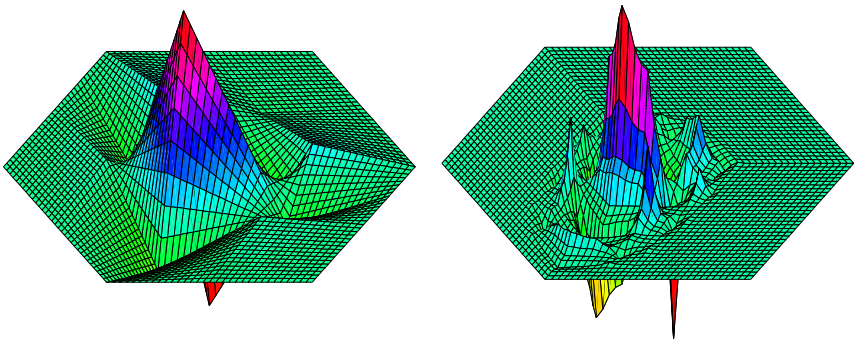


Fig. 17 A globally continuous wavelet $\psi^{(2,2)}$ located on an edge (left) and its corresponding dual (right)

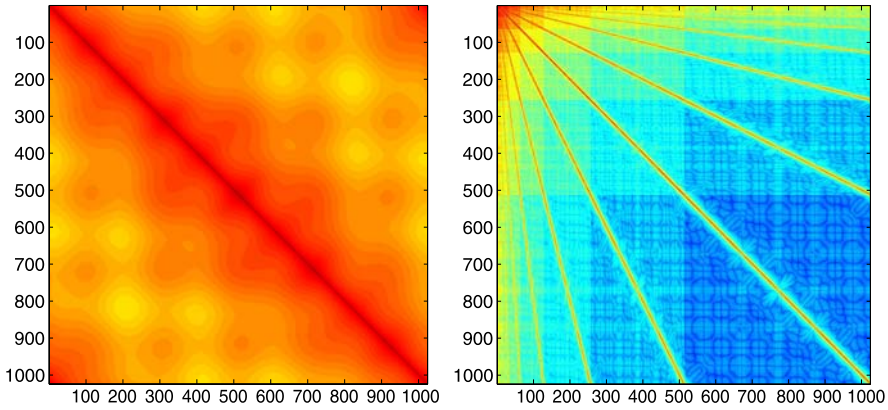


Fig. 18 The system matrix with respect to the single-scale basis (left) and the wavelet basis (right)

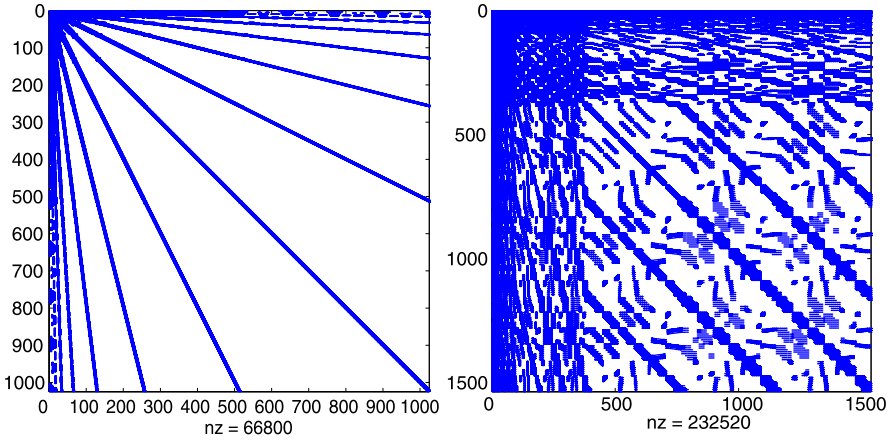


Fig. 19 The finger structure of the compressed system matrix with respect to a circle (left) and a sphere (right)

References

1. A. Barinka, W. Dahmen, and R. Schneider. Fast Computation of Adaptive Wavelet Expansions. *Numer. Math.*, 105(4):549–589, 2007.
2. M. Bebendorf. Approximation of Boundary Element Matrices. *Numer. Math.*, 86:565–589, 2000.
3. M. Bebendorf and S. Rjasanow. Adaptive low-rank approximation of collocation matrices. *Computing*, 70:1–24, 2003.
4. G. Beylkin, R. Coifman, and V. Rokhlin. The fast wavelet transform and numerical algorithms. *Comm. Pure and Appl. Math.*, 44:141–183, 1991.
5. J. Bramble, J. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55:1–22, 1990.
6. C. Canuto, A. Tabacco, and K. Urban. The wavelet element method, part I: Construction and analysis. *Appl. Comput. Harm. Anal.*, 6:1–52, 1999.
7. J. Carnicer, W. Dahmen, and J. Peña. Local decompositions of refinable spaces. *Appl. Comp. Harm. Anal.*, 3:127–153, 1996.
8. C. Carstensen and D. Praetorius. Averaging Techniques for the Effective Numerical Solution of Symm's Integral Equation of the First Kind. *SIAM J. Sci. Comput.*, 27:1226–1260, 2006.
9. C. Carstensen and E.P. Stephan. Adaptive boundary element methods for some first kind integral equations. *SIAM J. Numer. Anal.*, 33:2166–2183, 1996.
10. Z. Chen, C.A. Micchelli, and Y. Xu. A construction of interpolating wavelets on invariant sets. *Math. Comp.*, 68:1569–1587, 1999.
11. Z. Chen, C.A. Micchelli, and Y. Xu. Fast collocation methods for second kind integral equations. *SIAM J. Numer. Anal.*, 40:344–375, 2002.
12. A. Cohen. *Numerical Analysis of Wavelet Methods*. Studies in Mathematics and Its Applications 32, North-Holland Publishing Co., Amsterdam, 2003.
13. A. Cohen, W. Dahmen, and R. DeVore. Adaptive Wavelet Methods for Elliptic Operator Equations – Convergence Rates. *Math. Comp.*, 70:27–75, 2001.
14. A. Cohen, W. Dahmen, and R. DeVore. Adaptive Wavelet Methods II – Beyond the Elliptic Case. *Found. Comput. Math.*, 2:203–245, 2002.
15. A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet schemes for nonlinear variational problems. *SIAM J. Numer. Anal.*, 41:1785–1823, 2003.
16. A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Pure Appl. Math.*, 45:485–560, 1992.
17. A. Cohen and R. Masson. Wavelet adaptive method for second order elliptic problems – boundary conditions and domain decomposition. *Numer. Math.*, 86:193–238, 2000.
18. R. Coifman, G. David, Y. Meyer, and S. Semmes. ω-Calderón-Zygmund operators. Harmonic analysis and partial differential equations (El Escorial, 1987), 132–145, Lecture Notes in Math., 1384, Springer, Berlin, 1989.
19. R.R. Coifman and Y. Meyer. A simple proof of a theorem by G. David and J.-L. Journé on singular integral operators. *Probability theory and harmonic analysis*. (Cleveland, Ohio, 1983), 61–65, Monogr. Textbooks Pure Appl. Math., 98, Dekker, New York, 1986.
20. M. Costabel. Boundary integral operators on Lipschitz domains: Elementary results. *SIAM J. Math. Anal.*, 19:613–626, 1988.
21. W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numerica*, 6:55–228, 1997.
22. W. Dahmen, H. Harbrecht, and R. Schneider. Compression techniques for boundary integral equations – optimal complexity estimates. *SIAM J. Numer. Anal.*, 43:2251–2271, 2006.
23. W. Dahmen, H. Harbrecht, and R. Schneider. Adaptive Methods for Boundary Integral Equations – Complexity and Convergence Estimates. *Math. Comput.*, 76:1243–1274, 2007.
24. W. Dahmen, B. Kleemann, S. Prößdorf, and R. Schneider. A multiscale method for the double layer potential equation on a polyhedron. In H.P. Dikshit and C.A. Micchelli, editors, *Advances in Computational Mathematics*, pages 15–57, World Scientific Publ., Singapore, 1994.

25. W. Dahmen and A. Kunoth. Multilevel preconditioning. *Numer. Math.*, 63:315–344, 1992.
26. W. Dahmen, A. Kunoth, and K. Urban. Biorthogonal spline-wavelets on the interval – stability and moment conditions. *Appl. Comp. Harm. Anal.*, 6:259–302, 1999.
27. W. Dahmen, S. Prößdorf, and R. Schneider. Multiscale methods for pseudodifferential equations. In L.L. Schumaker and G. Webb, editors, *Wavelet Analysis and its Applications*, volume 3, pages 191–235, 1993.
28. W. Dahmen, S. Prößdorf, and R. Schneider. Wavelet approximation methods for periodic pseudodifferential equations. Part II – Matrix compression and fast solution. *Adv. Comput. Math.*, 1:259–335, 1993.
29. W. Dahmen, S. Prößdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations. Part I – Stability and convergence. *Mathematische Zeitschrift*, 215:583–620, 1994.
30. W. Dahmen, S. Prößdorf, and R. Schneider. Multiscale methods for pseudo-differential equations on smooth closed manifolds. In C.K. Chui, L. Montefusco, and L. Puccio, editors, *Proceedings of the International Conference on Wavelets: Theory, Algorithms, and Applications*, pages 385–424, 1994.
31. W. Dahmen and R. Schneider. Composite wavelet bases for operator equations. *Math. Comput.*, 68:1533–1567, 1999.
32. W. Dahmen and R. Schneider. Wavelets on manifolds I. Construction and domain decomposition. *Math. Anal.*, 31:184–230, 1999.
33. W. Dahmen, R. Schneider, and Y. Xu. Nonlinear functionals of wavelet expansions—adaptive reconstruction and fast evaluation. *Numer. Math.*, 86:49–101, 2000.
34. A. Desimone, R.V. Kohn, S. Müller, and F. Otto. A reduced theory for thin-film micromagnetics. *Comm. Pure Appl. Math.*, 55(11):1408–1460, 2002.
35. A. Desimone, R.V. Kohn, S. Müller, F. Otto, and R. Schäfer. Two-dimensional modelling of soft ferromagnetic films. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 457A:2983–2991, 2001.
36. M. Duffy. Quadrature over a pyramid or cube of integrands with a singularity at the vertex. *SIAM J. Numer. Anal.*, 19:1260–1262, 1982.
37. K. Eppler and H. Harbrecht. Second Order Shape Optimization using Wavelet BEM. *Optim. Methods Softw.*, 21:135–153, 2006.
38. K. Eppler and H. Harbrecht. Wavelet based boundary element methods in exterior electromagnetic shaping. *Engineering Analysis with Boundary Elements*, 32:645–657, 2008.
39. B. Faermann. Local a-posteriori error indicators for the Galerkin discretization of boundary integral equations. *Numer. Math.*, 79:43–76, 1998.
40. T. Gantumur, H. Harbrecht, and R. Stevenson. An Optimal Adaptive Wavelet Method for Elliptic Equations Without Coarsening of Iterands. *Math. Comput.*, 76:615–629, 2007.
41. G.N. Gatica, H. Harbrecht, and R. Schneider. Least squares methods for the coupling of FEM and BEM. *SIAM J. Numer. Anal.*, 41(5):1974–1995, 2003.
42. L. Greengard and V. Rokhlin. A fast algorithm for particle simulation. *J. Comput. Phys.*, 73:325–348, 1987.
43. M. Griebel. Multilevel algorithms considered as iterative methods on semidefinite systems. *SIAM J. Sci. Comput.*, 15:547–565, 1994.
44. M. Griebel. *Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen*. Teubner Skripten zur Numerik. B.G. Teubner, Stuttgart, 1994.
45. M. Griebel, P. Oswald, and T. Schiekofer. Sparse grids for boundary integral equations. *Numer. Math.*, 83(2):279–312, 1999.
46. W. Hackbusch. *Integral equations. Theory and numerical treatment*. International Series of Numerical Mathematics, volume 120, Birkhäuser Verlag, Basel, 1995.
47. W. Hackbusch. A sparse matrix arithmetic based on $\mathscr{H}$-matrices. Part I: Introduction to $\mathscr{H}$-matrices. *Computing*, 64:89–108, 1999.
48. W. Hackbusch and Z.P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.*, 54:463–491, 1989.
49. H. Harbrecht. Wavelet Galerkin schemes for the boundary element method in three dimensions. *PHD Thesis*, Technische Universität Chemnitz, Germany, 2001.

50. H. Harbrecht. A Newton method for Bernoulli's free boundary problem in three dimensions. *Computing*, 82:11–30, 2008.
51. H. Harbrecht and T. Hohage. Fast Methods for Three-Dimensional Inverse Obstacle Scattering. *J. Integral Equations Appl.*, 19(3):237–260, 2007.
52. H. Harbrecht and T. Hohage. A Newton method for reconstructing non star-shaped domains in electrical impedance tomography. NAM Preprint 2009-01, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen, 2009. (to appear in Inverse Problems and Imaging).
53. H. Harbrecht, U. Kähler, and R. Schneider. Wavelet Galerkin BEM on unstructured meshes. *Comput. Vis. Sci.*, 8:189–199, 2005.
54. H. Harbrecht, F. Paiva, C. Pérez, and R. Schneider. Biorthogonal wavelet approximation for the coupling of FEM-BEM. *Numer. Math.*, 92:325–356, 2002.
55. H. Harbrecht, F. Paiva, C. Pérez, and R. Schneider. Wavelet preconditioning for the coupling of FEM-BEM. *Num. Lin. Alg. Appl.*, 10:197–222, 2003.
56. H. Harbrecht and M. Randrianarivony. From Computer Aided Design to Wavelet BEM. *Bericht 07-18*, Berichtsreihe des Mathematischen Seminars, Christian-Albrechts-Universität zu Kiel, 2007. (to appear in *Comput. Vis. Sci.*).
57. H. Harbrecht and R. Schneider. Wavelet Galerkin Schemes for 2D-BEM. In J. Elschner, I. Gohberg and B. Silbermann, editors, *Problems and methods in mathematical physics*, Operator Theory: Advances and Applications, volume 121, pages 221–260, Birkhäuser Verlag, Basel, 2001.
58. H. Harbrecht and R. Schneider. Adaptive Wavelet Galerkin BEM. In Computational Fluid and Solid Mechanics 2003, vol. 2, edited by K.J. Bathe, Elsevier, 1982–1986 (2003).
59. H. Harbrecht and R. Schneider. Biorthogonal wavelet bases for the boundary element method. *Math. Nachr.*, 269–270:167–188, 2004.
60. H. Harbrecht and R. Schneider. Wavelet based fast solution of boundary integral equations. In *Abstract and Applied Analysis* (Proceedings of the International Conference in Hanoi, 2002), edited by N.M. Chuong et al., World Scientific Publishing Company, 139–162 (2004).
61. H. Harbrecht and R. Schneider. Wavelet Galerkin Schemes for Boundary Integral Equations – Implementation and Quadrature. *SIAM J. Sci. Comput.*, 27(4):1347–1370, 2006.
62. H. Harbrecht, R. Schneider, and C. Schwab. Sparse second moment analysis for elliptic problems in stochastic domains. *Numer. Math.*, 109(3):167–188, 2008.
63. H. Harbrecht and R. Stevenson. Wavelets with patchwise cancellation properties. *Math. Comp.*, 75:1871–1889, 2006.
64. N. Hoang and R. Stevenson. Finite element wavelets on manifolds. *IMA J. Numer. Math.*, 23:149–173, 2003.
65. J. Hoschek and D. Lasser. *Grundlagen der geometrischen Datenverarbeitung*. Teubner, Stuttgart, 1989.
66. S. Jaffard. Wavelet methods for fast resolution of elliptic equations. *SIAM J. Numer. Anal.*, 29:965–986, 1992.
67. U. Kähler. $\mathscr{H}^2$-wavelet Galerkin BEM and its application to the radiosity equation. *PHD Thesis*, Technische Universität Chemnitz, Germany, 2007.
68. S. Knapek and K. Foster. Integral operators on sparse grids. *SIAM J. Numer. Anal.*, 39:1794–1809, 2001/02.
69. A. Kunoth and J. Sahner. Wavelets on Manifolds: An Optimized Construction. *Math. Comp.*, 75:1319–1349, 2006.
70. C. Lage and C. Schwab. A wavelet-Galerkin boundary element method on polyhedral surfaces in $\mathbb{R}^3$. Numerical treatment of multi-scale problems (Kiel, 1997), 104–118, *Notes Numer. Fluid Mech.* vol. 70, Vieweg, Braunschweig, 1999.
71. M. Maischak, P. Mund, and E.P. Stephan. Adaptive multilevel BEM for acoustic scattering. Symposium on Advances in Computational Mechanics, Vol. 2, (Austin, TX, 1997). *Comput. Methods Appl. Mech. Engrg.*, 150:351–367, 1997.
72. Y. Meyer. *Ondelettes et Opérateurs 2: Opérateurs de Calderón-Zigmund*. Hermann, Paris, 1990.

73. P. Mund, and E.P. Stephan. An adaptive two-level method for hypersingular integral equations in R^3. Proceedings of the 1999 International Conference on Computational Techniques and Applications (Canberra). *ANZIAM J.*, 42:C1019–C1033, 2000.
74. J.-C. Nedelec. *Acoustic and Electromagnetic Equations — Integral Representations for Harmonic Problems*. Springer Verlag, 2001.
75. A. Novruzi. Contribution en Optimisation des Formes et Applications. *PHD Thesis*, Nancy, France, 1997.
76. A. Novruzi and J.-R. Roche. Newton's method in shape optimisation: a three-dimensional case. *BIT*, 40:102–120, 2000.
77. P. Oswald. *Multilevel finite element approximation. Theory and applications*. Teubner Skripten zur Numerik. B.G. Teubner, Stuttgart, 1994.
78. P. Oswald. On function spaces related to Finite Element Approximation Theory. *Z. Anal. Anwendungen*, 9:43–64, 1990.
79. T. von Petersdorff, R. Schneider. and C. Schwab Multiwavelets for second kind integral equations. *SIAM J. Num. Anal.*, 34:2212–2227, 1997.
80. T. von Petersdorff and C. Schwab. Wavelet approximation for first kind integral equations on polygons. *Numer. Math.*, 74:479–519, 1996.
81. T. von Petersdorff and C. Schwab. Fully discretized multiscale Galerkin BEM. In W. Dahmen, A. Kurdila, and P. Oswald, editors, *Multiscale wavelet methods for PDEs*, pages 287–346, Academic Press, San Diego, 1997.
82. T. von Petersdorff and C. Schwab. Sparse finite element methods for operator equations with stochastic data. *Appl. Math.*, 51:145–180, 2006.
83. T. von Petersdorff and E.P. Stephan. Decompositions in edge and corner singularities for the solution of the Dirichlet problem of the Laplacian in a polyhedron. *Math. Nachr.*, 149:71–104, 1990.
84. T. von Petersdorff and E.P. Stephan. Regularity of mixed boundary value problems in $\mathbb{R}^3$ and boundary element methods on graded meshes. *Math. Meth. Appl. Sci.*, 12:229–249, 1990.
85. M. Pierre and J.-R. Roche. Computation of free surfaces in the electromagnetic shaping of liquid metals by optimization algorithms. *Eur. J. Mech., B/Fluids*, 10:489–500, 1991.
86. A. Rathsfeld. A wavelet algorithm for the solution of a singular integral equation over a smooth two-dimensional manifold. *J. Integral Equations Appl.*, 10:445–501, 1998.
87. A. Rathsfeld and R. Schneider. On a quadrature algorithm for the piecewise linear wavelet collocation applied to boundary integral equations. *Math. Methods Appl. Sci.*, 26:937–979, 2003.
88. N. Reich. Wavelet Compression of Anisotropic Integrodifferential Operators on Sparse Tensor Product Spaces. *PHD Thesis 17661*, ETH Zurich, Switzerland, 2008.
89. N. Reich. Wavelet Compression of Integral Operators on Sparse Tensor Spaces: Construction, Consistency and Asymptotically Optimal Complexity. *Report 2008-24*, Seminar of Applied Mathematics, ETH Zurich, 2008.
90. N. Reich. Wavelet compression of anisotropic integrodifferential operators on sparse tensor product spaces. *Report 2008-26*, Seminar of Applied Mathematics, ETH Zurich, 2008.
91. V. Rokhlin. Rapid solution of integral equations of classical potential theory. *J. Comput. Phys.*, 60:187–207, 1985.
92. S. Sauter and C. Schwab. Quadrature for the hp-Galerkin BEM in $\mathbb{R}^3$. *Numer. Math.*, 78:211–258, 1997.
93. S. Sauter and C. Schwab. *Randelementmethoden. Analyse, Numerik und Implementierung schneller Algorithmen*. B.G. Teubner, Stuttgart, 2005.
94. G. Schmidlin and C. Schwab. Wavelet Galerkin BEM on unstructured meshes by aggregation. *Multiscale and multiresolution methods*, pages 359–378, Lect. Notes Comput. Sci. Eng., 20, Springer, Berlin, 2002.
95. R. Schneider. *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme*. B.G. Teubner, Stuttgart, 1998.
96. H. Schulz and O. Steinbach. A new a posteriori error estimator in adaptive direct boundary element methods. The Dirichlet problem. *Calcolo*, 37:79–96, 2000.

97. C. Schwab. Variable order composite quadrature of singular and nearly singular integrals. *Computing*, 53:173–194, 1994.
98. O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems. Finite and Boundary Elements*. Springer, New York, 2008.
99. R. Stevenson. On the compressibility of operators in wavelet coordinates. *SIAM J. Math. Anal.*, 35:1110–1132, 2004.
100. R. Stevenson. Composite wavelet bases with extended stability and cancellation properties. *SIAM J. Math. Anal.*, 45:133–162, 2007.
101. W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, 3:186–200, 1996.
102. J. Tausch. A variable order wavelet method for the sparse representation of layer potentials in the non-standard form. J. Numer. Math., 12:233–254, 2004.
103. J. Tausch and J. White. Multiscale bases for the sparse representation of boundary integral operators on complex geometries. *SIAM J. Sci. Comput.*, 24:1610–1629, 2003.
104. E.E. Tyrtyshnikov. Mosaic sceleton approximation. *Calcolo*, 33:47–57, 1996.
105. L. Villemoes. Wavelet analysis of refinement equations. *SIAM J. Math. Anal.*, 25:1433–1460, 1994.
106. W.L. Wendland. On asymptotic error analysis and underlying mathematical principles for boundary element methods. In C.A. Brebbia, editor, *Boundary Element Techniques in Computer Aided Engineering, NATO ASI Series E-84*, pages 417–436, Martinus Nijhoff Publ., Dordrecht-Boston-Lancaster, 1984.
107. J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.* 34:581–613, 1992.

Learning out of leaders

Gérard Kerkyacharian, Mathilde Mougeot, Dominique Picard and Karine Tribouley

Abstract In this paper we investigate the problem of supervised learning. We are interested in universal procedures producing exponential bounds. One main purpose is to link this problem to a general approach on high dimensional linear models in statistics and to propose some tools resulting from a combination of inspirations: many of them coming from previous works of Wolfgang Dahmen and coauthors combined with regression and thresholding techniques. We present different types of algorithms initiated in Wolfgang Dahmen's (and coauthors) work and provide a new algorithm: the LOL procedure. We prove that it has optimal exponential rate of convergence. We also study the practical behavior of the procedure: our simulation study confirms its very good properties.

Key words: Learning Theory, non-linear methods, thresholding, wavelets.

AMS Subject Classification: 62G08.

1 Introduction

In this paper, we are interested in the problem of learning an unknown function defined on a set $\mathbb{X}$ which takes values in a set $\mathscr{Y}$. We assume that $\mathbb{X}$ is a compact domain in $\mathbb{R}^d$ and $\mathscr{Y}$ is a finite interval in $\mathbb{R}$.

One main purpose is to link this problem to a general approach on high dimensional linear models in statistics and to propose some tools resulting from a com-

Gérard Kerkyacharian, Dominique Picard
Université Paris-Diderot, CNRS LPMA, 175 rue du Chevaleret, 75013 Paris, France,
e-mail: picard@math.jussieu.fr

Mathilde Mougeot, Karine Tribouley
MODALX, Université Paris Ouest Nanterre, 200 avenue de la République, 92001 Nanterre Cedex, France

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_9,

bination of inspirations: many of them coming from previous works of Wolfgang Dahmen and coauthors combined with regression and thresholding techniques.

Let $Z_1,\dots,Z_n$ be an observed $n-$sample of $Z=(X,Y)$ whose distribution is denoted by ρ. Our aim is to recover the function f

$$f(x)=\mathbb{E}_\rho[Y|X=x].$$

We measure the error of estimation in the $L_2(\mathbb{X},\rho_X)$ norm or in the $L_2(\mathbb{X},\hat{\rho}_X)$ norm. Each of them are defined by

$$\|g\|^2_{\rho_X}=\int_{\mathbb{X}} g(x)^2 d\rho_X(x) \quad \text{and} \quad \|g\|^2_{\widehat{\rho_X}}=\frac{1}{n}\sum_{i=1}^n g(X_i)^2,$$

ρ_X being the distribution of X and $\widehat{\rho_X}$ being the empirical measure calculated on the data X_i's. Let $\hat{f}$ be an estimator of f i.e. a measurable function of $Z_1,\dots,Z_n$ taking its values in the set, say, of bounded functions. Given any $\eta>0$, the quantity

$$\rho^{\otimes n}\{\mathbf{z}:\|\hat{f}-f\|_{\rho_X}>\eta\} \tag{1}$$

measures the confidence we have that the estimator $\hat{f}$ is accurate to tolerance η.

One main goal in learning theory is to obtain results with almost no assumptions on the distribution ρ. However, it is known that it is not possible to have fast rates of convergence without assumptions and a large portion of statistics and learning theory proceeds under the condition that f belongs to a known set Θ or a family of such sets. Typical choices of Θ are compact sets determined by some smoothness condition or by some prescribed rate of decay for a specific approximation process. Another standard condition is to assume that the function f can be expressed in a dictionary using only a small number of coefficients (sparsity property).

Given our prior space Θ and the associated class $\mathbb{M}(\Theta)$ of probability measures ρ, it has been defined in DeVore et al.[12] the **accuracy confidence function of the procedure** $\hat{f}$:

$$\mathbf{AC}_n(\Theta,\hat{f},\eta):=\sup_{\rho\in\mathbb{M}(\Theta)}\rho^{\otimes n}\{\mathbf{z}:\|f-\hat{f}\|_{\rho_X}>\eta\} \tag{2}$$

for each $\eta>0$. This quantity measures a uniform confidence (over the space $\mathbb{M}(\Theta)$) that the estimator $\hat{f}$ is accurate to the tolerance η. Upper and lower bounds for **AC** have been proved in [12]. These lower bounds for **AC** are our vehicle for proving expectation lower bounds. In most examples, there exists a phase transition and a critical value η_n depending on n and Θ such that for any $\eta>\eta_n$, $\mathbf{AC}_n(\Theta,\hat{f},\eta)$ decreases exponentially. This critical value η_n (highly linked with the sparsity of Θ and often directly expressed in terms of the entropy of Θ) is essential since it yields, as a consequence, bounds of type

$$e(\Theta,\hat{f})\le C\eta_n^q \tag{3}$$

which have been extensively studied in statistics for

$$e(\Theta, \hat{f}) = \sup_{\rho \in \mathbb{M}(\Theta)} \mathbb{E}_{\rho^{\otimes n}} \|\hat{f} - f\|_{\rho_X}^q \tag{4}$$

where $q \geq 1$. For instance, when Θ is the Besov space $B_q^s(L_\infty(\mathbb{R}^d))$, $\eta_n = n^{-\frac{s}{2s+d}}$ is the minimax rate in the sense that there exist two constants $0 < c \leq C$ such that

$$cn^{-\frac{s}{2s+d}} \leq \inf_{\hat{f}} \sup_{f \in B_q^s(L_\infty(\mathbb{R}^d))} \mathbb{E}_{\rho^{\otimes n}} \|f - \hat{f}\|_{dx} \leq Cn^{-\frac{s}{2s+d}},$$

where the infimum above is taken over all possible estimate. More stringent conditions on the measure ρ are needed to prove this kind of results. Note here an important difference with the statistics framework where more often the loss function is given by $\|g\|_{dx}^2 = \int_{\mathbb{X}} g(x)^2 dx$ replacing the measure ρ_X by the Lebesgue measure. For more details see, for instance (and among many others) Ibraguimov and Hasminski [17], Stone [24], Nemirovski [23] for a slightly more restricted context than Besov spaces, and Donoho et al. [13].

Concerning upper bounds for $\mathbf{AC}_n(\Theta, \eta)$, many much properties have been established: see for instance Yang and Barron [28] in statistical context, [10], [12], Konyagyn and Temlyakov [20] in learning theory. These upper bounds are generally proved using estimation methods based on empirical mean square minimization. The estimator is obtained by the following minimization problem

$$\hat{f} = \text{Arg} \min_{f \in \mathbb{H}_n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \tag{5}$$

where $\mathbb{H}_n$ is a functional set associated to the method. These estimation rules raise two important problems. First, they generally require heavy computation times. The second serious problem lies in the fact that their construction (the choice of $\mathbb{H}_n$) is, most of the time, highly depending on the knowledge of the prior Θ. There also exist universal estimates which are generally obtained by adding a selection step to the estimation process using penalization, cross validation or aggregation (see for instance Yang and Barron [28], Temlyakov [25], Dalalyan and Tsybakov [11]). However these rules are up to now prohibitive in terms of computation time.

In Section 2, we present different types of algorithms initiated in Wolfgang Dahmen's (and coauthors) work (see [1], [5], [3],[4]). All these algorithms differ from the other universal constructions in that they rely on fast algorithms, which may be implemented by simple on-line updates. They have been a source of inspiration for our estimation method presented in Section 3, with its theoretical performances. The practical performances of the LOL procedure are investigated in Section 4. The proofs are given in section 5.

2 Various learning algorithms in Wolfgang Dahmen's work

2.1 Greedy learning algorithms

The first of these algorithms is linked with greedy algorithms and detailed in [1]. Basically, we consider a dictionary $\mathscr{D}$ included in a Hilbert space $\mathbb{H}$ and operate the following steps:

• **Preparation step for the dictionary.** We normalize the elements of the dictionary $\mathscr{D}$ for the norm $L_2(\mathbb{X}, \widehat{\rho_X})$. We truncate the dictionary $\mathscr{D}$ to $\mathscr{D}_m$ i.e. we introduce a fixed exhaustion of $\mathscr{D}$

$$\mathscr{D}_1 \subset \ldots \subset \mathscr{D}_m \subset \mathscr{D}, \quad \#(\mathscr{D}_m) = m.$$

Let $a \geq 1$ be a fixed number chosen once for all. We choose m such that $m = [n^a]$.

• **Algorithm step.** We choose a greedy algorithm among those which are detailed in the sequel. We perform this algorithm for the dictionary $\mathscr{D}_m$ to $\widetilde{Y} = \{Y_1, \ldots, Y_n\}$ considered as a function. We obtain a sequence of functions $(\hat{f}_k)_{k=0,\ldots}$ defined on $\mathbb{X}$ corresponding to the different steps of the algorithm.

• **Final estimation.** Let T be the truncation operator at the level t defined by $Tu = Sgn(u)(t \vee |u|)$. Choose the index k^* to minimize a penalized empirical least square criterion

$$k^* := \mathrm{Arg}\min_{k \in \mathbb{N}} \{ \|\widetilde{Y} - T\hat{f}_k\|^2_{\widehat{\rho_X}} + \kappa \frac{k \log n}{n} \}$$

where κ is the tuning constant of the method. Define the final estimator of f as $\hat{f} = T\hat{f}_{k^*}$.

For a complete overview on the various algorithms called 'greedy algorithms', we refer to [26]. Let us detail those used in [1]: OGA, SPA and RGA. They can be described using the following steps

$$\text{Set } \hat{f}_0 := 0. \text{ Define recursively } \hat{f}_{k+1}, \text{ based on } \hat{f}_k \text{ and } r_k := \widetilde{Y} - \hat{f}_k.$$

• In the **OGA (Orthogonal Greedy)**, a member of the dictionary is selected as

$$g_{k+1} := \mathrm{Arg} \max_{g \in \mathscr{D}_m} |\langle r_k, g \rangle_{\widehat{\rho_X}}|$$

and the recursive steps of estimation are defined by

$$\hat{f}_{k+1} := P_{k+1} Y$$

where P_{k+1} denotes the orthogonal projection on the space spanned by $\{g_1, \ldots, g_{k+1}\}$.

• In the **SPA (Stepwise Greedy)**, $\hat{f}_{k+1}$ is still selected using a projector P_V as previously but the space of projection V spanned by a subset of $\mathscr{D}_m$ is chosen so that $\|f - P_V f\|_n$ is minimum among all possible V.

- In the **RGA (Ridge Greedy)**, take $\alpha_1 = 0$ and $\alpha_k = 1 - 2/k$ for $k > 1$, and

$$(\beta_k, g_k) = \text{Arg} \min_{\beta \in \mathbb{R}, g \in \mathscr{D}_m} \| \widetilde{Y} - \alpha_k f_{k-1} - \beta g \|_{\widehat{\rho_X}}$$

Then define

$$\hat{f}_k = \alpha_k \hat{f}_{k-1} + \beta_k g_k.$$

To express the results associated to these procedures, we need to define functional spaces linked to the dictionary. For any $N > 0$, we denote

$$\|h\|_{\mathscr{L}_1(\mathscr{D}_N)} := \inf_{\alpha_g} \| \alpha_g \|_{l_1}$$

where the infimum is taken over all the vectors $\alpha_g \in \mathbb{R}^N$ such that $h = \sum_{g \in \mathscr{D}_N} \alpha_g g$. Let $r > 0$ and define the following functional space

$$\mathscr{L}_{1,r} = \left\{ f, \quad \forall N, \exists h, \ \|h\|_{\mathscr{L}_1(\mathscr{D}_N)} \leq C \text{ and } \|f - h\|_{\mathscr{L}_1(\mathscr{D})} \leq C N^{-r} \text{ for some } C > 0 \right\}$$

Then the results obtained in [1] can be summarized in the following theorem:

Theorem 2.1. *For any h belonging to the functional space spanned by $\mathscr{D}_m$, we have*

$$\forall k > 0, \quad \mathbb{E} \, \|\hat{f} - f\|^2_{\rho_X} \leq 8 \frac{1}{k} \|h\|_{\mathscr{L}_1(\mathscr{D}_m)} + 2\|f - h\|^2_{\rho_X} + C k \frac{\log n}{n}$$

as soon as $\kappa \geq \kappa_0$ where κ_0 is only depending on a and on the threshold t and the positive constant C only depends on κ, t, a.

This theorem has as a consequence that, if the function f belongs to the space $\mathscr{L}_{1,r}$ defined above with $r > 1/2a$, then there exists $C > 0$ such that

$$\mathbb{E} \, \|\hat{f} - f\|^2_{\rho_X} \leq C(1 + \|f\|_{\mathscr{L}_{1,r}}) \left(\frac{\log n}{n} \right)^{1/2}.$$

Although no exponential bounds are produced, refined moments such as (3) are established.

2.2 Tree thresholding procedures

Two others methods developed in [5] and [3] are linked with an adapted tree partition. In this approach, the set of considered functions (the basic Hilbert space) is the space of piecewise constant functions or polynomials associated to partitions Λ. We describe here the procedure associated to piecewise constant functions. These partitions are chosen in a set of admissible partitions based on a tree structured split-

ting role. Among all the partitions, the procedure selects using an algorithm much lighter than would be a systematic penalized least square estimate which makes the procedure truly implementable.

The procedure follows the following steps:

• **Prepare the partition.** Here $\mathbb{X}$ is $[0,1]^d$ and we denote by $(\mathscr{D}_j)$ the collection of dyadic sub cubes of $\mathbb{X}$ of side length 2^{-j} with $\mathscr{D} = \cup_j \mathscr{D}_j$. These cubes are naturally associated with a tree $\mathscr{T} = \mathscr{T}(\mathscr{D})$ where each node of the tree is identified with a cube of $\mathscr{D}$. If $I \in \mathscr{D}_j$, its children are the 2^d cubes of $\mathscr{D}_{j+1}$ which are included in I. We denote by $\mathscr{C}(I)$ the set of children of I. Of course, each child J has I for parent: this is denoted by $I = \mathscr{P}(J)$. A proper subtree $\mathscr{T}_0$ of $\mathscr{T}$ is a collection of nodes of $\mathscr{T}$ with the following properties:

- (i) The root node $I = X$ belongs to $\mathscr{T}_0$,
- (ii) if $I \neq X$ and $I \in \mathscr{T}_0$ then its parent is also in $\mathscr{T}_0$.

Given a proper subtree $\mathscr{T}_0$, we call outer leaves of $\mathscr{T}_0$ the nodes of $J \in \mathscr{T}$ which are not in $\mathscr{T}_0$, but such that their parents are in $\mathscr{T}_0$ ($J \notin \mathscr{T}_0$, $\mathscr{P}(J) \in \mathscr{T}_0$). The collection of all outer leaves of a proper subtree $\mathscr{T}_0$ forms a partition.

This is the prototypal dyadic partition but the authors can work in a more general setting. Fix $a \geq 2$. We assume that if X is to be refined then its children consist in a partition of a subsets. Such a refinement strategy also results in a tree called the master tree. The refinement level of a node is the smallest number of refinements (starting from the root) to create this node. We denote by $\mathscr{T}_j$ the proper subtree consisting of all nodes with level $\leq j$, and by Λ_j the corresponding partition.

• **Approximation on a partition.** Given a partition Λ, we are interested into approximations of the function f with functions of the form $\sum_{I\in\Lambda} \alpha_I \chi_I$ where χ_I is the indicator function of the set χ.

The best approximation of the function f in terms of $\mathbb{L}_2(\rho_X)$ in the sense that $\|f - \sum_{I\in\Lambda} \alpha_I \chi_I\|_{\rho_X}$ is minimum as a function of the α_I's, is obtained by

$$P_\Lambda f = \sum_{I\in\Lambda} c_I \chi_I \quad \text{where} \quad c_I = \frac{\int_I f d\rho_X}{\int_I d\rho_X}.$$

The convention that the ratio above is zero whenever the denominator is zero is adopted. Notice that the observation $\widetilde{Y}$ can be viewed as a function of X setting $Y_i = "f"(X_i)$ for $i = 1,\ldots,n$ and can be approximated using the empirical measure $\widehat{\rho_X}$. The best approximation is then

$$\widehat{P_\Lambda} = \sum_{I\in\Lambda} \widehat{c_I} \chi_I \quad \text{where} \quad \widehat{c_I} = \frac{\sum_{i=1}^n Y_i \chi_I(X_i)}{\sum_{i=1}^n \chi_I(X_i)}$$

which minimizes the quantity

$$\|f - \sum_{I\in\Lambda} \alpha_I \chi_I\|^2_{\widehat{\rho_X}} = \frac{1}{n} \sum_{i=1}^n (Y_i - \sum_{I\in\Lambda} \alpha_I \chi_I(X_i))^2.$$

Of course, only the last approximation enters an estimation algorithm. The first one is used to measure the performances of the algorithm.

• **Choose adaptively the partition.** The algorithm is based on a stepwise refinement of the partition with a decision of subdividing or not at each node of the tree. More precisely for each node, we compute

$$\varepsilon_I(Z)^2 = \|\widetilde{Y} - \widehat{c_I}\|^2_{\mathbb{L}_2(I,\widehat{\rho_X})} - \sum_{J\in C(I)} \|\widetilde{Y} - \widehat{c_J}\|^2_{\mathbb{L}_2(J,\widehat{\rho_X})} = \sum_{J\in C(I)} \hat{\rho}_X(J)\,[\widehat{c_J} - \widehat{c_I}]^2 \quad (6)$$

The following threshold is chosen

$$\tau_n = \kappa\sqrt{\frac{\log n}{n}},$$

where κ is a tuning constant of the procedure. We fix $\gamma > 0$ and define j_0 as the largest integer j such that $a^j \le \tau_n^{-1/\gamma}$ and consider the smallest tree $\mathscr{T}(Z,n)$ which contains the set of nodes

$$\Sigma(Z,n) := \{I \in \cup_{j\le j_0}\Lambda_j\,;\, \varepsilon_I(Z) \ge \tau_n\}.$$

To this tree is associated the partition $\Lambda(Z,n)$ and the final estimator is

$$\widehat{f} = \widehat{P_{\Lambda(Z,n)}} = \sum_{I\in\Lambda(Z,n)} \widehat{c_I}\,\chi_I. \quad (7)$$

To state the results, let us define some functional spaces naturally associated with the procedure. Recall that $\mathscr{T}_j$ is the proper subtree consisting of all nodes with level smaller than j and Λ_j is the corresponding partition. We measure the approximation error by $\|f - \sum_{I\in\Lambda_j} c_I\chi_I\|_{\rho_X}$ and denote by $\mathscr{A}^s(M)$ for some $M > 0$ the ball

$$\mathscr{A}^s(M) = \left\{ f \in \mathbb{L}_2(\rho_X), \quad \|f\|_{\mathscr{A}^s} := \sup_{j=0,1,\dots} a^{-js}\|f - \sum_{I\in\Lambda_j} c_I\chi_I\|_{\rho_X} \le M, \right\}.$$

Let us also define in the same way as in (6)

$$\varepsilon_I^2 = \|f - \widehat{c_I}\|^2_{\mathbb{L}_2(I,\widehat{\rho_X})} - \sum_{J\in C(I)} \|f - \widehat{c_J}\|^2_{\mathbb{L}_2(J,\widehat{\rho_X})} = \sum_{J\in C(I)} \rho_X(J)\,[\widehat{c_J} - \widehat{c_I}]^2$$

and consider the smallest tree $\mathscr{T}(\eta)$ which contains the set of nodes

$$\Sigma(\eta) := \{I \in \cup_{j\le j_0}\Lambda_j\,;\, \varepsilon_I \ge \eta\}.$$

Corresponding to this tree, we have the partition $\Lambda(f,\eta)$ consisting of the outer leaves of $\mathscr{T}(f,\eta)$. Let $s > 0$, we denote $\mathscr{B}^s$ the set

$$\mathscr{B}^s = \left\{ f \in \mathbb{L}_2(\rho_X), \quad \|f\|^p_{\mathscr{B}^s} := \sup_{\eta>0} \eta^p \#(\mathscr{T}(f,\eta)) < \infty \right\}.$$

where $p = (s+1/2)^{-1}$. It is possible to prove that there exists a constant C depending on s such that (see [9])

$$\|f - P_{\Lambda(\eta)} f\|_{\rho_X} \leq C \|f\|_{\mathscr{B}^s} \eta^{\frac{2s}{1+2s}} \leq C\|f\|_{\mathscr{B}^s} (\#(\mathscr{T}(f,\eta)))^{-s}.$$

It is easy to see that $\mathscr{A}^s \subset \mathscr{B}^s$. The distinction between the two forms of approximation is that in the first one the partitions are fixed in advance regardless of f whereas in the second form, the partition can adapt to f.

The results of the procedure are summarized in the following theorems [5].

Theorem 2.2. *Let $s > 0$, choose*

$$j^* = \inf \left\{ j \in \mathbb{N},\ a^{j(1+2s)} \geq \frac{m}{\log m} \right\}$$

where m is the size of the dictionary. Define $\widehat{f} = \widehat{P_{\Lambda^*(Z,n)}}$ where $\Lambda^*(Z,n)$ is the partition associated to the tree

$$\Sigma^*(Z,n) = \{ I \in \cup_{j \leq j^*} \Lambda_j \,;\, \varepsilon_I(Z) \geq \tau_n \}.$$

If $f \in \mathscr{A}^s(M)$ for $M > 0$ then, for any $\alpha > 0$, there exists a constant $\widetilde{c} := \widetilde{c}(M,\alpha,a)$ such that

$$\rho^{\otimes n} \left\{ \|f - \widehat{f}\|^2_{\rho_X} > (\widetilde{c} + \|f\|_{\mathscr{A}^s}) \left(\frac{m}{\log m} \right)^{-\frac{2s}{1+2s}} \right\} \leq C m^{-\alpha}$$

and

$$E\|f - \widehat{f}\|^2_{\rho_X} \leq (C + \|f\|_{\mathscr{A}^s}) \left(\frac{m}{\log m} \right)^{-\frac{2s}{1+2s}}$$

where C depends only on a and M.

If we now turn to the adaptive procedure described above (7), it is proved in [5] the following result:

Theorem 2.3. *Let $\alpha, \gamma > 0$. Let $\widehat{f}$ be the adaptive estimator described in (7) for some $\kappa \geq \kappa_0$ where κ_0 is only depending on γ. Let $s > 0, M > 0$ and assume that $f \in \mathscr{A}^\gamma(M) \cap \mathscr{B}^s$. Then there exists a constant $\widetilde{c} > 0$ such that*

$$\rho^{\otimes n} \left\{ \|f - \widehat{f}\|^2_{\rho_X} > \widetilde{c} \left(\frac{m}{\log m} \right)^{-\frac{2s}{1+2s}} \right\} \leq C m^{-\alpha}$$

and

$$E\|f - \widehat{f}\|^2_{\rho_X} \leq (C + \|f\|_{\mathscr{A}^s}) \left(\frac{m}{\log m} \right)^{-\frac{2s}{1+2s}}$$

where C depends only on a and M.

If we relate these results to our introduction, the conclusion is that in a non adaptive (for the first one) and adaptive way, the authors obtain nearly exponential bounds if η is of the order of the critical value η_n described in the introduction.

3 Learning out leaders: LOL

3.1 Gaussian regression model

We assume the following model which is a bit more restrictive than the 'learning' model described above. We still observe $Z_1,\dots,Z_n$, for $Z_i = (X_i, Y_i)$ with now

$$Y_i = f(X_i) + \varepsilon_i,\ i = 1 \dots n$$

where f is the unknown function to be estimated and

1. the $X_i's$ are i.i.d. random variables on $\mathbb{X}$ which is a compact domain of $\mathbb{R}^d$. Let ρ be the unknown common law of the vector $Z = (X,Y)$.
2. the $\varepsilon_i's$ are i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$ random variables independent of the $X_i's$ for some unknown positive constant σ^2.

In addition, we assume that f is bounded. In view to estimate f, we consider a dictionary $\mathscr{D}$ of size $\#\mathscr{D} = p$ included in a Hilbert space $\mathbb{H}$ as in the previous sections

$$\mathscr{D} = \{g_1, \dots g_p\}.$$

We norm the g_ℓ's in the dictionary with respect to the empirical measure such a way that

$$\forall \ell = 1,\dots,p, \quad \frac{1}{n}\sum_{i=1}^{n} g_\ell^2(X_i) = 1.$$

Let τ_n be the **coherence** of the dictionary, again with respect to the empirical measure

$$\tau_n = \sup_{\ell \neq \ell' = 1,\dots,p} \left| \frac{1}{n}\sum_{i=1}^{n} g_\ell(X_i) g_{\ell'}(X_i) \right|. \tag{8}$$

Let us fix once for all $\delta \in]0,1[$. δ is linked with the precision of our procedure. Let N be an integer less or equal to δ/τ_n and suppose that $N > 1$. A simple consequence of the definition of the coherence is the following: for any subset $\mathscr{D}'$ of size m of the dictionary $\mathscr{D}$, we defined the $m \times m$ Gram matrix $M(\mathscr{D}')$

$$M(\mathscr{D}') = \left(\frac{1}{n}\sum_{i=1}^{n} g_\ell(X_i) g_{\ell'}(X_i) \right)_{g_\ell, g_{\ell'} \in \mathscr{D}'}.$$

If the size $m = \#\mathscr{D}'$ of $\mathscr{D}'$ verifies $m \leq N$ then $M(\mathscr{D}')$ is almost diagonal, in the sense that: **(Restricted Isometry Property)**

$$\forall x \in \mathbb{R}^m,\ \|x\|^2_{l_2(m)}(1-\delta) \leq x^t M(\mathscr{D}')x \leq \|x\|^2_{l_2(m)}(1+\delta).$$

or equivalently

$$\forall x \in \mathbb{R}^m,\ \|x\|^2_{l_2(m)}(1+\delta)^{-1} \leq x^t M(\mathscr{D}')^{-1}x \leq \|x\|^2_{l_2(m)}(1-\delta)^{-1}. \tag{9}$$

This is due to the fact that

$$|x^t M(\mathscr{D}')x - x^t x| \leq \tau_n \sum_{k\neq\ell=1}^{m} |x_k\, x_\ell\,| \leq N\tau_n \|x\|^2_{l_2(m)}$$

and this proves in particular that the matrix $M(\mathscr{D}')$ is invertible.

3.2 LOL procedure

Once τ_n (or a bound for τ_n) is evaluated and N is available, this procedure has three steps: Find N 'leaders', Regress on the leaders, Threshold.

1. Find the leaders:
   • For some constant $T_1 > 0$, fix a threshold

   $$\lambda_n(1) = T_1 \left(\frac{\log p}{n}\right)^{1/2}$$

   • Compute the correlations

   $$K_\ell = |\frac{1}{n}\sum_{i=1}^{n} g_\ell(X_i)Y_i|$$

   and consider the ordered truncated sequence $K_{(1)} \geq K_{(2)} \geq \ldots \geq K_{(N)}$, and the associated set of indices $\mathscr{K} = \{\kappa_{(1)}, \kappa_{(2)}, \ldots, \kappa_{(N)}\}$.
   • The final set of the leaders is defined by

   $$B = \{g_\ell, \ell \in \mathscr{K} \text{ and } K_\ell \geq \lambda_n(1)\}$$

   and we denote $\mathscr{B}$ the set of their indices (which might be different of $\mathscr{K}$). It is clear that by construction, N appears as a bound for the number of leaders i.e. the cardinal of $\mathscr{B}$.
2. Regress on the leaders
   • Consider the pseudo-regression model:

   $$Y_i = \sum_{\ell\in\mathscr{B}} \alpha_\ell g_\ell(X_i) + \varepsilon_i$$

and define the matrix $G_{\mathscr{B}}$ by

$$(G_{\mathscr{B}})_{\ell,i} = g_\ell(X_i) \quad \text{for any } \ell \in \mathscr{B} \text{ and } i \in \{1,\dots,n\}.$$

• Let $\widehat{\alpha} = (\widehat{\alpha}_\ell,\ \ell \in \mathscr{B})$ be the minimum least square error in this model:

$$\hat{\alpha} = \mathrm{Arg}\min_{\alpha=(\alpha_\ell)_{\ell\in\mathscr{B}}} \left(\sum_{i=1}^{n} (Y_i - \sum_{\ell\in\mathscr{B}} \alpha_\ell g_\ell(X_i))^2 \right) = (G_{\mathscr{B}} G_{\mathscr{B}}^t)^{-1} G_{\mathscr{B}} (Y_1,\dots,Y_n)^t .$$

3. For some constant $T_2 > 0$, fix a threshold

$$\lambda_n(2) = T_2 \left(\frac{\log n}{n} \right)^{1/2}$$

and threshold the estimated coefficients

$$\widetilde{\alpha}_\ell = \widehat{\alpha}_\ell \, I\{|\widehat{\alpha}_\ell| \geq \lambda_n(2)\}.$$

Define

$$\hat{f}(x) := \sum_{\ell\in\mathscr{B}} \widetilde{\alpha}_\ell \, g_\ell(x).$$

3.3 Sparsity conditions on the target function f

We assume the following sparsity conditions on the function f. There exist $S \leq N$, a sequence $(\alpha_\ell)_{\ell=1,\dots,p}$ and constants M, c_t, c_0 such that f can be written

$$f = \sum_{\ell=1}^{p} \alpha_\ell \, g_\ell + h$$

with

$$\sum_{\ell=1}^{p} |\alpha_\ell| \leq M, \tag{10}$$

$$\|h\|_{\hat{\rho}_X}^2 \leq c_1 \frac{S}{n} \tag{11}$$

$$\#\{\ell \in \{1,\dots,p\},\ |\alpha_\ell| \geq \lambda_n(2)/2\} \ \leq S \tag{12}$$

$$\sum_{(\ell)>N} |\alpha_{(\ell)}| \leq c_t \sqrt{\frac{S}{n}} \tag{13}$$

$$\sum_{\ell=1}^{p} |\alpha_\ell|^2 \, I\{|\alpha_\ell| \leq 2\lambda_n(2)\} \leq c_0 \frac{S}{n} \tag{14}$$

Recall that $(\alpha_{(\ell)})$ is the ordered sequence $|\alpha_{(1)}| \geq |\alpha_{(2)}| \geq \ldots |\alpha_{(p)}|$. For $S, M > 0$, we denote $V(S,M)$ the space of functions f satisfying the sparsity conditions (10), (11), (12), (13), (14). An example of such a space occurs when we suppose that all the coefficients of α are 0 but S coefficients (with $S \leq N$) which are of modulus greater that $\lambda_n(2)/2$.

Another type of assumptions ensuring the conditions above are the following: f can be written as

$$f = \sum_{\ell=1}^{p} \alpha_\ell \, g_\ell$$

with

$$\forall \lambda > 0, \, \#\{|\alpha_\ell| \geq \lambda\} \leq c\lambda^{-\frac{2}{1+2s}} \tag{15}$$

and

$$\forall k \geq 0, \, \sum_{\ell > k} |\alpha_{(\ell)}| \leq c' k^{-\tau} \tag{16}$$

for constants $b, c, c' > 0$ and $0 < b < 1$ and

$$N = n^b, \quad S = n^{\frac{1}{1+2s}}, \quad \tau \geq \frac{s}{b(1+2s)}$$

As discussed in [18], Condition (16) reflects a 'minimal compacity condition' which does not really interfere in the entropy calculations of the set (for instance) neither in the minimax rates of convergence. Condition (15) does drive the rates. It is given here with a Lorentz type constraint on the α_ℓ's. These conditions are obviously implied if the sequence α belongs to l_r for $r = \frac{2}{1+2s}$ which then looks very much like Besov-type conditions.

3.4 Results

The performances of the LOL procedure are summarized in the following theorem.

Theorem 3.1. *Let $S, M > 0$ and fix δ in $]0,1[$. Choose a dictionary $\mathscr{D}$ such that $\#(\mathscr{D}) \leq n^a$ for some $a > 0$. Choose the constants appearing in the thresholds such that*

$$T_1 \geq O(\delta, a, M) \text{ and } T_2 \geq O(\delta).$$

If the coherence satisfies $\tau_n \leq c\sqrt{\frac{\log n}{n}}$ then there exist positive constants D, c and γ, such that

$$\sup_{\rho} \sup_{f \in V(S,M)} \rho^{\otimes}\{\|f - \hat{f}\|_{\widehat{\rho X}} > \eta\} \leq \begin{cases} c\exp(-\gamma n \eta^2) & \text{for } \eta \geq D\eta_n\sqrt{\log n}, \\ 1 & \text{for } \eta \leq D\eta_n\sqrt{\log n} \end{cases} \tag{17}$$

where

$$\eta_n^2 = \frac{S}{n}.$$

As mentioned in the introduction, these results prove that the behavior of the LOL estimator is optimal in terms of the critical value η_n as predicted in [12]. It is also optimal in terms of exponential rates. An elementary consequence of Theorem 3.1 is

Corollary 3.1. *Let $q > 0$. Under the same assumptions as in Theorem 3.1, we get*

$$\mathbb{E}\|f - \hat{f}\|^q_{\widehat{\rho_X}} \leq D' \eta_n^q [\log n]^{\frac{q}{2}}$$

for some positive constant D'.

3.5 Discussion

As mentioned in the previous section, our procedure finds its inspiration especially in [1], [5], [3],[4]. In all these papers, the results are obtained under fewer assumptions but with no exponential bounds and a cost in implementation a little higher.

Temlyakov[25] provides optimal critical value η_n as well as exponential bounds with fewer assumptions: there is no coherence restriction and the setting is not the gaussian regression framework but the learning. Our gain is in the simplicity of implementation of our procedure.

Fan and Lv [16] also provide a search among leaders. Their results are basically concerning a linear model in ultra high dimension, with assumptions on the matrix of the model making it difficult to directly apply to a dictionary.

Our procedure also has to be compared with the various procedures affiliated to the Lasso or Dantzig selectors: see among many others Tibshirani [27], Candes and Tao [8], Bickel et al. [2], Bunea et al. [6][7]. The assumptions thereby are also expressed in terms of Gaussian linear model in high dimension. The restrictions of coherence types are depending on the papers and may be lighter than ours. The great advantage of our procedure LOL is to produce very simple algorithm. Moreover our assumptions are quite elementary and leads to optimal exponential rates. Nevertheless, our condition on τ_n is probably too severe. The practical results do not seem to reflect such a strict condition as it is shown via the simulation study in the following section. Last, let us emphasize that it would be interesting to obtain results also with the theoretical norm $\|.\|_\rho$ as well as the empirical norm. Passing from one norm to the other is reasonably simple when the estimators are bounded. However, here bounding the estimators seriously changes their nature at least theoretically.

3.6 Restricted LOL

Notice that the LOL procedure has two thresholding steps. In view to understand how the two steps are working separately, we isolate the second one which seems to be the more critical one. For that, let us suppose that an oracle is selecting N leaders. The first step of the algorithm becomes useless, but we can still perform the second and third step of the procedure: projection on the leaders and thresholding. We call this procedure **restricted LOL**. Observe that this generally does not provide a procedure since it depends on the oracle. However, in the specific case where the cardinal of the dictionary is small enough - less than N-, it is indeed a procedure. So studying this restricted LOL procedure has also an interest per se. A very interesting consequence of the following Theorem 3.2 is that we need much less conditions for restricted LOL. For instance, we do not need the restrictive condition on the coherence $\tau_n \leq c\sqrt{\frac{\log n}{n}}$ and the sparsity conditions become wider: Assumption (10) is useless and in the case where the size of the dictionary is less than N, (13) is not necessary. Another interesting remark is that the restricted LOL procedure has the same optimal exponential bounds.

Theorem 3.2. *Let $N > 1$ and assume that the dictionary $\mathscr{D}$ satisfies $\#(\mathscr{D}) \leq n^a$ for some $a > 0$. Choose the thresholding constant such that $T_2 \geq O(\delta)$. Let $S, M > 0$ and assume that $f \in V(S,M)$. Suppose that an oracle is able to select N elements $\{g_{\ell_1}, \ldots, g_{\ell_N}\}$ of the dictionary such that the coefficients α_{ℓ_j} of f on g_{ℓ_j} for $j = 1, \ldots, N$ are the N largest among all the coefficients $\alpha_\ell, \ell = 1, \ldots, p$. Define the restricted LOL procedure as:*

$$\hat{f}'(x) := \sum_{\ell=1}^{N} \widehat{\alpha}_\ell \, I\{|\widehat{\alpha}_\ell| \geq \lambda_n(2)\} \, g_\ell(x).$$

Then, there exist positive constants c, γ, such that

$$\sup_{\rho} \sup_{f \in V(S,M)} \rho^{\otimes}\{\|f - \hat{f}'\|_{\widehat{\rho_X}} > \eta\} \leq \begin{cases} c \exp\left(-\gamma n \eta^2\right) & \text{for } \eta \geq D\eta_n, \\ 1 & \text{for } \eta \leq D\eta_n \end{cases} \tag{18}$$

for some positive constant D depending on c_t, c_0, T_2, δ and

$$\eta_n^2 = \frac{S}{n}.$$

In this paper, we give the proof of Theorem 3.2. The proof of Theorem 3.1 as well as a more detailed study of the procedure in the case of particular dictionaries as well as in general ultra high dimension linear models are postponed in the companion paper [19].

4 Practical performances of the LOL procedure

To study the performance of the LOL method, we now present an experimental design, results of simulations and a practical example of reconstruction.

4.1 Experimental design

Dictionaries: Five dictionaries ($\mathscr{D}_1, \ldots \mathscr{D}_5$) with different coherences are used to illustrate the performances of the LOL procedure. Each dictionary is composed of functions of the trigonometric or the Haar bases and is characterized by a given coherence. The two first dictionaries are homogenous, composed exclusively of the trigonometric ($\mathscr{D}_1$) or the Haar base ($\mathscr{D}_2$). Both are of equal size ($p = 512$) and are characterized by a weak coherence, empirically close to zero. The three other dictionaries are mixed, composed of an equal part of the trigonometric and the Haar bases. $\mathscr{D}_3$ is the union of the two homogeneous dictionaries: $\mathscr{D}_3 = \mathscr{D}_1 \cup \mathscr{D}_2$, $p = 1024$. $\mathscr{D}_4$ and $\mathscr{D}_5$ are respectively composed of an half ($p = 512$) or a quarter ($p = 256$) of the two homogeneous dictionaries. Figure 1 shows the empirical distribution of the absolute values correlations of the functions of the dictionaries. As the coherence is defined as the maximum value of the correlations distribution, for both homogeneous dictionaries, the coherence is almost zero and is equal to $\tau_n = 0.637$ for the mixed dictionaries.

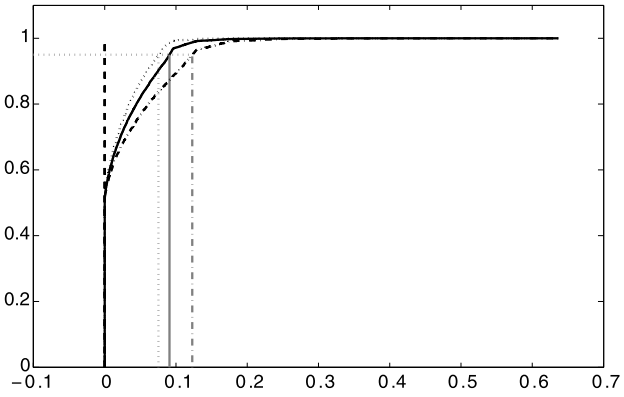


Fig. 1 Empirical distribution of the correlations (absolute value) for each dictionary. $\mathscr{D}_1$- dashed line, $\mathscr{D}_2$- dashed line ($\mathscr{D}_1$, $\mathscr{D}_2$ vertically confounded), $\mathscr{D}_3$- dotted line, $\mathscr{D}_4$- solid line, $\mathscr{D}_5$-dashed dot line. For each dictionary quantiles values are indicated by vertically gray lines using the same drawing code

As introduced in section 3, the coherence is theoretically defined as the maximum of the correlations computed between the dictionary functions. As we observe

in the previous figure, some dictionaries can have the same coherence but with various statistical distributions. Practically, we find more appropriate to characterize the internal relations between the functions of the dictionary as the value of the quantile at 95% of the set of correlations. This quantity, we called "internal coherence" better takes into account the empirical relations inside each dictionary as the maximum value does. For mixed dictionaries, the internal coherence is then equaled to $\tau_{\mathscr{D}_3} = 0.075$, $\tau_{\mathscr{D}_4} = 0.091$ and $\tau_{\mathscr{D}_5} = 0.123$.

Observations: We considered here a model with a fixed design, $X_{i,1\leq i\leq n} = i/n$. Given X_i, the observation Y_i is defined by $Y_i = f(X_i) + \varepsilon_i$, where ε_i are i.i.d with a normal distribution $N(0,\sigma_\varepsilon^2)$. Each function f, considered as unknown in our procedure, is defined as a randomly weighted sum of the functions of a dictionary $\mathscr{D}$, $f(X_i) = \sum_{l=1}^{p} \alpha_l g_l(X_i)$. We are interested in studying sparse functions with a weak number of non-zero coefficients ($\alpha_l \neq 0$). Practically, to simulate a function with S non-zero components, we draw randomly with a uniform law, S different functions of the dictionary $\mathscr{D}$, then S non-zero coefficients are randomly chosen as follows: $\alpha_l = (-1)^b|z|$ and affected to the previous selected functions g. b is drawn from a Bernouilli distribution with parameter 0.5 and z is drawn from a normal distribution $N(m,\sigma_z^2)$. Practically, we choose $m = 2$, $\sigma_z^2 = 4$ and a signal over noise ratio equaled to $\sigma_\varepsilon^2 = 5$.

Performances indicators: The practical behavior of LOL procedure is evaluate using four performances indicators computed, after each run of the algorithm, given the observations $(Y_i,X_i)_{1\leq i\leq n}$: the number of non-zero coefficients $\hat{S}$ estimated by LOL (1), the relative l_2 reconstruction error E_f (2), the relative error for the coefficients, E_α (3), a flag if the reconstruction is declared successful meaning that E_f is less than 1 percent (4):

$$\begin{aligned} \hat{S} &= \sharp\{l \in \{1\dots p\}, \widetilde{\alpha}_l \neq 0\} \\ E_f &= ||f - \widetilde{f}||_2^2/||f||^2, \\ E_\alpha &= ||\alpha - \widetilde{\alpha}||_2^2/||\alpha||^2 \end{aligned}$$

Sparsity and problem indeterminacy: For a fixed $n = 512$ number of observations, we varied the normalized measure of sparsity, $\rho = S/n$ from 0.01 to 0.15 in 29 steps. The problem indeterminacy is measured by $\delta = n/p$. In this experimental study, three cases of indeterminacy are studied: $\delta = 1.0$ ($\mathscr{D}_1$, $\mathscr{D}_2$,$\mathscr{D}_4$), $\delta = 0.5$ ($\mathscr{D}_3$) and $\delta = 2$ ($\mathscr{D}_5$).

For each value of ρ, we solved K=500 different random problem instances.

4.2 Algorithm

The LOL procedure works in tree steps: Find N 'leaders', Regress on leaders, Threshold. The selection of the leaders and of the regression coefficients depends on two thresholds $\lambda_n(1)$ and $\lambda_n(2)$ previously defined in section 3. Thresholds are critical values, very often hard to tune.

Selection of leaders: In the LOL procedure, the first threshold $\lambda_n(1)$ is used to select the candidates to the regression. Considering the set of correlations K_ℓ, the objective is to split the correlations values in two clusters to pick up the regression candidates in one group. Here, we use the property that the solution is supposed to be sparse. That means, that some functions of the dictionary are more correlated to the target Y than some others associated to a weak correlation value, close to zero. This remark implies that the distribution of correlations (in absolute value) should be distributed in two clusters: one for the leaders and one for the other functions of the dictionary. The cluster with the high correlations defines the cluster of leaders. Taking this remark into account, we adaptively compute the frontier between the clusters by minimizing the deviance of the absolute value correlations for two classes as described in [22]. For a sparse solution, when the coherence of the dictionary is weak, this method tends to automatically select the N 'leaders', with $N << n$. When the coherence of the dictionary is high, the set of selected leaders can be greater than the number of observations n and then too big to remove the indeterminacy. In order to avoid this situation, we choose $\lambda_n(1)$ in such a way that $N \leq \sqrt{nlog(n)}$.

Selection of regressors: The same procedure as described above is used to select adaptively the non zero coefficients of the model. After the regression on the N leaders, the distribution of the regression coefficients provides two clusters: one cluster associated to the largest coefficients (in absolute value) corresponding to the S non zero coefficients and one cluster composed of coefficients closed to zero, which should not be involved in the model. The frontier between the two clusters, which defines $\lambda_n(2)$, is computed by minimizing the deviance between two classes of regression coefficients. The sparse coefficients are then defined by the cluster associated with the strongest coefficients.

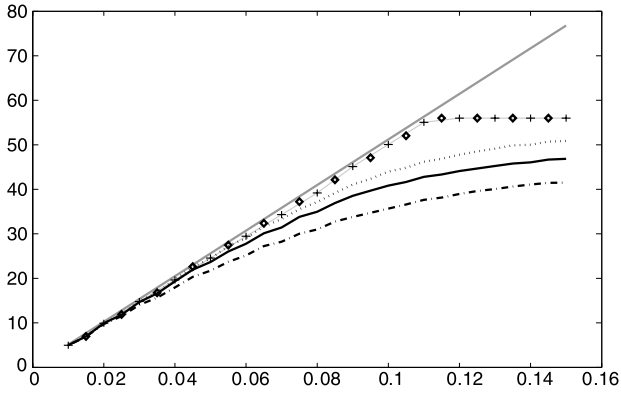


Fig. 2 Sparsity estimation using LOL for 5 studied dictionaries function of sparsity rate. $\mathscr{D}_1$- $'+'$ symbols line, $\mathscr{D}_2$-diamond symbols line ($\mathscr{D}_1$, $\mathscr{D}_2$ confounded), $\mathscr{D}_3$- dotted line, $\mathscr{D}_4$- solid line, $\mathscr{D}_5$-dashed dot line. The gray line represents the sparsity level S function of ρ

4.3 Simulation results

Figure 2 presents for the five studied dictionaries the estimation of the sparsity $\hat{S}$ function of ρ ($n = 512$). Each point of the curve is an averaged value computed over 500 instances. We observe that the results computed for the two homogeneous dictionaries $\mathscr{D}_1$ and $\mathscr{D}_2$ exclusively composed of bases are confounded. In this case, from a weak sparsity rate of $\rho = 1\%$ to a sparsity rate of $\rho = 11.5\%$, $\hat{S}$ is equaled to S. At $\rho = 11.5\%$, a sudden break is observed and the number of non-zero estimated coefficients is under estimated. This break corresponds to the phase transition already observed by Donoho et al. (see [14], [15]). Concerning the mixed dictionaries, we observe that for low sparsity levels, the results are as good as for the homogeneous dictionaries. For high sparsity levels, the number of non-zero components are more under estimated compared to the homogeneous dictionaries.

It can be point out that the internal coherence of the dictionaries has an influence on the results. The number of miss detection increases with the internal coherence and the sparsity measure. For mixed dictionaries with high internal coherence, we do not observe a sharp phase transition as for dictionaries associated with a weak coherence.

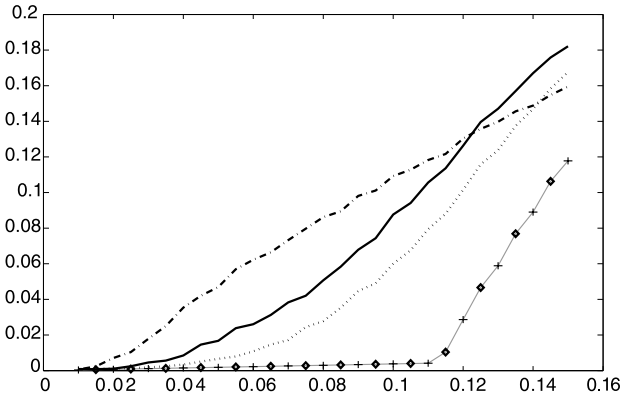


Fig. 3 Reconstruction error function of sparsity rate. $\mathscr{D}_1$-$'+'$ symbols line, $\mathscr{D}_2$-diamond symbols line ($\mathscr{D}_1$, $\mathscr{D}_2$ confounded), $\mathscr{D}_3$-dotted line, $\mathscr{D}_4$-solid line, $\mathscr{D}_5$-dashed dot line

Figure 3 and 4 show the reconstruction error and the relative error l_2 for the coefficients. We observe that for the dictionaries composed of one base, the error is almost null and increases suddenly just after the phase transition at $\rho = 11\%$.

For a fixed sparsity level ρ, the error increases as the value of the internal coherence of the dictionaries does. In the case where $\rho = 2$ with non indeterminacy in the solution, we observe that the reconstruction error curve crosses the other curves, which is not yet theoretically explained.

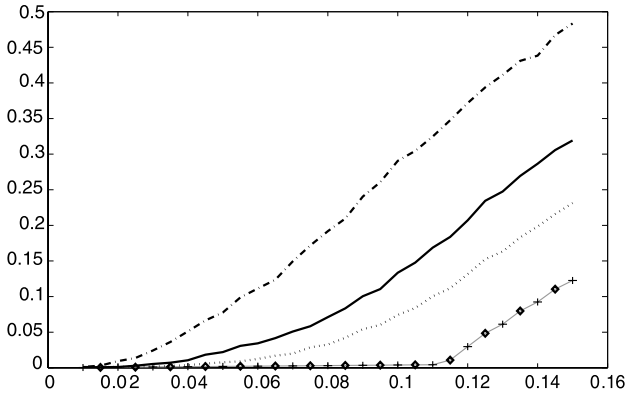


Fig. 4 Quadratic error of the coefficients function of sparsity rate. $\mathscr{D}_1$- $'+'$ symbols line, $\mathscr{D}_2$-diamond symbols line ($\mathscr{D}_1$, $\mathscr{D}_2$ confounded), $\mathscr{D}_3$-dotted line, $\mathscr{D}_4$-solid line, $\mathscr{D}_5$-dashed dot line

Figures 5 illustrates the rate of success when the reconstruction error is less than 1% as proposed by Donoho et al. (see [15]). We observe that the rate decreases as the coherence of the dictionaries increases.

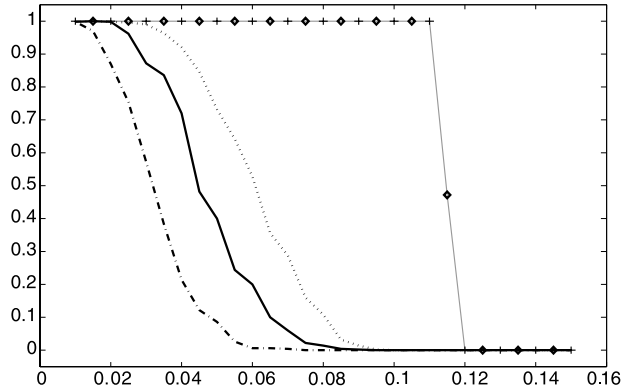


Fig. 5 Successful reconstruction rate. $\mathscr{D}_1$- $'+'$ symbols line, $\mathscr{D}_2$-diamond symbols line, $\mathscr{D}_3$ -dotted line, $\mathscr{D}_4$ - solid line, $\mathscr{D}_5$ -dashed dot line

4.4 Quality reconstruction

If we consider only the results of the simulations presented above, the advantage to use a mixed dictionary compared to an homogeneous one is not clear and the inter-

nal coherence of a dictionary seems to be a drawback rather than an added value. However, mixed dictionary can offer a complementarity which is now illustrate. Figure 6 presents a very simple example to illustrate the benefits of using a mixed dictionary. The signal is of size $p = 512$, composed of three area: two waves area at two different frequencies split by an area where the signal is constant. This signal does not belong to any previous dictionaries and is not a weighted sum of functions of any dictionary. This signal is successively compressed then restored with both homogeneous dictionaries $\mathscr{D}_1$ and $\mathscr{D}_2$ and then with the mixed dictionary $\mathscr{D}_3$.

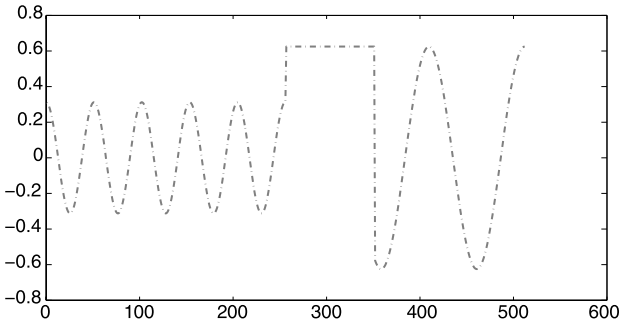


Fig. 6 Original Signal

Figure 7 shows the reconstruction of the signal, using the LOL procedure, which automatically selects in the dictionary the best functions and estimates the associated coefficients. The trigonometric dictionary is used in the first graph. The LOL procedure automatically select $k = 14$ non zero coefficients. The reconstruction error is $E_f = 14.23\%$. We observe that the reconstructed errors are at most localized on the horizontal line. The middle graph presents the results using the Haar dictionary. The LOL procedure automatically select $k = 13$ non zero coefficients which is similar to $\mathscr{D}_1$. The reconstruction error is similar to the one obtained with the trigonometric dictionary $E_f = 15.67\%$: the reconstructed errors are localized in the waves in the signal.

The last graph presents the results using the mixed dictionary. The LOL procedure automatically select $k = 17$ non zero coefficients and the reconstruction error is quite low: $E_f = 4.04\%$. The reconstructed errors seem to be spread out all over the signal and not localized as for the previous dictionaries.

4.5 Discussion

This practical study shows, that some very good results can be obtained with the LOL procedure when the sparsity level ρ is lower to 11% and for a indeterminacy level $\delta < 1$. The performances of the procedure measured by the reconstruction error

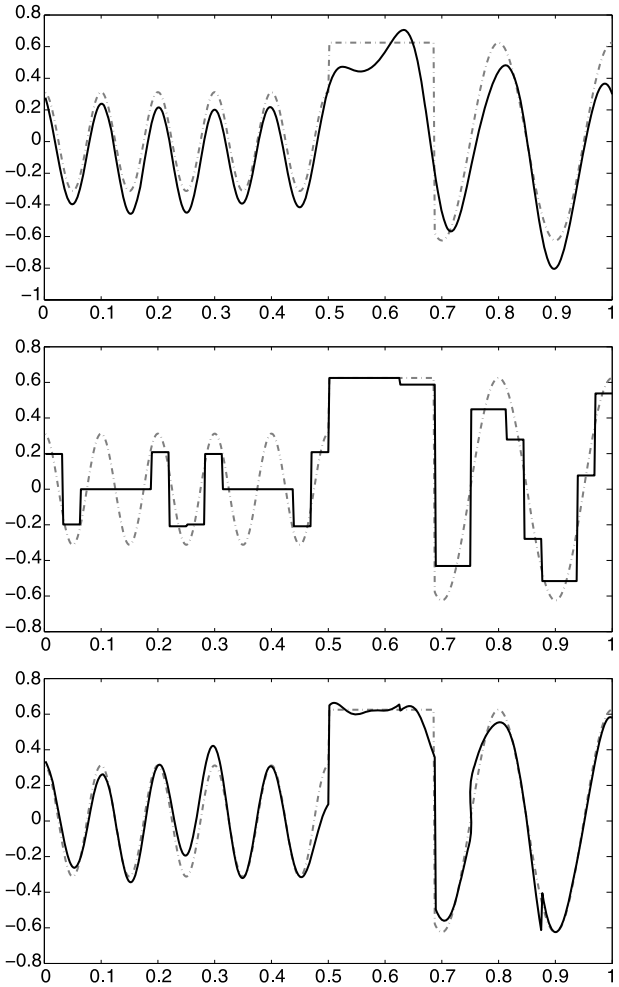


Fig. 7 Comparison of the original (gray dashed color) and the restored signal (solid black line) using various dictionaries ($\mathscr{D}_1$-upper figure; $\mathscr{D}_2$-middle figure; $\mathscr{D}_3$-lower figure)

increase as the internal coherence of the dictionary decreases to zero. However, we show that mixed dictionaries, can bring some benefits through the complementarity of their functions in the reconstruction for specific signals using the LOL procedure.

For weak sparsity levels $\rho < 11\%$, and for an indeterminacy level below $\delta < 1$, the LOL procedure shows very good performances.

5 Proofs

5.1 Preliminaries

Let us start with some notation. For any subset of indices $\mathscr{C} \subset \{1,\dots,p\}$, $V_{\mathscr{C}}$ denotes the subspace spanned by the functions $\{g_\ell, \ell \in \mathscr{C}\}$ and $P_{V_{\mathscr{C}}}$ denotes the projection (in the $\mathbb{L}_2(\hat{\rho})$ sense) over $V_{\mathscr{C}}$. For any function $\varphi = \sum_\ell \alpha_\ell g_\ell$, set $\bar{\alpha}^{\mathscr{C}}$ such that $G^t_{\mathscr{C}}\bar{\alpha}^{\mathscr{C}} := P_{V_{\mathscr{C}}}\bar{\varphi}$ where $\bar{\varphi} = (\varphi(X_1),\dots,\varphi(X_n))^t$ and define the matrix $G_{\mathscr{C}}$ by

$$((G_{\mathscr{C}})_{\ell\, i})_{\ell\in\mathscr{C}, i\in\{1,\dots,n\}} = (g_\ell(X_i))_{\ell\in\mathscr{C}, i\in\{1,\dots,n\}}\,.$$

Since the empirical norm only concerns the values of the function at the points $(X_i)_{i=1\dots,n}$, one can identify φ and $\bar{\varphi}$ (with a slight abuse of notation). Our assumptions above and standard calculations prove that

$$\bar{\alpha}^{\mathscr{C}} = (G_{\mathscr{C}}G^t_{\mathscr{C}})^{-1}G_{\mathscr{C}}\bar{\varphi}$$

if $\#(\mathscr{C}) \le N$. As well, set $G^t_{\mathscr{C}}\hat{\alpha}^{\mathscr{C}} := P_{V_{\mathscr{C}}}(\widetilde{Y}) = P_{V_{\mathscr{C}}}[\bar{f} + \varepsilon]$ where $\widetilde{Y} := (Y_1,\dots,Y_n)^t$ and $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^t$.

Recall that an oracle providing the addresses of the N largest (in modulus) coefficients α_l is available. For sake of simplicity, we suppose that their indices are $1,\dots,N$, and denote by $\bar{\alpha}$ (resp. $\hat{\alpha}$) $\bar{\alpha}^{\{1,\dots,N\}}$ (resp. $\hat{\alpha}^{\{1,\dots,N\}}$). Let us begin with the following proposition.

Proposition 5.1. *If $\mathscr{C} \subset \{1,\dots,N\}$ then*

$$\|\sum_{\ell\in C}(\widehat{\alpha}_\ell - \alpha_\ell)\, g_\ell\|_{\widehat{\rho_X}} \le \sqrt{\frac{(1+\delta)\delta}{(1-\delta)}}c_l\,\eta_n + \sqrt{\frac{1}{n(1-\delta)}}\|P_{V_C}\varepsilon\|_{l^2(n)} \tag{19}$$

Before proving the proposition we need to establish the following lemmas.

Lemma 5.1. $\forall\, x = (x_1,\dots,x_N)\ \in \mathbb{R}^N$,

$$(1-\delta)\|x\|^2_{l^2(N)} \le \|\sum_{\ell=1}^{N} x_\ell\, g_\ell\,\|^2_{\widehat{\rho_X}} \le (1+\delta)\|x\|^2_{l^2(N)}.$$

The proof of Lemma 5.1 is elementary and only relies on (9).

Lemma 5.2. *If $\#(\mathscr{C}) \le N$ then*

$$\|\sum_{\ell\in\mathscr{C}}(\alpha_l - \bar{\alpha}^{\mathscr{C}}_\ell)g_\ell\|^2_{\widehat{\rho_X}} \le \#(C)\ \tau_n^2\ \left(\sum_{\ell\in\mathscr{C}^c}|\alpha_\ell|\right)^2 \tag{20}$$

Moreover if $\mathscr{C}$ is included in $\{1,\dots,N\}$ then

$$\| \sum_{\ell \in \mathscr{C}} (\alpha_\ell - \bar{\alpha}_\ell)\, g_\ell \|^2_{\widehat{\rho_X}} \leq (1-\delta)^{-1} \sum_{\ell=1}^{N} (\alpha_\ell - \bar{\alpha}_\ell)^2 \leq c_t^2 \delta (1-\delta)^{-1} \eta_n^2 \tag{21}$$

Lemma 5.3. *If* $\#(\mathscr{C}) \leq N$, *then, for any* $u \in \mathbb{R}^n$,

$$(1+\delta)^{-1} \sum_{\ell \in \mathscr{C}} \frac{1}{n} \left(\sum_{i=1}^{n} u_i g_\ell(X_i) \right)^2 \leq \|P_{V_\mathscr{C}} u\|^2_{l^2(n)} \leq (1-\delta)^{-1} \sum_{\ell \in \mathscr{C}} \frac{1}{n} \left(\sum_{i=1}^{n} u_i g_\ell(X_i) \right)^2 \tag{22}$$

Proof. (*Lemma 5.2*) Let us denote $M(\mathscr{C})$ the Gram matrix

$$M(\mathscr{C}) = \frac{1}{n} G_\mathscr{C} G_\mathscr{C}^t.$$

$$\begin{aligned} \| \sum_{\ell \in \mathscr{C}} (\alpha_\ell - \bar{\alpha}_\ell^\mathscr{C})\, g_\ell \|^2_{\hat{\rho}} &= \| \sum_{\ell \in \mathscr{C}} \alpha_\ell g_\ell - P_{V_\mathscr{C}} \left[\sum_{\ell \in \mathscr{C}} \alpha_\ell g_\ell + \sum_{\ell \in \mathscr{C}^c} \alpha_\ell g_\ell \right] \|^2_{\hat{\rho}} \\ &= \| P_{V_\mathscr{C}} \left[\sum_{\ell \in \mathscr{C}^c} \alpha_\ell g_\ell \right] \|^2_{\hat{\rho}}. \end{aligned}$$

Denote

$$\bar{h}_C = \left(\sum_{\ell \in C^c} \alpha_\ell\, g_\ell(X_1), \ldots, \sum_{l \in C^c} \alpha_\ell\, g_\ell(X_n) \right).$$

Then

$$P_{V_C} \left[\sum_{\ell \in \mathscr{C}^c} \alpha_\ell\, g_\ell \right] = G_\mathscr{C}^t \left(n M(\mathscr{C}) \right)^{-1} G_\mathscr{C} \bar{h}_C$$

implying that

$$\begin{aligned} \| \sum_{\ell \in C} (\alpha_\ell - \bar{\alpha}_\ell^\mathscr{C})\, g_\ell \|^2_{\hat{\rho}} &= \frac{1}{n} \left\| G_\mathscr{C}^t \left(n M(\mathscr{C}) \right)^{-1} G_\mathscr{C} \bar{h}_\mathscr{C} \right\|^2_{l^2} \\ &= \frac{1}{n^2} \left(G_\mathscr{C} \bar{h}_\mathscr{C} \right)^t \left(M(\mathscr{C}) \right)^{-1} \left(G_\mathscr{C} \bar{h}_\mathscr{C} \right). \end{aligned}$$

Using Property (9), we deduce

$$\begin{aligned} \| \sum_{\ell \in \mathscr{C}} (\alpha_\ell - \bar{\alpha}_\ell^\mathscr{C})\, g_\ell \|^2_{\hat{\rho}} &\leq (1-\delta)^{-1} \frac{1}{n^2} \left(G_\mathscr{C} \bar{h}_\mathscr{C} \right)^t \left(G_\mathscr{C} \bar{h}_\mathscr{C} \right) \\ &= (1-\delta)^{-1} \frac{1}{n^2} \sum_{\ell \in \mathscr{C}} \left(\sum_{i=1}^{n} g_\ell(X_i) \sum_{\ell' \in \mathscr{C}^c} \alpha_{\ell'}\, g_{\ell'}(X_i) \right)^2 \\ &\leq (1-\delta)^{-1} \#(\mathscr{C}) \left(\tau_n \sum_{\ell \in \mathscr{C}^c} |\alpha_\ell| \right)^2 \end{aligned}$$

which gives (20). (21) is a consequence of (20), property (9) and the definition of N. □

Proof. (*Lemma 5.3*) Since

$$P_{V_{\mathscr{C}}} u = G^t_{\mathscr{C}} (nM(\mathscr{C}))^{-1} G_{\mathscr{C}} u,$$

we obtain

$$\|P_{V_C} u\|^2_{l^2(n)} = (G_{\mathscr{C}} u)^t \, (nM(\mathscr{C}))^{-1} \, (G_{\mathscr{C}} u).$$

Applying Property (9) and observing that

$$\|G_{\mathscr{C}} u\|^2_{l_2} = (G_{\mathscr{C}} u)^t \, (G_{\mathscr{C}} u) = \sum_{\ell \in \mathscr{C}} \left(\sum_{i=1}^{n} u_i g_\ell(X_i) \right)^2,$$

we obtain the announced result. □

Proof. (*Proposition 5.1*) Notice that

$$\| \sum_{\ell \in \mathscr{C}} (\hat{\alpha}_\ell - \alpha_\ell) \, g_\ell \|_{\hat{\rho}} \leq \| \sum_{\ell \in \mathscr{C}} (\hat{\alpha}_\ell - \bar{\alpha}_\ell) \, g_\ell \|_{\hat{\rho}} + \| \sum_{\ell \in \mathscr{C}} (\bar{\alpha}_\ell - \alpha_\ell) \, g_\ell \|_{\hat{\rho}}.$$

We bound the second term using Lemma 5.2. For the first one, we have

$$\sum_{\ell \in \{1,\ldots,N\}} \widehat{\alpha}_\ell \, g_\ell = \sum_{\ell \in \{1,\ldots,N\}} \bar{\alpha}_\ell \, g_\ell + P_{V_N} \varepsilon.$$

Hence

$$\sum_{\ell \in \mathscr{C}} (\widehat{\alpha}_\ell - \bar{\alpha}_\ell) g_\ell = \sum_{\ell \in \mathscr{C}} [P_{V_N} \varepsilon]_\ell \, g_\ell.$$

We finish the proof using Lemma 5.3. □

5.2 *Concentration lemma 5.4*

The following lemma will give the concentration inequality used in the sequel to obtain exponential bounds.

Lemma 5.4. *Let U be a χ^2_k variable. Then*

$$\forall u^2 \geq 4\frac{k}{n}, \quad P(U \geq nu^2) \leq \exp\left(-nu^2/8\right).$$

Proof. Recall the following result by [21]. If X_t is be a centered gaussian process such that $\sigma^2 := \sup_t \mathbb{E} X_t^2$, then

$$\forall y > 0, \quad P\left(\sup_t X_t - \mathbb{E}\sup_t X_t \geq y\right) \leq \exp -\frac{y^2}{2\sigma^2}. \tag{23}$$

Let $Z_1, \ldots, Z_k$ i.i.d. standard Gaussian variables such that

$$\begin{aligned} P(U \geq nu^2) &= P(\sum_{i=1}^k Z_i^2 \geq nu^2) = P(\sup_{a\in S_1} \sum_{i=1}^k a_i Z_i \geq (nu^2)^{1/2}) \\ &= P\left(\sup_{a\in S_1} \sum_{i=1}^k a_i Z_i - \mathbb{E}\sup_{a\in S_1} \sum_{i=1}^k a_i Z_i \geq (nu^2)^{1/2} - \mathbb{E}\sup_{a\in S_1} \sum_{i=1}^k a_i Z_i\right) \end{aligned}$$

where $S_1 = \{a \in R^k, \|a_i\|_{l^2(k)} = 1\}$. Denote

$$X_a = \sum_{i=1}^k a_i Z_i \quad \text{and} \quad y = (nu^2)^{1/2} - \mathbb{E}\sup_{a\in S_1} \sum_{i=1}^k a_i Z_i.$$

Notice that

$$a \in S_1 \Rightarrow \mathbb{E}\,(X_a)^2 = 1$$

as well as

$$\mathbb{E}\sup_{a\in S_1} X_a = \mathbb{E}\left[\sum_{i=1}^k Z_i^2\right]^{1/2} \leq \left[\mathbb{E}\sum_{i=1}^k Z_i^2\right]^{1/2} = k^{1/2}.$$

Since $u^2 \geq 4\frac{k}{n}$, the announced result is proved as soon as $y > (nu^2)^{1/2}/2$. □

5.3 *Proof of Theorem 3.2*

Since $f = \sum_{\ell=1}^p \alpha_\ell\, g_\ell + h$, we get

$$\begin{aligned} \|f - \hat{f}\|_{\widehat{\rho_X}} &\leq \|f - \sum_{\ell=1}^N \alpha_\ell\, g_\ell\|_{\hat{\rho}} + \|\sum_{\ell=1}^N \alpha_\ell\, g_\ell - \hat{f}\|_{\widehat{\rho_X}} \\ &\leq \|\sum_{\ell=N}^p \alpha_\ell\, g_\ell\|_{\hat{\rho}} + \|h\|_{\hat{\rho}} + \|\sum_{\ell=1}^N (\alpha_\ell - \widetilde{\alpha}_\ell)\; g_\ell\|_{\widehat{\rho_X}}. \end{aligned}$$

Using Hypothesis (11), we get $\|h\|_{\hat{\rho}} \leq \sqrt{c_1 \frac{S}{n}}$. Using the oracle property and Condition (13), we get

$$\|\sum_{l=N}^p \alpha_l g_l\|_{\hat{\rho}} \leq \sum_{l\geq N} |\alpha_{(l)}| \leq c_t \sqrt{\frac{S}{n}}.$$

In the case where the dictionary is such that $p \leq N$ (where we do not need the oracle), this term is simply zero. Next, we have the following decomposition:

$$
\begin{aligned}
\|\sum_{\ell=1}^{N}(\alpha_\ell-\widetilde{\alpha}_\ell)\,g_\ell\|_{\widehat{\rho_X}} \;\le\; & \|\sum_{l=1}^{N}(\alpha_l-\hat{\alpha}_l)\,g_\ell\,\mathbb{I}\{|\hat{\alpha}_\ell|\ge\lambda_n(2)\}\mathbb{I}\{|\alpha_\ell|\ge\lambda_n(2)/2\}\|_{\widehat{\rho_X}}\\
&+\|\sum_{\ell=1}^{N}(\alpha_\ell-\hat{\alpha}_\ell)\,g_\ell\,\mathbb{I}\{|\hat{\alpha}_\ell|\ge\lambda_n(2)\}\mathbb{I}\{|\alpha_\ell|<\lambda_n(2)/2\}\|_{\widehat{\rho_X}}\\
&+\|\sum_{l=1}^{N}\alpha_\ell\,g_\ell\,\mathbb{I}\{|\hat{\alpha}_\ell|<\lambda_n(2)\}\,\mathbb{I}\{|\alpha_\ell|\ge 2\lambda_n(2)\}\|_{\widehat{\rho_X}}\\
&+\|\sum_{\ell=1}^{N}\alpha_\ell\,g_\ell\,\mathbb{I}\{|\hat{\alpha}_\ell|<\lambda_n(2)\}\mathbb{I}\{|\alpha_\ell|<2\lambda_n(2)\}\|_{\widehat{\rho_X}}\\
:=\; & BB+BS+SB+SS.
\end{aligned}
$$

Using Lemma 5.1 and Hypothesis (14) :

$$
SS^2\le(1+\delta)\sum_{\ell=1}^{N}|\alpha_\ell|^2\,\mathbb{I}\{|\alpha_\ell|<2\lambda_n(2)\}\le(1+\delta)\,c_0\frac{S}{n}\le c_0(1+\delta)\eta_n^2
$$

We obviously have

$$
BB^2\le(1+\delta)\sum_{\ell=1}^{N}(\alpha_l-\hat{\alpha}_\ell)^2\,\mathbb{I}\{|\alpha_\ell|\ge\lambda_n(2)/2\}
$$

and using the triangular inequality, we obtain the same bound for SB

$$
\begin{aligned}
SB^2 \le\; & \|\sum_{l=1}^{N}\widehat{\alpha}_\ell\,g_\ell\,\mathbb{I}\{|\widehat{\alpha}_\ell|<\lambda_n(2)\}\,\mathbb{I}\{|\alpha_\ell|\ge 2\lambda_n(2)\}\|_{\widehat{\rho_X}}\\
&+\|\sum_{l=1}^{N}(\alpha_\ell-\widehat{\alpha}_\ell)\,g_\ell\,\mathbb{I}\{|\widehat{\alpha}_\ell|<\lambda_n(2)\}\,\mathbb{I}\{|\alpha_\ell|\ge 2\lambda_n(2)\}\|_{\widehat{\rho_X}}\\
\le\; & (1+\delta)\sum_{\ell=1}^{N}\widehat{\alpha}_\ell^{\,2}\,\mathbb{I}\{|\hat{\alpha}_\ell|<\lambda_n(2)<|\hat{\alpha}_\ell-\alpha_\ell|\}\,\mathbb{I}\{|\alpha_\ell|\ge 2\lambda_n(2)\}\\
&+(1+\delta)\sum_{\ell=1}^{N}(\hat{\alpha}_\ell-\alpha_\ell)^2\,\mathbb{I}\{|\hat{\alpha}_\ell|<\lambda_n(2)\}\,\mathbb{I}\{|\alpha_\ell|\ge 2\lambda_n(2)\}\\
\le\; & 2(1+\delta)\sum_{l=1}^{N}(\alpha_l-\hat{\alpha}_l)^2\mathbb{I}\{|\alpha_l|\ge 2\lambda_n(2)\}\\
\le\; & 2(1+\delta)\sum_{l=1}^{N}(\alpha_l-\hat{\alpha}_l)^2\mathbb{I}\{|\alpha_l|\ge\lambda_n(2)/2\}.
\end{aligned}
$$

Remember the notations given in the preliminaries. V_N is the space spanned by the functions $\{g_1,\dots,g_N\}$ and $\bar{\alpha}$ is such that $G_N^t\bar{\alpha}:=P_{V_N}f$. We denote $\bar{f}=(f(X_1),\dots,f(X_n))$. As well, set $G_N^t\hat{\alpha}:=P_{V_N}[\bar{f}+\varepsilon]$. We can write

$$\left[\sum_{\ell=1}^{N}(\alpha_l-\widehat{\alpha}_\ell)^2\,\mathbb{I}\{|\alpha_\ell|\geq\lambda_n(2)/2\}\right]^{1/2}\leq\left[\sum_{l=1}^{N}(\alpha_\ell-\bar{\alpha}_\ell)^2\,\mathbb{I}\{|\alpha_\ell|\geq\lambda_n(2)/2\}\right]^{1/2}+\left[\sum_{\ell=1}^{N}(\bar{\alpha}_\ell-\widehat{\alpha}_\ell)^2\mathbb{I}\{|\alpha_\ell|\geq\lambda_n(2)/2\}\right]^{1/2}.$$

The first term of the RHS can be bounded using Lemma 5.2. Denote by $\mathscr{L}$ the set of indices

$$\mathscr{L}=\{\ell\in\ \{1,\ldots,N\},\ |\alpha_\ell|\geq\lambda_n(2)/2\},$$

S the subset of the dictionary $S=\{g_\ell,\ell\in\mathscr{L}\}$ and V_S the space spanned by the functions of S. Using again Lemma 5.1,

$$\begin{aligned}\sum_{\ell=1}^{N}(\bar{\alpha}_\ell-\widehat{\alpha}_\ell)^2\,\mathbb{I}\{|\alpha_\ell|\geq\lambda_n(2)/2\}&\leq(1-\delta)^{-1}\|\sum_{\ell=1}^{N}\mathbb{I}\{|\alpha_\ell|\geq\lambda_n(2)/2\}(\widehat{\alpha}_\ell-\bar{\alpha}_\ell)g_\ell\|_{\hat{\rho}}^2\\&=(1-\delta)^{-1}\frac{1}{n}\,\|P_{V_S}[G_N(\widehat{\alpha}-\bar{\alpha})]\,\|_{l^2(n)}^2\\&=(1-\delta)^{-1}\frac{1}{n}\|P_{V_S}\varepsilon\|_{l^2(n)}^2.\end{aligned}$$

Using this bound, we can summarize the different inequalities by:

$$\begin{aligned}SS+SB+BB\leq\{c_0(1+\delta)^{\frac{1}{2}}+3c_t[(1+\delta)(1-\delta)^{-1}\delta]^{\frac{1}{2}}\}\eta_n\\+3[(1+\delta)(1-\delta)^{-1}]^{\frac{1}{2}}\frac{1}{n^{1/2}}\|P_{V_S}\varepsilon\|_{l^2(n)}\end{aligned}$$

as soon as

$$D\geq\sqrt{3}\{c_0(1+\delta)^{\frac{1}{2}}+3c_t[(1+\delta)(1-\delta)^{-1}\delta]^{\frac{1}{2}}\}.\tag{24}$$

Now, applying the technical Lemma 5.4 we can bound this last term by

$$P(\|P_{V_S}\varepsilon\|_{l^2(n)}^2\geq n[27(1-\delta)(1+\delta)]^{-1}\eta^2)\leq\exp\{-c_1n\eta^2\}$$

with $c_1=[216(1-\delta)(1+\delta)]^{-1}$. The last term to investigate is BS. Notice that

$$P(BS\geq\eta^2/3)\leq\sum_{k\leq N}P\left(BS\geq\eta^2\ \text{ and }\ \#\{|\alpha_l-\hat{\alpha}_l|\geq\frac{\lambda_n(2)}{2}\}=k\right)$$

Using Condition (9), observe that

$$\#\left\{|\alpha_l-\hat{\alpha}_l|\geq\frac{\lambda_n(2)}{2}\right\}=k\Longrightarrow BS\geq(1-\delta)k\lambda_n(2)^2/4.$$

Fix $z>0$ and introduce $K_0=\inf\{k,\ k\lambda_n(2)^2\geq z\eta^2\}$. We get

$$P(BS \geq \eta^2/3) \leq \sum_{k=1}^{K_0} \sum_{A\subset\{1,\dots,N\},\#A=k} P\left(\| \sum_{\ell\in A} (\alpha_\ell - \widehat{\alpha}_\ell)\, g_\ell \|^2_{\widehat{\rho_X}} \geq \eta^2/3 \right)$$

$$+ \sum_{k=1+K_0}^{N} \sum_{A\subset\{1,\dots,N\},\#A=k} P\left(\| \sum_{\ell\in A} (\widehat{\alpha}_\ell - \alpha_\ell)\, g_\ell \mathbb{I}\{|\alpha_\ell - \widehat{\alpha}_\ell| \geq \frac{\lambda_n(2)}{2}\} \|^2_{\widehat{\rho_X}} \geq (1-\delta)k\frac{\lambda_n(2)^2}{4} \right)$$

$$:= BS_1 + BS_2.$$

Applying Proposition 5.1, we have

$$\| \sum_{\ell\in A} (\widehat{\alpha}_\ell - \alpha_\ell)\, g_\ell \|_{\widehat{\rho_X}} \leq \sqrt{\frac{(1+\delta)\delta}{(1-\delta)}} c_t \eta_n + (1-\delta)^{-1/2} \frac{1}{n^{1/2}} \|P_{V_A}\varepsilon\|_{l^2(\#A)}$$

$$\leq \frac{\eta}{2\sqrt{3}} + (1-\delta)^{-1/2} \frac{1}{n^{1/2}} \|P_{V_A}\varepsilon\|_{l^2(\#A)}$$

as soon as

$$D \geq 2\sqrt{3}\sqrt{\frac{(1+\delta)\delta}{(1-\delta)}} c_t. \tag{25}$$

We bound BS_1 using Lemma 5.4

$$BS_1 \leq \sum_{k=1}^{K_0} \sum_{A\subset\{1,\dots,N\},\#A=k} P\left(\|P_{V_A}\varepsilon\|_{l^2(k)^2} \geq n\frac{1-\delta}{12}\eta^2 \right)$$

$$\leq N^{K_0} \exp\left(-c_2 n \eta^2\right)$$

with $c_2 = (1-\delta)/96$. For BS_2, we proceed as above:

$$\| \sum_{l\in A} (\hat{\alpha}_l - \alpha_l) g_l \|_{\widehat{\rho_X}} \leq \sqrt{\frac{(1+\delta)\delta}{(1-\delta)}} c_t \eta_n + (1-\delta)^{-1/2} \frac{1}{n^{1/2}} \|P_{V_A}\varepsilon\|_{l^2(\#A)}$$

$$\leq \sqrt{\frac{(1-\delta)k\lambda_n^2(2)}{8}} + (1-\delta)^{-1/2} \frac{1}{n^{1/2}} \|P_{V_A}\varepsilon\|_{l^2(\#A)}$$

as soon as

$$zD^2 \geq 8(1+\delta)\delta(1-\delta)^{-2} c_t^2. \tag{26}$$

Using again concentration lemma 5.4 (and because $(1-\delta)k\lambda_n^2(2)/8 \geq 2k/n$),

$$BS_2 \leq \sum_{k=1+K_0}^{N} N^k \exp\left(-(1-\delta)nk\frac{\lambda_n(2)^2}{8} \right)$$

$$\leq \sum_{k=1+K_0}^{N} \exp\left(k\left[\log N - (1-\delta)\frac{T_2^2 \log n}{8} \right] \right).$$

Recall that $N \leq n$, take $T_2^2 \geq 16/(1-\delta)$. We get

$$BS_2 \leq \exp\left(-K_0(1-\delta)\frac{T_2^2 \log n}{16}\right).$$

If we summarize, we obtain

$$\begin{aligned} BS &\leq N^{K_0} \exp\left(-c_2 n\eta^2\right) + \exp\left(-K_0(1-\delta)\frac{T_2^2 \log n}{16}\right) \\ &\leq \exp\left(-n\eta^2 \, \frac{(1-\delta)^2 T_2^2}{76[(1-\delta)T_2^2 + 16]}\right). \end{aligned}$$

Taking $z = (1-\delta)^2 T_2^2 \left(76[(1-\delta)T_2^2 + 16]\right)^{-1}$ ends the proof for D such that

$$\begin{aligned} D^2 \geq \Bigg[&\frac{608(1+\delta)\delta c_t^2[(1-\delta)T_2^2 + 16]}{(1-\delta)^4 T_2^2} \\ &\vee 3\left(c_0^2(1+\delta) + 9c_t^2 \frac{(1+\delta)\delta}{1-\delta}\right) \vee 12 \frac{(1+\delta)\delta}{(1-\delta)} c_t^2 \Bigg]. \end{aligned}$$

References

1. Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94, 2008.
2. P.J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. Preprint Submitted to the Annals of Statistics, 2007.
3. Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Universal algorithms for learning theory. II. Piecewise polynomial functions. *Constr. Approx.*, 26(2):127–152, 2007.
4. Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore Universal piecewise polynomial estimators for machine learning. In *Curve and surface design: Avignon 2006*, Mod. Methods Math., pages 48–77. Nashboro Press, Brentwood, TN, 2007.
5. Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, and Vladimir Temlyakov. Universal algorithms for learning theory. I. Piecewise constant functions. *J. Mach. Learn. Res.*, 6:1297–1321 (electronic), 2005.
6. Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007. (electronic)
7. Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 530–543. Springer, Berlin, 2007.
8. Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.
9. Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, and Ronald DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.
10. Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.
11. A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.
12. Ronald DeVore, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov. Approximation methods for supervised learning. *Found. Comput. Math.*, 6(1):3–58, 2006.

13. D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995. With Discussion.
14. David L. Donoho, I. Droro, Y. Tsaig, and J.L. Stark. Sparse solution of underdeterminded linear equations by stagewise orthogonal matching pursuit. Technical report, Stanford University, 2009.
15. David L. Donoho and A. Maleki. Freely available, optimally tuned iterative reconstruction algorithms for compressed sensing. Technical report, Stanford University, 2009.
16. Jianqing Fan and Lv Jinchi. Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, 70:849–911, 2008.
17. I. Ibragimov and R. Hasminskii. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981.
18. G. Kerkyacharian and D. Picard. Thresholding algorithms and well-concentrated bases. *Test*, 9(2), 2000.
19. Gerard Kerkyacharian, Mathilde Mougeot, Dominique Picard, and Karine Tribouley. Learning out leaders: exponential rates of convergence in high dimensional regression. 2009.
20. S.V. Konyagyn and V.N. Temlyakov Some error estimates in learning theory. In *Approximation theory: a volume dedicated to Borislav Bojanov*, pages 126–144. Prof. M. Drinov Acad. Publ. House, Sofia, 2004.
21. Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
22. Mathilde Mougeot. Choosing adaptive thresholds for diagnosis problems. Technical report, Nanterre University, 2009.
23. A.S. Nemirovskiy. Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, 235(3):50–60, 1985.
24. Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.
25. V.N. Temlyakov. Approximation in learning theory. *Constr. Approx.*, 27(1):33–74, 2008.
26. V.N. Temlyakov. Greedy approximation. *Acta Numer.*, 17:235–409, 2008.
27. Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
28. Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

Optimized wavelet preconditioning

Angela Kunoth

Abstract The numerical solution of linear stationary variational problems involving elliptic partial differential operators usually requires iterative solvers on account of their problem size. Our guiding principle is to devise theoretically and practically efficient iterative solution schemes which are optimal in the number of arithmetic operations, i.e., of linear complexity in the total number of unknowns. For these algorithms, asymptotically optimal preconditioners are indispensable. This article collects the main ingredients for multilevel preconditioners based on wavelets for certain systems of elliptic PDEs with smooth solutions. Specifically, we consider problems from optimal control with distributed or Dirichlet boundary control constrained by elliptic PDEs. Moreover, the wavelet characterization of function space norms will also be used in modelling the control functional, thereby extending the range of applicability over conventional methods. The wavelet preconditioners are optimized for these PDE systems to exhibit small absolute condition numbers and consequently entail absolute low iteration numbers, as numerical experiments show.

1 Introduction

For variational systems involving linear elliptic partial differential equations (PDEs) with smooth solutions, standard finite element or finite difference discretizations on uniform grids lead to the problem to solve a large ill-conditioned system of linear equations, due to the fact that PDE operators have positive order. Any iterative solution scheme will therefore become prohibitively slow since its speed depends on the spectral condition number, and the effect becomes even worse when the grid becomes finer and the number of unknowns increases. But since solutions typically exhibit a multiscale behaviour, enhancing iterative methods by multilevel ingredi-

Angela Kunoth
Institut für Mathematik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany,
e-mail: kunoth@math.uni-paderborn.de,
URL: http://www2.math.uni-paderborn.de/ags/kunoth/group/angelakunoth.html

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_10,

ents have proved to achieve much more efficient solution schemes. Naturally, one strives for an 'optimally efficient scheme', meaning that one can solve the problem with fine grid accuracy with an amount of arithmetic operations that is proportional to the number of unknowns on this grid. The first such methods which were proven to provide an asymptotically optimal iterative scheme were geometric multigrid algorithms [BH]. The basic idea of these schemes is to successively solve smaller versions of the linear system which can often be interpreted as discretizations with respect to coarser grids, thereby reducing the spectral condition number of the original system matrix and, hence, suggesting the term 'preconditioner'.

The search for optimal preconditioners was a major topic for numerical elliptic PDE solvers in the '80's. The goal was to better understand the ingredients which made a preconditioner optimal and, specifically, to find directly applicable versions which could be interpreted as a change of basis. With the arrival of the hierarchical basis preconditioner [Y], extending an idea of Babuška from the univariate case, a simple preconditioner became available. Although it is not optimal — the system matrix still exhibits a logarithmically growing spectral condition number in the bivariate case and exponential growth in three spatial dimensions — its simplicity still makes it popular up to now [MB]. During this time, a new methodology to derive preconditioners via space decomposition and subspace corrections was developed by Jinchao Xu [X1, X2]. The BPX preconditioner proposed first in [BPX] was numerically observed to be optimal; it is based on a weighted hierarchical generator system. With techniques from Approximation Theory, its optimality was theoretically established in [DK1, O]. Since then, its range of application has been widened extensively. For example, for second and fourth order elliptic problems on the sphere a BPX-type preconditioner has been developed and its optimality proved recently in [MKB]. The survey article by Jinchao Xu and coauthors in this volume records extensions of the BPX and of multigrid preconditioners to $H(\mathrm{grad})$, $H(\mathrm{curl})$, and $H(\mathrm{div})$ systems on adaptive and unstructured grids.

At about the same time, wavelets as a special example of a multiscale basis of $L_2(\mathbb{R})$ with compact support were constructed [Dau]. While initially mainly developed and used for signal analysis and image compression, wavelets were soon discovered to also provide optimal preconditioners in the above sense for second order elliptic boundary value problems [DK1, J]. However, the fact that one cannot really exploit L_2-orthogonality for elliptic boundary value problems together with the difficulty that the L_2-orthogonal Daubechies wavelets are only given implicitly led to the search for variants which are more practical for numerical PDEs. It was soon realized that biorthogonal spline-wavelets as developed in [CDF] are better suited since they allow one to work with piecewise polynomials for the actual discretization.

The principal and crucial property to prove optimality of a wavelet preconditioner are norm equivalences between Sobolev norms and sequence norms of weighted wavelet expansion coefficients. On this basis, optimal conditioning of the resulting linear system of equations can be achieved by applying the Fast Wavelet Transform to a single-scale discretization on a uniform grid, together with an application of an appropriate diagonal matrix.

Nowadays, the terminology 'wavelets' is used in a more general sense that originally in [Dau]: we rather consider classes of multiscale bases with three main features:

(R) Riesz basis property for the underlying function spaces,
(L) locality of the basis functions, and
(CP) cancellation properties.

These will be detailed in Section 3.

After the initial results concerning optimal preconditioning with functions of local support in [DK1], research on using wavelets for numerically solving elliptic PDEs went into different directions. One problem was that the original constructions in [Dau, CDF] and many others were based on employing the Fourier transform so that these constructions provide bases only for function spaces on all of $\mathbb{R}$, on the torus or, by tensorization, on $\mathbb{R}^n$. In contrast, PDEs naturally live on a bounded domain $\Omega \subset \mathbb{R}^n$. In order for wavelets to be employed for numerical PDEs, there arose the need for constructions of wavelets on bounded intervals and domains without, of course, loosing the crucial properties (R), (L) and (CP). The first such systematic construction of biorthogonal spline-wavelets on $[0,1]$ and, by tensor products, on $[0,1]^n$, was provided in [DKU]. Different domain decomposition approaches yield constructions of biorthogonal wavelets on domains which can be represented as unions of parametric mappings of $[0,1]^n$ [CTU, DS2, DS3, KS], see the article by Helmut Harbrecht and Reinhold Schneider in this volume and also [U] for details. Once such bases are available, the absolute value of the condition numbers of (systems of) elliptic PDEs can be ameliorated significantly by further inexpensive linear transformations taking into account a setup of the system matrices on the coarsest grid called operator–based preconditioning [Bu1, Pa].

Aside from optimal preconditioning, the built-in potential of local adaptivity for wavelets is playing a prominent role when solving stationary PDEs with non–smooth solutions, on account of the fact that wavelets provide a locally supported Riesz basis for a whole function space. This issue is extensively addressed in the article by Rob Stevenson in this volume.

In addition to the material in this volume, there are at least four extensive surveys on wavelet and multiscale methods for more general PDEs addressing, among other things, the connection between adaptivity and nonlinear approximation and the evaluation of nonlinearities [Co, D2, D3, D4].

In my article, I want to remain focussed on discretizations for smooth solutions (for which uniform grids give desired accuracy) since ideally an adaptive scheme should also perform numerically well for this case. Thus, in order to asses numerical tests, results for reference schemes on uniform grids should be available.

Another extremely useful application of the Riesz basis property (R) of wavelets concerns PDE-constrained control problems guided by elliptic boundary value problems. Here a quadratic optimization functional involving Sobolev norms of the state and the control of a system is to be minimized, subject to an elliptic PDE in variational form which couples state and control variables. In wavelet bases, the numerical evaluation of Sobolev norms even with fractional smoothness indices amounts to a multiplication with a diagonal basis and can be realized fast [Bu2].

This allows one to efficiently evaluate natural function space norms as they arise for PDE–constrained control, or different norms in the control functional more adequate for modelling purposes [Bu1, BK]. Conventional discretizations based on finite elements have concentrated here on evaluating function space norms with integer smoothness. Also for linear–quadratic elliptic control problems with non–smooth solutions, adaptive wavelets provide most efficient solution schemes. Convergence and optimal complexity estimates of respective adaptive wavelet methods were established in [DK3]. Among such optimization problems, boundary control problems where the control is exerted through essential boundary conditions, appear practically most often. Formulating the elliptic PDE as a saddle point problem by introducing Lagrange multipliers for the boundary conditions allows one to handle changing boundary controls in a flexible manner. Wavelet approaches for treating such more involved systems of elliptic PDEs in saddle point form have been investigated in [K1, K4] and numerically optimized in [Pa].

This paper is of a more introductory nature: its purpose is to collect the basic ingredients for wavelet preconditioners, apply them to (systems of) linear elliptic PDEs in variational form and provide some numerical results on their performance. Specifically, some effort will be spent on the description of nested iterative solution schemes for systems from PDE-constrained control problems.

The structure of this paper is as follows. First, in Section 2, some well–posed variational problems are compiled. The simplest example is a linear second order elliptic boundary value problem for which we derive two forms of an operator equation, once as a single equation and once as a saddle point system where nonhomogeneous boundary conditions are handled by means of Lagrange multipliers. Both formulations are then used as basic systems for PDE-constrained control problems, one with control through the right hand side and one involving a Dirichlet boundary control. In Section 3 necessary ingredients and basic properties of wavelets are assembled. In particular, Section 3.4 collects the essential construction principles for wavelets on bounded domains which do not rely on Fourier techniques, namely, multiresolution analyses of function spaces and the concept of stable completions. In Section 4 we formulate the problem classes introduced in Section 2 in wavelet coordinates and derive in particular for the control problems the resulting systems of linear equations arising from the optimality conditions. Section 5 is devoted to the iterative solution of these systems. We investigate fully iterative schemes on uniform grids and show that the resulting systems can be solved in the wavelet framework together with a nested iteration strategy with an amount of arithmetic operations which is proportional to the total number of unknowns on the finest grid. Numerical experiments on the performance of the solvers as well as on the modelling issue round off this contribution.

The following notations are used frequently. The relation $a \sim b$ always stands for $a \lesssim b$ and $b \lesssim a$ where the latter inequality means that b can be bounded by some constant times a uniformly in all parameters on which a and b may depend. Norms and inner products are indexed by the corresponding function space. $L_p(\Omega)$ are for $1 \le p \le \infty$ the usual Lebesgue spaces on a domain $\Omega \subset \mathbb{R}^n$, and $W_p^k(\Omega) \subset L_p(\Omega)$

denote for $k \in \mathbb{N}$ the Sobolev spaces of functions whose weak derivatives up to order k are bounded in $L_p(\Omega)$. For $p = 2$, we abbreviate $H^k(\Omega) = W_2^k(\Omega)$.

2 Systems of elliptic partial differential equations (PDEs)

We first formulate the classes of variational problems which will be investigated here in an abstract form.

2.1 Abstract operator systems

Let $\mathscr{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathscr{H}}$ with normed dual $\mathscr{H}'$ endowed with the norm

$$\|w\|_{\mathscr{H}'} := \sup_{v \in \mathscr{H}} \frac{\langle v, w \rangle}{\|v\|_{\mathscr{H}}} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dual pairing between $\mathscr{H}$ and $\mathscr{H}'$.

Given $F \in \mathscr{H}'$, the goal is to find a solution to the operator equation

$$\mathscr{L} U = F \tag{2}$$

where $\mathscr{L} : \mathscr{H} \to \mathscr{H}'$ is a linear operator which is assumed to be a bounded bijection,

$$\|\mathscr{L} V\|_{\mathscr{H}'} \sim \|V\|_{\mathscr{H}}, \qquad V \in \mathscr{H}. \tag{3}$$

The operator equation (2) is *well-posed* since (3) implies for any given data $F \in \mathscr{H}'$ the existence and uniqueness of the solution $U \in \mathscr{H}$ which depends continuously on the data. Property (3) is also called *mapping property* of $\mathscr{L}$.

The examples that we consider will be such that $\mathscr{H}$ is a product space

$$\mathscr{H} := H_{1,0} \times \cdots \times H_{m,0}, \tag{4}$$

where each of the $H_{i,0} \subseteq H_i$ is a Hilbert space or a closed subspace of a Hilbert space H_i determined, for instance, by homogeneous boundary conditions. The spaces H_i will be Sobolev spaces living on a bounded domain $\Omega \subset \mathbb{R}^n$ or on (part of) its boundary. In view of the definition of $\mathscr{H}$, the elements $V \in \mathscr{H}$ will consist of m components $V = (v_1, \ldots, v_m)^T$ for which we define $\|V\|_{\mathscr{H}}^2 := \sum_{i=1}^m \|v_i\|_{H_i}^2$. The dual space $\mathscr{H}'$ is then endowed with the norm

$$\|W\|_{\mathscr{H}'} := \sup_{V \in \mathscr{H}} \frac{\langle V, W \rangle}{\|V\|_{\mathscr{H}}} \tag{5}$$

where $\langle V, W \rangle := \sum_{i=1}^m \langle v_i, w_i \rangle_i$ in terms of the dual pairing $\langle \cdot, \cdot \rangle_i$ between H_i and H_i'.

We will formulate four classes of problems which fit into this format. A recurring theme in the derivation of the system of operator equations (2) is the minimization of a quadratic functional.

2.2 A scalar elliptic boundary value problem

Denote by $\partial\Omega := \Gamma \cup \Gamma_N$ the boundary of Ω which is assumed to be piecewise smooth. We consider the scalar second order boundary value problem

$$\begin{aligned} -\nabla\cdot(\mathbf{a}\nabla y) + cy &= f \quad \text{in } \Omega, \\ y &= g \quad \text{on } \Gamma, \\ (\mathbf{a}\nabla y)\cdot\mathbf{n} &= 0 \quad \text{on } \Gamma_N, \end{aligned} \tag{6}$$

where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is the outward normal at $\mathbf{x} \in \Gamma$, $\mathbf{a} = \mathbf{a}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is symmetric, uniformly positive definite and bounded on Ω, and $c \in L_\infty(\Omega)$. Furthermore, f and g are some given right hand side and boundary data. With the usual definition of the bilinear form

$$a(v,w) := \int_\Omega (\mathbf{a}\nabla v\cdot\nabla w + cvw)\, d\mathbf{x}, \tag{7}$$

the weak formulation of (6) requires in the case $g \equiv 0$ to find $y \in \mathscr{H}$ where

$$\mathscr{H} := H^1_{0,\Gamma}(\Omega) := \{v \in H^1(\Omega) :\ v|_\Gamma = 0\}, \tag{8}$$

or

$$\mathscr{H} := \{v \in H^1(\Omega) : \int_\Omega v(\mathbf{x})\, d\mathbf{x} = 0\} \quad \text{when } \Gamma = \emptyset, \tag{9}$$

such that

$$a(y,v) = \langle v, f\rangle, \quad v \in \mathscr{H}. \tag{10}$$

Neumann–type boundary conditions on Γ_N are implicitly contained in the weak formulation (10), therefore called *natural boundary conditions*. In contrast, Dirichlet boundary conditions on Γ have to be posed explicitly, therefore called *essential boundary conditions*. The easiest way to achieve this for homogeneous Dirichlet boundary conditions when $g \equiv 0$ is to include them into the solution space as above in (8). In the nonhomogeneous case $g \not\equiv 0$ on Γ in (6) and $\Gamma \neq \emptyset$, one can reduce this to a problem with homogeneous boundary conditions by *homogenization* as follows. Let $w \in H^1(\Omega)$ be such that $w \equiv g$ on Γ. Then $\tilde{y} := y - w$ satisfies $a(\tilde{y},v) = a(y,v) - a(w,v) = \langle v, f\rangle - a(w,v) =: \langle v, \tilde{f}\rangle$ for all $v \in \mathscr{H}$ defined in (8), and on Γ one has $\tilde{y} = g - w \equiv 0$ yielding $\tilde{y} \in \mathscr{H}$. Therefore, it suffices to consider the weak form (10) with a perhaps modified right hand side. Another possibility which allows to treat the case $g \not\equiv 0$ explicitly is discussed in the next section.

The crucial properties are that the bilinear form defined in (7) is symmetric, continuous and elliptic on $\mathscr{H}$,

$$a(v,v) \sim \|v\|^2_{\mathscr{H}} \qquad \text{for any } v \in \mathscr{H}, \tag{11}$$

see, e.g., [B]. By Riesz' representation theorem, the bilinear form defines a linear operator $A : \mathscr{H} \to \mathscr{H}'$ by

$$\langle w, Av\rangle := a(v,w), \qquad v,w \in \mathscr{H}, \tag{12}$$

which is under the above assumptions an isomorphism,

$$c_A\|v\|_{\mathscr{H}} \;\le\; \|Av\|_{\mathscr{H}'} \;\le\; C_A\|v\|_{\mathscr{H}} \qquad \text{for any } v \in \mathscr{H}. \tag{13}$$

Relation (13) entails that given any $f \in \mathscr{H}'$, there exists a unique $y \in \mathscr{H}$ which solves the linear operator equation

$$Ay = f \qquad \text{in } \mathscr{H}' \tag{14}$$

derived from (10). This linear system where the operator defines a bounded bijection in the sense of (13) is the simplest case of a well-posed variational problem (2). In the notation from Section 2.1, we have here $m = 1$ and $\mathscr{L} = A$.

2.3 Saddle point problems involving essential boundary conditions

A particular saddle point problem derived from (6) shall be considered next. Since it is particularly appropriate to handle essential non–homogeneous Dirichlet boundary conditions, it will also be employed later in the context of boundary control problems.

Recall, e.g., from [B] that the solution $y \in \mathscr{H}$ of (10) is also the unique minimizer of the minimization problem

$$\inf_{v\in\mathscr{H}} \mathscr{J}(v), \qquad \mathscr{J}(v) := \frac{1}{2}a(v,v) - \langle v, f\rangle. \tag{15}$$

This means that y is a critical point for the first order variational derivative of $\mathscr{J}$, i.e., $\delta\mathscr{J}(y;v) = 0$. Here $\delta^s\mathscr{J}(v;w_1,\ldots,w_s)$ denotes the s–th variation of $\mathscr{J}$ at v in directions $w_1,\ldots,w_s$. In particular, for $s = 1$

$$\delta\mathscr{J}(v;w) := \lim_{t\to 0} \frac{\mathscr{J}(v+tw) - \mathscr{J}(v)}{t} \tag{16}$$

is the (Gateaux) derivative of $\mathscr{J}$ at v in direction w.

Generalizing (15) to the case of nonhomogeneous Dirichlet boundary conditions g, we want to minimize $\mathscr{J}$ over $v \in H^1(\Omega)$ subject to constraints in form of the essential boundary conditions $v = g$ on Γ. A standard technique from nonlinear optimization is to employ a *Lagrange multiplier* p to append the constraints to the optimization functional $\mathscr{J}$ defined in (15). Satisfying the constraint is guaranteed

by taking the supremum over all such Lagrange multipliers before taking the infimum. Thus, minimization subject to a constraint leads to the problem of finding a *saddle point* (y,p) of the *saddle point problem*

$$\inf_{v\in H^1(\Omega)} \sup_{q\in(H^{1/2}(\Gamma))'} \mathscr{J}(v) + \langle v-g, q\rangle_\Gamma. \tag{17}$$

The choice of the Lagrange multiplier space and the dual form $\langle\cdot,\cdot\rangle_\Gamma$ in (17) can be explained as follows. The boundary expression $v = g$ actually means taking the *trace* of $v \in H^1(\Omega)$ to $\Gamma \subseteq \partial\Omega$ which we explicitly write as $\gamma v := v|_\Gamma$. Classical trace theorems from, e.g., [Gr] state that for any $v \in H^1(\Omega)$ one looses '$\frac{1}{2}$ order of smoothness' when taking traces, therefore yielding $\gamma v \in H^{1/2}(\Gamma)$. Thus, when the data g is such that $g \in H^{1/2}(\Gamma)$, the expression in (17) involving the dual form $\langle\cdot,\cdot\rangle_\Gamma := \langle\cdot,\cdot\rangle_{H^{1/2}(\Gamma)\times(H^{1/2}(\Gamma))'}$ is well–defined, and so is the selection of the multiplier space $(H^{1/2}(\Gamma))'$. In case of Dirichlet boundary conditions on the whole boundary of Ω, i.e., the case $\Gamma \equiv \partial\Omega$, one can identify $(H^{1/2}(\Gamma))' = H^{-1/2}(\Gamma)$.

The above formulation (17) was first investigated in [Ba]. Another standard technique from optimization to handle minimization problems under constraints is to append the constraints to $\mathscr{J}(v)$ by means of a *penalty parameter* ε. For this approach, however, the system matrix depends on ε. So far, no optimal preconditioners have been established for this case so that we do not discuss this method here any further.

The method of appending essential boundary conditions by Lagrange multipliers is particularly appealing in connection with *fictitious domain methods* which may be used for problems with changing boundaries such as shape optimization problems. There one embeds the domain Ω into a larger, simple domain $\square$, and formulates (17) with respect to $H^1(\square)$ and dual form on the boundary Γ [K2]. One should note, however, that in the case that Γ is a proper subset of $\partial\Omega$, there may occur some ambiguity in the relation between the fictitious domain formulation and the corresponding strong form (6). In fact, without further assumptions, one cannot establish that the infimum of (17) with respect to $H^1(\square)$, when restricted to Ω, is the same as taking the infimum of (17) with respect to $H^1(\Omega)$. This is indeed guaranteed by using another set of Lagrangian multipliers. We currently investigate this for a problem from electrical impedance tomography in [KK].

In order to bring out the role of the trace operator, we define in addition to (7) a second bilinear form on $H^1(\Omega)\times(H^{1/2}(\Gamma))'$ by

$$b(v,q) := \int_\Gamma (\gamma v)(s)\, q(s)\, ds \tag{18}$$

so that the saddle point problem (17) may be rewritten as

$$\inf_{v\in H^1(\Omega)} \sup_{q\in(H^{1/2}(\Gamma))'} \mathscr{J}(v,q), \qquad \mathscr{J}(v,q) := \mathscr{J}(v) + b(v,q) - \langle g, q\rangle_\Gamma. \tag{19}$$

Determining the critical points of first order variations of $\mathscr{J}$, now with respect to both v and q, yields the system of equations that a saddle point (y,p) has to satisfy

$$\begin{aligned} a(y,v)+b(v,p) &= \langle v,f\rangle, && v\in H^1(\Omega),\\ b(y,q) &= \langle g,q\rangle_\Gamma, && q\in (H^{1/2}(\Gamma))'. \end{aligned} \tag{20}$$

Defining the linear operator $B: H^1(\Omega)\to H^{1/2}(\Gamma)$ and its adjoint $B': (H^{1/2}(\Gamma))'\to (H^1(\Omega))'$ by $\langle Bv,q\rangle_\Gamma = \langle v,B'q\rangle_\Gamma := b(v,q)$, this can be rewritten as a linear operator equation from $\mathscr{H} := H^1(\Omega)\times (H^{1/2}(\Gamma))'$ to $\mathscr{H}'$ as follows:
Given $(f,g)\in\mathscr{H}'$, find $(y,p)\in\mathscr{H}$ that solves

$$\begin{pmatrix} A & B'\\ B & 0\end{pmatrix}\begin{pmatrix} y\\ p\end{pmatrix} = \begin{pmatrix} f\\ g\end{pmatrix}. \tag{21}$$

It can be shown in the present context that the Lagrange multiplier can be determined by $p=-\mathbf{n}\cdot\mathbf{a}\nabla y$ which can be interpreted as *stress force* on the boundary [Ba].

We briefly discuss the properties of B representing the trace operator. Classical trace theorems from, e.g., [Gr], state that for any $f\in H^s(\Omega)$, $1/2<s<3/2$, one has

$$\|f|_\Gamma\|_{H^{s-1/2}(\Gamma)} \lesssim \|f\|_{H^s(\Omega)}. \tag{22}$$

Conversely, for every $g\in H^{s-1/2}(\Gamma)$, there exists some $f\in H^s(\Omega)$ such that $f|_\Gamma = g$ and

$$\|f\|_{H^s(\Omega)} \lesssim \|g\|_{H^{s-1/2}(\Gamma)}. \tag{23}$$

Note that the range of s extends accordingly if Γ is more regular. Estimate (22) immediately entails for $s=1$ that $B: H^1(\Omega)\to H^{1/2}(\Gamma)$ is continuous. Moreover, the second property (23) means B is surjective, i.e., $\operatorname{range} B = H^{1/2}(\Gamma)$ and $\ker B' = \{0\}$. This yields that the *inf–sup condition*

$$\inf_{q\in(H^{1/2}(\Gamma))'}\sup_{v\in H^1(\Omega)}\frac{\langle Bv,q\rangle_\Gamma}{\|v\|_{H^1(\Omega)}\,\|q\|_{(H^{1/2}(\Gamma))'}} \gtrsim 1 \tag{24}$$

is satisfied.

In the notation from Section 2.1, the system (21) is a saddle point problem on $\mathscr{H} = Y\times Q$. Thus, we identify $Y = H^1(\Omega)$ and $Q = (H^{1/2}(\Gamma))'$ and linear operators $A: Y\to Y'$ and $B: Y\to Q'$.

The abstract theory of saddle point problems states that existence and uniqueness of a solution pair $(y,p)\in\mathscr{H}$ of (21) holds if and only if A and B are continuous, A is invertible on $\ker B\subseteq Y$ and the range of B is closed in Q', see, e.g., [B, BF] and the article in this volume by Ricardo Nochetto and coauthors. The properties for B and the continuity for A have been assured above. In addition, we will always deal here with operators A which are invertible on $\ker B$ which cover the standard cases of the Laplacian ($\mathbf{a}=I$ and $c\equiv 0$) and the Helmholtz operator ($\mathbf{a}=I$ and $c=1$).

Consequently, the operator

$$\mathscr{L} := \begin{pmatrix} A & B'\\ B & 0\end{pmatrix}: \mathscr{H}\to\mathscr{H}' \tag{25}$$

is linear bijection, and one has the mapping property

$$\left\| \mathscr{L} \begin{pmatrix} v \\ q \end{pmatrix} \right\|_{\mathscr{H}'} \sim \left\| \begin{pmatrix} v \\ q \end{pmatrix} \right\|_{\mathscr{H}} \tag{26}$$

for any $(v,q) \in \mathscr{H}$ with constants depending on upper and lower bounds for A, B. Finally, the operator equation (21) is established as a well-posed variational problem in the sense of Section 2.1: for given $(f,g) \in \mathscr{H}'$, there exists a unique solution $(y,p) \in \mathscr{H} = Y \times Q$ which depends continuously on the data.

2.4 PDE-constrained control problems: Distributed control

An important class of problems where the numerical solution of systems (14) or (21) is required repeatedly are control problems with PDE-constraints. Using the notation from Section 2.2, consider as a guiding model the objective to minimize a quadratic functional of the form

$$\mathscr{J}(y,u) = \frac{1}{2}\|y - y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2, \tag{27}$$

subject to the linear constraints

$$Ay = f + u \qquad \text{in } H' \tag{28}$$

where $A : H \to H'$ is defined as above in (12) satisfying (13), and $f \in H$ is given. The space H is in this subsection defined as in (8) or in (9), and we reserve the symbol $\mathscr{H}$ for a resulting product space later.

In order for a solution y of (28), the *state* of the system, to be well–defined, the problem formulation has to ensure that the unknown *control* u appearing in the right hand side of (28) is at least in H'. This can be achieved by choosing the *control space* $\mathscr{U}$ whose norm appears in (27) such that it is as least as smooth as H'. The second ingredient in the optimization functional (27) is a data fidelity term which enforces a match of the system state y to some prescribed target state y_*, measured in some norm which is typically weaker than $\|\cdot\|_H$. Thus, we require that the *observation space* $\mathscr{Z}$ and the control space $\mathscr{U}$ are such that the continuous embeddings

$$\|v\|_{H'} \lesssim \|v\|_{\mathscr{U}}, \quad v \in \mathscr{U}, \qquad \|v\|_{\mathscr{Z}} \lesssim \|v\|_H, \quad v \in H, \tag{29}$$

hold. Mostly one has investigated the simplest cases of norms which occur for $\mathscr{U} = \mathscr{Z} = L_2(\Omega)$ and which are covered by these assumptions [Li]. The parameter $\omega > 0$ balances the norms in (27), the data fidelity term and the control.

Since the control appears in all of the right hand side of (28), such control problems are called *distributed* control problems. Although their practical value is of a rather limited nature, distributed control problems help to bring out the basic mechanisms. Note that when the observed data is *compatible* in the sense that $y_* \equiv A^{-1}f$, the control problem yields the trivial control $u \equiv 0$ which implies $\mathscr{J}(y,u) \equiv 0$.

Solution schemes for the control problem (27) subject to the constraints (28) can be based on the system of operator equations derived next by the same variational principles as employed in the previous section, using a Lagrange multiplier p to enforce the constraints. Defining the Lagrangian functional

$$\mathrm{Lagr}(y,p,u) := \mathscr{J}(y,u) + \langle p, Ay - f - u\rangle \tag{30}$$

on $H \times H \times H'$, the first order necessary conditions or *Karush-Kuhn-Tucker (KKT) conditions* $\delta\,\mathrm{Lagr}(x) = 0$ for $x = p, y, u$ can be derived as

$$\begin{aligned} Ay &= f + u \\ A'p &= -S(y - y_*) \\ \omega R u &= p. \end{aligned} \tag{31}$$

Here the linear operators S and R can be interpreted as Riesz operators defined by the inner products $(\cdot,\cdot)_{\mathscr{Z}}$ and $(\cdot,\cdot)_{\mathscr{U}}$. The system (31) may be written in saddle point form as

$$\mathscr{L}V := \begin{pmatrix} \mathscr{A} & \mathscr{B}' \\ \mathscr{B} & 0 \end{pmatrix} V := \begin{pmatrix} S & 0 & A' \\ 0 & \omega R & -I \\ A & -I & 0 \end{pmatrix} \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} Sy_* \\ 0 \\ f \end{pmatrix} =: F \tag{32}$$

on $\mathscr{H} := H \times H \times H'$. Here we can also allow that $\mathscr{Z}$ in (27) is a *trace space* on part of the boundary $\partial\Omega$ as long as the corresponding condition (29) is satisfied [K3]. The class of control problems where the control is exerted through Neumann boundary conditions can also be written in this form since in this case the control still appears on the right hand side of a single operator equation of a form like (28), see [DK3]. Well-posedness of the system (32) can now be established by applying the conditions for saddle point problems stated in Section 2.3.

A few statements on the *model* of the control problem should be made. While the PDE constraints (28) that govern the system are fixed, there is in many applications some ambiguity with respect to the choice of the spaces $\mathscr{Z}$ and $\mathscr{U}$. L_2 norms are easily realized in finite element discretizations, although in some applications smoother norms for the observation $\|\cdot\|_{\mathscr{Z}}$ or for the control $\|\cdot\|_{\mathscr{U}}$ are desirable. This is the case, for instance, in temperature cooling processes where also the gradient of the temperature of a material is to be controlled. Once $\mathscr{Z}$ and $\mathscr{U}$ are fixed, there is only a single parameter ω to balance the two norms in (27). *Modelling* the objective functional is therefore an issue where more flexibility may be advantageous. Specifically in a multiscale setting, one may want to weight contributions on different scales by multiple parameters.

The wavelet setting which we describe in Section 3 allows for this flexibility. It is based on formulating the norms in the objective functional in terms of weighted wavelet coefficient sequences which are equivalent to the norms for $\mathscr{Z}$, $\mathscr{U}$ and which, in addition, support an efficient numerical implementation. Once wavelet discretizations are introduced, we formulate below control problems with such objective functionals.

2.5 PDE-constrained control problems: Dirichlet boundary control

Practically the most relevant control problems are problems with Dirichlet boundary control. They can be posed using the saddle point formulation from Section 2.3.

We consider as an illustrative guiding model the problem to minimize for some given data y_* the quadratic functional

$$\mathscr{J}(y,u) = \frac{1}{2}\|y-y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2, \tag{33}$$

where, adhering to the notation in Section 2.2 the state y and the control u are coupled through the linear second order elliptic boundary value problem

$$\begin{aligned} -\nabla\cdot(\mathbf{a}\nabla y) + ky &= f \qquad \text{in } \Omega,\\ y &= u \qquad \text{on } \Gamma,\\ (\mathbf{a}\nabla y)\cdot\mathbf{n} &= 0 \qquad \text{on } \Gamma_N. \end{aligned} \tag{34}$$

The appearance of the control u as a Dirichlet boundary condition in (34) is referred to as a *Dirichlet boundary control*. In view of the treatment of essential Dirichlet boundary conditions in the context of saddle point problems derived in Section 2.3, we write the PDE constraints (34) in the operator form (21) on $Y\times Q$ where $Y = H^1(\Omega)$ and $Q = (H^{1/2}(\Gamma))'$.

The model control problem with Dirichlet boundary control then reads as follows: Minimize for given data $y_*\in\mathscr{Z}$ and $f\in Y'$ the quadratic functional

$$\mathscr{J}(y,u) = \frac{1}{2}\|y-y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2 \tag{35}$$

subject to

$$\begin{pmatrix} A & B'\\ B & 0\end{pmatrix}\begin{pmatrix} y\\ p\end{pmatrix} = \begin{pmatrix} f\\ u\end{pmatrix}. \tag{36}$$

In view of the problem formulation in Section 2.4 and the discussion of the choice of the observation space $\mathscr{Z}$ and the control space, here we require analogously that $\mathscr{Z}$ and $\mathscr{U}$ are such that the continuous embeddings

$$\|v\|_{Q'} \lesssim \|v\|_{\mathscr{U}}, \quad v\in\mathscr{U}, \qquad \|v\|_{\mathscr{Z}} \lesssim \|v\|_Y, \quad v\in Y, \tag{37}$$

hold. Also the case of observations on part of the boundary $\partial\Omega$ can be taken into account [K4]. It should be mentioned that the simple choice $\mathscr{U} = L_2(\Gamma)$ which is used in many applications of Dirichlet control problems is *not* covered here. Indeed, there may arise the problem of well-posedness as follows. The constraints (34) or, in weak form (21), guarantee a unique weak solution $y\in Y = H^1(\Omega)$ *provided* that the boundary term u satisfies $u\in Q' = H^{1/2}(\Gamma)$. In the framework of control problems, this smoothness of u therefore has to be required either by the choice of $\mathscr{U}$ or by the choice of $\mathscr{Z}$ (such as $\mathscr{Z} = H^1(\Omega)$) which would assure $By\in Q'$. In the latter case, we could relax condition (37) on $\mathscr{U}$.

By variational principles, we can derive as before the first order necessary conditions for a coupled *system* of saddle point problems. Well-posedness of this system can again be established by applying the conditions for saddle point problems from Section 2.3 where the inf-sup condition for the saddle point problem (21) yields an inf-sup condition for the exterior saddle point problem of interior saddle point problems [K1].

3 Wavelets

The numerical solution of the afore-mentioned classes of problems hinges on the availability of appropriate wavelet bases for the function spaces under consideration which are all specific Hilbert spaces on the domain or on (part of) its boundary.

3.1 Basic properties

For the above classes of problems, we need to have a wavelet basis at our disposal for each occurring function space. A *wavelet basis* for a Hilbert space H is here understood as a collection of functions

$$\Psi_H := \{\psi_{H,\lambda} : \lambda \in \mathcal{I}_H\} \subset H \tag{38}$$

which are indexed by elements λ from an infinite index set $\mathcal{I}_H$. Each of the indices λ comprises different information $\lambda = (j, \mathbf{k}, \mathbf{e})$ such as the *refinement scale* or *level of resolution* j and a spatial location $\mathbf{k} = \mathbf{k}(\lambda) \in \mathbb{Z}^n$. In more than one space dimensions, the basis functions are built from taking tensor products of certain univariate functions, and in this case the third index $\mathbf{e}$ contains information on the *type* of wavelet. We will frequently use the symbol $|\lambda| := j$ to access the resolution level j. In the univariate case on all of $\mathbb{R}$, $\psi_{H,\lambda}$ is typically generated by means of shifts and dilates of a single function ψ, i.e., $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$, $j,k \in \mathbb{Z}$, normalized with respect to $\|\cdot\|_{L_2}$. On bounded domains, the structure of the functions is essentially the same up to modifications near the boundary.

The three crucial properties that we will assume the wavelet basis to have for the sequel are the following.

Riesz basis property (R): Every $v \in H$ has a unique expansion in terms of Ψ_H,

$$v = \sum_{\lambda \in \mathcal{I}_H} v_\lambda \, \psi_{H,\lambda} =: \mathbf{v}^T \Psi_H, \quad \mathbf{v} := (v_\lambda)_{\lambda \in \mathcal{I}_H}, \tag{39}$$

and its expansion coefficients satisfy a *norm equivalence*: for any $\mathbf{v} = \{v_\lambda : \lambda \in \mathcal{I}_H\}$ one has

$$c_H \, \|\mathbf{v}\|_{\ell_2(\mathcal{I}_H)} \;\le\; \|\mathbf{v}^T \Psi_H\|_H \;\le\; C_H \, \|\mathbf{v}\|_{\ell_2(\mathcal{I}_H)}, \quad \mathbf{v} \in \ell_2(\mathcal{I}_H), \tag{40}$$

where $0 < c_H \le C_H < \infty$. This means that wavelet expansions induce *isomorphisms* between certain function spaces and sequence spaces. We write ℓ_2 norms without subscripts as $\|\cdot\| := \|\cdot\|_{\ell_2(\mathcal{I}_H)}$ when the index set is clear from the context. If the precise constants do not matter, we write the norm equivalence (40) shortly as

$$\|\mathbf{v}\| \ \sim \ \|\mathbf{v}^T \Psi_H\|_H, \quad \mathbf{v} \in \ell_2(\mathcal{I}_H). \tag{41}$$

Locality (L): The functions $\psi_{H,\lambda}$ have compact support which decreases with increasing level $j = |\lambda|$, i.e.,

$$\operatorname{diam}(\operatorname{supp} \psi_{H,\lambda}) \ \sim \ 2^{-|\lambda|}. \tag{42}$$

Cancellation property (CP): There exists an integer $\tilde{m} = \tilde{m}_H$ such that

$$\langle v, \psi_{H,\lambda} \rangle \lesssim 2^{-|\lambda|(n/2-n/p+\tilde{m})} |v|_{W_p^{\tilde{m}}(\operatorname{supp} \psi_{H,\lambda})}. \tag{43}$$

This means that integrating against a wavelet has the effect of taking an $\tilde{m}$th order difference which annihilates the smooth part of v. This property is for wavelets defined on Euclidean domains typically realized by constructing Ψ_H in such a way that it possesses a *dual* or *biorthogonal* basis $\tilde{\Psi}_H \subset H'$ such that the multiresolution spaces $\tilde{S}_j := \operatorname{span}\{\tilde{\psi}_{H,\lambda} : |\lambda| < j\}$ contain all polynomials of order $\tilde{m}$. Here *dual basis* means that $\langle \psi_{H,\lambda}, \tilde{\psi}_{H,\nu} \rangle = \delta_{\lambda,\nu}$, $\lambda, \nu \in \mathcal{I}_H$.

A few remarks on these properties should be made. In (R), the norm equivalence (41) is crucial since it means complete control over a function measured in $\|\cdot\|_H$ from above and below by its expansion coefficients: small changes in the coefficients only cause small changes in the function. Together with the locality (L), this also means that local changes stay local. This stability is an important feature which is used for deriving optimal preconditioners. Finally, the cancellation property (CP) entails that smooth functions have small wavelet coefficients which, on account of (40) may be neglected in a controllable way. Moreover, (CP) can be used to derive quasi–sparse representations of a wide class of operators, see the article by Rob Stevenson in this volume.

By duality arguments one can show that (40) is equivalent to the existence of a biorthogonal collection which is *dual* or *biorthogonal* to Ψ_H,

$$\tilde{\Psi}_H := \{\tilde{\psi}_{H,\lambda} : \lambda \in \mathcal{I}_H\} \subset H', \quad \langle \psi_{H,\lambda}, \tilde{\psi}_{H,\mu} \rangle = \delta_{\lambda,\mu}, \qquad \lambda, \mu \in \mathcal{I}_H, \tag{44}$$

which is a Riesz basis for H', that is, for any $\tilde{v} = \tilde{\mathbf{v}}^T \tilde{\Psi}_H \in H'$ one has

$$C_H^{-1} \|\tilde{\mathbf{v}}\| \ \le \ \|\tilde{\mathbf{v}}^T \tilde{\Psi}_H\|_{H'} \ \le \ c_H^{-1} \|\tilde{\mathbf{v}}\|, \tag{45}$$

see [D1, D3]. Here and in the sequel the tilde expresses that the collection $\tilde{\Psi}_H$ is a dual basis to a primal one for the space identified by the subscript, so that $\tilde{\Psi}_H = \Psi_{H'}$.

Above in (40), we have already introduced the following shorthand notation which simplifies the presentation of many terms. We will view Ψ_H both as in (38) as a *collection* of functions as well as a (possibly infinite) column *vector* containing all functions always assembled in some fixed unspecified order. For a countable collection of functions Θ and some single function σ, the term $\langle \Theta, \sigma \rangle$ is to be understood as the column vector with entries $\langle \theta, \sigma \rangle$, $\theta \in \Theta$, and correspondingly $\langle \sigma, \Theta \rangle$ the row vector. For two collections Θ, Σ, the quantity $\langle \Theta, \Sigma \rangle$ is then a (possibly infinite) matrix with entries $(\langle \theta, \sigma \rangle)_{\theta \in \Theta,\ \sigma \in \Sigma}$ for which $\langle \Theta, \Sigma \rangle = \langle \Sigma, \Theta \rangle^T$. This also implies for a (possibly infinite) matrix $\mathbf{C}$ that $\langle \mathbf{C}\Theta, \Sigma \rangle = \mathbf{C}\langle \Theta, \Sigma \rangle$ and $\langle \Theta, \mathbf{C}\Sigma \rangle = \langle \Theta, \Sigma \rangle \mathbf{C}^T$.

In this notation, the *biorthogonality* or *duality conditions* (44) can be expressed shortly as

$$\langle \Psi, \tilde{\Psi} \rangle = \mathbf{I} \tag{46}$$

with the infinite identity matrix $\mathbf{I}$.

Wavelets with the above properties can actually be obtained in the following way. In particular, this concerns a scaling depending on the regularity of the space under consideration. In our case, H will always be a Sobolev space $H^s = H^s(\Omega)$ or a closed subspace of $H^s(\Omega)$ determined by homogeneous boundary conditions, or its dual. For $s < 0$, H^s is interpreted as above as the dual of H^{-s}.

We typically obtain the wavelet basis Ψ_H for H from an *anchor basis* $\Psi = \{\psi_\lambda : \lambda \in I\!I = I\!I_H\}$ which is a Riesz basis for $L_2(\Omega)$, meaning that Ψ is scaled such that $\|\psi_\lambda\|_{L_2(\Omega)} \sim 1$. Moreover, its dual basis $\tilde{\Psi}$ is also a Riesz basis for $L_2(\Omega)$. Ψ and $\tilde{\Psi}$ are constructed in such a way that rescaled versions of *both bases* $\Psi, \tilde{\Psi}$ form Riesz bases for a whole range of (closed subspaces of) Sobolev spaces H^s, for $0 < s < \gamma, \tilde{\gamma}$, respectively. Consequently, one can derive that for each $s \in (-\tilde{\gamma}, \gamma)$ the collection

$$\Psi_s := \{2^{-s|\lambda|}\psi_\lambda : \lambda \in I\!I\} =: \mathbf{D}^{-s}\Psi \tag{47}$$

is a Riesz basis for H^s [D1]. This means that there exist positive finite constants c_s, C_s such that

$$c_s \|\mathbf{v}\| \ \le\ \|\mathbf{v}^T \Psi_s\|_{H^s} \ \le\ C_s \|\mathbf{v}\|, \quad \mathbf{v} \in \ell_2(I\!I), \tag{48}$$

holds for each $s \in (-\tilde{\gamma}, \gamma)$. Such a scaling represented by a diagonal matrix $\mathbf{D}^s$ introduced in (47) will play an important role later on. The analogous expression in terms of the dual basis reads

$$\tilde{\Psi}_s := \{2^{s|\lambda|}\,\tilde{\psi}_\lambda : \lambda \in I\!I\} = \mathbf{D}^s \tilde{\Psi}, \tag{49}$$

where $\tilde{\Psi}_s$ forms a Riesz basis of H^s for $s \in (-\gamma, \tilde{\gamma})$. This entails the following fact. For $t \in (-\tilde{\gamma}, \gamma)$ the mapping

$$D^t : v = \mathbf{v}^T \Psi \mapsto (\mathbf{D}^t \mathbf{v})^T \Psi = \mathbf{v}^T \mathbf{D}^t \Psi = \sum_{\lambda \in I\!I} v_\lambda\, 2^{t|\lambda|}\, \psi_\lambda \tag{50}$$

acts as a shift operator between Sobolev scales which means that

$$\|D^t v\|_{H^s} \sim \|v\|_{H^{s+t}} \sim \|\mathbf{D}^{s+t}\mathbf{v}\|, \quad \text{if } s,\ s+t \in (-\tilde{\gamma},\gamma). \tag{51}$$

Concrete constructions of wavelet bases with the above properties for parameters $\gamma,\tilde{\gamma} \le 3/2$ on a bounded Lipschitz domain Ω can be found in [DKU, DSt]. This suffices for the above mentioned examples where the relevant Sobolev regularity indices range between -1 and 1.

3.2 Norm equivalences and Riesz maps

As we have seen, the scaling provided by $\mathbf{D}^{-s}$ is an important feature to establish norm equivalences (48) for the range $s \in (-\tilde{\gamma},\gamma)$ of Sobolev spaces H^s. However, there are several other norms which are *equivalent* to $\|\cdot\|_{H^s}$ which may later be used in the objective functional (27) in the context of control problems. This issue addresses the *mathematical model* which we briefly discuss now.

We first consider norm equivalences for the L_2 norm. Let as before Ψ be the anchor wavelet basis for L_2 for which the *Riesz operator* $\mathbf{R} = \mathbf{R}_{L_2}$ is the (infinite) Gramian matrix with respect to the inner product $(\cdot,\cdot)_{L_2}$ defined as

$$\mathbf{R} := (\Psi,\Psi)_{L_2} = \langle \Psi,\Psi\rangle. \tag{52}$$

Expanding Ψ in terms of $\tilde{\Psi}$ and recalling the duality (46), this yields

$$\mathbf{I} = \langle \Psi,\tilde{\Psi}\rangle = \big\langle \langle \Psi,\Psi\rangle\tilde{\Psi},\tilde{\Psi}\big\rangle = \mathbf{R}\langle\tilde{\Psi},\tilde{\Psi}\rangle \qquad \text{or} \qquad \mathbf{R}^{-1} = \langle\tilde{\Psi},\tilde{\Psi}\rangle. \tag{53}$$

$\mathbf{R}$ may be interpreted as the transformation matrix for the change of basis from $\tilde{\Psi}$ to Ψ, that is, $\Psi = \mathbf{R}\tilde{\Psi}$. For any $w = \mathbf{w}^T\Psi \in L_2$, we now obtain the identities

$$\|w\|_{L_2}^2 = (\mathbf{w}^T\Psi,\mathbf{w}^T\Psi)_{L_2} = \mathbf{w}^T\langle\Psi,\Psi\rangle\,\mathbf{w} = \mathbf{w}^T\mathbf{R}\mathbf{w} = \|\mathbf{R}^{1/2}\mathbf{w}\|^2 =: \|\hat{\mathbf{w}}\|^2. \tag{54}$$

Expanding w with respect to the basis $\hat{\Psi} := \mathbf{R}^{-1/2}\Psi = \mathbf{R}^{1/2}\tilde{\Psi}$, that is, $w = \hat{\mathbf{w}}^T\hat{\Psi}$, yields $\|w\|_{L_2} = \|\hat{\mathbf{w}}\|$. On the other hand, we obtain from (48) with $s = 0$

$$c_0^2\,\|\mathbf{w}\|^2 \ \le\ \|w\|_{L_2}^2 \ \le\ C_0^2\,\|\mathbf{w}\|^2. \tag{55}$$

From this we can derive the *condition number* $\kappa(\Psi)$ of the wavelet basis in terms of the extreme eigenvalues of $\mathbf{R}$ by defining

$$\kappa(\Psi) := \left(\frac{C_0}{c_0}\right)^2 = \frac{\lambda_{\max}(\mathbf{R})}{\lambda_{\min}(\mathbf{R})} = \kappa(\mathbf{R}) \sim 1, \tag{56}$$

where $\kappa(\mathbf{R})$ also denotes the spectral condition number of $\mathbf{R}$ and where the last relation is assured by the asymptotic estimate (55). However, the absolute constants will have an impact on numerical results in each individual case.

For a Hilbert space H denote by Ψ_H a wavelet basis for H satisfying (R), (L), (CP) with a corresponding dual basis $\tilde{\Psi}_H$. The (infinite) Gramian matrix with respect to the inner product $(\cdot,\cdot)_H$ inducing $\|\cdot\|_H$ which is defined by

$$\mathbf{R}_H := (\Psi_H, \Psi_H)_H \tag{57}$$

will be also called *Riesz operator*. The space L_2 is covered trivially by $\mathbf{R}_0 = \mathbf{R}$. For any function $v := \mathbf{v}^T\Psi_H \in H$ we have then the identity

$$\begin{aligned}\|v\|_H^2 = (v,v)_H &= (\mathbf{v}^T\Psi_H, \mathbf{v}^T\Psi_H)_H = \mathbf{v}^T(\Psi_H,\Psi_H)_H\,\mathbf{v}\\ &= \mathbf{v}^T\mathbf{R}_H\mathbf{v} = \|\mathbf{R}_H^{1/2}\mathbf{v}\|^2.\end{aligned} \tag{58}$$

Note that in general $\mathbf{R}_H$ may not be explicitly computable, in particular, when H is a fractional Sobolev space.

Again referring to (48), we obtain as in (56) for the more general case

$$\kappa(\Psi_s) := \left(\frac{C_s}{c_s}\right)^2 = \frac{\lambda_{\max}(\mathbf{R}_{H^s})}{\lambda_{\min}(\mathbf{R}_{H^s})} = \kappa(\mathbf{R}_{H^s}) \sim 1 \quad \text{for each } s \in (-\tilde{\gamma},\gamma). \tag{59}$$

Thus, all Riesz operators on the applicable scale of Sobolev spaces are spectrally equivalent. Moreover, comparing (59) with (56), we get

$$\frac{c_s}{C_0}\,\|\mathbf{R}^{1/2}\mathbf{v}\| \;\le\; \|\mathbf{R}_{H^s}^{1/2}\mathbf{v}\| \;\le\; \frac{C_s}{c_0}\,\|\mathbf{R}^{1/2}\mathbf{v}\|. \tag{60}$$

Of course, in practice, the constants appearing in this equation may be much sharper, as the bases for Sobolev spaces with different exponents are only obtained by a diagonal scaling which preserves much of the structure of the original basis for L_2.

We summarize these results for further reference.

Proposition 3.1. *In the above notation, we have for any $v = \mathbf{v}^T\Psi_s \in H^s$ the norm equivalences*

$$\|v\|_{H^s} = \|\mathbf{R}_{H^s}^{1/2}\mathbf{v}\| \sim \|\mathbf{R}^{1/2}\mathbf{v}\| \sim \|\mathbf{v}\| \qquad \textit{for each } s \in (-\tilde{\gamma},\gamma). \tag{61}$$

3.3 Representation of operators

A final ingredient concerns the *wavelet representation* of linear operators in terms of wavelets. Let H, V be Hilbert spaces with wavelet bases Ψ_H, Ψ_V and corresponding duals $\tilde{\Psi}_H$, $\tilde{\Psi}_V$, and suppose that $\mathscr{L} : H \to V$ is a linear operator with dual $\mathscr{L}' : V' \to H'$ defined by $\langle v, \mathscr{L}'w\rangle := \langle \mathscr{L}v, w\rangle$ for all $v \in H$, $w \in V$.

We shall make frequent use of this representation and its properties.

Remark 3.1. The wavelet representation of $\mathscr{L} : H \to V$ with respect to the bases $\Psi_H, \tilde{\Psi}_V$ of H, V', respectively, is given by

$$\mathbf{L} := \langle \tilde{\Psi}_V, \mathscr{L}\Psi_H \rangle, \quad \mathscr{L}v = (\mathbf{L}\mathbf{v})^T \Psi_V. \tag{62}$$

Thus, the expansion coefficients of $\mathscr{L}v$ in the basis that spans the range space of $\mathscr{L}$ are obtained by applying the *infinite* matrix $\mathbf{L} = \langle \tilde{\Psi}_V, \mathscr{L}\Psi_H \rangle$ to the coefficient vector of v. Moreover, boundedness of $\mathscr{L}$ implies boundedness of $\mathbf{L}$ in ℓ_2, i.e.,

$$\|\mathscr{L}v\|_V \lesssim \|v\|_H, \quad v \in H, \quad \text{implies} \quad \|\mathbf{L}\| := \sup_{\|\mathbf{v}\|_{\ell_2(I\!I_H)} \le 1} \|\mathbf{L}\mathbf{v}\|_{\ell_2(I\!I_V)} \lesssim 1. \tag{63}$$

Proof. Any image $\mathscr{L}v \in V$ can naturally be expanded with respect to Ψ_V as $\mathscr{L}v = \langle \mathscr{L}v, \tilde{\Psi}_V \rangle \Psi_V$. Expanding in addition v in the basis Ψ_H, $v = \mathbf{v}^T \Psi_H$ yields

$$\mathscr{L}v = \mathbf{v}^T \langle \mathscr{L}\Psi_H, \tilde{\Psi}_V \rangle \Psi_V = (\langle \mathscr{L}\Psi_H, \tilde{\Psi}_V \rangle^T \mathbf{v})^T \Psi_V = (\langle \tilde{\Psi}_V, \mathscr{L}\Psi_H \rangle \mathbf{v})^T \Psi_V. \tag{64}$$

As for (63), we can infer from (40) and (62) that

$$\|\mathbf{L}\mathbf{v}\|_{\ell_2(I\!I_V)} \sim \|(\mathbf{L}\mathbf{v})^T \Psi_V\|_V = \|Lv\|_V \lesssim \|v\|_H \sim \|\mathbf{v}\|_{\ell_2(I\!I_H)},$$

which confirms the claim. □

3.4 Multiscale decomposition of function spaces

In this section, the basic construction principles of the biorthogonal wavelets with properties (R), (L) and (CP) are summarized, see, e.g., [D2]. Their cornerstones are *multiresolution analyses* of the function spaces under consideration and the concept of *stable completions*. These concepts are free of Fourier techniques and can therefore be applied to derive constructions of wavelets on domains or manifolds which are subsets of $\mathbb{R}^n$.

Multiresolution of L_2 (univariate case). Practical constructions of wavelets typically start out with multiresolution analyses of function spaces. Consider a *multiresolution* $\mathscr{S}$ of L_2 which consists of closed subspaces S_j of L_2, called *trial spaces*, such that they are nested and their union is dense in L_2,

$$S_{j_0} \subset S_{j_0+1} \subset \ldots \subset S_j \subset S_{j+1} \subset \ldots L_2, \qquad \operatorname{clos}_{L_2}\Big(\bigcup_{j=j_0}^{\infty} S_j \Big) = L_2. \tag{65}$$

The index j is the refinement level which appeared already in the elements of the index set $I\!I$ in (38), starting with some coarsest level $j_0 \in \mathbb{N}_0$. We abbreviate for a finite subset $\Theta \subset L_2$ the linear span of Θ as

$$S(\Theta) = \operatorname{span}\{\Theta\}.$$

Typically the multiresolution spaces S_j have the form

$$S_j = S(\Phi_j), \qquad \Phi_j = \{\phi_{j,k} : k \in \Delta_j\}, \tag{66}$$

for some finite index set Δ_j, where the set $\{\Phi_j\}_{j=j_0}^{\infty}$ is *uniformly stable* in the sense that

$$\|\mathbf{c}\|_{\ell_2(\Delta_j)} \sim \|\mathbf{c}^T \Phi_j\|_{L_2}, \qquad \mathbf{c} = \{c_k\}_{k\in\Delta_j} \in \ell_2(\Delta_j), \tag{67}$$

holds uniformly in j. Here we have used again the shorthand notation

$$\mathbf{c}^T \Phi_j = \sum_{k\in\Delta_j} c_k \phi_{j,k}$$

and Φ_j denotes both the (column) vector containing the functions $\phi_{j,k}$ as well as the set of functions (66).

The collection Φ_j is called *single scale basis* since all its elements live only on one scale j. In the present context of multiresolution analysis, Φ_j is also called *generator basis* or shortly *generators* of the multiresolution. We assume that the $\phi_{j,k}$ are compactly supported with

$$\operatorname{diam}(\operatorname{supp}\phi_{j,k}) \sim 2^{-j}. \tag{68}$$

It follows from (67) that they are scaled such that

$$\|\phi_{j,k}\|_{L_2} \sim 1 \tag{69}$$

holds. It is known that nestedness (65) together with stability (67) implies the existence of matrices $\mathbf{M}_{j,0} = (m^j_{r,k})_{r\in\Delta_{j+1},k\in\Delta_j}$ such that the two-scale relation

$$\phi_{j,k} = \sum_{r\in\Delta_{j+1}} m^j_{r,k}\phi_{j+1,r}, \quad k \in \Delta_j, \tag{70}$$

is satisfied. We can essentially simplify the subsequent presentation of the material by viewing (70) as a matrix–vector equation which then attains the compact form

$$\Phi_j = \mathbf{M}^T_{j,0}\Phi_{j+1}. \tag{71}$$

Any set of functions satisfying an equation of this form, the *refinement* or *two–scale relation*, will be called *refinable*.

Denoting by $[X,Y]$ the space of bounded linear operators from a normed linear space X into the normed linear space Y, one has that

$$\mathbf{M}_{j,0} \in [\ell_2(\Delta_j), \ell_2(\Delta_{j+1})]$$

is *uniformly sparse* which means that the number of entries in each row or column is uniformly bounded. Furthermore, one infers from (67) that

$$\|\mathbf{M}_{j,0}\| = \mathcal{O}(1), \quad j \geq j_0, \tag{72}$$

where the corresponding operator norm is defined as

$$\|\mathbf{M}_{j,0}\| := \sup_{\mathbf{c}\in\ell_2(\Delta_j),\ \|\mathbf{c}\|_{\ell_2(\Delta_j)}=1} \|\mathbf{M}_{j,0}\mathbf{c}\|_{\ell_2(\Delta_{j+1})}.$$

Since the union of $\mathscr{S}$ is dense in L_2, a basis for L_2 can be assembled from functions which span any complement between two successive spaces S_j and S_{j+1}, i.e.,

$$S(\Phi_{j+1}) = S(\Phi_j) \oplus S(\Psi_j) \tag{73}$$

where

$$\Psi_j = \{\psi_{j,k} : k \in \nabla_j\}, \qquad \nabla_j := \Delta_{j+1} \setminus \Delta_j. \tag{74}$$

The functions Ψ_j are called *wavelet functions* or shortly *wavelets* if, among other conditions detailed below, the union $\{\Phi_j \cup \Psi_j\}$ is still uniformly stable in the sense of (67). Since (73) implies $S(\Psi_j) \subset S(\Phi_{j+1})$, the functions in Ψ_j must also satisfy a matrix–vector relation of the form

$$\Psi_j = \mathbf{M}_{j,1}^T \Phi_{j+1} \tag{75}$$

with a matrix $\mathbf{M}_{j,1}$ of size $(\#\Delta_{j+1}) \times (\#\nabla_j)$. Furthermore, (73) is equivalent to the fact that the linear operator composed of $\mathbf{M}_{j,0}$ and $\mathbf{M}_{j,1}$,

$$\mathbf{M}_j = (\mathbf{M}_{j,0}, \mathbf{M}_{j,1}), \tag{76}$$

is *invertible* as a mapping from $\ell_2(\Delta_j \cup \nabla_j)$ onto $\ell_2(\Delta_{j+1})$. One can also show that the set $\{\Phi_j \cup \Psi_j\}$ is uniformly stable if and only if

$$\|\mathbf{M}_j\|, \|\mathbf{M}_j^{-1}\| = \mathscr{O}(1), \quad j \to \infty. \tag{77}$$

The particular cases that will be important for practical purposes are when not only $\mathbf{M}_{j,0}$ and $\mathbf{M}_{j,1}$ are uniformly sparse but also the inverse of $\mathbf{M}_j$. We denote this inverse by $\mathbf{G}_j$ and assume that it is split into

$$\mathbf{G}_j = \mathbf{M}_j^{-1} = \begin{pmatrix} \mathbf{G}_{j,0} \\ \mathbf{G}_{j,1} \end{pmatrix}. \tag{78}$$

A special situation occurs when $\mathbf{M}_j$ is an orthogonal matrix,

$$\mathbf{G}_j = \mathbf{M}_j^{-1} = \mathbf{M}_j^T$$

which corresponds to the case of L_2 *orthogonal wavelets* [Dau]. A systematic construction of more general $\mathbf{M}_j$, $\mathbf{G}_j$ for spline-wavelets can be found in [DKU], see also [D2] for more examples, including the hierarchical basis.

Thus, the identification of the functions Ψ_j which span the complement of $S(\Phi_j)$ in $S(\Phi_{j+1})$ is equivalent to completing a given refinement matrix $\mathbf{M}_{j,0}$ to an invertible matrix $\mathbf{M}_j$ in such a way that (77) is satisfied. Any such completion $\mathbf{M}_{j,1}$ is called *stable completion* of $\mathbf{M}_{j,0}$. In other words, the problem of the construction of compactly supported wavelets can equivalently be formulated as an algebraic prob-

lem of finding the (uniformly) sparse completion of a (uniformly) sparse matrix $\mathbf{M}_{j,0}$ in such a way that its inverse is also (uniformly) sparse. The fact that inverses of sparse matrices are usually dense elucidates the difficulties in the constructions.

The concept of stable completions has been introduced in [CDP] for which a special case is known as Sweldens' *lifting scheme*. Of course, constructions that yield compactly supported wavelets are particularly suited for computations in numerical analysis.

Combining the two–scale relations (71) and (75), one can see that $\mathbf{M}_j$ performs a change of bases in the space S_{j+1},

$$\begin{pmatrix} \Phi_j \\ \Psi_j \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{j,0}^T \\ \mathbf{M}_{j,1}^T \end{pmatrix} \Phi_{j+1} = \mathbf{M}_j^T \Phi_{j+1}. \tag{79}$$

Conversely, applying the inverse of $\mathbf{M}_j$ to both sides of (79) results in the *reconstruction identity*

$$\Phi_{j+1} = \mathbf{G}_j^T \begin{pmatrix} \Phi_j \\ \Psi_j \end{pmatrix} = \mathbf{G}_{j,0}^T \Phi_j + \mathbf{G}_{j,1}^T \Psi_j. \tag{80}$$

An example of the structure of the matrices $\mathbf{M}_j$ and $\mathbf{G}_j$ is given in Figure 1.

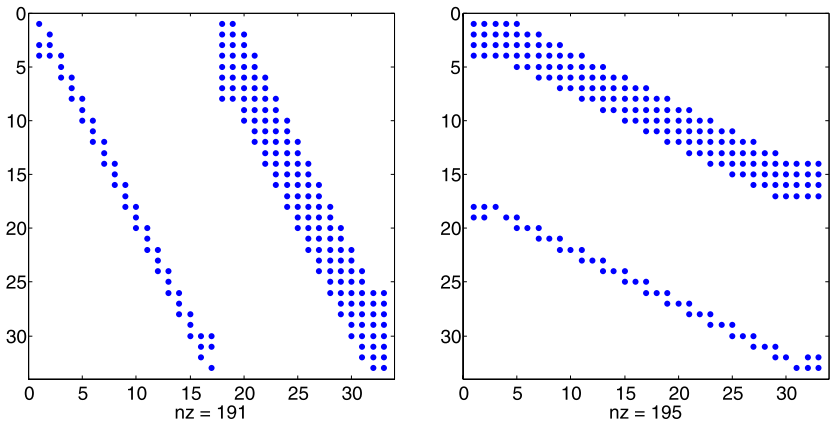


Fig. 1 Nonzero pattern of matrices $\mathbf{M}_j$ (left) and $\mathbf{G}_j$ (right) for boundary-adapted B-splines of order $d = 2$ as generators and duals of order $\tilde{d} = 4$

Fixing a *finest resolution level* J, one can repeat the decomposition (73) so that $S_J = S(\Phi_J)$ can be written in terms of the functions from the coarsest space supplied with the complement functions from all intermediate levels,

$$S(\Phi_J) = S(\Phi_{j_0}) \oplus \bigoplus_{j=j_0}^{J-1} S(\Psi_j). \tag{81}$$

Thus, every function $v \in S(\Phi_J)$ can be written in its *single–scale representation*

$$v = (\mathbf{c}_J)^T \Phi_J = \sum_{k \in \Delta_J} c_{J,k} \phi_{J,k} \tag{82}$$

as well as in its *multi–scale form*

$$v = (\mathbf{c}_{j_0})^T \Phi_{j_0} + (\mathbf{d}_{j_0})^T \Psi_{j_0} + \cdots + (\mathbf{d}_{J-1})^T \Psi_{J-1} \tag{83}$$

with respect to the *multiscale* or *wavelet basis*

$$\Psi^J := \Phi_{j_0} \cup \bigcup_{j=j_0}^{J-1} \Psi_j =: \bigcup_{j=j_0-1}^{J-1} \Psi_j \tag{84}$$

Often the single–scale representation of a function may be easier to compute and evaluate while the multi–scale representation allows one to separate features of the underlying function characterized by different length scales. Since therefore both representations are advantageous, it is useful to determine the transformation between the two representations, commonly referred to as the *Wavelet Transform*,

$$\mathbf{T}_J : \ell_2(\Delta_j) \to \ell_2(\Delta_j), \qquad \mathbf{d}^J \mapsto \mathbf{c}_J, \tag{85}$$

where

$$\mathbf{d}^J := (\mathbf{c}_{j_0}, \mathbf{d}_{j_0}, \ldots, \mathbf{d}_{J-1})^T.$$

The previous relations (79) and (80) indicate that this will involve the matrices $\mathbf{M}_j$ and $\mathbf{G}_j$. In fact, $\mathbf{T}_J$ has the representation

$$\mathbf{T}_J = \mathbf{T}_{J,J-1} \cdots \mathbf{T}_{J,j_0}, \tag{86}$$

where each factor has the form

$$\mathbf{T}_{J,j} := \begin{pmatrix} \mathbf{M}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^{(\#\Delta_J - \#\Delta_{j+1})} \end{pmatrix} \in \mathbb{R}^{(\#\Delta_J) \times (\#\Delta_J)}. \tag{87}$$

Schematically $\mathbf{T}_J$ can be visualized as a pyramid scheme

$$\begin{array}{ccccccccc} \mathbf{c}_{j_0} & \xrightarrow{\mathbf{M}_{j_0,0}} & \mathbf{c}_{j_0+1} & \xrightarrow{\mathbf{M}_{j_0+1,0}} & \mathbf{c}_{j_0+2} & \longrightarrow \cdots & \mathbf{c}_{J-1} & \xrightarrow{\mathbf{M}_{J-1,0}} & \mathbf{c}_J \\ & \nearrow_{\mathbf{M}_{j_0,1}} & & \nearrow_{\mathbf{M}_{j_0+1,1}} & & \nearrow \cdots & & \nearrow_{\mathbf{M}_{J-1,1}} & \\ \mathbf{d}_{j_0} & & \mathbf{d}_{j_0+1} & & \mathbf{d}_{j_0+2} & & \mathbf{d}_{J-1} & & \end{array} \tag{88}$$

Accordingly, the inverse transform $\mathbf{T}_J^{-1}$ can be written also in product structure (86) in reverse order involving the matrices $\mathbf{G}_j$ as follows:

$$\mathbf{T}_J^{-1} = \mathbf{T}_{J,j_0}^{-1} \cdots \mathbf{T}_{J,J-1}^{-1}, \tag{89}$$

where each factor has the form

$$\mathbf{T}_{J,j}^{-1} := \begin{pmatrix} \mathbf{G}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^{(\#\Delta_J - \#\Delta_{j+1})} \end{pmatrix} \in \mathbb{R}^{(\#\Delta_J)\times(\#\Delta_J)}. \tag{90}$$

The corresponding pyramid scheme is then

$$\begin{array}{ccccccccc} & & \mathbf{G}_{J-1,0} & & \mathbf{G}_{J-2,0} & & & & \mathbf{G}_{j_0,0} \\ \mathbf{c}_J & \longrightarrow & \mathbf{c}_{J-1} & \longrightarrow & \mathbf{c}_{J-2} & \longrightarrow & \cdots & \longrightarrow & \mathbf{c}_{j_0} \\ & & \searrow & & \searrow & & \searrow \quad \cdots \quad \searrow & & \\ & & \mathbf{d}_{J-1} & & \mathbf{d}_{J-2} & & \mathbf{d}_{J-1} & & \mathbf{d}_{j_0} \end{array} \tag{91}$$

Remark 3.2. Property (77) and the fact that $\mathbf{M}_j$ and $\mathbf{G}_j$ can be applied in $(\#\Delta_{j+1})$ operations uniformly in j entails that the complexity of applying $\mathbf{T}_J$ or $\mathbf{T}_J^{-1}$ using the pyramid scheme is of order $\mathcal{O}(\#\Delta_J) = \mathcal{O}(\dim S_J)$ uniformly in J. For this reason, $\mathbf{T}_J$ is called the *Fast Wavelet Transform* (FWT). Note that one should *not* explicitly assemble $\mathbf{T}_J$ or $\mathbf{T}_J^{-1}$. In fact, due to the particular band structure of $\mathbf{M}_j$ and $\mathbf{G}_j$, this would result in matrices with $\mathcal{O}(J\#\Delta_J)$ entries.

In Table 1 at the end of this section, spectral condition numbers for the Fast Wavelet Transform for different constructions of biorthogonal wavelets on the interval computed in [Pa] are displayed.

Since $\cup_{j\geq j_0} S_j$ is dense in L_2, a basis for the whole space L_2 is obtained when letting $J \longrightarrow \infty$ in (84),

$$\begin{aligned} \Psi &:= \bigcup_{j=j_0-1}^{\infty} \Psi_j = \{\psi_{j,k} : (j,k) \in I\!I\}, \qquad \Psi_{j_0-1} := \Phi_{j_0} \\ I\!I &:= \big\{\{j_0\} \times \Delta_{j_0}\big\} \cup \bigcup_{j=j_0}^{\infty} \big\{\{j\} \times \nabla_j\big\}. \end{aligned} \tag{92}$$

The next theorem from [D1] illustrates the relation between Ψ and $\mathbf{T}_J$.

Theorem 3.1. *The multiscale transformations $\mathbf{T}_J$ are well–conditioned in the sense*

$$\|\mathbf{T}_J\|, \|\mathbf{T}_J^{-1}\| = \mathcal{O}(1), \quad J \geq j_0, \tag{93}$$

if and only if the collection Ψ defined by (92) is a Riesz basis for L_2, i.e., every $v \in L_2$ has unique expansions

$$v = \sum_{j=j_0-1}^{\infty} \langle v, \tilde{\Psi}_j \rangle \Psi_j = \sum_{j=j_0-1}^{\infty} \langle v, \Psi_j \rangle \tilde{\Psi}_j, \tag{94}$$

where $\tilde{\Psi}$ defined analogously as in (92) is also a Riesz basis for L_2 which is biorthogonal or dual to Ψ,

$$\langle \Psi, \tilde{\Psi} \rangle = \mathbf{I} \tag{95}$$

such that

$$\|v\|_{L_2} \sim \|\langle \tilde{\Psi}, v\rangle\|_{\ell_2(I)} \sim \|\langle \Psi, v\rangle\|_{\ell_2(I)}. \tag{96}$$

We briefly explain next how the functions in $\tilde{\Psi}$, denoted as *wavelets dual to* Ψ, or *dual wavelets*, can be determined. Assume that there is a second multiresolution $\tilde{\mathcal{S}}$ of L_2 satisfying (65) where

$$\tilde{S}_j = S(\tilde{\Phi}_j), \qquad \tilde{\Phi}_j = \{\tilde{\phi}_{j,k} : k \in \Delta_j\} \tag{97}$$

and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ is uniformly stable in j in the sense of (67). Let the functions in $\tilde{\Phi}_j$ also have compact support satisfying (68). Furthermore, suppose that the biorthogonality conditions

$$\langle \Phi_j, \tilde{\Phi}_j\rangle = \mathbf{I} \tag{98}$$

hold. We will often refer to Φ_j as the *primal* and to $\tilde{\Phi}_j$ as the *dual generators*. The nestedness of the $\tilde{S}_j$ and the stability again implies that $\tilde{\Phi}_j$ is refinable with some matrix $\tilde{\mathbf{M}}_{j,0}$, similar to (71),

$$\tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^T \tilde{\Phi}_{j+1}. \tag{99}$$

The problem of determining biorthogonal wavelets now consists in finding bases $\Psi_j, \tilde{\Psi}_j$ for the complements of $S(\Phi_j)$ in $S(\Phi_{j+1})$, and of $S(\tilde{\Phi}_j)$ in $S(\tilde{\Phi}_{j+1})$, such that

$$S(\Phi_j) \perp S(\tilde{\Psi}_j), \qquad S(\tilde{\Phi}_j) \perp S(\Psi_j) \tag{100}$$

and

$$S(\Psi_j) \perp S(\tilde{\Psi}_r), \quad j \neq r, \tag{101}$$

holds. The connection between the concept of stable completions and the dual generators and wavelets is made by the following result which is a special case from [CDP].

Proposition 3.2. *Suppose that the biorthogonal collections* $\{\Phi_j\}_{j=j_0}^{\infty}$ *and* $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ *are both uniformly stable and refinable with refinement matrices* $\mathbf{M}_{j,0}$, $\tilde{\mathbf{M}}_{j,0}$, *i.e.,*

$$\Phi_j = \mathbf{M}_{j,0}^T \Phi_{j+1}, \qquad \tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^T \tilde{\Phi}_{j+1}, \tag{102}$$

and satisfy the duality condition (98). Assume that $\check{\mathbf{M}}_{j,1}$ *is any stable completion of* $\mathbf{M}_{j,0}$ *such that*

$$\check{\mathbf{M}}_j := (\mathbf{M}_{j,0}, \check{\mathbf{M}}_{j,1}) = \check{\mathbf{G}}_j^{-1} \tag{103}$$

satisfies (77).
Then

$$\mathbf{M}_{j,1} := (\mathbf{I} - \mathbf{M}_{j,0}\tilde{\mathbf{M}}_{j,0}^T)\check{\mathbf{M}}_{j,1} \tag{104}$$

is also a stable completion of $\mathbf{M}_{j,0}$, *and* $\mathbf{G}_j = \mathbf{M}_j^{-1} = (\mathbf{M}_{j,0}, \mathbf{M}_{j,1})^{-1}$ *has the form*

$$\mathbf{G}_j = \begin{pmatrix} \tilde{\mathbf{M}}_{j,0}^T \\ \check{\mathbf{G}}_{j,1} \end{pmatrix}. \tag{105}$$

Moreover, the collections of functions

$$\Psi_j := \mathbf{M}_{j,1}^T \Phi_{j+1}, \qquad \tilde{\Psi}_j := \check{\mathbf{G}}_{j,1} \tilde{\Phi}_{j+1} \tag{106}$$

form biorthogonal systems,

$$\langle \Psi_j, \tilde{\Psi}_j \rangle = \mathbf{I}, \qquad \langle \Psi_j, \tilde{\Phi}_j \rangle = \langle \Phi_j, \tilde{\Psi}_j \rangle = \mathbf{0}, \tag{107}$$

so that

$$S(\Psi_j) \perp S(\tilde{\Psi}_r), \quad j \neq r, \qquad S(\Phi_j) \perp S(\tilde{\Psi}_j), \quad S(\tilde{\Phi}_j) \perp S(\Psi_j). \tag{108}$$

In particular, the relations (98), (107) imply that the collections

$$\Psi = \bigcup_{j=j_0-1}^{\infty} \Psi_j, \qquad \tilde{\Psi} := \bigcup_{j=j_0-1}^{\infty} \tilde{\Psi}_j := \tilde{\Phi}_{j_0} \cup \bigcup_{j=j_0}^{\infty} \tilde{\Psi}_j \tag{109}$$

are biorthogonal,

$$\langle \Psi, \tilde{\Psi} \rangle = \mathbf{I}. \tag{110}$$

Remark 3.3. Note that the properties needed in addition to (110) to ensure (96) are neither properties of the complements nor of their bases $\Psi, \tilde{\Psi}$ but of the multiresolution sequences $\mathscr{S}$ and $\tilde{\mathscr{S}}$. These can be phrased as approximation and regularity properties and appear in Theorem 3.2.

We briefly recall yet another useful point of view. The operators

$$\begin{aligned} P_j v &:= \langle v, \tilde{\Phi}_j \rangle \Phi_j = \langle v, \tilde{\Psi}^j \rangle \Psi^j = \langle v, \tilde{\Phi}_{j_0} \rangle \Phi_{j_0} + \sum_{r=j_0}^{j-1} \langle v, \tilde{\Psi}_r \rangle \Psi_r \\ P'_j v &:= \langle v, \Phi_j \rangle \tilde{\Phi}_j = \langle v, \Psi^j \rangle \tilde{\Psi}^j = \langle v, \Phi_{j_0} \rangle \tilde{\Phi}_{j_0} + \sum_{r=j_0}^{j-1} \langle v, \Psi_r \rangle \tilde{\Psi}_r \end{aligned} \tag{111}$$

are projectors onto

$$S(\Phi_j) = S(\Psi^j) \qquad \text{and} \qquad S(\tilde{\Phi}_j) = S(\tilde{\Psi}^j) \tag{112}$$

respectively, which satisfy

$$P_r P_j = P_r, \quad P'_r P'_j = P'_r, \qquad r \leq j. \tag{113}$$

Remark 3.4. Let $\{\Phi_j\}_{j=j_0}^{\infty}$ be uniformly stable. The P_j defined by (111) are uniformly bounded if and only if $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ is also uniformly stable. Moreover, the P_j satisfy (113) if and only if the $\tilde{\Phi}_j$ are refinable as well. Note that then (98) implies

$$\mathbf{M}_{j,0}^T \tilde{\mathbf{M}}_{j,0} = \mathbf{I}. \tag{114}$$

In terms of the projectors, the uniform stability of the complement bases Ψ_j, $\tilde{\Psi}_j$ means that

$$\|(P_{j+1}-P_j)v\|_{L_2} \sim \|\langle \tilde{\Psi}_j, v\rangle\|_{\ell_2(\nabla_j)}, \quad \|(P'_{j+1}-P'_j)v\|_{L_2} \sim \|\langle \Psi_j, v\rangle\|_{\ell_2(\nabla_j)}, \quad (115)$$

so that the L_2 norm equivalence (96) is equivalent to

$$\|v\|_{L_2}^2 \sim \sum_{j=j_0}^{\infty} \|(P_j - P_{j-1})v\|_{L_2}^2 \sim \sum_{j=j_0}^{\infty} \|(P'_j - P'_{j-1})v\|_{L_2}^2 \quad (116)$$

for any $v \in L_2$, where $P_{j_0-1} = P'_{j_0-1} := 0$.

The whole concept derived so far lives from both Φ_j *and* $\tilde{\Phi}_j$. It should be pointed out that in the algorithms one actually does not need $\tilde{\Phi}_j$ explicitly for computations.

We recall next results that guarantee norm equivalences of the type (40) for Sobolev spaces.

Multiresolution of Sobolev spaces. Let now $\mathscr{S}$ be a multiresolution sequence consisting of closed subspaces of H^s with the property (65) whose union is dense in H^s. The following result from [D1] ensures under which conditions norm equivalences hold for the H^s–norm.

Theorem 3.2. *Let $\{\Phi_j\}_{j=j_0}^{\infty}$ and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ be uniformly stable, refinable, biorthogonal collections and let the $P_j : H^s \to S(\Phi_j)$ be defined by (111).*
If the Jackson-type estimate

$$\inf_{v_j \in S_j} \|v - v_j\|_{L_2} \lesssim 2^{-sj}\|v\|_{H^s}, \quad v \in H^s,\ 0 < s \le \bar{d}, \quad (117)$$

and the Bernstein inequality

$$\|v_j\|_{H^s} \lesssim 2^{sj}\|v_j\|_{L_2}, \quad v_j \in S_j,\ s < \bar{t}, \quad (118)$$

hold for

$$S_j = \left\{ \begin{matrix} S(\Phi_j) \\ S(\tilde{\Phi}_j) \end{matrix} \right\} \text{ with order } \bar{d} = \left\{ \begin{matrix} d \\ \tilde{d} \end{matrix} \right\} \text{ and } \bar{t} = \left\{ \begin{matrix} t \\ \tilde{t} \end{matrix} \right\}, \quad (119)$$

then for

$$0 < \sigma := \min\{d, t\}, \qquad 0 < \tilde{\sigma} := \min\{\tilde{d}, \tilde{t}\}, \quad (120)$$

one has

$$\|v\|_{H^s}^2 \sim \sum_{j=j_0}^{\infty} 2^{2sj}\|(P_j - P_{j-1})v\|_{L_2}^2, \quad s \in (-\tilde{\sigma}, \sigma). \quad (121)$$

Recall that we always write $H^s = (H^{-s})'$ for $s < 0$.

The regularity of $\mathscr{S}$ and $\tilde{\mathscr{S}}$ is characterized by

$$t := \sup\{s : S(\Phi_j) \subset H^s,\ j \ge j_0\}, \qquad \tilde{t} := \sup\{s : S(\tilde{\Phi}_j) \subset H^s,\ j \ge j_0\} \quad (122)$$

Recalling the representation (115), we can immediately derive the following fact.

Corollary 3.1. *Suppose that the assumptions in Theorem 3.2 hold. Then we have the norm equivalence*

$$\|v\|_{H^s}^2 \sim \sum_{j=j_0-1}^{\infty} 2^{2sj} \|\langle \tilde{\Psi}_j, v\rangle\|_{\ell_2(\nabla_j)}^2, \quad s \in (-\tilde{\sigma}, \sigma). \tag{123}$$

In particular for $s = 0$ the Riesz basis property of the Ψ, $\tilde{\Psi}$ relative to L_2(96) is recovered. For many applications it suffices to have (121) or (123) only for certain $s > 0$ for which one only needs to require (117) and (118) for $\{\Phi_j\}_{j=j_0}^{\infty}$. The Jackson estimates (117) of order $\tilde{d}$ for $S(\tilde{\Phi}_j)$ imply the cancellation properties (CP) (43), see, e.g., [D4].

Remark 3.5. When the wavelets live on $\Omega \subset \mathbb{R}^n$, (117) means that all polynomials up to order $\tilde{d}$ are contained in $S(\tilde{\Phi}_j)$. One also says that $S(\tilde{\Phi}_j)$ is *exact* of order $\tilde{d}$. On account of (95), this implies that the wavelets $\psi_{j,k}$ are orthogonal to polynomials up to order $\tilde{d}$ or have $\tilde{d}$th order *vanishing moments*. By Taylor expansion, this in turn yields (43).

The following generalizations of the discrete norms (116) are useful. Let for $s \in \mathbb{R}$

$$|||v|||_s := \left(\sum_{j=j_0}^{\infty} 2^{2sj} \|(P_j - P_{j-1})v\|_{L_2}^2 \right)^{1/2} \tag{124}$$

which by the relations (115) is also equivalent to

$$|v|_s := \left(\sum_{j=j_0-1}^{\infty} 2^{2sj} \|\langle \tilde{\Psi}_j, v\rangle\|_{\ell_2(\nabla_j)}^2 \right)^{1/2}. \tag{125}$$

In this notation, (121) and (123) read

$$\|v\|_{H^s} \sim |||v|||_s \sim |v|_s. \tag{126}$$

In terms of such discrete norms, Jackson and Bernstein estimates hold with constants equal to one.

Lemma 3.1. *[K1] Let $\{\Phi_j\}_{j=j_0}^{\infty}$ and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ be uniformly stable, refinable, biorthogonal collections and let the P_j be defined by (111). Then the estimates*

$$|v - P_j v|_{s'} \le 2^{-(j+1)(s-s')} |v|_s, \qquad v \in H^s, \ s' \le s \le d, \tag{127}$$

and

$$|v_j|_s \le 2^{j(s-s')} |v_j|_{s'}, \qquad v_j \in S(\Phi_j), \ s' \le s \le d, \tag{128}$$

are valid, and correspondingly for the dual side.

The same results hold for the norm $|||\cdot|||$ defined in (124).

Reverse Cauchy–Schwarz Inequalities. The biorthogonality condition (98) implies together with direct and inverse estimates the following reverse Cauchy–Schwarz inequalities for finite–dimensional spaces [DK2]. This is one essential ingredient in proving a sufficient condition for satisfying the LBB condition in Section 4.2.

Lemma 3.2. *Let the assumptions in Theorem 3.2 be valid such that the norm equivalence (121) holds for $(-\tilde\sigma,\sigma)$ with $\sigma,\tilde\sigma$ defined in (120). Then for any $v\in S(\Phi_j)$ there exists some $\tilde v^*=\tilde v^*(v)\in S(\tilde\Phi_j)$ such that*

$$\|v\|_{H^s}\,\|\tilde v^*\|_{H^{-s}} \lesssim \langle v,\tilde v^*\rangle \tag{129}$$

for any $0\le s<\min(\sigma,\tilde\sigma)$.

The proof of this result given in [DK2] for $s=1/2$ in terms of the projectors P_j defined in (111) and corresponding duals P_j' immediately carries over to more general s. Recalling the representation (112) in terms of wavelets, the reverse Cauchy inequality (129) attains the following sharp form.

Lemma 3.3. *[K1] Let the assumptions of Lemma 3.1 hold. Then for every $v\in S(\Phi_j)$ there exists some $\tilde v^*=\tilde v^*(v)\in S(\tilde\Phi_j)$ such that*

$$|v|_s\,|\tilde v^*|_{-s} = \langle v,\tilde v^*\rangle \tag{130}$$

for any $0\le s\le\min(\sigma,\tilde\sigma)$.

Proof. Every $v\in S(\Phi_j)$ can be written as

$$v=\sum_{r=j_0-1}^{j-1}2^{sr}\sum_{k\in\nabla_r}v_{r,k}\psi_{r,k}.$$

Setting now

$$\tilde v^*:=\sum_{r=j_0-1}^{j-1}2^{-sr}\sum_{k\in\nabla_r}v_{r,k}\tilde\psi_{r,k}$$

with the same coefficients $v_{j,k}$, the definition of $|\cdot|_s$ yields by biorthogonality (110)

$$|v|_s\,|\tilde v^*|_{-s} = \sum_{r=j_0-1}^{j-1}\sum_{k\in\nabla_r}|v_{j,k}|^2.$$

Combining this with the observation

$$\langle v,\tilde v^*\rangle = \sum_{r=j_0-1}^{j-1}\sum_{k\in\nabla_r}|v_{j,k}|^2$$

confirms (130). □

Remark 3.6. The previous proof reveals that the identity (130) is also true for elements from infinite–dimensional spaces H^s and $(H^s)'$ for which Ψ and $\tilde\Psi$ are Riesz bases.

Biorthogonal wavelets on $\mathbb{R}$. The construction of biorthogonal spline-wavelets on $\mathbb{R}$ from [CDF] for $L_2=L_2(\mathbb{R})$ employs the multiresolution framework introduced

at the beginning of this section. There the $\phi_{j,k}$ are generated through the dilates and translates of a single function $\phi \in L_2$,

$$\phi_{j,k} = 2^{j/2}\phi(2^j \cdot -k). \tag{131}$$

This corresponds to the idea of a *uniform* virtual underlying grid, explaining the terminology *uniform refinements*. B–Splines on uniform grids are known to satisfy refinement relations (70) in addition to being compactly supported and having L_2–stable integer translates. For computations, they have the additional advantage that they can be expressed as piecewise polynomials. In the context of variational formulations for second order boundary value problems, a well–used example are the nodal finite elements $\phi_{j,k}$ generated by the cardinal B–Spline of order two, i.e., the piecewise linear continuous function commonly called the 'hat function'. For cardinal B–Splines as generators, a whole class of dual generators $\tilde{\phi}_{j,k}$ (of arbitrary smoothness at the expense of larger supports) can be constructed which are also generated by one single function $\tilde{\phi}$ through translates and dilates. By Fourier techniques, one can construct from $\phi, \tilde{\phi}$ then a pair of biorthogonal wavelets $\psi, \tilde{\psi}$ whose dilates and translates built as in (131) constitute Riesz bases for $L_2(\mathbb{R})$.

By taking tensor products of these functions, of course, one can generate biorthogonal wavelet bases for $L_2(\mathbb{R}^n)$.

Biorthogonal wavelets on domains. Some constructions that exist by now have as a core ingredient tensor products of one-dimensional wavelets on an *interval* derived from the biorthogonal wavelets from [CDF] on $\mathbb{R}$. On finite intervals in $\mathbb{R}$, the corresponding constructions are usually based on keeping the elements of $\Phi_j, \tilde{\Phi}_j$ supported *inside* the interval while modifying those translates overlapping the end points of the interval so as to preserve a desired degree of polynomial exactness. A general detailed construction satisfying all these requirements has been proposed in [DKU]. Here just the main ideas for constructing a biorthogonal pair $\Phi_j, \tilde{\Phi}_j$ and corresponding wavelets satisfying the above requirements are sketched, where we apply the techniques derived at the beginning of this section.

We start out with those functions from two collections of biorthogonal generators $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ for some fixed $j \geq j_0$ living on the whole real line whose support has nonempty intersection with the interval $(0,1)$. In order to treat the boundary effects separately, we assumed that the coarsest resolution level j_0 is large enough so that, in view of (68), functions overlapping one end of the interval vanish at the other. One then leaves as many functions from the collection $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ living in the interior of the interval untouched and modifies only those near the interval ends. Note that keeping just the restrictions to the interval of those translates overlapping the end points would destroy stability (and also the cardinality of the primal and dual basis functions living on $(0,1)$ since their supports do not have the same size). Therefore, modifications at the end points are necessary; also, just discarding them from the collections (66), (97) would produce an error near the end points. The basic idea is essentially the same for all constructions of orthogonal and biorthogonal wavelets on $\mathbb{R}$ adapted to an interval. Namely, one takes *fixed* linear combinations of all functions in $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ living near the ends of the interval in such a way that monomials

up to the exactness order are reproduced there and such that the generator bases have the same cardinality. Because of the boundary modifications, the collections of generators are there no longer biorthogonal. However, one can show in the case of cardinal B–Splines as primal generators (which is a widely used class for numerical analysis) that biorthogonalization is indeed possible. This yields collections denoted by $\Phi_j^{(0,1)}, \tilde{\Phi}_j^{(0,1)}$ which then satisfy (98) on $(0,1)$ and all assumptions required in Proposition 3.2.

For the construction of corresponding wavelets, first an *initial* stable completion $\check{\mathbf{M}}_{j,1}$ is computed by applying Gaussian eliminations to factor $\mathbf{M}_{j,0}$ and then to find a uniformly stable inverse of $\check{\mathbf{M}}_j$. Here we exploit that for cardinal B–Splines as generators the refinement matrices $\mathbf{M}_{j,0}$ are totally positive. Thus, they can be stably decomposed by Gaussian elimination without pivoting. Application of Proposition 3.2 then gives the corresponding biorthogonal wavelets $\Psi_j^{(0,1)}, \tilde{\Psi}_j^{(0,1)}$ on $(0,1)$ which satisfy the requirements in Corollary 3.1. It turns out that these wavelets coincide in the interior of the interval again with those on all of $\mathbb{R}$ from [CDF]. An example of the primal wavelets for $d = 2$ generated by piecewise linear continuous functions is displayed in Figure 2 on the left.

After constructing these basic versions, one can then perform local transformations near the ends of the interval in order to improve the condition or L_2 stability constants, see [Bu1, Pa] for corresponding results and numerical examples.

We display spectral condition numbers for the FWT for two different constructions of biorthogonal wavelets on the interval in Table 1. The first column denotes the finest level on which the spectral condition numbers of the FWT are computed. The next column contains the numbers for the construction of biorthogonal spline-wavelets on the interval from [DKU] for the case $d = 2, \tilde{d} = 4$ while the last column displays the condition numbers for a scaled version derived in [Bu1]. We observe that the absolute numbers stay constant and low even for high levels j. We will see later in Section 4.1 how the transformation $\mathbf{T}_J$ is used for preconditioning.

j	$\kappa_2(\mathbf{T}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{T}_{\mathrm{B}})$
4	4.743e+00	4.640e+00
5	6.221e+00	6.024e+00
6	8.154e+00	6.860e+00
7	9.473e+00	7.396e+00
8	1.023e+01	7.707e+00
9	1.064e+01	7.876e+00
10	1.086e+01	7.965e+00

j	$\kappa_2(\mathbf{T}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{T}_{\mathrm{B}})$
11	1.097e+01	8.011e+00
12	1.103e+01	8.034e+00
13	1.106e+01	8.046e+00
14	1.107e+01	8.051e+00
15	1.108e+01	8.054e+00
16	1.108e+01	8.056e+00

Table 1 Computed spectral condition numbers for the Fast Wavelet Transform on $[0,1]$ for different constructions of biorthogonal wavelets on the interval [Pa]

Along these lines, also biorthogonal generators and wavelets with homogeneous (Dirichlet) boundary conditions can be constructed. Since the $\Phi_j^{(0,1)}$ are locally near the boundary monomials which all vanish at $0,1$ except for one, removing the one from $\Phi_j^{(0,1)}$ which corresponds to the constant function produces a collection of generators with homogeneous boundary conditions at $0,1$. In order for the moment conditions (43) still to hold for the Ψ_j, the dual generators have to have *complementary* boundary conditions. A corresponding construction has been carried out in [DS1] and implemented in [Bu1]. Homogeneous boundary conditions of higher order can be generated accordingly.

By taking tensor products of the wavelets on $(0,1)$, in this manner biorthogonal wavelets for Sobolev spaces on $(0,1)^n$ with or without homogeneous boundary conditions are obtained. This construction can be further extended to any other domain or manifold which is the image of a regular parametric mapping of the unit cube. Some results on the construction of wavelets on manifolds are summarized in [D3]. There are essentially two approaches. The first idea is based on domain decomposition and consists in 'glueing' generators across interelement boundaries, see, e.g., [CTU, DS2]. These approaches all have in common that the norm equivalences (123) for $H^s = H^s(\Gamma)$ can be shown to hold only for the range $-1/2 < s < 3/2$, due to the fact that duality arguments apply only for this range because of the nature of a modified inner product to which biorthogonality refers. The other approach which overcomes the above limitations on the ranges for which the norm equivalences hold has been developed in [DS3] based on previous characterizations of function spaces as Cartesian products from [CF]. The construction in [DS3] has been optimized and implemented to construct biorthogonal wavelet bases on the sphere in [KS], see the right graphic in Figure 2. More on such constructions for boundary integral operators can be found in the article by Helmut Harbrecht and Reinhold Schneider in this volume.

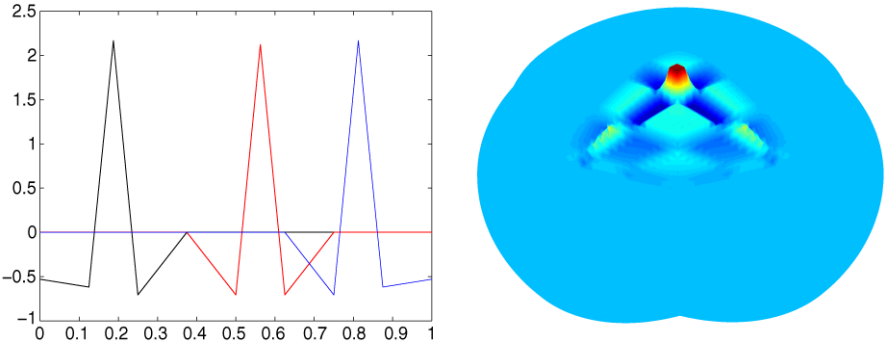


Fig. 2 Primal wavelets for $d = 2$ on $[0,1]$ (left) and on a sphere as constructed in [KS] (right)

Of course, there are also different attempts to construct wavelet bases with the above properties without using tensor products. A construction of biorthogonal spline-wavelets on triangles introduced by [Stv] has been implemented in two spa-

tial dimensions with an application to the numerical solution of a linear second order elliptic boundary value problem in [Kr].

4 Problems in wavelet coordinates

4.1 Elliptic boundary value problems

We now derive a representation of the elliptic boundary value problem from Section 2.2 in terms of (infinite) wavelet coordinates.

Let for $\mathscr{H}$ given by (8) or (9) $\Psi_{\mathscr{H}}$ be a wavelet basis with corresponding dual $\tilde{\Psi}_{\mathscr{H}}$ which satisfies the properties (R), (L) and (CP) from Section 3.1. Following the recipe from Section 3.3, expanding $y = \mathbf{y}^T \Psi_{\mathscr{H}}$, $f = \mathbf{f}^T \tilde{\Psi}_{\mathscr{H}}$ and recalling (12), the wavelet representation of the elliptic boundary value problem (14) is given by

$$\mathbf{A}\mathbf{y} = \mathbf{f} \tag{132}$$

where

$$\mathbf{A} := a(\Psi_{\mathscr{H}}, \Psi_{\mathscr{H}}), \qquad \mathbf{f} := \langle \Psi_{\mathscr{H}}, f \rangle. \tag{133}$$

Then the mapping property (13) and the Riesz basis property (R) yield the following fact.

Proposition 4.1. *The infinite matrix* $\mathbf{A}$ *is a boundedly invertible mapping from* $\ell_2 = \ell_2(\mathbb{I}_{\mathscr{H}})$ *into itself, and there exists finite positive constants* $c_{\mathbf{A}} \le C_{\mathbf{A}}$ *such that*

$$c_{\mathbf{A}} \|\mathbf{v}\| \le \|\mathbf{A}\mathbf{v}\| \le C_{\mathbf{A}} \|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2(\mathbb{I}_{\mathscr{H}}). \tag{134}$$

Proof. For any $v \in \mathscr{H}$ with coefficient vector $\mathbf{v} \in \ell_2$, we have by the lower estimates in (40), (13) and the upper inequality in (45), respectively,

$$\|\mathbf{v}\| \le c_{\mathscr{H}}^{-1} \|v\|_{\mathscr{H}} \le c_{\mathscr{H}}^{-1} c_A^{-1} \|Av\|_{\mathscr{H}'} = c_{\mathscr{H}}^{-1} c_A^{-1} \|(\mathbf{A}\mathbf{v})^T \tilde{\Psi}_{\mathscr{H}}\|_{\mathscr{H}'} \le c_{\mathscr{H}}^{-2} c_A^{-1} \|\mathbf{A}\mathbf{v}\|$$

where we have used the wavelet representation (62) for A. Likewise, the converse estimate

$$\|\mathbf{A}\mathbf{v}\| \le C_{\mathscr{H}} \|Av\|_{\mathscr{H}'} \le C_{\mathscr{H}} C_A \|v\|_{\mathscr{H}} \le C_{\mathscr{H}}^2 C_A \|\mathbf{v}\|$$

follows by the lower inequality in (45) and the upper estimates in (13) and (40). The constants appearing in (134) are therefore identified as $c_{\mathbf{A}} := c_{\mathscr{H}}^2 c_A$ and $C_{\mathbf{A}} := c_{\mathscr{H}}^2 C_A$. □

In the present situation where $\mathbf{A}$ is defined via the elliptic bilinear form $a(\cdot,\cdot)$, Proposition 4.1 entails the following result with respect to *preconditioning*. Let for $\mathbb{I} = \mathbb{I}_{\mathscr{H}}$ the symbol Λ denote *any* finite subset of the index set $\mathbb{I}$. For the corresponding set of wavelets $\Psi_\Lambda := \{\psi_\lambda : \lambda \in \Lambda\}$ denote by $S_\Lambda := \operatorname{span}\Psi_\Lambda$ the respective finite-dimensional subspace of $\mathscr{H}$. For the wavelet representation of A in terms of Ψ_Λ,

$$\mathbf{A}_\Lambda := a(\Psi_\Lambda, \Psi_\Lambda), \tag{135}$$

we obtain the following result.

Proposition 4.2. *If $a(\cdot,\cdot)$ is $\mathscr{H}$-elliptic according to (11), the finite matrix $\mathbf{A}_\Lambda$ is symmetric positive definite and its spectral condition number is bounded uniformly in Λ, i.e.,*

$$\kappa_2(\mathbf{A}_\Lambda) \le \frac{C_{\mathbf{A}}}{c_{\mathbf{A}}}, \tag{136}$$

where $c_{\mathbf{A}}, C_{\mathbf{A}}$ are the constants from (134).

Proof. Clearly, since $\mathbf{A}_\Lambda$ is just a finite section of $\mathbf{A}$, we have $\|\mathbf{A}_\Lambda\| \le \|\mathbf{A}\|$. On the other hand, by assumption, $a(\cdot,\cdot)$ is $\mathscr{H}$-elliptic which entails that $a(\cdot,\cdot)$ is also elliptic on every finite subspace $S_\Lambda \subset \mathscr{H}$. Thus, we infer $\|\mathbf{A}_\Lambda^{-1}\| \le \|\mathbf{A}^{-1}\|$, and we have

$$c_{\mathbf{A}}\|\mathbf{v}_\Lambda\| \le \|\mathbf{A}_\Lambda \mathbf{v}_\Lambda\| \le C_{\mathbf{A}}\|\mathbf{v}_\Lambda\|, \qquad \mathbf{v}_\Lambda \in S_\Lambda. \tag{137}$$

Together with the definition $\kappa_2(\mathbf{A}_\Lambda) := \|\mathbf{A}_\Lambda\|\,\|\mathbf{A}_\Lambda^{-1}\|$ we obtain the claimed estimate. □

In other words, representations of A with respect to properly scaled wavelet bases for $\mathscr{H}$ entail well-conditioned system matrices $\mathbf{A}_\Lambda$ independent of Λ. This in turn means that the convergence speed of an iterative solver applied to the corresponding finite system

$$\mathbf{A}_\Lambda \mathbf{y}_\Lambda = \mathbf{f}_\Lambda \tag{138}$$

does not deteriorate as $|\Lambda| \to \infty$.

In summary, ellipticity implies stability of the Galerkin discretizations for any set $\Lambda \subset \mathbb{I}$. This is not automatically the case for any finite versions of the saddle point problems, as we will see in Section 4.2.

Fast wavelet transform. We briefly summarize how in the situation of uniform refinements, i.e., when $S(\Phi_J) = S(\Psi^J)$, the Fast Wavelet Transformation (FWT) $\mathbf{T}_J$ can be used for preconditioning linear elliptic operators, together with a a diagonal scaling induced by the norm equivalence (123) [DK1]. We recall the notation from Section 3.4 where the wavelet basis is in fact the (unscaled) anchor basis from Section 3.1. Thus, the norm equivalence (40) using the scaled wavelet basis Ψ_H is the same as (123) in the anchor basis. Recall that the norm equivalence (123) implies that every $v \in H^s$ can be expanded uniquely in terms of the Ψ and its expansion coefficients $\mathbf{v}$ satisfy

$$\|v\|_{H^s} \sim \|\mathbf{D}^s \mathbf{v}\|_{\ell_2}$$

where $\mathbf{D}^s$ is a diagonal matrix with entries $\mathbf{D}^s_{(j,k),(j',k')} = 2^{sj}\delta_{j,j'}\delta_{k,k'}$. For $\mathscr{H} \subset H^1(\Omega)$, the case $s = 1$ is relevant.

In a stable Galerkin scheme for (10) with respect to $S(\Psi^J) = S(\Psi_\Lambda)$, we have therefore already identified the diagonal (scaling) matrix $\mathbf{D}_J$ consisting of the finite portion of the matrix $\mathbf{D} = \mathbf{D}^1$ for which $j_0 - 1 \le j \le J - 1$. The representation of A with respect to the (unscaled) wavelet basis Ψ^J can be expressed in terms of the Fast Wavelet Transform $\mathbf{T}_J$, that is,

$$\langle \Psi^J, A\Psi^J \rangle \;=\; \mathbf{T}_J^T \, \langle \Phi_J, A\Phi_J \rangle \, \mathbf{T}_J, \tag{139}$$

where Φ_J is the single–scale basis for $S(\Psi^J)$. Thus, we first set up the operator equation as in finite element settings in terms of the single–scale basis Φ_J. Applying the Fast Wavelet Transform $\mathbf{T}_J$ together with $\mathbf{D}_J$ yields that the operator

$$\mathbf{A}_J \;:=\; \mathbf{D}_J^{-1} \mathbf{T}_J^T \, \langle \Phi_J, A\Phi_J \rangle \, \mathbf{T}_J \mathbf{D}_J^{-1} \tag{140}$$

has uniformly bounded condition numbers independent of J. This can be seen by combining the properties of A according to (13) with the norm equivalences (40) and (45).

It is known that the boundary adaptations of the generators and wavelets aggravate the absolute values of the condition numbers. Nevertheless, these constants can be substantially reduced by an operator-adapted transformation which takes into account only the coarsest discretization level and, thus, is inexpensive [Bu1]. Numerical tests confirm that the absolute constants can further be improved by taking instead of $\mathbf{D}_J^{-1}$ the inverse of the diagonal of $\langle \Psi^J, A\Psi^J \rangle$ for the scaling in (140) [Bu1, Pa].

In Table 2 we display the condition numbers for discretizations using the weak form of the elliptic operator $-\Delta + \mathrm{id}$ on $(0,1)^n$ in up to three dimensions using boundary adapted biorthogonal spline-wavelets in the case $d = 2$, $\tilde{d} = 4$ with such a scaling and additional shifts of small eigenvalues which is an inexpensive operation [Bu1].

j	$n=1$	$n=2$	$n=3$
3	22.3	9.6	18.3
4	23.9	11.8	37.1
5	25.0	14.3	39.8
6	25.7	16.0	40.9
8	26.6	18.4	
10	27.1		
12	27.3		

Table 2 Optimized spectral condition numbers of the operator $\mathbf{A}$ using tensor products of biorthogonal wavelets on the interval for space dimensions $n = 1,2,3$ [Bu1]

4.2 Saddle point problems

As in the previous situation, we derive a representation of the saddle point problem introduced in Section 2.3 in terms of (infinite) wavelet coordinates.

Let for $\mathscr{H} = Y \times Q$ with $Y = H^1(\Omega)$, $Q = (H^{1/2}(\Gamma))'$ two collections of wavelet bases Ψ_Y, Ψ_Q be available, each satisfying (R), (L) and (CP), with respective duals

$\tilde{\Psi}_Y$, $\tilde{\Psi}_Q$. Like before, we expand $y = \mathbf{y}^T \Psi_Y$ and $p = \mathbf{p}^T \Psi_Q$ and test with the elements from Ψ_Y, Ψ_Q. Then (21) attains the form

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} := \begin{pmatrix}\mathbf{A} & \mathbf{B}^T\\ \mathbf{B} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}, \tag{141}$$

where

$$\begin{aligned}\mathbf{A} &:= \langle \Psi_Y, A\Psi_Y\rangle \qquad & \mathbf{f} &:= \langle \Psi_Y, f\rangle,\\ \mathbf{B} &:= \langle \Psi_Q, B\Psi_Y\rangle, & \mathbf{g} &:= \langle \Psi_Q, g\rangle.\end{aligned} \tag{142}$$

In view of the above assertions, the operator $\mathbf{L}$ is an ℓ_2–automorphism, i.e., for every $(\mathbf{v},\mathbf{q}) \in \ell_2(\mathcal{I}) = \ell_2(\mathcal{I}_Y \times \mathcal{I}_Q)$ we have

$$c_{\mathbf{L}} \left\|\begin{pmatrix}\mathbf{v}\\ \mathbf{q}\end{pmatrix}\right\| \le \left\|\mathbf{L}\begin{pmatrix}\mathbf{v}\\ \mathbf{q}\end{pmatrix}\right\| \le C_{\mathbf{L}} \left\|\begin{pmatrix}\mathbf{v}\\ \mathbf{q}\end{pmatrix}\right\| \tag{143}$$

with constants $c_{\mathbf{L}}, C_{\mathbf{L}}$ only depending on $c_{\mathscr{L}}, C_{\mathscr{L}}$ from (26) and the constants in the norm equivalences (40) and (45).

For saddle point problems with an operator $\mathbf{L}$ satisfying (143), finite sections are in general not uniformly stable in the sense of (137). In fact, for discretizations on uniform grids, the validity of the corresponding mapping property relies on a suitable stability condition, see,e.g., [BF] or the article by Ricardo Nochetto and coauthors in this volume. Corresponding results derived in [DK2] are as follows.

The bilinear form $a(\cdot,\cdot)$ defined in (7) is for $c > 0$ elliptic on all of $Y = H^1(\Omega)$ and, hence, also on any finite–dimensional subspace of Y. Let there be two multiresolution analyses $\mathscr{Y}$ of $H^1(\Omega)$ and $\mathscr{Q}$ of Q where the discrete spaces are $Y_j \subset H^1(\Omega)$ and $Q_\Lambda =: Q_\ell \subset (H^{1/2}(\Gamma))'$. With the notation from Section 3.4 and in addition superscripts referring to the domain on which the functions live, these spaces are represented by

$$\begin{aligned}Y_j &= S(\Phi_j^\Omega) = S(\Psi^{j,\Omega}), \qquad & \tilde{Y}_j &= S(\tilde{\Phi}_j^\Omega) = S(\tilde{\Psi}^{j,\Omega}),\\ Q_\ell &= S(\Phi_\ell^\Gamma) = S(\Psi^{\ell,\Gamma}), & \tilde{Q}_\ell &= S(\tilde{\Phi}_\ell^\Gamma) = S(\tilde{\Psi}^{\ell,\Gamma}).\end{aligned} \tag{144}$$

Here the indices j and ℓ refer to mesh sizes on the domain and the boundary,

$$h_\Omega \sim 2^{-j} \qquad \text{and} \qquad h_\Gamma \sim 2^{-\ell}.$$

The discrete inf–sup condition, the *LBB condition*, for the pair Y_j, Q_ℓ requires that there exists a constant $\beta_1 > 0$ *independent* of j and ℓ such that

$$\inf_{q\in Q_\ell} \sup_{v\in Y_j} \frac{b(v,q)}{\|v\|_{H^1(\Omega)}\,\|q\|_{(H^{1/2}(\Gamma))'}} \ge \beta_1 > 0 \tag{145}$$

holds. We have investigated in [DK2] the general case in arbitrary spatial dimensions where the Q_ℓ are *not* trace spaces of Y_j. Employing the reverse Cauchy-Schwarz inequalities from Section 3.4, one can show that (145) is satisfied provided

that $h_\Gamma(h_\Omega)^{-1} = 2^{j-\ell} \geq c_\Omega > 1$. This is similar to the condition which was known for bivariate polygons and particular finite elements [Ba]. Although the theoretical estimates are quite pessimistic, numerical experiments for a circular boundary within a square show that the spectral condition numbers of $\mathbf{B}\mathbf{B}^T$ are still well-behaved even when this sufficient condition is violated.

It should be mentioned that the obstructions caused by the LBB condition can be avoided by means of stabilization techniques proposed, where, however, the location of the boundary of Ω relative to the mesh is somewhat constrained. A related approach which systematically avoids restrictions of the LBB type is based on least squares techniques [DKS].

It is particularly noteworthy that adaptive schemes based on wavelets can be designed in such a way that the LBB condition is *automatically* enforced. This was first observed in [DDU]. More on this subject can be found in [D4].

In order to get an impression of the value of the constants for the condition numbers for $\mathbf{A}$ in (136) and the corresponding ones for the saddle point operator on uniform grids (143), an example with $\Omega = (0,1)^2$ and Γ chosen as one face of its boundary was implemented and investigated in [Pa]. In Table 3, the spectral condition numbers of $\mathbf{A}$ and $\mathbf{L}$ with respect to two different constructions of wavelets for the case $d = 2$ and $\tilde{d} = 4$ are displayed. We see next to the first column in which the refinement level j is listed the spectral condition numbers of $\mathbf{A}$ with the wavelet construction from [DKU] denoted by $\mathbf{A}_{\mathrm{DKU}}$ and with the modification introduced in [Bu1] and a further transformation [Pa] denoted by $\mathbf{A}_{\mathrm{B}}$. The last columns contain the respective numbers for the saddle point matrix $\mathbf{L}$ where $\kappa_2(\mathbf{L}) := \sqrt{\kappa(\mathbf{L}^T\mathbf{L})}$. We observe that the spectral condition numbers stay uniformly bounded and small as j increases.

j	$\kappa_2(\mathbf{A}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{A}_{\mathrm{B}})$	$\kappa_2(\mathbf{L}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{L}_{\mathrm{B}})$
3	5.195e+02	1.898e+01	1.581e+02	4.147e+01
4	6.271e+02	1.066e+02	1.903e+02	1.050e+02
5	6.522e+02	1.423e+02	1.997e+02	1.399e+02
6	6.830e+02	1.820e+02	2.112e+02	1.806e+02
7	7.037e+02	2.162e+02	2.318e+02	2.145e+02
8	7.205e+02	2.457e+02	2.530e+02	2.431e+02
9	7.336e+02	2.679e+02	2.706e+02	2.652e+02

Table 3 Spectral condition numbers of the operators $\mathbf{A}$ and $\mathbf{L}$ on $\Omega = (0,1)^2$ for different constructions of biorthogonal wavelets on the interval [Pa]

4.3 Control problems: Distributed control

After these preparations, we can now discuss appropriate wavelet formulations for PDE-constrained control problems with distributed control as introduced in Section 2.4. Let for $\mathscr{V} \in \{H, \mathscr{Z}, \mathscr{U}\}$ $\Psi_{\mathscr{V}}$ denote a wavelet basis with the properties (R), (L), (CP) for $\mathscr{V}$ with dual basis $\tilde{\Psi}_{\mathscr{V}}$.

Let $\mathscr{Z}, \mathscr{U}$ satisfy the embedding (29). In terms of wavelet bases, the corresponding canonical injections correspond in view of (47) to a multiplication by a diagonal matrix. That is, let $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ be such that

$$\Psi_{\mathscr{Z}} = \mathbf{D}_{\mathscr{Z}} \Psi_H, \quad \tilde{\Psi}_H = \mathbf{D}_H \Psi_{\mathscr{U}}. \tag{146}$$

Since $\mathscr{Z}$ possibly induces a weaker and $\mathscr{U}$ a stronger topology, the diagonal matrices $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ are such that their entries are nondecreasing in scale, and there is a finite constant C such that

$$\|\mathbf{D}_{\mathscr{Z}}^{-1}\|, \|\mathbf{D}_H^{-1}\| \le C. \tag{147}$$

For instance, for $H = H^\alpha, \mathscr{Z} = H^\beta$, or for $H' = H^{-\alpha}$, $\mathscr{U} = H^{-\beta}$, $0 \le \beta \le \alpha$, $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ have entries $(\mathbf{D}_{\mathscr{Z}})_{\lambda,\lambda} = (\mathbf{D}_H)_{\lambda,\lambda} = (\mathbf{D}^{\alpha-\beta})_{\lambda,\lambda} = 2^{(\alpha-\beta)|\lambda|}$.

We expand y in Ψ_H and u in a wavelet basis $\Psi_{\mathscr{U}}$ for $\mathscr{U} \subset H'$,

$$u = \mathbf{u}^T \Psi_U = (\mathbf{D}_H^{-1}\mathbf{u})^T \Psi_{H'}. \tag{148}$$

Following the derivation in Section 4.1, the linear constraints (28) attain the form

$$\mathbf{A}\mathbf{y} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u} \tag{149}$$

where

$$\mathbf{A} := a(\Psi_H, \Psi_H), \qquad \mathbf{f} := \langle \Psi_H, f \rangle. \tag{150}$$

Recall that $\mathbf{A}$ has been assumed to be symmetric. The objective functional (33) is stated in terms of the norms $\|\cdot\|_{\mathscr{Z}}$ and $\|\cdot\|_{\mathscr{U}}$. For an exact representation of these norms, corresponding Riesz operators $\mathbf{R}_{\mathscr{Z}}$ and $\mathbf{R}_{\mathscr{U}}$ defined analogously to (57) would come into play which may not be explicitly computable if $\mathscr{Z}, \mathscr{U}$ are fractional Sobolev spaces. On the other hand, as mentioned before, such a cost functional in many cases serves the purpose of yielding unique solutions while there is some ambiguity in its exact formulation. Hence, in search for a formulation which best supports numerical realizations, it is often sufficient to employ norms which are *equivalent* to $\|\cdot\|_{\mathscr{Z}}$ and $\|\cdot\|_{\mathscr{U}}$. In view of the discussion in Section 3.2, we can work for the norms $\|\cdot\|_{\mathscr{Z}}$, $\|\cdot\|_{\mathscr{U}}$ only with the diagonal scaling matrices $\mathbf{D}^s$ induced by the regularity of $\mathscr{Z}, \mathscr{U}$, or we can in addition include the Riesz map $\mathbf{R}$ defined in (52). In the numerical studies in [Bu1], a somewhat better quality of the solution is observed when $\mathbf{R}$ is included. In order to keep track of the appearance of the Riesz maps in the linear systems derived below, we choose here the latter variant.

Moreover, we expand the given observation function $y_* \in \mathscr{Z}$ as

$$y_* = \langle y_*, \tilde{\Psi}_{\mathscr{Z}} \rangle \Psi_{\mathscr{Z}} =: (\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_*)^T \Psi_{\mathscr{Z}} = \mathbf{y}_*^T \Psi_H. \tag{151}$$

The way the vector $\mathbf{y}_*$ is defined here for notational convenience may by itself actually have infinite norm in ℓ_2. However, its occurrence will always include premultiplication by $\mathbf{D}_{\mathscr{Z}}^{-1}$ which is therefore always well–defined. In view of (61), we obtain the relations

$$\|y - y_*\|_{\mathscr{Z}} \sim \|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)\| \sim \|\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*\|. \tag{152}$$

Note that here $\mathbf{R} = \langle \Psi, \Psi \rangle$ (and not $\mathbf{R}^{-1}$) comes into play since y, y_* have been expanded in a scaled version of the primal wavelet basis Ψ. Hence, equivalent norms for $\|\cdot\|_{\mathscr{Z}}$ may involve $\mathbf{R}$. As for describing equivalent norms for $\|\cdot\|_{\mathscr{U}}$, recall that u is expanded in the basis Ψ_U for $U \subset H'$. Consequently, $\mathbf{R}^{-1}$ is the natural matrix to take into account when considering equivalent norms, i.e., we choose here

$$\|u\|_{\mathscr{U}} \sim \|\mathbf{R}^{-1/2}\mathbf{u}\|. \tag{153}$$

Finally, we formulate the following control problem in (infinite) wavelet coordinates.

(DCP) *For given data* $\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_* \in \ell_2(\mathbb{I}_{\mathscr{Z}})$, $\mathbf{f} \in \ell_2(\mathbb{I}_H)$, *and weight parameter* $\omega > 0$, *minimize the quadratic functional*

$$\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) := \tfrac{1}{2}\,\|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2 \tag{154}$$

over $(\mathbf{y}, \mathbf{u}) \in \ell_2(\mathbb{I}_H) \times \ell_2(\mathbb{I}_H)$ *subject to the linear constraints*

$$\mathbf{A}\mathbf{y} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u}. \tag{155}$$

Remark 4.1. Problem (DCP) can be viewed as (discretized yet still infinite–dimensional) *representation* of the linear–quadratic control problem (27) together with (28) in wavelet coordinates in the following sense. The functional $\check{\mathbf{J}}(\mathbf{y}, \mathbf{u})$ defined in (154) is equivalent to the functional $\mathscr{J}(y, u)$ from (27) in the sense that there exist constants $0 < c_J \le C_J < \infty$ such that

$$c_J\,\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \;\le\; \mathscr{J}(y, u) \;\le\; C_J\,\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \tag{156}$$

holds for any $y = \mathbf{y}^T\Psi_H \in H$, given $y_* = (\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_*)^T\Psi_{\mathscr{Z}} \in \mathscr{Z}$ and any $u = \mathbf{u}^T\Psi_{\mathscr{U}} \in \mathscr{U}$. Moreover, in the case of compatible data $y_* = A^{-1}f$ yielding $\mathscr{J}(y, u) \equiv 0$, the respective minimizers coincide, and $\mathbf{y}_* = \mathbf{A}^{-1}\mathbf{f}$ yields $\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \equiv \mathbf{0}$. In this sense the new functional (154) captures the essential features of the model minimization functional.

Once problem (DCP) is posed, we can apply variational principles to derive necessary and sufficient conditions for a unique solution. All control problems considered here are in fact simple in this regard, as we have to minimize a quadratic functional subject to linear constraints, for which the first order necessary conditions are also sufficient. In principle, there are two ways to derive the optimality conditions for (DCP). We have encountered in Section 2.4 already the technique via the Lagrangian.

We define for (DCP) the *Lagrangian* introducing the *Lagrange multiplier*, *adjoint variable* or *adjoint state* $\mathbf{p}$ as

$$\mathbf{Lagr}(\mathbf{y},\mathbf{p},\mathbf{u}) := \check{\mathbf{J}}(\mathbf{y},\mathbf{u}) + \langle \mathbf{p}, \mathbf{A}\mathbf{y} - \mathbf{f} - \mathbf{D}_H^{-1}\mathbf{u}\rangle. \tag{157}$$

Then the KKT conditions $\delta\,\mathbf{Lagr}(\mathbf{w}) = \mathbf{0}$ for $\mathbf{w} = \mathbf{p},\mathbf{y},\mathbf{u}$ are, respectively,

$$\mathbf{A}\mathbf{y} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u}, \tag{158a}$$

$$\mathbf{A}^T\mathbf{p} = -\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*) \tag{158b}$$

$$\omega\mathbf{R}^{-1}\mathbf{u} = \mathbf{D}_H^{-1}\mathbf{p}. \tag{158c}$$

The first system resulting from the variation with respect to the Lagrange multiplier always recovers the original constraints (155) and will be referred to as the *primal system* or the *state equation*. Accordingly, we call (158b) the *adjoint* or *dual system*, or the *costate equation*. The third equation (158c) is sometimes denoted as the *design equation*. Although $\mathbf{A}$ is symmetric, we continue to write $\mathbf{A}^T$ for the operator of the adjoint system to distinguish it from the primal system.

The coupled system (158) is to be solved later. However, in order to derive convergent iterations and deduce complexity estimates, a different formulation will be advantageous. It is based on the fact that $\mathbf{A}$ is according to Proposition 4.1 a boundedly invertible mapping on ℓ_2. Thus, we can formally invert (149) to obtain $\mathbf{y} = \mathbf{A}^{-1}\mathbf{f} + \mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u}$. Substitution into (154) yields a functional depending only on $\mathbf{u}$,

$$\mathbf{J}(\mathbf{u}) := \tfrac{1}{2}\,\|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\left(\mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u} - (\mathbf{y}_* - \mathbf{A}^{-1}\mathbf{f})\right)\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2. \tag{159}$$

Employing the abbreviations

$$\mathbf{Z} := \mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{A}^{-1}\mathbf{D}_H^{-1}, \tag{160a}$$

$$\mathbf{G} := -\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f} - \mathbf{y}_*), \tag{160b}$$

the functional simplifies to

$$\mathbf{J}(\mathbf{u}) = \tfrac{1}{2}\|\mathbf{Z}\mathbf{u} - \mathbf{G}\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2. \tag{161}$$

Proposition 4.3. *[K3] The functional* $\mathbf{J}$ *is twice differentiable with first and second variation*

$$\delta\mathbf{J}(\mathbf{u}) = (\mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1})\mathbf{u} - \mathbf{Z}^T\mathbf{G}, \qquad \delta^2\mathbf{J}(\mathbf{u}) = \mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1}. \tag{162}$$

In particular, $\mathbf{J}$ *is convex so that a unique minimizer exists.*

Setting

$$\mathbf{Q} := \mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1}, \qquad \mathbf{g} := \mathbf{Z}^T\mathbf{G}, \tag{163}$$

the unique minimizer $\mathbf{u}$ of (161) is given by solving

$$\delta \mathbf{J}(\mathbf{u}) = \mathbf{0} \tag{164}$$

or, equivalently, the system

$$\mathbf{Q}\mathbf{u} = \mathbf{g}. \tag{165}$$

By definition (163), $\mathbf{Q}$ is a symmetric positive definite (infinite) matrix. Hence, finite versions of (165) could be solved by gradient or conjugate gradient iterative schemes. As the convergence speed of any such iteration depends on the spectral condition number of $\mathbf{Q}$, it is important to note that the following result.

Proposition 4.4. *The (infinite) matrix* $\mathbf{Q}$ *is uniformly bounded on* ℓ_2*, i.e., there exist constants* $0 < c_{\mathbf{Q}} \le C_{\mathbf{Q}} < \infty$ *such that*

$$c_{\mathbf{Q}} \|\mathbf{v}\| \le \|\mathbf{Q}\mathbf{v}\| \le C_{\mathbf{Q}} \|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2. \tag{166}$$

The proof follows from (13) and (147) [DK3]. Of course, in order to make such iterative schemes for (165) practically feasible, the explicit inversion of $\mathbf{A}$ in the definition of $\mathbf{Q}$ has to be avoided and replaced by an iterative solver in turn. This is where the system (158) will come into play. In particular, the third equation (158c) has the following interpretation which will turn out to be very useful later.

Proposition 4.5. *If we solve for a given control vector* $\mathbf{u}$ *successively (155) for* $\mathbf{y}$ *and (158b) for* $\mathbf{p}$*, then the* residual *for (165) attains the form*

$$\mathbf{Q}\mathbf{u} - \mathbf{g} = \omega \mathbf{R}^{-1}\mathbf{u} - \mathbf{D}_U^{-1}\mathbf{p}. \tag{167}$$

Proof. Solving consecutively (155) and (158b) and recalling the definitions of $\mathbf{Z}$, $\mathbf{g}$ (160a), (163) we obtain

$$\begin{aligned}
\mathbf{D}_H^{-1}\mathbf{p} &= -\mathbf{D}_H^{-1}(\mathbf{A}^{-T}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y}-\mathbf{y}_*)) \\
&= -\mathbf{Z}^T\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f}+\mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u}-\mathbf{y}_*) \\
&= \mathbf{Z}^T\mathbf{G} - \mathbf{Z}^T\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u} \\
&= \mathbf{g} - \mathbf{Z}^T\mathbf{Z}\mathbf{u}.
\end{aligned}$$

Hence, the residual $\mathbf{Q}\mathbf{u} - \mathbf{g}$ attains the form

$$\mathbf{Q}\mathbf{u} - \mathbf{g} = (\mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1})\mathbf{u} - \mathbf{g} = \omega\mathbf{R}^{-1}\mathbf{u} - \mathbf{D}_H^{-1}\mathbf{p},$$

where we have used the definition of $\mathbf{Q}$ from (163). □

Having derived the optimality conditions (158), the next issue is their efficient numerical solution. In view of the fact that the system (158) still involves infinite matrices and vectors, this also raises the question how to derive computable finite versions. By now we have investigated two scenarios.

The first version with respect to *uniform discretizations* is based on choosing finite–dimensional subspaces of the function spaces under consideration. The second version which deals with *adaptive discretizations* is actually based on the infi-

nite system (158). In both scenarios, a fully iterative numerical scheme for the solution of (158) can be designed along the following lines. The basic iteration scheme is a *gradient* or *conjugate gradient iteration* for (165) as an *outer iteration* where each application of $\mathbf{Q}$ is in turn realized by solving the primal and the dual system (155) and (158b) also by a gradient or conjugate gradient method as *inner iterations*.

For *uniform* discretizations for which we wanted to test numerically the role of equivalent norms and the influence of Riesz maps in the cost functional (154), we have used in [BK] as central iterative scheme the conjugate gradient (CG) method. Since the interior systems are only solved up to discretization error accuracy, the whole procedure may therefore be viewed as an *inexact conjugate gradient (CG) method*. We stress already at this point that the iteration numbers of such a method do *not* depend on the discretization level as finite versions of all involved operators are also uniformly well–conditioned in the sense of (166). In each step of the outer iteration, the error will be reduced by a fixed factor ρ. Combined with a *nested iteration strategy*, it will be shown that this yields an asymptotically optimal method in the amount of arithmetic operations.

Starting from the infinite coupled system (158), we have investigated in [DK3] *adaptive schemes* which, given any prescribed accuracy $\varepsilon > 0$, solve (158) such that the error for $\mathbf{y}, \mathbf{u}, \mathbf{p}$ is controlled by ε. There we have used for a simpler analysis a *gradient scheme* as basic iterative scheme.

4.4 Control problems: Dirichlet boundary control

Having derived a representation in wavelet coordinates for both the saddle point problem from Section 2.3 and the PDE-constrained control problem in the previous section, an appropriate representation of the control problem with Dirichlet boundary control introduced in Section 2.5 is straightforward. In order not to be overburdened with notation, we specifically choose the control space on the boundary as $\mathscr{U} := Q(= (H^{1/2}(\Gamma))')$. For the more general situation covered by (37), a diagonal matrix with nondecreasing entries like in (146) would come into play to switch between $\mathscr{U}$ and Q. Thus, the exact wavelet representation of the constraints (36) is given by the system (141), where we exchange the given Dirichlet boundary term $\mathbf{g}$ by $\mathbf{u}$ in the present situation to express the dependence on the control in the right hand side, i.e.,

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} := \begin{pmatrix}\mathbf{A} & \mathbf{B}^T\\ \mathbf{B} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix}. \tag{168}$$

The derivation of a representer of the initial objective functional (35) is under the embedding condition (37) $\|v\|_{\mathscr{Z}} \lesssim \|v\|_Y$ for $v \in Y$ now the same as in the previous section, where all reference to the space H is to be exchanged by reference to Y. We end up with the following minimization problem in wavelet coordinates for the case of Dirichlet boundary control.

(DCP) *For given data* $\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_* \in \ell_2(I_{\mathscr{Z}})$, $\mathbf{f} \in \ell_2(I_Y)$, *and weight parameter* $\omega > 0$, *minimize the quadratic functional*

$$\check{\mathbf{J}}(\mathbf{y},\mathbf{u}) := \tfrac{1}{2}\,\|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y}-\mathbf{y}_*)\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2 \tag{169}$$

over $(\mathbf{y},\mathbf{u}) \in \ell_2(I_Y) \times \ell_2(I_Y)$ *subject to the linear constraints (168),*

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix}.$$

The corresponding Karush-Kuhn-Tucker conditions can be derived by the same variational principles as in the previous section by defining a Lagrangian in terms of the functional $\check{\mathbf{J}}(\mathbf{y},\mathbf{u})$ and appending the constraints (149) with the help of additional Lagrange multipliers $(\mathbf{z},\mu)^T$, see [K3]. We obtain in this case a system of coupled saddle point problems

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix} \tag{170a}$$

$$\mathbf{L}^T\begin{pmatrix}\mathbf{z}\\ \mu\end{pmatrix} = \begin{pmatrix}-\omega\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y}-\mathbf{y}_*)\\ \mathbf{0}\end{pmatrix} \tag{170b}$$

$$\mathbf{u} = \mu. \tag{170c}$$

Again, the first system appearing here, the *primal system*, are just the constraints (149) while (46) will be referred to as the *dual* or *adjoint system*. The specific form of the right hand side of the dual system emerges from the particular formulation of the minimization functional (169). The (here trivial) equation (170c) stems from measuring $\mathbf{u}$ just in ℓ_2, representing measuring the control in its natural trace norm. Instead of replacing μ by $\mathbf{u}$ in (46) and trying to solve the resulting equations, (170c) will be essential to devise an inexact gradient scheme. In fact, since $\mathbf{L}$ in (149) is an invertible operator, we can rewrite $\check{\mathbf{J}}(\mathbf{y},\mathbf{u})$ by formally inverting (149) as a functional of $\mathbf{u}$, that is, $\mathbf{J}(\mathbf{u}) := \check{\mathbf{J}}(\mathbf{y}(\mathbf{u}),\mathbf{u})$ as above. The following result will be very useful for the design of the outer–inner iterative solvers

Proposition 4.6. *The first variation of* $\mathbf{J}$ *satisfies*

$$\delta\mathbf{J}(\mathbf{u}) \;=\; \mathbf{u}-\mu, \tag{171}$$

where $(\mathbf{u},\mu)$ *are part of the solution of (170). Moreover,* $\mathbf{J}$ *is convex so that a unique minimizer exists.*

Hence, equation (170c) is just $\delta\mathbf{J}(\mathbf{u}) = \mathbf{0}$. For a unified treatment below of both control problems considered in these notes, it will be useful to rewrite (170c) like in (165) as a condensed equation for the control $\mathbf{u}$ alone. We formally invert (168) and (170b) and obtain

$$\mathbf{Q}\mathbf{u} = \mathbf{g} \tag{172}$$

with the abbreviations

$$\mathbf{Q} := \mathbf{Z}^T\mathbf{Z} + \omega\mathbf{I}, \quad \mathbf{g} := \mathbf{Z}^T(\mathbf{y}_* - \mathbf{T}_\square\mathbf{L}^{-1}\mathbf{I}_\square\mathbf{f}) \tag{173}$$

and

$$\mathbf{Z} := \mathbf{T}_\square\mathbf{L}^{-1}\mathbf{I}_\square, \qquad \mathbf{I}_\square := \begin{pmatrix}\mathbf{0}\\ \mathbf{I}\end{pmatrix}, \qquad \mathbf{T}_\square := (\mathbf{T}\ \ \mathbf{0}). \tag{174}$$

Proposition 4.7. *The vector* $\mathbf{u}$ *as part of the solution vector* $(\mathbf{y},\mathbf{p},\mathbf{z},\mu,\mathbf{u})$ *of (170) coincides with the unique solution* $\mathbf{u}$ *of the condensed equations (172).*

5 Iterative solution

Each of the four problem classes discussed above lead to the problem to finally solve a system

$$\delta\mathbf{J}(\mathbf{q}) \ = \ \mathbf{0} \tag{175}$$

or, equivalently, a linear system

$$\mathbf{M}\mathbf{q} = \mathbf{b}, \tag{176}$$

where $\mathbf{M} : \ell_2 \to \ell_2$ is a (possibly infinite) symmetric positive definite matrix satisfying

$$c_\mathbf{M}\|\mathbf{v}\| \le \|\mathbf{M}\mathbf{v}\| \le C_\mathbf{M}\|\mathbf{v}\|, \quad \mathbf{v} \in \ell_2, \tag{177}$$

for some constants $0 < c_\mathbf{M} \le C_\mathbf{M} < \infty$ and where $\mathbf{b} \in \ell_2$ is some given right hand side.

A simple *gradient method* for solving (175) is

$$\mathbf{q}_{k+1} := \mathbf{q}_k - \alpha\,\delta\mathbf{J}(\mathbf{q}_k), \qquad k = 0,1,2,\ldots \tag{178}$$

with some initial guess $\mathbf{q}_0$. In all of the previously considered situations, it has been asserted that there exists a fixed parameter α, depending on bounds for the second variation of $\mathbf{J}$, such that (178) converges and reduces the error in each step by at least a fixed factor $\rho < 1$, i.e.,

$$\|\mathbf{q} - \mathbf{q}_{k+1}\| \le \rho\|\mathbf{q} - \mathbf{q}_k\|, \quad k = 0,1,2,\ldots, \tag{179}$$

where ρ is determined by

$$\rho := \|\mathbf{I} - \alpha\mathbf{M}\| < 1.$$

Hence, the scheme (178) is a convergent iteration for the possibly infinite system (176). Next we will need to discuss how to reduce the infinite systems to computable finite versions.

5.1 Finite systems on uniform grids

We consider finite-dimensional trial spaces with respect to uniform discretizations. For each of the Hilbert spaces H, this means in the wavelet setting to pick the index set of all indices up to some *highest refinement level* J, i.e.,

$$\mathcal{I}_{J,H} := \{\lambda \in \mathcal{I}_H : |\lambda| \le J\} \subset \mathcal{I}_H$$

satisfying $N_{J,H} := \#\mathcal{I}_{J,H} < \infty$. The representation of operators is then built as in Section 3.3 with respect to this truncated index set which corresponds to deleting all rows and columns that refer to indices λ such that $|\lambda| > J$, and correspondingly for functions. There is by construction also a *coarsest level* of resolution denoted by j_0.

Computationally the representation of operators according to (62) is in the case of uniform grids always realized as follows. First, the operator is set up in terms of the *generator basis* on the finest level J. This generator basis simply consists of tensor products of B-Splines, or linear combinations of these near the boundaries. The representation of an operator in the *wavelet basis* is then achieved by applying the Fast Wavelet Transform (FWT) which needs $\mathcal{O}(N_{J,H})$ arithmetic operations and is therefore asymptotically optimal, see, e.g., [D2, DKU, K1] and Section 3.4.

In order not to overburden the notation, let in this subsection the resulting system for $N = N_{J,H}$ unknowns again be denoted by

$$\mathbf{M}\mathbf{q} = \mathbf{b}, \tag{180}$$

where now $\mathbf{M} : \mathbb{R}^N \to \mathbb{R}^N$ is a symmetric positive definite matrix satisfying (177) on $\mathbb{R}^N$. It will be convenient to abbreviate the residual using an approximation $\tilde{\mathbf{q}}$ to $\mathbf{q}$ for (180) as

$$\text{RESD}(\tilde{\mathbf{q}}) := \mathbf{M}\tilde{\mathbf{q}} - \mathbf{b}. \tag{181}$$

We will employ a basic conjugate gradient method that iteratively computes an approximate solution $\mathbf{q}_K$ to (180) with given initial vector $\mathbf{q}_0$ and given tolerance $\varepsilon > 0$ such that

$$\|\mathbf{M}\mathbf{q}_K - \mathbf{b}\| = \|\text{RESD}(\mathbf{q}_K)\| \le \varepsilon, \tag{182}$$

where K denotes the number of iterations used. Later we specify ε depending on the discretization for which (180) is set up. The following scheme CG contains a routine $\text{APP}(\eta_k, \mathbf{M}, \mathbf{d}_k)$ which in view of the problem classes discussed above is to have the property that it approximately computes the product $\mathbf{M}\mathbf{d}_k$ up to a tolerance $\eta_k = \eta_k(\varepsilon)$ depending on ε, i.e., the output ${}_k$ of $\text{APP}(\eta_k, \mathbf{M}, \mathbf{d}_k)$ satisfies

$$\|\mathbf{m}_k - \mathbf{M}\mathbf{d}_k\| \le \eta_k. \tag{183}$$

For the cases where $\mathbf{M} = \mathbf{A}$, this is simply the matrix-vector multiplication $\mathbf{M}\mathbf{d}_k$. For the situations where $\mathbf{M}$ may involve the solution of an additional system, this multiplication will be only approximative.

CG $[\varepsilon, \mathbf{q}_0, \mathbf{M}, \mathbf{b}] \to \mathbf{q}_K$

(I) SET $\mathbf{d}_0 := \mathbf{b} - \mathbf{M}\mathbf{q}_0$ AND $\mathbf{r}_0 := -\mathbf{d}_0$. LET $k = 0$.
(II) WHILE $\|\mathbf{r}_k\| > \varepsilon$

$$\begin{aligned} \mathbf{m}_k &:= \mathrm{APP}(\eta_k(\varepsilon), \mathbf{M}, \mathbf{d}_k) \\ \alpha_k &:= \frac{(\mathbf{r}_k)^T \mathbf{r}_k}{(\mathbf{d}_k)^T \mathbf{m}_k} \qquad\qquad \mathbf{q}_{k+1} := \mathbf{q}_k + \alpha_k \mathbf{d}_k \\ \mathbf{r}_{k+1} &:= \mathbf{r}_k + \alpha_k \mathbf{m}_k \qquad\qquad \beta_k := \frac{(\mathbf{r}_{k+1})^T \mathbf{r}_{k+1}}{(\mathbf{r}_k)^T \mathbf{r}_k} \\ \mathbf{d}_{k+1} &:= -\mathbf{r}_{k+1} + \beta_k \mathbf{d}_k \\ k &:= k+1 \end{aligned} \tag{184}$$

(III) SET $K := k - 1$.

Let us briefly discuss in the case $\mathbf{M} = \mathbf{A}$ that the final iterate $\mathbf{q}_K$ indeed satisfies (182). From the newly computed iterate $\mathbf{q}_{k+1} = \mathbf{q}_k + \alpha_k \mathbf{d}_k$ it follows by applying $\mathbf{M}$ on both sides that $\mathbf{M}\mathbf{q}_{k+1} - \mathbf{b} = \mathbf{M}\mathbf{q}_k - \mathbf{b} + \alpha_k \mathbf{M}\mathbf{d}_k$ which is the same as $\mathrm{RESD}(\mathbf{q}_{k+1}) = \mathrm{RESD}(\mathbf{q}_k) + \alpha_k \mathbf{M}\mathbf{d}_k$. By the initialization for $\mathbf{r}_k$ used above, this in turn is the updating term for $\mathbf{r}_k$, hence, $\mathbf{r}_k = \mathrm{RESD}(\mathbf{q}_k)$. After the stopping criterion based on $\mathbf{r}_k$ is met, the final iterate $\mathbf{q}_K$ observes (182).

The routine CG computes the *residual* up to the stopping criterion ε. From the residual, we can in view of (177) estimate the *error* in the solution as

$$\|\mathbf{q} - \mathbf{q}_K\| = \|\mathbf{M}^{-1}(\mathbf{b} - \mathbf{M}\mathbf{q}_K)\| \le \|\mathbf{M}^{-1}\| \, \|\mathrm{RESD}(\mathbf{q}_K)\| \le \frac{\varepsilon}{c_{\mathbf{M}}}, \tag{185}$$

that is, it may deviate from the norm of the residual from a factor proportional to the smallest eigenvalue of $\mathbf{M}$.

Distributed control. Let us now apply the solution scheme to the situation from Section 4.3 where $\mathbf{Q}$ now involves the inversion of finite-dimensional systems (158a) and (158b). The material in the remainder of this subsection is essentially contained in [BK].

We begin with a specification of the approximate computation of the right hand side $\mathbf{b}$ which also contains applications of $\mathbf{A}^{-1}$.

RHS$[\zeta, \mathbf{A}, \mathbf{f}, \mathbf{y}_*] \to \mathbf{b}_\zeta$

(I) CG $[\frac{c_{\mathbf{A}}}{2C} \frac{c_{\mathbf{A}}}{C^2 C_0^2} \zeta, \mathbf{0}, \mathbf{A}, \mathbf{f}] \to \mathbf{b}_1$
(II) CG $[\frac{c_{\mathbf{A}}}{2C} \zeta, \mathbf{0}, \mathbf{A}^T, -\mathbf{D}_{\mathscr{Z}}^{-1} \mathbf{R} \mathbf{D}_{\mathscr{Z}}^{-1} (\mathbf{b}_1 - \mathbf{y}_*)] \to \mathbf{b}_2$
(III) $\mathbf{b}_\zeta := \mathbf{D}_H^{-1} \mathbf{b}_2$.

The tolerances used within the two conjugate gradient methods depend on the constants $c_{\mathbf{A}}, C, C_0$ from (13), (147) and (55), respectively. Since the additional factor $c_{\mathbf{A}}(CC_0)^{-2}$ in the stopping criterion in step (I) in comparison to step (II) is in general smaller than one, this means that the primal system needs to be solved more accurately than the adjoint system in step (II).

Proposition 5.1. *The result* $\mathbf{b}_\zeta$ *of* $\mathrm{RHS}[\zeta,\mathbf{A},\mathbf{f},\mathbf{y}_*]$ *satisfies*

$$\|\mathbf{b}_\zeta-\mathbf{b}\|\le\zeta. \tag{186}$$

Proof. Recalling the definition (163) of $\mathbf{b}$, step (III) and step (II) yield

$$\begin{aligned}\|\mathbf{b}_\zeta-\mathbf{b}\| &\le \|\mathbf{D}_H^{-1}\|\,\|\mathbf{b}_2-\mathbf{D}_H\mathbf{b}\|\\ &\le C\|\mathbf{A}^{-T}\|\,\|\mathbf{A}^T\mathbf{b}_2-\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f}-\mathbf{b}_1+\mathbf{b}_1-\mathbf{y}_*)\|\\ &\le \frac{C}{c_\mathbf{A}}\left(\frac{c_\mathbf{A}}{2C}\zeta+\|\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f}-\mathbf{b}_1)\|\right).\end{aligned} \tag{187}$$

Employing the upper bounds for $\mathbf{D}_{\mathscr{Z}}^{-1}$ and $\mathbf{R}$, we arrive at

$$\begin{aligned}\|\mathbf{b}_\zeta-\mathbf{b}\| &\le \frac{C}{c_\mathbf{A}}\left(\frac{c_\mathbf{A}}{2C}\zeta+C^2C_0^2\,\|\mathbf{A}^{-1}\|\,\|\mathbf{f}-\mathbf{A}\mathbf{b}_1\|\right)\\ &\le \frac{C}{c_\mathbf{A}}\left(\frac{c_\mathbf{A}}{2C}\zeta+\frac{C^2C_0^2}{c_\mathbf{A}}\frac{c_\mathbf{A}}{2C}\frac{c_\mathbf{A}}{C^2C_0^2}\zeta\right)\ =\zeta.\end{aligned} \tag{188}$$

□

Accordingly, an approximation $\mathbf{m}_\eta$ to the matrix-vector product $\mathbf{Q}\mathbf{d}$ is the output of the following routine APP.

APP$[\eta,\mathbf{Q},\mathbf{d}]\to\mathbf{m}_\eta$

(I) CG $[\frac{c_\mathbf{A}}{3C}\frac{c_\mathbf{A}}{C^2C_0^2}\eta,\mathbf{0},\mathbf{A},\mathbf{f}+\mathbf{D}_H^{-1}\mathbf{d}]\to\mathbf{y}_\eta$

(II) CG $[\frac{c_\mathbf{A}}{3C}\eta,\mathbf{0},\mathbf{A}^T,-\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_Z^{-1}(\mathbf{y}_\eta-\mathbf{y}_*)]\to\mathbf{p}_\eta$

(III) $\mathbf{m}_\eta:=\mathbf{g}_{\eta/3}+\omega\mathbf{R}^{-1}\mathbf{d}-\mathbf{D}_H^{-1}\mathbf{p}_\eta$.

The choice of the tolerances for the interior application of CG in steps (I) and (II) will become clear from the following result.

Proposition 5.2. *The result* $\mathbf{m}_\eta$ *of* APP$[\eta,\mathbf{Q},\mathbf{d}]$ *satisfies*

$$\|\mathbf{m}_\eta-\mathbf{Q}\mathbf{d}\|\le\eta. \tag{189}$$

Proof. Denote by $\mathbf{y}_\mathbf{d}$ the exact solution of (158a) with $\mathbf{d}$ in place of $\mathbf{u}$ on the right hand side, and by $\mathbf{p}_\mathbf{d}$ the exact solution of (158b) with $\mathbf{y}_\mathbf{d}$ on the right hand side. Then we deduce from step (III) and (167) combined with (55) and (147)

$$\begin{aligned}\|\mathbf{m}_\eta-\mathbf{Q}\mathbf{d}\| &= \|\mathbf{g}_{\eta/3}-\mathbf{g}+\omega\mathbf{R}^{-1}\mathbf{d}-\mathbf{D}_U^{-1}\mathbf{p}_\eta\ -(\mathbf{Q}\mathbf{d}-\mathbf{g})\|\\ &\le \frac{1}{3}\eta+\|\omega\mathbf{R}^{-1}\mathbf{d}-\mathbf{D}_U^{-1}\mathbf{p}_\eta-(\omega\mathbf{R}^{-1}\mathbf{d}-\mathbf{D}_U^{-1}\mathbf{p}_\mathbf{d})\|\\ &\le \frac{1}{3}\eta+C\|\mathbf{p}_\mathbf{d}-\mathbf{p}_\eta\|.\end{aligned} \tag{190}$$

Denote by $\hat{\mathbf{p}}$ the exact solution of (158b) with $\mathbf{y}_\eta$ on the right hand side. Then we have $\mathbf{p}_\mathbf{d}-\hat{\mathbf{p}}=-\mathbf{A}^{-T}\mathbf{D}_Z^{-1}\mathbf{R}\mathbf{D}_Z^{-1}(\mathbf{y}_\mathbf{d}-\mathbf{y}_\eta)$. It follows by (13), (55) and (147) that

$$\|\mathbf{p_d} - \hat{\mathbf{p}}\| \leq \frac{C^2 C_0^2}{c_\mathbf{A}} \|\mathbf{y_d} - \mathbf{y}_\eta\| \leq \frac{1}{3C}\eta, \tag{191}$$

where the last estimate follows by the choice of the threshold in step (I). Finally, the combination(190) and (191) together with (186) and the stopping criterion in step (II) readily confirms that

$$\begin{aligned}\|\mathbf{m}_\eta - \mathbf{Qd}\| &\leq \frac{1}{3}\eta + C\left(\|\mathbf{p_d} - \hat{\mathbf{p}}\| + \|\hat{\mathbf{p}} - \mathbf{p}_\eta\|\right) \\ &\leq \frac{1}{3}\eta + C\left(\frac{1}{3C}\eta + \frac{1}{3C}\eta\right) = \eta.\end{aligned}$$

□

The effect of perturbed applications of $\mathbf{M}$ in CG and more general Krylov subspace schemes with respect to convergence has been investigated in a numerical linear algebra context for a given linear system (180) in several papers. Here we have chosen the η_i to be proportional to the outer accuracy ε incorporating a safety factor accounting for the values of β_i and $\|\mathbf{r}_i\|$.

Finally, we can formulate a full nested iteration strategy for finite systems (158) on uniform grids which employs outer and inner CG routines as follows. The scheme starts at the coarsest level of resolution j_0 with some initial guess $\mathbf{u}_0^{j_0}$ and successively solves (165) with respect to each level j until the norm of the current residual is below the discretization error on that level.

In wavelet coordinates, $\|\cdot\|$ corresponds to the energy norm. If we employ on the primal side for approximation linear combinations of B–splines of order d, the discretization error is for smooth solutions expected to be proportional to $2^{-(d-1)j}$. Then the refinement level is successively increased until on the finest level J a prescribed tolerance proportional to the discretization error $2^{-(d-1)J}$ is met. In the following, superscripts on vectors denote the refinement level on which this term is computed. The given data $\mathbf{y}_*^j$, $\mathbf{f}^j$ are supposed to be accessible on all levels. On the coarsest level, the solution of (165) is computed exactly up to double precision by QR decomposition. Subsequently, the results from level j are prolongated onto the next higher level $j+1$. Using wavelets, this is accomplished by simply adding zeros: wavelet coordinates have the character of differences so that this prolongation corresponds to the exact representation in higher resolution wavelet coordinates. The resulting *Nested–Iteration–Incomplete–Conjugate–Gradient* Algorithm is the following.

NIICG$[J] \to \mathbf{u}^J$

(I) INITIALIZATION FOR COARSEST LEVEL $j := j_0$

(1) COMPUTE RIGHT HAND SIDE $\mathbf{g}^{j_0} = (\mathbf{Z}^T\mathbf{G})^{j_0}$ BY QR DECOMPOSITION USING (160).

(2) COMPUTE SOLUTION $\mathbf{u}^{j_0}$ OF (165) BY QR DECOMPOSITION.

(II) WHILE $j < J$

(1) PROLONGATE $\mathbf{u}^j \to \mathbf{u}_0^{j+1}$ BY ADDING ZEROS, SET $j := j+1$.
(2) COMPUTE RIGHT HAND SIDE USING RHS $[2^{-(d-1)j}, \mathbf{A}, \mathbf{f}^j, \mathbf{y}_*^j] \to \mathbf{g}^j$.
(3) COMPUTE SOLUTION OF (165) USING CG $[2^{-(d-1)j}, \mathbf{u}_0^j, \mathbf{Q}, \mathbf{g}^j] \to \mathbf{u}^j$.

Recall that step (II.3) requires multiple calls of APP$[\eta, \mathbf{Q}, \mathbf{d}]$, which in turn invokes both CG $[\ldots, \mathbf{A}, \ldots]$ as well as CG $[\ldots, \mathbf{A}^T, \ldots]$ in each application.

On account of (13) and (166), finite versions of the system matrices $\mathbf{A}$ and $\mathbf{Q}$ have uniformly bounded condition numbers, entailing that each CG routine employed in the process reduces the error by a fixed rate $\rho < 1$ in each iteration step. Let $N_J \sim 2^{nJ}$ be the total number of unknowns (for $\mathbf{y}^J, \mathbf{u}^J$ and $\mathbf{p}^J$) on the highest level J. Employing the CG method only on the highest level, one needs $\mathcal{O}(J) = \mathcal{O}(\log \varepsilon)$ iterations to achieve the prescribed discretization error accuracy $\varepsilon_J = 2^{-(d-1)J}$. As each application of $\mathbf{A}$ and $\mathbf{Q}$ requires $\mathcal{O}(N_J)$ operations, the solution of (165) by CG only on the finest level requires $\mathcal{O}(J N_J)$ arithmetic operations.

Proposition 5.3. *If the residual (167) is computed up to discretization error proportional to $2^{-(d-1)j}$ on each level j and the corresponding solutions are taken as initial guesses for the next higher level,* NIICG *is an asymptotically optimal method in the sense that it provides the solution $\mathbf{u}^J$ up to discretization error on level J in an overall amount of $\mathcal{O}(N_J)$ arithmetic operations.*

Proof. In the above notation, nested iteration allows one to get rid of the factor J in the total amount of operations. Starting with the exact solution on the coarsest level j_0, in view of the uniformly bounded condition numbers of $\mathbf{A}$ and $\mathbf{Q}$, one needs only a fixed amount of iterations to reduce the error up to discretization error accuracy $\varepsilon_j = 2^{-(d-1)j}$ on each subsequent level j, taking the solution from the previous level as initial guess. Thus, on each level, one needs $\mathcal{O}(N_j)$ operations to realize discretization error accuracy. Since the spaces are nested and the number of unknowns on each level grows like $N_j \sim 2^{nj}$, by a geometric series argument the total number of arithmetic operations stays proportional to $\mathcal{O}(N_J)$. □

5.2 Numerical examples

5.2.1 Distributed control problem

As an illustration of the issue which norms to choose in the control functional, we consider the following example of a one-dimensional distributed control problem with the Helmholtz operator in (6) ($\mathbf{a} = I, c = 1$) and homogeneous Dirichlet boundary condition. A non–constant right hand side $f(x) := 1 + 2.3\exp(-15|x-0.5|)$ is chosen, and the target state is set to a constant $y_* \equiv 1$. We first investigate the role the different norms $\|\cdot\|_{\mathscr{Z}}$ and $\|\cdot\|_{\mathscr{U}}$ in (27), which is encoded in the diagonal matrices $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ from (146), have on the solution. We see in Figure 3 for the choice $\mathscr{U} = L_2(0,1)$ and $\mathscr{Z} = H^s(0,1)$ for different values of s varying between 0 and 1

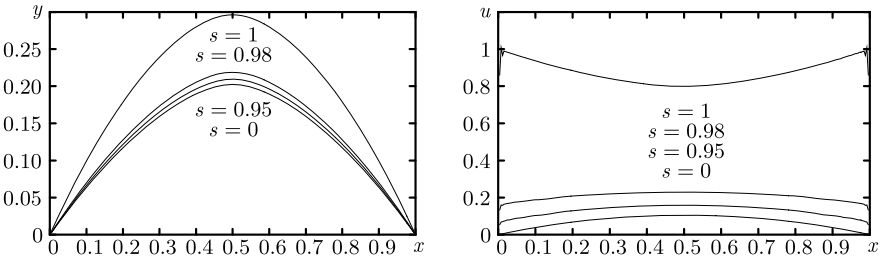


Fig. 3 Distributed control problem for elliptic PDE with Dirichlet boundary conditions, a peak as right hand side f, $y_* \equiv 1$, $\omega = 0$, $\mathscr{U} = L_2(0,1)$ and varying $\mathscr{Z} = H^s(0,1)$. Left: state y, right: control u

the solution y (left) and the corresponding control u (right) for fixed weight $\omega = 1$. As s is increased, a stronger tendency of y towards the prescribed state $y_* \equiv 1$ can be observed which is, however, deterred from reaching this state by the homogeneous boundary conditions. Extensive studies of this type can be found in [Bu1, BK].

An example displaying the performance of the proposed fully iterative scheme NIICG is shown in Table 4 for $n = 2$ and in Table 5 for $n = 3$.

j	$\|\mathbf{r}_K^j\|$	#O	#E	#A	#R	$\|R(\mathbf{y}^J) - \mathbf{y}^j\|$	$\|\mathbf{y}^J - P(\mathbf{y}^j)\|$	$\|R(\mathbf{u}^J) - \mathbf{u}^j\|$	$\|\mathbf{u}^J - P(\mathbf{u}^j)\|$
3						6.86e-03	1.48e-02	1.27e-04	4.38e-04
4	1.79e-05	5	12	5	8	2.29e-03	7.84e-03	4.77e-05	3.55e-04
5	1.98e-05	5	14	6	9	6.59e-04	3.94e-03	1.03e-05	2.68e-04
6	4.92e-06	7	13	5	9	1.74e-04	1.96e-03	2.86e-06	1.94e-04
7	3.35e-06	7	12	5	9	4.55e-05	9.73e-04	9.65e-07	1.35e-04
8	2.42e-06	7	11	5	10	1.25e-05	4.74e-04	7.59e-07	8.88e-05
9	1.20e-06	8	11	5	10	4.55e-06	2.12e-04	4.33e-07	5.14e-05
10	4.68e-07	9	10	5	9	3.02e-06	3.02e-06	2.91e-07	2.91e-07

Table 4 Iteration history for a two-dimensional distributed control problem with Neumann boundary conditions, $\omega = 1$, $\mathscr{Z} = H^1(\Omega)$, $\mathscr{U} = (H^{0.5}(\Omega))'$

j	$\|\mathbf{r}_K^j\|$	#O	#E	#A	#R	$\|R(\mathbf{y}^J) - \mathbf{y}^j\|$	$\|\mathbf{y}^J - P(\mathbf{y}^j)\|$	$\|R(\mathbf{u}^J) - \mathbf{u}^j\|$	$\|\mathbf{u}^J - P(\mathbf{u}^j)\|$
3						1.41e-04	2.92e-04	1.13e-05	2.36e-05
4	6.09e-06	10	9	1	49	1.27e-04	1.78e-04	3.46e-06	3.79e-06
5	3.25e-06	10	7	1	58	1.11e-05	6.14e-05	9.47e-07	9.53e-07
6	1.71e-06	7	6	1	57	1.00e-05	2.86e-05	5.03e-07	5.03e-07
7	8.80e-07	6	6	1	53	9.19e-06	9.19e-06	3.72e-07	3.72e-07

Table 5 Iteration history for a three-dimensional distributed control problem with Neumann boundary conditions, $\omega = 1$, $\mathscr{Z} = H^1(\Omega)$, $\mathscr{U} = (H^1(\Omega))'$

This is an example of a control problem for the Helmholtz operator with Neumann boundary conditions. The stopping criterion for the outer iteration (relative to

$\|\cdot\|$ which corresponds to the energy norm) on level j is chosen to be proportional to 2^{-j}. The second column displays the final value of the residual of the outer CG scheme on this level, i.e., $\|\mathbf{r}_K^j\| = \|\text{RESD}(\mathbf{u}_K^j)\|$. The next three columns show the number of outer CG iterations (#O) for $\mathbf{Q}$ according to the APP scheme followed by the maximum number of inner iterations for the primal system (#E), the adjoint system (#A) and the design equation (#R). We see very well the effect of the uniformly bounded condition numbers of all involved operators. The last columns display different versions of the actual error in the state $\mathbf{y}$ and the control $\mathbf{u}$ when compared to the fine grid solution (R denotes restriction of the fine grid solution to the actual grid, and P denotes prolongation). Here we can see the slight effect of the constants appearing in (185). Nevertheless the error is very well controlled by the residual.

More results for up to three spatial dimensions can be found in [Bu1, BK]. All numbers were obtained on a 3.2GHz Pentium IV computer (family 15, model 4, stepping 1, with 1MB L2 Cache).

5.2.2 Dirichlet boundary control

For the system of saddle point problems (170) arising from the control problem with Dirichlet boundary control in Section 2.5, also a fully iterative algorithm NI-ICG can be designed along the above lines with yet another level of inner iteration. Again the design equation (170c) for $\mathbf{u}$ serves as the equation for which a basic iterative scheme (178) can be posed. Of course, the CG method for $\mathbf{A}$ then has to be replaced by a convergent iterative scheme for saddle point operators $\mathbf{L}$ like Uzawa's algorithm. Also the discretization has to be chosen such that the LBB condition is satisfied, see Section 4.2. Details can be found in [K3]. Alternatively, since $\mathbf{L}$ has a uniformly bounded condition number, the CG scheme can, in principle, also be applied to $\mathbf{L}^T\mathbf{L}$. The performance of wavelet schemes on uniform grids for such systems of saddle point problems arising from optimal control has been investigated systematically in [Pa].

For illustration of the choice of different norms for the Dirichlet boundary control problem, consider the following example. We control the system through the (green) control boundary Γ in Figure 4 while a prescribed state $y_* \equiv 1$ on the (red) observation boundary Γ_y opposite the control boundary is to be achieved. The right hand side is chosen as constant $f \equiv 1$, and $\omega = 1$. Each layer in Figure 4 corresponds to the state y for different values of s when the observation term is measured in $H^s(\Gamma_y)$, that is, the objective functional (35) contains a term $\|y - y_*\|^2_{H^s(\Gamma_y)}$ for increasing s from bottom to top. We see that as the smoothness index s for the observation increases, the state moves towards the target state at the observation boundary. In comparison, in Figure 5 the weight parameter ω balancing the two terms in the functional is modified. We observe that the effect on the solution of varying s corresponds to a similar behaviour of varying the weight. However, as ω directly influences the conditioning of the system of saddle point operators, a solution scheme with fixed ω and varying s can be considered numerically more stable.

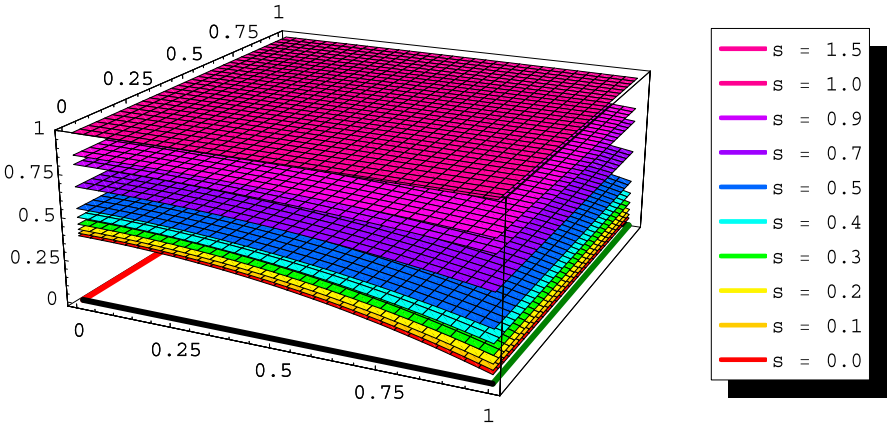


Fig. 4 State y of the Dirichlet boundary control problem using the objective functional $\mathcal{J}(y,u) = \frac{1}{2}\|y - y_*\|^2_{H^s(\Gamma_y)} + \frac{1}{2}\|u\|^2_{H^{1/2}(\Gamma)}$ for control boundary Γ (green) and observation boundary Γ_y (red) for different values of the Sobolev smoothness index s on resolution level $J = 5$ [Pa]

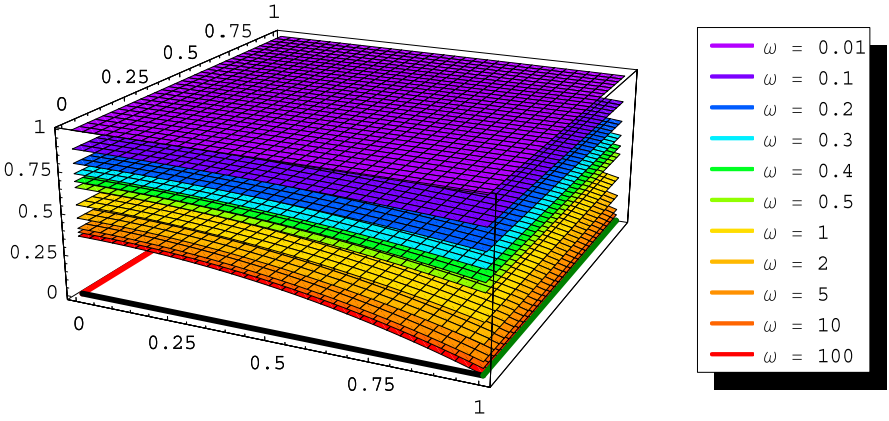


Fig. 5 State y of the Dirichlet boundary control problem using the objective functional $\mathcal{J}(y,u) = \frac{1}{2}\|y - y_*\|^2_{H^s(\Gamma_y)} + \frac{\omega}{2}\|u\|^2_{H^{1/2}(\Gamma)}$ for control boundary Γ (green) and observation boundary Γ_y (red) for different values of the weight parameter ω [Pa]

Finally, we display in Table 6 some numerical results for an elliptic control problem with Dirichlet boundary control in two spatial dimensions. Among the various iteration schemes tested, the best results with a minimal amount of iteration numbers (here: at most 2) were obtained for an inexact gradient iteration on $\mathbf{u}$ and Uzawa–type schemes with conjugate directions for each of the saddle point problems together with nested iteration.

j	$\|\mathbf{r}^j\|$	$\|\mathbf{y}-\mathbf{y}^J\|$	k_j	$\frac{\#\text{Int-It}}{k_j}$
4	1.6105e-02	7.7490e-00	0	–
5	1.6105e-02	7.7506e-00	0	–
6	6.3219e-03	1.7544e-02	2	1
7	5.8100e-03	3.3873e-02	0	–
8	1.6378e-03	3.4958e-03	2	1
9	1.8247e-03	7.4741e-03	0	–
10	4.3880e-04	9.2663e-04	2	1
11	4.6181e-04	1.8486e-03	0	–

Table 6 Dirichlet boundary control problem in two spatial dimensions with $y_{\Gamma_y} \equiv 1$, $f \equiv 1$, $\omega = 1$, $s = t = 0.5$. The table shows the number of iterations k_j needed to reduce the $\mathscr{L}$-error of $\mathbf{r}^j$ by a factor of 0.5 after prolongation of all final vectors from the previous level [Pa]

References

[Ba] I. Babuška, The finite element method with Lagrange multipliers, Numer. Math. **20** (1973), 179–192.

[B] D. Braess, Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics, 2nd ed., Cambridge University Press, Cambridge, 2001.

[BH] D. Braess, W. Hackbusch, A new convergence proof for the multigrid method including the V-cycle, SIAM J. Numer. Anal. **20** (1983), 967–975.

[BPX] J.H. Bramble, J.E. Pasciak, J. Xu, Parallel multilevel preconditioners, Math. Comp. **55** (1990), 1–22.

[BF] F. Brezzi, M. Fortin, Mixed and Hybrid Finite Element Methods, Springer, 1991.

[Bu1] C. Burstedde, Fast Optimised Wavelet Methods for Control Problems Constrained by Elliptic PDEs, PhD Dissertation, Mathematisch-Naturwissenschaftliche Fakultät, Universität Bonn, Germany, 2005.

[Bu2] C. Burstedde, On the numerical evaluation of fractional Sobolev norms, Comm. Pure Appl. Anal. **6**(3) (2007), 587–605.

[BK] C. Burstedde, A. Kunoth, Fast iterative solution of elliptic control problems in wavelet discretizations, J. Comp. Appl. Math. **196**(1) (2006), 299–319.

[CTU] C. Canuto, A. Tabacco, K. Urban, The wavelet element method, part I: Construction and analysis, Appl. Comput. Harm. Anal. **6** (1999), 1–52.

[CDP] J.M. Carnicer, W. Dahmen, J.M. Peña, Local decomposition of refinable spaces, Appl. Comp. Harm. Anal. **3** (1996), 127–153.

[CF] Z. Ciesielski, T. Figiel, Spline bases in classical function spaces on compact C^∞ manifolds: Part I and II, Studia Mathematica (1983), 1–58 and 95–136.

[Co] A. Cohen, Numerical Analysis of Wavelet Methods, Studies in Mathematics and its Applications 32, Elsevier, 2003.

[CDF] A. Cohen, I. Daubechies, J.-C. Feauveau, Biorthogonal bases of compactly supported wavelets, Comm. Pure Appl. Math. **45** (1992), 485–560.

[DDU] S. Dahlke, W. Dahmen, K. Urban, Adaptive wavelet methods for saddle point problems — Optimal convergence rates, SIAM J. Numer. Anal. **40** (2002), 1230–1262.

[D1] W. Dahmen, Stability of multiscale transformations, J. Four. Anal. Appl. **2** (1996), 341–361.

[D2] W. Dahmen, Wavelet and multiscale methods for operator equations, Acta Numerica (1997), 55–228.

[D3] W. Dahmen, Wavelet methods for PDEs – Some recent developments, J. Comput. Appl. Math. **128** (2001), 133–185.
[D4] W. Dahmen, Multiscale and wavelet methods for operator equations, in: Multiscale Problems and Methods in Numerical Simulation, C. Canuto (ed.), C.I.M.E. Lecture Notes in Mathematics 1825, Springer Heidelberg (2003), 31–96.
[DK1] W. Dahmen, A. Kunoth, Multilevel preconditioning, Numer. Math. **63** (1992), 315–344.
[DK2] W. Dahmen, A. Kunoth, Appending boundary conditions by Lagrange multipliers: Analysis of the LBB condition, Numer. Math. **88** (2001), 9–42.
[DK3] W. Dahmen, A. Kunoth, Adaptive wavelet methods for linear–quadratic elliptic control problems: Convergence Rates, SIAM J. Contr. Optim. **43**(5) (2005), 1640–1675.
[DKS] W. Dahmen, A. Kunoth, R. Schneider, Wavelet least squares methods for boundary value problems, SIAM J. Numer. Anal. **39** (2002), 1985–2013.
[DKU] W. Dahmen, A. Kunoth, K. Urban, Biorthogonal spline wavelets on the interval – Stability and moment conditions, Appl. Comput. Harm. Anal. **6** (1999), 132–196.
[DS1] W. Dahmen, R. Schneider, Wavelets with complementary boundary conditions — Function spaces on the cube, Results in Mathematics **34** (1998), 255–293.
[DS2] W. Dahmen, R. Schneider, Composite wavelet bases for operator equations, Math. Comp. **68** (1999), 1533–1567.
[DS3] W. Dahmen, R. Schneider, Wavelets on manifolds I: Construction and domain decomposition, SIAM J. Math. Anal. **31** (1999), 184–230.
[DSt] W. Dahmen, R. Stevenson, Element–by–element construction of wavelets satisfying stability and moment conditions, SIAM J. Numer. Anal. **37** (1999), 319–325.
[Dau] I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. **41** (1988), 909–996.
[Gr] P. Grisvard, Elliptic Problems in Nonsmooth Domains, Pitman, 1985.
[J] S. Jaffard, Wavelet methods for fast resolution of elliptic problems, Siam J. Numer. Anal. **29** (1992), 965–986.
[KK] P. Kantartzis, A. Kunoth, A wavelet approach for a problem in electrical impedance tomography formulated by means of a domain embedding method, Manuscript, in preparation.
[Kr] J. Krumsdorf, Finite Element Wavelets for the Numerical Solution of Elliptic Partial Differential Equations on Polygonal Domains, Diploma Thesis, Universität Bonn, 2004.
[K1] A. Kunoth, Wavelet Methods — Elliptic Boundary Value Problems and Control Problems, Advances in Numerical Mathematics, Teubner, 2001.
[K2] A. Kunoth, Wavelet techniques for the fictitious domain—Lagrange multiplier approach, Numer. Algor. **27** (2001), 291–316.
[K3] A. Kunoth, Fast iterative solution of saddle point problems in optimal control based on wavelets, Comput. Optim. Appl. **22** (2002), 225–259.
[K4] A. Kunoth, Adaptive wavelet schemes for an elliptic control problem with Dirichlet boundary control, Numer. Algor. **39**(1-3) (2005), 199–220.
[KS] A. Kunoth, J. Sahner, Wavelets on manifolds: An optimized construction, Math. Comp. **75** (2006), 1319–1349.
[Li] J.L. Lions, Optimal Control of Systems Governed by Partial Differential Equations, Springer, Berlin, 1971.
[MB] J. Maes, A. Bultheel, A hierarchical basis preconditioner for the biharmonic equation on the sphere, IMA J. Numer. Anal. **26**(3) (2006), 563–583.
[MKB] J. Maes, A. Kunoth, A. Bultheel, BPX-type preconditioners for 2nd and 4th order elliptic problems on the sphere, SIAM J. Numer. Anal. **45**(1) (2007), 206–222.
[O] P. Oswald, On discrete norm estimates related to multilevel preconditioners in the finite element method, in: Constructive Theory of Functions, K.G. Ivanov, P. Petrushev, B. Sendov, (eds.), Proc. Int. Conf. Varna 1991, Bulg. Acad. Sci., Sofia (1992), 203–214.
[Pa] R. Pabel, Wavelet Methods for PDE Constrained Elliptic Control Problems with Dirichlet Boundary Control, Diploma Thesis, Universität Bonn, 2006. doi: 10.2370/236_232
[Stv] R. Stevenson, Locally supported, piecewise polynomial biorthogonal wavelets on nonuniform meshes, Constr. Approx. **19** (2003), 477–508.

[U] K. Urban, Wavelet Methods for Elliptic Partial Differential Equations, Oxford University Press, 2009.

[X1] J. Xu, Theory of multilevel methods, Report AM 48, Department of Mathematics, Pennsylvania State University, 1989.

[X2] J. Xu, Iterative methods by space decomposition and subspace correction, SIAM Review **34**(4) (1992), 581–613.

[Y] H. Yserentant, On the multilevel splitting of finite element spaces, Numer. Math. **49** (1986), 379–412.

Multiresolution schemes for conservation laws

Siegfried Müller

Abstract The concept of fully adaptive multiresolution finite volume schemes has been developed and investigated during the past decade. By now it has been successfully employed in numerous applications arising in engineering. In the present work a review on the methodology is given that aims to summarize the underlying concepts and to give an outlook on future developments.

1 Introduction

Nowadays scientific computing has become an indispensable tool in engineering. For instance, numerical simulations of fluid flow can help to shorten the development cycle of new airplanes. In the near future, it might be even possible to perform real-time simulations of flying airplanes, to determine aerodynamical loads for the entire flight regime, to numerically predict the performance and the flight quality of an airplane before the maiden flight, as well as to do the certification before the airplane construction on the basis of numerical data. These are the challenging goals of the new Center for Computer Applications in Aerospace Science and Engineering ($C^2A^2S^2E$) funded in 2007 at the DLR Braunschweig.

Typically the numerical simulation of such real-world applications requires meshes with several millions of cells. This poses enormous challenges to computing resources and data management strategies. Improved hardware or purely data oriented strategies such as parallel computing are not sufficient to overcome the arising difficulties. In the long run, they have to be complemented by mathematical concepts that aim at minimizing the size of the resulting discrete problems and, thus, to keep the computational complexity tractable. One promising approach in

Siegfried Müller
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, D-52056 Aachen, Germany, e-mail: mueller@igpm.rwth-aachen.de

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_11,

this direction is based on local grid adaptation which aims to adjust the resolution of the discretization to the local regularity of the underlying solution.

This paper summarizes some recent work on grid adaptation in the context of hyperbolic conservation laws that arise, for instance, from the balance equations derived in continuum mechanics and modeling fluid flow. Currently, several different adaptive concepts for conservation laws are being discussed and investigated in the literature. A standard approach is based on error indicators, for instance gradient-based indicators [10, 8] or local residuals [43, 66, 67]. In practice, these approaches turned out to be very efficient. However, the error indicator is highly case dependent, i.e., it needs a lot of parameter tuning to avoid excessive mesh growth or missing refinement of important flow features. In particular, it does not estimate the local discretization error and, hence, it provides no reliable error control. Here a-posteriori error estimators offer an alternative that aims at the equidistribution of the error, cf. [48]. These rely on L^1-error-estimates. In particular, they are based on Kruzkov's entropy condition [49] and Kuznetsov's a-priori estimates [50] that are only available for scalar multidimensional conservation laws. If only a functional of the solution is of interest rather than the solution in the entire flow field, then another approach is of interest based on the solution of an adjoint problem [5, 42, 70, 69]. Here grid adaptation is tuned with respect to the efficient and accurate computation of a target quantity, e.g. drag or lift. Since this approach requires to store to some extent the time history of the evolution, this certainly poses a considerable challenge to computational resources in case of 3D unsteady problems.

In recent years, the new concept of multiscale-based grid adaptation has been developed and applied to complex multidimensional flow problems. The main distinction from previous work lies in the fact that we employ *multiresolution techniques*. The starting point is a proposal by Harten [38] to transform the arrays of cell averages associated with any given finite volume discretization of the underlying conservation laws into a different format that reveals insight into the local behavior of the solution. The cell averages on a given highest level of resolution (*reference mesh*) are represented as cell averages on some coarse level where the fine scale information is encoded in arrays of *detail coefficients* of ascending resolution. This requires a *hierarchy of meshes*.

In Harten's original approach [39, 40, 12], the multiscale analysis is used to control a hybrid flux computation which can save CPU time for the flux evaluation. However, the overall computational complexity is not reduced but still stays proportional to the number of cells on the uniformly fine reference mesh which in 3D calculations is prohibitive. Alternatively to this strategy, threshold techniques are applied to the multiresolution decomposition in [54, 25], where detail coefficients below a threshold value are discarded. By means of the remaining significant details, a locally refined mesh is determined whose complexity is significantly reduced in comparison to the underlying reference mesh. Thus a principal objective is to extract the inherent complexity of the problem by placing as few degrees of freedom as possible while the features of the solution are still captured within a given tolerance. A central mathematical problem is to show that the essential information to be propagated in time is still kept with sufficient accuracy when working on locally

coarser meshes. This has been proven for scalar onedimensional conservation laws in [25, 44].

The fully adaptive concept has turned out to be highly efficient and reliable. So far, it has been applied with great success to different applications, e.g., 2D/3D–steady and unsteady computations of compressible fluids around airfoils modeled by the Euler and Navier–Stokes equations, respectively, on block–structured curvilinear grid patches [15], backward–facing step on 2D triangulations [26] and simulation of a flame ball modeled by reaction–diffusion equations on 3D Cartesian grids [63]. These applications have been performed for compressible single-phase fluids. More recently, this concept has been extended to two-phase fluid flow of compressible gases, and applied to the investigation of non–stationary shock–bubble interactions on 2D Cartesian grids for the Euler equations [1, 55]. By now, there are several groups working on this subject: Postel et al. [28], Schneider et al. [61, 62], Bürger et al. [20, 19] and Domingues et al. [30].

The aim of the present work is to give an overview on the concept of multiscale-based grid adaptation. For this purpose, we first summarize the basic ingredients of the grid adaptation concept starting with the underlying equations and their discretization using finite volume schemes, see Section 2. This is followed by the multiscale analysis of the discrete cell averages resulting from the finite volume discretization, see Section 3, and the construction of locally refined grids using data compression techniques, see Section 4. Applying the multiscale analysis to the original finite volume discretization on the uniform grid we obtain multiscale evolution equations, see Section 5. The crucial point is then to perform the time evolution on the adaptive grid where the accuracy of the uniform discretization is maintained but the computational complexity is proportional only to the number of cells of the adaptive grid, see Section 5.5. For this purpose, the computation of the local flux balances and sources has to be performed judiciously, see Section 5.2, and the adaptive grid has to be predicted appropriately from the data of the previous time step, see Section 5.3. The resulting adaptive multiresolution scheme is further accelerated using multilevel time stepping strategies, see Section 5.4. In order to confirm that the multiresolution grid adaptation concept can deal with challenging applications in engineering, we present in Section 6 numerical simulations of two vortices generated at an airplane wing and moving in the wake of the airplane. The computations have been performed with the adaptive, parallel Quadflow solver [15]. In Section 7, we conclude with some remarks on future trends of adaptive multiresolution schemes.

2 Governing equations and finite volume schemes

The fluid equations are determined by the balance equations

$$\frac{\partial}{\partial t}\int_V \mathbf{u}\ dV + \oint_{\partial V} \mathbf{f}(\mathbf{u})\cdot\mathbf{n}\ dS = \int_V \mathbf{s}(\mathbf{u})\ dV\,, \tag{1}$$

where $\mathbf{u}$ is the array of the mean conserved quantities, e.g., density of mass, momentum, specific total energy, $\mathbf{f}$ is the array of the corresponding convective and diffusive fluxes, and $\mathbf{s}$ denotes a source term that may occur, for instance, in turbulence modeling. For simplicity of representation, we will always assume that V is time-independent. In principle, the concepts presented below can easily be extended to moving boundaries, cf. [15, 52].

The balance equations (1) are approximated by a finite volume scheme. For this purpose the finite fluid domain $\Omega \subset \mathbf{R}^d$ is split into a finite set of subdomains, the cells V_i, such that all V_i are disjoint and their union covers Ω. According to our simplifying assumption, the grid does not move in time. Furthermore let $N(i)$ be the set of cells that have a common edge with the cell i, and for $j \in N(i)$ let $\Gamma_{ij} := \partial V_i \cap \partial V_j$ be the interface between the cells i and j and $\mathbf{n}_{ij}$ the outer normal of Γ_{ij} corresponding to the cell i. In time we use a global time step τ^n for all cells that might change due to the Courant-Friedrich-Levy (CFL) condition, i.e., $t^{n+1} = t^n + \tau^{n+1}$, $t^0 = 0$. For the time discretization in (1) we confine to an explicit time discretization of the approximated cell averages $v_i^n \approx |V_i|^{-1} \int_{V_i} u(t^n, x)\, dx$ that can be written in the form

$$v_i^{n+1} = v_i^n - \frac{\tau_i^{n+1}}{|V_i|} (B_i^n + |V_i| S_i^n). \tag{2}$$

By this discrete evolution equation the approximated cell averages of the conserved variables are updated on the new time step. Here the fluxes and the source terms are approximated by

$$B_i^n := \sum_{j \in N(i)} |\Gamma_{ij}| F(v_{ij}^n, v_{ji}^n, \mathbf{n}_{ij}), \quad S_i^n := S(v_i^n), \tag{3}$$

where the numerical flux function $F(\mathbf{u}, \mathbf{w}, \mathbf{n})$ is an approximation for the flux $f(\mathbf{u}, \mathbf{n}) := \mathbf{f} \cdot \mathbf{n}$ in outer normal direction $\mathbf{n}_{ij}$ on the edge Γ_{ij}. The numerical flux is assumed to be *consistent*, i.e., $F(\mathbf{u}, \mathbf{u}, \mathbf{n}) = f(\mathbf{u}, \mathbf{n})$. For simplicity of presentation we neglect the fact that, due to higher order reconstruction, F usually depends on an enlarged stencil of cell averages. Moreover, to preserve a constant flow field we assume that the geometric consistency condition $\sum_{j \in N(i)} |\Gamma_{ij}| \mathbf{n}_{ij} = \mathbf{0}$ holds. This condition is easy to satisfy in case of planar faces. However, for more general discretizations, e.g. curvilinear grid patches [51], it imposes a constraint on the approximation of the normal vector $\mathbf{n}_{ij}$.

We want to remark that the finite volume discretization (2) is just a simplified prototype. More advanced discretizations can be considered where (i) the time discretization is replaced by some implicit scheme, cf. [15, 56, 57], or a Runge-Kutta scheme, cf. [63], (ii) the time stepsize is changing locally for each cell, cf. [15, 56], and (iii) the source term approximation is based on some higher order approximation.

3 Multiscale analysis

A finite volume discretization typically works on an array of cell averages. In order to realize a certain target accuracy at the expense of a possibly low number of degrees of freedom, viz. a possibly low computational effort, one should keep the size of the cells large wherever the data exhibit little variation, reflecting a high regularity of the searched solution components. Our analysis of the local regularity behavior of the data is based on the concept of biorthogonal wavelets [21]. This approach may be seen as a natural generalization of Harten's discrete framework [41]. The core ingredients are (i) a hierarchy of nested grids, (ii) biorthogonal wavelets and (iii) the multiscale decomposition. In what follows we will only summarize the basic ideas. For the realization and implementation see [54].

Grid hierarchy. Let $\Omega_l := \{V_\lambda\}_{\lambda \in I_l}$ be a sequence of different meshes corresponding to different resolution levels $l \in \mathbf{N}_0$, where the mesh size decreases with increasing refinement level. The grid hierarchy is assumed to be *nested*. This implies that each cell $\lambda \in I_l$ on level l is the union of cells $\mu \in M_\lambda^0 \subset I_{l+1}$ on the next higher refinement level $l+1$, i.e.,

$$V_\lambda = \bigcup_{\mu \in M_\lambda^0 \subset I_{l+1}} V_\mu, \quad \lambda \in I_l, \tag{4}$$

where $M_\lambda^0 \subset I_{l+1}$ is the refinement set. A simple example is shown in Figure 1 for a dyadic grid refinement of Cartesian meshes. Note that the framework presented here is not restricted to this simple configuration but can also be applied to *unstructured* grids and *irregular* grid refinements, cf. [54].

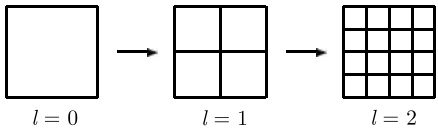


Fig. 1 Sequence of nested grids

Example. In the sequel, the concept will be illustrated for 1D dyadic grid refinements on the real axis. Then a nested grid hierarchy is determined by $\mathscr{G}_l := \{V_{l,k}\}_{k \in I_l}$, $l \in \mathbf{N}_0$, $I_l = \mathbf{Z}$. These meshes are composed of the intervals $V_{l,k} = [x_{l,k}, x_{l,k+1}]$ determined by the grid points $x_{l,k} = 2^{-l}k$, $k \in I_l$, with interval length $h_l = 2^{-l}$. Due to the subdivision $V_{l,k} = V_{l+1,2k} \cup V_{l+1,2k+1}$ the refinement set is determined by $M_{l,k}^0 = \{2k, 2k+1\}$. Here the index λ is identified by (l,k).

Box function and cell averages. With each cell V_λ in the partitions Ω_l we associate the so–called *box function*

$$\tilde{\phi}_\lambda(x) := \frac{1}{|V_\lambda|}\chi_{V_\lambda}(x) = \begin{cases} 1/|V_\lambda|, & x \in V_\lambda \\ 0, & x \notin V_\lambda \end{cases}, \quad \lambda \in I_l \tag{5}$$

defined as the L^1–normalized characteristic function of V_λ. By $|V|$ we denote the volume of a cell V. Then the averages of a scalar, integrable function $u \in L^1(\Omega)$ can be interpreted as an inner product, i.e.,

$$\hat{u}_\lambda := \langle u, \tilde{\phi}_\lambda \rangle_\Omega \quad \text{with} \quad \langle u, v \rangle_\Omega := \int_\Omega u\, v\, dx. \tag{6}$$

Obviously, the nestedness of the grids as well as the linearity of integration imply the two–scale relations

$$\tilde{\phi}_\lambda = \sum_{\mu \in M_\lambda^0 \subset I_l} m_{\mu,\lambda}^{l,0}\, \tilde{\phi}_\mu \quad \text{and} \quad \hat{u}_\lambda = \sum_{\mu \in M_\lambda^0 \subset I_l} m_{\mu,\lambda}^{l,0}\, \hat{u}_\mu, \quad \lambda \in I_{l-1}, \tag{7}$$

where the mask coefficients turn out to be $m_{\mu,\lambda}^{l,0} := |V_\mu|/|V_\lambda|$ for each cell $\mu \in M_\lambda^0$ in the refinement set.

Example. In case of the 1D dyadic grid refinement the box function is just $\tilde{\phi}_{l,k}(x) := 2^{-l}$ for $x \in V_{l,k}$ and zero elsewhere, see Figure 2 (left). The corresponding mask coefficients are $m_{r,k}^{l,0} := |V_{l+1,r}|/|V_{l,k}| = 0.5$ for $r \in M_{l,k}^0 \subset I_{l+1}$, $k \in I_l$. For a general grid hierarchy the mask coefficients may depend on the level and the position.

Wavelets and details. In order to detect singularities of the solution we consider the difference of the cell averages corresponding to different resolution levels. For this purpose we introduce the wavelet functions $\tilde{\psi}_\lambda$ as linear combinations of the box functions, i.e.,

$$\tilde{\psi}_\lambda := \sum_{\mu \in M_\lambda^1 \subset I_{l+1}} m_{\mu,\lambda}^{l,1}\, \tilde{\phi}_\mu, \quad \lambda \in J_l, \tag{8}$$

with mask coefficients $m_{\mu,\lambda}^{l,1}$ that only depend on the grids. Here the wavelet functions $\tilde{\Psi}_l := (\tilde{\psi}_\lambda)_{\lambda \in J_l}$ build an appropriate completion of the basis system $\tilde{\Phi}_l := (\tilde{\phi}_\lambda)_{\lambda \in I_l}$. By this we mean (i) they are locally supported, (ii) provide vanishing moments of a certain order and (iii) there exists a biorthogonal system Φ_l and Ψ_l of primal functions satisfying two-scale relations similar to (7) and (8). The last requirement is typically the hardest to satisfy. It is closely related to the Riesz basis property of the infinite collection $\tilde{\Phi}_0 \cup \bigcup_{l=0}^\infty \tilde{\Psi}_l$ of $L_2(\Omega)$. For details we refer to the *concept of stable completions*, see [21].

Aside from these stability aspects, the biorthogonal framework allows for an efficient change of basis. While the relations (7) and (8) provide expressions of the coarse scale box functions and detail functions as linear combinations of fine scale box functions, the mask coefficients in the analogous two-scale relations for the dual system Φ_l, Ψ_l give rise to the reverse change of basis between $\tilde{\Phi}_l \cup \tilde{\Psi}_l$ and $\tilde{\Phi}_{l+1}$, i.e.,

$$\tilde{\phi}_\lambda = \sum_{\mu \in G_\lambda^0 \subset I_l} g_{\mu,\lambda}^{l,0}\, \tilde{\phi}_\mu + \sum_{\mu \in G_\lambda^1 \subset J_l} g_{\mu,\lambda}^{l,1}\, \tilde{\psi}_\mu, \quad \lambda \in I_{l+1}, \tag{9}$$

where we rewrite the basis function $\tilde{\phi}_\lambda$ on level $l+1$ by the scaling functions $\tilde{\phi}_\mu$ and the wavelet functions $\tilde{\psi}_\mu$ on the next coarser scale l. Here again the mask coefficients $g^{l,0}_{\mu,\lambda}$ and $g^{l,1}_{\mu,\lambda}$ depend only on the grid geometry.

Biorthogonality also yields a data representation in terms of the primal system Ψ. The expansion coefficients d_λ with respect to the basis Ψ are obtained by testing u with the elements from $\tilde{\Psi}$, i.e.,

$$d_\lambda := \langle u, \tilde{\psi}_\lambda \rangle_\Omega = \sum_{\mu \in M^1_\lambda} m^{l,1}_{\mu,\lambda}\, \hat{u}_\mu, \quad \lambda \in J_l. \tag{10}$$

These are referred to as the *detail coefficients*. Their two-scale format follows from the functional counterpart of (8).

Note that the dual system $\tilde{\Psi}$ is used to expand the cell averages which are *functionals* of the solution u whose propagation in time gives rise to the finite volume scheme. The primal basis itself will actually never be used to represent the solution u. Instead, the enhanced accuracy of the approximate cell averages can be used for higher order reconstructions commonly used in finite volume schemes.

Example. In case of the 1D dyadic grid refinement, the L^1-normalized Haar wavelet $\tilde{\psi}^H_{l,k} := (\tilde{\phi}_{l+1,2k} + \tilde{\phi}_{l+1,2k+1})/2$ can be used, see Figure 2 (middle). The corresponding mask coefficients are $m^{l,1}_{r,k} := 0.5$ for $r \in M^1_{l,k} \equiv M^0_{l,k} \subset I_{l+1}$, $k \in I_l$. For a general grid hierarchy the mask coefficients may depend on the level and the position.

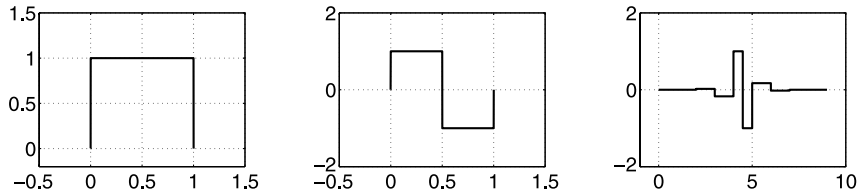


Fig. 2 Box function $\tilde{\phi}_{0,0}$ (left), Haar wavelet $\tilde{\psi}^H_{0,0}$ (middle), and modified Haar wavelet $\tilde{\psi}_{0,0}$ with $s = 2$ (right)

Cancellation Property. It can be shown that the details become small with increasing refinement level when the underlying function is smooth

$$|d_\lambda| \le C 2^{-lM} \|u^{(M)}\|_{L^\infty(V_\lambda)} \tag{11}$$

in the support of the wavelet $\tilde{\psi}_\lambda$. More precisely, the details decay at a rate of at least 2^{-lM}, provided that the function u is sufficiently differentiable and the wavelets have vanishing moments of order M, i.e.,

$$\langle p, \tilde{\psi}_\lambda \rangle_\Omega = 0 \tag{12}$$

for all polynomials p of degree less than M. Here we assume that the grid hierarchy is quasi-uniform in the sense that the diameters of the cells on each level l are proportional to 2^{-l}.

If coefficient and function norms behave essentially the same, as asserted by the Riesz basis property, (11) suggests to neglect all sufficiently small details in order to compress the original data. In fact, the higher M the more details may be discarded in smooth regions.

Example. The Haar wavelet has only one vanishing moment as can be easily checked from its definition. Then (10) implies that the corresponding details vanish when the function u is locally constant.

Higher vanishing moments. In order to realize a better compression by exploiting a higher order smoothness we have to raise the order of vanishing polynomial moments. The basic idea is first to construct the box wavelets $\tilde{\psi}_\lambda^H$, $\lambda \in I_l$, cf. [37, 54], and then to modify the box wavelet by some coarse grid box functions $\tilde{\phi}_\mu$, $\mu \in I_l$, leading to the ansatz

$$\tilde{\psi}_\lambda := \tilde{\psi}_\lambda^H + \sum_{\mu \in L_\lambda} l_\mu^\lambda \tilde{\phi}_\mu, \tag{13}$$

with parameters l_μ^λ that are still to be determined. Here the stencil $L_\lambda \subset I_l$ denotes a finite number of cells V_μ in the local neighborhood of the cell V_λ. Then the parameters l_μ^λ are chosen such that (12) holds for all polynomials p of degree less than M. This will lead to a linear system of equations for the coefficients l_μ^λ. In higher dimensions, the cardinality of the stencil L_λ is typically chosen larger than the number of conditions imposed by (12). Then the under-determined system can be solved using the Moore-Penrose inverse.

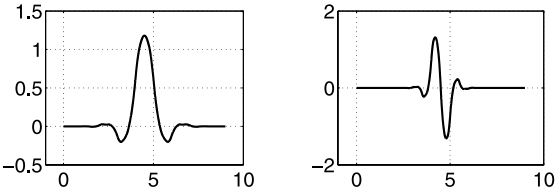


Fig. 3 Primal scaling function $\phi_{0,0}$ (left) and primal wavelet $\psi_{0,0}$ (right) corresponding to the modified Haar wavelet $\tilde{\psi}_{0,0}$ with $s = 2$

Example. Modified Haar wavelets with higher vanishing moments $M = 2s+1$ can be obtained according to the above procedure where we choose $L_{l,k} = \{k - s, \ldots, k+s\}$. In this particular case, the resulting linear system has a unique solution. Furthermore, there exists a primal system of scaling and wavelet functions that is biorthogonal to the dual system of the box function and the modified Haar wavelet. For $s = 2$ the modified Haar wavelet and the corresponding primal functions are shown in Figures 2 (right) and 3, respectively. The biorthogonal system coincides with the system derived from the pair ${}_1\tilde{\Phi}$, ${}_{1,\tilde{N}}\tilde{\Psi}$ and ${}_1\Phi$, ${}_{1,\tilde{N}}\Psi$ corresponding to the B-spline function ${}_1\tilde{\Phi} = \chi_{[0,1]}$ of order 1 with $\tilde{N} = M = 2s+1$ as constructed in [24]. Note that for our purposes the dual and the primal functions are normalized with respect to L^1 and L^∞, respectively, instead of L^2 in [24].

Multiscale Transformation. In order to exploit the above compression potential, the idea is to transform the array of cell averages $u_L := (\hat{u}_\lambda)_{\lambda \in I_L}$ corresponding to a finest uniform discretization level into a sequence of coarse grid data $u_0 := (\hat{u}_\lambda)_{\lambda \in I_0}$

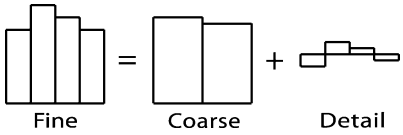


Fig. 4 Two-scale Transformation

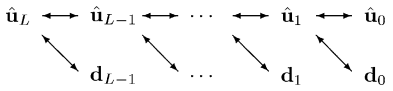


Fig. 5 Multiscale transformation

and details $d_l := (d_\lambda)_{\lambda \in J_l}$, $l = 0, \ldots, L-1$, representing the successive update from a coarser resolution to a higher resolution.

In summary, according to (7) and (10), the change of basis provides two–scale relations for the coefficients inherited from the two–scale relations of the box functions and the wavelet functions

$$\hat{u}_\lambda = \sum_{\mu \in M_\lambda^0 \subset I_{l+1}} m_{\mu,\lambda}^{l,0}\, \hat{u}_\mu, \quad \lambda \in I_l, \qquad d_\lambda = \sum_{\mu \in M_\lambda^1 \subset I_{l+1}} m_{\mu,\lambda}^{l,1}\, \hat{u}_\mu, \quad \lambda \in J_l, \tag{14}$$

and, conversely,

$$\hat{u}_\lambda = \sum_{\mu \in G_\lambda^0 \subset I_l} g_{\mu,\lambda}^{l,0}\, \hat{u}_\mu + \sum_{\mu \in G_\lambda^1 \subset J_l} g_{\mu,\lambda}^{l,1}\, d_\mu, \qquad \lambda \in I_{l+1}, \tag{15}$$

which reflects the typical cascadic format of a wavelet transform. The two-scale relations are illustrated for the 1D case in Figure 4.

A successive application of the relations (14), see Figure 5, decomposes the array $\hat{u}_L$ into coarse scale averages and higher level fluctuations. We refer to this transformation as the *multiscale transformation*. It is inverted by the *inverse multiscale transformation* (15).

4 Multiscale-based spatial grid adaptation

To determine a locally refined grid we employ the above multiscale decomposition. The basic idea is to perform data compression on the vector of detail coefficients using hard thresholding as suggested by the cancellation property. This will significantly reduce the complexity of the data. Based on the thresholded array we then perform local grid adaptation where we refine a cell whenever there exists a significant detail, i.e. a detail coefficient with absolute value above the given threshold. The main steps in this procedure are summarized in the following:

Step 1: Multiscale analysis. Let v_L^n be the cell averages representing the discretized flow field at some fixed time step t^n on a given locally refined grid with highest level of resolution $l = L$. This sequence is encoded in arrays of *detail coefficients* d_l^n, $l = 0, \ldots, L-1$ of ascending resolution, see Figure 5, and cell averages on some coarsest level $l = 0$. For this purpose the multiscale transformation (14) needs to be performed *locally* which is possible due to the locality of the mask coefficients.

Step 2: Thresholding. In order to compress the original data we discard all detail coefficients d_λ whose absolute values fall below a level-dependent threshold value $\varepsilon_l = 2^{l-L}\varepsilon$. Let

$$D^n_{L,\varepsilon} := \left\{\lambda \ ; \ |d^n_\lambda| > \varepsilon_l, \ \lambda \in I_l, \ l \in \{0,\ldots,L-1\}\right\}$$

be the set of *significant details*. The ideal strategy would be to determine the threshold value ε such that the *discretization error* of the reference scheme, i.e., difference between exact solution and reference scheme, and the *perturbation error*, i.e., the difference between the reference scheme and the adaptive scheme, are balanced. For a detailed treatment of this issue we refer to [25].

Step 3: Prediction and grading. Since the flow field evolves in time, grid adaptation is performed after each evolution step to provide the adaptive grid at the *new* time step. In order to guarantee the adaptive scheme to be *reliable* in the sense that no significant future feature of the solution is missed, we have to *predict* all significant details at the new time step $n+1$ by means of the details at the *old* time step n. Let $\tilde{D}^{n+1}_{L,\varepsilon}$ be the prediction set satisfying the *reliability condition*

$$D^n_{L,\varepsilon} \cup D^{n+1}_{L,\varepsilon} \subset \tilde{D}^{n+1}_{L,\varepsilon}. \tag{16}$$

Basically there are two prediction strategies (i.e. ways of choosing $\tilde{D}^{n+1}_{L,\varepsilon}$) discussed in the literature, see [40, 25]. Moreover, in order to perform the grid adaptation process, this set is additionally inflated somewhat such that the grid refinement history, i.e., the parent-child relations of subdivided cells, corresponds to a *graded* tree. Then the set of significant details can be interpreted as a graph where all details are connected by an edge in the graph.

Step 4: Grid adaptation. By means of the set $\tilde{D}^{n+1}_{L,\varepsilon}$ a locally refined grid is determined along the following lines. We check for the transformed flow data represented on $\tilde{D}^{n+1}_{L,\varepsilon}$ proceeding levelwise from coarse to fine whether the detail associated with any cell marked by the prediction set is significant or not. If it is, we refine the respective cell. We finally obtain the locally refined grid with hanging nodes represented by the index set $\tilde{G}^{n+1}_{L,\varepsilon}$. The flow data on the new grid can be computed from the detail coefficients in the same loop where we locally apply the inverse multiscale transformation (15).

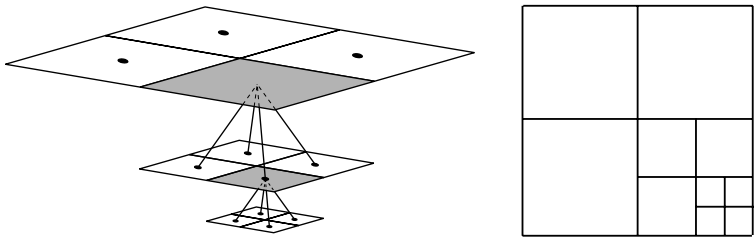


Fig. 6 Grid adaptation: refinement tree (left) and corresponding adaptive grid (right)

5 Adaptive multiresolution finite volume schemes

The rationale behind our design of adaptive multiresolution finite volume schemes (MR-FVS) is to accelerate a given finite volume scheme (reference scheme) on a uniformly refined mesh (reference mesh) by computing actually *only* on a locally refined adapted subgrid, while preserving (up to a fixed constant multiple) the accuracy of the discretization on the full uniform grid. We shall briefly indicate now how to realize this strategy with the aid of the ingredients discussed in the previous section.

5.1 From the reference scheme to an adaptive scheme

The conceptual starting point is to rewrite the evolution equations (2) for the cell averages v_λ, $\lambda \in I_L$, of the reference scheme in terms of evolution equations for the multiscale coefficients. For this purpose we apply the multiscale transformation (14) to the set of evolution equations (2). Then we discard all equations that do not correspond to the prediction set $\tilde{D}^{n+1}_{L,\varepsilon}$ of significant details. Finally we apply locally the inverse multiscale transformation (15) and obtain the evolution equations for the cell averages on the adaptive grid $\tilde{G}^{n+1}_{L,\varepsilon}$ which is obtained from $\tilde{D}^{n+1}_{L,\varepsilon}$ as explained before:

$$v^{n+1}_\lambda = v^n_\lambda - \lambda_\lambda \, (\overline{B}^n_\lambda + |V_\lambda| \overline{S}^n_\lambda), \tag{17}$$

for all $\lambda \in \tilde{G}^{n+1}_{L,\varepsilon}$ where $\lambda_\lambda := \Delta t^{n+1}/|V_\lambda|$. Here the flux balances $\overline{B}^n_\lambda$, the numerical fluxes $\overline{F}^n_\lambda$ and the source terms $\overline{S}^n_\lambda$ are recursively defined from fine to coarse scale via

$$\overline{B}^n_\lambda = \sum_{\Gamma^l_{\lambda,\mu} \subset \partial V_\lambda} |\Gamma^l_{\lambda,\mu}| \overline{F}^{l,n}_{\lambda,\mu}, \tag{18}$$

$$\overline{F}^{l,n}_{\lambda,\mu} = \sum_{\Gamma^{l+1}_{\mu,\nu} \subset \Gamma^l_{\lambda,\mu}} |\Gamma^{l+1}_{\mu,\nu}| \overline{F}^{l+1,n}_{\mu,\nu} = \ldots = \sum_{\Gamma^L_{\mu,\nu} \subset \Gamma^l_{\lambda,\mu}} |\Gamma^L_{\mu,\nu}| F(v^n_{L,\mu\nu}, v^n_{L,\nu\mu}, \mathbf{n}_{L,\mu\nu}), \tag{19}$$

$$\overline{S}^n_\lambda = \sum_{V_\mu \subset V_\lambda, \mu \in I_{l+1}} \frac{|V_\mu|}{|V_\lambda|} \overline{S}^n_\mu = \ldots = \sum_{V_\mu \subset V_\lambda, \mu \in I_L} \frac{|V_\mu|}{|V_\lambda|} S(v^n_\mu). \tag{20}$$

We refer to (19) and (20) as *exact flux and source reconstruction*, respectively. Since in (20) we have to compute *all* sources on the finest scale, there is no complexity reduction, i.e., we still have the complexity $\#I_L$ of the reference mesh. In order to gain efficiency we therefore have to replace the exact flux and source reconstruction by some approximation such that the overall accuracy is maintained. The local flux and source computation and the choice of the prediction set $\tilde{D}^{n+1}_{L,\varepsilon}$ will be discussed in detail in Section 5.2 and 5.3, respectively.

The complete adaptive scheme consists now of the following three steps:

Step 1. (*Refinement*) Determine the prediction set $\tilde{D}^{n+1}_{L,\varepsilon}$ from the data of the old

time step t^n and project the data of the old time step onto the pre-refined grid $\tilde{G}_{L,\varepsilon}^{n+1}$ of the new time step, i.e.,

$$\{v_\lambda^n\}_{\lambda\in G^n} \to \{v_\lambda^n\}_{\lambda\in\tilde{G}^{n+1}}.$$

Step 2. (*Evolution*) Evolve the cell averages associated to the pre-refined grid $\tilde{G}_{L,\varepsilon}^{n+1}$ according to (17), where the numerical fluxes and sources are not necessarily determined by (19) and (20), respectively, i.e.,

$$\{v_\lambda^n\}_{\lambda\in\tilde{G}_{L,\varepsilon}^{n+1}} \to \{v_\lambda^{n+1}\}_{\lambda\in\tilde{G}_{L,\varepsilon}^{n+1}}.$$

Step 3. (*Coarsening*) Compress the data of the new time step by thresholding the corresponding detail coefficients and project the data to the (somewhat coarsened new) adaptive grid $G_{L,\varepsilon}^{n+1}$, i.e.,

$$\{v_\lambda^{n+1}\}_{\lambda\in\tilde{G}_{L,\varepsilon}^{n+1}} \to \{v_\lambda^{n+1}\}_{\lambda\in G_{L,\varepsilon}^{n+1}}.$$

5.2 Approximate flux and source approximation strategies

As already mentioned above, the adaptive MR-FVS with exact flux and source reconstruction (19) and (20) will have the same complexity as the reference scheme performed on the reference mesh. If there is no inhomogeneity, i.e., $s \equiv 0$, then the complexity of the resulting algorithm might be significantly reduced from the cardinality of the reference mesh to the cardinality of the refined mesh. To see this we note that, due to the nestedness of the grid hierarchy and the conservation property of the numerical fluxes, the coarse-scale flux balances are only computed by the fine-scale fluxes corresponding to the edges of the coarse cells, see (19). Those in turn, have to be determined by the fine scale data. However, the internal fluxes are canceled and, hence, the overall complexity is reduced. For instance, for a d-dimensional Cartesian grid hierarchy we would have to compute $2d\,2^{(L-l)(d-1)}$ fluxes corresponding to all fine-scale interfaces $\mu \in I_L$ with $\partial V_\mu \subset \partial V_\lambda$ where $\lambda \in I_l$, $l \le L$, due to the subdivision of the cell faces. Note that in both cases missing data on the finest scale have to be determined by locally applying the inverse two-scale transformation. This is illustrated in Figure 7. On the other hand, the coarse scale sources can be computed similarly with the aid of the recursive formulae (20). Here, however, we have to compute *all* sources on the finest scale which at the first glance prevents the desired complexity reduction.

Hence the adaptive scheme with both exact flux and source reconstruction is useless for practical purposes. However, in the reliability analysis one may perform the adaptive scheme with some approximate flux and source reconstruction to be considered as a further perturbation of the "exact" adaptive scheme.

Fig. 7 Exact (left) versus local (right) flux and source computation

In order to retain efficiency we therefore have to replace the exact flux and source reconstruction by some approximation such that the overall accuracy is maintained. A naive approach would be to use the local data provided by the adaptive grid, i.e.,

$$\overline{F}^{l,n}_{\lambda,\mu} = F(v^n_{l,\lambda\mu}, v^n_{l,\mu\lambda}, \mathbf{n}_{l,\lambda\mu}), \quad \overline{S}^n_\lambda = S(v^n_\lambda) \tag{21}$$

for $\lambda, \mu \in I_l$.

So far, this approach is applied in Quadflow. Obviously, the complexity of the resulting adaptive MR-FVS is reduced to the cardinality of the adaptive grid. Unfortunately, this approach may suffer from serious loss in accuracy in comparison with the reference scheme.

Recently, in [44] a new approach was suggested using an approximate flux and source reconstruction strategy that are discussed along the following lines:

Step 1. Determine for each cell V_λ, $\lambda \in \tilde{G}^{n+1}_{L,\varepsilon}$, a higher order reconstruction polynomial R^N_λ of degree N using only local data corresponding to the adaptive grid.
Step 2. Approximate the boundary and volume integrals in (19) and (20) by some appropriate quadrature rules.
Step 3. Compute fluxes and source terms in quadrature nodes by determining point-values or cell averages on level L of the local reconstruction polynomial R^N_λ, respectively.

This concept has been analyzed in detail for the 1D case, cf. [44]. In particular, it was proven that the accuracy of the reference scheme can be maintained when using the prediction strategy in [25] and appropriately tuning the parameters such as the reconstruction order and the quadrature rules. Computations verify the analytical results. Therefore the new approach seems to be superior to the naive approach with respect to accuracy and efficiency.

5.3 Prediction strategies

The accuracy of the adaptive scheme crucially relies on the grid refinement process. In our case it is triggered by the details. In order to guarantee that all significant flow features are always adequately resolved, we have to pre-refine the grid before performing the time evolution. For this purpose, we have to *predict* all details $\tilde{D}_{L,\varepsilon}^{n+1}$ on the *new* time step that may become significant due to the evolution by means of the details $D_{L,\varepsilon}^{n}$ on the *old* time step. We consider the prediction set $\tilde{D}_{L,\varepsilon}^{n+1}$ to be reliable, if the reliability condition (16) is satisfied in each time step where, of course, $D_{L,\varepsilon}^{n+1}$ is not known yet. Then no significant future feature of the solution is missed on the old and the new time step, respectively.

Harten's strategy. A first strategy was proposed by Harten in [40]. The basic idea of his heuristic approach is based on two characteristic features of hyperbolic conservation laws: (i) details in a local neighborhood $N_\lambda^q := \{\mu \in I_l\,;\, \|\mu - \lambda\|_\infty \le q\}$ of a significant detail $\lambda \in I_l$ may also become significant within one time step, i.e.,

$$\lambda \in D_{L,\varepsilon}^{n} \quad \Rightarrow \quad \tilde{D}_{L,\varepsilon}^{n+1} = \tilde{D}_{L,\varepsilon}^{n+1} \cup N_\lambda^q, \tag{22}$$

due to the finite speed of propagation, and (ii) gradients may become steeper causing significant details on a higher refinement level due to the developing of discontinuities, i.e.,

$$\lambda \in D_{L,\varepsilon}^{n} \quad \Rightarrow \quad \tilde{D}_{L,\varepsilon}^{n+1} = \tilde{D}_{L,\varepsilon}^{n+1} \cup M_\lambda^0, \tag{23}$$

where $M_\lambda^0 \subset I_{l+1}$ is the refinement set of cell V_λ, $\lambda \in I_l$. Note that the choice of q in (22) depends on the CFL number. If the CFL number is less than 1, that is reasonable for explicit time discretizations, we may choose $q = 1$. However, in case of an implicit time discretization higher CFL numbers might be admissible. In this case an information could move by more than one cell and we have to adjust q accordingly. In general, the range of influence of an information within one time step depends on the configuration at hand. If the flow field is weakly instationary, cf. [69], or even stationary, cf. [15], then an information will not move by as many cells as is indicated by the CFL number, cf. [27]. This also holds in case of small parabolic perturbations due to viscosity terms, cf. [11].

So far Harten's approach could not be rigorously verified to satisfy (16). Nevertheless, it is frequently used in applications and turned out to give good results.

Strategy by Cohen et al.. A slight modification of Harten's prediction strategy has been shown to lead to a reliable prediction strategy in the sense of (16). This was rigorously proven for a certain class of *explicit* finite volume schemes applied to *one–dimensional scalar* conservation laws *without source terms* on *uniform dyadic grids* as base hierarchies, using exact flux reconstruction, cf. [25]. Recently, the proof has been extended for conservation laws *with* source term using approximate flux and source reconstruction, cf. [44]. In the following we briefly summarize the strategy. For simplicity of representation, we first introduce the convention $d_\lambda := v_\lambda$ for $\lambda \in I_{-1}$, where we identify I_{-1} with I_0 but replace the level $l = 0$ by $l = -1$. Then the prediction set can be determined in three steps:

Step 1: First of all, we determine the influence set D_λ that contains all coefficients d_μ^{n+1} on the new time step which are influenced by a coefficient d_λ^n on the old time step. For this purpose, we first have to compute the *range of influence* Σ_λ of the coefficient d_λ^n and the *domain of dependence* $\tilde{\Sigma}_\mu$ of the coefficient d_μ^{n+1}. In the range of influence we collect the indices of all averages v_ν^n, $\nu \in I_L$, that are influenced by the detail d_λ^n whereas the domain of dependence contains the indices of all averages v_ν^{n+1}, $\nu \in I_L$, that are needed to compute the coefficient d_μ^{n+1}. Note that the index sets $\tilde{\Sigma}_\mu \subset I_L$ and $\Sigma_\lambda \subset I_L$ correspond to data on the reference mesh but for different time steps, $n+1$ and n, respectively. By the evolution process (17) with exact reconstruction (19) and (20), the domain of dependence has to be extended taking into account the stencil $S_\lambda \subset I_l$ of the numerical flux F and source S associated to the cell $\lambda \in I_l$, i.e., $\tilde{\Sigma}_\mu^- := \bigcup_{\lambda \in \tilde{\Sigma}_\mu} S_\lambda$. Then the influence set is determined by

$$D_\lambda = \{\mu \,;\, \tilde{\Sigma}_\mu^- \cap \Sigma_\lambda \neq \emptyset\}.$$

Step 2: The prediction strategy has to take into account that the coefficients d_λ^n may not only cause a perturbation in the neighborhood of the cell V_λ, $\lambda \in I_l$, because of the time evolution but may also influence coefficients d_μ^{n+1}, $\mu \in I_{l'}$, on higher scales, i.e., $l' \geq l+1$. Since the additional higher levels inflate the influence set, we would like to bound the number of higher levels to a minimum number. For this purpose, we fix some $\sigma > 1$ and assign to each coefficient corresponding to $\lambda \in D^n$ a unique index $\nu = \nu(\lambda)$ such that

$$\begin{aligned} 2^{\nu(\lambda)\sigma}\varepsilon_l < |d_\lambda^n| \leq 2^{(\nu(\lambda)+1)\sigma}\varepsilon_l, \quad \lambda \in I_l,\ l \in \{0,\dots,L-1\},\\ 2^{\nu(\lambda)\sigma}\varepsilon_0 < |v_\lambda^n| \leq 2^{(\nu(\lambda)+1)\sigma}\varepsilon_0, \quad \lambda \in I_0. \end{aligned}$$

This process is referred to as *nesting of details*. The parameter σ is linked to the smoothness of the primal wavelet functions, cf. [25]. Since the index $\nu(\lambda)$ becomes smaller the larger σ is, it is convenient to choose σ as large as possible.

Step 3: Finally, we determine the prediction set from the influence set D_λ and the nesting of coefficients

$$\tilde{D}_{L,\varepsilon}^{n+1} := D_{L,\varepsilon}^n \cup \bigcup_{\lambda \in D_{L,\varepsilon}^n \cup I_{-1}} \{\mu \,;\, \mu \in D_\lambda \backslash I_{-1} \text{ and } l' \leq l + \nu(\lambda)\}. \tag{24}$$

Note that opposite to Harten's original prediction strategy, a significant detail might affect cells not only at one higher level but up to $\nu(\lambda)$ additional scales.

5.4 Multilevel time stepping

For instationary problems, the time step is typically restricted for stability reasons by some CFL condition. This holds true even for implicit time discretizations due to nonlinear stability criteria, e.g., total variation diminishing (TVD) property. There-

fore the time stepsize has to be bounded by the smallest cell in the grid. Hence τ is determined by the CFL condition on the highest refinement level L, i.e., $\tau = \tau_L$. For reasons of simplicity, we neglect the time index n here. However, for cells on the coarser scales $l = 0, \ldots, L-1$ we may use $\tau_l = 2^{L-l}\,\tau_L$ to locally satisfy the CFL condition.

In [56] a local time stepping strategy has been incorporated into the adaptive multiresolution finite volume scheme as presented in previous sections. This strategy has been extended to multidimensional problems in [53, 52]. Here ideas similar to the predictor-corrector scheme [58] and the adaptive mesh refinement (AMR) technique [10, 9] are used. The differences between classical approaches and the multilevel strategy are discussed in [56] in detail.

Time evolution. The basic idea is to evolve each cell on level l with the level-dependent time discretization $\tau_l = 2^{L-l}\,\tau_L$, $l = 0, \ldots, L$. Obviously, after having performed 2^l time steps with τ_l, all cell averages correspond to the same integration time, i.e., the cells are *synchronized*. Therefore one macro time step with $\tau_0 = 2^L\,\tau_L$ consists of 2^L intermediate time steps with step size τ_L. Obviously, at time $t_{n+i2^{-L}}$ all cells on levels $l = l_i, \ldots, L$ are synchronized. Here l_i denotes the *smallest synchronization level* that is determined by

$$l_i := \min\{l;\, 0 \le l \le L,\ i \bmod 2^{L-l} = 0\}, \quad i = 0, \ldots, 2^L - 1.$$

Then the time evolution for the intermediate time steps $i = 0, \ldots, 2^L - 1$, takes the form

$$v_\lambda^{n+(i+1)2^{-L}} = v_\lambda^{n+i2^{-L}} - \lambda_\lambda\,(\overline{B}_\lambda^{n+i2^{-L}} + |V_\lambda|\,\overline{S}_\lambda^{n+i2^{-L}}), \tag{25}$$

for any cell $\lambda \in \tilde{G}_{L,\varepsilon}$ of the current locally adapted grid. Similar to (18) the numerical flux balance is determined by

$$\mathbf{B}_\lambda^{n+i2^{-L}} = \sum_{\Gamma_{\lambda,\mu}^l \subset \partial V_\lambda} |\Gamma_{\lambda,\mu}^l|\,\overline{F}_{\lambda,\mu}^{l,n+i2^{-L}}.$$

However, the numerical flux computation is performed differently. Here the basic idea is (i) to update the fluxes on the synchronized levels $l_i \le l \le L$, whereas (ii) for all other interfaces we do not update the numerical flux but use the same value as in the previous intermediate time step. In detail, we proceed as follows: (i) if the neighbor cell V_μ is living on the same level l, then we apply the flux computation strategy as in case of global time stepping, where we either use the exact strategy (19), the naive strategy (21), or the reconstruction strategy according to Section 5.2, respectively. Alternatively, the neighboring cell could live on the finer level $l+1$ due to grid refinement. Then there exist hanging nodes at the interface $\Gamma_{\lambda,\mu}^l$. and we compute the numerical flux by the fluxes on the finer scale, i.e.,

$$\overline{F}_{\lambda,\mu}^{l,n+i2^{-L}} = \sum_{\Gamma_{\mu,\nu}^{l+1} \subset \Gamma_{\lambda,\mu}^l} |\Gamma_{\mu,\nu}^{l+1}|\,\overline{F}_{\mu,\nu}^{l+1,n+i2^{-L}}. \tag{26}$$

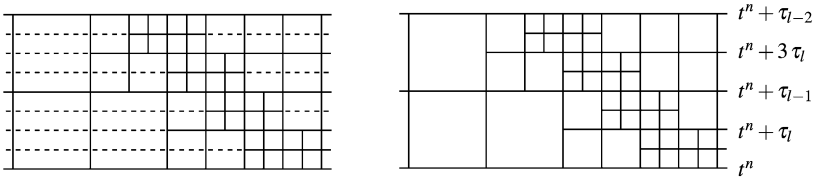


Fig. 8 Synchronized time evolution on space-time grid, horizontal axis: 1D space, vertical axis: time

This is motivated by (19) and immediately implies the conservation property of the scheme. Note that the refinement level of two adjacent cells differs by at most one, i.e., there is at most one hanging node at one edge. This can be ensured by a grading process of the adaptive grid, cf. Section 4. (ii) For all other interfaces in the adaptive grid we use the flux of the previous intermediate time step, i.e.,

$$\overline{F}_{\lambda,\mu}^{l,n+i2^{-L}} = |\Gamma_{\mu,\nu}^{l}|\,\mathbf{F}_{\lambda}^{n+(i-1)2^{-L}}. \tag{27}$$

To ensure that the fluxes at an interface with hanging nodes have already been computed when determining the corresponding flux on the coarser level, we perform in each intermediate time step the time evolution first for the cells on the highest level and then successively for the coarser levels. This procedure is similar to the predictor-corrector method in [58].

The source terms are updated accordingly, where we either apply the naive strategy (21) or the reconstruction strategy, cf. Section 5.2, respectively, on the synchronized levels $l_i \le l \le L$ or use the source term from the previous intermediate time step for the non-synchronized levels $l < l_i$, i.e.,

$$\mathbf{S}_{\lambda}^{n+i2^{-L}} = \mathbf{S}_{\lambda}^{n+(i-1)2^{-L}}. \tag{28}$$

Note that for the lower levels $0,\dots,l_i-1$ we do not compute new fluxes or source terms. This makes the local time stepping version of the adaptive multiresolution concept more efficient then the standard approach using a global time stepsize. However, book-keeping of the interfaces with hanging nodes is time consuming and the algorithms become hard to read and to implement, cf. [56, 27]. In practice, it is more convenient to perform the time evolution (25) for *all* cells of the adaptive grid for *all* intermediate time steps. Then all data are synchronized at any time. Of course, there is a small overhead to perform (25) for non-synchronized level $l < l_i$. However, this is negligible in comparison to the time needed to evaluate the original numerical fluxes that typically requires the solution of some Riemann problem. Then only few changes are needed to embed the multilevel time stepping into an existing code.

In Figure 8 the time evolution algorithm is schematically illustrated in the one-dimensional case: In a global time stepping, i.e., using $\Delta t = \tau_L$ for all cells, each vertical line section appearing in Fig. 8 (left) represents a flux evaluation and each horizontal line (dashed or solid) represents a cell update of the cell average due to

the fluxes. In the multilevel time stepping a flux evaluation is only performed at vertical line sections that emanate from a point where at least one solid horizontal line section is attached. If a vertical line section emanates from a point, where two dashed horizontal sections are attached, then we do not recompute the flux, but keep the flux value from the preceeding vertical line section. Hence fluxes are only computed for the vertical edges in Fig. 8 (right).

Intermediate grid adaptation. Finally, we have to comment on the grid adaptation step. The ultimate goal is to provide an approximation after one macro time step with $\tau_0 = 2^L \tau_L$ as good as having performed 2^L time steps with the reference scheme on the reference mesh using the time step size τ_L. Therefore we have to make sure that the solution is adequately resolved at each intermediate time step.

For the original adaptive multiresolution scheme this is ensured by the prediction step of the grid adaption, see Section 5.3. The prediction of the details ensures that a significant information can only move by at most one cell on the *finest* level, e.g. controlled by parameter q in (22) typically set to 1. However, by employing the same strategy for the local time stepping this information could move up to one cell on the *coarsest* mesh or 2^L cells on the *finest* mesh, respectively. This would result in a completely underresolution of discontinuities on the new time step. To account for this we have to modify the prediction set $\tilde{D}_{L,\varepsilon}^{n+1}$ such that the modified reliability condition

$$\bigcup_{i=0}^{2^L} D_{L,\varepsilon}^{n+i2^{-L}} \subset \tilde{D}_{L,\varepsilon}^{n+1}, \tag{29}$$

holds where the sets $D_{L,\varepsilon}^{n+i2^{-L}}$ correspond to the significant details of the solution at the intermediate times $t_{n+i2^{-L}} = t_n + i\,\tau_L$, $i = 0, \ldots, 2^L$.

Obviously, using $q = 2^L$ would ensure that all effects are properly resolved on the new time step after having performed the macro time step. However, the efficiency degrades tremendously. A very efficient and reliable alternative was suggested in [56]. The idea is to perform additional grid adaptation steps according to Section 4 before each even intermediate time step, i.e., $i = 0, 2, \ldots, 2^L - 2$. However, we do not apply the adaptation process for the whole computational domain, but only for the cells on the levels $l = l_i, \ldots, L$, i.e., level l_i is considered to be the coarsest scale in the multiscale analysis. Note, that only for this range of scales new fluxes and sources have to be recomputed. This process provides us with the sets $G_{L,\varepsilon}^{n+(i+1)2^{-L}}$ for which we perform the evolution step (25). For the odd intermediate time steps we use the same grid as in the previous step, i.e., $G_{L,\varepsilon}^{n+i2^{-L}} = G_{L,\varepsilon}^{n+(i-1)2^{-L}}$, $i = 1, 3, \ldots, 2^L - 1$. Hence, it is possible to track, for instance, the shock position on the intermediate time steps instead of a-priori refining the whole range of influence, see Fig. 8 (right).

5.5 Error analysis

The performance of the adaptive MR-FVS crucially depends on the threshold parameter ε. With decreasing value the adaptive grid becomes richer and, finally, if ε tends to zero, we obtain the uniform reference mesh, i.e., the adaptive scheme coincides with the reference scheme. On the other hand, the adaptive grid becomes coarser with increasing threshold value, i.e., the computation becomes faster but provides a less accurate solution. An ideal choice would maintain the accuracy of the reference scheme at reduced computational cost. For a detailed analysis we refer to [25, 44] and explain only the main ideas here.

In order to estimate the error, we introduce the averages $\hat{u}_L^n$ of the exact solution, the averages v_L^n determined by the reference FVS and the averages $\overline{v}_L^n$ of the adaptive scheme prolongated to the reference mesh by means of the inverse multiscale transformation where non-significant details are simply set to zero. Ideally one would like to choose the threshold ε so as to guarantee that $\|\hat{u}_L^n - \overline{v}_L^n\| \leq tol$ where tol is a given target accuracy and $\|\cdot\|$ denotes the standard weighted l^1-norm. Since $\overline{v}_L^n$ can be regarded as a perturbation of v_L^n, this is only possible if L is chosen so as to ensure that the reference scheme is sufficiently accurate, i.e. one also has $\|\hat{u}_L^n - v_L^n\| \leq tol$. Again ideally, a possibly low number of refinement levels L should be determined during the computation such that the error meets the desired tolerance $\|\hat{u}_L^n - \overline{v}_L^n\| \leq tol$. Since no explicit error estimator is available for the adaptive scheme, we try to assess the error by splitting the error into two parts corresponding to the *discretization error* $\tau_L^n := \hat{u}_L^n - v_L^n$ of the reference FVS and the *perturbation error* $e_L^n := v_L^n - \overline{v}_L^n$. We now assume that there is an a priori error estimate of the discretization error, i.e., $\tau_L^n \sim h_L^\alpha$ where h_L denotes the spatial step size and α the convergence order. Then, ideally we would determine the number of refinement levels L such that $h_L^\alpha \sim tol$. In order to preserve the accuracy of the reference FVS, we may now admit a perturbation error which is proportional to the discretization error, i.e., $\|e_L^n\| \sim \|\tau_L^n\|$. From this, one can derive a suitable level $L = L(tol, \alpha)$ and $\varepsilon = \varepsilon(L)$.

Therefore it remains to verify that the perturbation error can be controlled. To this end, note that in each time step we introduce an error due to the thresholding procedure. Obviously, this error accumulates in each step, i.e., the best we can hope for is an estimate of the form $\|e_L^n\| \leq C n \varepsilon$. However, the threshold error may be amplified in addition by the evolution step. In order to control the cumulative perturbation error, we have to prove that the constant C is independent of L, n, τ and ε. For a simplified model problem this was rigorously done in [25] for homogeneous problems and exact reconstruction and, recently, in [44] for inhomogeneous problems using approximate flux and source reconstruction.

6 Numerical results

Finally, we would like to demonstrate that the multiscale-based grid adaptation concept has been developed beyond pure academic investigations and can be applied

to real-world problems. For this purpose, we present the results of 3D simulations that have been recently performed with the new solver Quadflow for a challenging problem arising in aerodynamics.

6.1 The solver Quadflow

The above multiscale-based grid adaptation concept has been integrated into the new adaptive and parallel solver Quadflow [14, 15]. This solver has been developed for more than one decade within the collaborative research center SFB 401 *Modulation of Flow and Fluid-Structure Interaction at Airplane Wings*, cf. [3, 65]. In order to exploit synergy effects, it has been designed as an *integrated* tool where each of the core ingredients, namely, (i) the flow solver concept based on a finite volume discretization [13], (ii) the grid adaptation concept based on wavelet techniques [54], and (iii) the grid generator based on B-spline mappings [51] is adapted to the needs of the others. In particular, the three tools are not just treated as independent black boxes communicating via interfaces. Instead, they are highly intertwined on a conceptual level mainly linking (i) the multiresolution-based grid adaption that reliably detects and resolves all physical relevant effects, and (ii) the B-spline grid generator which reduces grid changes to just moving a few control points whose number is, in particular, independent of any local grid refinement. The mathematical concepts have been complemented recently by parallelization techniques that are indispensable for further reducing the computational time to an affordable order of magnitude when dealing with realistic 3D computations for complex geometries, cf. [18, 4].

6.2 Application

The efficiency of an airport is strongly influenced by the takeoff and landing frequency that is determined by the system of vortices generated at the wing tips. These vortices continue to exist for a long period of time in the wake of an airplane, see Figure 9. It is possible to detect wake vortices as far as 100 wing spans behind the airplane, which are a hazard to following airplanes. In the SFB 401, the research aimed to induce instabilities into the system of vortices to accelerate their collapse. The effects of different measures, e.g. additional flaps installed at each airfoil, taken in order to destabilize the vortices have been examined in a water tunnel. A model of a wing was mounted in a water tunnel and the velocity components in the area behind the wing were measured using particle image velocimetry It was possible to conduct measurements over a length of 4 wing spans. The experimental analysis of a system of vortices far behind the wing poses great difficulties due to the size of the measuring system. Numerical simulations are not subject to such severe constraints and therefore Quadflow is used to examine the behavior of vortices far behind the wing. To minimize the computational effort, the grid adaptation adjusts the refine-

ment of the grid with the goal to resolve all important flow phenomena, while using as few cells as possible.

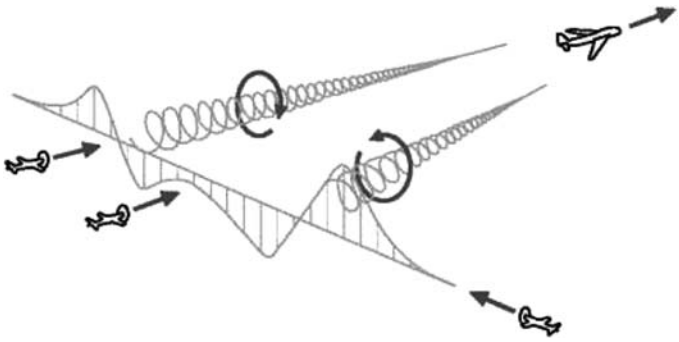


Fig. 9 System of wing tip vortices in the wake of an airplane. (Courtesy of Institute of Aeronautics and Astronautics, RWTH Aachen)

In the present study instationary, quasi incompressible, inviscid fluid flow described by the Euler equations is considered. An assessment is presented to validate the ability of Quadflow to simulate the behavior of the wake of an airplane. A velocity field based on the experimental measurements is prescribed as boundary condition in the inflow plane. Here the measured velocity fields at the wing tip and at the flap, respectively, have been used to generate two different Lamb-Oseen vortices. These vortices are used to specify the circumferential part of the velocity distribution $v_\Theta(r)$. The circumferential velocity distribution of one Lamb-Oseen vortex is computed by

$$v_\Theta(r) = \frac{\Gamma}{2\pi r}\left(1 - e^{-\left(\frac{r}{d_0}\right)^2}\right). \tag{30}$$

The radius r is the distance from the center of a boundary face in the inflow plane to the vortex core. The two parameters of the Lamb-Oseen vortices, circulation Γ and core radius d_0 are chosen in such a way that the models fit to the measured velocity field of the wing tip vortex and the flap vortex as close as possible, respectively. As observed in the experiment, both vortices are rotating in the same direction. The circumferential part of the velocity distribution at the inflow boundary is computed by the superposition of the velocity distribution of both vortices. The axial velocity component in the inflow direction is set to the constant inflow velocity of the water tunnel.

Instead of water, which is used as fluid in the experiment to visualize the vortices, the computation relies on air as fluid. This is justified because of the low Mach number $Ma = 0.05$ and, hence, compressibility effects are negligible. The inflow velocity in the x-direction, u_∞, is computed to fulfill the condition that the Reynolds number in the computational test case is the same as in the experiment.

The experimental conditions are a flow velocity $u_w = 1.1\,\mathrm{m\,s^{-1}}$ and a Reynolds number $Re_w = 1.9 \times 10^5$. From the condition $Re_{air} = Re_w$ the inflow velocity in the x-direction has been determined as $u_\infty = 16.21\,\mathrm{m\,s^{-1}}$. For purpose of consistency, the circumferential velocity v_θ has also been multiplied by the factor u_∞/u_w. The velocity of the initial solution is set to parallel, uniform flow $u_0 = u_\infty$, $v_0 = w_0 = 0.0$.

The computation[1] has been performed on 32 Intel Xeon E5450 processors running at 3 GHz clock speed. The CPU time spent was about 214 hours. The computational domain matches the experimental setup which extents $l = 6\,\mathrm{m}$ in the x-direction, $b = 1.5\,\mathrm{m}$ in the y-direction and $h = 1.1\,\mathrm{m}$ in the z-direction. The boundaries parallel to the x-direction have been modeled as symmetry walls. This domain is discretized by a coarse grid with 40 cells in flow direction, 14 cells in y-direction and 10 cells in the z-direction. The maximum number of refinement levels has been set to $L = 6$. With this setting, both vortices can be resolved on the finest level by about 80 points in the y-z-plane.

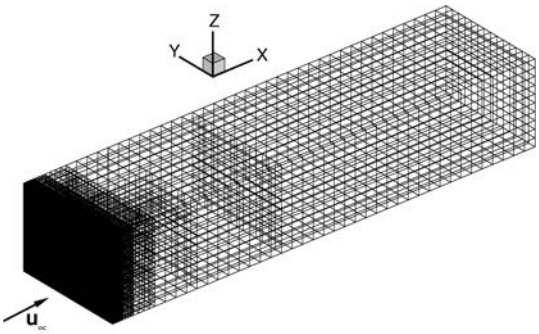


Fig. 10 Initial computational grid

Since Quadflow solves the compressible Euler equations, a preconditioner for low Mach numbers was applied in a dual-time framework acting only on the dual time-derivatives. It has been used for the purposes of numerical discretization and iterative solution, cf. [59]. The spatial discretization of the convective fluxes is based on the AUSMDV(P) flux vector splitting method [32]. For time integration the implicit midpoint rule is applied. In each time step the unsteady residual of the Newton iterations is reduced by about three orders of magnitude. The physical time step is uniformly set to $\Delta t = 5 \times 10^{-5}\,s$ which corresponds to a maximum CFL number of about $CFL_{max} = 28.0$ in the domain. Grid adaptation is performed after each time step. After every 100th time step the load balancing is repeated.

[1] The computations have been performed by Gero Schieffer. They have been made possible by the parallelization concept of space-filling curves embedded in the multiscale library by Silvia-Sorana Mogosan and Kolja Brix.

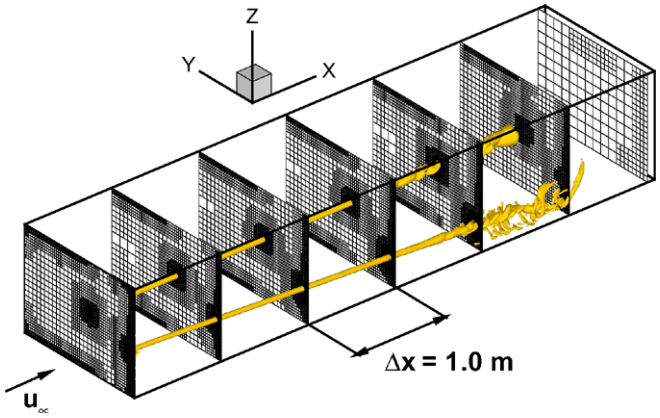


Fig. 11 Slices of the computational grid after 6046 time steps at seven different positions and the distribution of $\lambda_2 = -100$

When the computation starts, the vortices have to be resolved properly on a sufficiently refined grid. For this purpose, the grid on the inflow plane is pre-refined to the maximum level, see Figure 10. Due to this procedure the first grid contains 384000 cells. When the information at the inlet has crossed the first cell layer, the pre-adaptation of the cells at the inlet is no longer needed and then the grid is only adapted according to the adaptation criterion based on the multiscale analysis. For the multiscale analysis we use modified box wavelets with $M = 2$ vanishing moments, see Section 3. The threshold value is set to $\varepsilon = 2.5 \times 10^{-4}$. For the prediction step we apply Harten's original strategy summarized in Section 5.3.

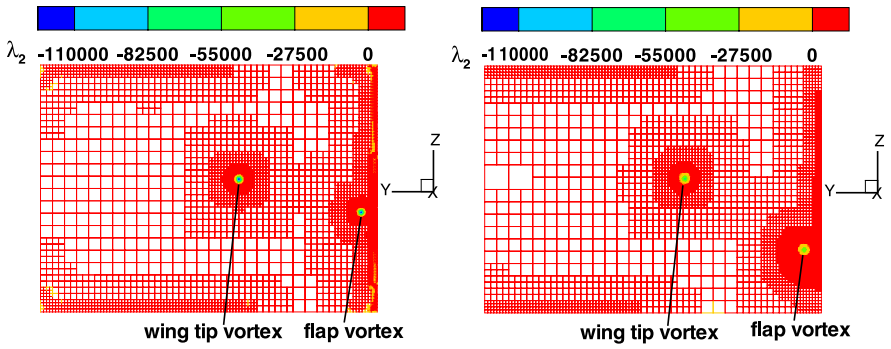


Fig. 12 Slices of the computational grid at two different positions in x-direction, the grid color is consistent with the value of λ_2. Left Figure: Slice of the computational grid at $x = 0.0\,\mathrm{m}$. Right Figure: Slice of the computational grid at $x = 3.0\,\mathrm{m}$

After 6046 time steps, which corresponds to a computed real time of $t = 0.3023\,\mathrm{s}$, the grid contains 3.04×10^6 cells in total. This is about 0.2 % of the uniformly refined reference mesh, i.e., by grid adaptation the computational complexity is reduced by a factor of about 500. Figure 11 shows seven cross sections of the mesh, which are equally spaced in x-direction with distances $\Delta x = 1.0\,\mathrm{m}$. In addition, the isosurface of the λ_2-criterion is presented with the value $\lambda_2 = -100$. The λ_2-criterion has been proposed by Jeong et al. [45] to detect vortices. A negative value of λ_2 identifies a vortex, whereas the smallest of these negative values marks the core of the vortex. As can be seen from Figure 11, the vortices are transported through the computational domain. The locally adapted grid exhibits high levels of refinement only in the vicinity of the vortices. A more detailed view of the grid for the cross sections at $x = 0.0\,\mathrm{m}$ and $x = 3.0\,\mathrm{m}$ is presented in Figure 12.

From the engineering point of view, the interaction of the two vortices is of special interest. The central question is whether the strong wing tip vortex can be destabilized by the flap vortex. For this purpose, the computation has to be continued. This is subject of current research. Nevertheless, the computations performed so far verify that the presented concepts are sustainable and necessary in order to investigate this challenging problem.

7 Conclusion and trends

Adaptive multiresolution schemes have turned out be very efficient in numerous applications. In particular, the adaptation process is only controlled by the threshold parameter. The choice of this parameter seems to be very robust with respect to varying configurations and applications. Ideally, it depends on the discretization error of the reference finite volume scheme. This was confirmed by rigorous mathematical estimates for scalar model problems.

Originally, the multiresolution-based grid adaptation technique was kept separate from the treatment of discrete evolution equations. However, the multiresolution analysis offers a much higher potential when applying it directly to the (discrete) evolution equations. Therefore we would like to conclude with some comments on the future development of adaptive multiresolution schemes that is beyond mere grid adaptation.

Trend 1: Adaptive mesh refinement and multiresolution analysis. In order to optimize computational resources, AMR techniques have become a standard way to optimize computational resources. These techniques have been originally developed in the 1980's by Berger et al. [10, 8, 7]. Typically, the refinement process is triggered by gradients [60] or higher order interpolation [2]. Recent investigations by [31] show that using a discrete multiresolution analysis instead leads to a much more efficient refinement criteria. In particular, in areas of partial smoothness such as rarefaction waves. It turned out that only minimal changes in the existing AMR code were necessary to embed the multiresolution-based refinement criterion.

In principle, it would be possible to embed the multiresolution-based grid adaptation concept to any AMR code as a black box, where the data have to be transferred between the two tools. This has been realized in the Quadflow solver [14, 15]. However, this requires some computational overhead in terms of memory and CPU time. In particular, the multiresolution-based grid adaptation technique is kept separate from the treatment of discrete evolution equations and therefore we could not employ the much higher potential of the multiresolution analysis when applying it directly to the discrete evolution.

Trend 2: Implicit time discretization. In Section 5, adaptive multiresolution finite volume schemes have been derived only for explicit time discretizations. We may proceed similarly in case of an implicit time discretization, cf. [15]. These are of interest when dealing with stationary flow problems, weakly instationary problems or models that exhibit some stiffness due to relaxation processes, e.g. chemical reactions, or dissipation, e.g. diffusion, viscosity and heat conduction, resulting in anisotropic flow structures such as boundary layers. For these types of problems an explicit time discretization would lead to very small time steps in order to meet the CFL condition. Although the derivation is straight-forward, several new questions arise:

(i) In each time step the implicit time discretization results in a nonlinear system of discrete evolution equations. Typically this system is solved by Newton-Krylov methods. For steady state problems, only one Newton step is performed, because the time plays only the role of a relaxation parameter and there is no need to be accurate in each time step. However, for instationary problems several Newton steps are needed to maintain the accuracy in each time step. In recent work by Steiner et al. [69, 68], it was possible to design a break condition for the Newton methods that relies on the threshold value of the multiscale method.

To improve the efficiency of the solution of the nonlinear system one might employ the multilevel structure of the underlying grid hierarchy in the multiscale analysis similar to adaptive multigrid techniques such as Brandt's so-called multilevel adaptive technique (MLAT), cf. [16, 17], that is an adaptive generalization of the full approximation scheme (FAS). The efficiency of these methods crucially relies on the proper choice of problem-dependent transfer and relaxation operators. First investigations in [56] and [57] for unsteady state and steady state flow problems, respectively, show that opposite to classical adaptive multigrid schemes we may employ the multiresolution analysis using biorthogonal wavelets to define the restriction and prolongation operators. Since the underlying problem is nonlinear, the FAS [16] is used for the coarse grid correction. Further investigations are needed to fully employ the high potential of the multiresolution analysis when applying it directly to the discrete evolution equations arising from the finite volume discretization rather than just using it as a data compression tool for the set of discrete cell data.

(ii) By the implicit time discretization, the data in all cells are coupled and, hence, an information could propagate throughout the entire computational domain in one time step. Since the prediction strategy in Section 5.3 relies on the fact that the information propagates at most by one cell, the prediction has to be adjusted. Typ-

ically, for convection-dominated problems such as compressible fluid flow at high Reynolds numbers the influence of a local perturbation decays rapidly in space and stays more or less local. In [11], a heuristic approach has been developed for viscous problems where the parameter q in Harten's strategy has been coupled with the viscosity parameter. However, a rigorous mathematical justification of its reliability in the sense of the condition (16) is still missing.

Trend 3: Time step adaptation. The crux of adaptive multiresolution schemes is the multiresolution analysis of data corresponding to an arbitrary but fixed time. Therefore the local time variation is not directly accessible from the analysis of the spatial variation. In recent years, there have been several attempts to develop time adaptive scheme where the time step is controlled. This is not to be confused with multilevel time stepping as presented in Section 5.4.

A possible strategy has been investigated by Ferm and Lotstedt [36] based on time step control strategies for ODEs. Here a Runge-Kutta-Fehlberg method is applied to the semi-discretized flow equations by which the local spatial and temporal errors are estimated. These errors determine the local stepsize in time and space. Later on, this idea was also embedded in fully adaptive multiresolution finite volume schemes, cf. [30]. Alternatively, Kröner and Ohlberger [48] based their space-time adaptivity upon Kuznetsov-type a-posteriori L_1-error-estimates for scalar conservation laws.

More recently, explicit and implicit finite volume solvers on adaptively refined meshes have been coupled with adjoint techniques to control the time stepsizes for the solution of weakly instationary compressible inviscid flow problems like transonic flight. These can be considered perturbations of stationary flows. While time accuracy is still needed to study phenomena like aero-elastic interactions, large time steps may be possible when the perturbations have passed. Here the time step control is based on a space-time-splitting of the adjoint error representation, cf. [33, 5, 6]. In [68, 69] the multiscale-based grid adaptation was combined with these adjoint techniques to solve efficiently instationary problems. The advantage of this space adaptive method is that it also provides an efficient break condition for the Newton iteration in the implicit time integration.

Trend 4: Parallelization. Although multiscale-based grid adaptation leads to a significant reduction of the computational complexity (CPU time and memory) in comparison to computations on uniform meshes, this is not sufficient to perform 3D computations for complex geometries efficiently. In addition, we need parallelization techniques in order to further reduce the computational time to an affordable order of magnitude. On a distributed memory architecture, the performance of a parallelized code crucially depends on the load-balancing and the interprocessor communication. Since the underlying adaptive grids are unstructured due to hanging nodes, this task cannot be considered trivial. For this purpose, graph partitioning methods are frequently employed using the Metis software [47, 46]. An alternative approach is based on space-filling curves, cf. [71]. Here the basic idea is to map level-dependent multiindices identifying the cells in a dyadic grid hierarchy of nested grids to a onedimensional line. The interval is then split into different parts each containing the same number of entries. In the context of adaptive multiresolu-

tion schemes both the graph-partitioning and the space-filling curve approach have been used, cf. [61, 62] and [18, 4], respectively.

Nowadays more and more powerful parallel hardware architectures based on clusters of shared memory machines are being developed. Therefore the above concepts have to be reconsidered. In order to fully employ the power of the machines, a redesign of algorithms and data structures seems to be indispensable taking into account issues such as caching and threading.

Trend 5: Turbulence Modeling. The potential of the multiresolution analysis is not only restricted to pure data analysis but can be used, for instance, to model turbulent flow. The inherent problem of simulating turbulent flows comes from the number of degrees of freedom needed to resolve turbulent structures. This number is proportional to $Re^{9/4}$ and becomes dramatically large with increasing Reynolds number Re, e.g. in aerodynamics $Re \sim 10^6$, that makes a direct numerical simulation (DNS) impossible in many applications. In general, the interest is not in the fully resolved turbulent flow field but in some macroscopic quantities such as lift and drag coefficients. At the macroscale the quantities can be resolved. However, they are influenced by the non-resolved fluctuations. Typically, the influence of the fluctuations is described using some algebraic models, the Reynolds-averaged Navier-Stokes equations (RANS) or large eddy simulations (LES). Alternatively, the coherent vortex simulation (CVS) developed by Farge et al. [35, 64, 34] for incompressible flows has been designed to compute this problem with a reduced number of degrees of freedom. This methodology is based on the wavelet representation of the vorticity. The basic idea is to extract the coherent vortex structures from the noise which will then be modeled to compute the flow evolution.

Up to now, it is not apriorily known whether the choice of degrees of freedom corresponding to the resolved macroscale is sufficient to capture adequately the influence of the small scales on the macroscale. Using multiresolution techniques in combination with recent quantitative estimates for the action of the nonlinearity on different scales of the flow field, cf. [22, 23], seem to offer a promising possibility to investigate more rigorously the effect of the fluctuations on the coarse scales. In particular, it will be interesting to adjust the local scale of resolution adaptively at run time instead of fixing it before starting the computation. Work in this regard is done in [29].

Acknowledgements The author would like to express his deepest gratitude to Wolfgang Dahmen who has been supporting and inspiring the author's scientific work for many years.

References

1. S. Andreae, J. Ballmann, and S. Müller. Wave processes at interfaces. In G. Warnecke, editor, *Analysis and numerics for conservation laws*, pages 1–25. Springer, Berlin, 2005.
2. A. Baeza and P. Mulet. Adaptive mesh refinement techniques for high-order shock capturing schemes for multi-dimensional hydrodynamic simulations. *Int. Journal for Numerical Methods in Fluids*, 52(4):455–471, 2006.

3. J. Ballmann. *Flow Modulation and Fluid-Structure-Interaction at Airplane Wings*, volume 84 of *Numerical Notes on Fluid Mechanics*. Springer Verlag, 2003.
4. J. Ballmann, K. Brix, W. Dahmen, Ch. Hohn, S. Mogosan, S. Müller, and G. Schieffer. Parallel and adaptive methods for fluid-structure-interactions. Numerical Notes on Fluid Mechanics, 2009. Submitted.
5. R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Num. Math.*, 4:237–264, 1996.
6. R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001.
7. J. Bell, M.J. Berger, J. Saltzman, and M. Welcome. Three-dimensional adaptive mesh refinement for hyperbolic conservation laws. *SIAM J. Sci. Comput.*, 15(1):127–138, 1994.
8. M.J. Berger and P. Colella. Local adaptive mesh refinement for shock hydrodynamics. *J. Comp. Physics*, 82:64–84, 1989.
9. M.J. Berger and R.J. LeVeque. Adaptive mesh refinement using wave-propagation algorithms for hyperbolic systems. *SIAM J. Numer. Anal.*, 35(6):2298–2316, 1998.
10. M.J. Berger and J. Oliger. Adaptive mesh refinement for hyperbolic partial differential equations. *J. Comp. Physics*, 53:484–512, 1984.
11. B. Bihari. Multiresolution schemes for conservation laws with viscosity. *J. Comp. Phys.*, 123(1):207–225, 1996.
12. B. Bihari and A. Harten. Multiresolution schemes for the numerical solution of 2–D conservation laws I. *SIAM J. Sci. Comput.*, 18(2):315–354, 1997.
13. F. Bramkamp. *Unstructured h-Adaptive Finite-Volume Schemes for Compressible Viscous Fluid Flow*. PhD thesis, RWTH Aachen, 2003. http://darwin.bth.rwth-aachen.de/opus3/volltexte/2003/725/03_255.pdf.
14. F. Bramkamp, B. Gottschlich-Müller, M. Hesse, Ph. Lamby, S. Müller, J. Ballmann, K.-H. Brakhage, and W. Dahmen. *H*-adaptive Multiscale Schemes for the Compressible Navier–Stokes Equations — Polyhedral Discretization, Data Compression and Mesh Generation. In J. Ballmann, editor, *Flow Modulation and Fluid-Structure-Interaction at Airplane Wings*, volume 84 of *Numerical Notes on Fluid Mechanics*, pages 125–204. Springer, 2003.
15. F. Bramkamp, Ph. Lamby, and S. Müller. An adaptive multiscale finite volume solver for unsteady an steady state flow computations. *J. Comp. Phys.*, 197(2):460–490, 2004.
16. A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31:333–390, 1977.
17. A. Brandt. Multi-level adaptive techniques (mlat) for partial differential equations: Ideas and software. In *Mathematical software III, Proc. Symp., Madison 1977*, pages 277–318, 1977.
18. K. Brix, S. Mogosan, S. Müller, and G. Schieffer. Parallelization of multiscale-based grid adaptation using space-filling curves. IGPM–Report 299, RWTH Aachen, 2009.
19. R. Bürger, R. Ruiz, and K. Schneider. Fully adaptive multiresolution schemes for strongly degenerate parabolic equations with discontinuous flux. *J. Eng. Math.*, 60(3-4):365–385, 2008.
20. R. Bürger, R. Ruiz, K. Schneider, and M.A. Sepulveda. Fully adaptive multiresolution schemes for strongly degenerate parabolic equations in one space dimension. *ESAIM, Math. Model. Numer. Anal.*, 42(4):535–563, 2008.
21. J.M. Carnicer, W. Dahmen, and J.M. Peña. Local decomposition of refinable spaces and wavelets. *Appl. Comput. Harmon. Anal.*, 3:127–153, 1996.
22. A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet schemes for nonlinear variational schemes. *Numer. Anal.*, 41(5):1785–1823, 2003.
23. A. Cohen, W. Dahmen, and R. DeVore. Sparse evaluation of compositions of functions using multiscale expansions. *SIAM J. Math. Anal.*, 35(2):279–303, 2003.
24. A. Cohen, I. Daubechies, and J. Feauveau. Bi–orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.
25. A. Cohen, S.M. Kaber, S. Müller, and M. Postel. Fully Adaptive Multiresolution Finite Volume Schemes for Conservation Laws. *Math. Comp.*, 72(241):183–225, 2003.
26. A. Cohen, S.M. Kaber, and M. Postel. Multiresolution Analysis on Triangles: Application to Gas Dynamics. In G. Warnecke and H. Freistühler, editors, *Hyperbolic Problems: Theory, Numerics, Applications*, pages 257–266. Birkhäuser, 2002.

27. F. Coquel, Q.L. Nguyen, M. Postel, and Q.H. Tran. Local time stepping applied to implicit-explicit methods for hyperbolic systems. *SIAM Multiscale Modeling and Simulation*, 2009. Accepted for publication.
28. F. Coquel, M. Postel, N. Poussineau, and Q.H. Tran. Multiresolution technique and explicit-implicit scheme for multicomponent flows. *J. Numer. Math.*, 14:187–216, 2006.
29. W. Dahmen and S. Müller. Multiscale techniques for high-resolved vortex strcutures, 2008. DFG project *Multiresolution and Adaptive Methods for Convection-Dominated Problems*, http://www.sfbtr40.de/index.php?option=com_content&view=category&layout=blog&id=34&Itemid=58&lang=en
30. M. Domingues, O. Roussel, and K. Schneider. On space-time adaptive schemes for the numerical solution of partial differential equations. *ESAIM: Proceedings*, 16:181–194, 2007.
31. R. Donat. Using Harten's multiresolution framework on existing high resolution shock capturing schemes, 2009. Presentation at Workshop on *Multiresolution and Adaptive Methods for Convection-Dominated Problems*, http://www.ann.jussieu.fr/mamcdp09/slides/RosaDonatMAMCDP09.pdf.
32. J. Edwards and M.S. Liou. Low-diffusion flux-splitting methods for flows at all speeds. *AIAA Journal*, 36:1610–1617, 1998.
33. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. IV. Nonlinear problems. *SIAM J. Numer. Anal.*, 32:1729–1749, 1995.
34. M. Farge and K. Schneider. Coherent vortex simulation (cvs), a semi-deterministic turbulence model using wavelets. *Turbulence and Combustion*, 66(4):393–426, 2001.
35. M. Farge, K. Schneider, and N. Kevlahan. Non-gaussianity and coherent vortex simulation for two-dimensional turbulence using an orthogonal wavelet basis. *Phys. Fluids*, 11(8):2187–2201, 1999.
36. L. Ferm and P. Lötstedt. Space-time adaptive solution of first order PDEs. *J. Sci. Comput.*, 26(1):83–110, 2006.
37. B. Gottschlich–Müller. *Multiscale Schemes for Conservation Laws*. PhD thesis, RWTH Aachen, 1998.
38. A. Harten. Discrete multi–resolution analysis and generalized wavelets. *J. Appl. Num. Math.*, 12:153–193, 1993.
39. A. Harten. Adaptive multiresolution schemes for shock computations. *J. Comp. Phys.*, 115:319–338, 1994.
40. A. Harten. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Comm. Pure Appl. Math.*, 48(12):1305–1342, 1995.
41. A. Harten. Multiresolution representation of data: A general framework. *SIAM J. Numer. Anal.*, 33(3):1205–1256, 1996.
42. R. Hartmann and R. Rannacher. Adaptive FE-methods for conservation laws. In G. Warnecke and H. Freistühler, editors, *Hyperbolic Problems: Theory, Numerics, Applications*, pages 495–504. Birkhäuser, 2002.
43. P. Houston, J.A. Mackenzie, E. Süli, and G. Warnecke. A posteriori error analysis for numerical approximations of Friedrichs systems. *Numer. Math.*, 82:433–470, 1999.
44. N. Hovhannisyan and S. Müller. On the stability of fully adaptive multiscale schemes for conservation laws using approximate flux and source reconstruction strategies. IGPM–Report 284, RWTH Aachen, 2008. Accepted for publication in IMA Journal of Numerical Analysis.
45. J. Jeong and F. Hussain. On the identification of a vortex. *Journal of Fluid Mechanics*, 285:69–94, 1995.
46. G. Karypis and V. Kumar. Multilevel algorithms for multi-constraint graph partitioning. *Supercomputing*, 1998.
47. G. Karypis and V. Kumar. A parallel algorithm for multilevel graph partitioning and sparse matrix ordering. *Journal of Parallel and Distributed Computing*, 48:71–85, 1998.
48. D. Kröner and M. Ohlberger. A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions. *Math. Comp.*, 69(229):25–39, 2000.
49. S.N. Kruzhkov. First order quasilinear equations with several space variables. *Math. USSR Sb.*, 10:217–243, 1970.

50. N.N. Kuznetsov. The weak solution of the Cauchy problem for a multi–dimensional quasilinear equation. *Mat. Zametki*, 2:401–410, 1967. In Russian.
51. Ph. Lamby. *Parametric Multi-Block Grid Generation and Application To Adaptive Flow Simulations*. PhD thesis, RWTH Aachen, 2007. http://darwin.bth.rwth-aachen.de/opus3/volltexte/2007/1999/pdf/Lamby_Philipp.pdf.
52. Ph. Lamby, R. Massjung, S. Müller, and Y. Stiriba. Inviscid flow on moving grids with multiscale space and time adaptivity. In *Numerical Mathematics and Advanced Applications: Proceedings of Enumath 2005 the 6th European Conference on Numerical Mathematics and Advanced Mathematics*, pages 755–764. Springer, 2006.
53. Ph. Lamby, S. Müller, and Y. Stiriba. Solution of shallow water equations using fully adaptive multiscale schemes. *Int. Journal for Numerical Methods in Fluids*, 49(4):417–437, 2005.
54. S. Müller. *Adaptive Multiscale Schemes for Conservation Laws*, volume 27 of *Lecture Notes on Computational Science and Engineering*. Springer, 2002.
55. S. Müller, Ph. Helluy, and J. Ballmann. Numerical simulation of a single bubble by compressible two-phase fluids. *Int. Journal for Numerical Methods in Fluids*, 2009. DOI 10.1002/fld.2033.
56. S. Müller and Y. Stiriba. Fully adaptive multiscale schemes for conservation laws employing locally varying time stepping. *Journal for Scientific Computing*, 30(3):493–531, 2007.
57. S. Müller and Y. Stiriba. A multilevel finite volume method with multiscale-based grid adaptation for steady compressible flow. *Journal of Computational and Applied Mathematics, Special Issue: Emergent Applications of Fractals and Wavelets in Biology and Biomedicine*, 227(2):223–233, 2009. DOI 10.1016/j.cam.2008.03.035.
58. S. Osher and R. Sanders. Numerical approximations to nonlinear conservation laws with locally varying time and space grids. *Math. Comp.*, 41:321–336, 1983.
59. S.A. Pandya, S. Venkateswaran, and T.H. Pulliam. Implementation of preconditioned dual-time procedures in overflow. *AIAA Paper 2003-0072*, 2003.
60. J.J. Quirk. *An adaptive grid algorithm for computational shock hydrodynamics*. PhD thesis, Cranfield Institute of Technology, 1991.
61. O. Roussel and K. Schneider. Adaptive multiresolution method for combustion problems: Application to flame ball-vortex interaction. *Computers and Fluids*, 34(7):817–831, 2005.
62. O. Roussel and K. Schneider. Numerical studies of spherical flame structures interacting with adiabatic walls using an adaptive multiresolution scheme. *Combust. Theory Modelling*, 10(2):273–288, 2006.
63. O. Roussel, K. Schneider, A. Tsigulin, and H. Bockhorn. A conservative fully adaptive multiresolution algorithm for parabolic PDEs. *J. Comp. Phys.*, 188(2):493–523, 2003.
64. K. Schneider and M. Farge. Numerical simulation of a mixing layer in an adaptive wavelet basis. *C.R. Acad. Sci. Paris Série II b*, 328:263–269, 2001.
65. W. Schröder. *Flow Modulation and Fluid-Structure-Interaction at Airplane Wings II*. Numerical Notes on Fluid Mechanics. Springer Verlag, 2009. In preparation.
66. T. Sonar, V. Hannemann, and D. Hempel. Dynamic adaptivity and residual control in unsteady compressible flow computation. *Math. and Comp. Modelling*, 20:201–213, 1994.
67. T. Sonar and E. Süli. A dual graph–norm refinement indicator for finite volume approximations of the Euler equations. *Numer. Math.*, 78:619–658, 1998.
68. Ch. Steiner. *Adaptive timestepping for conservation laws via adjoint error representation*. PhD thesis, RWTH Aachen, 2008. http://darwin.bth.rwth-aachen.de/opus3/volltexte/2009/2679/.
69. Ch. Steiner, S. Müller, and S. Noelle. Adaptive timestep control for weakly instationary solutions of the Euler equations. IGPM–Report 292, RWTH Aachen, 2009.
70. Ch. Steiner and S. Noelle. On adaptive timestepping for weakly instationary solutions of hyperbolic conservation laws via adjoint error control. *Communications in Numerical Methods in Engineering*, 2008. Accepted for publication.
71. G. Zumbusch. *Parallel multilevel methods. Adaptive mesh refinement and loadbalancing*. Advances in Numerical Mathematics. Teubner, Wiesbaden, 2003.

Theory of adaptive finite element methods: An introduction

Ricardo H. Nochetto, Kunibert G. Siebert and Andreas Veeser

Abstract This is a survey on the theory of adaptive finite element methods (AFEM), which are fundamental in modern computational science and engineering. We present a self-contained and up-to-date discussion of AFEM for linear second order elliptic partial differential equations (PDEs) and dimension $d > 1$, with emphasis on the differences and advantages of AFEM over standard FEM. The material is organized in chapters with problems that extend and complement the theory. We start with the functional framework, inf-sup theory, and Petrov-Galerkin method, which are the basis of FEM. We next address four topics of essence in the theory of AFEM that cannot be found in one single article: mesh refinement by bisection, piecewise polynomial approximation in graded meshes, a posteriori error analysis, and convergence and optimal decay rates of AFEM. The first topic is of geometric and combinatorial nature, and describes bisection as a rather simple and efficient technique to create conforming graded meshes with optimal complexity. The second topic explores the potentials of FEM to compensate singular behavior with local resolution and so reach optimal error decay. This theory, although insightful, is insufficient to deal with PDEs since it relies on knowing the exact solution. The third topic provides the missing link, namely a posteriori error estimators, which hinge exclusively on accessible data: we restrict ourselves to the simplest residual-type estimators and present a complete discussion of upper and lower bounds, along with the concept of oscillation and its critical role. The fourth topic refers to the convergence of adaptive loops and its comparison with quasi-uniform refinement. We first show, under rather modest assumptions on the problem class and AFEM, convergence in the natural norm associated to the variational formulation. We next restrict the problem class to coercive symmetric bilinear forms, and show that AFEM is a contraction for a suitable error notion involving the induced energy norm. This property is then instrumental to prove optimal cardinality of AFEM for a class of singular functions, for which the standard FEM is suboptimal.

Ricardo H. Nochetto
Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742, USA, e-mail: rhn@math.umd.edu. Partially supported by NSF grant DMS-0807811.

Kunibert G. Siebert
Fakultät für Mathematik, Universität Duisburg-Essen, Forsthausweg 2, D-47057 Duisburg, Germany, e-mail: kg.siebert@uni-due.de

Andreas Veeser
Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, I-20133 Milano, Italy, e-mail: andreas.veeser@unimi.it

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_12,

1 Introduction

Adaptive finite element methods are a fundamental numerical instrument in science and engineering to approximate partial differential equations. In the 1980s and 1990s a great deal of effort was devoted to the design of a posteriori error estimators, following the pioneering work of Babuška. These are computable quantities, depending on the discrete solution(s) and data, that can be used to assess the approximation quality and improve it adaptively. Despite their practical success, adaptive processes have been shown to converge, and to exhibit optimal complexity, only recently and for linear elliptic PDE.

This survey presents an up-to-date discussion of adaptive finite element methods encompassing its design and basic properties, convergence, and optimality.

1.1 Classical vs adaptive approximation in 1d

We start with a simple motivation in 1d for the use of adaptive procedures, due to DeVore [28]. Given $\Omega = (0,1)$, a partition $\mathscr{T}_N = \{x_i\}_{n=0}^N$ of Ω

$$0 = x_0 < x_1 < \cdots < x_n < \cdots < x_N = 1$$

and a continuous function $u : \Omega \to \mathbb{R}$, we consider the problem of *interpolating* u by a *piecewise constant* function U_N over $\mathscr{T}_N$. To quantify the difference between u and U_N we resort to the *maximum norm* and study two cases depending on the regularity of u.

Case 1: W^1_∞-Regularity. Suppose that u is Lipschitz in $[0,1]$. We consider the approximation

$$U_N(x) := u(x_{n-1}) \qquad \text{for all } x_{n-1} \le x < x_n.$$

Since

$$|u(x) - U_N(x)| = |u(x) - u(x_{n-1})| = \left| \int_{x_{n-1}}^{x} u'(t)dt \right| \le h_n \|u'\|_{L^\infty(x_{n-1},x_n)}$$

we conclude that

$$\|u - U_N\|_{L^\infty(\Omega)} \le \frac{1}{N} \|u'\|_{L^\infty(\Omega)}, \tag{1}$$

provided the local mesh-size h_n is about constant (*quasi-uniform* mesh), and so proportional to N^{-1} (the reciprocal of the number of degrees of freedom). Note that the same integrability is used on both sides of (1). A natural question arises: *Is it possible to achieve the same asymptotic decay rate N^{-1} with weaker regularity demands?*

Case 2: W^1_1-Regularity. To answer this question, we suppose $\|u'\|_{L^1(\Omega)} = 1$ and consider the non-decreasing function

$$\phi(x) := \int_0^x |u'(t)|dt$$

which satisfies $\phi(0) = 0$ and $\phi(1) = 1$. Let $\mathscr{T}_N = \{x_i\}_{n=0}^N$ be the partition given by

$$\int_{x_{n-1}}^{x_n} |u'(t)|dt = \phi(x_n) - \phi(x_{n-1}) = \frac{1}{N}.$$

Then, for $x \in [x_{n-1}, x_n]$,

$$|u(x) - u(x_{n-1})| = \Big| \int_{x_{n-1}}^x u'(t)dt \Big| \leq \int_{x_{n-1}}^x |u'(t)|dt \leq \int_{x_{n-1}}^{x_n} |u'(t)|dt = \frac{1}{N},$$

whence

$$\|u - U_N\|_{L^\infty(\Omega)} \leq \frac{1}{N} \|u'\|_{L^1(\Omega)}. \tag{2}$$

We thus conclude that we could achieve the same rate of convergence N^{-1} for rougher functions with just $\|u'\|_{L^1(\Omega)} < \infty$. The following comments are in order for Case 2.

Remark 1.1 (Equidistribution). The optimal mesh $\mathscr{T}_N$ *equidistributes* the max-error. This mesh is graded instead of uniform but, in contrast to a uniform mesh, such a partition may not be adequate for another function with the same basic regularity as u. It is instructive to consider the singular function $u(x) = x^\gamma$ with $\gamma = 0.1$ and error tolerance 10^{-2} to quantify the above computations: if N_1 and N_2 are the number of degrees of freedom with uniform and graded partitions, we obtain $N_1/N_2 = 10^{18}$.

Remark 1.2 (Nonlinear Approximation). The regularity of u in (2) is measured in $W_1^1(\Omega)$ instead of $W_\infty^1(\Omega)$ and, consequently, the fractional γ regularity measured in $L^\infty(\Omega)$ increases to one full derivative when expressed in $L^1(\Omega)$. This exchange of integrability between left and right-hand side of (2), and gain of differentiability, is at the heart of the matter and the very reason why suitably graded meshes achieve optimal asymptotic error decay for singular functions. By those we mean functions which are not in the usual linear Sobolev scale, say $W_\infty^1(\Omega)$ in this example, but rather in a nonlinear scale [28]. We will get back to this issue in Chap. 5.

1.2 Outline

The function U_N may be the result of a minimization process. If we wish to minimize the norm $\|u - v\|_{L^2(\Omega)}$ within the space $\mathbb{V}_N$ of piecewise constant functions over $\mathscr{T}_N$, then it is easy to see that the solution U_N satisfies the orthogonality relation

$$U_N \in \mathbb{V}_N : \qquad \langle u - U_N, v \rangle = 0 \qquad \text{for all } v \in \mathbb{V}_N \tag{3}$$

and is given by the explicit local expression

$$U_N(x) = \frac{1}{h_n} \int_{x_{n-1}}^{x_n} u \qquad \text{for all } x_{n-1} < x < x_n.$$

The previous comments apply to this U_N as well even though U_N coincides with u at an unknown point in each interval $[x_{n-1}, x_n]$.

The latter example is closer than the former to the type of approximation issues discussed in this survey. A brief summary along with an outline of this survey follows:

PDE: The function u is not directly accessible but rather it is the solution of an elliptic PDE. Its approximation properties are intimately related to its regularity. In Chap. 2 we review briefly Sobolev spaces and the variational formulation of elliptic PDE, a present a full discussion of the inf-sup theory. We show the connection between approximability and regularity in Chap. 5, when we assess constructive approximation and use this later in Chap. 9 to derive rates of convergence.

FEM: To approximate u we need a numerical method which is sufficiently flexible to handle both geometry and accuracy (local mesh refinement); the method of choice for elliptic PDEs is the finite element method. We present its basic theory in Chap. 3, with emphasis on piecewise linear elements. We discuss the refinement of simplicial meshes in any dimension by bisection in Chap. 4, and address its complexity. This allows us to shed light on the geometric aspects of FEM that make them so flexible and useful in practice. The complexity analysis of bisection turns out to be crucial to construct optimal approximations in graded meshes in Chap. 5 and to derive convergence rates in Chap. 9 for AFEM.

Approximation: We briefly recall polynomial interpolation theory in Chap. 5 as well as the principle of error equidistribution. The latter is a concept that leads to optimal graded meshes and suggests that FEM might be able to approximate singular functions with optimal rate. We conclude Chap. 5 with the construction of optimal meshes via bisection for functions in a certain regularity class relevant to elliptic PDE. We emphasize the energy norm.

A Posteriori Error Estimation: To extract the local errors incurred by FEM, and thus be able to equidistribute them, we present residual-type a posteriori error estimators in Chap. 6. These are computable quantities in terms of the discrete solution and data which encode the correct information about the error distribution. They are the simplest but not the most accurate ones. Therefore, we also present alternative estimators, which are equivalent to the residual estimators. The discussion of Chap. 6 includes the appearence of an oscillation term and a proof that it cannot be avoided for the estimator to be practical. We show both upper and lower bounds between the energy error and the residual estimator. The former is essential for convergence and the latter for optimality.

Adaptivity: This refers to the use and study of loops to the form

$$\mathsf{SOLVE} \longrightarrow \mathsf{ESTIMATE} \longrightarrow \mathsf{MARK} \longrightarrow \mathsf{REFINE} \tag{4}$$

to iteratively improve the approximation of the solution of a PDE while keeping an optimal distribution of computational resources (degrees of freedom). The design of each module, along with some key properties, is discussed in Chap. 7 and 8. We emphasize the standard AFEM employed in practice which employs the estimator exclusively to make refinement decisions and never uses coarsening.

Convergence: This issue has been largely open until recently. In Chap. 7 we present a basic convergence theory for most linear elliptic PDEs, including saddle point problems, under rather modest assumptions and valid for all existing marking strategies. The final result is rather general but does not, and cannot, provide a convergence rate.

Optimality: We restrict ourselves to a model problem, which is symmetric and coercive, to investigate the convergence rate of AFEM. In Chap. 8 we derive a contraction property of AFEM for the so-called quasi-error, which is a scaled sum of the energy error and the estimator. In Chap. 9 we prove that AFEM converges with optimal rate as dictated by approximation theory even though the adaptive loop (4) does not use any regularity information but just the estimator. This analysis leads to approximation classes adequate for FEM, and so to the geometric restrictions caused by conforming grids, which are not the usual ones in nonlinear approximation theory.

2 Linear boundary value problems

In this section we examine the variational formulation of elliptic partial differential equations (PDE). We start with a brief review of Sobolev spaces and their properties and continue with several boundary value problems with main emphasis on a model problem that plays a relevant role in the subsequent analysis. Then we present the so-called inf-sup theory that characterizes existence and uniqueness of variational problems, and conclude by reviewing the applications in light of the inf-sup theory.

2.1 Sobolev spaces

The variational formulation of elliptic PDEs is based on Sobolev spaces. Moreover, approximability and regularity of functions are intimately related concepts. Therefore we briefly review definitions, basic concepts and properties of L^p-based Sobolev spaces for $1 \leq p \leq \infty$ and dimension $d \geq 1$. For convenience we restrict ourselves to bounded domains $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary.

Definition 2.1 (Sobolev Space). Given $k \in \mathbb{N}$ and $1 \leq p \leq \infty$, we define

$$W_p^k(\Omega) := \{v\colon \Omega \to \mathbb{R} \mid D^\alpha v \in L^p(\Omega) \text{ for all } |\alpha| \leq k\}$$

where $D^\alpha v = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d} v$ stands for the weak derivative of order α. The corresponding norm and seminorm are for $1 \le p < \infty$

$$\|v\|_{W_p^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha v\|_{L^p(\Omega)}^p \right)^{1/p}, \quad |v|_{W_p^k(\Omega)} := \left(\sum_{|\alpha| = k} \|D^\alpha v\|_{L^p(\Omega)}^p \right)^{1/p},$$

and for $p = \infty$

$$\|v\|_{W_p^\infty(\Omega)} := \sup_{|\alpha| \le k} \|D^\alpha v\|_{L^\infty(\Omega)}, \quad |v|_{W_p^\infty(\Omega)} := \sup_{|\alpha| = k} \|D^\alpha v\|_{L^\infty(\Omega)}.$$

For $p = 2$ the spaces $W_2^k(\Omega)$ are Hilbert spaces and we denote them by $H^k(\Omega) = W_2^k(\Omega)$. The scalar product inducing the norm $\|\cdot\|_{H^k(\Omega)} = \|\cdot\|_{W_2^k(\Omega)}$ is given by

$$\langle u, v \rangle_{H^k(\Omega)} = \sum_{|\alpha| \le k} \int_\Omega D^\alpha u D^\alpha v \qquad \text{for all } u, v \in H^k(\Omega).$$

We let $H_0^k(\Omega)$ be the completion of $C_0^\infty(\Omega)$ within $H^k(\Omega)$. The space $H_0^k(\Omega)$ is a strict subspace $H^k(\Omega)$ because $1 \in H^k(\Omega) \setminus H_0^k(\Omega)$.

There is a natural scaling of the seminorm in $W_p^k(\Omega)$. Consider for $h > 0$ the change of variables $\widehat{x} = x/h$ for all $x \in \Omega$, which transforms the domain Ω into $\widehat{\Omega}$ and functions v defined over Ω into functions $\widehat{v}$ defined over $\widehat{\Omega}$. Then

$$|\widehat{v}|_{W_p^k(\widehat{\Omega})} = h^{k-d/p} |v|_{W_p^k(\Omega)}.$$

This motivates the following definition, which turns out to be instrumental.

Definition 2.2 (Sobolev Number). The Sobolev number of $W_p^k(\Omega)$ is defined by

$$\mathrm{sob}(W_p^k) := k - d/p. \tag{5}$$

2.1.1 Properties of Sobolev Spaces

We summarize now, but not prove, several important properties of Sobolev spaces which play a key role later. We refer to [35, 38, 39] for details.

Embedding Theorem. Let $m > k \ge 0$ and assume $\mathrm{sob}(W_p^m) > \mathrm{sob}(W_q^k)$. Then the embedding

$$W_p^m(\Omega) \hookrightarrow W_q^k(\Omega)$$

is compact.

The assumption on the Sobolev number cannot be relaxed. To see this, consider Ω to be the unit ball of $\mathbb{R}^d$ for $d \ge 2$ and set $v(x) = \log\log \frac{|x|}{2}$ for $x \in \Omega \setminus \{0\}$. Then there holds $v \in W_d^1(\Omega)$ and $v \notin L^\infty(\Omega)$, but

$$\operatorname{sob}(W_d^1) = 1 - d/d = 0 = 0 - d/\infty = \operatorname{sob}(L^\infty).$$

Therefore, equality cannot be expected in the embedding theorem.

Density. The space $C^\infty(\overline{\Omega})$ is dense in $W_p^k(\Omega)$, i. e.,

$$W_p^k(\Omega) = \overline{C^\infty(\overline{\Omega})}^{\|\cdot\|_{\mathbb{V}}}.$$

Poincaré Inequality. The following inequality holds

$$\Big\| v - |\Omega|^{-1} \int_\Omega v \Big\|_{L^2(\Omega)} \le C(\Omega) \|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in W_2^1(\Omega) \tag{6}$$

with a constant $C(\Omega)$ depending on the shape of Ω. The best constant within the class of convex domains is

$$C(\Omega) = \frac{1}{\pi} \operatorname{diam}(\Omega);$$

see [60, 11].

Poincaré-Friedrichs Inequality. There is a constant $C_d > 0$ depending only on the dimension such that [38, p. 158]

$$\|v\|_{L^2(\Omega)} \le C_d |\Omega|^{1/d} \|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega). \tag{7}$$

Trace Theorem. Functions in $H^1(\Omega)$ have 'boundary values' in $L^2(\Omega)$, called *trace*, in that there exists a unique linear operator $T : H^1(\Omega) \to L^2(\partial\Omega)$ such that

$$\begin{aligned} \|Tv\|_{L^2(\partial\Omega)} &\le c(\Omega) \|v\|_{H^1(\Omega)} && \text{for all } v \in H^1(\Omega), \\ Tv &= v && \text{for all } v \in C^0(\overline{\Omega}) \cap H^1(\Omega). \end{aligned}$$

Since $Tv = v$ for continuous functions we write v for Tv. For a simplex we give an explicit construction of the constant $c(\Omega)$ in Sect. 6.2. The image of T is a strict subspace of $L^2(\partial\Omega)$, the so-called $H^{1/2}(\partial\Omega)$. The definition of $H_0^1(\Omega)$ can be reconciled with that of traces because

$$H_0^1(\Omega) = \big\{ v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega \big\}.$$

The operator T is also well defined on $W_p^1(\Omega)$ for $1 \le p \le \infty$.

Green's Formula. Given functions $v, w \in H^1(\Omega)$, the following fundamental Green's formula

$$\int_\Omega \partial_i w \, v = -\int_\Omega w \, \partial_i v + \int_{\partial\Omega} w \, v \, n_i \tag{8}$$

holds for any $i = 1, \dots, d$, where $\boldsymbol{n}(x) = [n_1(x), \dots, n_d(x)]^T$ is the outer unit normal of $\partial\Omega$ at x. Equivalently, if $v \in H^1(\Omega)$ and $\boldsymbol{w} \in H^1(\Omega; \mathbb{R}^d)$ then there holds

$$\int_\Omega \operatorname{div} \boldsymbol{w}\, v = -\int_\Omega \boldsymbol{w}\cdot\nabla v + \int_{\partial\Omega} v\,\boldsymbol{w}\cdot\boldsymbol{n}. \tag{9}$$

Green's formula is a direct consequence of Gauß' Divergence Theorem

$$\int_\Omega \operatorname{div} \boldsymbol{w} = \int_{\partial\Omega} \boldsymbol{w}\cdot\boldsymbol{n} \qquad \text{for all } \boldsymbol{w}\in W_1^1(\Omega;\mathbb{R}^d).$$

2.2 Variational formulation

We consider elliptic PDEs that can be formulated as the following variational problem: Let $(\mathbb{V},\langle\cdot,\cdot\rangle_{\mathbb{V}})$ be an Hilbert space with induced norm $\|\cdot\|_{\mathbb{V}}$ and denote by $\mathbb{V}^*$ its dual space equipped with the norm

$$\|f\|_{\mathbb{V}^*} = \sup_{v\in\mathbb{V}} \frac{\langle f, v\rangle}{\|v\|_{\mathbb{V}}} \qquad \text{for all } f\in\mathbb{V}^*.$$

Consider a continuous bilinear form $\mathscr{B}\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$ and $f\in\mathbb{V}^*$. Then we seek a solution $u\in\mathbb{V}$ of

$$u\in\mathbb{V}: \qquad \mathscr{B}[u,v] = \langle f, v\rangle \qquad \text{for all } v\in\mathbb{V}. \tag{10}$$

We first look at several examples that are relevant for the rest of the presentation.

2.2.1 Model Problem

The model problem of this survey is the following 2nd order elliptic PDE

$$-\operatorname{div}(\boldsymbol{A}(x)\nabla u) = f \qquad \text{in } \Omega, \tag{11a}$$
$$u = 0 \qquad \text{on } \partial\Omega, \tag{11b}$$

where $f\in L^2(\Omega)$ and $\boldsymbol{A}\in L^\infty(\Omega;\mathbb{R}^{d\times d})$ is uniformly symmetric positive definite (SPD) over Ω, i. e., there exists constants $0<\alpha_1\le\alpha_2$ such that

$$\alpha_1|\boldsymbol{\xi}|^2 \le \boldsymbol{\xi}^T\boldsymbol{A}(x)\boldsymbol{\xi} \le \alpha_2|\boldsymbol{\xi}|^2 \qquad \text{for all } x\in\Omega, \boldsymbol{\xi}\in\mathbb{R}^d. \tag{12}$$

For the variational formulation of (11) we let $\mathbb{V}=H_0^1(\Omega)$ and denote its dual by $\mathbb{V}^*=H^{-1}(\Omega)$. Since $H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ of functions with vanishing trace, asking for $u\in\mathbb{V}$ accounts for the homogeneous Dirichlet boundary values in (11b).

We next multiply (11a) with a test function $v\in H_0^1(\Omega)$, integrate over Ω and use Green's formula (9), provided $\boldsymbol{w}=-\boldsymbol{A}\nabla u\in H^1(\Omega;\mathbb{R}^d)$, to derive the variational formulation

$$u \in \mathbb{V}: \qquad \int_\Omega \nabla v \cdot \boldsymbol{A}(x) \nabla u = \int_\Omega f v \qquad \text{for all } v \in \mathbb{V}, \tag{13}$$

because the boundary term is zero thanks to $v = 0$ on $\partial\Omega$. However, problem (13) makes sense with much less regularity of the flux $\boldsymbol{w}$. Setting

$$\mathscr{B}[w, v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x) \nabla w \qquad \text{for all } v, w \in H^1_0(\Omega),$$
$$\langle f, v \rangle := \int_\Omega f v \qquad \text{for all } v \in H^1_0(\Omega),$$

(13) formally reads as (10). In Sect. 2.5.1 we analyze further $\mathscr{B}$ and $\langle f, \cdot \rangle$.

2.2.2 Other Boundary Value Problems

We next introduce several elliptic boundary value problems that also fit within the present theory.

General 2nd Order Elliptic Operator. Let $\boldsymbol{A} \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ be uniformly SPD as above, $\boldsymbol{b} \in L^\infty(\Omega; \mathbb{R}^d)$, $c \in L^\infty(\Omega)$, and $f \in L^2(\Omega)$. We now consider the general 2nd order elliptic equation

$$\begin{aligned} -\operatorname{div}(\boldsymbol{A}(x)\nabla u) + \boldsymbol{b}(x) \cdot \nabla u + c(x) u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

The variational formulation utilizes $\mathbb{V} = H^1_0(\Omega)$, as in Sect. 2.2.1. We again multiply the PDE with a test function $v \in H^1_0(\Omega)$, integrate over Ω, and use Green's formula (9) provided $\boldsymbol{A}(x)\nabla u \in H^1(\Omega; \mathbb{R}^d)$. This gives the bilinear form

$$\mathscr{B}[w, v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x) \nabla w + v\, \boldsymbol{b} \cdot \nabla w + c\, v w \qquad \text{for all } v, w \in H^1_0(\Omega)$$

and $\langle f, v \rangle = \int_\Omega f v$ in (10). We examine $\mathscr{B}$ further in Sect. 2.5.2.

The Biharmonic Equation. The vertical displacement u of the mid-surface $\Omega \subset \mathbb{R}^2$ of a clamped plate under a vertical acting force $f \in L^2(\Omega)$ can be modeled by the *biharmonic equation*

$$\Delta^2 u = f \qquad \text{in } \Omega, \tag{14a}$$
$$u = \partial_{\boldsymbol{n}} u = 0 \qquad \text{on } \partial\Omega, \tag{14b}$$

where $\partial_{\boldsymbol{n}} u = \nabla u \cdot \boldsymbol{n}$ is the normal derivative of u on $\partial\Omega$.

For the variational formulation we let $\mathbb{V} = H^2_0(\Omega)$, and note that

$$H^2_0(\Omega) = \left\{ v \in H^2(\Omega) \mid v = \partial_{\boldsymbol{n}} v = 0 \text{ on } \partial\Omega \right\}$$

also accounts for the boundary values (14b). Here, we use Green's formula (9) twice to deduce for all $u \in H^4(\Omega)$ and $v \in H^2(\Omega)$

$$\int_\Omega \Delta^2 u\, v = \int_\Omega \Delta u\, \Delta v + \int_{\partial\Omega} \partial_{\boldsymbol{n}} \Delta u\, v + \int_{\partial\Omega} \Delta u\, \partial_{\boldsymbol{n}} v.$$

Multiplying (14a) with $v \in H^2_0(\Omega)$, integrating over Ω, and using the above formula (without boundary terms), we derive the bilinear form of (10)

$$\mathscr{B}[w, v] := \int_\Omega \Delta v \Delta w \qquad \text{for all } v, w \in \mathbb{V},$$

and set $\langle f, v\rangle := \int_\Omega f\, v$ for $v \in \mathbb{V}$.

The 3d Eddy Current Equations. Given constant material parameters $\mu, \kappa > 0$ and $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^3)$ we next consider the *3d eddy current equations*

$$\operatorname{curl}(\mu \operatorname{curl} \boldsymbol{u}) + \kappa \boldsymbol{u} = \boldsymbol{f} \qquad \text{in } \Omega, \tag{15a}$$

$$u \wedge \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega, \tag{15b}$$

with the curl operator

$$\operatorname{curl} \boldsymbol{v} := \nabla \wedge \boldsymbol{v} = \left[\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}\right]$$

and the vector product $\wedge$ in $\mathbb{R}^3$.

The variational formulation is based on the Sobolev space

$$H(\operatorname{curl};\Omega) := \left\{\boldsymbol{v} \in L^2(\Omega;\mathbb{R}^3) \mid \operatorname{curl} \boldsymbol{v} \in L^2(\Omega;\mathbb{R}^3)\right\}$$

equipped with the norm $\|\boldsymbol{v}\|^2_{H(\operatorname{curl};\Omega)} := \|\boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\operatorname{curl} \boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)}$. This is a Hilbert space and is larger than $H^1(\Omega;\mathbb{R}^3)$. The weak formulation of (15) utilizes the subspace of functions with vanishing tangential trace on $\partial\Omega$

$$\mathbb{V} := H_0(\operatorname{curl};\Omega) = \left\{\boldsymbol{v} \in H(\operatorname{curl};\Omega) \mid \boldsymbol{v} \wedge \boldsymbol{n} = 0 \text{ on } \partial\Omega\right\} = \overline{C_0^\infty(\Omega;\mathbb{R}^3)}^{\|\cdot\|_{H(\operatorname{curl};\Omega)}},$$

which thereby incorporates the boundary values of (15b). This space is a closed and proper subspace of $H(\operatorname{curl};\Omega)$.

From Green's formula (8) with proper choices of $\boldsymbol{v}$ and $\boldsymbol{w}$ it is easy to derive the following formula for all $\boldsymbol{v}, \boldsymbol{w} \in H(\operatorname{curl};\Omega)$

$$\int_\Omega \operatorname{curl} \boldsymbol{w} \cdot \boldsymbol{v} = \int_\Omega \boldsymbol{w} \cdot \operatorname{curl} \boldsymbol{v} + \int_{\partial\Omega} \boldsymbol{w} \cdot (\boldsymbol{v} \wedge \boldsymbol{n}).$$

Multiplying (15a) with a test function $\boldsymbol{v} \in H_0(\operatorname{curl};\Omega)$, integrating over Ω and using the above formula with $\boldsymbol{w} = \mu \operatorname{curl} \boldsymbol{u} \in H(\operatorname{curl};\Omega)$, we end up with the bilinear form and right hand side of (10)

$$\mathscr{B}[\boldsymbol{w},\boldsymbol{v}] := \int_\Omega \mu \operatorname{curl}\boldsymbol{v}\cdot\operatorname{curl}\boldsymbol{w} + \kappa\boldsymbol{v}\cdot\boldsymbol{w} \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in\mathbb{V},$$
$$\langle \boldsymbol{f},\boldsymbol{v}\rangle := \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v} \qquad \text{for all } \boldsymbol{v}\in\mathbb{V}.$$

The Stokes System. Given an external force $\boldsymbol{f}\in L^2(\Omega;\mathbb{R}^d)$, let the velocity-pressure pair $(\boldsymbol{u},p)$ satisfy the momemtum and incompressibility equations with no-slip boundary condition:

$$\begin{aligned} -\Delta\boldsymbol{u}+\nabla p &= \boldsymbol{f} && \text{in } \Omega,\\ \operatorname{div}\boldsymbol{u} &= 0 && \text{in } \Omega,\\ \boldsymbol{u} &= 0 && \text{on } \partial\Omega. \end{aligned}$$

For the variational formulation we consider two Hilbert spaces $\mathbb{V}=H^1_0(\Omega;\mathbb{R}^d)$ and $\mathbb{Q}=L^2_0(\Omega)$, where $L^2_0(\Omega)$ is the space of L^2 functions with zero mean value. The space $H^1_0(\Omega;\mathbb{R}^d)$ takes care of the no-slip boundary values of the velocity. Proceeding as in Sect. 2.2.1, this time using component-wise integration by parts for $\int_\Omega v_i\Delta w_i$ and assuming $\boldsymbol{w}\in H^2(\Omega;\mathbb{R}^d)$, we obtain the bilinear form $a\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$

$$a[\boldsymbol{w},\boldsymbol{v}] := \int_\Omega \nabla\boldsymbol{v}:\nabla\boldsymbol{w} = \sum_{i=1}^d \int_\Omega \nabla v_i\cdot\nabla w_i \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in\mathbb{V}.$$

Likewise, integration by parts of $\int_\Omega \boldsymbol{v}\nabla q$ yields the bilinear form $b\colon \mathbb{Q}\times\mathbb{V}\to\mathbb{R}$

$$b[q,\boldsymbol{v}] := -\int_\Omega q\operatorname{div}\boldsymbol{v} \qquad \text{for all } q\in\mathbb{Q},\ \boldsymbol{v}\in\mathbb{V}.$$

The variational formulation then reads: find $(\boldsymbol{u},p)\in\mathbb{V}\times\mathbb{Q}$ such that

$$\begin{aligned} a[\boldsymbol{u},\boldsymbol{v}]+b[p,\boldsymbol{v}] &= \langle\boldsymbol{f},\boldsymbol{v}\rangle && \text{for all } \boldsymbol{v}\in\mathbb{V},\\ b[q,\boldsymbol{u}] &= 0 && \text{for all } q\in\mathbb{Q}. \end{aligned}$$

We will see in Sect. 2.4.2 how this problem can be formulated in the form (10).

2.3 The inf-sup theory

In this subsection we present a functional analytic theory, the so-called inf-sup theory, that characterizes existence, uniqueness, and continuous dependence on data of the variational problem (10).

Throughout this section we let $(\mathbb{V},\langle\cdot,\cdot\rangle_{\mathbb{V}})$ and $(\mathbb{W},\langle\cdot,\cdot\rangle_{\mathbb{W}})$ be a pair of Hilbert spaces with induced norms $\|\cdot\|_{\mathbb{V}}$ and $\|\cdot\|_{\mathbb{W}}$. We denote by $\mathbb{V}^*$ and $\mathbb{W}^*$ their respective dual spaces equipped with norms

$$\|f\|_{\mathbb{V}^*} = \sup_{v\in\mathbb{V}} \frac{\langle f, v\rangle}{\|v\|_{\mathbb{V}}} \qquad \text{and} \qquad \|g\|_{\mathbb{W}^*} = \sup_{v\in\mathbb{W}} \frac{\langle g, v\rangle}{\|v\|_{\mathbb{W}}}.$$

We write $L(\mathbb{V};\mathbb{W})$ for the space of all linear and continuous operators from $\mathbb{V}$ into $\mathbb{W}$ with operator norm

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \sup_{v\in\mathbb{V}} \frac{\|Bv\|_{\mathbb{W}}}{\|v\|_{\mathbb{V}}}.$$

The following result relates a continuous bilinear form $\mathscr{B}\colon \mathbb{V}\times\mathbb{W}\to\mathbb{R}$ with an operator $B\in L(\mathbb{V};\mathbb{W})$.

Theorem 2.1 (Banach-Nečas). *Let $\mathscr{B}:\mathbb{V}\times\mathbb{W}\to\mathbb{R}$ be a continuous bilinear form with norm*

$$\|\mathscr{B}\| := \sup_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}. \tag{16}$$

Then there exists a unique linear operator $B\in L(\mathbb{V},\mathbb{W})$ such that

$$\langle Bv, w\rangle_{\mathbb{W}} = \mathscr{B}[v,w] \qquad \text{for all } v\in\mathbb{V}, w\in\mathbb{W}$$

with operator norm

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \|\mathscr{B}\|.$$

Moreover, the bilinear form $\mathscr{B}$ satisfies

$$\text{there exists } \alpha>0 \text{ such that } \alpha\|v\|_{\mathbb{V}} \le \sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|w\|_{\mathbb{W}}} \quad \text{for all } v\in\mathbb{V}, \tag{17a}$$

$$\text{for every } 0\neq w\in\mathbb{W} \text{ there exists } v\in\mathbb{V} \text{ such that } \mathscr{B}[v,w]\neq 0, \tag{17b}$$

if and only if $B:\mathbb{V}\to\mathbb{W}$ is an isomorphism with

$$\|B^{-1}\|_{L(\mathbb{W},\mathbb{V})} \le \alpha^{-1}. \tag{18}$$

Proof. 1 *Existence of B.* For fixed $v\in\mathbb{V}$, the mapping $\mathscr{B}[v,\cdot]$ belongs to $\mathbb{W}^*$ by linearity of $\mathscr{B}$ in the second component and continuity of $\mathscr{B}$. Applying the Riesz Representation Theorem (see for instance [16, (2.4.2) Theorem], [38, Theorem 5.7]), we deduce the existence of an element $Bv\in\mathbb{W}$ such that

$$\langle Bv, w\rangle_{\mathbb{W}} = \mathscr{B}[v,w] \quad \text{for all } w\in\mathbb{W}.$$

Linearity of $\mathscr{B}$ in the first argument and continuity of $\mathscr{B}$ imply $B\in L(\mathbb{V};\mathbb{W})$. In view of (16), we get

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \sup_{v\in\mathbb{V}} \frac{\|Bv\|_{\mathbb{W}}}{\|v\|_{\mathbb{V}}} = \sup_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}} \frac{\langle Bv, w\rangle}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} = \sup_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} = \|\mathscr{B}\|.$$

2 *Closed Range of B.* The inf-sup condition (17a) implies

$$\alpha \|v\|_{\mathbb{V}} \leq \sup_{w \in \mathbb{W}} \frac{\langle Bv, w\rangle}{\|w\|_{\mathbb{W}}} = \|Bv\|_{\mathbb{W}} \quad \text{for all } v \in \mathbb{V}, \tag{19}$$

whence B is *injective*. To prove that the range $B(\mathbb{V})$ of B is closed in $\mathbb{W}$, we let $w_k = Bv_k$ be a sequence such that $w_k \to w \in \mathbb{W}$ as $k \to \infty$. We need to show that $w \in B(\mathbb{V})$. Invoking (19), we have

$$\alpha \|v_k - v_j\|_{\mathbb{V}} \leq \|B(v_k - v_j)\|_{\mathbb{W}} = \|w_k - w_j\|_{\mathbb{W}} \to 0$$

as $k, j \to \infty$. Thus $\{v_k\}_{k=0}^{\infty}$ is a Cauchy sequence in $\mathbb{V}$ and so it converges $v_k \to v \in \mathbb{V}$ as $k \to \infty$. Continuity of B yields

$$Bv = \lim_{k \to \infty} Bv_k = w \in B(\mathbb{V}),$$

which shows that $B(\mathbb{V})$ is closed.

3 *Surjectivity of B.* We argue by contradiction, i. e., assume $B(\mathbb{V}) \neq \mathbb{W}$. Since $B(\mathbb{V})$ is closed we can decompose $\mathbb{W} = B(\mathbb{V}) \oplus B(\mathbb{V})^{\perp}$, where $B(\mathbb{V})^{\perp}$ is the orthogonal complement of $B(\mathbb{V})$ in $\mathbb{W}$ (see for instance [16, (2.3.5) Proposition], [38, Theorem 5.6]). By assumption $B(\mathbb{V})^{\perp}$ is non-trivial, i. e., there exists $0 \neq w_0 \in B(\mathbb{V})^{\perp}$. This is equivalent to

$$w_0 \neq 0 \quad \text{and} \quad \langle w, w_0\rangle = 0 \quad \text{for all } w \in B(\mathbb{V}),$$

or

$$w_0 \neq 0 \quad \text{and} \quad 0 = \langle Bv, w_0\rangle = \mathscr{B}[v, w_0] \quad \text{for all } v \in \mathbb{V}.$$

This in turn contradicts (17b) and shows that $B(\mathbb{V}) = \mathbb{W}$. Therefore, we conclude that B is an isomorphism from $\mathbb{V}$ onto $\mathbb{W}$.

4 *Property* (18). We rewrite (19) as follows:

$$\alpha \|B^{-1}w\|_{\mathbb{V}} \leq \|w\|_{\mathbb{W}} \quad \text{for all } w \in \mathbb{W},$$

which is (18) in disguise.

5 *Property* (18) *implies* (17a) *and* (17b). Compute

$$\begin{aligned}\inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathscr{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} &= \inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\langle Bv, w\rangle}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \inf_{v \in \mathbb{V}} \frac{\|Bv\|_{\mathbb{W}}}{\|v\|_{\mathbb{V}}} \\ &= \inf_{w \in \mathbb{W}} \frac{\|w\|_{\mathbb{W}}}{\|B^{-1}w\|_{\mathbb{V}}} = \frac{1}{\sup_{w \in \mathbb{W}} \frac{\|B^{-1}w\|_{\mathbb{V}}}{\|w\|_{\mathbb{W}}}} = \frac{1}{\|B^{-1}\|} \geq \alpha\end{aligned}$$

which shows (17a). Property (17b) is a consequence of B being an isomorphism: there exists $0 \neq v \in \mathbb{V}$ such that $Bv = w$ and

$$\mathscr{B}[v, w] = \langle Bv, w\rangle = \|w\|_{\mathbb{W}}^2 \neq 0.$$

This concludes the theorem. □

We are now in the position to characterize properties of the bilinear form $\mathscr{B}$ in (10) that imply that the variational problem (10) is well-posed. This result from 1962 is due to Nečas [56, Theorem 3.3].

Theorem 2.2 (Nečas Theorem). *Let $\mathscr{B}\colon \mathbb{V}\times\mathbb{W}\to\mathbb{R}$ be a continuous bilinear form. Then the variational problem*

$$u\in\mathbb{V}: \qquad \mathscr{B}[u,v]=\langle f,v\rangle \qquad \text{for all } v\in\mathbb{W}, \tag{20}$$

admits a unique solution $u\in\mathbb{V}$ for all $f\in\mathbb{W}^*$, which depends continuously on f, if and only if the bilinear form $\mathscr{B}$ satisfies one of the equivalent inf-sup conditions*:*

(1) *There exists $\alpha>0$ such that*

$$\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|w\|_{\mathbb{W}}}\geq\alpha\|v\|_{\mathbb{V}} \quad \text{for some } \alpha>0; \tag{21a}$$

$$\text{for every } 0\neq w\in\mathbb{W} \text{ there exists } v\in\mathbb{V} \text{ such that } \mathscr{B}[v,w]\neq 0. \tag{21b}$$

(2) *There holds*

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}>0, \qquad \inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}>0. \tag{22}$$

(3) *There exists $\alpha>0$ such that*

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\alpha. \tag{23}$$

In addition, the solution u of (20) satisfies the stability estimate

$$\|u\|_{\mathbb{V}}\leq\alpha^{-1}\|f\|_{\mathbb{W}^*}. \tag{24}$$

Proof. 1 Denote by $J\colon\mathbb{W}\to\mathbb{W}^*$ the isometric Riesz isomorphism between $\mathbb{W}$ and $\mathbb{W}^*$; see [16, (2.4.2) Theorem], [38, Theorem 5.7]. Let $B\in L(\mathbb{V};\mathbb{W})$ be the linear operator corresponding to $\mathscr{B}$ introduced in Theorem 2.1. Then (20) is equivalent to

$$u\in\mathbb{V}: \qquad Bu=J^{-1}f \qquad \text{in } \mathbb{W}.$$

Assume that (21) is satisfied. Then, according to Theorem 2.1, the operator B is invertible. For any $f\in\mathbb{W}^*$ the unique solution $u\in\mathbb{V}$ is given by $u=B^{-1}J^{-1}f$ and u depends continuously on f with

$$\|u\|_{\mathbb{V}}\leq\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}\|J^{-1}f\|_{\mathbb{W}}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}\|f\|_{\mathbb{W}^*}\leq\alpha^{-1}\|f\|_{\mathbb{W}^*}.$$

Conversely, if (20) admits a unique solution u for any $f\in\mathbb{W}^*$, then B has to be invertible, which implies (21) by Theorem 2.1.

2 To show the equivalence of the inf-sup conditions (21), (22), and (23) we rewrite Step 5 of the proof of Theorem 2.1:

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}.$$

Furthermore,

$$\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\langle Bv,w\rangle_{\mathbb{W}}}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\langle v,B^{*}w\rangle_{\mathbb{V}}}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-*}\|^{-1}_{L(\mathbb{V};\mathbb{W})},$$

where $B^{*}\colon \mathbb{W}\to\mathbb{V}$ is the adjoint operator of B and $B^{-*}\colon \mathbb{V}\to\mathbb{W}$ is its inverse. Recalling that $\|B^{*}\|_{L(\mathbb{W};\mathbb{V})}=\|B\|_{L(\mathbb{V};\mathbb{W})}$ and $\|B^{-*}\|_{L(\mathbb{V};\mathbb{W})}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}$ we deduce the desired expression

$$\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}.$$

and conclude the proof. □

The equality in (23) might seem at first surprising but is just a consequence of $\|B^{-*}\|_{L(\mathbb{V};\mathbb{W})}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}$. In general, (21) is simpler to verify than (23) and α of (23) is the largest possible α in (21a). Moreover, the above proof readily gives the following result.

Corollary 2.1 (Well Posedness vs. Inf-Sup). *Assume that the variational problem* (20) *admits a unique solution* $u\in\mathbb{V}$ *for all* $f\in\mathbb{W}^{*}$ *so that*

$$\|u\|_{\mathbb{V}}\le C\|f\|_{\mathbb{W}^{*}}.$$

Then $\mathscr{B}$ *satisfies the inf-sup condition* (23) *with* $\alpha\ge C^{-1}$.

Proof. Since (20) admits a unique solution u for all f, we conclude that the operator $B\in L(\mathbb{V};\mathbb{W})$ of Theorem 2.1 is invertible and the solution operator $B^{-1}\in L(\mathbb{W};\mathbb{V})$ is bounded with norm $\|B^{-1}\|_{L(\mathbb{W};\mathbb{W})}\le C$, thanks to $\|u\|_{\mathbb{V}}\le C\|f\|_{\mathbb{W}^{*}}$. On the other hand, Step 2 in the proof of Theorem 2.2 shows that $\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}$ is the optimal inf-sup constant α for $\mathscr{B}$, which yields $\alpha\ge C^{-1}$. □

2.4 Two special problem classes

We next study two special cases included in the inf-sup theory. The first class are problems with coercive bilinear form and the second one comprises problems of saddle point type.

2.4.1 Coercive Bilinear Forms

An existence and uniqueness result for coercive bilinear forms was established by Lax and Milgram eight years prior to the result by Nečas [45]. Coercivity of $\mathscr{B}$ is a sufficient condition for existence and uniqueness but it is not necessary.

Corollary 2.2 (Lax-Milgram Theorem). *Let $\mathscr{B}\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$ be a continuous bilinear form that is* coercive*, namely there exists $\alpha>0$ such that*

$$\mathscr{B}[v,v]\geq\alpha\|v\|_{\mathbb{V}}^2 \qquad \text{for all } v\in\mathbb{V}. \tag{25}$$

Then (10) has a unique solution that satisfies (24).

Proof. Since (25) implies $\sup_{w\in\mathbb{V}}\mathscr{B}[v,w]\geq\mathscr{B}[v,v]\geq\alpha\|v\|_{\mathbb{V}}^2$ for all $0\neq v\in\mathbb{V}$, both (21a) and (21b) follow immediately, whence Theorem 2.2 implies the assertion. □

If the bilinear form $\mathscr{B}$ is also symmetric, i. e.,

$$\mathscr{B}[v,w]=\mathscr{B}[w,v] \qquad \text{for all } v,w\in\mathbb{V},$$

then $\mathscr{B}$ is a scalar product on $\mathbb{V}$. The norm induced by $\mathscr{B}$ is the so-called *energy norm*

$$|||v|||_{\Omega}:=\mathscr{B}[v,v]^{1/2}.$$

Coercivity and continuity of $\mathscr{B}$ in turn imply that $|||\cdot|||_{\Omega}$ is equivalent to the natural norm $\|\cdot\|_{\mathbb{V}}$ in $\mathbb{V}$ since

$$\alpha\|v\|_{\mathbb{V}}^2\leq|||v|||_{\Omega}^2\leq\|\mathscr{B}\|\|v\|_{\mathbb{V}}^2 \qquad \text{for all } v\in\mathbb{V}. \tag{26}$$

Moreover, it is rather easy to show that for symmetric and coercive $\mathscr{B}$ the solution u of (10) is the unique minimizer of the quadratic energy

$$J[v]:=\frac{1}{2}\mathscr{B}[v,v]-\langle f,v\rangle \qquad \text{for all } v\in\mathbb{V},$$

i. e., $u=\operatorname{argmin}_{v\in\mathbb{V}}J[v]$. The energy norm and the quadratic energy play a relevant role in both Chap. 8 and Chap. 9.

2.4.2 Saddle Point Problems

Given a pair of Hilbert spaces $(\mathbb{V},\mathbb{Q})$, we consider two continuous bilinear forms $a\colon\mathbb{V}\times\mathbb{V}\to\mathbb{R}$ and $b\colon\mathbb{Q}\times\mathbb{V}\to\mathbb{R}$. If $f\in\mathbb{V}^*$ and $g\in\mathbb{Q}^*$, then we seek a pair $(u,p)\in\mathbb{V}\times\mathbb{Q}$ solving the *saddle point problem*

$$a[u,v]+b[p,v]=\langle f,v\rangle \qquad \text{for all } v\in\mathbb{V}, \tag{27a}$$
$$b[q,u]=\langle g,q\rangle \qquad \text{for all } q\in\mathbb{Q}. \tag{27b}$$

Problem (27) is a variational problem which can of course be stated in the form (10). In doing so we define the product space $\mathbb{W} := \mathbb{V} \times \mathbb{Q}$, which is a Hilbert space with scalar product

$$\langle (v,q), (w,r) \rangle_{\mathbb{W}} := \langle v, w \rangle_{\mathbb{V}} + \langle q, r \rangle_{\mathbb{Q}} \qquad \text{for all } (v,q), (w,r) \in \mathbb{W}$$

and induced norm $\|(v,q)\|_{\mathbb{W}} := (\|v\|_{\mathbb{V}}^2 + \|q\|_{\mathbb{Q}}^2)^{1/2}$. From the bilinear forms a and b we define the bilinear form $\mathscr{B}\colon \mathbb{W} \times \mathbb{W} \to \mathbb{R}$ by

$$\mathscr{B}[(v,q), (w,r)] := a[v,w] + b[q,w] + b[r,v] \qquad \text{for all } (v,q), (w,r) \in \mathbb{W}.$$

Then, (27) is equivalent to the problem

$$(u,p) \in \mathbb{W}: \qquad \mathscr{B}[(u,p), (v,q)] = \langle f, v \rangle + \langle g, q \rangle \qquad \text{for all } (v,q) \in \mathbb{W}. \tag{28}$$

To see this, test (28) first with $(v,0)$, which gives (27a), and then utilizing $(0,q)$ yields (27b). Obviously, a solution (u,p) to (27) is a solution to (28) and vice versa.

Therefore, the saddle point problem (27) is well-posed if and only if $\mathscr{B}$ satisfies the inf-sup condition (23). Since $\mathscr{B}$ is defined via the bilinear forms a and b and due to the degenerate structure of (27) it is not that simple to show (23). However it is a direct consequence of the inf-sup theorem for saddle point problems given by Brezzi in 1974 [17].

Theorem 2.3 (Brezzi Theorem). *The saddle point problem (27) has a unique solution $(u,p) \in \mathbb{V} \times \mathbb{Q}$ for all data $(f,g) \in \mathbb{V}^* \times \mathbb{Q}^*$, that depends continuously on data, if and only if there exist constants $\alpha, \beta > 0$ such that*

$$\inf_{v \in \mathbb{V}_0} \sup_{w \in \mathbb{V}_0} \frac{a[v,w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}}} = \inf_{w \in \mathbb{V}_0} \sup_{v \in \mathbb{V}_0} \frac{a[v,w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}}} = \alpha > 0, \tag{29a}$$

$$\inf_{q \in \mathbb{Q}} \sup_{v \in \mathbb{V}} \frac{b[q,v]}{\|q\|_{\mathbb{Q}} \|v\|_{\mathbb{V}}} = \beta > 0, \tag{29b}$$

where

$$\mathbb{V}_0 := \{ v \in \mathbb{V} \mid b[q,v] = 0 \text{ for all } q \in \mathbb{Q} \}.$$

In addition, there exists $\gamma = \gamma(\alpha, \beta, \|a\|)$ such that the solution (u,p) is bounded by

$$\left(\|u\|_{\mathbb{V}}^2 + \|p\|_{\mathbb{Q}}^2 \right)^{1/2} \leq \gamma \left(\|f\|_{\mathbb{V}^*}^2 + \|g\|_{\mathbb{Q}^*}^2 \right)^{1/2}. \tag{30}$$

Proof. 1 Continuity of b implies that the subspace $\mathbb{V}_0$ of $\mathbb{V}$ is closed. We therefore can decompose $\mathbb{V} = \mathbb{V}_0 \oplus \mathbb{V}_\perp$ where $\mathbb{V}_\perp$ is the orthogonal complement of $\mathbb{V}_0$ in $\mathbb{V}$; see [16, (2.3.5) Proposition], [38, Theorem 5.6]. Both $\mathbb{V}_0$ and $\mathbb{V}_\perp$ are Hilbert spaces. 2 The inf-sup condition (29b) is (21a) for $\mathscr{B} = b$. On the other hand, by definition of $\mathbb{V}_0$, for every $v \in \mathbb{V}_\perp$ there exists a $q \in \mathbb{Q}$ with $b[q,v] \neq 0$, which is (21b). Hence, the equivalence of (21) and (23) implies that the operators $B : \mathbb{Q} \to \mathbb{V}_\perp$ and $B^* : \mathbb{V}_\perp \to \mathbb{Q}$ defined by

$$\langle Bq, v\rangle_{\mathbb{V}} = \langle B^* v, q\rangle_{\mathbb{Q}} = b[q,v] \qquad \text{for all } q \in \mathbb{Q}, v \in \mathbb{V}_\perp,$$

are isomorphisms.

3 We write the solution $u = u_0 + u_\perp$ with $u_0 \in \mathbb{V}_0$ and $u_\perp \in \mathbb{V}_\perp$ to be determined as follows. Since B^* is an isomorphism, the problem

$$u_\perp \in \mathbb{V}_\perp : \qquad b[q,u_\perp] = \langle B^* v, q\rangle_{\mathbb{Q}} = \langle g, q\rangle \qquad \text{for all } q \in \mathbb{Q} \tag{31}$$

is well-posed for all $g \in \mathbb{Q}^*$, and selects $u_\perp$ uniquely. We next consider

$$u_0 \in \mathbb{V}_0 : \qquad a[u_0, v] = \langle f, v\rangle - a[u_\perp, v] \qquad \text{for all } v \in \mathbb{V}_0. \tag{32}$$

This problem admits a unique solution u_0 thanks to (29b), which is (23) with $\mathscr{B} = a$.

4 Upon setting

$$\langle F, v\rangle := \langle f, v\rangle - a[u,v] \qquad \text{for all } v \in \mathbb{V}$$

we see that $F \in \mathbb{V}_\perp^*$ because $\langle F, v\rangle = 0$ for all $v \in \mathbb{V}_0$ by (32). Since B is an isomorphism, there is a unique solution of

$$p \in \mathbb{Q} : \qquad b[p,v] = \langle Bp, v\rangle_{\mathbb{V}} = \langle F, v\rangle \qquad \text{for all } v \in \mathbb{V}_\perp. \tag{33}$$

This construction yields the desired pair (u,p) and shows that problems (31), (32), and (33) are well-posed if and only if b satisfies (29b) and a fulfills (29a).

5 We conclude by estimating (u,p). In view of (29b), $u_\perp$ is bounded by

$$\|u_\perp\|_{\mathbb{V}} \le \beta^{-1}\|g\|_{\mathbb{Q}^*}$$

which, in conjunction with (29a), implies for u_0

$$\|u_0\|_{\mathbb{V}} \le \alpha^{-1}\big(\|f\|_{\mathbb{V}^*} + \|a\|\|u_\perp\|_{\mathbb{V}}\big) \le \alpha^{-1}\|f\|_{\mathbb{V}^*} + \|a\|(\alpha\beta)^{-1}\|g\|_{\mathbb{Q}^*}.$$

Hence,

$$\|u\|_{\mathbb{V}} \le \|u_0\|_{\mathbb{V}} + \|u_\perp\|_{\mathbb{V}} \le \alpha^{-1}\|f\|_{\mathbb{V}^*} + \big(1 + \alpha^{-1}\|a\|\big)\beta^{-1}\|g\|_{\mathbb{Q}^*}.$$

Finally, using $\|F\|_{\mathbb{V}_\perp^*} = \|F\|_{\mathbb{V}^*} \le \|f\|_{\mathbb{V}^*} + \|a\|\|u\|_{\mathbb{V}}$, (29b) gives the bound for p

$$\|p\|_{\mathbb{Q}} \le \beta^{-1}\|F\|_{\mathbb{V}^*} \le \beta^{-1}\big(1 + \alpha^{-1}\|a\|\big)\,\big(\|f\|_{\mathbb{V}^*} + \beta^{-1}\|a\|\|g\|_{\mathbb{Q}^*}\big).$$

Adding the two estimates gives the stability bound (30) with $\gamma = \gamma(\alpha, \beta, \|a\|)$. □

Remark 2.1 (Optimal constant). A better bound of the stability constant γ in terms of α, β and $\|a\|$ is available. Setting

$$\kappa := \frac{\|a\|}{\beta}, \quad \kappa_{11} := \frac{1+\kappa^2}{\alpha^2}, \quad \kappa_{22} := \kappa^2\kappa_{11} + \frac{1}{\beta^2}, \quad \kappa_{12} := \kappa\kappa_{11},$$

Xu and Zikatanov have derived the bound [79]

$$\gamma \leq \kappa_{12} + \max(\kappa_{11}, \kappa_{22}).$$

For establishing this improved bound one has to make better use of the orthogonal decomposition $\mathbb{V} = \mathbb{V}_0 \oplus \mathbb{V}_\perp$ when estimating $u = u_0 + u_\perp$ and one has to resort to a result of Kato for non-trivial idempotent operators [42].

Combining the Brezzi theorem with Corollary 2.1 we infer the inf-sup condition for the bilinear form $\mathscr{B}$ in (28).

Corollary 2.3 (Inf-Sup of $\mathscr{B}$). *Let the bilinear form $\mathscr{B}\colon \mathbb{W} \to \mathbb{W}$ be defined by (28).*

Then there holds

$$\inf_{(v,q)\in\mathbb{W}} \sup_{(w,r)\in\mathbb{W}} \frac{\mathscr{B}[(v,q),(w,r)]}{\|(v,q)\|_{\mathbb{W}}\|(w,r)\|_{\mathbb{W}}} = \inf_{(w,r)\in\mathbb{W}} \sup_{(v,q)\in\mathbb{W}} \frac{\mathscr{B}[(v,q),(w,r)]}{\|(v,q)\|_{\mathbb{W}}\|(w,r)\|_{\mathbb{W}}} \geq \gamma^{-1},$$

where γ is the stability constant from Theorem 2.3.

Assume that $a\colon \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ is symmetric and let (u,p) be the solution to (27). Then u is the unique minimizer of the energy $J[v] := \frac{1}{2}a[v,v] - \langle f, v\rangle$ under the constraint $b[\cdot,u] = g$ in $\mathbb{Q}^*$. In view of this, p is the corresponding Lagrange multiplier and the pair (u,p) is the unique saddle point of the Lagrangian

$$L[v,q] := J[v] + b[q,v] - \langle g, q\rangle \qquad \text{for all } v\in\mathbb{V}, q\in\mathbb{Q}.$$

The Brezzi theorem also applies to non-symmetric a, in which case the pair (u,p) is no longer a saddle point.

2.5 Applications

We now review the examples introduced in Sect. 2.2 in light of the inf-sup theory.

2.5.1 Model Problem

Since $\boldsymbol{A}$ is symmetric, the variational formulation of the model problem in Sect. 2.2.1 leads to the symmetric bilinear form $\mathscr{B}\colon H^1_0(\Omega)\times H^1_0(\Omega) \to \mathbb{R}$ defined by

$$\mathscr{B}[w,v] := \int_\Omega \nabla v\cdot\boldsymbol{A}(x)\nabla w, \qquad \text{for all } v,w\in H^1_0(\Omega).$$

We have to decide which norm to use on $H^1_0(\Omega)$. The Poincaré-Friedrichs inequality (7) implies the equivalence of $\|\cdot\|_{H^1(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$ on $H^1_0(\Omega)$ because

$$|v|_{H^1(\Omega)} \leq \|v\|_{H^1(\Omega)} \leq \left(1 + C_d^2\,|\Omega|^{2/d}\right)^{1/2} |v|_{H^1(\Omega)} \qquad \text{for all } v\in H^1_0(\Omega). \quad (34)$$

On the other hand, assumption (12) on the eigenvalues of $\boldsymbol{A}$ directly leads to

$$\alpha_1 |v|^2_{H^1(\Omega)} \leq \mathscr{B}[v,v] \leq \alpha_2 |v|^2_{H^1(\Omega)} \qquad \text{for all } v \in H^1_0(\Omega).$$

Therefore, $|\cdot|_{H^1_0(\Omega)}$ is a convenient norm on $\mathbb{V} = H^1_0(\Omega)$ for the model problem, for which $\mathscr{B}$ is coercive with constant $\alpha = \alpha_1$ and continuous with norm $\|\mathscr{B}\| = \alpha_2$.

To apply the Lax-Milgram theorem it remains to show that $f \in L^2(\Omega)$ implies $f \in \mathbb{V}^* = H^{-1}(\Omega)$, in the sense that $v \mapsto \int_\Omega f v$ belongs to $H^1 - 1(\Omega)$. Recalling

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v \in H^1_0(\Omega)} \frac{\langle f, v\rangle}{|v|_{H^1(\Omega)}}.$$

and using the Poincaré-Friedrichs inequality (7) once more we estimate

$$|\langle f, v\rangle| = \left| \int_\Omega f v \right| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq C_d\, |\Omega|^{1/d}\, \|f\|_{L^2(\Omega)}\, |v|_{H^1(\Omega)},$$

and therefore $\|f\|_{H^{-1}(\Omega)} \leq C_d\, |\Omega|^{1/d}\, \|f\|_{L^2(\Omega)}$. In view of Corollary 2.2, we have the stability bound

$$|u|_{H^1(\Omega)} \leq \frac{C_d |\Omega|^{1/d}}{\alpha_1^{1/2}} \|f\|_{L^2(\Omega)}.$$

Since $\mathscr{B}$ is symmetric and coercive, it defines a scalar product in $H^1_0(\Omega)$. Consequently, an even more convenient choice of norm on $\mathbb{V}$ is the energy norm $|||\cdot|||_\Omega = \mathscr{B}[\cdot,\cdot]^{1/2}$. In this case we have $\alpha = \|\mathscr{B}\| = 1$ and

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v \in H^1_0(\Omega)} \frac{\langle f, v\rangle}{|||v|||_\Omega} \leq \frac{C_d\, |\Omega|^{1/d}}{\alpha_1^{1/2}} \|f\|_{L^2(\Omega)}$$

whence we obtain the same stability estimate as above.

2.5.2 Other Boundary Value Problems

We now review the examples from Sect. 2.2.2.

General 2nd Order Elliptic Operator. We take $\mathbb{V} = H^1_0(\Omega)$ and the bilinear form

$$\mathscr{B}[w,v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x) \nabla w + v\, \boldsymbol{b} \cdot \nabla w + c\, v\, w \qquad \text{for all } v, w \in H^1_0(\Omega).$$

A straightforward estimate shows continuity of $\mathscr{B}$ with respect to the norm $\|\cdot\|_{H^1(\Omega)}$

$$|\mathscr{B}[w,v]| \leq \|\mathscr{B}\|\, \|v\|_{H^1(\Omega)}\, \|w\|_{H^1(\Omega)} \qquad \text{for all } v, w \in H^1(\Omega)$$

with operator norm $\|\mathscr{B}\| \leq \alpha_2 + \|\boldsymbol{b}\|_{L^\infty(\Omega;\mathbb{R}^d)} + \|c\|_{L^\infty(\Omega)}$.

Assume now that $\operatorname{div} \boldsymbol{b}$ is bounded and $c - \frac{1}{2}\operatorname{div} \boldsymbol{b} \geq 0$ in Ω. In light of Green's formula (9) we get the identity

$$\int_\Omega v\,\boldsymbol{b}\cdot\nabla w = -\int_\Omega \nabla v\cdot\boldsymbol{b}\,w - \int_\Omega \operatorname{div}\boldsymbol{b}\,v\,w \qquad \text{for all } v,w\in H_0^1(\Omega),$$

whence $\int_\Omega v\,\boldsymbol{b}\cdot\nabla v = -\frac{1}{2}\int_\Omega \operatorname{div}\boldsymbol{b}\,v^2$. If $C = C_d\,|\Omega|^{1/d}$ is the Poincarè-Friedrichs constant for Ω, then we deduce as in Sect. 2.2.1 for any $v\in H_0^1(\Omega)$

$$\mathscr{B}[v,v] \geq \alpha_1\,|v|^2_{H^1(\Omega)} + \int_\Omega \big(c - \tfrac{1}{2}\operatorname{div}\boldsymbol{b}\big)v^2 \geq \alpha_1\,|v|^2_{H^1(\Omega)} \geq \frac{\alpha_1}{1+C^2}\|v\|^2_{H^1(\Omega)},$$

thanks to the norm equivalence (34). Using $\|\cdot\|_{H^1(\Omega)}$ as norm on $\mathbb{V}$ we have

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v\in H_0^1(\Omega)} \frac{\langle f, v\rangle}{\|v\|_{H^1(\Omega)}} \leq \|f\|_{L^2(\Omega)}.$$

Assuming only $c\geq 0$ the bilinear form $\mathscr{B}$ is no longer coercive. Nevertheless, for any bounded $\boldsymbol{b}$ and $c\geq 0$ it can be shown that $\mathscr{B}$ satisfies the inf-sup condition (23) but the proof is not elementary; see for instance [9].

The Biharmonic Equation. For the variational formulation of the biharmonic equation we use the Hilbert space $\mathbb{V} = H_0^2(\Omega)$ and claim that $\|\Delta\cdot\|_{L^2(\Omega)}$ is a norm on $H_0^2(\Omega)$ that is equivalent to $\|\cdot\|_{H^2(\Omega)}$. From Green's formula we deduce for $v\in C_0^\infty(\Omega)$

$$|v|^2_{H^2(\Omega)} = \sum_{i,j=1}^d \int_\Omega \big(\partial^2_{ij} v\big)^2 = -\sum_{i,j=1}^d \int_\Omega \partial_i v\,\partial^3_{ijj} v = \sum_{i,j=1}^d \int_\Omega \partial^2_{ii} v\,\partial^2_{jj} v = \|\Delta v\|^2_{L^2(\Omega)}.$$

Using density we thus conclude $|v|_{H^2(\Omega)} = \|\Delta v\|_{L^2(\Omega)}$ for all $v\in H_0^2(\Omega)$. For those functions v the Poincaré-Friedrichs inequality (7) implies $|v|_{H^1(\Omega)} \leq c(\Omega)\,|v|_{H^2(\Omega)}$ which, in conjunction with the norm equivalence (34), yields

$$\|\Delta v\|_{L^2(\Omega)} \leq \|v\|_{2,\Omega} \leq C(\Omega)\,|v|_{H^2(\Omega)} = C(\Omega)\,\|\Delta v\|_{L^2(\Omega)}. \tag{35}$$

The bilinear form $\mathscr{B}$ given by

$$\mathscr{B}[w,v] = \int_\Omega \Delta v\,\Delta w$$

is symmetric and the energy norm $|||\cdot|||_\Omega$ coincides with the norm $\|\Delta\cdot\|_{L^2(\Omega)}$. Therefore, $\mathscr{B}$ is continuous and coercive on $H_0^2(\Omega)$ with constants $\|B\| = \alpha = 1$.

We denote by $H^{-2}(\Omega)$ the dual space of $H_0^2(\Omega)$. The norm equivalence (35) implies $\|f\|_{H^{-2}(\Omega)} \leq C(\Omega)\|f\|_{L^2(\Omega)}$ for $f\in L^2(\Omega)$.

The 3d Eddy Current Equations. We take $\mathbb{V} = H_0(\operatorname{curl};\Omega)$ along with the symmetric bilinear form

$$\mathscr{B}[\boldsymbol{w},\boldsymbol{v}] := \int_\Omega \mu \operatorname{curl}\boldsymbol{v}\cdot\operatorname{curl}\boldsymbol{w} + \kappa\boldsymbol{v}\cdot\boldsymbol{w} \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in\mathbb{V}$$

and subordinate energy norm

$$|||\boldsymbol{v}|||_\Omega^2 = \|\mu^{1/2}\operatorname{curl}\boldsymbol{v}\|_{L^2(\Omega;\mathbb{R}^3)}^2 + \|\kappa^{1/2}\boldsymbol{v}\|_{L^2(\Omega;\mathbb{R}^3)}^2.$$

Since $\mu,\kappa>0$, this norm and the corresponding $H(\operatorname{curl};\Omega)$ norm (i.e. $\mu=\kappa=1$) are equivalent. Accordingly, $\mathscr{B}$ is continuous and coercive with respect to $|||\cdot|||_\Omega$ with $\|\mathscr{B}\|=\alpha=1$.

Furthermore, any $\boldsymbol{f}\in L^2(\Omega;\mathbb{R}^3)$ belongs to the dual space $\mathbb{V}^* = (H_0(\operatorname{curl};\Omega))^*$ and $\|\boldsymbol{f}\|_{\mathbb{V}^*}\le\kappa^{-1/2}\|\boldsymbol{f}\|_{L^2(\Omega;\mathbb{R}^3)}$.

The Stokes System. We use the Hilbert spaces $\mathbb{V}=H_0^1(\Omega;\mathbb{R}^d)$ equipped with the norm $|\cdot|_{H_0^1(\Omega;\mathbb{R}^d)}$ and $\mathbb{Q}=L_0^2(\Omega)$ equipped with $\|\cdot\|_{L^2(\Omega)}$. With this choice, $\|\cdot\|_{\mathbb{V}}$ is the energy norm associated with the bilinear form $a[\boldsymbol{w},\boldsymbol{v}]=\int_\Omega\nabla\boldsymbol{v}:\nabla\boldsymbol{w}$. Therefore, a is continuous and coercive on $\mathbb{V}$ with $\|a\|=\alpha=1$. This implies the inf-sup condition (29a).

Using integration by parts one can show $\|\operatorname{div}\boldsymbol{v}\|_{L^2(\Omega)}\le|\boldsymbol{v}|_{H_0^1(\Omega;\mathbb{R}^d)}$, whence the bilinear form $b[q,\boldsymbol{v}]=-\int_\Omega\operatorname{div}\boldsymbol{v}\,q$ is continuous with norm $\|b\|=1$. In addition, for any $q\in L_0^2(\Omega)$ there exists a $\boldsymbol{w}\in H_0^1(\Omega;\mathbb{R}^d)$ such that

$$-\operatorname{div}\boldsymbol{w}=q \quad \text{in } \Omega \qquad \text{and} \qquad |\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)}\le C(\Omega)\|q\|_{L^2(\Omega)}.$$

This non-trivial result goes back to Nečas [19] and a proof can for instance be found in [36, Theorem III.3.1]. This implies

$$\sup_{\boldsymbol{v}\in H_0^1(\Omega;\mathbb{R}^d)}\frac{b[q,\boldsymbol{v}]}{|\boldsymbol{v}|_{H^1(\Omega;\mathbb{R}^d)}} \ge \frac{b[q,\boldsymbol{w}]}{|\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)}} = \frac{\|q\|_{L^2(\Omega)}^2}{|\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)}} \ge C(\Omega)^{-1}\|q\|_{L^2(\Omega)}.$$

Therefore, (29b) holds with $\beta\ge C(\Omega)^{-1}$ and Theorem 2.3 applies for all $\boldsymbol{f}\in L^2(\Omega;\mathbb{R}^d)$ and gives existence, uniqueness and stability of the solution $(\boldsymbol{u},p)\in\mathbb{V}\times\mathbb{Q}$ of the Stokes system.

2.6 Problems

Problem 2.1. Let $\Omega=(0,1)$ and $u\in W_p^1(\Omega)$ with $1<p\le\infty$. Prove that the function u is $(p-1)/p$-Hölder continuous, namely

$$|u(x)-u(y)|\le|x-y|^{(p-1)/p}\|u'\|_{L^p(\Omega)} \qquad \text{for all } x,y\in\Omega.$$

If $p=1$, then $u\in W_1^1(\Omega)$ is uniformly continuous in $\overline{\Omega}$ because of the absolute continuity of the integral.

Problem 2.2. Find the weak gradient of $v(x) = \log\log(|x|/2)$ in the unit ball Ω, and show that $v \in W_d^1(\Omega)$ for $d \geq 2$. This shows that functions in $W_d^1(\Omega)$, and in particular in $H^1(\Omega)$, may not be continuous, and even bounded, in dimension $d \geq 2$.

Problem 2.3. Prove the following simplified version of the Poincaré-Friedrichs inequality (7): let Ω be contained in the strip $\{x \in \mathbb{R}^d \mid 0 < x_d < h\}$; then

$$\|v\|_{L^2(\Omega)} \lesssim h\|\nabla v\|_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

To this end, take $v \in C_0^\infty(\Omega)$, write for $0 < s < h$

$$v^2(x,s) = v^2(x,0) + 2\int_0^s \partial_d v \cdot v, \tag{36}$$

integrate, and use Cauchy-Schwarz inequality to prove (7). Next use a density argument, based on the definition of $H_0^1(\Omega)$, to extend the inequality to $H_0^1(\Omega)$.

Problem 2.4. Let $\Omega_h = \{(x', x_d) \mid |x| < h, x_d > 0\}$ be the upper half ball in $\mathbb{R}^d$ of radius $h > 0$ centered at the origin. Let Γ_h be the flat part of $\partial\Omega_h$.

(a) Let $\zeta \geq 0$ be a C_0^∞ cut-off function in the unit ball that equals 1 in the ball of radius $1/2$. Use the identity (36) for $v\zeta$, followed by a density argument, to derive the trace inequality

$$\|v\|^2_{L^2(\Gamma_{1/2})} \lesssim \|v\|^2_{L^2(\Omega_1)} + \|\nabla v\|^2_{L^2(\Omega_1)} \qquad \text{for all } \; v \in H^1(\Omega_1).$$

(b) Use a scaling argument to Ω_h to deduce the scaled trace inequality

$$\|v\|^2_{L^2(\Gamma_{h/2})} \lesssim h^{-1}\|v\|^2_{L^2(\Omega_h)} + h\|\nabla v\|^2_{L^2(\Omega_h)} \qquad \text{for all } \; v \in H^1(\Omega_h).$$

Problem 2.5. Show that $\operatorname{div} \boldsymbol{q} \in H^{-1}(\Omega)$ for $\boldsymbol{q} \in L^2(\Omega;\mathbb{R}^d)$. Compute the corresponding H^{-1}-norm.

Problem 2.6. (a) Find a variational formulation which amounts to solving

$$-\Delta u = f \quad \text{in } \Omega, \qquad \partial_\nu u + pu = g \quad \text{on } \partial\Omega,$$

where $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$, $0 < p_1 \leq p \leq p_2$ on $\partial\Omega$. Show that the bilinear form is coercive in $H^1(\Omega)$.

(b) Suppose that $p = \varepsilon^{-1} \to \infty$ and denote the corresponding solution by u_ε. Determine the boundary value problem satisfied by $u_0 = \lim_{\varepsilon \downarrow 0} u_\varepsilon$.

(c) Derive an error estimate for $\|u_0 - u_\varepsilon\|_{H^1(\Omega)}$.

Problem 2.7. Let $\boldsymbol{A}$ be uniformly SPD and $c \in L^\infty(\Omega)$ satisfy $c \geq 0$. Consider the quadratic functional

$$I[v] = \frac{1}{2}\int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla v + c(x)v^2 - \langle f, v\rangle \quad \text{for all } v \in H_0^1(\Omega),$$

where $f \in H^{-1}(\Omega)$. Show that $u \in H^1_0(\Omega)$ is a minimizer of $I[v]$ if and only if u satisfies the Euler-Lagrange equation

$$\mathscr{B}[u, v] = \int_\Omega \nabla v \cdot \boldsymbol{A} \nabla u + c\, uv = \langle f, v \rangle \quad \text{for all } v \in H^1_0(\Omega).$$

Problem 2.8. Consider the model problem with Neumann boundary condition

$$-\operatorname{div}(\boldsymbol{A}\nabla u) = f \quad \text{in } \Omega, \qquad \boldsymbol{n} \cdot \boldsymbol{A} \nabla u = g \quad \text{on } \partial\Omega$$

(a) Derive the variational formulation in $\mathbb{V} = H^1(\Omega)$ and show that the bilinear form $\mathscr{B}$ is continuous and symmetric but not coercive.
(b) Let $\mathbb{V}$ be the subspace of $H^1(\Omega)$ of functions with vanishing mean value. Show that $\mathscr{B}$ is coercive.
(c) Derive a compatibility condition between f and g for existence of a weak solution.

Problem 2.9. Consider the space $\mathbb{V} = H(\operatorname{div}; \Omega) = \{\boldsymbol{q} \in L^2(\Omega; \mathbb{R}^d) \,|\, \operatorname{div} \boldsymbol{q} \in L^2(\Omega)\}$, and the bilinear form

$$\mathscr{B}[\boldsymbol{p}, \boldsymbol{q}] = \int_\Omega \operatorname{div} \boldsymbol{p} \operatorname{div} \boldsymbol{q} + \boldsymbol{p} \cdot \boldsymbol{q} \qquad \text{for all } \boldsymbol{p}, \boldsymbol{q} \in \mathbb{V}.$$

(a) Show that $\mathbb{V}$ is a Hilbert space and that $\mathscr{B}$ is symmetric, continuous and coercive in $H(\operatorname{div}; \Omega)$.
(b) Determine the strong form of the PDE and implicit boundary condition corresponding to the variational formulation

$$\boldsymbol{p} \in \mathbb{V}: \qquad \mathscr{B}[\boldsymbol{p}, \boldsymbol{q}] = \langle \mathbf{f}, \boldsymbol{q} \rangle \qquad \text{for all } \boldsymbol{q} \in \mathbb{V}.$$

Problem 2.10. Let $\boldsymbol{\sigma} := -\boldsymbol{A}\nabla u$ be the flux of the model problem, which can be written equivalently as

$$\boldsymbol{A}^{-1}\boldsymbol{\sigma} + \nabla u = 0, \qquad \operatorname{div} \boldsymbol{\sigma} = -f.$$

(a) Let $\mathbb{V} = H(\operatorname{div}; \Omega)$ and $\mathbb{Q} = L^2_0(\Omega)$. Multiply the first equation by $\boldsymbol{\tau} \in \mathbb{V}$ and integrate by parts using Green's formula (9). Multiply the second equation by $v \in \mathbb{Q}$. Write the resulting variational formulation in the form (27) and show that (29) is satisfied.
(b) Apply Theorem 2.3 to deduce existence, uniqueness, and stability.

3 The Petrov-Galerkin method and finite element bases

The numerical approximation of boundary value problems is typically an effective way, and often the only one available, to extract quantitive information about their solutions. In this chapter we introduce the finite element method (FEM) which,

due to its geometric flexibility, practical implementation, and powerful and elegant theory, is one of the most successful discretization methods for this task.

Roughly speaking, a finite element method consists in computing the Petrov-Galerkin solution with respect to a finite-dimensional space and that space is constructed from local function spaces (finite elements), which are glued together by some continuity condition.

We first analyze Petrov-Galerkin approximations and then review Lagrange elements, the most basic and common finite element spaces; for other finite element spaces, we refer to the standard finite element literature, e.g. [15, 16, 18, 25, 51].

3.1 Petrov-Galerkin solutions

The solution of a boundary value problem cannot be computed, since the solution is characterized by an infinite number of (linearly-independent) conditions. To overcome this principal obstacle, we replace the boundary value problem by its Petrov-Galerkin discretization.

3.1.1 Definition, Existence and Uniqueness

To obtain a computable approximation to a solution to the variational problem (10) we simply restrict the continuous spaces $\mathbb{V}, \mathbb{W}$ in (10) to finite dimensional subspaces of equal dimension $N < \infty$. As we shall see, this leads to a linear system in $\mathbb{R}^{N\times N}$ which can be solved by standard methods.

Definition 3.1 (Discrete Solution). For $N \in \mathbb{N}$ let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension N. Then a solution U_N to

$$U_N \in \mathbb{V}_N : \qquad \mathscr{B}[U_N, W] = \langle f, W\rangle \qquad \text{for all } W \in \mathbb{W}_N \tag{37}$$

is called *Petrov-Galerkin Solution*.

Remark 3.1. For $\mathbb{V} \neq \mathbb{W}$ the test functions $W \in \mathbb{W}_N$ in (37) are different from the ansatz functions $V \in \mathbb{V}_N$ which results in the naming *Petrov-Galerkin discretization*. If the continuous spaces $\mathbb{V} = \mathbb{W}$ are equal, then we will choose also the same discrete space $\mathbb{V}_N = \mathbb{W}_N$. In this case, (37) is called *Galerkin discretization* and, if additionally $\mathscr{B}$ is symmetric and coercive, it is called *Ritz-Galerkin discretization*. In any case, the discrete spaces are subsets of the continuous ones, and thus all discrete functions belong to the continuous function spaces. For this reason, the method is called a *conforming discretization* of (10).

For any conforming discretization, the bilinear form $\mathscr{B}$ is well defined and continuous on the discrete pair $\mathbb{V}_N \times \mathbb{W}_N$. The continuity constant is bounded by $\|\mathscr{B}\|$. This can easily be seen, since all discrete functions $V \in \mathbb{V}_N$ and $W \in \mathbb{W}_N$ are admissible in (16). In the same vain, for a coercive form $\mathscr{B}\colon \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ we are allowed

to use any discrete function $V \in \mathbb{V}_N$ in (25) yielding

$$\mathscr{B}[V,V] \geq c_{\mathscr{B}} \|V\|_{\mathbb{V}}^2 \qquad \text{for all } V \in \mathbb{V}_N.$$

Therefore coercivity of $\mathscr{B}$ is inherited for conforming discretizations from the continuous space to the discrete one with the same coercivity constant $c_{\mathscr{B}} > 0$. This in turn implies the existence and uniqueness of the Galerkin solution $U_N \in \mathbb{V}_N$.

Recalling the theorem of Lax-Milgram, stated as Corollary 2.2, we know that a coercive form $\mathscr{B}$ satisfies the inf-sup condition (23). Since coercivity is inherited to subspaces we can conclude in this case the discrete counterpart of (23), namely

$$\inf_{V \in \mathbb{V}_N} \sup_{W \in \mathbb{W}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}} \|W\|_{\mathbb{W}}} = \inf_{W \in \mathbb{W}_N} \sup_{V \in \mathbb{V}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}} \|W\|_{\mathbb{W}}} = \beta_N \tag{38}$$

with a constant $\beta_N \geq c_{\mathscr{B}}$.

For general $\mathscr{B}$, the continuous inf-sup (23) does not imply the discrete one. In order to state a simple as possible criterion for the existence and uniqueness or a discrete solution, we consider the discrete operators $B_N \in L(\mathbb{V}_N; \mathbb{W}_N^*)$ and $B_N^* \in L(\mathbb{W}_N; \mathbb{V}_N^*)$, defined in the same way as B and B^* in Sect. 2.3 by

$$\langle B_N V, W \rangle = \langle B_N^* W, V \rangle = \mathscr{B}[V,W] \qquad \text{for all } V \in \mathbb{V}_N, W \in \mathbb{W}_N.$$

The discrete problem (37) is well-posed if and only if the operator B_N is an isomorphism from $\mathbb{V}_N$ to $\mathbb{W}_N^*$. Since we deal with finite dimensional spaces, a necessary condition for B_N being invertible is $\dim \mathbb{V}_N = \dim \mathbb{W}_N^* = \dim \mathbb{W}_N$, which we assume in the definition of the Petrov-Galerkin solution. Hence a neccessary and sufficient condition for invertibility of B_N is injectivity of B_N, which can be characterized by

$$\textit{for every } 0 \neq V \in \mathbb{V}_N \textit{ there exists } W \in \mathbb{W}_N \textit{ such that } \mathscr{B}[V,W] \neq 0. \tag{39}$$

As a direct consequence we can characterize the existence and uniqueness of the discrete solution.

Theorem 3.1 (Existence and Uniqueness of the Petrov-Galerkin Solution). *Let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension.*

Then for any $f \in \mathbb{W}_N^*$ there exists a unique Petrov-Galerkin solution $U_N \in \mathbb{V}_N$, i. e.,

$$U_N \in \mathbb{V}_N : \qquad \mathscr{B}[U_N, W] = \langle f, W \rangle \qquad \textit{for all } W \in \mathbb{W}_N,$$

if and only if (39) is satisfied.

Proof. As for the continuous problem (10) the existence and uniqueness of a discrete solution U_N for any $f \in \mathbb{W}_N^*$ is equivalent to the invertibility of the operator $B_N \colon \mathbb{V}_N \to \mathbb{W}_N^*$. The latter is equivalent to (39). □

Proposition 3.1. *Let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension.*

Then the following statements are equivalent:

(1) *The discrete inf-sup condition (38) holds for some $\beta_N > 0$;*

(2) $\inf_{V\in\mathbb{V}_N}\sup_{W\in\mathbb{W}_N}\dfrac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}}>0;$

(3) $\inf_{W\in\mathbb{W}_N}\sup_{V\in\mathbb{V}_N}\dfrac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}}>0;$

(4) *condition* (39) *is satisfied;*

(5) *for every* $0\neq W\in\mathbb{W}_N$ *there exists* $V\in\mathbb{V}_N$ *such that* $\mathscr{B}[V,W]\neq 0$.

Proof. Obviously, (1) implies (2) and (3). The inf-sup condition (2) implies (4) and (3) yields (5). Statement (4) is equivalent to invertibility of $B_N\in L(\mathbb{V}_N,\mathbb{W}_N^*)$ and in the same way (5) is equivalent to invertibility of $B_N^*\in L(\mathbb{W}_N,\mathbb{V}_N^*)$, whence (4) and (5) are equivalent. Recalling Theorem 3.1, statement (4) is equivalent to existence and uniqueness of a discrete solution for any $f\in\mathbb{W}_N^*$. Applying Theorem 2.2 with $\mathbb{V},\mathbb{W}$ replaced by $\mathbb{V}_N,\mathbb{W}_N$ the latter is equivalent with the inf-sup condition on $\mathbb{V}_N,\mathbb{W}_N$, i. e., (4) is equivalent to (1). □

This proposition allows for different conditions that imply existence and uniqueness of a discrete solution. Conditions (2)–(5) of Proposition 3.1 seem to be more convenient than (1) since we do not have to specify the discrete inf-sup constant β_N. However, the *value* of this constant is critical, as we shall see from the following section.

3.1.2 Stability and Quasi-Best Approximation

In this section we investigate the stability and approximation properties of Petrov-Galerkin solutions. In doing so, we explore properties that are uniform in the dimension N of the discrete spaces.

We start with the stability properties.

Corollary 3.1 (Stability of the Discrete Solution). *If* (38) *holds, then the Petrov-Galerkin solution* U_N *satisfies*

$$\|U_N\|_{\mathbb{V}}\leq\frac{1}{\beta_N}\|f\|_{\mathbb{W}^*}. \tag{40}$$

Proof. Use the same arguments as in the proof of Theorem 2.2 for the stability estimate of the true solution. □

We next relate the Petrov-Galerkin solution to the best possible approximation to the true solution u in $\mathbb{V}_N$ and show that U_N is up to a constant as close to u as the best approximation. For coercive forms this is Cea's Lemma [22]. For general $\mathscr{B}$ this follows from the theories of Babuška [8, 9] and Brezzi [17].

The key for the best approximation property of the Petrov-Galerkin solution is the following relationship, which holds for all conforming discretizations and is usually referred to as *Galerkin orthogonality*:

$$\mathscr{B}[u-U_N,W]=0\qquad\text{for all } W\in\mathbb{W}_N. \tag{41}$$

If $\mathbb{V} = \mathbb{W}$, $\mathscr{B}$ symmetric and coercive, then this means that the error $u - U_N$ is orthogonal to $\mathbb{V}_N = \mathbb{W}_N$ in the energy norm $|||\cdot|||_\Omega$. To prove (41), simply observe that we are allowed to use any $W \in \mathbb{W}_N$ as a test function in the definition of the continuous solution (10), which gives

$$\mathscr{B}[u, W] = \langle f, W \rangle \qquad \text{for all } W \in \mathbb{W}_N.$$

Then recalling the definition of the Petrov-Galerkin solution and taking the difference yields (41).

Theorem 3.2 (Quasi-Best-Approximation Property). *Let $\mathscr{B}\colon \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ be continuous and assume (38) is satisfied. Let u be the solution to (10) and let $U_N \in \mathbb{V}_N$ be the Petrov-Galerkin solution.*

Then the error $u - U_N$ satisfies the bound

$$\|u - U_N\|_{\mathbb{V}} \leq \frac{\|\mathscr{B}\|}{\beta_N} \min_{V \in \mathbb{V}_N} \|u - V\|_{\mathbb{V}}.$$

Proof. We give a simplified proof, which follows Babuška [8, 9] and yields the constant $1 + \frac{\|\mathscr{B}\|}{\beta_N}$. The asserted constant is due to Xu and Zikatanov [79].

Combining (38), (41), and the continuity of $\mathscr{B}$, we derive for all $V \in \mathbb{V}_N$

$$\beta_N \|U_N - V\|_{\mathbb{V}} \leq \sup_{W \in \mathbb{W}_N} \frac{\mathscr{B}[U_N - V, W]}{\|W\|_{\mathbb{W}}} = \sup_{W \in \mathbb{W}_N} \frac{\mathscr{B}[u - V, W]}{\|W\|_{\mathbb{W}}} \leq \|\mathscr{B}\| \|u - V\|_{\mathbb{V}},$$

whence

$$\|U_N - V\|_{\mathbb{V}} \leq \frac{\|\mathscr{B}\|}{\beta_N} \|u - V\|_{\mathbb{V}}.$$

Using the triangle inequality yields

$$\|u - U_N\|_{\mathbb{V}} \leq \|u - V\|_{\mathbb{V}} + \|V - U_N\|_{\mathbb{V}} \leq \left(1 + \frac{\|\mathscr{B}\|}{\beta_N}\right) \|u - V\|_{\mathbb{V}}$$

for all $V \in \mathbb{V}_N$. It just remains to minimize in $\mathbb{V}_N$. □

The last two results reveal the critical role of the discrete inf-sup constant β_N. If a sequence of spaces $\{(\mathbb{V}_N, \mathbb{W}_N)\}_{N \geq 1}$ approximates the pair $(\mathbb{V}, \mathbb{W})$ with deteriorating $\beta_N \to 0$ as $N \to \infty$, then the sequence of discrete solutions $\{U_N\}_{N \geq 1}$ is not guaranteed to be uniformly bounded. Furthermore, the discrete solutions in general approximate the true solution with a reduce rate as compared to the best approximation within $\mathbb{V}_N$. For these reasons a lower bound for the discrete inf-sup constants becomes highly desirable.

Definition 3.2 (Stable Discretization). We call a sequence $\{(\mathbb{V}_N, \mathbb{W}_N)\}_{N \geq 1}$ of discrete spaces with inf-sup constants $\{\beta_N\}_{N \geq 1}$ stable if and only if there exists $\beta > 0$ such that

$$\inf_{N \geq 1} \beta_N \geq \beta > 0.$$

In contrast to the continuous inf-sup condition where one has to prove

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}>0 \qquad \textit{and} \qquad \inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}>0$$

it suffices in the discrete setting to show one

$$\inf_{V\in\mathbb{V}_N}\sup_{W\in\mathbb{W}_N}\frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}}\geq\beta \qquad \textit{or} \qquad \inf_{W\in\mathbb{W}_N}\sup_{V\in\mathbb{V}_N}\frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}}\geq\beta$$

in order to furnish a uniform lower bound for the discrete inf-sup constant $\beta_N\geq\beta$. This simplification stems from the assumption $\dim\mathbb{V}_N=\dim\mathbb{W}_N<\infty$. Allowing for infinite dimensional spaces $\mathbb{V}_N$ and $\mathbb{W}_N$ gives rise to both inf-sup conditions as in the continuous case.

3.1.3 Computation

In view of Theorem 3.2, the *quality* of the Petrov-Galerkin solution depends in particular on the approximation properties of the discrete spaces. Before embarking on the construction of suitable spaces, it is useful to see how a Petrov-Galerkin solution can be computed. This will reveal that the real task is the construction of a suitable basis and it will give hints towards what affects the *cost* of a Petrov-Galerkin solution.

Let $\phi_1,\dots,\phi_N$ and $\psi_1,\dots,\psi_N$ be bases of $\mathbb{V}_N$ and $\mathbb{W}_N$, respectively. Writing

$$\begin{aligned} U_N&=\sum_{j=1}^{N}\alpha_j\phi_j,\\ K=(k_{ij})_{i,j=1,\dots,N}\quad&\text{with}\quad k_{ij}=\mathscr{B}[\phi_j,\psi_i],\\ F=(F_1,\dots,F_N)\quad&\text{with}\quad F_i=\langle f,\psi_i\rangle \end{aligned}$$

the definition of the Petrov-Galerkin solution (37) is equivalent to the linear system

$$\alpha\in\mathbb{R}^N: \qquad K\alpha=F. \tag{42}$$

Its solution can be computed by various methods from numerical linear algebra. The method of choice as well as the cost is affected by the properties of the matrix. Of course these properties depend on the bilinear form $\mathscr{B}[\cdot,\cdot]$ and on the chosen bases $\phi_1,\dots,\phi_N$ and $\psi_1,\dots,\psi_N$.

For example, in the case of the model problem of Sect. 2.2.1, $\mathbb{V}_N=\mathbb{W}_N$ and $\phi_i=\psi_i$ for $i=1,\dots,N$, the matrix K is symmetric positive definite, irrespective of the choice of $\phi_1,\dots,\phi_N$. The linear system (42) gets trivial if we take $\phi_1,\dots,\phi_N$ to be the eigenvectors of K. However, finding the eigenvectors of K is a nonlinear problem and typically more expensive than solving linear systems. On the other hand, taking

the easily available polynomials for $\phi_1, \ldots, \phi_N$ will lead to full and ill-conditioned matrices in general.

Finite element bases provide a compromise between these two extremes. The basis functions can be relatively easily constructed and are locally supported. The latter leads to sparse matrices for bilinear forms associated with boundary value problems.

3.2 Finite element spaces

The choice, or better the construction, of suitable finite element spaces in the Petrov-Galerkin discretization is the subject of this section. We shall discuss here only the spaces of Lagrange elements, emphasizing the case of polynomial degree $n = 1$. These spaces are appropriate for our model problem of Sect. 2.2.1.

3.2.1 Simplices and Triangulations

As already mentioned, a key property of finite element bases is that there are locally supported. This is achieved with the help of a decomposition of the domain of the boundary value problem. Here we consider triangulations, which are build from simplices.

Definition 3.3 (Simplex and Subsimplices). Let $d \in \mathbb{N}$. A subset T of $\mathbb{R}^d$ is an *n-simplex in* $\mathbb{R}^d$ if there exist $n+1$ points $z_0, \ldots, z_n \in \mathbb{R}^d$ such that

$$T = \text{conv hull}\{z_0, \ldots, z_n\} = \left\{ \sum_{i=1}^{n} \lambda_i z_i \mid \lambda_i \geq 0 \text{ for } i = 0, \ldots, d, \sum_{i=0}^{n} \lambda_i = 1 \right\}$$

and $z_1 - z_0, \ldots, z_n - z_0$ are linearly independent vectors in $\mathbb{R}^d$. By convention, we refer to points as 0-simplices. A subset T' of T is a (proper) *k-subsimplex* of T if T' is a k-simplex such that

$$T' = \text{conv hull}\{z'_0, \ldots, z'_k\} \subset \partial T$$

with $k < n$ and $z'_0, \ldots, z'_k \in \{z_0, \ldots, z_d\}$.

The 0-subsimplices are the vertices of a simplex. Moreover, 1-subsimplices are edges and 2-subsimplices of 3-simplices are faces. We shall refer to $(n-1)$-subsimplices of n-simplices as sides.

Two d-simplices in $\mathbb{R}^d$ are always affine equivalent, meaning that one can be mapped onto the other by an affine bijection. This fact is useful for implementation and also for the theory that follows. The following lemma fixes a reference simplex and controls the affine bijection in terms of geometric quantities of the generic simplex.

Lemma 3.1 (Reference and Generic Simplex). *Let the* reference simplex *in* $\mathbb{R}^d$ *be defined as*

$$\hat{T} = \text{conv hull}\{0, e_1, \dots, e_d\},$$

where $e_1, \dots, e_d$ *denotes the canonical basis in* $\mathbb{R}^d$*. For any* d*-simplex* T *in* $\mathbb{R}^d$*, there exists a bijective affine map*

$$F_T : \hat{T} \to T, \quad \hat{x} \mapsto A_T \hat{x} + b_T$$

where $A_T \in \mathbb{R}^{d\times d}$ *and* $b_T \in \mathbb{R}^d$*. If we define*

$$\begin{aligned} \overline{h}_T &:= \sup\{|x-y| \mid x, y \in T\}, \\ \underline{h}_T &:= \sup\{2r \mid B_r \subset T \text{ is a ball of radius } r\}, \\ h_T &:= |T|^{1/d}, \end{aligned}$$

there holds

$$\|A_T\| \le \overline{h}_T, \qquad \|A_T^{-1}\| \le \frac{C_d}{\underline{h}_T}, \qquad |\det A_T| = \frac{h_T^d}{d!}. \tag{43}$$

Proof. See Problem 3.4.

All three quantities in (43) measure somehow the size of the given simplex. In view of

$$\underline{h}_T \le h_T \le \overline{h}_T$$

they are equivalent up to the following quantity.

Definition 3.4 (Shape Coefficient). The *shape coefficient* of a d-simplex T in $\mathbb{R}^d$ is the ratio of the diameter and the inball diameter of T,

$$\sigma_T := \frac{\overline{h}_T}{\underline{h}_T}.$$

Of course this notion becomes useful when it refers to many simplices. This brings us to the notion of triangulation.

Definition 3.5 (Triangulation). Let $\Omega \subset \mathbb{R}^d$ be a bounded, polyhedral domain. A finite set $\mathcal{T}$ of d-simplices in $\mathbb{R}^d$ with

$$\overline{\Omega} = \bigcup_{T\in\mathcal{T}} T \qquad \text{and} \qquad |\Omega| = \sum_{T\in\mathcal{T}} |T| \tag{44}$$

is a *triangulation of* Ω. We denote the set of all vertices of $\mathcal{T}$ by $\mathcal{V}_\mathcal{T}$ and the set of all sides by $\mathcal{S}_\mathcal{T}$. The *shape coefficient* of a triangulation $\mathcal{T}$ is the quantity $\sigma_\mathcal{T} := \max_{T\in\mathcal{T}} \sigma_T$. A triangulation $\mathcal{T}$ is *conforming* if it satisfies the following property: if any two simplices $T_1, T_2 \in \mathcal{T}$ have a nonempty intersection $S = T_1 \cap T_2 \neq \emptyset$, then S is a k-subsimplex of both T_1 and T_2 with $k \in \{0, \dots, d\}$.

A sequence of triangulations $\{\mathcal{T}_k\}_{k\ge 0}$ is *shape regular* if $\sup_{T\in\mathcal{T}_k} \sigma_\mathcal{T} \le C$. It is called *quasi-uniform* if there exists a constant C such that, for all k, there holds

$\max_{T\in\mathscr{T}_k}\overline{h}_T \le C\min_{T\in\mathscr{T}_k}\underline{h}_T$. In both cases we assume tacitly that the constant C is of moderate size.

The first condition in (44) ensures that $\mathscr{T}$ is a covering of the closure of Ω, while the second requires that there is no overlapping. Notice that the latter is required not in a set-theoretic but in a measure-theoretic manner. Conformity will turn out to be a very useful property when constructing bases that are regular across simplex boundaries.

3.2.2 Lagrange Elements

The purpose of this section is to show that the following finite-dimensional space is appropriate for our model problem in Sect. 2.2.1:

$$\mathbb{V}(\mathscr{T}) := \{v \in C(\overline{\Omega}) \mid v_{|T} \in \mathbb{P}_n(T) \quad \text{for all } T \in \mathscr{T} \ \text{ and } \ v_{|\partial\Omega} = 0\}$$

where $\mathscr{T}$ is a conforming triangulation of $\Omega \subset \mathbb{R}^d$ and $\mathbb{P}_n(T)$ stands for the space of polynomials with degree $\le n$ over T. More precisely, we will show that $\mathbb{V}(\mathscr{T}) \subset H^1_0(\Omega)$ possesses a basis which is locally supported and easy to implement, and conclude with approximation properties of $\mathbb{V}(\mathscr{T})$. In what follows, this will be called the *standard discretization* of the model problem.

Lemma 3.2 (H^1_0-Conformity). *If $\mathscr{T}$ is a conforming triangulation of a bounded, polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^d$, then $\mathbb{V}(\mathscr{T}) \subset H^1_0(\Omega)$.*

Proof. Let $v \in \mathbb{V}(\mathscr{T})$. We start by checking that v has a weak derivative. For any test function $\eta \in C^\infty_0(\Omega)$ and $i \in \{1,\dots,d\}$ there holds

$$\int_\Omega v\partial_i\eta = \sum_{T\in\mathscr{T}}\int_T v\partial_i\eta = \sum_{T\in\mathscr{T}}\int_T(\partial_i v)\eta + \sum_{T\in\mathscr{T}}\sum_{S\subset\partial T}\int_S v\eta n_{T,i},$$

where $n_{T,i}$ is the i-th coordinate of the exterior normal to ∂T. The second sum on the right hand side vansishes for the following reasons: if $S \subset \partial\Omega$, then there holds $\eta_{|S} = 0$; otherwise there exists a unique simplex $T' \in \mathscr{T}$ such that $S = T \cap T'$ and $n_{T',i} = -n_{T,i}$. Consequently, $w \in L^\infty(\Omega)$ given by $w_{|T} = \partial_i v_{|T}$ for all $T \in \mathscr{T}$ is the i-th weak derivate of v. In particular, we have $v \in H^1(\Omega)$. In view of the characterization

$$H^1_0(\Omega) = \{v \in H^1(\Omega) \mid v_{|\partial\Omega} = 0\}$$

and the definition of $\mathbb{V}(\mathscr{T})$, we conclude that $v \in H^1_0(\Omega)$. □

Next, we construct a suitable basis of

$$S^{n,0}(\mathscr{T}) := \{v \in C(\overline{\Omega}) \mid \forall T \in \mathscr{T}\ v_{|T} \in \mathbb{P}_n(T)\},$$

which yields immediately one for $\mathbb{V}(\mathscr{T})$. We first consider the case $n = 1$ and, in view of the piecewise structure, start with the following result on $\mathbb{P}_1(T)$.

Lemma 3.3 (Local $\mathbb{P}_1$-Basis). *Let $T = \text{conv hull}\,\{z_0,\dots,z_d\}$ be a d-simplex in $\mathbb{R}^d$. The barycentric coordinates $\lambda_0,\dots,\lambda_d : T \to \mathbb{R}$ on T defined by*

$$T \ni x = \sum_{i=0}^{d} \lambda_i(x) z_i, \quad \text{and} \quad \sum_{i=0}^{d} \lambda_i(x) = 1, \tag{45}$$

are a basis of $\mathbb{P}_1(T)$ such that

$$\lambda_i(z_j) = \delta_{ij} \qquad \text{for all } i,j \in \{0,\dots,d\}. \tag{46}$$

For each $p \in \mathbb{P}_1(T)$, there holds the representation formula

$$p = \sum_{i=0}^{d} p(z_i)\lambda_i. \tag{47}$$

Proof. We first check that the barycentric coordinates $\lambda_0,\dots,\lambda_d$ are well-defined. To this end, fix $x \in T$ for a moment and observe that (45) for $\lambda_i = \lambda_i(x)$ can be rewritten as

$$\begin{bmatrix} | & & | \\ z_0 & \cdots & z_d \\ | & & | \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_d \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ 1 \end{bmatrix}.$$

If we choose F_T in Lemma 3.1 such that $F_T(0) = z_0$, $F_T(e_i) = F(z_i)$ for $i = 1,\dots,d$, we easily see that the above matrix has the same determinant as A_T, which is different from 0.

Consequently, the functions $\lambda_0,\dots,\lambda_d$ are well-defined and, varying x, we see that $\lambda_i \in \mathbb{P}_1(T)$ for $i = 0,\dots,d$. Property (46) is now readily verified and ensures that the $(d+1)$ functions $\lambda_0,\dots,\lambda_d$ are linearly independent. From the definition of $\mathbb{P}_1(T)$ it is immediate that $\dim \mathbb{P}_1(T) = d+1$, whence $\lambda_0,\dots,\lambda_d$ has to be a basis. Writing $p = \sum_{i=0}^{d} \alpha_i \lambda_i$ for $p \in \mathbb{P}_1(T)$, and using (46), yields (47) and finishes the proof. □

Property (46) means that $\lambda_0,\dots,\lambda_d$ is the basis in $\mathbb{P}_1(T)$ that is dual to the basis $\mathcal{N}_1(T) = \{N_1,\dots,N_d\}$ of $\mathbb{P}_1(T)^*$ given by $p \mapsto p(z_i)$ for $i = 0,\dots,d$. By the Riesz representation theorem in $L^2(T)$, we can associate a function $\lambda_i^* \in \mathbb{P}_1(T)$ to each functional N_i such that

$$\int_T \lambda_i \lambda_j^* = \delta_{ij} \qquad \text{for all } i,j \in \{0,\dots,d\}. \tag{48}$$

A simple computation using [25, Exercise 4.1.1] reveals that

$$\lambda_i^* = \frac{(1+d)^2}{|T|}\lambda_i - \frac{1+d}{|T|}\sum_{j\neq i}\lambda_j \qquad \text{for all } i \in \{1,\dots,d\}.$$

Since $\mathcal{N}_1(T)$ is a basis of $\mathbb{P}_1(T)^*$, the triple

$$(T, \mathbb{P}_1(T), \mathcal{N}_1(T))$$

is a finite element; for the definition of a finite element see, e.g., [16, Ch. 3]. The elements of $\mathcal{N}_1(T)$ are its *nodal variables* and $\lambda_0, \dots, \lambda_d$ its *nodal basis*.

Theorem 3.3 (Courant Basis). *A function $v \in S^{1,0}(\mathcal{T})$ is characterized by its values at the nodes $\mathcal{N}_1(\mathcal{T}) := \mathcal{V}_{\mathcal{T}}$. The functions ϕ_z, $z \in \mathcal{N}_1(\mathcal{T})$, defined by*

$$\phi_z \in S^{1,0}(\mathcal{T}) \qquad \text{and} \qquad \phi_z(y) = \delta_{yz} \qquad \text{for all } y \in \mathcal{N}_1(\mathcal{T})$$

are a basis of $S^{1,0}(\mathcal{T})$ such that, for every $v \in S^{1,0}(\mathcal{T})$,

$$v = \sum_{z \in \mathcal{N}_1(\mathcal{T})} v(z)\phi_z.$$

In particular, $\{\phi_z\}_{z \in \mathcal{N}_1(\mathcal{T}) \cap \Omega}$ is a basis of $S^{1,0}(\Omega) \cap H^1_0(\Omega)$.

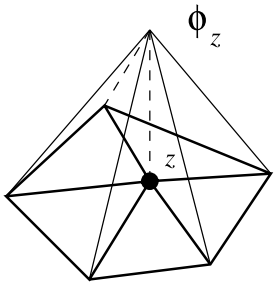


Fig. 1 Courant basis function ϕ_z for an interior vertex $z \in \mathcal{N}_1(\mathcal{T})$

Proof. Let $T_1, T_2 \in \mathcal{T}$ be two distinct simplices such that $T_1 \cap T_2 \neq \emptyset$; then $S := T_1 \cap T_2$ is a k-subsimplex with $0 \le k < d$ because $\mathcal{T}$ is conforming. Let $w_i \in \mathbb{P}_1(T_i)$, for $i = 1, 2$, be two affine functions with the same nodal values $w_1(z) = w_2(z)$ at all vertices $z \in S$. We assert that $w_1 = w_2$ on S. Since this is obvious for $k = 0$, we consider $k > 0$, recall that S is isomorphic to the reference simplex $\hat{T}_k$ in $\mathbb{R}^k$ and apply Lemma 3.3 to deduce $w_1 = w_2$ on S. This shows that any continuous piecewise affine function $v \in S^{1,0}(\mathcal{T})$ can be built by pasting together local affine functions with the restriction of having the same nodal values, or equivalently to coincide at all vertices $z \in \mathcal{N}_1(\mathcal{T})$. Moreover, v is characterized by its nodal values $\{v(z)\}_{z \in \mathcal{N}_1(\mathcal{T})}$.

Therefore, the functions ϕ_z are well-defined for all $z \in \mathcal{N}_1(\mathcal{T})$. In addition, for all $v \in S^{1,0}(\mathcal{T})$ the function $\sum_{z \in \mathcal{N}_1(\mathcal{T})} v(z)\phi_z$ equals v at the nodes, whence they coincide everywhere and $S^{1,0}(\mathcal{T}) = \operatorname{span}\{\phi_z\}_{z \in \mathcal{N}_1(\mathcal{T})}$. Since $\{\phi_z\}_{z \in \mathcal{N}_1(\mathcal{T})}$ are linearly independent, they form a basis of $S^{1,0}(\mathcal{T})$.

Finally, to prove that $\{\phi_z\}_{z \in \mathcal{N}_1(\mathcal{T}) \cap \Omega}$ is a basis of $S^{1,0}(\Omega) \cap H^1_0(\Omega)$ we observe that if $v \in S^{1,0}(\Omega)$ vanishes at the vertices of a side $S \in \mathcal{S}$ contained in $\partial\Omega$ then

v vanishes in S, again as a consequence of Lemma 3.3. Therefore, $v \in S^{1,0}(\Omega) \cap H^1_0(\Omega)$ if and only if the nodal values $v(z) = 0$ for all $z \in \mathcal{N}_1(\mathcal{T}) \cap \partial\Omega$. □

Remark 3.2 (Representation of Courant Basis). The proof of Theorem 3.3 shows that the global basis functions are given in terms of local basis functions. More precisely, if λ_z^T denotes the barycentric coordinate of $T \in \mathcal{T}$ associated with the vertex $z \in T$, there holds

$$\phi_z = \begin{cases} \lambda_z^T & \text{if } z \in T, \\ 0 & \text{otherwise} \end{cases}$$

for any node $z \in \mathcal{N}$.

We thus now have a basis of $\mathbb{V}(\mathcal{T}) = S^{1,0}(\mathcal{T}) \cap H^1_0(\Omega)$ that can be implemented relatively easily. Its basis functions are locally supported and the corresponding matrix in (42) is sparse in the case of our model problem in Sect. 2.2.1; see Problem 3.7.

Remark 3.3 (Dual of Courant Basis). Let $v_z \in \mathbb{N}$ be the valence of z for each node $z \in \mathcal{N}_1(\mathcal{T})$, namely the number of elements $T \in \mathcal{T}$ containing z as a vertex. The *discontinuous* piecewise linear functions $\phi_z^* \in S^{1,-1}(\mathcal{T})$ defined by

$$\phi_z^* = \frac{1}{v_z} \sum_{T \ni z} (\lambda_z^T)^* \, \chi_T \qquad \text{for all } z \in \mathcal{N}_1(\mathcal{T}), \tag{49}$$

with χ_T being the characteristic function of T, are (global) dual functions to the Courant basis $\{\phi_z\}_{z \in \mathcal{N}}$ in that they satisfy

$$\int_\Omega \phi_z \phi_y^* = \delta_{yz} \qquad \text{for all } y, z \in \mathcal{N}_1(\mathcal{T}). \tag{50}$$

We briefly comment on the generalization to arbitray polynomial degree $n \in \mathbb{N}$. Given a d-simplex $T = \text{ conv hull } \{z_0, \dots, z_d\}$ and identifying nodal variables and nodes, we set

$$\mathcal{N}_n(T) := \left\{ z_\alpha = \sum_{i=0}^{d+1} \frac{\alpha_i}{n} z_i \mid \alpha \in \mathbb{N}_0^{d+1}, \sum_{i=0}^{d+1} \alpha_i = n \right\} \tag{51}$$

The number of elements in $\mathcal{N}_n(T)$ coincides with the number of coefficients of polynomial in $\mathbb{P}_n(T)$. This is necessary for the existence of the corresponding nodal basis. The construction, see e.g. [16, Chapt. 3], reveals that also the location of the nodes plays some role. The latter implies also that restricting $\mathcal{N}_n(T)$ to a k-subsimplex and transforming to $\hat{T}_k$ yields $\mathcal{N}_n(\hat{T}_k)$. Consequently, the following theorem can be proven in the same way as Theorem 3.3.

Theorem 3.4 (Lagrange Basis). *A function $v \in S^{n,0}(\mathcal{T})$ is characterized by its values at the nodes $\mathcal{N}_n(\mathcal{T}) := \cup_{T \in \mathcal{T}} \mathcal{N}_n(T)$. The functions ϕ_z, $z \in \mathcal{N}_n(\mathcal{T})$, defined by*

$$\phi_z \in S^{n,0}(\mathscr{T}) \qquad \text{and} \qquad \phi_z(y) = \delta_{yz} \qquad \text{for all } y \in \mathscr{N}_n(\mathscr{T})$$

are a basis of $S^{n,0}(\mathscr{T})$ *such that, for every* $v \in S^{n,0}(\mathscr{T})$,

$$v = \sum_{z \in \mathscr{N}_n(\mathscr{T})} v(z)\phi_z.$$

In particular, $(\phi_z)_{z \in \mathscr{N}_n(\mathscr{T}) \cap \Omega}$ *is a basis of* $S^{n,0}(\Omega) \cap H^1_0(\Omega)$.

Remark 3.4 (Dual of Lagrange Basis). The construction of local and global piecewise linear dual functions extends to any polynomial degree $n \geq 1$; see Problem 3.9 for $k = 2$. Consequently, there exist discontinuous functions $\phi_z^* \in S^{n,-1}(\mathscr{T})$ such that $\operatorname{supp} \phi_z^* = \operatorname{supp} \phi_z$ and

$$\int_\Omega \phi_z \phi_y^* = \delta_{yz} \qquad \text{for all } y, z \in \mathscr{N}_n(\mathscr{T}). \tag{52}$$

Remark 3.5 (Barycentric Coordinates). For linear finite elements the basis functions on a single element T are the barycentric coordinates on T. The barycentric coordinates play also an important role for higher degree. First we observe that any point $z_\alpha \in \mathscr{N}_n(T)$ is determined from the barycentric coordinates $\frac{1}{n}(\alpha_1, \ldots, \alpha_d)$. Secondly, using the $(d+1)$ barycentric coordinates as a local coordinate system on T is a rather convenient choice for the explicit construction of a local basis on T; compare with Problem 3.8 as well as [63, Sect. 1.4.1] for a more detailed description. This is one reason that local basis functions are defined in the finite element toolbox ALBERTA in terms of the barycentric coordinates [63, Sect. 3.5].

3.2.3 Looking Ahead

We close this section with a few comments about fundamental issues of finite elements that will be addressed later in this survey.

Mesh Construction. The formalism above relies on a conforming mesh $\mathscr{T}$. Its practical construction is a rather delicate matter, especially if it will be successively refined as part of an adaptive loop. We study mesh refinement by *bisection* in Chap. 4 in any dimension and assess the complexity of such process. This study involves basic geometry and graph theory as well as combinatorics.

Piecewise Polynomial Interpolation. As established in Theorem 3.2, the performance of the FEM hinges on the quality of piecewise polynomial approximation. We discuss this topic in Chap. 5, where we construct a *quasi interpolation* operator to approximate rough functions and introduce the concept of mesh optimality; Remark 3.4 will be crucial in this respect. We present an algorithm that builds quasi-optimal meshes by thresholding for a rather large class of rough functions. This hints at the potentials of FEM to approximate singular solutions.

A Posteriori Error Analysis. Thresholding assumes to have full access to the function in question, which is not realistic when dealing with PDE. The missing

item is the design of a posteriori error estimators that extract the desired information from the discrete solution rather than the exact one. We present *residual estimators* in Chap. 6 and discuss their basic properties. They are instrumental.

Adaptivity. The fact that we learn about the approximation quality via a posteriori error estimators rather than directly from the function being approximated makes the study of AFEM quite different from classical approximation theory. This interplay between discrete and continuum will permeate the subsequent discussion in Chap. 7–Chap. 9.

In this survey, particularly when studying a posteriori error estimators and adaptivity, we assume that we have the *exact* Petrov-Galerkin solution U at hand. In doing this we ignore two important aspects of a practical finite element method: numerical integration and inexact solution of the resulting linear system. We close this chapter with two remarks concerning these issues.

Remark 3.6 (Numerical Integration). In contrast to the a priori error analysis of quadrature [25, Chapter 4.1], its treatment within an a posteriori context is a delicate matter, especially if one is not willing to assume regularity a priori and accept asymptotic results as the mesh size goes to zero. This seems to be largely open.

Remark 3.7 (Multilevel Solvers). For a hierarchy of quasi-uniform meshes, V-cycle multigrid and BPX-preconditioned conjugate gradient methods can approximate the Ritz-Galerkin solution U of our model problem (13) to a desired accuracy with a number of operations proportional to $\#\mathscr{T}$ [15, 16]. This, however, entails some restrictions on the coefficient matrix $\boldsymbol{A}$. Much less is known for graded meshes such as those generated by an adaptive method. For graded bisection meshes, we quote the results of Wu and Chen [77] for the V-cycle multigrid for $d = 2, n = 1$, and the recent results of Chen et al. [23, 78] for multigrid methods and multilevel preconditioners for $d \geq 2, n \geq 1$: they both show linear complexity in terms of $\#\mathscr{T}$. The latter exploits the geometric properties of bisection grids explained in Chap. 4.

3.3 Problems

Problem 3.1. Prove *Cea's Lemma:* Let $\mathscr{B}\colon \mathbb{V} \times \mathbb{V}$ be a continuous and coercive form. Let u be the true solution and $U_N \in \mathbb{V}_N$ be the Galerkin solution. Then U_N is a quasi-best approximation to u in $\mathbb{V}_N$, i. e.,

$$\|u - U_N\|_{\mathbb{V}} \leq \frac{\|\mathscr{B}\|}{c_{\mathscr{B}}} \min_{V \in \mathbb{V}_N} \|u - V\|_{\mathbb{V}}.$$

If, in addition, $\mathscr{B}$ is symmetric, then U_N is the best approximation to u in $\mathbb{V}_N$ with respect to the energy norm $|||\cdot|||_{\Omega}$, i. e.,

$$|||u - U_N|||_{\Omega} = \min_{V \in \mathbb{V}_N} |||u - V|||_{\Omega}$$

and the error in the $\mathbb{V}$-norm can be estimated by

$$\|u-U_N\|_{\mathbb{V}} \le \sqrt{\frac{\|\mathscr{B}\|}{c_{\mathscr{B}}}} \min_{V\in\mathbb{V}_N} \|u-V\|_{\mathbb{V}}.$$

Problem 3.2. Let $\{\mathbb{V}_N, \mathbb{W}_N\}_{N\in\mathbb{N}}$ be a sequence of nested subspaces of $\mathbb{V}, \mathbb{W}$ of equal dimension N, i. e.,

$$\mathbb{V}_M \subset \mathbb{V}_N \qquad \text{and} \qquad \mathbb{W}_M \subset \mathbb{W}_N \qquad \text{for all } M \le N,$$

such that

$$\overline{\bigcup_{N\in\mathbb{N}} \mathbb{V}_N}^{\|\cdot\|_{\mathbb{V}}} = \mathbb{V} \qquad \text{and} \qquad \overline{\bigcup_{N\in\mathbb{N}} \mathbb{W}_N}^{\|\cdot\|_{\mathbb{W}}} = \mathbb{W}.$$

Suppose that, for every $f \in \mathbb{W}^*$, the sequence of discrete Petrov-Galerkin solutions $\{U_N\}_{\mathbb{N}}$ defined by

$$U_N \in \mathbb{V}_N : \qquad \mathscr{B}[U_N, W] = \langle f, W\rangle \qquad \text{for all } W \in \mathbb{W}_N$$

satisfies

$$\lim_{N\to\infty} \|u-U_N\|_{\mathbb{V}} = 0.$$

Show that there holds

$$\inf_{N\in\mathbb{N}} \inf_{V\in\mathbb{V}_N} \sup_{W\in\mathbb{W}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} > 0.$$

Problem 3.3. Verify that the matrix K in (42) is symmetric positive definite for the model problem of Sect. 2.2.1, $\mathbb{V}_N = \mathbb{W}_N$ and $\phi_i = \psi_i$ for $i = 1,\dots,N$, irrespective of the choice of $\phi_1,\dots,\phi_N$.

Problem 3.4. Prove Lemma 3.1. Start by expressing A_T and b_T in terms of the vertices of T.

Problem 3.5. Prove Lemma 3.2 for a not necessarily conforming triangulation.

Problem 3.6. Given a d-simplex $T =$ conv hull $\{z_0,\dots,z_d\}$ in $\mathbb{R}^d$, construct a basis $\bar\lambda_0,\dots,\bar\lambda_d$ of $\mathbb{P}_1(T)$ such that

$$\bar\lambda_i(\bar z_j) = \delta_{ij} \qquad \text{for all } i,j \in \{1,\dots,d\},$$

where $\bar z_j$ denotes the barycenter of the face opposite to the vertex z_j. Does this local basis also lead to a global one in $S^{1,0}(\mathscr{T})$?

Problem 3.7. Determine the support of a basis function ϕ_z, $z \in \mathscr{N}$, in Theorem 3.3. Show that, with this basis, the matrix K in (42) is sparse for the model problem in Sect. 2.2.1.

Problem 3.8. Express the nodal basis of $(T, \mathbb{P}_2(T), \mathscr{N}_2(T))$ in terms of barycentric coordinates.

Problem 3.9. Derive expressions for the dual functions of the quadratic local Lagrange basis of $P_2(T)$ for each element $T \in \mathcal{T}$. Construct a global discontinuous dual basis $\phi_z^* \in S^{2,-1}(\mathcal{T})$ of the global Lagrange basis $\phi_z \in S^{2,0}(\mathcal{T})$ for all $z \in \mathcal{N}_2(\mathcal{T})$.

4 Mesh refinement by bisection

In this section we discuss refinement of a given initial triangulation consisting of d simplices using bisection, i. e., any selected simplex is divided into two subelements of same size. Refinement by bisection in 2d can be traced back to Sewell in the early 1970s [66]. In the mid of the 1980s Rivara introduced the longest edge bisection [61] and Mitchell formulated a recursive algorithm for the newest vertex bisection [49, 50]. In the beginning of the 1990s Bänsch was the first to present a generalization of the newest vertex bisection to 3d [10]. A similar approach was published by Liu and Joe [46] and later on by Arnold et al. [2]. A recursive variant of the algorithm by Bänsch was derived by Kossaczký [44]. He formulated the bisection rule for tetrahedra using a local order of their vertices and their element type. This concept is very convenient for implementation. In addition, it can be generalized to any space dimension which was done independently by Maubach [47] and Traxler [72].

Asking for conformity of locally refined meshes has the unalterable consequence that refinement propagates, i. e., besides the selected elements additional simplices have to be refined in order to maintain conformity. Although practical experience clearly suggests that local refinement stays local, the first theoretical foundation was given by Binev, Dahmen, and DeVore [13] in 2d in 2004. We summarize in this chapter the generalization to any space dimension by Stevenson [70].

4.1 Subdivision of a single simplex

We first describe how a single d-simplex is bisected, along with the concepts of vertex order and type. We then turn to recurrent bisection of a given initial element and the problem of shape regularity.

Bisection Rule based on Vertex Order and Type. We identify a simplex T with the set of its *ordered vertices* and its *type* t by

$$T = \{z_0, \ldots, z_d\}_t, \quad t \in \{0, \ldots, d-1\}.$$

Given such a d-simplex T we use the following bisection rule to split it in a unique fashion and to impose both vertex order and type to its children. The edge $\overline{z_0 z_d}$ connecting the first and last vertex of T is the *refinement edge* of T and its midpoint $\bar{z} = \frac{z_0 + z_d}{2}$ becomes the new vertex. Connecting the new vertex $\bar{z}$ with the vertices of T other than z_0, z_d determines the common side $S = \{\bar{z}, z_1, \ldots, z_{d-1}\}$ shared by the

two children T_1, T_2 of T. The *bisection rule* dictates the following vertex order and type for T_1, T_2

$$\begin{aligned} T_1 &:= \{z_0, \bar{z}, \underbrace{z_1, \dots, z_t}_{\rightarrow}, \underbrace{z_{t+1}, \dots, z_{d-1}}_{\rightarrow}\}_{(t+1) \bmod d}, \\ T_2 &:= \{z_d, \bar{z}, \underbrace{z_1, \dots, z_t}_{\rightarrow}, \underbrace{z_{d-1}, \dots, z_{t+1}}_{\leftarrow}\}_{(t+1) \bmod d}, \end{aligned} \tag{53}$$

with the convention that arrows point in the direction of increasing indices and $\{z_1, \dots, z_0\} = \emptyset$, $\{z_d, \dots, z_{d-1}\} = \emptyset$.

In 2d the bisection rule does not depend on the element type and we get for $T = \{z_0, z_1, z_2\}$ the two children

$$T_1 = \{z_0, \bar{z}, z_1\} \qquad \text{and} \qquad T_2 = \{z_2, \bar{z}, z_1\}.$$

As depicted in Fig. 2, the refinement edge of the two children is opposite to the

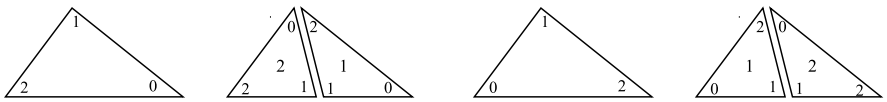


Fig. 2 Refinement of a single triangle $T = \{z_0, z_1, z_2\}$ and its reflected triangle $T_R = \{z_2, z_1, z_0\}$

new vertex $\bar{z}$, whence this procedure coincides with *the newest vertex bisection* for $d = 2$. For $d \geq 3$ the bisection of an element does depend on its type, and, as we shall see below, this is important for preserving shape regularity. For instance, in 3d the children of $T = \{z_0, z_1, z_2, z_3\}_t$ are (see Fig. 3)

$$\begin{aligned} t = 0: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_1 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_2, z_1\}_1, \\ t = 1: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_2 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_1, z_2\}_2, \\ t = 2: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_0 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_1, z_2\}_0. \end{aligned}$$

Note that the vertex labeling of T_1 is type-independent, whereas that of T_2 is the same for type 1 and 2. To account for this fact the vertices z_1 and z_2 of T are tagged $(3,2,2)$ and $(2,3,3)$ in Fig. 3. The type of T then dictates which component of the triple is used to label the vertex.

Any different labeling of an element's vertices does not change its geometric shape but applying the above bisection rule it does change the shape and vertex order of its two children. This holds true for any relabeling except one. An element with this special relabeling of vertices is called *reflected element*. We state next its precise definition.

Definition 4.1 (Reflected Element). Given an element $T = \{z_0, \cdots, z_d\}_t$, the *reflected element* is given by

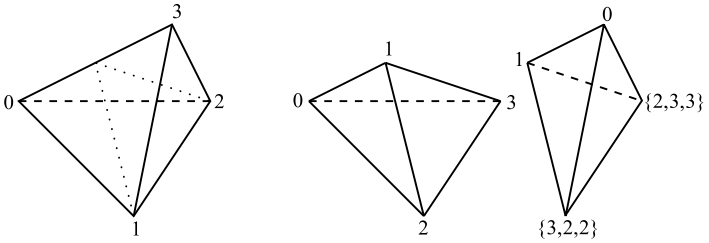


Fig. 3 Refinement of a single tetrahedron T of type t. The child T_1 in the middle has the same node ordering regardless of type. In contrast, for the child T_2 on the right a triple is appended to two nodes. The local vertex index is given for these nodes by the t-th component of the triple

$$T_R := \{z_d, \underbrace{z_1, \ldots, z_t}_{\rightarrow}, \underbrace{z_{d-1}, \ldots, z_{t+1}}_{\leftarrow}, z_0\}_t.$$

Fig. 2 depicts for 2d $T = \{z_0, z_1, z_2\}$ and $T_R = \{z_2, z_1, z_0\}$. It shows that the children of T and T_R are the same. This property extends to $d \geq 3$; compare with Problem 4.2. Any other relabeling of vertices leads to different shapes of the children, in fact as many as $\frac{1}{2}(d+1)!$

Recurrent Bisection and Binary Tree. We next turn towards the recurrent bisection of a given *initial* simplex $T_0 = \{z_0, \ldots, z_d\}_{t_0}$. We let $\{T_1, T_2\} = \mathsf{BISECT}(T)$ be a function that implements the above bisection rule and outputs the two children of T. The input of BISECT can be T_0 or any element of the output from a previous application of BISECT.

This procedure of recurrent bisection of T_0 is associated with an *infinite binary tree* $\mathbb{F}(T_0)$. The nodes $T \in \mathbb{F}(T_0)$ correspond to simplices generated by repeated application of BISECT. The two successors of a node T are the two children $\{T_1, T_2\} = \mathsf{BISECT}(T)$. Note that $\mathbb{F}(T_0)$ strongly depends on the vertex order of T_0 and its type t_0. Once this is set for T_0 the associated binary tree is completely determined by the bisection rule. Recalling that the children of an element and its reflected element are the same this gives in total $\frac{d(d+1)!}{2}$ different binary trees that can be associated with T_0 by the bisection procedure.

The binary tree $\mathbb{F}(T_0)$ holds full information about the shape, ordering of vertices, type, etc. of any element T that can be generated by recurrent bisection of T_0. Important in this context is the distance of T to T_0 within $\mathbb{F}(T_0)$, which we call *generation*.

Definition 4.2 (Generation). The *generation* $g(T)$ of a node/element $T \in \mathbb{F}(T_0)$ is the number of its ancestors in the tree, or, equivalently, the number of bisections needed to create T from T_0.

Using the notion of generation, some information about T can uniquely be deduced from $g(T)$. For instance, for an element $T \in \mathbb{F}(T_0)$, its type is $(g(T) + t_0) \bmod d$, and, in view of the definition $h_T = |T|^{1/d}$, its size is

$$h_T = 2^{-g(T)/d} h_{T_0}. \tag{54}$$

Shape Regularity. We next analyse the shape coefficients of descendants of a given simplex T_0. A uniform bound on the shape coefficients σ_T for all $T \in \mathbb{F}(T_0)$ plays a crucial role in the interpolation estimates derived in Sect. 5.1. When turning towards shape regularity the dependence of the bisection rule on the element type for $d \geq 3$ becomes indispensable. The fact that the type t increases by 1 and the vertex ordering changes with t implies that after d recurrent bisections of T all its edges are bisected; compare with Problem 4.1.

We first consider a so-called *Kuhn-simplex*, i. e., a simplex with (ordered) vertices

$$z_0^\pi = 0, \quad z_i^\pi := \sum_{j=1}^{i} e_{\pi(j)} \quad \text{for all } i = 1, \dots, d,$$

where π is a permutation of $\{1, \dots, d\}$. Note, that $z_d^\pi = (1, \dots, 1)^T$ for any permutation π. Therefore, the refinement edge $\overline{z_0^\pi, z_d^\pi}$ of any Kuhn-simplex is always the longest edge. If T_0 is a type 0 Kuhn-simplex, recurrent bisection always cuts the longest edge. This is the key property for obtaining uniform bound on the shape coefficients [47, 72].

Theorem 4.1 (Shape Regularity for a Kuhn-Simplex). *All 2^g descendants of generation g of a Kuhn-simplex $T_\pi = \{z_0^\pi, \dots, z_d^\pi\}_0$ are mutually congruent with at most d different shapes. Moreover, the descendants of generation d are congruent to T_0 up to a scaling with factor $\frac{1}{2}$.*

In two dimensions, all descendants of a Kuhn-triangle belong to one similarity class; see Figure 4. Using an affine transformation we conclude from Theorem 4.1 shape

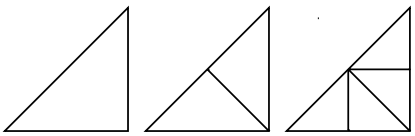


Fig. 4 Recurrent bisection of a Kuhn-triangle generates only one similarity class

regularity for all descendants of an arbitrary simplex.

Corollary 4.1 (Shape Regularity). *Let $T_0 = \{z_0, \dots, z_d\}_t$ be an arbitrary d-simplex. Then all descendants of T generated by bisection are shape regular, i. e.,*

$$\sup_{T \in \mathbb{F}(T_0)} \sigma_T = \sup_{T \in \mathbb{F}(T_0)} \frac{\overline{h}_T}{\underline{h}_T} \leq C(T_0) < \infty.$$

Proof. Consider first a simplex T_0 of type 0 and let $\hat{T}_0 := \{\hat{z}_0, \dots, \hat{z}_d\}_0$ be the a Kuhn-simplex of type 0. From Lemma 3.1 we know that there exists a bijective affine mapping $F : \hat{T}_0 \to T_0$.

Recurrent refinement by bisection implies that for any $T \in \mathbb{F}(T_0)$ there exists a unique $\hat{T} \in \mathbb{F}(\hat{T}_0)$ such that $T = F(\hat{T})$. Since all descendants of $\hat{T}_0$ belong to at most d similarity classes, this implies that the minimal angle of all descendants of T_0 is uniformly bounded from below by a constant solely depending on the shape of T_0.

The same is valid for a simplex T_0 of type $t \in \{1, \dots, d-1\}$ because its 2^{d-t} descendants of generation $d-t$ are all of type 0. □

Note, that for a general d-simplex, the number of similarity classes for the descendants is larger than for a Kuhn d-simplex. This number is 4 for $d = 2$; compare Figures 4 and 5.

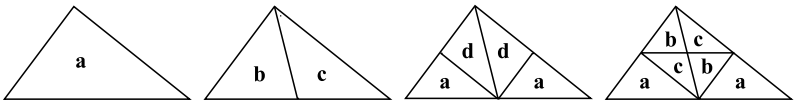


Fig. 5 Bisection produces at most 4 similarity classes for any initial triangle

4.2 Mesh refinement by bisection

After discussing the refinement of a single simplex, we next turn to the refinement of a given initial conforming triangulation $\mathscr{T}_0$ by bisection. For recurrent refinement of a single element T_0 we are free to choose any order of its vertices and element type. The requirement to produce conforming refinements of $\mathscr{T}_0$ results in restrictions on local vertex order and type of the elements in $\mathscr{T}_0$. We first introduce the binary forest associated to triangulations generated by bisection and then elaborate on conformity and basic properties of the refined triangulations.

Master Forest and Forest. We recall that recurrent bisection of an element $T_0 \in \mathscr{T}_0$ is uniquely associated with an infinite binary tree $\mathbb{F}(T_0)$; see Sect. 4.1. In the same way we can identify all possible refinements of $\mathscr{T}_0$ with a *master forest* of binary trees.

Definition 4.3 (Forest and Refinement). Let $\mathscr{T}_0$ be an initial conforming triangulation. Then

$$\mathbb{F} = \mathbb{F}(\mathscr{T}_0) := \bigcup_{T_0 \in \mathscr{T}_0} \mathbb{F}(T_0).$$

is the associated *master forest* of binary trees. For a node $T \in \mathbb{F}$ so that $T \in \mathbb{F}(T_0)$ with $T_0 \in \mathscr{T}_0$, the generation $g(T)$ is the generation of T within $\mathbb{F}(T_0)$.

A subset $\mathscr{F} \subset \mathbb{F}$ is called *forest* iff

(1) $\mathscr{T}_0 \subset \mathscr{F}$;
(2) all nodes of $\mathscr{F} \setminus \mathscr{T}_0$ have a predecessor;
(3) all nodes of $\mathscr{F}$ have either two successors or none.

A forest $\mathscr{F}$ is called *finite*, if $\max_{T\in\mathscr{F}} g(T) < \infty$. The nodes with no successors are called *leaves* of $\mathscr{F}$.

Any finite forest $\mathscr{F}$ is uniquely associated with a triangulation $\mathscr{T} = \mathscr{T}(\mathscr{F})$ of Ω by defining $\mathscr{T}$ to be the set of all leaves in $\mathscr{F}$. Given two finite forests $\mathscr{F}, \mathscr{F}_* \in \mathbb{F}$ with associated triangulations $\mathscr{T}, \mathscr{T}_*$ we call $\mathscr{T}_*$ *refinement* of $\mathscr{T}$ iff $\mathscr{F} \subset \mathscr{F}_*$ and we denote this by $\mathscr{T} \leq \mathscr{T}_*$ or, equivalently, $\mathscr{T}_* \geq \mathscr{T}$.

Note that the definition of a finite forest $\mathscr{F}$ implies that the leaf nodes cover Ω, whence the associated triangulation $\mathscr{T}(\mathscr{F})$ is a partition of Ω. In general, this triangulation is not conforming and it is a priori not clear that conforming refinements of $\mathscr{T}_0$ exist.

Conforming Refinements. We next wonder about the properties of $\mathscr{T}_0$ that allow for conforming refinements. This brings us to the notion of *neighboring elements*.

Definition 4.4 (Neighboring Elements). Two elements $T_1, T_2 \in \mathscr{T}$ are called *neighboring elements* if they share a common side, namely a full $(d-1)$-simplex.

In 2d new vertices are always midpoints of edges. Generating the descendants of generation 2 for all elements of a given conforming triangulation $\mathscr{T}$ bisects all edges of $\mathscr{T}$ exactly once and all midpoints of the edges are vertices of the grandchildren. This implies conformity for $d = 2$. For $d > 2$ the situation is completely different.

Assume $d = 3$ and let $T_1, T_2 \in \mathscr{T}$ be two neighboring elements with common side $S = T_1 \cap T_2$. Denote by E_1, E_2 their respective refinement edges and assume that they belong to S. The 3d bisection of T_1 leads to a 2d bisection of S with E_1 being the refinement edge of S induced by T_1. The same holds true for T_2. If $E_1 \neq E_2$ the new edges in S created by refinement of T_1 and T_2 are not identical but do intersect. This leads to a non-conformity that cannot be cured by any further bisection of S. The same holds true for $d > 3$ upon replacing the newly created edge by the newly created $(d-2)$-simplex inside the common side. This yields for $d \geq 3$ a *necessary* condition for constructing a conforming refinement:

> *Whenever the refinement edges of two neighboring elements are both on the common side they have to coincide.*

For $d = 3$ this condition has been shown to also be *sufficient* for obtaining conforming refinements. It is also known that for any initial conforming triangulation $\mathscr{T}_0$ there exists a local labeling of the vertices satisfying this condition [10, 46, 2].

For $d > 3$ the above condition is not known to be sufficient. In addition, for proving the complexity result in Sect. 4.5 we need stronger assumptions on the distribution of refinement edges on $\mathscr{T}_0$. For the general case $d \geq 2$, we therefore formulate an assumption on the labeling of $\mathscr{T}_0$ given by Stevenson that ensures conformity of any *uniform* refinement of $\mathscr{T}_0$. This condition relies on the notion of reflected neighbor.

Definition 4.5 (Reflected Neighbors). Two neighboring elements $T = \{z_0, \ldots, z_d\}_t$ and $T' = \{z'_0, \ldots, z'_d\}_t$ are called *reflected neighbors* iff the ordered vertices of T or T_R coincide exactly with those of T' at all but one position.

We are now in the position to pose the assumptions on the initial triangulation $\mathcal{T}_0$.

Assumption 11.1 (Admissibility of the Initial Grid). *Let $\mathcal{T}_0$ be a conforming triangulation that fulfills*

(1) *all elements are of the same type $t \in \{0,\dots,d-1\}$;*
(2) *all neighboring elements $T = \{z_0,\dots,z_d\}_t$ and $T' = \{z'_0,\dots,z'_d\}_t$ with common side S are* matching neighbors *in the following sense: if $\overline{z_0 z_d} \subset S$ or $\overline{z'_0 z'_d} \subset S$ then T and T' are reflected neighbors; otherwise the pair of neighboring children of T and T' are reflected neighbors.*

For instance, the set of the $d!$ Kuhn-simplices of type 0 is a conforming triangulation of the unit cube in $\mathbb{R}^d$ satisfying Assumption 11.1; see Problem 4.3. We also refer to Fig. 6 and Problem 4.4 to explore this concept for $d = 2$.

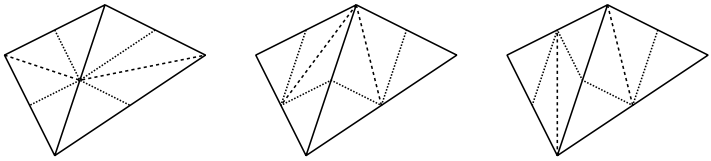


Fig. 6 Matching neighbors in 2d and their descendants of generation 1 and 2. The elements in the left and middle picture are reflected neighbors. The elements in the rightmost picture are not reflected neighbours, but the pair of their neighboring children are

Uniform Refinements. We next state the following important implication of this structural assumption on $\mathcal{T}_0$. The proof is a combination of [72, Sect. 4] and [70, Theorem 4.3].

Theorem 4.2 (Uniform Refinement). *Let $\mathcal{T}_0$ be a conforming triangulation and for $g \in \mathbb{N}_0$ denote by*

$$\mathcal{T}_g := \{T \in \mathbb{F}(\mathcal{T}_0) \mid g(T) = g\}$$

the uniform refinement of $\mathcal{T}_0$ with elements of generation exactly g.

If Assumption 11.1 is satisfied, then $\mathcal{T}_g$ is conforming for any $g \in \mathbb{N}_0$. In addition, if all elements in $\mathcal{T}_0$ are of the same type, then condition (2) is necessary for $\mathcal{T}_g$ to be conforming for all g.

To interpret Theorem 4.2 we introduce the following useful definition.

Definition 4.6 (Compatible Bisection). We say that two elements $T, T' \in \mathbb{F}$ are *compatibly divisible* if they have the same refinement edge. If all elements sharing an edge are compatibly divisible, then they form a *bisection patch*.

Using this notion, Theorem 4.2 states that two elements $T, T' \in \mathbb{F}$ of the same generation sharing a common edge are either compatibly divisible, or the refinement of T does not affect T' and vice versa. In the latter case any common edge is neither the refinement edge of T nor of T'.

Let $d = 2$ and $T = \{z_0, z_1, z_2\}_t$ and $T' = \{z'_0, z'_1, z'_2\}_t$ be neighboring elements with common side S. If $\overline{z_0 z_d} = \overline{z'_0 z'_d}$ then T and T' are compatibly divisible and thus form a bisection patch: they can be refined without affecting any other element. If $z_1, z'_1 \in S$, then the pair of neighboring children of T and T' are compatibly divisible and thus form a bisection patch; compare with Fig. 6 and Problem 4.4.

Remark 4.1 (Discussion of Assumption 11.1). Assumption 11.1, given by Stevenson [70], is weaker than the condition required by Maubach [47] and Traxler [72]: they asked that all neighboring elements are reflected neighbors. It is an important open question whether for any conforming triangulation $\mathscr{T}_0$ there exists a suitable labeling of the element's vertices such that Assumption 11.1 is satisfied.

For dimension $d = 2$ such a result has been shown by Mitchell [49, Theorem 2.9] as well as Binev et al. [13, Lemma 2.1]. Both proofs are based on graph theory and they are not constructive. It can be shown that the problem of finding a suitable labeling of the vertices, the so-called *perfect matching*, is NP-complete.

For dimension $d > 2$ this is an open problem. In 3d Kossaczký has constructed a conforming refinement of any given coarse grid into an initial grid $\mathscr{T}_0$ that satisfies Assumption 11.1. This construction has been generalized by Stevenson to any space dimension. [70, Appendix A].

As mentioned above, the conditions of Bänsch [10], Liu and Joe [46], and Arnold et al. [2] on the initial tetrahedral mesh can be satisfied for any given conforming triangulation. But then it can only be shown that uniform refinements $\mathscr{T}_g$ with $g \bmod d = 0$ are conforming [2, 10, 46]. The property that any uniform refinement $\mathscr{T}_g$ for $g \in \mathbb{N}_0$ is conforming is the key tool for the complexity proof in Sect. 4.5.

We next define the class of *conforming* refinements of $\mathscr{T}_0$ to be

$$\mathbb{T} = \{\mathscr{T} = \mathscr{T}(\mathscr{F}) \mid \mathscr{F} \subset \mathbb{F} \text{ is finite and } \mathscr{T}(\mathscr{F}) \text{ is conforming}\}.$$

Then Theorem 4.2 has two direct consequences.

(a) The class $\mathbb{T}$ contains an infinite number of conforming refinements of $\mathscr{T}_0$.
(b) There exists a function $\mathsf{REFINE}(\mathscr{T}, \mathscr{M})$ that, given a conforming triangulation $\mathscr{T} \in \mathbb{T}$ and a subset $\mathscr{M} \subset \mathscr{T}$ of marked elements, bisects all simplices in $\mathscr{M}$ at least once, and outputs the smallest conforming refinement $\mathscr{T}_* \in \mathbb{T}$ of $\mathscr{T}$ with $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Before constructing such function REFINE we analyze some basic properties of triangulations.

4.3 Basic properties of triangulations

In this section we analyze basic properties of refinement by bisection, namely uniform shape regularity, convergence of mesh-size functions, and the cardinality of an

overlay of two triangulations. The results can be easily derived using the structure of the master forest $\mathbb{F}$.

Uniform Shape Regularity. A direct consequence of Corollary 4.1 is that refinement by bisection only produces elements T with shape coefficient σ_T uniformly bounded by a constant solely depending on $\mathscr{T}_0$; recall Definition 3.4.

Lemma 4.1. *All elements in $\mathbb{F}$ are uniformly shape regular, i. e.,*

$$\sup_{T\in\mathbb{F}} \sigma_T = \sup_{T\in\mathbb{F}} \frac{\overline{h}_T}{\underline{h}_T} \leq C(\mathscr{T}_0) < \infty.$$

For any conforming mesh $\mathscr{T} \in \mathbb{T}$, the discrete neighborhood of $T \in \mathscr{T}$ is given by

$$N_{\mathscr{T}}(T) := \{T' \in \mathscr{T} \,|\, T' \cap T \neq \emptyset\}.$$

Lemma 4.1 implies that the cardinality of this patch is bounded uniformly and the measure of all its elements is comparable

$$\max_{T\in\mathscr{T}} \#N_{\mathscr{T}}(T) \leq C(\mathscr{T}_0), \qquad \max_{T'\in N_{\mathscr{T}}(T)} \frac{|T|}{|T'|} \leq C(\mathscr{T}_0), \tag{55}$$

with $C(\mathscr{T}_0)$ only depending on $\mathscr{T}_0$. This is usually called *local quasi-uniformity*.

Convergence of Mesh-Size Functions. Let $\{\mathscr{T}_k\}_{k\geq 0} \subset \mathbb{T}$ be any sequence of nested refinements, i. e., $\mathscr{T}_k \leq \mathscr{T}_{k+1}$ for $k \geq 0$. This sequence is accompanied by the sequence of mesh-size functions $\{h_k\}_{k\geq 0}$, defined as $h_k \in L^\infty(\Omega)$ with

$$h_{k|T} = h_T = |T|^{1/d} \qquad \text{for all } T \in \mathscr{T}_k.$$

If the sequence is produced by uniform refinement then we easily obtain from (54)

$$\lim_{k\to\infty} \|h_k\|_{L^\infty(\Omega)} = 0. \tag{56}$$

However, this may not hold when the sequence $\mathscr{T}_k$ is generated adaptively, i. e., we allow for local refinement. Therefore we have to generalize it appropriately. For a first generalization of (56), we observe that the skeleton $\Gamma_k := \bigcup\{\partial T \cap \Omega : T \in \mathscr{T}_k\}$ of $\mathscr{T}_k$ has d-dimensional Lebesgue measure zero. We may thus interpret h_k as a piecewise constant function in $L^\infty(\Omega)$. Moreover, the limiting skeleton $\Gamma_\infty := \cup_{k\geq 0}\Gamma_k$ has also d-dimensional Lebesgue measure zero. Since, for every $x \in \Omega \setminus \Gamma_\infty$, the sequence $h_k(x)$ is monotonically decreasing and bounded from below by 0,

$$h_\infty(x) := \lim_{k\to\infty} h_k(x) \tag{57}$$

is well-defined for $x \in \Omega \setminus \Gamma_\infty$ and defines a function in $L^\infty(\Omega)$. As the next lemma shows, the pointwise convergence in (57) holds actually in $L^\infty(\Omega)$. Another generalization of (56), where the limit function is 0, will be provided in Corollary 7.1 in Chap. 7.

Lemma 4.2 (Uniform Convergence of Mesh-Size Functions). *For any sequence $\{\mathscr{T}_k\}_{k\geq 0} \subset \mathbb{T}$ of nested refinements the corresponding sequence $\{h_k\}_{k\geq 0}$ of mesh-size functions converges uniformly in $\Omega \setminus \Gamma_\infty$ to h_∞, i. e.,*

$$\lim_{k\to\infty} \|h_k - h_\infty\|_{L^\infty(\Omega)} = 0.$$

Proof. 1 Denote by $\mathscr{F}_k = \mathscr{F}(\mathscr{T}_k)$ the corresponding forest of $\mathscr{T}_k$. From $\mathscr{T}_k \leq \mathscr{T}_{k+1}$ we conclude $\mathscr{F}_k \subset \mathscr{F}_{k+1}$ and thus the forest

$$\mathscr{F}_\infty := \bigcup_{k\geq 0} \mathscr{F}_k$$

is well defined. Note that in general $\mathscr{F}_\infty$ is infinite.

2 For arbitrary $\varepsilon > 0$, let $g = g(\varepsilon) \in \mathbb{N}$ be the smallest number such that

$$g \geq \log(\varepsilon^d/M)/\log(\tfrac{1}{2})$$

with $M = \max\{|T| \mid T \in \mathscr{T}_0\}$. Obviously, $\hat{\mathscr{F}} := \{T \in \mathscr{F}_\infty \mid g(T) \leq g\}$ is a finite forest and $\mathscr{T}(\hat{\mathscr{F}})$ is a triangulation of Ω. Since $\hat{\mathscr{F}} \subset \mathscr{F}_\infty$ is finite there exists $k = k(\varepsilon) \geq 0$ with $\hat{\mathscr{F}} \subset \mathscr{F}_k$.

3 Let $T \in \mathscr{T}_k$ be any leaf node of $\mathscr{F}_k$ and let $T \in \mathscr{F}(T_0)$ for some $T_0 \in \mathscr{T}_0$. To estimate $h_k - h_\infty$ on T, we distinguish the following two cases:

Case 1: $g(T) < g$. This implies that T is a leaf node of $\mathscr{F}_\infty$ and thus $h_{k|T} = h_{\infty|T}$ or, equivalently, $(h_k - h_\infty)_{|T} = 0$.

Case 2: $g(T) \geq g$. Hence, T is generated by at least g bisections of T_0. By (54), the monotonicity of the mesh-size functions, and the choice of g, we obtain

$$0 \leq (h_k - h_\infty)_{|T} \leq h_{k|T} = h_T \leq 2^{-g(T)/d} h_{T_0} \leq 2^{-g/d} M^{1/d} \leq \varepsilon.$$

Combining the two cases we end up with $0 \leq (h_k - h_\infty)_{|T} \leq \varepsilon$ for all $T \in \mathscr{T}_k$. Since ε is arbitrary and $0 \leq h_\ell - h_\infty \leq h_k - h_\infty$ in Ω for all $\ell \geq k$, this finishes the proof. □

Overlay of Triangulations. Let $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ be conforming triangulations with corresponding finite forests $\mathscr{F}_1$ and $\mathscr{F}_2$. Then $\mathscr{F}_1 \cup \mathscr{F}_2$ is also a finite forest and we call the unique triangulation

$$\mathscr{T}_1 \oplus \mathscr{T}_2 := \mathscr{T}(\mathscr{F}_1 \cup \mathscr{F}_2) \tag{58}$$

the *overlay of* $\mathscr{T}_1$ *and* $\mathscr{T}_2$. The name overlay is motivated by printing 2d triangulations $\mathscr{T}_1$ and $\mathscr{T}_2$ at the same position on two slides. The overlay is then the triangulation that can be seen when putting one slide on top of the other. It turns out that the overlay is the smallest conforming triangulation with $\mathscr{T}_1, \mathscr{T}_2 \leq \mathscr{T}_1 \oplus \mathscr{T}_2$ and its cardinality can be estimated by the ones of $\mathscr{T}_1$ and $\mathscr{T}_2$.

Lemma 4.3 (Overlay of Meshes). *For $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ the overlay $\mathscr{T} := \mathscr{T}_1 \oplus \mathscr{T}_2$ is the smallest common refinement of $\mathscr{T}_1$ and $\mathscr{T}_2$ and satisfies*

$$\#\mathscr{T} \leq \#\mathscr{T}_1 + \#\mathscr{T}_2 - \#\mathscr{T}_0.$$

Proof. Argue by contradiction and assume that $\mathscr{T}$ contains a non-conforming vertex z. That is, there exist $T_1, T_2 \in \mathscr{T}$ such that z is a vertex of T_1 and $z \in T_2$ is not a vertex of T_2. Without loss of generality let $T_1 \in \mathscr{T}_1$. Since $\mathscr{T}_1$ is conforming, there exists a $T' \in \mathscr{T}_1$, $T' \subset T_2$ such that z is a vertex of T'. Hence, T' is a descendant of T_2 in $\mathscr{F}_1$ and thus T_2 cannot be a leaf node of $\mathscr{F}(\mathscr{T})$, i.e., $T_2 \notin \mathscr{T}$, a contradiction. Since the overlay only contains elements from $\mathscr{T}_1$ or $\mathscr{T}_2$ and is conforming, it is the smallest conforming refinement.

For $T \in \mathscr{T}_0$ and $i = 1,2$ we denote by $\mathscr{F}_i(T) \subset \mathscr{F}(\mathscr{T})$ the binary trees with root T corresponding to $\mathscr{T}_i$ and let $\mathscr{T}_i(T)$ be the triangulation given by the leaf nodes of $\mathscr{F}_i(T)$. Since $\mathscr{T}(T) \subset \mathscr{T}_1(T) \cup \mathscr{T}_2(T)$, we infer that $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T)$. We now show that $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$ by distinguishing two cases.

Case 1: $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) \neq \emptyset$. Then there exists $T' \in \mathscr{T}_1(T) \cap \mathscr{T}_2(T)$, and so $T' \in \mathscr{T}(T)$. By counting T' only once in $\#(\mathscr{T}_1(T) \cup \mathscr{T}_2(T))$ we get $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$.

Case 2: $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) = \emptyset$. Then there exists $T' \in \mathscr{T}_1(T)$ (resp., $T' \in \mathscr{T}_2(T)$) so that $T' \notin \mathscr{T}(T)$, for otherwise $T' \in \mathscr{T}_2(T)$ (resp., $T' \in \mathscr{T}_1(T)$), thereby contradicting the assumption. We obtain again $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$.

Finally, since $\mathscr{T}_i = \bigcup_{T \in \mathscr{T}_0} \mathscr{T}_i(T)$, the assertion follows by adding over the elements in $\mathscr{T}_0$. □

4.4 Refinement algorithms

We discuss two refinement algorithms based on the bisection rule introduced in Sect. 4.1. Given a conforming triangulation $\mathscr{T}$ and a subset of marked elements $\mathscr{M}$ both variants output the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ such that all elements of $\mathscr{M}$ are bisected, i. e., $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Iterative Refinement. The basic idea is to first bisect all marked elements in $\mathscr{T}$ leading to a non-conforming grid $\mathscr{T}_*$. In order to restore conformity, we identify all elements $T \in \mathscr{T}$ containing a so-called *irregular (or hanging) node* $z \in T$, namely a vertex $z \in \mathscr{V}_{\mathscr{T}_*}$ which is not a vertex of T. These elements are then also scheduled for refinement. This procedure has to be iterated until all irregular nodes are removed and this step is called *completion*. The core of iterative refinement is a routine that bisects all marked elements in a possibly non-conforming triangulation:

REFINE_MARKED($\mathscr{T}, \mathscr{M}$)
for all $T \in \mathscr{M}$ do
  $\{T_0, T_1\} = \mathsf{BISECT}(T)$;
  $\mathscr{T} := \mathscr{T} \setminus \{T\} \cup \{T_0, T_1\}$;
end for
return($\mathscr{T}$)

The refinement of a given conforming grid $\mathscr{T}$ with a subset of marked elements $\mathscr{M}$ into a new conforming refinement is then executed by

REFINE$(\mathscr{T},\mathscr{M})$
while $\mathscr{M} \neq \emptyset$ do
  $\mathscr{T} :=$ REFINE_MARKED$(\mathscr{T},\mathscr{M})$;
  $\mathscr{M} := \{T \in \mathscr{T} \mid T$ contains an irregular node$\}$;
end while
return$(\mathscr{T})$

We let $\mathscr{T}_*$ be the output of REFINE_MARKED$(\mathscr{T},\mathscr{M})$ on its first call. Since non-conforming situations can only be cured by refining all elements containing an irregular node, the above algorithm outputs the smallest conforming refinement of $\mathscr{T}_*$ if the while-loop terminates. We let g be the maximal generation of any element in $\mathscr{T}_*$. By Theorem 4.2 the uniform refinement $\mathscr{T}_g$ is conforming, and by construction it satisfies $\mathscr{T}_* \leq \mathscr{T}_g$. Since REFINE$(\mathscr{T},\mathscr{M})$ only refines elements to remove non-conforming situations, any intermediate grid $\mathscr{T}$ produced by REFINE_MARKED$(\mathscr{T},\mathscr{M})$ satisfies $\mathscr{T} \leq \mathscr{T}_g$ and this implies that the while loop in the above algorithm terminates.

We point out that this algorithm works without any assumption on the ordering of vertices in $\mathscr{T}_0$ in 2d and with the less restrictive assumptions in [2, 10, 46] in 3d. This follows from the fact that $\mathscr{T}_g$ with $g \bmod d = 0$ is conforming and thus one can choose a suitable $\mathscr{T}_g \geq \mathscr{T}_*$; compare with Remark 4.1.

The above implementation of iterative refinement is not efficient since there are too many loops in the completion step. We observe that the bisection of a single element T enforces the bisection of all elements at its refinement edge. Some of these elements may also be marked for refinement and will directly be refined. Other elements have to be refined in the completion step. The algorithm can be speeded up by directly scheduling those elements for refinement.

This motivates the simultaneous bisection of all elements meeting at the refinement edge. This variant is discussed next.

Recursive Refinement. Let $\mathscr{T}$ be a given conforming grid and let $T \in \mathscr{T}$ be an element with refinement edge E. We define the *refinement patch of* T to be

$$R(\mathscr{T};T) := \{T' \in \mathscr{T} \mid E \subset T'\}.$$

As mentioned above, a bisection of T enforces a refinement of all elements in $R(\mathscr{T};T)$ for regaining conformity. We could avoid non-conforming situations by a simultaneous refinement of the whole refinement patch. This is only possible if all elements in $R(\mathscr{T};T)$ are compatibly divisible, i. e., E is the refinement edge of all $T' \in R(\mathscr{T};T)$ and $R(\mathscr{T};T)$ is a bisection patch. This is called the *atomic refinement operation* and is depicted in Fig. 7 for $d = 2$ (top) and $d = 3$ (bottom).

If there are elements in $R(\mathscr{T};T)$ that are not compatibly divisible with T, the basic idea is to recursively refine these elements first. This builds up the new refinement patch around E that in the end allows for the atomic refinement operation.

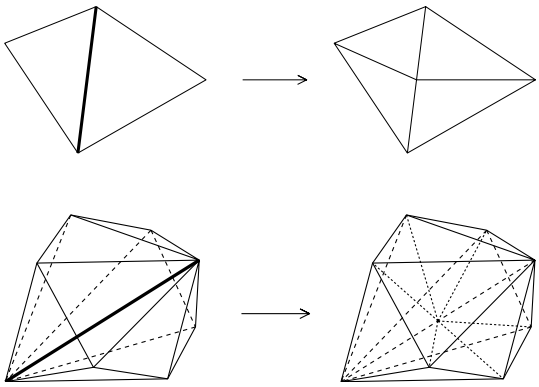


Fig. 7 Atomic refinement operation in 2d (top) and 3d (bottom): The common edge is the refinement edge for all elements

In 2d, there is one neighbor sharing the refinement edge E in case E is interior. Either this neighbor is compatibly divisible, or the neighboring child is compatibly divisible after bisection of the neighbor. If E lies on the boundary, instead, bisection can be executed directly. Fig. 8 illustrates a situation that requires recursion.

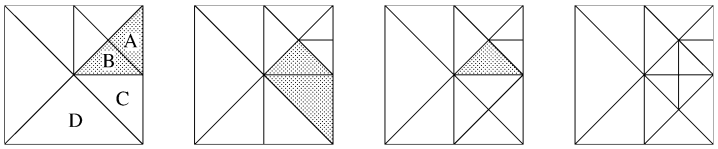


Fig. 8 Recursive refinement in 2d: Triangles A and B are initially marked for refinement

In higher dimension there is in general a whole bunch of elements in $R(\mathscr{T};T)$. Since $R(\mathscr{T};T) \subset N_{\mathscr{T}}(T)$, (55) implies that the cardinality of $R(\mathscr{T};T)$ is uniformly bounded depending only on $\mathscr{T}_0$. Here it may happen that several elements have to be refined before we can perform the atomic refinement operation. It may also happen that an element inside the refinement patch has to be refined several times but the number of bisections is bounded by $d-1$; see Lemma 4.4 below. This lemma also allows for an elegant formulation of the recursive algorithm.

Lemma 4.4. *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 11.1 and let $\mathscr{T} \in \mathbb{T}$ be a conforming refinement.*

Then any $T \in \mathscr{T}$ is of locally highest generation in $R(\mathscr{T};T)$, i. e.,

$$g(T) = \max\{g(T') \mid T' \in R(\mathscr{T};T)\}$$

and $T' \in R(\mathscr{T};T)$ is compatibly divisible with T if and only if $g(T') = g(T)$.

Furthermore, $\min\{g(T') \mid T' \in R(\mathscr{T};T)\} \geq g(T) - d + 1$ *and a non-compatibly divisible neighboring element of T has generation $g(T)-1$.*

Proof. Denote by E the refinement edge of T and set $g := g(T)$. The uniform refinement $\mathscr{T}_{g+1}$ of $\mathscr{T}_0$ contains the midpoint $\bar{z}$ of E as a vertex and is a conforming refinement of $\mathscr{T}$. For any $T' \in R(\mathscr{T};T)$ the new vertex $\bar{z}$ is an irregular node on T', whence $T' \notin \mathscr{T}_{g+1}$. Since $\mathscr{T}_{g+1}$ is a conforming refinement of $\mathscr{T}$ we know that descendants of T' belong to $\mathscr{T}_{g+1}$ and thus $g(T') \le g$ for all $T' \in R(\mathscr{T};T)$.

If T' is compatibly divisible with T, then $\bar{z}$ is the new vertex of the two children, which belong to $\mathscr{T}_{g+1}$; hence, $g(T') = g$. If T' is not compatibly divisible with T, then $\bar{z}$ is the new vertex of descendants of one child of T', whence $g(T') < g$.

The refinement rule (53) implies that after d recurrent bisections all edges of the original simplex are bisected (see Problem 4.1). Consequently, any $T' \in R(\mathscr{T};T)$ has descendants of generation at most $g(T') + d$ that have $\bar{z}$ as a vertex and belong to $\mathscr{T}_{g+1}$. This yields $g(T') \ge g - d + 1$.

If T' is a non-compatibly divisible neighbor of T, then the refinement rule (53) implies that the refinement edge of one child T'' of T' is contained in the common side of T and T'. Since $\mathscr{T}_{g+1}$ is conforming this implies that T and T'' are compatibly divisible, and thus $g(T') = g - 1$. □

For $\mathscr{T} \in \mathbb{T}$ the recursive refinement of a single element $T \in \mathscr{T}$ now reads:

```
REFINE_RECURSIVE(𝒯, T)
do forever
    get refinement patch R(𝒯, T);
    access T' ∈ R(𝒯, T) with g(T') = min{g(T'') | T'' ∈ R(𝒯;T)};
    if g(T') < g(T) then
        𝒯 := REFINE_RECURSIVE(𝒯, T');
    else
        break;
    end if
end do
get refinement patch R(𝒯, T);
for all T' ∈ R(𝒯, T) do
    {T'_0, T'_1} = BISECT(T');
    𝒯 := 𝒯 \ {T'} ∪ {T'_0, T'_1};
end for
return(𝒯)
```

Lemma 4.4 implies that only elements T' with $g(T') < g(T)$ are not compatibly divisible with T. Hence, recursion is only applied to elements with $g(T') < g(T)$ and thus the maximal depth of recursion is $g(T)$ and recursion terminates. Recursive refinement of an element T' may affect other elements of $R(\mathscr{T};T)$ with same generation $g(T')$. When the do-loop aborts, all elements in the refinement patch $R(\mathscr{T};T)$ are compatibly divisible, and the atomic refinement operation is executed in the for-loop: all elements $T' \in R(\mathscr{T};T)$ are refined, removed from $R(\mathscr{T};T)$, and replaced by the respective children sharing the refinement edge of T. Those children are all of generation $\le g(T) + 1$. Since $\#R(\mathscr{T};T) \le C(\mathscr{T}_0)$, all elements in $R(\mathscr{T};T)$ are of the same generation $g(T)$ after a finite number of iterations. Observe that, except for

T, elements in $R(\mathscr{T};T)$ are only refined to avoid a non-conforming situation. This in summary yields the following result.

Lemma 4.5 (Recursive Refinement). *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 11.1 and let $\mathscr{T} \in \mathbb{T}$ be any conforming refinement.*

Then, for any $T \in \mathscr{T}$ a call of REFINE_RECURSIVE($\mathscr{T}$,T) terminates and outputs the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ where T is bisected. All newly created elements $T' \in \mathscr{T}_* \setminus \mathscr{T}$ satisfy $g(T') \leq g(T)+1$.

Remark 4.2. Assumption 11.1 is a sufficient condition for recursion to terminate but it is not necessary. Such a characterization of recursive bisection is not known. Obviously, termination of the recursion for all elements of $\mathscr{T}_0$ is necessary. Practical experience shows that in 2d this is also sufficient, whereas this is not true in 3d.

We next formulate the algorithm for refining a given conforming grid $\mathscr{T}$ with marked elements $\mathscr{M}$ into a new conforming triangulation:

REFINE($\mathscr{T}$,$\mathscr{M}$)
for all $T \in \mathscr{M} \cap \mathscr{T}$ do
  $\mathscr{T}$:= REFINE_RECURSIVE($\mathscr{T}$,T);
end
return($\mathscr{T}$)

Let T be an element of the input set of marked elements $\mathscr{M}$. Then it may happen that there is an element $T_* \in \mathscr{M}$ scheduled prior to T for refinement and so that the refinement of T_* enforces the refinement of T, for instance $T \in R(\mathscr{T};T_*)$. In the bisection step T is replaced by its two children in $\mathscr{T}$ and thus $T \notin \mathscr{M} \cap \mathscr{T}$. This avoids to refine T twice. In addition, since REFINE_RECURSIVE($\mathscr{T}$,T) outputs the smallest refinement such that T is bisected, REFINE($\mathscr{T}$,$\mathscr{M}$) outputs the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ with $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Remark 4.3 (Iterative vs Recursive Refinement). The iterative and recursive variant of REFINE produce the same output mesh whenever they both terminate. Proposition 4.1 in Sect. 4.5 makes use of the recursive refinement algorithm but is also valid for the iterative variant.

We concluded successful termination of both variants from the fact that the output grid $\mathscr{T}_*$ satisfies $\mathscr{T}_* \leq \mathscr{T}_g$ with g sufficiently large. Therefore, the used arguments do not imply that local refinement stays local. This property is an implication of Theorem 4.3 below.

On a first glance, the iterative variant seems to be easier to implement. But it turns out that handling non-conforming situations can become rather knotty, especially for $d \geq 3$. The implementation of the recursive variant avoids any non-conforming situation by performing the atomic refinement operation, which, as a consequence, simplifies the implementation drastically. The drawback of recursive refinement are stronger assumptions on the distribution of refinement edges on the initial grid.

4.5 Complexity of refinement by bisection

In this section we analyze the cardinality of conforming triangulations produced by adaptive iterations of the form (4). Assuming that a function $\mathsf{REFINE}(\mathscr{T},\mathscr{M})$ outputs the smallest conforming refinement of $\mathscr{T}$ with all elements in $\mathscr{M}$ bisected, we study a sequence of conforming refinements $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \cdots \leq \mathscr{T}_k \leq \ldots$ generated by an iteration of the form

for $k \geq 0$ do
  determine a suitable subset $\mathscr{M}_k \subset \mathscr{T}_k$;
  $\mathscr{T}_{k+1} := \mathsf{REFINE}(\mathscr{T}_k, \mathscr{M}_k)$;
end

The main result is the following theorem.

Theorem 4.3 (Complexity of Refinement by Bisection). *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 11.1.*

Then there exists a constant $\Lambda > 0$ solely depending on $\mathscr{T}_0$, such that for any $K \geq 0$ the conforming triangulation $\mathscr{T}_K$ produced by the above iteration verifies

$$\#\mathscr{T}_K - \#\mathscr{T}_0 \leq \Lambda \sum_{k=0}^{K-1} \#\mathscr{M}_k.$$

The proof of this theorem is split into several steps. Before embarking on it we want to remark that an estimate of the form

$$\#\mathscr{T}_{k+1} - \#\mathscr{T}_k \leq \Lambda \, \#\mathscr{M}_k \tag{59}$$

would imply Theorem 4.3 by summing up over $k = 0, \ldots, K-1$. But such a bound does not hold for refinement by bisection. To see this, consider the initial grid $\mathscr{T}_0$

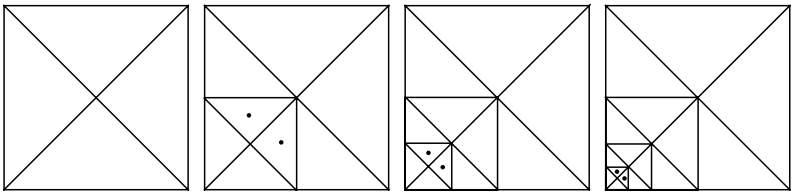


Fig. 9 An example showing that the depth of recursion in is only bounded by the generation of the selected element. Initial triangulation in the leftmost picture and grids $\mathscr{T}_K$ for $K = 2,4,6$. Recursion has depth K for the refinement of the elements marked with bullets

depicted as the leftmost picture in Fig. 9. For all elements the boundary edge is selected as refinement edge and this choice satisfies Assumption 11.1. Pick up any even $K \in \mathbb{N}$ and let

$$\mathscr{M}_k := \{T \in \mathscr{T}_k \mid 0 \in T\} \quad \text{for } k = 0, \dots, K-1$$

and

$$\mathscr{M}_K := \{T \in \mathscr{T}_K \mid g(T) = K \text{ and } 0 \notin T\}.$$

In Fig. 9 we show the grids $\mathscr{T}_K$ for $K = 2,4,6$ and the two elements in $\mathscr{M}_K$ are indicated by a bullet. For $k \leq K$ we only refine marked elements, whence $\#\mathscr{T}_{k+1} - \#\mathscr{T}_k = \#\mathscr{M}_k = 2$ for $k = 0 \dots, K-1$. When refining $\mathscr{T}_K$ into $\mathscr{T}_{K+1}$ we have to recursively refine elements of generation $K-1, K-2, \dots, 0$ for both elements in $\mathscr{M}_K$. From this it is easy to verify that $\#\mathscr{T}_{K+1} - \#\mathscr{T}_K = 4K+2$. Since $\#\mathscr{M}_K = 2$ and K is an arbitrary even number it is obvious that (59) can not hold. On the other hand,

$$\sum_{k=0}^{K} \#\mathscr{T}_{k+1} - \#\mathscr{T}_k = (4K+2) + \sum_{k=0}^{K-1} 2 = 6K+2 \leq 3\,(2K+2) = 3 \sum_{k=0}^{K} \#\mathscr{M}_k.$$

This shows that Theorem 4.3 holds true for this example.

The proof of the theorem can be heuristically motivated as follows. Consider the set $\mathscr{M} := \bigcup_{k=0}^{K-1} \mathscr{M}_k$ used to generate the sequence $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \cdots \leq \mathscr{T}_K =: \mathscr{T}$. Suppose that each element $T_* \in \mathscr{M}$ is assigned a fixed amount C_1 of money to spend on refined elements in $\mathscr{T}$, i. e., on $T \in \mathscr{T} \setminus \mathscr{T}_0$. Assume further that $\lambda(T, T_*)$ is the portion of money spent by T_* on T. Then it must hold

$$\sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \leq C_1 \qquad \text{for all } T_* \in \mathscr{M}. \tag{60a}$$

In addition, we suppose that the investment of all elements in $\mathscr{M}$ is fair in the sense that each $T \in \mathscr{T} \setminus \mathscr{T}_0$ gets at least a fixed amount C_2, whence

$$\sum_{T_* \in \mathscr{M}} \lambda(T, T_*) \geq C_2 \qquad \text{for all } T \in \mathscr{T} \setminus \mathscr{T}_0. \tag{60b}$$

Therefore, summing up (60b) and using the upper bound (60a) we readily obtain

$$C_2(\#\mathscr{T} - \#\mathscr{T}_0) \leq \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \sum_{T_* \in \mathscr{M}} \lambda(T, T_*) = \sum_{T_* \in \mathscr{M}} \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \leq C_1 \#\mathscr{M},$$

which proves the theorem for $\mathscr{T}$ and $\mathscr{M}$. In the remainder of this section we design such an allocation function $\lambda \colon \mathscr{T} \times \mathscr{M} \to \mathbb{R}^+$ in several steps and prove that recurrent refinement by bisection yields (60) provided $\mathscr{T}_0$ satisfies Assumption 11.1.

In view of (54), measure and diameter of an element are related to its generation:

$$D_1 2^{-g(T)} \leq |T| \quad \text{and} \quad \operatorname{diam}(T) \leq D_2 2^{-g(T)/d} \qquad \text{for all } T \in \mathbb{F}, \tag{61}$$

with $D_1 = \min\{|T_0| \mid T_0 \in \mathscr{T}_0\}$ and $D_2 \approx \max\{|T_0| \mid T_0 \in \mathscr{T}_0\}$. The constant hidden in $\approx$ solely depends on the shape regularity of $\mathbb{F}$ (and thus on $\mathscr{T}_0$).

Suppose now that T' is generated by REFINE_RECURSIVE($\mathscr{T}$,T). The constant D_2 enables us to relate the distance of T' to T with its generation $g(T')$, where

$$\operatorname{dist}(T,T') = \inf_{x\in T, x'\in T'} |x-x'|.$$

Proposition 4.1 (Distance and Generation). *Let $\mathscr{T}\in\mathbb{T}$, $T\in\mathscr{T}$ and assume that T' is created by* REFINE_RECURSIVE($\mathscr{T}$,T)*. Then there holds*

$$\operatorname{dist}(T,T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d} < D_2 \frac{2^{1/d}}{1-2^{-1/d}}\, 2^{-g(T')/d}.$$

Proof. We prove $\operatorname{dist}(T,T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d}$ by induction over the generation of T. The rightmost inequality is a direct consequence of the geometric sum.

1 If $g(T)=0$, then the refinement patch $R(\mathscr{T};T)$ is compatibly divisible thanks to Lemma 4.4. Consequently REFINE_RECURSIVE($\mathscr{T}$,T) only creates elements T' with $\operatorname{dist}(T,T')=0$ and the assertion follows trivially.

2 Let now $g(T)>0$ and assume that the assertion holds for any $T''\in\mathscr{T}$ with $0\le g(T'')<g(T)$. We only need to consider $\operatorname{dist}(T,T')>0$, whence T' is created by a recursive call REFINE_RECURSIVE($\mathscr{T}$,T'') for an element $T''\in R(\mathscr{T};T)$ that is not compatibly divisible with T; thus $g(T'')<g(T)$ by Lemma 4.4. The induction hypothesis yields

$$\operatorname{dist}(T'',T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')} 2^{-g/d}.$$

Since $T''\in R(\mathscr{T},T)$, and so T'' contains the refinement edge of T, we realize that $\operatorname{dist}(T'',T)=0$. Combining the last estimate with (61), we deduce

$$\begin{aligned}\operatorname{dist}(T,T') &\le \operatorname{dist}(T'',T') + \operatorname{diam}(T'') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')} 2^{-g/d} + D_2\, 2^{-g(T'')/d} \\ &= D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')+1} 2^{-g/d} \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d},\end{aligned}$$

where we have used $g(T'')<g(T)$ in the last step. This finishes the proof. □

We next construct the allocation function λ. The construction is based on two sequences $\{a(\ell)\}_{\ell=-1}^{\infty}, \{b(\ell)\}_{\ell=0}^{\infty} \subset \mathbb{R}^+$ of positive numbers satisfying

$$\sum_{\ell\ge -1} a(\ell) = A < \infty, \qquad \sum_{\ell\ge 0} 2^{-\ell/d}\, b(\ell) = B < \infty, \qquad \inf_{\ell\ge 1} b(\ell)\, a(\ell) = c_* > 0,$$

and $b(0)\ge 1$. Valid instances are $a(\ell)=(\ell+2)^{-2}$ and $b(\ell)=2^{\ell/(d+1)}$.

With these settings we are prepared to define $\lambda : \mathscr{T}\times\mathscr{M} \to \mathbb{R}^+$ by

$$\lambda(T,T_*) := \begin{cases} a(g(T_*)-g(T)), & \operatorname{dist}(T,T_*) < D_3 B\, 2^{-g(T)/d} \text{ and } g(T)\le g(T_*)+1 \\ 0, & \text{else,} \end{cases}$$

where $D_3 := D_2\big(1+2^{1/d}(1-2^{-1/d})^{-1}\big)$. Therefore, the investment of money by $T_* \in \mathscr{M}$ is restricted to cells T that are sufficiently close and are of generation $g(T) \le g(T_*)+1$. Only elements of such generation can be created during refinement of T_* according to Lemma 4.4.

The following lemma shows that the total amount of money spend by this allocation function per marked element is bounded.

Lemma 4.6 (Upper Bound). *There exists a constant $C_1 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies (60a), i. e.,*

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) \le C_1 \qquad \text{for all } T_* \in \mathscr{M}.$$

Proof. 1 Given $T_* \in \mathscr{M}$ we set $g_* = g(T_*)$ and we let $0 \le g \le g_*+1$ be a generation of interest in the definition of λ. We claim that for such g the cardinality of the set

$$\mathscr{T}(T_*,g) = \{T \in \mathscr{T} \mid \operatorname{dist}(T,T_*) < D_3 B 2^{-g/d} \text{ and } g(T) = g\}$$

is uniformly bounded, i. e., $\#\mathscr{T}(T_*,g) \le C$ with C solely depending on D_1, D_2, D_3, B.

From (61) we learn that $\operatorname{diam}(T_*) \le D_2 2^{-g_*/d} \le 2D_2 2^{-(g_*+1)/d} \le 2D_2 2^{-g/d}$ as well as $\operatorname{diam}(T) \le D_2 2^{-g/d}$ for any $T \in \mathscr{T}(T_*,g)$. Hence, all elements of the set $\mathscr{T}(T_*,g)$ lie inside a ball centered at the barycenter of T_* with radius $(D_3B+3D_2)2^{-g/d}$. Again relying on (61) we thus conclude

$$\#\mathscr{T}(T_*,g) D_1 2^{-g} \le \sum_{T\in\mathscr{T}(T_*,g)} |T| \le c(d)(D_3B+3D_2)^d 2^{-g},$$

whence $\#\mathscr{T}(T_*,g) \le c(d)\, D_1^{-1}\,(D_3B+3D_2)^d =: C$.

2 Accounting only for non-zero contributions $\lambda(T,T_*)$ we deduce

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) = \sum_{g=0}^{g_*+1} \sum_{T\in\mathscr{T}(T_*,g)} a(g_*-g) \le C \sum_{\ell=-1}^{\infty} a(\ell) = CA =: C_1,$$

which is the desired upper bound. □

The definition of λ also implies that each refined element receives a fixed amount of money.

Lemma 4.7 (Lower Bound). *There exists a constant $C_2 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies (60b), i. e.,*

$$\sum_{T_*\in\mathscr{M}} \lambda(T,T_*) \ge C_2 \qquad \text{for all } T \in \mathscr{T}\setminus\mathscr{T}_0.$$

Proof. 1 Fix an arbitrary $T_0 \in \mathscr{T}\setminus\mathscr{T}_0$. Then there is an iteration count $1 \le k_0 \le K$ such that $T_0 \in \mathscr{T}_{k_0}$ and $T_0 \notin \mathscr{T}_{k_0-1}$. Therefore there exists an $T_1 \in \mathscr{M}_{k_0-1} \subset \mathscr{M}$ such that T_0 is generated during REFINE_RECURSIVE$(\mathscr{T}_{k_0-1}, T_1)$. Iterating this

process we construct a sequence $\{T_j\}_{j=1}^J \subset \mathscr{M}$ with corresponding iteration counts $\{k_j\}_{j=1}^J$ such that T_j is created by REFINE_RECURSIVE$(\mathscr{T}_{k_j-1}, T_{j+1})$. The sequence is finite since the iteration counts are strictly decreasing and thus $k_J = 0$ for some $J > 0$, or equivalently $T_J \in \mathscr{T}_0$.

Since T_j is created during refinement of T_{j+1} we infer from Lemma 4.5 that

$$g(T_{j+1}) \geq g(T_j) - 1.$$

Accordingly, $g(T_{j+1})$ can decrease the previous value of $g(T_j)$ at most by 1. Since $g(T_J) = 0$ there exists a smallest value s such that $g(T_s) = g(T_0) - 1$. Note that for $j = 1, \dots, s$ we have $\lambda(T_0, T_j) > 0$ if $\operatorname{dist}(T_0, T_j) \leq D_3 B g^{-g(T_0)/d}$.

2 We next estimate the distance $\operatorname{dist}(T_0, T_j)$. For $1 \leq j \leq s$ and $\ell \geq 0$ we define the set

$$\mathscr{T}(T_0, \ell, j) := \{T \in \{T_0, \dots, T_{j-1}\} \mid g(T) = g(T_0) + \ell\}$$

and denote by $m(\ell, j)$ its cardinality. The triangle inequality combined with an induction argument yields

$$\begin{aligned} \operatorname{dist}(T_0, T_j) &\leq \operatorname{dist}(T_0, T_1) + \operatorname{diam}(T_1) + \operatorname{dist}(T_1, T_j) \\ &\leq \sum_{i=1}^{j} \operatorname{dist}(T_{i-1}, T_i) + \sum_{i=1}^{j-1} \operatorname{diam}(T_i). \end{aligned}$$

We apply Proposition 4.1 for the terms of the first sum and (61) for the terms of the second sum to obtain

$$\begin{aligned} \operatorname{dist}(T_0, T_j) &< D_2 \frac{2^{1/d}}{1 - 2^{-1/d}} \sum_{i=1}^{j} 2^{-g(T_{i-1})/d} + D_2 \sum_{i=1}^{j-1} 2^{-g(T_i)/d} \\ &= D_2 \left(1 + \frac{2^{1/d}}{1 - 2^{-1/d}}\right) \sum_{i=0}^{j-1} 2^{-g(T_i)/d} \\ &= D_3 \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-(g(T_0)+\ell)/d} \\ &= D_3 2^{-g(T_0)/d} \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-\ell/d}. \end{aligned}$$

For establishing the lower bound we distinguish two cases depending on the size of $m(\ell, s)$. This is done next.

3 *Case 1:* $m(\ell, s) \leq b(\ell)$ for all $\ell \geq 0$. From this we conclude

$$\operatorname{dist}(T_0, T_s) < D_3 2^{-g(T_0)/d} \sum_{\ell=0}^{\infty} b(\ell)\, 2^{-\ell/d} = D_3 B 2^{-g(T_0)/d}$$

and the definition of λ then readily implies

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_s) = a(g(T_s)-g(T_0)) = a(-1) > 0.$$

4 *Case 2:* There exists $\ell \geq 0$ such that $m(\ell,s) > b(\ell)$. For each of these ℓ's there exists a smallest $j = j(\ell)$ such that $m(\ell,j(\ell)) > b(\ell)$. We let ℓ^* be the index ℓ that gives rise to the smallest $j(\ell)$, and set $j^* = j(\ell^*)$. Consequently

$$m(\ell,j^*-1) \leq b(\ell) \quad \text{for all } \ell \geq 0 \qquad \text{and} \qquad m(\ell^*,j^*) > b(\ell^*).$$

As in Case 1 we see $\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/d}$ for all $i \leq j^*-1$, or equivalently

$$\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/d} \qquad \text{for all } T_i \in \mathscr{T}(T_0,\ell,j^*).$$

We next show that the elements in $\mathscr{T}(T_0,\ell^*,j^*)$ spend enough money on T_0. We first consider $\ell^* = 0$ and note that $T_0 \in \mathscr{T}(T_0,0,j^*)$. Since $m(0,j^*) > b(0) \geq 1$ we discover $j^* \geq 2$. Hence, there is an $T_i \in \mathscr{T}(T_0,0,j^*) \cap \mathscr{M}$, which yields the estimate

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_i) = a(g(T_i)-g(T_0)) = a(0) > 0.$$

For $\ell^* > 0$ we see that $T_0 \notin \mathscr{T}(T_0,\ell^*,j^*)$, whence $\mathscr{T}(T_0,\ell^*,j^*) \subset \mathscr{M}$. In addition, $\lambda(T_0,T_i) = a(\ell^*)$ for all $T_i \in \mathscr{T}(T_0,\ell^*,j^*)$. From this we conclude

$$\begin{aligned}\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) &\geq \sum_{T_*\in\mathscr{T}(T_0,\ell^*,j^*)} \lambda(T_0,T_*) = m(\ell^*,j^*)\,a(\ell^*) \\ &> b(\ell^*)\,a(\ell^*) \geq \inf_{\ell\geq 1} b(\ell)\,a(\ell) = c_* > 0.\end{aligned}$$

5 In summary we have proved the assertion since for any $T_0 \in \mathscr{T} \setminus \mathscr{T}_0$

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \min\{a(-1),a(0),c_*\} =: C_2 > 0. \qquad \square$$

Lemmas 4.6 and 4.7 show that the allocation function λ satisfies (60), which implies Theorem 4.3.

Remark 4.4 (Several Bisections). In practice, one often likes to bisect selected elements several times, for instance each marked element is scheduled for $b \geq 1$ bisections. This can be done by assigning the number $b(T) = b$ of bisections that have to be executed for each marked element T. If T is bisected then we assign $(b(T)-1)$ as the number of pending bisections to its children and the set of marked elements is $\mathscr{M} := \{T \in \mathscr{T} \mid b(T) > 0\}$.

To show the complexity estimate when REFINE performs $b > 1$ bisections, the set $\mathscr{M}_k$ is to be understood as a sequence of *single* bisections recorded in sets $\{\mathscr{M}_k(j)\}_{j=1}^b$, which belong to intermediate triangulations between $\mathscr{T}_k$ and $\mathscr{T}_{k+1}$ with $\#\mathscr{M}_k(j) \leq 2^{j-1}\#\mathscr{M}_k$, $j = 1,\dots,b$. Then we also obtain Theorem 4.3 because

$$\sum_{j=1}^{b} \#\mathscr{M}_k(j) \le \sum_{j=1}^{b} 2^{j-1}\#\mathscr{M}_k = (2^b-1)\#\mathscr{M}_k.$$

Remark 4.5 (Optimal Constant). Trying to trace the value of the constant Λ one realizes that Λ becomes rather large since it depends via C_1 and C_2 on the constants A, B, c_*. Experiments suggest that $\Lambda \approx 14$ in 2d and $\Lambda \approx 180$ in 3d when $\mathscr{T}_0$ is the initial triangulation of the d-dimension cube $(0,1)^d$ build from the $d!$ Kuhn-simplices of type 0. According to Problem 4.3, $\mathscr{T}_0$ satisfies Assumption 11.1.

There is an interesting connection to a result by Atalay and Mount that can be formulated as follows [3]: there exists a constant $C_d \le 3^d d!$ such that the smallest conforming refinement $\mathscr{T}_* \in \mathbb{T}$ of any non-conforming refinement $\mathscr{T}$ of $\mathscr{T}_0$ satisfies

$$\#\mathscr{T}_* \le C_d \#\mathscr{T}.$$

For 2d the optimal constant is shown to be $C_2 = 14$ and the constant $C_3 = 162$ for $d = 3$ is quite close to the constant observed in experiments. The agreement between theory and experiments for 2d is quite exiting, but nevertheless the estimate by Atalay and Mount cannot be used to show Theorem 4.3.

4.6 Problems

Problem 4.1. Show that after d recurrent bisections of a simplex T all edges of T are bisected exactly once. To this end, let first $T = \{z_0, \dots, z_d\}_0$ be of type 0 and show by induction that any sub-simplex T' of T with generation $t = g < d$ has the structure

$$T' = \left\{ z_{k_0}, \bar{z}_t, \bar{z}_{t-1}, \dots, \bar{z}_1, z_{k_1}, z_{k_2}, \dots, z_{k_{d-t}} \right\}_t,$$

where $\bar{z}_i$ are the new vertices of the bisection step i, $i = 1, \dots, t$, and $k_0, \dots, k_{d-t}$ are consecutive natural numbers, for instance $0, 1, 2, \dots, d-1$ or $d, d-1, \dots, 1$ for $t = 1$. Then generalize the claim to a simplex T of type $t \in \{0, \dots, d-1\}$.

Problem 4.2. Show that the output of $\mathsf{BISECT}(T)$ and $\mathsf{BISECT}(T_R)$ is the same, i. e., the children of T and its reflected element T_R are identical.

Problem 4.3. Show that the set of the $d!$ Kuhn-simplices of type 0 is a conforming triangulation of the unit cube $(0,1)^d \subset \mathbb{R}^d$ satisfying Assumption 11.1.

Problem 4.4. Let $d = 2$ and $T = \{z_0, z_1, z_2\}_t, T' = \{z'_0, z'_1, z'_2\}_t$ be neighboring elements with common side $S = T \cap T'$. Show that

(a) T and T' are reflected neighbors if and only if $\overline{z_0 z_2} = \overline{z'_0 z'_2}$ or $z_1 = z'_1$.
(b) If T and T' are reflected neighbors, then so are their neighboring children.
(c) If $z_1 = z'_2$ and $z_2 = z'_1$, then T and T' are not reflected neighbors but their neighboring children are.
(d) If $S = \overline{z_0 z_2} = \overline{z'_0 z'_2}$ or $z_1, z'_1 \in S$, then T and T' are matching neighbors.

5 Piecewise polynomial approximation

The numerical solution of a boundary value problem may be seen as a special approximation problem where the target function is not given explicitly but implicitly. Theorem 3.2 shows that the error of a Petrov-Galerkin solution of a stable discretization is dictated by the best approximation from the discrete space. In this chapter we investigate approximation properties of continuous piecewise polynomials, the standard discretization for the model problem in Sect. 2.2.1. We do not strive for completeness but rather want to provide some background and motivation for the successive chapters. To this end, we depart from classical finite element approximation and end up with a result on nonlinear or adaptive approximation.

For more information about nonlinear and constructive approximation, we refer to the survey [28] and the book [29].

5.1 Quasi-interpolation

We start with a brief discussion on piecewise polynomial interpolation of rough functions, namely those without point values as we expect H^1-functions to be. This leads to the concept of quasi-interpolation and to a priori error estimates for the standard discretization of our model problem in Sect. 2.2.1.

Using the Lagrange basis $\{\phi_z\}_{z\in\mathcal{N}_n(\mathcal{T})} \subset S^{n,0}(\mathcal{T})$ from Theorem 3.4 we have for any $v \in S^{n,0}(\mathcal{T})$ the representation $v = \sum_{z\in\mathcal{N}_n(\mathcal{T})} v(z)\phi_z$. This may suggest to use for given v the *Lagrange interpolant*

$$I_{\mathcal{T}} v(x) := \sum_{z\in\mathcal{N}_n(\mathcal{T})} v(z)\phi_z(x). \tag{62}$$

However, this operator requires that point values of v are well-defined. If $v \in W_p^s(\Omega)$, this entails the condition $\mathrm{sob}(W_p^s) > 0$, which in turn requires regularity beyond the trial space $H_0^1(\Omega)$ when $d \geq 2$.

Quasi-interpolants, like those in Clément [26] or Scott-Zhang [65], replace $v(z)$ in (62) by a suitable local average and so are well-defined also for rough functions, e.g. from $H_0^1(\Omega)$. For any conforming refinement $\mathcal{T} \geq \mathcal{T}_0$ of $\mathcal{T}_0$, the averaging process extends beyond nodes and so brings up the discrete neigborhood

$$N_{\mathcal{T}}(T) := \{T' \in \mathcal{T} \mid T' \cap T \neq \emptyset\}$$

for each element $T \in \mathcal{T}$ along with the uniform properties (55), namely,

$$\max_{T\in\mathcal{T}} \#N_{\mathcal{T}}(T) \leq C(\mathcal{T}_0), \qquad \max_{T'\in N_{\mathcal{T}}(T)} \frac{|T|}{|T'|} \leq C(\mathcal{T}_0),$$

where $C(\mathcal{T}_0)$ depends only on the shape coefficient of $\mathcal{T}_0$. We shall make use of the following estimate of the local interpolation error; see [16, 65].

Proposition 5.1 (Local Error Estimate for Quasi-Interpolant). *Let s be the regularity index with $0 \le s \le n+1$, and $1 \le p \le \infty$ be the integrability index.*
(a) There exists an operator $I_{\mathscr{T}} : L^1(\Omega) \to S^{n,0}(\mathscr{T})$ such that for all $T \in \mathscr{T}$ we have

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W_p^s)-\mathrm{sob}(W_q^t)} \|D^s v\|_{L^p(N_{\mathscr{T}}(T))} \tag{63}$$

where $0 \le t \le s$, $1 \le q \le \infty$ are such that $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$. The hidden constant depends on the shape coefficient of $\mathscr{T}_0$ and d.
(b) There exists an operator $I_{\mathscr{T}} : W_1^1(\Omega) \to S^{n,0}(\mathscr{T})$ satisfying (63) for $s \ge 1$ and, in addition, if $v \in W_1^1(\Omega)$ has a vanishing trace on $\partial\Omega$, then so does $I_{\mathscr{T}} v$.

Both operators are invariant in $S^{n,0}(\mathscr{T})$, namely $I_{\mathscr{T}} V = V$ for all $V \in S^{n,0}(\mathscr{T})$.

Proof. We sketch the proof; see [16, 65] for details. Recall that $\{\phi_z\}_{z \in \mathscr{N}_n(\mathscr{T})}$ is the global Lagrange basis of $S^{n,0}(\mathscr{T})$ and $\{\phi_z^*\}_{z \in \mathscr{N}_n(\mathscr{T})}$ is the global dual basis and, according to Remark 3.4, $\operatorname{supp} \phi_z^* = \operatorname{supp} \phi_z$ for all $z \in \mathscr{N}_n(\mathscr{T})$. We thus define $I_{\mathscr{T}} : L^1(\Omega) \to S^{n,0}(\mathscr{T})$ to be

$$I_{\mathscr{T}} v = \sum_{z \in \mathscr{N}_n(\mathscr{T})} \langle v, \phi_z^* \rangle \phi_z,$$

and observe that by construction this operator is invariant in $S^{n,0}(\mathscr{T})$, namely,

$$I_{\mathscr{T}} P = P \qquad \text{for all } P \in S^{n,0}(\mathscr{T}).$$

In particular, the averaging process giving rise to the values of $I_{\mathscr{T}} v$ for each element $T \in \mathscr{T}$ takes place in the neighborhood $N_{\mathscr{T}}(T)$, whence we also deduce the local invariance

$$I_{\mathscr{T}} P|_T = P \qquad \text{for all } P \in \mathbb{P}_n(N_{\mathscr{T}}(T))$$

as well as the local stability estimate

$$\|I_{\mathscr{T}} v\|_{L^q(T)} \lesssim \|v\|_{L^q(N_{\mathscr{T}}(T))}.$$

We thus may write

$$v - I_{\mathscr{T}} v|_T = (v - P) - I_{\mathscr{T}}(v - P)|_T \qquad \text{for all } T \in \mathscr{T},$$

where $P \in \mathbb{P}_{s-1}$ is arbitrary. It suffices now to prove (63) in the reference element $\hat{T}$ and scale back and forth via Lemma 3.1; the definition (5) of Sobolev number accounts precisely for this scaling. We keep the notation T for $\hat{T}$, apply the inverse estimate for $\mathbb{P}_n$-polynomials $\|D^t(I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|I_{\mathscr{T}} v\|_{L^q(T)}$ to $v - P$ instead of v, and use the above local stability estimate, to infer that

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|v - P\|_{W_q^t(N_{\mathscr{T}}(T))} \lesssim \|v - P\|_{W_p^s(N_{\mathscr{T}}(T))}.$$

The last inequality is a consequence $W_p^s(N_{\mathscr{T}}(T)) \subset W_q^t(N_{\mathscr{T}}(T))$ because $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$. Estimate (63) now follows from the Bramble-Hilbert lemma [16, Lemma 4.3.8], [25, Theorem 3.1.1]

$$\inf_{P\in\mathbb{P}_{s-1}(N_{\mathscr{T}}(T))} \|v-P\|_{W_p^s(N_{\mathscr{T}}(T))} \lesssim \|D^s v\|_{L^p(N_{\mathscr{T}}(T))}. \tag{64}$$

This proves (a). To show (b) we modify the averaging process for boundary nodes and define a set of dual functions with respect to an L^2-scalar product over $(d-1)$-subsimplices contained on $\partial\Omega$; see again [16, 65] for details. This retains the invariance property of $I_{\mathscr{T}}$ on $S^{n,0}(\mathscr{T})$ and guarantees that $I_{\mathscr{T}}v$ has a zero trace if $v\in W_1^1(\Omega)$ does. Hence, the same argument as above applies and (63) follows. □

Remark 5.1 (Sobolev Numbers). We cannot expect (63) to be valid if $\mathrm{sob}(W_p^s) = \mathrm{sob}(W_q^t)$ since this may not imply $W_p^s(\Omega)\subset W_q^t(\Omega)$; recall the counterexample $W_p^s(\Omega)=W_d^1(\Omega)$ and $W_q^t(\Omega)=L^\infty(\Omega)$ of Sect. 2.1.1. However, equality of Sobolev numbers is allowed in (63) as long as the space embedding is valid.

Remark 5.2 (Fractional Regularity). We observe that (63) does not require the regularity indices t and s to be integer. The proof follows the same lines but replaces the polynomial degree $s-1$ by the greatest integer smaller that s; the generalization of (64) can be taken from [33].

Remark 5.3 (Local Error Estimate for Lagrange Interpolant). Let the regularity index s and integrability index $1\le p\le\infty$ satisfy $s-d/p>0$. This implies that $\mathrm{sob}(W_p^s)>\mathrm{sob}(L^\infty)$, whence $W_p^s(\Omega)\subset C(\overline{\Omega})$ and the Lagrange interpolation operator $I_{\mathscr{T}}: W_p^s(\Omega)\to S^{n,0}(\mathscr{T})$ is well defined and satisfies the fully local error estimate

$$\|D^t(v-I_{\mathscr{T}}v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W_p^s)-\mathrm{sob}(W_q^t)} \|D^s v\|_{L^p(T)}, \tag{65}$$

provided $0\le t\le s$, $1\le q\le\infty$ are such that $\mathrm{sob}(W_p^s)>\mathrm{sob}(W_q^t)$. We point out that $N_{\mathscr{T}}(T)$ in (63) is now replaced by T in (65). We also remark that if v vanishes on $\partial\Omega$ so does $I_{\mathscr{T}}v$. The proof of (65) proceeds along the same lines as that of Proposition 5.1 except that the nodal evaluation does not extend beyond the element $T\in\mathscr{T}$ and the inverse and stability estimates over the reference element are replaced by

$$\|D^t I_{\mathscr{T}}v\|_{L^q(\hat{T})} \lesssim \|I_{\mathscr{T}}v\|_{L^q(\hat{T})} \lesssim \|v\|_{L^\infty(\hat{T})} \lesssim \|v\|_{W_p^s(\hat{T})}.$$

Remark 5.4 (Boundary values). The procedure described at the end of the proof of Proposition 5.1 can be used to interpolate functions with boundary values different from zero while retaining invariance over the finite element space. We refer to [16, 65] for details.

Remark 5.5 (Localized Estimate). Suppose that $v\in W_1^1(\Omega)$ happens to be a piecewise polynomial of degree $\le n$ on a subdomain Ω_* of Ω. Let ω be a connected component of $\Omega\backslash\Omega_*$ and let the quasi-interpolant $I_{\mathscr{T}}v$ preserve the boundary values of v on $\partial\omega$, as indicated in Remark 5.4. If we repeat this construction for each connected component ω of $\Omega\backslash\Omega_*$ and define $I_{\mathscr{T}}v=v$ in Ω_*, then $I_{\mathscr{T}}v\in S^{n,0}(\mathscr{T})$ and we deduce the localized estimate for all $1\le p\le\infty$

$$\sum_{T\subset\omega} h_T^{-2}\|v-I_{\mathscr{T}}v\|_{L^p(T)}^p + h_T^{-2+2/p}\|v-I_{\mathscr{T}}v\|_{L^p(\partial T)}^p \lesssim \|\nabla v\|_{L^p(\omega)}^p. \tag{66}$$

This property will be crucial in Chap. 9 to prove quasi-optimality of AFEM.

The local interpolation error estimate in Proposition 5.1 implies a global one. The latter will be discussed as an upper bound for the error of the finite element solution in the next section.

5.2 A priori error analysis

Combining Theorem 3.2 with Proposition 5.1 we derive a so-called *a priori error estimate*, which bounds the error of the finite element solution in terms of the mesh-size function and regularity of the exact solution beyond $H^1(\Omega)$. We present a slightly more general variant than usual. This will help in the successive discussion on error reduction.

Theorem 5.1 (A Priori Error Estimate). *Let $1 \le s \le n+1, 1 \le p \le 2$, and let the solution u of the model problem (13) satisfy $u \in W_p^s(\Omega)$ with $r := \mathrm{sob}(W_p^s) - \mathrm{sob}(H^1) > 0$. Let $U \in \mathbb{V}(\mathscr{T}) = S^{n,0}(\mathscr{T}) \cap H_0^1(\Omega)$ be the corresponding discrete solution. If $h : \Omega \to \mathbb{R}$ denotes the piecewise constant mesh density function, then*

$$\|\nabla(u-U)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} \|h^r D^s u\|_{L^p(\Omega)}. \tag{67}$$

The hidden constant depends on shape coefficient of $\mathscr{T}_0$ and the dimension d.

Proof. Theorem 3.2 and Proposition 5.1 yield

$$\|\nabla(u-U)\|_{L^2(\Omega)}^2 \lesssim \frac{\alpha_2}{\alpha_1} \|\nabla(u-I_{\mathscr{T}} u)\|_{L^2(\Omega)}^2 \lesssim \frac{\alpha_2}{\alpha_1} \sum_{T \in \mathscr{T}} h_T^{2r} \|D^s u\|_{L^p(N_{\mathscr{T}}(T))}^2.$$

In order to sum up the right-hand side we need to accumulate in ℓ^p rather than ℓ^2. We recall the elementary property of series $\sum_n a_n \le (\sum_n a_n^q)^{1/q}$ for $0 < q \le 1$. We take $q = p/2$ and apply this property, in conjunction with (55), to arrive at

$$|u - I_{\mathscr{T}} u|_{H^1(\Omega)}^2 \lesssim \left(\sum_{T \in \mathscr{T}} h_T^{rp} \|D^s u\|_{L^p(N_{\mathscr{T}}(T))}^p \right)^{\frac{2}{p}} \lesssim \left(\int_\Omega h(x)^{rp} \, |D^s u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

This is the asserted estimate (67). □

Notice that in Theorem 5.1 the exploitable number of derivatives of the exact solution is limited by the polynomial degree

$$1 \le s \le 1 + n.$$

Moreover, decreasing the mesh-size function reduces the upper bound (67). The reduction rate is dictated by the difference of the Sobolev numbers

$$r = \mathrm{sob}(W_p^s) - \mathrm{sob}(H^1),$$

and is thus sensitive to the integrability of the relevant derivatives in both left and right-hand sides of (67). The best rate is obtained for integrability index $p = 2$, which coincides with the integrability of the error notion.

Relying solely on decreasing of the mesh-size function, and thus ignoring the local distribution of the derivative $D^s u$ of the exact solution u, leads to uniform refinement or quasi-uniform meshes. The specialization of Theorem 5.1 to this case reads as follows:

Corollary 5.1 (Quasi-Uniform Meshes). *Let $1 \le s \le n+1$, and let the solution u of the model problem (13) satisfy $u \in H^s(\Omega)$. Let $\mathscr{T}_N$ be a quasi-uniform partition of Ω with N interior nodes and let $U_N \in \mathbb{V}(\mathscr{T}_N)$ be the discrete solution corresponding to the model problem (13). Then*

$$\|\nabla(u - U_N)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} |u|_{H^s(\Omega)} N^{-(s-1)/d}. \tag{68}$$

Proof. Quasi-uniformity of $\mathscr{T}_N$ implies

$$\max_{T\in\mathscr{T}_N} h_T^d \le \max_{T\in\mathscr{T}_N} \overline{h}_T^d \lesssim \min_{T\in\mathscr{T}_N} \underline{h}_T{}^d \le \frac{1}{N} \sum_{T\in\mathscr{T}_N} h_T^d = \frac{|\Omega|}{N}$$

Since $r = (s - d/2) - (1 - d/2) = s - 1$, the assertion follows (67). □

A simple consequence of (68), under full regularity $u \in H^{n+1}(\Omega)$ is the maximal decay rate in terms of degrees of freedom

$$\|\nabla(u - U_N)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} |u|_{H^{n+1}(\Omega)} N^{-n/d}. \tag{69}$$

One may wonder whether (68) is sharp whenever $s < n+1$. The following example addresses this question.

Example 5.1 (Corner Singularity). We consider the Dirichlet problem for $-\Delta u = f$, for which $\alpha_1 = \alpha_2 = 1$, with exact solution (in polar coordinates)

$$u(r,\theta) = r^{\frac{2}{3}} \sin(2\theta/3) - r^2/4,$$

on an L-shaped domain Ω; this function satisfies $u \in H^s(\Omega)$ for $s < 5/3$. Recall that even though s is fractional, the error estimates are still valid; see Remark 5.2. In particular, (68) can be derived by space interpolation between $H^1(\Omega)$ and $H^{n+1}(\Omega)$. In Figure 5.1 we depict the sequence of *uniform* meshes, for which $N \approx h^{-2}$, h being the mesh-size. In Table 1 we report the order of convergence for polynomial degrees $n = 1, 2, 3$. The asymptotic rate is about $h^{2/3}$, or equivalently $N^{-1/3}$, regardless of n and is consistent with the estimate (68). This indicates that (68) is sharp.

The question arises whether the rate $N^{-1/3}$ in Example 5.1 is just a consequence of uniform refinement or unavoidable. It is important to realize that $u \notin H^s(\Omega)$ for

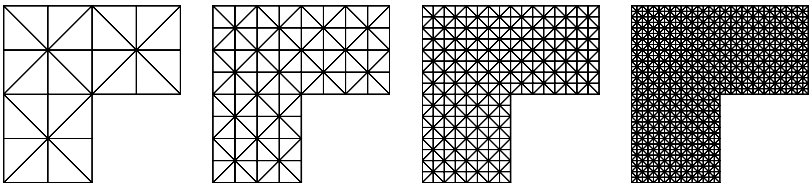


Fig. 10 Sequence of uniform meshes for L-shaped domain Ω

h	linear $(n=1)$	quadratic $(n=2)$	cubic $(n=3)$
1/4	1.14	9.64	9.89
1/8	0.74	0.67	0.67
1/16	0.68	0.67	0.67
1/32	0.66	0.67	0.67
1/64	0.66	0.67	0.67
1/128	0.66	0.67	0.67

Table 1 The asymptotic rate of convergence is about $h^{2/3}$, or equivalently $N^{-1/3}$, irrespective of the polynomial degree n as predicted by (68)

$s \geq 5/3$ and thus (68) is not applicable. However, the problem is not that second order derivatives of u do not exist but rather that they are not square-integrable. In particular, it is true that $u \in W_p^2(\Omega)$ if $1 \leq p < 3/2$. We therefore may apply Theorem 5.1 with, e.g., $n = 1$, $s = 2$, and $p \in [1, 3/2)$ and then ask whether the structure of (67) can be exploited, e.g., by compensating the local behavior of $D^s u$ with the local mesh-size h. If u is assumed to be known, this enterprise naturally leads to meshes *adapted to* u that may be *graded*. We discuss this possibility in Sect. 5.3 and propose a condition that should be satisfied by these meshes.

5.3 Principle of error equidistribution

For the model problem and its standard discretization, Theorem 3.2 and the considerations at the end of Sect. 5.2 suggest the optimization problem:

> *Given a function $u \in H^1(\Omega)$ and an integer $N > 0$ find conditions for a shape regular mesh $\mathscr{T}$ to minimize the error $|u - I_{\mathscr{T}} u|_{H^1(\Omega)}$ subject to the constraint that the number of degrees of freedom does not exceed N.*

In the framework of Chap. 4 this becomes a *discrete* optimization problem. Here we consider a simplified setting and, similar to Babuška and Rheinboldt [5], invoke a *continuous model*:

- The dimension is $d = 2$ and the regularity of $u \in C^2(\Omega) \cap W_p^2(\Omega)$ with $1 < p \leq 2$;
- There exists a C^1 function $h : \Omega \to \mathbb{R}$, a mesh density function, with the property that $h(x)$ is equivalent to h_T for all $T \in \mathscr{T}$ with equivalence constants only depending on shape regularity (thus on the shape coefficient of $\mathscr{T}_0$);

- The number of degrees of freedom and local mesh-size are related through the relation
$$N = \int_\Omega \frac{dx}{h(x)^2}.$$
- The mesh $\mathscr{T}$ is sufficiently fine so that D^2u is essentially constant within each element $T \in \mathscr{T}$;
- The error is given by the formula
$$\left(\int_\Omega h(x)^{2(p-1)} |D^2u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

A few comments about this model are in order. The first condition is motivated by the subsequent discussion and avoids dealing with Besov spaces with integrability index $p < 1$; in particular, all corner singularities for $d = 2$ are of the form $u(x) \approx |x|^\gamma$ and satisfy $u \in C^2(\Omega) \cap W^2_p(\Omega)$ for some $p > 1$. The second assumption is quite realistic since shape regularity is sufficient for the existence of a C^∞ mesh density with the property $D^t h \approx h^{1-t}$; see Nochetto et al. [57]. The third condition is based on the heuristics that the number of elements per unit of area is about $h(x)^{-2}$. The fourth assumption can be rephrased as follows: $\int_T |D^2u|^p \approx h_T^2 |D^2u(x_T)|^p$ where x_T is the barycenter of $T \in \mathscr{T}$. Finally, the fifth assumption replaces the error by an upper bound. In fact, if $I_\mathscr{T}$ is the Lagrange interpolation operator, we can use the local interpolation estimates (65) to write

$$|u - I_\mathscr{T} u|_{H^1(T)} \lesssim h_T^{\mathrm{sob}(W^2_p) - \mathrm{sob}(H^1)} |u|_{W^2_p(T)} \lesssim h_T^{2-\frac{2}{p}} |u|_{W^2_p(T)} \qquad \text{for all } T \in \mathscr{T}$$

and then argue as in the proof of Theorem 5.1 to derive the upper bound

$$\|\nabla(u - U)\|^2_{L^2(\Omega)} \lesssim \left(\int_\Omega h(x)^{2(p-1)} |D^2u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

Since we would like to minimize the error for a given number of degrees of freedom N, we propose the Lagrangian

$$\mathscr{L}[h, \lambda] = \int_\Omega \left(h(x)^{2(p-1)} |D^2u(x)|^p - \frac{\lambda}{h(x)^2} \right) dx,$$

with Lagrange multiplier $\lambda \in \mathbb{R}$. A stationary point of $\mathscr{L}$ satisfies (see Problem 5.2)

$$h(x)^{2(p-1)+2} |D^2u(x)|^p = \text{constant},$$

and thus requires a variable mesh-size $h(x)$ that compensates the local behavior of $D^2u(x)$. This relation can be interpreted as follows: since the error E_T associated with element $T \in \mathscr{T}$ satisfies

$$E_T = h_T^{2(p-1)} \int_T |D^2u|^p \approx h_T^{2(p-1)+2} |D^2u(x_T)|^p,$$

we infer that the *element error* is equidistributed.

Summarizing (and ignoring the asymptotic aspects of the above continuous model), a candidate for the sought condition is

$$E_T \approx \Lambda \quad \text{(constant)} \quad \text{for all } T \in \mathscr{T}.$$

Meshes satisfying this property have been constructed by Babuška et al [4] for corner singularities and $d = 2$; see also [39]. Problem 5.4 explores this matter and proposes a specific mesh grading towards the origin. However, what the above argument does not address is whether such meshes exist in general and whether they can be actually constructed upon bisecting the initial mesh $\mathscr{T}_0$, namely that $\mathscr{T} \in \mathbb{T}$.

5.4 Adaptive approximation

The purpose of this concluding section is to show that the maximum decay rate $N^{-n/d}$ in (69) can be reached under weaker regularity assumption when using suitably adapted meshes. Following the work of Binev et al. [14], we use an adaptive algorithm that is based on the knowledge of the element errors and on bisection.

The algorithm can be motivated with the above equidistribution principle in the following manner. Let $\delta > 0$ be a given tolerance and the polynomial degree $n = 1$. If the element error is equidistributed, that is $E_T \approx \delta^2$, and the global error decays with maximum rate $N^{-1/2}$, then

$$\delta^4 N \approx \sum_{T \in \mathscr{T}_N} E_T^2 = |u - I_{\mathscr{T}} u|^2_{H^1(\Omega)} \lesssim N^{-1}$$

that is $N \lesssim \delta^{-2}$. With this in mind, we impose $E_T \leq \delta^2$ as a common threshold to stop refining and expect $N \lesssim \delta^{-2}$.

The following algorithm implements this idea.

Algorithm (Thresholding). Given a tolerance $\delta > 0$ and a conforming mesh $\mathscr{T}_0$, THRESHOLD finds a conforming refinement $\mathscr{T} \geq \mathscr{T}_0$ of $\mathscr{T}_0$ by bisection such that $E_T \leq \delta^2$ for all $T \in \mathscr{T}$: let $\mathscr{T} = \mathscr{T}_0$ and

THRESHOLD$(\mathscr{T}, \delta)$
while $\mathscr{M} := \{T \in \mathscr{T} \,|\, E_T > \delta^2\} \neq \emptyset$
  $\mathscr{T} :=$ REFINE$(\mathscr{T}, \mathscr{M})$
end while
return$(\mathscr{T})$

We now discuss the situation mentioned above. Assume

$$u \in W^2_p(\Omega), \quad p > 1, d = 2, \tag{70}$$

which implies that u is uniformly continuous in Ω and we can take $I_{\mathscr{T}}$ to be the Lagrange interpolation operator. Since $p > 1$ we have $r = 2(1 - 1/p) > 0$, according

to (65), and

$$E_T \lesssim h_T^r \, \|D^2 u\|_{L^p(T)}. \tag{71}$$

Therefore, THRESHOLD *terminates* because h_T decreases monotonically to 0 with bisection. The quality of the resulting mesh is assessed next.

Theorem 5.2 (Thresholding). *If $u \in H_0^1(\Omega)$ verifies (70), then the output $\mathscr{T} \in \mathbb{T}$ of* THRESHOLD *satisfies*

$$|u - I_{\mathscr{T}} u|_{H^1(\Omega)} \leq \delta^2 (\#\mathscr{T})^{1/2}, \qquad \#\mathscr{T} - \#\mathscr{T}_0 \lesssim \delta^{-2} \, |\Omega|^{1-1/p} \|D^2 u\|_{L^p(\Omega)}.$$

Proof. Let $k \geq 1$ be the number of iterations of THRESHOLD before termination. Let $\mathscr{M} = \mathscr{M}_0 \cup \cdots \cup \mathscr{M}_{k-1}$ be the set of marked elements. We organize the elements in $\mathscr{M}$ by size in such a way that allows for a counting argument. Let $\mathscr{P}_j$ be the set of elements T of $\mathscr{M}$ with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-(j+1)/2} \leq h_T < h_T^{-j/2}.$$

We proceed in several steps.

1 We first observe that all T's in $\mathscr{P}_j$ are *disjoint*. This is because if $T_1, T_2 \in \mathscr{P}_j$ and $\mathring{T}_1 \cap \mathring{T}_2 \neq \emptyset$, then one of them is contained in the other, say $T_1 \subset T_2$, due to the bisection procedure. Thus

$$|T_1| \leq \frac{1}{2} |T_2|$$

contradicting the definition of $\mathscr{P}_j$. This implies

$$2^{-(j+1)} \, \#\mathscr{P}_j \leq |\Omega| \quad \Rightarrow \quad \#\mathscr{P}_j \leq |\Omega| \, 2^{j+1}. \tag{72}$$

2 In light of (71), we have for $T \in \mathscr{P}_j$

$$\delta^2 \leq E_T \lesssim 2^{-(j/2)r} \|D^2 u\|_{L^p(T)}.$$

Therefore

$$\delta^{2p} \, \#\mathscr{P}_j \lesssim 2^{-(j/2)rp} \sum_{T \in \mathscr{P}_j} \|D^2 u\|_{L^p(T)}^p \leq 2^{-(j/2)rp} \, \|D^2 u\|_{L^p(\Omega)}^p$$

whence

$$\#\mathscr{P}_j \lesssim \delta^{-2p} \, 2^{-(j/2)rp} \, \|D^2 u\|_{L^p(\Omega)}^p. \tag{73}$$

3 The two bounds for $\#\mathscr{P}$ in (72) and (73) are complementary. The first is good for j small whereas the second is suitable for j large (think of $\delta \ll 1$). The crossover takes place for j_0 such that

$$2^{j_0+1} |\Omega| = \delta^{-2p} \, 2^{-j_0(rp/2)} \|D^2 u\|_{L^p(\Omega)}^p \quad \Rightarrow \quad 2^{j_0} \approx \delta^{-2} \, \frac{\|D^2 u\|_{L^p(\Omega)}}{|\Omega|^{1/p}}.$$

4 We now compute

$$\#\mathscr{M} = \sum_j \#\mathscr{P}_j \lesssim \sum_{j\le j_0} 2^j|\Omega| + \delta^{-2p}\,\|D^2u\|^p_{L^p(\Omega)} \sum_{j>j_0} (2^{-rp/2})^j.$$

Since

$$\sum_{j\le j_0} 2^j \approx 2^{j_0}, \qquad \sum_{j>j_0} (2^{-rp/2})^j \lesssim 2^{-(rp/2)j_0} = 2^{-(p-1)j_0}$$

we can write

$$\#\mathscr{M} \lesssim \left(\delta^{-2} + \delta^{-2p}\delta^{2(p-1)}\right)|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} \approx \delta^{-2}\,|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}.$$

We finally apply Theorem 4.3 to arrive at

$$\#\mathscr{T} - \#\mathscr{T}_0 \lesssim \#\mathscr{M} \lesssim \delta^{-2}\,|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}.$$

5 It remains to estimate the energy error. We have, upon termination of THRESHOLD, that $E_T \le \delta^2$ for all $T \in \mathscr{T}$. Then

$$|u - I_{\mathscr{T}}u|^2_{H^1(\Omega)} = \sum_{T\in\mathscr{T}} E_T^2 \le \delta^4\,\#\mathscr{T}.$$

This concludes the Theorem. □

By relating the threshold value δ and the number of refinements N, we obtain a result about the convergence rate.

Corollary 5.2 (Convergence Rate). *Let $u \in H^1_0(\Omega)$ satisfy (70). Then for $N > \#\mathscr{T}_0$ integer there exists $\mathscr{T} \in \mathbb{T}$ such that*

$$|u - I_{\mathscr{T}}u|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}N^{-1/2}, \qquad \#\mathscr{T} - \#\mathscr{T}_0 \lesssim N.$$

Proof. Choose $\delta^2 = |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}N^{-1}$ in Theorem 5.2. Then, there exists $\mathscr{T} \in \mathbb{T}$ such that $\#\mathscr{T} - \#\mathscr{T}_0 \lesssim N$ and

$$\begin{aligned}|u - I_{\mathscr{T}}u|_{H^1(\Omega)} &\lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}N^{-1}\left(N + \#\mathscr{T}_0\right)^{1/2}\\ &\lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}N^{-1/2}\end{aligned}$$

because $N > \#\mathscr{T}_0$. This finishes the Corollary. □

Remark 5.6 (Piecewise smoothness). The global regularity (70) can be weakened to *piecewise W^2_p regularity* over the initial mesh $\mathscr{T}_0$, namely $W^2_p(\Omega;\mathscr{T}_0)$, and global $H^1_0(\Omega)$. This is because $W^2_p(T) \hookrightarrow C^0(\overline{T})$ for all $T \in \mathscr{T}_0$, whence $I_{\mathscr{T}}$ can be taken to be the Lagrange interpolation operator.

Remark 5.7 (Case $p < 1$). Consider either polynomial degree $n > 1$ and $d = 2$ or $n \ge 1$ for $d > 2$. The Sobolev number corresponding to a space with regularity of order $n+1$ is

$$n+1-\frac{d}{p}=\mathrm{sob}(H^1)=1-\frac{d}{2}\quad\Rightarrow\quad p=\frac{d}{n+d/2}.$$

For $d=2$ this implies $p<1$. Spaces based on $L^p(\Omega)$, $p<1$, are unusual in finite element theory but not in approximation theory [71, 30, 28]. The argument of Theorem 5.2 works provided we replace (71) by a modulus of regularity; in fact, $D^{n+1}u$ would not be locally integrable and so would fail to be a distribution. This requires two ingredients:

- The construction of a quasi-interpolation operator $I_{\mathscr{T}}: L^p(\Omega)\to S^{n,0}(\mathscr{T})$ for $p<1$ with optimal approximation properties; such operator $I_{\mathscr{T}}$ is inevitably nonlinear. We refer to [30, 28, 58], as well as [37] where the following key property is proven: $I_{\mathscr{T}}(v+P)=I_{\mathscr{T}}(v)+P$ for all $P\in S^{n,0}(\mathscr{T})$ and $v\in L^p(\Omega)$.
- Besov regularity properties of the solution u of an elliptic boundary value problem; we refer to [27] for such an endeavor for 2d Lipschitz domains and the Laplace operator. For the model problem with discontinuous coefficients as well as for $d>2$ this issue seems to be open in general.

Applying Corollary 5.2 to Example 5.1, we see that the maximum decay rate $N^{-1/2}$ for polynomial degree $n=1$ and dimension $d=2$, as well as $N^{-n/d}$ for $n\geq 1, d\geq 2$ when taking Remark 5.7 into account, can be reestablished by judicious mesh grading. Of course the thresholding algorithm cannot be applied directly within the finite element method because the exact solution u is typically unknown. In fact, we are only able to replace the element energy error by computable element error indicators, and thus gain access to u indirectly. This is the topic of a posteriori error analysis and is addressed in Chap. 6.

5.5 Problems

Problem 5.1. Let $\mathscr{T}$ be a shape regular and quasi-uniform triangulation of $\Omega\subset\mathbb{R}^d$. Let $\mathbb{V}_{\mathscr{T}}$ be the space of (possibly discontinuous) finite elements of degree $\leq n$. Given $u\in L^2(\Omega)$, the L^2-projection $U_{\mathscr{T}}\in\mathbb{V}_{\mathscr{T}}$ is defined by

$$\int_\Omega (u-U_{\mathscr{T}})V=0\quad\text{for all } V\in\mathbb{V}_{\mathscr{T}}.$$

Show

(a) $\|u-U_{\mathscr{T}}\|_{L^2(\Omega)}\lesssim h^{n+1}\,|u|_{H^{n+1}(\Omega)}$
(b) $\|u-U_{\mathscr{T}}\|_{H^{-m}(\Omega)}\lesssim h^{n+1+m}\,|u|_{H^{n+1}(\Omega)}$

for $0\leq m\leq n+1$ and h being the maximal mesh size of $\mathscr{T}$. The estimate in (b) ensures *superconvergence*.

Problem 5.2. Let $h(x)$ a smooth function locally equivalent to the mesh-size. Prove that a stationary point of the Lagrangian

$$\mathscr{L}[h,\lambda] = \int_\Omega \left(h(x)^{2(p-1)} |D^2 u(x)|^p - \frac{\lambda}{h(x)^2} \right) dx$$

satisfies the optimality condition

$$h^{2(p-1)+2} |D^2 u|^p = \text{constant}.$$

Problem 5.3. Consider the solution u of the model problem in Sect. 2.2.1 with corner singularity:

$$u(r,\theta) = r^\gamma \phi(\theta) \qquad 0 < \gamma < 1$$

in polar coordinates (r,θ). Show that $u \in W_p^2(\Omega) \setminus H^2(\Omega)$ for $1 \le p < 2/(2-\gamma)$.

Problem 5.4. Use the Principle of Equidistribution to determine the grading of an mesh for a corner singularity

$$u(r,\theta) = r^\gamma \phi(\theta) \qquad (0 < \gamma < 1).$$

In fact, show that

$$h_T = \Lambda \operatorname{dist}(T,0)^{1-\gamma/2} \qquad (\Lambda = \text{ constant}).$$

Count the number of elements using the expression $N \approx \int_\Omega \frac{dx}{h(x)^2}$ and derive an optimal bound $|u - I_{\mathscr{T}_N} u|_{H^1(\Omega)} \lesssim N^{-1/2}$ for polynomial degree $n = 1$.

Problem 5.5. Consider the function u of Problem 5.3.
(a) Examine the construction of an graded mesh via the Thresholding Algorithm.
(b) Repeat the proof of Theorem 5.2 replacing the W_p^2 regularity by the corresponding local H^2 regularity of u depending on the distance to the origin.

6 A posteriori error analysis

Suppose, as it is generically the case, that the solution of a boundary value problem is unknown. Then we may use a numerical method to compute an approximate solution. Of course, it is useful to have information about the error of such an approximation. Moreover, if the error is still to big, one would like to know how to modify the discretization so as to reduce the error effectively.

The results of the preceding chapters provide little such information, because they involve the exact solution and/or are of asymptotic nature. However, so-called *a posteriori error estimators* extract such information from the given problem and the approximate solution, without invoking the exact solution. Starting with the pioneering work [5] of Babuška and Rheinboldt , a great deal of work has been devoted to their derivation. We refer to [1, 6, 76] for an overview of the state-of-the-art.

This chapter is an introduction to a posteriori error estimators, providing the essentials for the following chapters about adaptive algorithms. To this end, we shall

mainly restrict ourselves to the model problem of Sect. 2.2.1 and we will drop the index N or $\mathscr{T}$, since it will be kept fixed during the whole chapter.

6.1 Error and residual

Let u be the exact solution of (10) and U be a corresponding Petrov-Galerkin solution as in (37). We want to obtain information about the error function $u-U$, which is typically unknown. The so-called *residual* $\mathscr{R} = \mathscr{R}(U,f) \in W^*$ given by

$$\langle \mathscr{R}, w\rangle := \langle f, w\rangle - \mathscr{B}[U,w] \quad \text{for all } w \in \mathbb{W}$$

depends only on data and the approximate solution U and is related to the error function by

$$\langle \mathscr{R}, w\rangle = \mathscr{B}[u-U,w] \quad \text{for all } w \in \mathbb{W}. \tag{74}$$

If the error notion of interest is $\|u-U\|_{\mathbb{V}}$, the following lemma determines a dual-norm of $\mathscr{R}$ that is equivalent to the error.

Lemma 6.1 (Abstract A posteriori Error Estimate). *There holds*

$$\alpha\|u-U\|_{\mathbb{V}} \le \|\mathscr{R}\|_{\mathbb{W}^*} \le \|\mathscr{B}\|\|u-U\|_{\mathbb{V}}, \tag{75}$$

where $0 < \alpha \le \|\mathscr{B}\|$ *are the inf-sup and continuity constants of* $\mathscr{B}$ *from* (21a) *and* (16).

Proof. The inf-sup condition (21a) and (74) imply

$$\alpha\|u-U\|_{\mathbb{V}} \le \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u-U,w] = \|\mathscr{R}\|_{\mathbb{W}^*},$$

while (74) and (16) imply

$$\|\mathscr{R}\|_{\mathbb{W}^*} = \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u-U,w] \le \|\mathscr{B}\|\|u-U\|_{\mathbb{V}}. \qquad \square$$

In view of the result, we are left with (approximately) evaluting $\|\mathscr{R}\|_{\mathbb{W}^*}$ at an acceptable cost. Notice that, while the quasi-best approximation property (3.2) relies on the stability of the discretization, Lemma 6.1 relies on the well-posedness of the continuous problem (10). It is thus a first, discretization-independent step. Here it was rather straight-forward, it can get more involved depending on problem and error notion.

There are various techniques for evaluting $\|\mathscr{R}\|_{\mathbb{W}^*}$. This second step depends on the discretization. In what follows, we present the most basic and common approach, *standard residual estimation*, in the case of our model problem of Sect. 2.2.1 and its standard discretization of Sect. 3.2.2.

Before embarking on it, it is instructive to analyze the structure of the residual for the model problem, where $\mathbb{W}^* = H^{-1}(\Omega)$, U is piecewise polynomial function

over a triangulation $\mathscr{T}$, and the residual is the distribution

$$\mathscr{R} = f + \operatorname{div}(\boldsymbol{A}\nabla U) \in H^{-1}(\Omega).$$

To this end, we suppose $f \in L^2(\Omega)$. This allows us to write $\langle \mathscr{R}, w \rangle$ as integrals over each $T \in \mathscr{T}$ and integration by parts yields the representation:

$$\begin{aligned} \langle \mathscr{R}, w \rangle &= \int_\Omega fw - \nabla U \cdot \boldsymbol{A}\nabla w = \sum_{T\in\mathscr{T}} \int_T fw - \nabla U \cdot \boldsymbol{A}\nabla w \\ &= \sum_{T\in\mathscr{T}} \int_T rw + \sum_{S\in\mathring{\mathscr{S}}} \int_S jw, \end{aligned} \tag{76}$$

with

$$\begin{aligned} r &= f + \operatorname{div}(\boldsymbol{A}\nabla U) \quad \text{in any simplex } T \in \mathscr{T}, \\ j &= [\![\boldsymbol{A}\nabla U]\!] = \boldsymbol{n}^+ \cdot \boldsymbol{A}\nabla U_{|T^+} + \boldsymbol{n}^- \cdot \boldsymbol{A}\nabla U_{|T^-} \quad \text{on any internal side } S \in \mathring{\mathscr{S}} \end{aligned}$$

and $\boldsymbol{n}^+$, $\boldsymbol{n}^-$ are unit normals pointing towards T^+, $T^- \in \mathscr{T}$. We see that the distribution $\mathscr{R}$ consists of a regular part r, called *interior or element residual*, and a singular part j, called *jump or interelement residual*. The regular part is absolutely continuous w.r.t. the d-dimensional Lebesgue measure and is related to the strong form of the PDE. The singular part is supported on the skeleton $\Gamma = \bigcup_{S\in\mathring{\mathscr{S}}} S$ of $\mathscr{T}$ and is absolutely continuous w.r.t. the $(d-1)$-dimensional Hausdorff measure.

We point out that this structure is not special to the model problem and its discretization but rather arises from the weak formulation of the PDE and the piecewise construction of finite element spaces.

6.2 Global upper bound

As already mentioned, we provide an a posteriori analysis for the model problem in Sect. 2.2.1 using standard residual estimation. This approach provides an upper bound $\|\mathscr{R}\|_{\mathbb{W}^*}$ with the help of suitably weighted Lebesgue norms (which are considered to be computable). We will see below that the weights are crucial for the sharpness of the derived bound.

In what follows, we shall write '$\lesssim$' instead of $\lesssim C$, where the constant C is bounded in terms of the shape coefficent $\sigma_{\mathscr{T}}$ of the triangulation $\mathscr{T}$ and the dimension d. The presentation here is a simplified version of [74], which has been influenced by [5, 20, 54] and provides in particular constants that are explicit in terms of local Poincaré constants.

6.2.1 Tools

For bounding $\|\mathscr{R}\|_{\mathbb{W}^*}$ we need two tools: a trace inequality that will help to bound the singular part with the jump residual and a Poincaré-type inequality that will take care of the lower order norms arising in the trace inequality and from the regular part with the element residual. We start by deriving the trace inequality.

Lemma 6.2 (Trace Identity). *Let T be a d-simplex, S a side of T, and z the vertex opposite to S. Defining the vector field $\boldsymbol{q}_S$ by*

$$\boldsymbol{q}_S(x) := x - z$$

the following equality holds

$$\frac{1}{|S|}\int_S v = \frac{1}{|T|}\int_T v + \frac{1}{d|T|}\int_T \boldsymbol{q}_S \cdot \nabla v \quad \text{for all } v \in W_1^1(T).$$

Proof. We start with properties of the vector field $\boldsymbol{q}_S$. Let S' be an arbitrary side of T and fix some $y \in S'$. We then see $\boldsymbol{q}_S(x)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T + (x-y)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T$ for any $x \in S'$ since $x-y$ is a tangent vector to S'. Therefore, on each side of T, the associated normal flux $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ is constant. In particular, we see $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ vanishes on $\partial T \setminus S$ by choosing $y = z$ for sides emanating from z. Moreover, $\operatorname{div}\boldsymbol{q}_S = d$. Thus, if $v \in C^1(\overline{T})$, the Divergence Theorem yields

$$\int_T \boldsymbol{q}_S\cdot\nabla v = -d\int_T v + (\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S}\int_S v.$$

Take $v = 1$ to show $(\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S} = d|T|/|S|$ and extend the result to $v \in W_1^1(T)$ by density. □

The following corollary is a ready-to-use form for our purposes.

Corollary 6.1 (Scaled Trace Inequality). *For any side $S \subset T$ the following inequality holds*

$$\|v\|_{L^2(S)} \lesssim h_S^{-1/2}\|v\|_{L^2(T)} + h_S^{1/2}\|\nabla v\|_{L^2(T)} \quad \text{for all } v \in H^1(T) \tag{77}$$

where $h_S =: |S|^{1/(d-1)}$.

Proof. Problem 6.2. □

We next present the Poincaré-type inequality. Let

$$\omega_z = \cup_{T\ni z} T$$

be the star (or patch) around a vertex $z \in \mathscr{V}$ of $\mathscr{T}$. We define

$$h_z := |\omega_z|^{1/d}$$

and notice that this quantity is, up to the shape coefficient of $\mathscr{T}$, equivalent to the diameter of ω_z, to h_T if $T \subset \omega_z$ and to h_S if $S \subset \omega_z$.

Lemma 6.3 (Local Poincaré-Type Inequality). *For any $v \in H^1_0(\Omega)$ and $z \in \mathscr{V}$ there exists $c_z \in \mathbb{R}$ such that*

$$\|v - c_z\|_{L^2(\omega_z)} \lesssim h_z \|\nabla v\|_{L^2(\omega_z)}. \tag{78}$$

If $z \in \partial\Omega$ is a boundary vertex, then we can take $c_z = 0$.

Proof. 1 In fact, for any $z \in \mathscr{V}$ the value

$$\bar{c}_z = \frac{1}{|\omega_z|} \int_{\omega_z} v$$

is an optimal choice and (78) can be shown with $c_z = \bar{c}_z$ as (64).
2 If $z \in \partial\Omega$, then we observe that there exists a side $S \subset \partial\omega_z \cap \partial\Omega$ such that $v = 0$ on S. We therefore can write

$$v = v - \frac{1}{|S|} \int_S v = (v - \bar{c}_z) - \frac{1}{|S|} \int_S (v - \bar{c}_z)$$

and thus, using Corollary 6.1 and Step 1 for the second term,

$$\|v\|_{L^2(\omega_z)} \lesssim \|v - \bar{c}_z\|_{L^2(\omega_z)} + h_z \|\nabla v\|_{L^2(\omega_z)} \lesssim h_z \|\nabla v\|_{L^2(\omega_z)},$$

which establishes the supplement for boundary vertices. □

6.2.2 Derivation of the Upper Bound

We now pass to the proper derivation of the upper bound. The following properties of the Courant basis $\{\phi_z\}_{z \in \mathscr{V}}$ from Theorem 3.3 are instrumental:

- It provides a discrete partition of unity:

$$\sum_{z \in \mathscr{N}} \phi_z = 1 \quad \text{in } \Omega. \tag{79}$$

- Each function ϕ_z is contained in $S^{n,0}(\mathscr{T})$ and so the residual is orthogonal to the interior contributions of the partition of unity:

$$\langle \mathscr{R}, \phi_z \rangle = 0 \quad \text{for all } z \in \mathring{\mathscr{V}} := \mathscr{V} \cap \Omega. \tag{80}$$

The second property corresponds to the Galerkin orthogonality. Notice that the first property involve all vertices, while in the second one the boundary vertices are excluded. For this reason, the supplement on boundary vertices in Lemma 78 is important.

For any $w \in H^1_0(\Omega)$ we start by applying (79) and then (80) with c_z from Lemma 6.3 for w to write

$$\langle \mathscr{R}, w\rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, w\,\phi_z\rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, (w-c_z)\phi_z\rangle,$$

where $c_z = 0$ whenever $z \in \partial\Omega$. In view of representation (76), we can write

$$|\langle \mathscr{R}, (w-c_z)\phi_z\rangle| \le \int_{\omega_z} |r||w-c_z|\phi_z + \int_{\gamma_z} |j||w-c_z|\phi_z$$

where γ_z is the skeleton of ω_z, i.e. the union of all sides emanating from z. We examine each term on the right hand side seperately. Invoking $\|\phi_z\|_{L^\infty(\omega_z)} \le 1$ and (78), we obtain

$$\int_{\omega_z} |r||w-c_z|\phi_z \le \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)}\|w-c_z\|_{L^2(\omega_z)} \lesssim h_z\|r\,\phi_z^{1/2}\|_{L^2(\omega_z)}\|\nabla w\|_{L^2(\omega_z)}.$$

Likewise, employing (77) and (78), we get

$$\int_{\gamma_z} |j||w-c_z|\phi_z \le \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\|w-c_z\|_{L^2(\gamma_z)} \lesssim h_z^{1/2}\|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\|\nabla w\|_{L^2(\omega_z)}.$$

Therefore,

$$|\langle \mathscr{R}, w\,\phi_z\rangle| \lesssim \Big(h_z\|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} + h_z^{1/2}\|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\Big)\|\nabla w\|_{L^2(\omega_z)}.$$

Summing over $z \in \mathscr{V}$ and using Cauchy-Schwarz in $\mathbb{R}^{\#\mathscr{T}}$ gives

$$|\langle \mathscr{R}, w\rangle| \lesssim \left(\sum_{z\in\mathscr{V}} h_z^2\|r\,\phi_z^{1/2}\|^2_{L^2(\omega_z)} + h_z\|j\,\phi_z^{1/2}\|^2_{L^2(\gamma_z)}\right)^{1/2}\left(\sum_{z\in\mathscr{V}} \|\nabla w\|^2_{L^2(\omega_z)}\right)^{1/2}.$$

Denote by $h\colon \Omega \to \mathbb{R}^+$ the mesh-size function given by $h(x) := |S|^{1/k}$ if x belongs to the interior of the k-subsimplex S of $\mathscr{T}$ with $k \in \{1,\dots,d\}$. Then for all $x \in \omega_z$ we have $h_z \lesssim h(x)$. Therefore employing (79) once more and recalling that Γ is the union of all interior sides of $\mathscr{T}$, we proceed by

$$\begin{aligned}\sum_{z\in\mathscr{V}} h_z^2\|r\,\phi_z^{1/2}\|^2_{L^2(\omega_z)} + h_z\|j\,\phi_z^{1/2}\|^2_{L^2(\gamma_z)} &\lesssim \sum_{z\in\mathscr{V}} \|h\,r\,\phi_z^{1/2}\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\,\phi_z^{1/2}\|^2_{L^2(\Gamma)}\\ &= \|h\,r\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\|^2_{L^2(\Gamma)}.\end{aligned}$$

We next resort to the finite overlapping property of stars, namely

$$\sum_{z\in\mathscr{V}} \chi_{\omega_z}(x) \le d+1$$

to deduce

$$\sum_{z\in\mathscr{V}} \|\nabla w\|^2_{L^2(\omega_z)} \lesssim \|\nabla w\|^2_{L^2(\Omega)}.$$

Thus, introducing the *element indicators*

$$\mathscr{E}^2(U,T) := h_T^2\|r\|^2_{L^2(T)} + h_T\|j\|^2_{L^2(\partial T\setminus\partial\Omega)} \tag{81}$$

and the *error estimator*

$$\mathscr{E}^2(U,\mathscr{T}) = \sum_{T\in\mathscr{T}} \mathscr{E}^2(U,T), \tag{82}$$

we have derived

$$\|\mathscr{R}\|_{\mathbb{W}^*} \lesssim \mathscr{E}(U,\mathscr{T}).$$

Combing this with the abstract a posteriori bound in Lemma 6.1, we obtain the main result of this section.

Theorem 6.1 (Upper Bound). *Let u and U be exact and Galerkin solution of the model problem and its standard discretization. Then there holds the following global upper bound:*

$$\|\nabla(u-U)\|_{L^2(\Omega)} \le \frac{C}{\alpha_1}\mathscr{E}(U,\mathscr{T}) \tag{83}$$

where α_1 is the global smallest eigenvalue of $\boldsymbol{A}(x)$ and C depends only on the shape coefficient $\sigma_{\mathscr{T}}$ and on the dimension d.

6.2.3 Sharpness of Weighted Lebesgue Norms

The indicators $\mathscr{E}(U,T)$, $T\in\mathscr{T}$, consists of weighted L^2-norms. The weights h_T and $h_T^{1/2}$ arise from the local Poincaré inequalities (78), which in turn rely on the orthogonality (80) of the residual. If we do not exploit orthogonality and use a global Poincaré-type inequality instead of the local ones, the resulting weights are 1 and $h_T^{-1/2}$ and the corresponding upper bound has a lower asymptotic decay rate. We wonder whether the ensuing weights h_T and $h_T^{1/2}$ are accurate and explore this issue for the first weight h_T of the element residual. The following discussion is a elaborated version of [62, Remark 3.1].

First we notice that the local counterpart of $\|\mathscr{R}\|_{H^{-1}(\Omega)}$ is $\|\mathscr{R}\|_{H^{-1}(T)}$ and observe

$$\|\mathscr{R}\|_{H^{-1}(T)} = \sup_{\|\nabla w\|_{L^2(T)}\le 1} \langle\mathscr{R}, w\rangle = \sup_{\|\nabla w\|_{L^2(T)}\le 1} \int_T rw = \|r\|_{H^{-1}(T)} \tag{84}$$

thanks to the representation (76). This suggests to compare the weighted norm $h_T\|r\|_{L^2(T)}$ in the indicator with the local negative norm $\|r\|_{H^{-1}(T)}$. Mimicking the local part in the argument of Sect. 6.2.2, we derive

$$\int_T rw \le \|r\|_{L^2(T)}\|w\|_{L^2(T)} \lesssim h_T\|r\|_{L^2(T)}\|\nabla w\|_{L^2(T)}$$

with the help of the Poincaré-Friedrichs inequality (7). Consequently there holds

$$\|r\|_{H^{-1}(T)} \lesssim h_T \|r\|_{L^2(T)}. \tag{85}$$

Since $L^2(\Omega)$ is a proper subspace of $H^{-1}(\Omega)$ the inverse inequality cannot hold for arbitrary r. Consequently, $h_T\|r\|_{L^2(T)}$ may overestimates $\|r\|_{H^{-1}(T)}$. On the other hand, if $r \in \mathbb{R}$ is *constant* and η denotes a non-negative function with properties

$$|T| \lesssim \int_T \eta, \quad \operatorname{supp}\eta = T, \quad \|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1} \tag{86}$$

(postpone the question of existence until (90) below), we deduce

$$\begin{aligned}\|r\|^2_{L^2(T)} &\lesssim \int_T r(r\eta) \le \|r\|_{H^{-1}(T)}\|\nabla(r\eta)\|_{L^2(T)} \\ &\le \|r\|_{H^{-1}(T)}\|r\|_{L^2(T)}\|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1}\|r\|_{H^{-1}(T)}\|r\|_{L^2(T)}.\end{aligned}$$

whence

$$h_T\|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}. \tag{87}$$

This shows that overestimation in (85) is caused by *oscillation* of r at a scale finer than the mesh-size. The estimate (87) is also valid for $r \in \mathbb{P}_l(T)$, but the constant deteriorates with the degree l; see Problem 6.6.

To conclude this discussion, we observe that $h_T\|r\|_{L^2(T)}$ can be easily approximated with the help of numerical integration, while this is not true for $\|r\|_{H^{-1}(T)}$. We therefore may say the weights are asymptotically acurate and that the possible overestimation of the weighted Lebesgue norms in (81) is the price for (almost) computability. This view is consistent with the fact that the indicators associated with the approximation of the Dirichlet boundary values in [62], which do not to invoke weighted Lebesgue norms, are overestimation-free.

6.3 Lower bounds

The discussion in Sect. 6.2.3 suggests that $h_T\|r\|_{L^2(T)}$ bounds 'asymptotically' $\|\mathscr{R}\|_{H^{-1}(T)}$ from below. This is the main step towards a *local* lower bound for the error. Such local lower bounds are the subject of this section. They do not contradict the global nature of the boundary value problem and their significance goes beyond a verification of the sharpness of the global upper bound (83).

For the sake of presentation, we present the case with polynomial degree $n = 1$ and leave the general case as problems to the reader.

6.3.1 Interior Residual

Let us start with a lower bound in terms of the interior residual and first check that $h_T\|r\|_{L^2(T)}$ bounds asymptotically $\|\mathscr{R}\|_{H^{-1}(T)}$ from below. To this end, we introduce the *oscillation of the interior residual* in T by

$$h_T\|r-\overline{r}_T\|_{L^2(T)},$$

where $\overline{r}_T$ denotes the mean value of r in T. Replacing r in (85) by $r-\overline{r}_T$ and in (87) by $\overline{r}_T$ as well as recalling (84), we derive

$$\begin{aligned} h_T\|r\|_{L^2(T)} &\le h_T\|\overline{r}_T\|_{L^2(T)} + h_T\|r-\overline{r}_T\|_{L^2(T)} \\ &\lesssim \|\overline{r}_T\|_{H^{-1}(T)} + h_T\|r-\overline{r}_T\|_{L^2(T)} \\ &\lesssim \|r\|_{H^{-1}(T)} + \|r-\overline{r}_T\|_{H^{-1}(T)} + h_T\|r-\overline{r}_T\|_{L^2(T)} \\ &\lesssim \|\mathscr{R}\|_{H^{-1}(T)} + h_T\|r-\overline{r}_T\|_{L^2(T)}. \end{aligned} \tag{88}$$

This is the desired statement because the oscillation $h_T\|r-\overline{r}_T\|_{L^2(T)}$ is expected to convergence faster than $h_T\|r\|_{L^2(T)}$ under refinement. In the case $n=1$ at hand there holds $r=f$ and, for example, there is one additional order if $f\in H^1(\Omega)$.

Since

$$\|\mathscr{R}\|_{H^{-1}(T)} = \sup_{w\in H^1_0(T)} \frac{\langle \mathscr{R}, w\rangle}{\|\nabla w\|_{L^2(T)}} = \sup_{w\in H^1_0(T)} \frac{\mathscr{B}[u-U,w]}{\|\nabla w\|_{L^2(T)}} \le \alpha_2\,\|\nabla(u-U)\|_{L^2(T)},$$

we have derived the following local lower bound

$$h_T\|r\|_{L^2(T)} \lesssim \alpha_2\,\|\nabla(u-U)\|_{L^2(T)} + h_T\|r-\overline{r}_T\|_{L^2(T)}, \tag{89}$$

which also holds with $\overline{r}_T$ chosen from $\mathbb{P}_l(T)$ at the price of a larger constant hidden in $\lesssim$.

Finally we comment on the choice of the cut-off function $\eta_T\in W^1_\infty(T)$ with (86). For example, we may take

$$\eta_T = (d+1)^{d+1}\prod_{z\in\mathscr{V}\cap T}\lambda_z, \tag{90}$$

where λ_z, $z\in\mathscr{V}\cap T$, are the barycentric coordinates of T; see Lemma 3.3. This choice is due to Verfürth [75, 76]. Another choice, due to Dörfler [32], can be defined as follows: refine T such that there appears an interior node and take the corresponding Courant basis function on the virtual triangulation of T; see Fig. 11 for the 2-dimensional case.

The Dörfler cut-off function has the additional property that it is an element of a refined finite element space. This is not important here but useful when proving lower bounds for the differences of two discrete solutions. Such estimates are

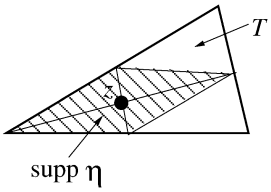


Fig. 11 Virtual refinement of a triangle for the Dörfler cut-off function

therefore called *discrete lower bound* whereas the bound for the true error is called *continuous lower bound*.

6.3.2 Jump Residual

We next strive for a local lower bound for the error in terms of the jump residual $h_S^{1/2}\|j\|_{L^2(S)}$, $S \in \mathring{\mathscr{S}}$, and use Sect. 6.3.1 on the interior residual as guideline.

We first notice that $j = [\![\boldsymbol{A}\nabla U]\!]$ is not necessary constant on an interior side $S \in \mathring{\mathscr{S}}$ due to the presence of A. We therefore introduce the oscillation of the jump residual in S:

$$h_S^{1/2}\|j - \overline{j}_S\|_{L^2(S)},$$

where $\overline{j}_S$ stands for the mean value of j on S. Notice that the important question about the order of this oscillation is not obvious because, in contrast to the oscillation of the element residual, the approximate solution U is involved. We postpone a corresponding discussion to Remark 6.1.

To choose a counterpart of η_T, let ω_S denote the patch composed of the two elements of $\mathscr{T}$ sharing S; see Fig. 12 for the 2-dimensional case. Obviously ω_S has a nonempty interior. Let $\eta_S \in W_\infty^1(\omega_S)$ be a cut-off function with the properties

$$|S| \lesssim \int_S \eta_S, \quad \operatorname{supp}\eta_S = \omega_S, \quad \|\eta_S\|_{L^\infty(\omega_S)} = 1, \quad \|\nabla\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{-1}. \tag{91}$$

Following Verfürth [75, 76] we may take η_S given by

$$\eta_{S|T} = d^d \prod_{z\in\mathscr{V}\cap S} \lambda_z^T, \tag{92}$$

where $T \subset \omega_S$ and λ_z^T, $z \in \mathscr{V} \cap T$, are the barycentric coordinates of T. Also here Dörfler [32] proposed an alternative which is obtained as follows: refine ω_S such that there appears an interior node of S and take the corresponding Courant basis function on the virtual triangualtion of ω_S; see Fig. 12 for the 2-dimensional case.

After these preparations we are ready to derive a counterpart of (88). In view of the properties of η_S, we have

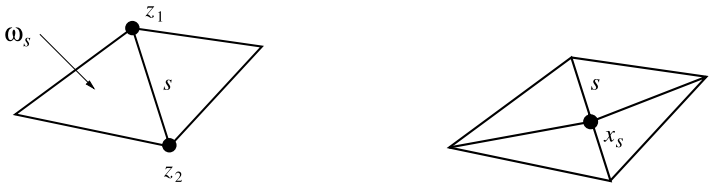


Fig. 12 Patch ω_S of triangles associated to interior side (left) and its refinement for Dörfler cut-off function (right)

$$\|\overline{j}_S\|^2_{L^2(S)} \lesssim \int_S \overline{j}_S(\overline{j}_S\eta_S) = \int_S j\,\psi_S + \int_S (\overline{j}_S - j)\,\psi_S \tag{93}$$

with $\psi_S = \overline{j}_S\eta_S$. We rewrite the first term on the right hand side with the representation formula (76) as follows:

$$\int_S j\,\psi_S = -\int_{\omega_S} r\,\psi_S + \langle \mathscr{R}, \psi_S\rangle,$$

where, in contrast to Sect. 6.3.1, the jump residual couples with the element residual. Hence

$$\left|\int_{\omega_S} j\,\psi_S\right| \le \|r\|_{L^2(\omega_S)}\|\psi_S\|_{L^2(\omega_S)} + \|\mathscr{R}\|_{H^{-1}(\omega_S)}\|\nabla\psi_S\|_{L^2(\omega_S)}.$$

In view of $|\omega_S| \lesssim h_S|S|$ and (91), we have

$$\|\psi_S\|_{L^2(\omega_S)} \le \|\overline{j}_S\|_{L^2(\omega_S)}\|\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{1/2}\|\overline{j}_S\|_{L^2(S)}$$

and

$$\|\nabla\psi_S\|_{L^2(\omega_S)} \le \|\overline{j}_S\|_{L^2(\omega_S)}\|\nabla\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{-1/2}\|\overline{j}_S\|_{L^2(S)}.$$

We infer that

$$\left|\int_{\omega_S} j\,\psi_S\right| \lesssim \left(h_S^{1/2}\|r\|_{L^2(\omega_S)} + h_S^{-1/2}\|\mathscr{R}\|_{H^{-1}(\omega_S)}\right)\|\overline{j}_S\|_{L^2(S)}.$$

In addition

$$\left|\int_S (\overline{j}_S - j)\psi_S\right| \le \|\overline{j}_S - j\|_{L^2(S)}\|\psi_S\|_{L^2(S)} \lesssim \|\overline{j}_S - j\|_{L^2(S)}\|\overline{j}_S\|_{L^2(S)}.$$

Inserting these estimates into (93) yields

$$\|\overline{j}_S\|^2_{L^2(S)} \lesssim \left(h_S^{1/2}\|r\|_{L^2(\omega_S)} + h_S^{-1/2}\|\mathscr{R}\|_{H^{-1}(\omega_S)} + \|\overline{j}_S - j\|_{L^2(S)}\right)\|\overline{j}_S\|_{L^2(S)}$$

whence, using (89) and $\|\mathscr{R}\|_{H^{-1}(T)} \le \|\mathscr{R}\|_{H^{-1}(\omega_S)}$ for $T \subset \omega_S$,

$$h_S^{1/2}\|j\|_{L^2(S)} \lesssim \|\mathscr{R}\|_{H^{-1}(\omega_S)} + \|h(r-\overline{r})\|_{L^2(\omega_S)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(S)}, \tag{94}$$

where h denotes the mesh-size function from Sect. 6.2.2 and $\overline{r}$ and $\overline{j}$ are given by

$$\overline{r}_{|T} = \overline{r}_T \quad \text{for all } T \in \mathscr{T} \qquad \text{and} \qquad \overline{j}_{|S} = \overline{j}_S \quad \text{for all } S \in \mathring{\mathscr{S}}.$$

Since

$$\|\mathscr{R}\|_{H^{-1}(\omega_S)} \leq \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_S)},$$

we obtain the local lower bound in terms of the jump residual:

$$h_S^{1/2}\|j\|_{L^2(S)} \lesssim \alpha_2\|\nabla(u-U)\|_{L^2(\omega_S)} + \|h(r-\overline{r})\|_{L^2(\omega_S)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(S)}. \tag{95}$$

Also this estimate holds with $\overline{j}$ piecewise polynomial of degree l; see Problem 6.11.

6.3.3 Local Lower Bound

We combine the two results on interior and jump residual and discuss its significance. To this end, we associate with each simplex $T \in \mathscr{T}$ the patch

$$\omega_T := \bigcup_{S \subset \partial T \setminus \partial\Omega} \omega_S,$$

see Fig. 13 for the 2-dimensional case, and define the oscillation in ω_T by

$$\operatorname{osc}(U, \omega_T) = \|h(r-\overline{r})\|_{L^2(\omega_T)} + \|h^{1/2}(j-\overline{j})\|_{L^2(\partial T \setminus \partial\Omega)}. \tag{96}$$

Recall that the higher order nature of $h_T\|r-\overline{r}_T\|_{L^2(T)}$ in (88) is crucial. We therefore compare the convergence order of (96) with that of the local error.

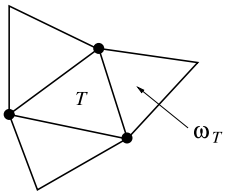


Fig. 13 Patch associated to a triangle in the local lower bound

Remark 6.1 (On Asymptotics of Oscillation). For simplicity, we consider only polynomial degree $n = 1$, maximum convergence rates and suppose that $\boldsymbol{A}$ and f are smooth. One then expects that the local error vanishes like

$$\|\nabla(u-U)\|^2_{L^2(T)} = \mathscr{O}(h_T^{d+2})$$

and interior and jump residual oscillations like

$$\|h(r-\overline{r})\|^2_{L^2(\omega_T)} + \|h^{1/2}(j-\overline{j})\|^2_{L^2(\partial T\setminus\partial\Omega)} = \mathscr{O}(h_T^{d+4}).$$

We already argued about the higher order of the interior residual after (88). Regarding the jump residual, the fact that ∇U is piecewise constant entails the identity

$$j-\overline{j}_S = [\![(\boldsymbol{A}-\overline{\boldsymbol{A}}_S)\nabla U]\!] = (\boldsymbol{A}-\overline{\boldsymbol{A}}_S)\boldsymbol{A}^{-1}j \quad \text{on an interior } S\in\mathring{\mathscr{S}},$$

which reveals the additional order for sufficiently smooth $\boldsymbol{A}$.

The oscillation $\operatorname{osc}(U,\omega_T)$ is therefore expected to be a higher order term for $h_T\downarrow 0$. However, as we shall see from the example in Remark 6.4 below, it may dominate on relatively coarse triangulations.

Similar arguments may be used to determine an appropriate polynomial degree of $\overline{j}_S$ and $\overline{r}_T$ in the case of general n. We do not insist on this and anticipate that in Chaps. 8 and 9 $\overline{r}_T$ will be the $L^2(T)$-best approximation in $P_{2n-2}(T)$ and $\overline{j}_S$ the $L^2(S)$-best approximation in $P_{2n-1}(S)$. This choices ensure, also for piecewise smooth A and f, that the oscillation is of higher order.

Since (89) and (95) hold also for piecewise polynomial $\overline{r}$ and $\overline{j}$, we have the following result for single indicators.

Theorem 6.2 (Local Lower Bound). *Let u and U be exact and Galerkin solution of the model problem and its standard discretization. Then, up to oscillation, each indicator is bounded by the local error:*

$$\mathscr{E}(U,T) \lesssim \alpha_2\|\nabla(u-U)\|_{L^2(\omega_T)} + \operatorname{osc}(U,\omega_T) \qquad \text{for all } T\in\mathscr{T}, \tag{97}$$

where α_2 is the largest eigenvalue of $\boldsymbol{A}(x)$ in ω_T and the hidden constant depends only on the shape coefficients of the simplices in ω_T, the dimension d and the polynomial degrees for $\overline{r}$ and $\overline{j}$.

Proof. Simply add the generalizations of (89) and (95) for all interior sides $S\in\mathring{\mathscr{S}}$ with $S\subset\partial T$. □

It is worthwhile to observe that in proving the local lower bound we have used the following the abstract notion of *local continuity* of the bilinear form $\mathscr{B}$. Let $\mathbb{V}$, $\mathbb{W}$ be normed spaces over Ω that are equipped with integral norms. If ω is a subdomain of Ω, then

$$\mathscr{B}[v,w] \le C_{\mathscr{B}}\|v\|_{\mathbb{V}(\omega)}\|w\|_{\mathbb{W}} \quad \text{for all } w \text{ with } w=0 \text{ in } \Omega\setminus\overline{\omega}, \tag{98}$$

where $\|\cdot\|_{\mathbb{V}(\omega)}$ stands for the restriction of $\|\cdot\|_{\mathbb{V}}$ to ω. Obviously, the continuity constant $\|\mathscr{B}\|$ satisfies $\|\mathscr{B}\|\le C_{\mathscr{B}}$ and therefore local continuity is stronger than global continuity. Property (98) readily implies an abstract local counterpart of the lower bound in Lemma 6.1.

We conclude this section with a remark about the importance of the fact that the lower bound in Theorem 6.2 is local and a remark about a simplifying setting in following chapters.

Remark 6.2 (Local Lower Bound and Marking). If $\operatorname{osc}(U,\omega_T) \ll \|\nabla(u-U)\|_{L^2(\omega_T)}$, as we expect asymptotically, then (97) translates into

$$\mathscr{E}(U,T) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_T)}.$$

This means that an element T with relatively large error indicator contains a large portion of the error. To improve the solution U effectively, such T must be split giving rise to a procedure that tries to equidistribute errors. This is consistent with the discussion of adaptive approximation in 1d of Sect. 1.1 and constructive approximation of Chap. 5.

Remark 6.3 (Oscillation vs Data Oscillation). The quantity (96) measures oscillations of both interior residual r and jump residual j beyond the local mesh scale. Note that if U is piecewise affine and $\boldsymbol{A}(x)$ is piecewise constant, then

$$r = f + \operatorname{div}(\boldsymbol{A}\nabla U) = f \quad \text{and} \quad j = [\![\boldsymbol{A}\nabla U]\!]_S = \overline{j}_S.$$

Consequently

$$\operatorname{osc}(U,\omega_T) = \|h(f-\overline{f})\|_{L^2(\omega_T)}$$

becomes *data oscillation*, which is independent of the discrete solution U. Otherwise, for variable $\boldsymbol{A}$, osc depends on the discrete solution U. This additional dependence creates a nonlinear interaction in the adaptive algorithm and so leads to difficulties in characterizing an appropriate approximation class for adaptive methods, see Chap. 9.

6.3.4 Global Lower Bound and Equivalence

We derive a global lower bound from Theorem 6.2 and summarize the achievements of global nature in this chapter.

To formulate the global lower bound, we introduce the global oscillation

$$\operatorname{osc}(U,\mathscr{T}) = \|h(r-\overline{r})\|_{L^2(\Omega)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(\Gamma)}, \tag{99}$$

recalling that Γ is the interior skeleton of $\mathscr{T}$. By summing (97) over all $T \in \mathscr{T}$ and taking into account (55), which entails a finite overlapping of the patches ω_T, we obtain the following global result.

Corollary 6.2 (Global Lower Bound). *Let u and U be exact and Ritz-Galerkin solutions of the model problem and its standard discretization. Then there holds the following global lower bound:*

$$\mathscr{E}(U,\mathscr{T}) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\Omega)} + \operatorname{osc}(U,\mathscr{T}) \tag{100}$$

where α_2 is the largest global eigenvalues of $\boldsymbol{A}$ and the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and the polynomial degrees for $\overline{r}$ and $\overline{j}$.

As already alluded to in Sect. 6.2.3, the presence of $\operatorname{osc}(U,\mathscr{T})$ in the lower bound is the price to pay for having a simple and computable estimator $\mathscr{E}(U,\mathscr{T})$. In the following remark, we present an example that shows that $\operatorname{osc}(U,\mathscr{T})$ cannot be removed from (100).

Remark 6.4 (Necessity of oscillation). Let $\varepsilon = 2^{-K}$ for K integer and extend the function $\frac{1}{2}x(\varepsilon - |x|)$ defined on $(-\varepsilon,\varepsilon)$ to a 2ε-periodic C^1 function u_ε on $\Omega = (-1,1)$. Moreover, let the forcing function be $f_\varepsilon = -u''$, which is 2ε-periodic and piecewise constant with values ± 1 that change at multiples of ε; see Fig. 14. Let

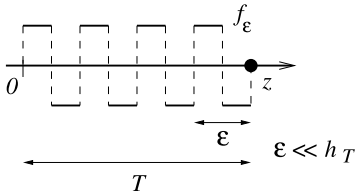


Fig. 14 An strongly oscillating forcing function

$\mathscr{T}$ be a uniform mesh with mesh-size $h = 2^{-k}$, with $k \ll K$. We consider piecewise linear finite elements $\mathbb{V}(\mathscr{T}_\varepsilon)$ and corresponding Galerkin solution $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$. It is easy to verify that f_ε is L^2-orthogonal to both the space of piecewise constants and linears over $\mathscr{T}_\varepsilon$, whence $U_\varepsilon = \bar{f}_\varepsilon = 0$ and

$$\begin{aligned}\|u_\varepsilon' - U_\varepsilon'\|_{L^2(\Omega)} = \|u_\varepsilon'\|_{L^2(\Omega)} &= \frac{\varepsilon}{\sqrt{6}} = \frac{2^{-K}}{\sqrt{6}} \\ &\ll 2^{-k} = h = \|hf_\varepsilon\|_{L^2(\Omega)} = \operatorname{osc}(U_\varepsilon,\mathscr{T}) = \mathscr{E}(U_\varepsilon,\mathscr{T}).\end{aligned}$$

Therefore, the ratio $\|u_\varepsilon' - U_\varepsilon'\|_{L^2(\Omega)}/\mathscr{E}(U_\varepsilon,\mathscr{T})$ can be made arbitrarily small by increasing K/k, and $\operatorname{osc}(U_\varepsilon,\mathscr{T})$ accounts for the discrepancy. On the other hand, measuring the oscillation in $H^{-1}(\Omega)$, as suggested in [13, 69],

$$\|f_\varepsilon - \bar{f}_\varepsilon\|_{H^{-1}(\Omega)} = \|f_\varepsilon\|_{H^{-1}(\Omega)} = \|u_\varepsilon'\|_{L^2(\Omega)} \approx \varepsilon,$$

would avoid overestimation but brings us back to the question how to (approximately) evaluate the $H^{-1}(\Omega)$-norm at acceptable cost.

This 1d example can be extended via a checkerboard pattern to any dimension.

We see that $\operatorname{osc}(U,\mathscr{T})$ may be dominant in early stages of the adaptive iteration (4). Therefore, it cannot be ignored in an optimality analysis without fineness assumptions on the initial mesh $\mathscr{T}_0$; compare with Example 9.1.

We conclude by combing the two global bounds in Theorem 6.1 and Corollary 6.2.

Theorem 6.3 (Asymptotic Equivalence). *Let u and U be exact and Galerkin solutions of the model problem and its standard discretization. Then the error estimator (82) is asymptotically equivalent to the error:*

$$\frac{1}{\alpha_2}\Big(\mathscr{E}(U,\mathscr{T})-\operatorname{osc}(U,\mathscr{T})\Big)\lesssim\|\nabla(u-U)\|_{L^2(\Omega)}\lesssim\frac{1}{\alpha_1}\mathscr{E}(U,\mathscr{T}) \tag{101}$$

where $0<\alpha_1\leq\alpha_2$ are the smallest and largest global eigenvalues of $\boldsymbol{A}$ and the hidden constants depend only on the shape coefficient of $\mathscr{T}$, the dimension d and the polynomial degrees for $\overline{r}$ and $\overline{j}$.

We thus have derived a computable quantity that may be used to stop the adaptive iteration (4) and, in view of the local lower bound in Sect. 6.3.3, the indicators may be used to provide the problem-specific information for local refinement.

6.4 Problems

Problem 6.1. The gap in (75) is dictated by $\|\mathscr{B}\|/\alpha$. Determine this quantity for the model problem in Sect. 2.2.1 and

(a) $\|v\|_{\mathbb{V}}=|v|_{1,\Omega}$,
(b) $\|v\|_{\mathbb{V}}=\big(\int_\Omega\nabla v\cdot A\nabla v\big)^{1/2}$.

Problem 6.2. Prove the scaled trace inequality (Corollary 6.1)

$$\|v\|_{L^2(S)}\lesssim h_S^{-1/2}\|v\|_{L^2(T)}+h_S^{1/2}\|\nabla v\|_{L^2(T)}\quad\text{for all } v\in H^1(T).$$

Problem 6.3. Show that, up to oscillation terms, the jump residual

$$\eta_{\mathscr{J}}(U,\mathscr{T})=\Big(\sum_{S\in\mathscr{S}}\|h^{1/2}j\|_{L^2(S)}^2\Big)^{1/2}$$

bounds $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, which entails that the estimator $\mathscr{E}(U,\mathscr{T})$ is dominated by $\eta_{\mathscr{J}}(U,\mathscr{T})$. To this end, revise the proof of the upper bound for $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, use

$$c_z=\frac{1}{\int_{\omega_z}\phi_z}\int_{\omega_z}r\,\phi_z.$$

and rewrite $\int_{\omega_z}r\,(w-c_z)\,\phi_z$ by exploiting this weighted L^2-orthogonality.

Problem 6.4. Considering the model problem with its standard discretization, derive the upper a posteriori error bound without using the discrete partition of unity.

To this end use (76) and combine the scaled trace inequality (77) with the local interpolation error estimate (63). Derive as an intermediate step the upper bound:

$$|\langle \mathscr{R}, w\rangle| \leq \sum_{T\in\mathscr{T}} \mathscr{E}(U,T)\|\nabla w\|_{L^2(N(T))},$$

with $N(T)$ from (55). Discuss the differences of the two derivations.

This form of the upper bound is useful in Chap. 7.

Problem 6.5. Verify that a suitable multiple of the Verfürth cut-off function (90) satisfies the properties (86). To this end, recall Lemma 3.1. Repeat for the Dörfler cut-off function.

Problem 6.6. (Try this problem after Problem 6.5.) Show that the choice (90) for η_T verifies, for all $p \in \mathbb{P}_l(T)$,

$$\int_T p^2 \lesssim \int_T p^2\eta_T, \qquad \|\nabla(p\eta_T)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on l and the shape coefficient of T. To this end, recall the equivalence of norms in finite-dimensional spaces. Derive the estimate

$$h_T\,\|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}$$

for $r \in \mathbb{P}_l(T)$.

Problem 6.7. Consider the model problem and its discretization for $d = 2$ and $n = 1$. Let U_1 be the solution over a triangulation $\mathscr{T}_1$ and U_2 the solution over $\mathscr{T}_2$, where $\mathscr{T}_2$ has been obtained by applying at least 3 bisections to every triangle of $\mathscr{T}_1$. Moreover, suppose that f is piecewise constant over $\mathscr{T}_1$. Show

$$\|\nabla(U_2 - U_1)\|_{L^2(\Omega)} \geq \|h_1 f\|_{L^2(\Omega)},$$

where h_1 is the mesh-size function of $\mathscr{T}_1$.

Problem 6.8. Verify that a suitable multiple of the Verfürth cut-off function (92) satisfies the properties (91). Repeat for the Dörfler cut-off function.

Problem 6.9. Let S be a side of a simplex T. Show that for each $q \in \mathbb{P}_l(S)$ there exists a $p \in \mathbb{P}_l(T)$ such that

$$p = q \text{ on } S \qquad \text{and} \qquad \|p\|_{L^2(T)} \lesssim h_T^{1/2}\|q\|_{L^2(S)}.$$

Problem 6.10. Let S be a side of a simplex T. Show that the choice (92) for η_S verifies, for all $q \in \mathbb{P}_k(\mathscr{S})$ and all $p \in \mathbb{P}_l(T)$,

$$\int_S q^2 \lesssim \int_S q^2\eta_S, \qquad \|\nabla(p\eta_S)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on l and the shape coefficient of T.

Problem 6.11. Derive the estimate (95), where $\overline{r}$ and $\overline{j}$ are piecewise polynomials of degree l_1 and l_2.

Problem 6.12. Generalize Remark 6.1 to polynomial degree $n \geq 2$.

Problem 6.13. Supposing (98), formulate and prove an abstract local lower bound in the spirit of Lemma 6.1.

Problem 6.14. Derive a posteriori error bounds for the energy norm

$$|||v|||_\Omega = \left(\int_\Omega \nabla v \cdot \boldsymbol{A} \nabla v \right)^{1/2}$$

and compare with Theorem 6.3.

7 Adaptivity: Convergence

The purpose of this chapter is to prove that the standard adaptive finite element method characterized by the iteration

$$\text{SOLVE} \longrightarrow \text{ESTIMATE} \longrightarrow \text{MARK} \longrightarrow \text{REFINE} \tag{102}$$

generates a sequence of discrete solutions converging to the exact one. This will be established under assumptions that are quite weak or even minimal. In particular, we will not suppose any regularity of the exact solution that goes beyond the natural one in the variational formulation. We therefore can expect only a plain convergence result that does not give any convergence rate in terms of degrees of freedom. The assumptions on the general variational problem allow for various examples that are of quite different from the model problem in Sect. 2.2.1. Examples are left as problems to the reader.

The presentation is based on the basic convergence result by Morin et al. [55] and the modifications by Siebert [67].

7.1 The adaptive algorithm

Given a continuous bilinear form $\mathscr{B}\colon \mathbb{V} \times \mathbb{W} \to \mathbb{R}$ and an element $f \in \mathbb{W}^*$ we consider the variational problem

$$u \in \mathbb{V}: \quad \mathscr{B}[u, w] = \langle f, w \rangle \qquad \text{for all } w \in \mathbb{W} \tag{103}$$

introduced in Chap. 2. We assume that $\mathscr{B}$ satisfies the inf-sup condition (21).

For the adaptive approximation of the solution u we consider a loop of the form (102). To be more precise, starting with an initial conforming triangulation $\mathscr{T}_0$ of the

underlying domain Ω and a refinement procedure REFINE as described in Sect. 4.4 we execute an iteration of the following main steps:

$$\begin{aligned}
&(1)\quad U_k := \mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T}_k), \mathbb{W}(\mathscr{T}_k)\big);\\
&(2)\quad \{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k} := \mathsf{ESTIMATE}\big(U_k, \mathscr{T}_k\big);\\
&(3)\quad \mathscr{M}_k := \mathsf{MARK}\big(\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k}, \mathscr{T}_k\big);\\
&(4)\quad \mathscr{T}_{k+1} := \mathsf{REFINE}\big(\mathscr{T}_k, \mathscr{M}_k\big), \text{ increment } k \text{ and go to Step (1).}
\end{aligned} \tag{104}$$

In practice, a stopping test is used after Step (2) for terminating the iteration; here we shall ignore it for notational convenience. Besides the initial grid $\mathscr{T}_0$ and the module REFINE from Sect. 4.4, the realization of these steps requires the following objects and modules:

- For any grid $\mathscr{T} \in \mathbb{T}$, there are finite element spaces $\mathbb{V}(\mathscr{T})$ and $\mathbb{W}(\mathscr{T})$ and the module SOLVE outputs the corresponding Petrov-Galerkin approximation $U_\mathscr{T}$ to u.
- A module ESTIMATE that, given a grid $\mathscr{T} \in \mathbb{T}$ and the corresponding discrete solution $U_\mathscr{T}$, outputs the a posteriori error estimator $\{\mathscr{E}_\mathscr{T}(U_\mathscr{T},T)\}_{T\in\mathscr{T}}$, where the so-called indicator $\mathscr{E}_\mathscr{T}(U_\mathscr{T},T) \geq 0$ is associated with the element $T \in \mathscr{T}$.
- A strategy in the module MARK that, based upon the a posteriori error indicators $\{\mathscr{E}_\mathscr{T}(U_\mathscr{T},T)\}_{T\in\mathscr{T}}$, collects elements of $\mathscr{T}$ in $\mathscr{M}$, which serves as input for REFINE.

Obviously, the modules SOLVE and ESTIMATE do strongly depend on the variational problem, i. e., on data $\mathscr{B}$ and f; compare with Sects. 3.1.3 and 6. For convenience of notation we have suppressed this dependence. The refinement module REFINE is problem independent and the same applies in general to the module MARK. We list the most popular marking strategies for (104):

(a) Maximum Strategy: For given parameter $\theta \in [0,1]$ we let

$$\mathscr{M} = \big\{T \in \mathscr{T} \mid \mathscr{E}_\mathscr{T}(U_\mathscr{T},T) \geq \theta \mathscr{E}_{\mathscr{T},\max}\big\} \quad \text{with} \quad \mathscr{E}_{\mathscr{T},\max} = \max_{T\in\mathscr{T}} \mathscr{E}_\mathscr{T}(U_\mathscr{T},T).$$

(b) Equidistribution Strategy: For given parameter $\theta \in [0,1]$ we let

$$\mathscr{M} = \Big\{T \in \mathscr{T} \mid \mathscr{E}_\mathscr{T}(U_k,T) \geq \theta \mathscr{E}_\mathscr{T}(U_\mathscr{T},\mathscr{T})/\sqrt{\#\mathscr{T}}\Big\}.$$

(c) Dörfler's Strategy: For given parameter $\theta \in (0,1]$ we let $\mathscr{M} \subset \mathscr{T}$ such that

$$\mathscr{E}_\mathscr{T}(U_\mathscr{T},\mathscr{M}) \geq \theta \mathscr{E}_\mathscr{T}(U_\mathscr{T},\mathscr{T}).$$

For efficiency reasons one wants to mark as few elements as possible. This can be achieved by selecting the elements holding the largest indicators, whence

$$\min_{T\in\mathscr{M}} \mathscr{E}_\mathscr{T}(U_\mathscr{T},T) \geq \max_{T\in\mathscr{T}\setminus\mathscr{M}} \mathscr{E}_\mathscr{T}(U_\mathscr{T},T).$$

The objective of this chapter is to prove that, under quite weak assumptions on the modules SOLVE, ESTIMATE, and MARK, the sequence $\{U_k\}_{k\geq 0}$ of discrete solutions converges to u, i. e.,

$$\lim_{k\to\infty} \|U_k - u\|_{\mathbb{V}} = 0. \tag{105}$$

This is a priori not clear, since the estimator only provides a global upper bound for the error. All the techniques used in Chap. 5 are based on completely local interpolation estimates and therefore cannot be used when working with an estimator. Then again, as long as $U_k \neq u$ the estimator is non-zero. This should lead to convergence provided that the indicators $\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k}$ pick up some local error information and the selection of elements in MARK accounts for that.

For convenience of notation we replace in what follows the argument $\mathscr{T}_k$ by a subscript k, for instance we set $\mathbb{V}_k := \mathbb{V}(\mathscr{T}_k)$.

7.2 Density and convergence

Plain convergence for a sequence of uniformly refined grids is a simple consequence of density. To see this, we set $\mathscr{M}_k = \mathscr{T}_k$ in each iteration of (104). Then (61) implies

$$h_{\max}(\mathscr{T}_k) := \max\{h_T \mid T \in \mathscr{T}_k\} \leq D_2\, 2^{-kb/d} \to 0 \qquad \text{as } k \to \infty,$$

if elements in $\mathscr{M}_k$ are scheduled for $b \geq 1$ bisections. Furthermore, let $\mathbb{V}^s \subset \mathbb{V}$ be a dense subspace and $I_k\colon \mathbb{V}^s \to \mathbb{V}_k$ be an interpolation operator with

$$\|I_k v - v\|_{\mathbb{V}} \leq C\, h^s_{\max}(\mathscr{T}_k)\|v\|_{\mathbb{V}^s} \qquad \text{for all } v \in \mathbb{V}^s \tag{106}$$

for $s > 0$. In case of the model problem with $\mathbb{V} = H^1_0(\Omega)$ we could take for instance $\mathbb{V}_k$ to be conforming Lagrange finite elements over $\mathscr{T}_k$, and I_k the Lagrange interpolant, which satisfies (106) with $s = 1$ on $\mathbb{V}^2 = H^2(\Omega)\cap H^1_0(\Omega)$; compare with Remark 5.3. For any $v \in \mathbb{V}$ and $\bar{v} \in \mathbb{V}^s$ we then derive

$$\|I_k\bar{v} - v\|_{\mathbb{V}} \leq \|I_k\bar{v} - \bar{v}\|_{\mathbb{V}} + \|\bar{v} - v\|_{\mathbb{V}} \leq C\, h^s_{\max}(\mathscr{T}_k)\|\bar{v}\|_{\mathbb{V}^s} + \|\bar{v} - v\|_{\mathbb{V}}.$$

For given v and ε we first can choose $\bar{v} \in \mathbb{V}^s$ such that $\|v - \bar{v}\|_{\mathbb{V}} \leq \varepsilon/2$ by density of $\mathbb{V}^s$ in $\mathbb{V}$. Then (106) implies $C\, h^s_{\max}(\mathscr{T}_k)\|\bar{v}\|_{\mathbb{V}^s} \leq \varepsilon/2$ provided k is sufficiently large, whence $\|I_k\bar{v} - v\|_{\mathbb{V}} \leq \varepsilon$. Therefore,

$$\lim_{k\to\infty} \min_{V_k\in\mathbb{V}_k} \|V_k - v\|_{\mathbb{V}} = 0 \qquad \text{for all } v \in \mathbb{V}$$

or, equivalently,

$$\mathbb{V} = \overline{\bigcup_{k\geq 0} \mathbb{V}_k}. \tag{107}$$

This density property already implies convergence if the sequence $\{\mathbb{V}_k, \mathbb{W}_k\}_{k\geq 0}$ is stable, i. e., it satisfies a uniform inf-sup condition. Recalling the quasi-best approximation property of the Petrov-Galerkin solution U_k established in Theorem 3.2, stability of the discretization yields

$$\|U_k - u\|_{\mathbb{V}} \leq \frac{\|\mathscr{B}\|}{\beta} \min_{V_k \in \mathbb{V}_k} \|V_k - u\|_{\mathbb{V}} \to 0 \qquad \text{as } k \to \infty, \tag{108}$$

thanks to density (107). Note, that this convergence result holds true irrespective of any regularity property of u beyond $\mathbb{V}$.

Assume now that the sequence $\{\mathscr{T}_k\}_{k\geq 0}$ is adaptively generated. We observe that (107) still holds whenever

$$\lim_{k\to\infty} h_{\max}(\mathscr{T}_k) = 0, \tag{109}$$

whence (108) is also true. But (109) does not hold in general for an adaptively generated sequence of meshes, as was already observed by Babuška and Vogelius [7]. Recalling the definition of the mesh-size function

$$h_k \in L^\infty(\Omega): \qquad h_{k|T} = |T|^{1/d}, \quad T \in \mathscr{T}_k$$

in Sect. 4.3 and its L^∞ limit $h_\infty \in L^\infty(\Omega)$ of Lemma 4.2, Eq. (109) is equivalent to $h_\infty \equiv 0$ in Ω. If $h_\infty \not\equiv 0$, then there exists an $x \in \Omega \setminus \Gamma_\infty$ with $h_\infty(x) > 0$. This implies that there is an element $T \ni x$ and an iteration counter $K = K(x)$ such that $T \in \mathscr{T}_k$ for all $k \geq K$.

This motivates to split the triangulations $\mathscr{T}_k$ into two classes of elements

$$\mathscr{T}_k^+ := \bigcap_{\ell\geq k} \mathscr{T}_k = \{T \in \mathscr{T}_k \mid T \in \mathscr{T}_\ell \ \forall \ell \geq k\}, \quad \text{and} \quad \mathscr{T}_k^0 := \mathscr{T}_k \setminus \mathscr{T}_k^+. \tag{110}$$

The set $\mathscr{T}_k^+$ contains all elements that are not refined after iteration k and we observe that the sequence $\{\mathscr{T}_k^+\}_{k\geq 0}$ is nested, i. e., $\mathscr{T}_\ell^+ \subset \mathscr{T}_k^+$ for all $k \geq \ell$. The set $\mathscr{T}_k^0$ contains all elements that are refined at least once more in a forthcoming step of the adaptive procedure. Since the sequence $\{\mathscr{T}_k^+\}_{k\geq 0}$ is nested the set

$$\mathscr{T}^+ := \bigcup_{k\geq 0} \mathscr{T}_k^+$$

is well-defined and we conclude

$$h_\infty \equiv 0 \qquad \text{if and only if} \qquad \mathscr{T}^+ = \emptyset.$$

If $\mathscr{T}^+ \neq \emptyset$ then the finite element spaces cannot be dense in $\mathbb{V}$ since inside $T \in \mathscr{T}^+$ we can only approximate discrete functions. Therefore, taking into accout the arguments at the beginning of the section, we have that (107) is equivalent to $h_\infty \equiv 0$.

On the other hand, when using adaptivity we do not aim at approximating *all* functions in $\mathbb{V}$ but rather one single function, namely the solution u to (103). A necessary condition for being able to approximate u is

$$\lim_{k\to\infty} \min_{V_k\in\mathbb{V}_k} \|u - V_k\|_{\mathbb{V}} = 0.$$

Assuming that the finite element spaces are nested, the space

$$\mathbb{V}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{V}_k}$$

is well-defined and we can approximate u by discrete functions if and only if $u \in \mathbb{V}_\infty$. We realize that $\mathbb{V}_\infty$ is defined via the adaptively generated spaces $\mathbb{V}_k$. Therefore, $u \in \mathbb{V}_\infty$ hinges on properties of the modules SOLVE, ESTIMATE, MARK, and REFINE. In addition, if $\mathbb{V}_\infty$ is a proper subspace of $\mathbb{V}$ and $u \in \mathbb{V}_\infty$ then u is locally a discrete function. This implies, that the adaptive method must only decide not to refine an element any more if u locally belongs to the finite element space, for instance u is affine in some part of the domain in case of Courant elements.

But this is not the generic case. If u is not locally discrete, then the decisions of the adaptive method have to yield $\mathscr{T}^+ = \emptyset$, and if so, convergence is a direct consequence of density as for uniform refinement. We aim at a convergence result for adaptive finite elements that just relies on this density argument in this case. In doing this we shall use a *local density* property of the finite element spaces in the region $\{h_\infty \equiv 0\}$ and properties of the adaptive method in its complement.

7.3 Properties of the problem and the modules

In this section we state structural assumptions on the Hilbert spaces $\mathbb{V}$ and $\mathbb{W}$ and the modules SOLVE, ESTIMATE, and MARK. For notational convenience we use '$a \lesssim b$' for '$a \leq Cb$' whenever the constant C only depends on $\mathscr{T}_0$ and data of (103) like $\mathscr{B}$ and f.

7.3.1 Properties of Hilbert spaces

We assume that $\mathbb{V}$ is a subspace of $L^2(\Omega;\mathbb{R}^m)$ with some $m \in \mathbb{N}$ and that $\|\cdot\|_{\mathbb{V}}$ is an L^2-type integral norm implying the following properties: The square of the norm $\|\cdot\|_{\mathbb{V}(\Omega)}$ is set-additive, i. e., for any subset $\omega \subset \Omega$ that is decomposed into $\omega = \omega_1 \cup \omega_2$ with $|\omega_1 \cap \omega_2| = 0$ there holds

$$\|v\|^2_{\mathbb{V}(\omega)} = \|v\|^2_{\mathbb{V}(\omega_1)} + \|v\|^2_{\mathbb{V}(\omega_2)} \qquad \text{for all } v \in \mathbb{V}(\omega). \tag{111}$$

In addition, we ask $\|\cdot\|_{\mathbb{V}}$ to be absolutely continuous with respect to the Lebesgue measure, this is, for any $v \in \mathbb{V}$ holds

$$\|v\|_{\mathbb{V}(\omega)} \to 0 \quad \text{as} \quad |\omega| \to 0.$$

Finally we require $\mathbb{W}$ to have the same properties.

7.3.2 Properties of SOLVE

For any grid $\mathscr{T} \in \mathbb{T}$ we assume the existence of a pair of finite element spaces $\{\mathbb{V}(\mathscr{T}), \mathbb{W}(\mathscr{T})\}$ and suppose the following properties:

(1) They are conforming

$$\mathbb{V}(\mathscr{T}) \subset \mathbb{V}, \quad \mathbb{W}(\mathscr{T}) \subset \mathbb{W} \qquad \text{for all } \mathscr{T} \in \mathbb{T} \tag{112a}$$

and nested

$$\mathbb{V}(\mathscr{T}) \subset \mathbb{V}(\mathscr{T}_*), \quad \mathbb{W}(\mathscr{T}) \subset \mathbb{W}(\mathscr{T}^*) \qquad \text{for all } \mathscr{T} \leq \mathscr{T}_* \in \mathbb{T}. \tag{112b}$$

(2) The finite element spaces are a stable discretization, i. e., there exists $\beta > 0$ such that for all $\mathscr{T} \in \mathbb{T}$

$$\dim \mathbb{V}(\mathscr{T}) = \dim \mathbb{W}(\mathscr{T}) \quad \text{and} \quad \inf_{\substack{V \in \mathbb{V}(\mathscr{T}) \\ \|V\|_{\mathbb{V}}=1}} \sup_{\substack{W \in \mathbb{W}(\mathscr{T}) \\ \|W\|_{\mathbb{W}}=1}} \mathscr{B}[V, W] \geq \beta. \tag{112c}$$

(3) Let $\mathbb{W}^s \subset \mathbb{W}$ be a dense sub-space with norm $\|\cdot\|_{\mathbb{W}^s}$ such that $\|\cdot\|^2_{\mathbb{W}^s}$ is set-additive and let $I_{\mathscr{T}} \in L(\mathbb{W}^s, \mathbb{W}(\mathscr{T}))$ be a continuous, linear interpolation operator such that

$$\|w - I_{\mathscr{T}} w\|_{\mathbb{W}(T)} \lesssim h_T^s \|w\|_{\mathbb{W}^s(T)} \qquad \text{for all } T \in \mathscr{T} \text{ and } w \in \mathbb{W}^s \tag{112d}$$

with $s > 0$.

(4) We suppose that $\mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T}), \mathbb{W}(\mathscr{T})\big)$ outputs the *exact* Petrov-Galerkin approximation of u, i. e.,

$$U_{\mathscr{T}} \in \mathbb{V}(\mathscr{T}): \qquad \mathscr{B}[U_{\mathscr{T}}, W] = \langle f, W \rangle \qquad \text{for all } w \in \mathbb{W}(\mathscr{T}). \tag{112e}$$

This entails exact integration and linear algebra; see Remarks 3.6 and 3.7.

Note, that for non-adaptive realizations of (104), condition (112c) is necessary for the well-posedness of (112e) and convergence irrespective of $f \in \mathbb{W}^*$; compare with Problem 3.2. Although phrasing the interpolation estimate (112d) as a condition on the choice of the finite element space, the construction of any finite element space is based on such a local approximation property.

7.3.3 Properties of ESTIMATE

Given a grid $\mathscr{T} \in \mathbb{T}$ and the Petrov-Galerkin approximation $U_{\mathscr{T}} \in \mathbb{V}_{\mathscr{T}}$ of (112e) we suppose that we can *compute* a posteriori error indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T)\}_{T \in \mathscr{T}}$ by

$$\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T)\}_{T \in \mathscr{T}} = \mathsf{ESTIMATE}(U_{\mathscr{T}}, \mathscr{T})$$

with the following properties:

(1) The estimator provides the following upper bound for the residual $\mathscr{R}_{\mathscr{T}} \in \mathbb{W}^*$ of $U_{\mathscr{T}}$:

$$|\langle \mathscr{R}_{\mathscr{T}}, w\rangle| \lesssim \sum_{T\in\mathscr{T}} \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\|w\|_{\mathbb{W}(N_{\mathscr{T}}(T))} \qquad \text{for all } w\in\mathbb{W}. \tag{113a}$$

(2) The estimator is efficient in that it satisfies the continuous local lower bound

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \lesssim \|U_{\mathscr{T}}-u\|_{\mathbb{V}(N_{\mathscr{T}}(T))} + \operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}},T) \qquad \text{for all } T\in\mathscr{T}, \tag{113b}$$

where the oscillation indicator $\operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}},T)$ satisfies

$$\operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}},T) \lesssim h_T^q\Big(\|U_{\mathscr{T}}\|_{\mathbb{V}(N_{\mathscr{T}}(T))} + \|D\|_{L^2(N_{\mathscr{T}}(T))}\Big). \tag{113c}$$

Hereafter, $q>0$ and $D\in L^2(\Omega)$ is given by data of (103).

The upper bound as stated in (113a) is usually an intermediate step when deriving a posteriori error estimates; compare with Problem 6.4. It allows us to extract *local* information about the residual. This is not possible when directly using the global upper bound $\|U_{\mathscr{T}}-u\|_{\mathbb{V}(\Omega)} \lesssim \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{T})$.

7.3.4 Properties of MARK

The last module for the adaptive algorithm is a function

$$\mathscr{M} = \mathsf{MARK}\big(\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}, \mathscr{T}\big)$$

that, given a mesh $\mathscr{T}\in\mathbb{T}$ and indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}$, selects elements subject to refinement. Given a fixed function $g:\mathbb{R}_0^+\to\mathbb{R}_0^+$ that is continuous at 0 with $g(0)=0$, we ask that the set $\mathscr{M}$ of marked elements has the property

$$\max\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \mid T\in\mathscr{T}\setminus\mathscr{M}\} \le g\big(\max\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \mid T\in\mathscr{M}\}\big). \tag{114}$$

Marking criterion (114) implies that all indicators in $\mathscr{T}$ are controlled by the maximal indicator in $\mathscr{M}$. Marking strategies that pick up the elements holding the largest indicator, as those from Sect. 7.1, satisfy (114) with $g(s)=s$.

7.4 Convergence

In this section we show that the realization of (104) generates a sequence of Petrov-Galerkin solutions that converges to the true solution in $\mathbb{V}$ under the above assumptions.

Theorem 7.1 (Convergence). *Let u be the exact solution of (103) and suppose that (21) holds. Let the finite element spaces and the functions* SOLVE, ESTIMATE, *and* MARK *satisfy (112), (113), and (114), respectively.*

Then the sequence of Galerkin approximations $\{U_k\}_{k\geq 0}$ generated by iteration (104) satisfies

$$\lim_{k\to\infty} \|U_k - u\|_{\mathbb{V}} = 0 \qquad \text{and} \qquad \lim_{k\to\infty} \mathscr{E}_k(U_k, \mathscr{T}_k) = 0.$$

In particular, any prescribed tolerance $\mathsf{TOL} > 0$ for the estimator is reached in a finite number of steps. In other words: there is an iteration k^* with

$$\|U_{k^*} - u\|_{\mathbb{V}} \lesssim \mathscr{E}_{k^*}(U_{k^*}, \mathscr{T}_{k^*}) \leq \mathsf{TOL}.$$

We split the proof in several steps.

7.4.1 Two limits

In this paragraph we give another generalization of (56) for a sequence of adaptively generated triangulations. In combination with the interpolation estimate (112d) this result yields a local density property of adaptively generated finite element spaces. Additionally we show that for any realization of (104) the Petrov-Galerkin solutions are a Cauchy-sequence in $\mathbb{V}$.

The uniform convergence $h_k \to h_\infty$ shown in Lemma 4.2 helps to locate the set $\{h_\infty \equiv 0\}$ in terms of the splitting $\mathscr{T}_k = \mathscr{T}_k^0 \cup \mathscr{T}_k^+$ introduced in (110). According to $\mathscr{T}_k^0$ and $\mathscr{T}_k^+$ we decompose the domain Ω into

$$\bar{\Omega} = \Omega(\mathscr{T}_k^+) \cup \Omega(\mathscr{T}_k^0) =: \Omega_k^+ \cup \Omega_k^0,$$

where for any sub-triangulation $\mathscr{T}_k' \subset \mathscr{T}_k$ we let

$$\Omega(\mathscr{T}_k') := \bigcup \{T : T \in \mathscr{T}_k'\}$$

be the part of Ω covered by $\mathscr{T}_k'$. A direct consequence of Lemma 4.2 is the following result.

Corollary 7.1 ($\{h_\infty \equiv 0\}$). *Denote by χ_k^0 the characteristic function of Ω_k^0.*

Then the definition of $\mathscr{T}_k^0$ implies

$$\lim_{k\to\infty} \left\| h_k \chi_k^0 \right\|_{L^\infty(\Omega)} = \lim_{k\to\infty} \|h_k\|_{L^\infty(\Omega_k^0)} = 0.$$

Proof. The definition of $\mathscr{T}_k^0$ implies that all elements in $\mathscr{T}_k^0$ are refined at least once. Hence, $h_\infty \leq 2^{-\frac{1}{d}} h_k$ in Ω_k^0, yielding $\left(1 - 2^{-1/d}\right) h_k \leq h_k - h_\infty$ in Ω_k^0. This in turn implies with $\gamma = 1 - 2^{1/d} > 0$

$$\left\| h_k \chi_k^0 \right\|_{L^\infty(\Omega)} \leq \gamma^{-1} \left\| (h_k - h_\infty) \chi_k^0 \right\|_{L^\infty(\Omega)} \leq \gamma^{-1} \|h_k - h_\infty\|_{L^\infty(\Omega)} \to 0$$

for $k \to \infty$ thanks to Lemma 4.2. □

Remark 7.1 (Local Density). We employ set-additivity of $\|\cdot\|^2_{\mathbb{W}^s}$ combined with the local approximation property (112d) to deduce for any sub-triangulation $\mathscr{T}_k' \subset \mathscr{T}_k$ and any $\bar{w} \in \mathbb{W}^s$ the local interpolation estimate

$$\|\bar{w} - I_k \bar{w}\|_{\mathbb{W}(\Omega(\mathscr{T}_k'))} \lesssim \|h_k^s\|_{L^\infty(\Omega(\mathscr{T}_k'))} \|\bar{w}\|_{\mathbb{W}^s(\Omega(\mathscr{T}_k'))}. \tag{115}$$

Using this estimate for $\mathscr{T}_k' = \mathscr{T}_k^0$ the above corollary implies

$$\|I_k \bar{w} - \bar{w}\|_{\mathbb{V}(\Omega_k^0)} \lesssim \|h_k^s\|_{L^\infty(\Omega_k^0)} \|\bar{w}\|_{\mathbb{W}^s(\Omega)} \qquad \text{for all } \bar{w} \in \mathbb{W}^s.$$

For any pair $w \in \mathbb{W}$ and $\bar{w} \in \mathbb{W}^s$ we then argue as in Sect. 7.2 for uniform refinement but restricted to subdomain $\Omega(\mathscr{T}_k^0)$ to conclude the 'local density'

$$\lim_{k\to\infty} \min_{W_k \in \mathbb{W}_k} \|w - W_k\|_{\mathbb{V}(\Omega_k^0)} = 0 \qquad \text{for all } w \in \mathbb{W}. \tag{116}$$

We use the interpolation estimate (115) in Proposition 7.1 below.

We next turn to the sequence $\{U_k\}_{k\geq 0}$ of approximate solutions. For characterizing the limit of this sequence we need the spaces

$$\mathbb{V}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{V}_k} \qquad \text{and} \qquad \mathbb{W}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{W}_k}.$$

Lemma 7.1 (Convergence of Petrov-Galerkin Approximations). *Assume that the sequence* $\{(\mathbb{V}_k, \mathbb{W}_k)\}_{k\geq 0}$ *satisfies* (112c) *and* (112b).

Then the sequence $\{U_k\}_{k\geq 0}$ *of approximate solutions converges in* $\mathbb{V}$ *to the solution* u_∞ *with respect to the pair* $(\mathbb{V}_\infty, \mathbb{W}_\infty)$*, which is characterized by*

$$u_\infty \in \mathbb{V}_\infty : \qquad \mathscr{B}[u_\infty, w] = f(w) \qquad \textit{for all } w \in \mathbb{W}_\infty. \tag{117}$$

Proof. 1 Let us first prove that the pair $(\mathbb{V}_\infty, \mathbb{W}_\infty)$ satisfies the inf-sup condition

$$\inf_{\substack{v\in\mathbb{V}_\infty \\ \|v\|_{\mathbb{V}}=1}} \sup_{\substack{w\in\mathbb{W}_\infty \\ \|w\|_{\mathbb{W}}=1}} \mathscr{B}[v,w] \geq \beta, \qquad \inf_{\substack{w\in\mathbb{W}_\infty \\ \|w\|_{\mathbb{W}}=1}} \sup_{\substack{v\in\mathbb{V}_\infty \\ \|v\|_{\mathbb{V}}=1}} \mathscr{B}[v,w] \geq \beta \tag{118}$$

with $\beta > 0$ from (112c).

To this end, fix first any $v \in \mathbb{V}_\infty \setminus \{0\}$ and choose a sequence $\{V_k\}_{k\geq 0}$ of functions $V_k \in \mathbb{V}_k$ such that $V_k \to v$ in $\mathbb{V}$ as $k \to \infty$. Moreover, with the help of (112c) choose a sequence $\{W_k\}_{k\geq 0}$ of functions $W_k \in \mathbb{W}_k$ such that

$$\|W_k\|_{\mathbb{W}} = 1 \qquad \text{and} \qquad \mathscr{B}[V_k, W_k] \geq \beta \|V_k\|_{\mathbb{V}}. \tag{119}$$

Thanks to (112a), the sequence $\{W_k\}_{k\geq 0}$ is in $\mathbb{W}$. Since the latter is reflexive, there exists a subsequence $\{W_{k_j}\}_{j\geq 0}$ and a function $w \in \mathbb{W}$ such that $W_{k_j} \rightharpoonup w$ weakly in $\mathbb{W}$ as $j \to \infty$. Since $\mathbb{W}_\infty$ is closed and convex as well as $\|\cdot\|_{\mathbb{W}}$ weakly lower

semicontinuous, we have $w \in \mathbb{W}_\infty$ and $\|w\|_{\mathbb{W}} \le \liminf_{j\to\infty} \|W_{k_j}\|_{\mathbb{W}} = 1$. Combing this with (112c) yields

$$\mathscr{B}[v, w] \ge \beta \|v\|_{\mathbb{V}} \ge \beta \|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}.$$

In view of the first inequality, $w \neq 0$ and the first part of (118) is proved.

Proposition 3.1 states that (112c) is equivalent to

$$\inf_{\substack{W\in\mathbb{W}(\mathscr{T})\\ \|W\|_{\mathbb{W}}=1}} \sup_{\substack{V\in\mathbb{V}(\mathscr{T})\\ \|V\|_{\mathbb{V}}=1}} \mathscr{B}[V, W] \ge \beta. \tag{120}$$

In the same way, but using (120) instead of (112c), we show that for any $w \in \mathbb{W}_\infty$ there exists $v \in \mathbb{V}_\infty \setminus \{0\}$ such that $\mathscr{B}[v, w] \ge \beta \|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}$. This shows (118).

2 The spaces $\mathbb{V}_\infty \subset \mathbb{V}$ and $\mathbb{W}_\infty \subset \mathbb{W}$ are closed and thus Hilbert spaces. The bilinear form $\mathscr{B}$ is continuous on $\mathbb{V}_\infty \times \mathbb{W}_\infty$ and satisfies the inf-sup condition (118). Therefore, by Theorem 2.2 there exits a unique $u_\infty \in \mathbb{V}_\infty$ with (117).

3 By construction, $\mathbb{V}_k \subset \mathbb{V}_\infty$, which implies that the Petrov-Galerkin solution U_k is a $\|\cdot\|_{\mathbb{V}}$-quasi-optimal choice in $\mathbb{V}_k$ with respect to u_∞, i. e., there holds

$$\|u_\infty - U_k\|_{\mathbb{V}} \le \frac{\|B\|}{\beta} \min_{V\in\mathbb{V}_k} \|u_\infty - V\|_{\mathbb{V}};$$

compare with Theorem 3.2. Besides that, $\bigcup_{k\ge 0} \mathbb{V}_k$ is dense in $\mathbb{V}_\infty$ and therefore

$$\lim_{k\to\infty} \|U_k - u_\infty\|_{\mathbb{V}} = 0. \qquad \square$$

In case of coercive $\mathscr{B}$ the proof is much simpler since coercivity is inherited from $\mathbb{V}$ to $\mathbb{V}_\infty$ and Step 1 of the proof is trivial. Existence of u_∞ is then a direct consequence of Corollary 2.2 (Lax-Milgram theorem). For symmetric and coercive $\mathscr{B}$ the above result has already been shown by Babuška and Vogelius [7].

Lemma 7.1 yields convergence of $U_k \to u_\infty$ in $\mathbb{V}$ as $k \to \infty$ irrespective of the decisions in the module MARK. We are going to prove below that the residual $\mathscr{R}_\infty$ of U_∞ satisfies $\mathscr{R}_\infty = 0$ in $\mathbb{W}^*$. The latter is equivalent to $u_\infty = u$ and thus shows Theorem 7.1. This, of course, hinges on the properties of ESTIMATE and MARK.

7.4.2 Auxiliary results

Next we prove two auxiliary results, namely boundedness of the estimator and convergence of the indicators. Before embarking on this, we observe that the set-additivity of $\|\cdot\|_{\mathbb{V}}^2$ allows us to sum over overlapping patches, if the overlap is finite; compare also with the proof of Theorem 5.1. To be more precise: Local quasi-uniformity of $\mathscr{T}_k$ (55) implies $\#N_k(T) \lesssim 1$ for all $T \in \mathscr{T}_k$. Thus set-additivity (111) of $\|\cdot\|_{\mathbb{V}}^2$ gives for any subset $\mathscr{T}_k' \subset \mathscr{T}_k$ and any $v \in \mathbb{V}$

$$\sum_{T\in\mathscr{T}_k'} \|v\|_{\mathbb{V}(N_k(T))}^2 = \sum_{T\in\mathscr{T}_k'} \sum_{T'\in N_k(T)} \|v\|_{\mathbb{V}(T')}^2 \lesssim \sum_{T\in\mathscr{T}_k^*} \|v\|_{\mathbb{V}(T)}^2 = \|v\|_{\mathbb{V}(\Omega_k^*)}^2 \tag{121}$$

with $\mathcal{T}_k^* = \{T' \in \mathcal{T}_k \mid T' \in N_k(T), T \subset \mathcal{T}_k'\}$ and $\Omega_k^* := \Omega(\mathcal{T}_k^*)$. The same argument applies to $\|\cdot\|_{\mathbb{W}}^2$, $\|\cdot\|_{\mathbb{W}^s}^2$, and $\|\cdot\|_{L^2(\Omega)}^2$.

In the next results we use the stability estimate

$$\mathcal{E}_{\mathcal{T}}(U_{\mathcal{T}}, T) \lesssim \|U_{\mathcal{T}}\|_{\mathbb{V}(N_{\mathcal{T}}(T))} + \|\tilde{D}\|_{L^2(N_{\mathcal{T}}(T))} \qquad \text{for all } T \in \mathcal{T}, \tag{122}$$

where $\tilde{D} = \tilde{D}(u, D) \in L^2(\Omega)$ with D from (113b). This bound can be derived as follows. Combining the lower bound (113b) and the triangle inequality we infer

$$\begin{aligned}\mathcal{E}_{\mathcal{T}}(U_{\mathcal{T}}, T) &\lesssim \|U_{\mathcal{T}} - u\|_{\mathbb{V}(N_{\mathcal{T}}(T))} + \operatorname{osc}_{\mathcal{T}}(U_{\mathcal{T}}, T) \\ &\lesssim \|U_{\mathcal{T}}\|_{\mathbb{V}(N_{\mathcal{T}}(T))} + \|u\|_{\mathbb{V}(N_{\mathcal{T}}(T))} + \|D\|_{L^2(N_{\mathcal{T}}(T))},\end{aligned}$$

where the constant in $\lesssim$ also depends on $\|h_0^q\|_{L^\infty(\Omega)}$ via (113c). Since $\|\cdot\|_{\mathbb{V}(\Omega)}$ is an L^2-type norm, the stability of the indicators (122) is a direct consequence of (113b) with $\tilde{D} = \tilde{D}(u, D) \in L^2(\Omega)$.

Lemma 7.2 (Stability). *Let the finite element spaces and the error the indicators satisfy* (112c) *respectively* (122).

Then the estimators $\mathcal{E}_k(U_k, \mathcal{T}_k)$ are uniformly bounded, i. e.,

$$\mathcal{E}_k(U_k, \mathcal{T}_k) \lesssim 1 \qquad \text{for all } k \geq 0.$$

Proof. Using (121) and the stability of the indicators (122) we derive for all $k \geq 0$

$$\mathcal{E}_k^2(U_k, \mathcal{T}_k) \lesssim \sum_{T \in \mathcal{T}_k} \|U_k\|_{\mathbb{V}(N_k(T))}^2 + \|\tilde{D}\|_{L^2(N_k(T))}^2 \lesssim \|U_k\|_{\mathbb{V}(\Omega)}^2 + \|\tilde{D}\|_{L^2(\Omega)}^2.$$

The uniform estimate $\|U_k\|_{\mathbb{V}(\Omega)} \leq \beta^{-1}\|f\|_{\mathbb{V}^*}$ implies the claim. □

We next investigate the maximal indicator in the set of marked elements. In addition to convergence of the discrete solutions and mesh-size functions we exploit stability of the indicators, and properties of REFINE.

Lemma 7.3 (Marking). *Suppose that the finite element spaces fulfill* (112) *and the estimator* (113b) *and* (113c).

Then the maximal indicator of the marked elements vanishes in the limit:

$$\lim_{k \to \infty} \max\{\mathcal{E}_k(U_k, T) \mid T \in \mathcal{M}_k\} = 0.$$

Proof. Let $T_k \in \mathcal{M}_k$ such that $\mathcal{E}_k(U_k, T_k) = \max\{\mathcal{E}_k(U_k, T) \mid T \in \mathcal{M}\}$. All elements in $\mathcal{M}_k$ are refined and therefore $T_k \in \mathcal{T}_k^0$. Local quasi-uniformity (55) of $\mathcal{T}_k$ implies

$$|N_k(T_k)| \lesssim |T_k| \leq \|h_k^d\|_{L^\infty(T_k)} \leq \|h_k^d\|_{L^\infty(\Omega_k^0)} \to 0 \tag{123}$$

as $k \to \infty$ by Corollary 7.1.

As shown above (113b) and (113c) imply the stability (122), whence we can proceed by the triangle inequality to estimate the maximal indicator by

$$\mathscr{E}_k(U_k,T_k) \lesssim \|U_k - u_\infty\|_{\mathbb{V}(\Omega)} + \|u_\infty\|_{\mathbb{V}(N_k(T_k))} + \|\tilde{D}\|_{L^2(N_k(T_k))}.$$

The first term on the right hand side converges to 0 as $k \to \infty$, thanks to Lemma 7.1 and the other terms vanish in the limit too, by continuity of $\|\cdot\|_{\mathbb{V}(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ with respect to the Lebesgue measure $|\cdot|$ and (123).

7.4.3 Convergence of the residuals

In this section we establish the weak convergence $\mathscr{R}_k \rightharpoonup 0$ in $\mathbb{W}^*$. In doing this, we distinguish two regions in Ω: in Ω_k^0 we use local density of the finite element spaces $\mathbb{W}_k$ in $\mathbb{W}$, and in Ω_k^+ we rely on properties of estimator and marking.

Proposition 7.1 (Weak Convergence of the Residuals). *Assume that* (112), (113), *and* (114) *are satisfied.*

Then the sequence of discrete solutions $\{U_k\}_{k\geq 0}$ *generated by iteration* (104) *verifies*

$$\lim_{k\to\infty} \langle \mathscr{R}_k, w\rangle = 0 \qquad \text{for all } w \in \mathbb{W}^s.$$

Proof. 1 For $k \geq \ell$ the inclusion $\mathscr{T}_\ell^+ \subset \mathscr{T}_k^+ \subset \mathscr{T}_k$ holds. Therefore, the sub-triangulation $\mathscr{T}_k \setminus \mathscr{T}_\ell^+$ of $\mathscr{T}_k$ covers the sub-domain Ω_ℓ^0, i. e., $\Omega_\ell^0 = \Omega(\mathscr{T}_k \setminus \mathscr{T}_\ell^+)$. We notice that any refinement of $\mathscr{T}_k$ does not affect any element in $\mathscr{T}_\ell^+$. Therefore, defining

$$\mathscr{T}_k^* = \{T' \mid T' \in N_k(T), T \in \mathscr{T}_k \setminus \mathscr{T}_\ell^+\}.$$

we also see that for $k \geq \ell$

$$\Omega_k^* = \Omega(\mathscr{T}_k^*) = \bigcup \{T' \colon T' \in N_\ell(T), T \in \mathscr{T}_\ell^0\}. \tag{124}$$

2 Let $w \in \mathbb{W}^s$ with $\|w\|_{\mathbb{W}^s(\Omega)} = 1$ be arbitrarily chosen. Since U_k is the Petrov-Galerkin solution we can employ Galerkin orthogonality (41) in combination with the upper bound (113a) to split for $k \geq \ell$

$$\begin{aligned}
|\langle \mathscr{R}_k, w\rangle| &= |\langle \mathscr{R}_k, w - I_k w\rangle| \\
&\lesssim \sum_{T\in\mathscr{T}_k\setminus\mathscr{T}_\ell^+} \mathscr{E}_k(U_k,T)\|w - I_k w\|_{\mathbb{V}(N_k(T))} + \sum_{T\in\mathscr{T}_\ell^+} \mathscr{E}_k(U_k,T)\|w - I_k w\|_{\mathbb{V}(N_k(T))} \\
&\lesssim \mathscr{E}_k(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+)\|w - I_k w\|_{\mathbb{V}(\Omega_k^*)} + \mathscr{E}_k(U_k,\mathscr{T}_\ell^+)\|w - I_k w\|_{\mathbb{V}(\Omega)},
\end{aligned}$$

by the Cauchy-Schwarz inequality and (121) for $\|\cdot\|_{\mathbb{V}}^2$. In view of Lemma 7.2 we bound $\mathscr{E}_k(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+) \leq \mathscr{E}_k(U_k,\mathscr{T}_k) \lesssim 1$. We next use (115) with $\mathscr{T}_k' = \mathscr{T}_k^*$ to obtain $\|w - I_k w\|_{\mathbb{W}(\Omega_k^*)} \lesssim \|h_k^s\|_{L^\infty(\Omega_k^*)}$, recalling $\|w\|_{\mathbb{W}^s(\Omega)} = 1$. From (124) we see that for any $T' \in \mathscr{T}_k^*$ we find $T \in \mathscr{T}_\ell^0$ with $T' \subset N_\ell(T)$. Local quasi-uniformity (55) of $\mathscr{T}_\ell$ and monotonicity of the mesh-size functions therefore imply

$$\|h_k\|_{L^\infty(\Omega_k^*)} \lesssim \|h_k\|_{L^\infty(\Omega_\ell^0)} \leq \|h_\ell\|_{L^\infty(\Omega_\ell^0)}.$$

In summary this yields

$$\|w - I_k w\|_{\mathbb{V}(\Omega_\ell^*)} \lesssim \|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} \qquad \text{and} \qquad \|w - I_k w\|_{\mathbb{V}(\Omega)} \lesssim 1,$$

which entails the existence of constants $0 \le C_1, C_2 < \infty$, such that

$$|\langle \mathscr{R}_k, w\rangle| \le C_1 \|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} + C_2 \mathscr{E}_k(U_k, \mathscr{T}_\ell^+) \qquad \text{for all } k \ge \ell. \tag{125}$$

[3] For any given $\varepsilon > 0$, convergence of the mesh-size function $\|h_\ell\|_{L^\infty(\Omega_\ell^0)} \to 0$ for $\ell \to \infty$, proven in Corollary 7.1, and $s > 0$ allows us to first choose $\ell \ge 0$ such that

$$\|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} \le \frac{\varepsilon}{2C_1}.$$

Employing the marking rule (114), we conclude

$$\lim_{k\to\infty} \max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{T}_k \setminus \mathscr{M}_k\} \le \lim_{k\to\infty} g\big(\max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{M}_k\}\big) = 0$$

by Lemma 7.3 and continuity of g in 0 with $g(0) = 0$. Since $\mathscr{T}_\ell^+ \cap \mathscr{M}_k = \emptyset$, this especially implies $\max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{T}_\ell^+\} \to 0$, whence we can next choose $K \ge \ell$ such that

$$\mathscr{E}_k(U_k,T) \le \frac{\varepsilon}{2C_2} (\#\mathscr{T}_\ell^+)^{-1/2} \qquad \text{for all } T \in \mathscr{T}_\ell^+ \text{ and all } k \ge K,$$

yielding $C_2 \mathscr{E}_k(U_k, \mathscr{T}_\ell^+) \le \varepsilon/2$ for those k. In summary, estimate (125) then implies $|\langle \mathscr{R}_k, w\rangle| \le \varepsilon$ for $k \ge K$. Since ε is arbitrary this finishes the proof. □

7.4.4 Proof of convergence

Collecting the auxillary results, we are in the position to prove the main result.

Proof of Theorem 7.1. [1] We first show convergence $U_k \to u$ in $\mathbb{V}$. For any $w \in \mathbb{W}^s$ we deduce

$$\begin{aligned}\langle \mathscr{R}_\infty, w\rangle &= \langle \mathscr{R}_\infty - \mathscr{R}_k, w\rangle + \langle \mathscr{R}_k, w\rangle = \mathscr{B}[u_\infty - U_k, w] + \langle \mathscr{R}_k, w\rangle \\ &\le \|\mathscr{B}\| \|u_\infty - U_k\|_{\mathbb{V}(\Omega)} \|w\|_{\mathbb{V}(\Omega)} + \langle \mathscr{R}_k, w\rangle \to 0\end{aligned}$$

as $k \to \infty$ by Lemma 7.1 and Proposition 7.1, whence $\langle \mathscr{R}_\infty, w\rangle = 0$ for all $w \in \mathbb{W}^s$. This implies $\mathscr{R}_\infty = 0$ in $\mathbb{W}^*$ since $\mathbb{W}^s$ is dense in $\mathbb{W}$. The continuous inf-sup condition (21) yields

$$\alpha \|u_\infty - u\|_{\mathbb{V}} \le \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u_\infty - u, w] = \sup_{\|w\|_{\mathbb{W}}=1} \langle \mathscr{R}_\infty, w\rangle = 0,$$

which shows $u = u_\infty$. Convergence of the Galerkin approximations finally implies

$$\lim_{k\to\infty} U_k = u_\infty = u \qquad \text{in } \mathbb{V}.$$

2 After proving $U_k \to u$ we next turn to the convergence of the estimators. Just like in the proof of Proposition 7.1 we split for $k \geq \ell$

$$\mathscr{E}_k^2(U_k, \mathscr{T}_k) = \mathscr{E}_k^2(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+) + \mathscr{E}_k^2(U_k, \mathscr{T}_\ell^+)$$

and we estimate the first term with the help of the local lower bound (113b), (113c) by

$$\begin{aligned}\mathscr{E}_k^2(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+) &\lesssim \sum_{T\in\mathscr{T}_k\setminus\mathscr{T}_\ell^+} \|U_k - u\|^2_{\mathbb{V}(N_k(T))} + h_T^{2q}\Big(\|U_k\|^2_{\mathbb{V}(N_k(T))} + \|D\|^2_{L^2(N_k(T))}\Big)\\ &\lesssim \|U_k - u\|^2_{\mathbb{V}} + \big\|h_k^{2q}\big\|_{L^\infty(\Omega_\ell^0)}\big(\beta^{-2}\|f\|^2_{\mathbb{V}^*} + \|D\|^2_{L^2(\Omega)}\big),\end{aligned}$$

where we have used (121) for $\|\cdot\|_{\mathbb{V}(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ as well as $\|U_k\|_{\mathbb{V}} \leq \beta^{-1}\|f\|_{\mathbb{W}^*}$ in the second step. Using once again monotonicity of the mesh-size functions we deduce for some constants C_1, C_2

$$\mathscr{E}_k^2(U_k, \mathscr{T}_k) \leq C_1\big\|h_\ell^{2q}\big\|_{L^\infty(\Omega_\ell^0)} + C_2\|U_k - u\|^2_{\mathbb{V}(\Omega)} + \mathscr{E}_k^2(U_k, \mathscr{T}_\ell^+).$$

By Corollary 7.1 we can make the first term small by choosing ℓ sufficiently large. In the proof of Proposition 7.1 we already have shown $\mathscr{E}_k(U_k, \mathscr{T}_\ell^+) \to 0$ for fixed ℓ and $k \to \infty$. Step 1 implies $\|U_k - u\|_{\mathbb{V}(\Omega)} \to 0$ as $k \to \infty$ which allows to make the last two terms small by choosing k large after fixing ℓ. This proves $\mathscr{E}_k(U_k, \mathscr{T}_k) \to 0$ as $k \to \infty$ and finishes the proof. □

Remark 7.2 (Lower Bound). For convergence $U_k \to u$ we have only utilized the stability (122) of the indicators, which is much weaker than efficiency (113b) because it allows for overestimation. Since most of the estimators for linear problems are shown to be reliable and efficient, we directly asked for efficiency of the estimator. For nonlinear problems this might be different and just asking for (122) may provide access for proving convergence for a larger problem class.

All convergence results but [21, 67] rely on a *discrete* local lower bound. For the model problem there is no difference in deriving the continuous or the discrete lower bound; compare with Sect. 6.3. In general, the derivation of a discrete lower bound is much more involved than its continuous counterpart. For instance, in Problem 7.2 below the discrete lower bound is not known and in Problem 7.3 it is only known for the lowest order elements. In respect thereof a convergence proof without lower bound enlarges the problem class where it applies to.

Yet, only asking for (122) yields convergence $U_k \to u$ but the progress without convergence $\mathscr{E}_k(U_k, \mathscr{T}_k) \to 0$ is not observable in the adaptive iteration. Therefore, a convergence result for non-efficient estimators is of little practical use.

Remark 7.3 (Characterization of Convergent Marking). The results in [55] and [67] also give a characterization of convergent marking. In our setting

$$\lim_{k\to\infty} \max\{\mathscr{E}_k(U_k,T) \mid T\in\mathscr{M}_k\} = 0 \quad \Longrightarrow \quad \lim_{k\to\infty} \mathscr{E}_k(U_k,T) = 0 \quad \text{for all } T\in\mathscr{T}^+ \tag{126}$$

is necessary and sufficient for convergence of (104). To see this, the hypothesis of (126) we have shown in Lemma 7.3 and the conclusion of (126) is obviously necessary for $\mathscr{E}_k(U_k,\mathscr{T}_k)\to 0$. If $\lim_{k\to\infty}\operatorname{osc}_k(U_k,T)=0$ for all $T\in\mathscr{T}^+$ then it is also necessary for $\|U_k-u\|_{\mathbb{V}}\to 0$ by the lower bound (113b), for instance in the model problem when $\boldsymbol{A}$ and f are piecewise constant over $\mathscr{T}_0$. Condition (114) on marking we only have used in Step 3 of the proof to Proposition 7.1 and there it can be replaced by (126), whence (126) is also sufficient.

On the one hand, this assumption is not 'a posteriori' in that it can not be checked at iteration k of the adaptive loop and thus seems of little practical use. On the other hand, being a characterization of convergent marking it may be used to treat marking strategies that are based on extrapolation techniques involving indicators from previous iterations [5], or that are based on some optimization procedure [41].

Similarly, the condition on marking can be generalized to marking procedures where a given tolerance of the adaptive method enters the selection of elements, for instance the original equidistribution strategy for parabolic problems in [34]. Such methods then in turn only aim at convergence into tolerance. For details we refer to [67, Sect. 5].

7.5 Problems

Problem 7.1. Consider the general 2^{nd} order elliptic problem from Sect. 2.2.2, where $\boldsymbol{A}$ piecewise Lipschitz over $\mathscr{T}_0$ with smallest eigenvalue strictly bounded away from 0 and $c-1/2\operatorname{div}\boldsymbol{b}\geq 0$. Therefore, the corresponding bilinear form $\mathscr{B}$ is coercive on $\mathbb{V}=H_0^1(\Omega)$; compare with Sect.2.5.2.

Show that a discretization with H_0^1 conforming Lagrange elements of order $n\geq 1$ introduced in Sect. 3.2.2 and the residual estimator from Sect. 6.2 satisfy the assumptions (112) and (113). This implies convergence of the adaptive iteration (104) for the general 2^{nd} order elliptic equation with any of the marking strategies from Sect. 7.3.4.

Problem 7.2. Consider the biharmonic equation in 2d from Sect. 2.2.2 which leads to a variational problem in $\mathbb{V}=H_0^2(\Omega)$ with a continuous and coercive bilinear form.

Show that the discretization with the Argyris triangle defined in [25, Theorems 2.2.11 and 2.2.13] of Ciarlet's book satisfies (112). In addition verify that the estimator derived by Verfürth in [76, Section 3.7] fulfills (113). This implies convergence of the adaptive iteration (104) for the biharmonic equation with any of the marking strategies from Sect. 7.3.4.

Problem 7.3. Consider the 3d Eddy Current Equations from Sect. 2.2.2 which leads to a variational problem in $\mathbb{V}=H_0(\operatorname{curl};\Omega)$ with a continuous and coercive bilinear form.

Show that the discretization with Nédélec finite elements of order $n \in \mathbb{N}$ comply with (112); compare with [51, Sect. 5.5]. Consider the estimator derived by Schöberl [64, Corollary 2] that has been shown to be efficient by Beck et al. [12, Theorem 3.3]. Show that it fulfills (113). This implies convergence of the adaptive iteration (104) for the 3d Eddy Current Equations with any of the marking strategies from Sect. 7.3.4.

Problem 7.4. Consider the Stokes problem from Sect. 2.2.2 that leads to a variational problem in $\mathbb{V} = H^1_0(\Omega;\mathbb{R}^d) \times L^2_0(\Omega)$ with a non-coercive bilinear form $\mathscr{B}$ that satisfies the inf-sup condition (23).

For the discretization with the Taylor-Hood element of order $n \geq 2$, this means we approximate the velocity with continuous piecewise polynomials of degree n and the pressure with continuous piecewise polynomials of degree $n-1$, Otto has shown (112c) in [59]. Prove that the Taylor-Hood element satisfies the other requirements of (112). Finally show that the estimator by Verfürth for the Stokes system [75] complies with (113). This implies convergence of the adaptive iteration (104) for the Stokes problem with any of the marking strategies from Sect. 7.3.4.

8 Adaptivity: Contraction property

This chapter discusses the contraction property of AFEM for the *model problem* of Sect. 2.2.1, namely

$$-\operatorname{div}(\boldsymbol{A}(x)\nabla u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{127}$$

The variational formulation of (127) from Sect. 2.5.1 reads with $\mathbb{V} = \mathbb{W} = H^1(\Omega)$

$$u \in \mathbb{V}: \quad \mathscr{B}[u,v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla u = \int_\Omega f\, v =: \langle f, v\rangle \qquad \text{for all } v \in \mathbb{V}.$$

We revisit the modules of the basic adaptive loop (4), i. e.,

$$\mathsf{SOLVE} \longrightarrow \mathsf{ESTIMATE} \longrightarrow \mathsf{MARK} \longrightarrow \mathsf{REFINE}.$$

Similar to Chap. 7, the outcome of each iteration with counter $k \geq 1$ is a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ of conforming bisection refinements $\mathscr{T}_k$ of $\mathscr{T}_0$, spaces of conforming finite element spaces $\mathbb{V}_k = \mathbb{W}_k = S^{n,0}(\mathscr{T}_k) \cap H^1_0(\Omega)$, i. e., C^0 continuous piecewise polynomials of degree $\leq n$ for both ansatz and test spaces, and Ritz-Galerkin solutions $U_k \in \mathbb{V}_k$.

Since error monotonicity is closely related to a minimization principle, we cannot in general expect a contraction property for problems governed by an inf-sup condition. We thus restrict ourselves to the special class of coercive and symmetric problems of the form (127). The first contribution in dimension $d > 1$ is due to Dörfler [32], who introduced a crucial marking, the so-called Dörfler marking of Sect. 7.1, and proved strict energy error reduction for the Laplacian provided

the initial mesh $\mathscr{T}_0$ satisfies a fineness assumption. The Dörfler marking will play an essential role in the present discussion, which does not seem to extend to other marking strategies such as those in Sect. 7.1. Morin, Nochetto, and Siebert [52, 53] showed that such strict energy error reduction does not hold in general even for (127). By introducing the concept of data oscillation and the interior node property, they proved convergence of the AFEM without restrictions on $\mathscr{T}_0$. The latter result, however, is valid only for $\boldsymbol{A}$ in (127) piecewise constant on $\mathscr{T}_0$. Inspired by the work of Chen and Feng [24], Mekchay and Nochetto [48] proved a contraction property for the total error, namely the sum of the energy error plus oscillation, for general second order elliptic operators such as those in Sect. 2.5.2. For non-symmetric $\mathscr{B}$ this requires a sufficient fineness of the initial grid $\mathscr{T}_0$. The total error will reappear in the study of convergence rates in Chap. 9.

Diening and Kreuzer proved a similar contraction property for the p-Laplacian replacing the energy norm by the so-called quasi-norm [31]. They were able to avoid marking for oscillation by using the fact that oscillation is dominated by the estimator. Most results for nonlinear problems utilize the equivalence of the energy error and error in the associated (nonlinear) energy; compare with Problem 8.3. This equivalence was first used by Veeser in a convergence analysis for the p-Laplacian [73] and later on by Siebert and Veeser for the obstacle problem [68].

The result of Diening and Kreuzer inspired the work by Cascón et al. [21], who proved a contraction property for the *quasi-error*:

$$|||u-U_k|||_\Omega^2+\gamma\mathscr{E}_k^2(U_k,\mathscr{T}_k),$$

where $\gamma>0$ is a suitable scaling constant. This approach hinges solely on a strict reduction of the mesh-size within refined elements, the upper a posteriori error bound, an orthogonality property natural for (127) in nested approximation spaces, and Dörfler marking. This appears to be the simplest approach currently available and is presented next.

8.1 The modules of AFEM for the model problem

We assume Ω is triangulated by some initial grid $\mathscr{T}_0$. We suppose that $\boldsymbol{A}$ is uniformly SPD so that (127) is *coercive* and in addition we ask $\boldsymbol{A}$ to be piecewise Lipschitz over $\mathscr{T}_0$. We next describe the modules of the adaptive algorithm.

Module SOLVE. For any $\mathscr{T}\in\mathbb{T}$ we set $\mathbb{V}(\mathscr{T})=S^{n,0}(\mathscr{T})\cap H_0^1(\Omega)$ and suppose that

$$U_{\mathscr{T}}=\mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T})\big)$$

outputs the *exact* Ritz-Galerkin approximation in $\mathbb{V}(\mathscr{T})$, namely,

$$U_{\mathscr{T}}\in\mathbb{V}(\mathscr{T}):\qquad \mathscr{B}[U_{\mathscr{T}},V]=\langle f,V\rangle\qquad\text{for all } V\in\mathbb{V}(\mathscr{T}).$$

This entails exact integration and linear algebra; see Remarks 3.6 and 3.7.

Module ESTIMATE. Given a grid $\mathscr{T} \in \mathbb{T}$ and the Ritz-Galerkin approximation $U_{\mathscr{T}} \in \mathbb{V}(\mathscr{T})$ the output

$$\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}} = \mathsf{ESTIMATE}\big(U_{\mathscr{T}},\mathscr{T}\big)$$

are the indicators of the residual estimator derived in Chap. 6. We recall that for a generic function $V \in \mathbb{V}(\mathscr{T})$ the *element* and *jump residuals* are defined by

$$\begin{aligned} r(V)|_T &= f + \operatorname{div}(\boldsymbol{A}\nabla V) = f && \text{for all } T \in \mathscr{T}, \\ j(V)|_S &= [\![\boldsymbol{A}\nabla V]\!]_S && \text{for all } S \in \mathring{\mathscr{S}} \end{aligned}$$

and the element indicator evaluated in V is then

$$\mathscr{E}^2_{\mathscr{T}}(V,T) = h_T^2 \, \|r(V)\|^2_{L^2(T)} + h_T \, \|j(V)\|^2_{L^2(\partial T\cap\Omega)} \quad \text{for all } T \in \mathscr{T}.$$

Module MARK. For any $\mathscr{T} \in \mathbb{T}$ and indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}$ the module MARK selects elements for refinement using *Dörfler Marking*, i. e., using a fixed parameter $\theta \in (0,1]$ the output

$$\mathscr{M} = \mathsf{MARK}\big(\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}},\mathscr{T}\big)$$

satisfies

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{M}) \geq \theta\, \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{T}).$$

Dörfler Marking guarantees that the total estimator is controlled up the constant θ^{-1} by the estimator on the marked elements. This is a crucial property in our arguments. The choice of $\mathscr{M}$ does not have to be minimal at this stage, that is, the marked elements $T \in \mathscr{M}$ do not necessarily must be those with largest indicators. However, minimality of $\mathscr{M}$ will be crucial to derive rates of convergence in Chap. 9.

Module REFINE. We fix the number $b \in \mathbb{N}$ of bisections and consider the module REFINE from Sect. 4.4 to refine all marked elements b times. Then for any $\mathscr{T} \in \mathbb{T}$ the output

$$\mathscr{T}_* = \mathsf{REFINE}\big(\mathscr{T},\mathscr{M}\big)$$

satisfies $\mathscr{T}_* \in \mathbb{T}$. Furthermore, if $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ is the set of refined elements of $\mathscr{T}$, then $\mathscr{M} \subset \mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ and $h_{\mathscr{T}_*} \leq 2^{-b/d} h_{\mathscr{T}}$ inside all elements of $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$.

8.2 Properties of the modules of AFEM

We next summarize some basic properties of the adaptive algorithm that emanate from the symmetry of the differential operator and features of the modules. In doing this, any explicit constant or hidden constant in $\lesssim$ must, apart from explicitly

stated other dependencies, only depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, and the (global) eigenvalues of $\boldsymbol{A}$, but not on a specific grid $\mathscr{T} \in \mathbb{T}$. Further on, u will always be the weak solution of (127).

Lemma 8.1 (Nesting of Spaces). *Any sequence* $\{\mathbb{V}_k = \mathbb{V}(\mathscr{T}_k)\}_{k\geq 0}$ *of discrete spaces generated by the basic adaptive loop* (4) *is nested, this is,*

$$\mathbb{V}_k \subset \mathbb{V}_{k+1} \qquad \text{for all } k \geq 0.$$

Proof. See Problem 8.1. □

The following property relies on the fact that $\mathscr{B}$ is coercive and symmetric, and so induces a scalar product in $\mathbb{V}$ equivalent to the H^1_0-scalar product.

Lemma 8.2 (Pythagoras). *Let* $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ *such that* $\mathscr{T} \leq \mathscr{T}_*$. *The respective Ritz-Galerkin solutions* $U \in \mathbb{V}(\mathscr{T})$ *and* $U_* \in \mathbb{V}(\mathscr{T}_*)$ *satisfy the following orthogonality property in the energy norm* $|||\cdot|||_\Omega$

$$|||u-U|||^2_\Omega = |||u-U_*|||^2_\Omega + |||U_*-U|||^2_\Omega\,. \tag{128}$$

Proof. See Problem 8.2. □

A by-product of (128) is the monotonicity property

$$|||U_*-U|||_\Omega \leq |||u-U|||_\Omega\,. \tag{129}$$

A perturbation of (128) is still valid for the general 2nd order elliptic operators of Sect. 2.5.2, as shown in [48], but not for non-coercive problems. Even for (127), property (128) is valid exclusively for the energy norm. This restricts the subsequent analysis to the energy norm, or equivalent norms, but does not extend to other, perhaps more practical, norms such as the maximum norm. This is an open problem.

We now continue the discussion of oscillation of Sect. 6.3.3. In view of (96), we denote by $\operatorname{osc}_{\mathscr{T}}(V,T)$ the *element oscillation* for any $V \in \mathbb{V}$

$$\operatorname{osc}_{\mathscr{T}}(V,T) = \|h(r(V)-\overline{r(V)})\|_{L^2(T)} + \|h^{1/2}(j(V)-\overline{j(V)})\|_{L^2(\partial T\cap\Omega)},$$

where $\overline{r(V)} = P_{2n-2}r(V)$ and $\overline{j(V)} = P_{2n-1}j(V)$ stand for L^2-projections of the residuals $r(V)$ and $j(V)$ onto the polynomials $\mathbb{P}_{2n-2}(T)$ and $\mathbb{P}_{2n-1}(S)$ defined on the element T or side $S \subset \partial T$, respectively. For variable $\boldsymbol{A}$, $\operatorname{osc}_{\mathscr{T}}(V,T)$ depends on the discrete function $V \in \mathbb{V}$, and its study is more involved than for piecewise constant $\boldsymbol{A}$. In the latter case, $\operatorname{osc}_{\mathscr{T}}(V,T)$ becomes *data oscillation* $\operatorname{osc}_{\mathscr{T}}(V,T) = \|h(f-\bar{f})\|_{L^2(T)}$; compare with Remark 6.3.

We now rewrite the a posteriori error estimates of Theorem 6.3 in the energy norm.

Lemma 8.3 (A Posteriori Error Estimates). *There exist constants* $0 < C_2 \leq C_1$, *such that for any* $\mathscr{T} \in \mathbb{T}$ *and the corresponding Ritz-Galerkin solution* $U \in \mathbb{V}(\mathscr{T})$ *there holds*

$$|||u-U|||_\Omega^2 \le C_1\,\mathscr{E}_\mathscr{T}^2(U,\mathscr{T}) \tag{130a}$$

$$C_2\,\mathscr{E}_\mathscr{T}^2(U,\mathscr{T}) \le |||u-U|||_\Omega^2 + \operatorname{osc}_\mathscr{T}^2(U,\mathscr{T}). \tag{130b}$$

The constants C_1 and C_2 depend on the smallest and largest global eigenvalues of $\boldsymbol{A}$. This dependence can be improved if the a posteriori analysis is carried out directly in the energy norm instead of the H_0^1-norm; see Problem 6.14. The definitions of $\overline{r(V)}$ and $\overline{j(V)}$, as well as the lower bound (130b), are immaterial for deriving a contraction property. However, they will be important for proving convergence rates in Chap. 9.

Lemma 8.4 (Lipschitz Property). *For any $\mathscr{T}\in\mathbb{T}$ and $T\in\mathscr{T}$, there holds*

$$|\mathscr{E}_\mathscr{T}(V,T)-\mathscr{E}_\mathscr{T}(W,T)| \lesssim \eta_\mathscr{T}(\boldsymbol{A},T)\|\nabla(V-W)\|_{L^2(\omega_T)} \quad \text{for all } V,W\in\mathbb{V}(\mathscr{T}).$$

By ω_T we again denote the union of elements sharing a side with T, $\operatorname{div}\boldsymbol{A}\in\mathbb{R}^d$ is the divergence of $\boldsymbol{A}$ computed by rows, and

$$\eta_\mathscr{T}(\boldsymbol{A},T) := h_T\|\operatorname{div}\boldsymbol{A}\|_{L^\infty(T)} + \|\boldsymbol{A}\|_{L^\infty(\omega_T)}.$$

Proof. Recalling the definition of the indicators, the triangle inequality yields

$$|\mathscr{E}_\mathscr{T}(V,T)-\mathscr{E}_\mathscr{T}(W,T)| \le h_T\|r(V)-r(W)\|_{L^2(T)} + h_T^{1/2}\|j(V)-j(W)\|_{L^2(\partial T)}.$$

We set $E:=V-W\in\mathbb{V}(\mathscr{T})$, and observe that

$$r(V)-r(W) = \operatorname{div}(\boldsymbol{A}\nabla E) = \operatorname{div}\boldsymbol{A}\cdot\nabla E + \boldsymbol{A}:D^2E,$$

where D^2E is the Hessian of E. Since E is a polynomial of degree $\le n$ in T, applying the inverse estimate $\|D^2E\|_{L^2(T)}\lesssim h_T^{-1}\|\nabla E\|_{L^2(T)}$, we deduce

$$h_T\|r(V)-r(W)\|_{L^2(T)} \lesssim \eta_\mathscr{T}(\boldsymbol{A},T)\|\nabla E\|_{L^2(T)}.$$

On the other hand, for any $S\subset\partial T$ applying the inverse estimate of Problem 8.4 gives

$$\|j(V)-j(W)\|_{L^2(S)} = \|j(E)\|_{L^2(S)} = \|\,[\![A\nabla E]\!]\,\|_{L^2(S)} \lesssim h_T^{-1/2}\|\nabla E\|_{L^2(\omega_T)}$$

where the hidden constant is proportional to $\eta_\mathscr{T}(\boldsymbol{A},T)$. This finishes the proof. □

One serious difficulty in dealing with AFEM is that one has access to the energy error $|||u-U|||_\Omega$ only through the estimator $\mathscr{E}_\mathscr{T}(U,\mathscr{T})$. The latter, however, fails to exhibit a monotonicity property such as (129) because it depends on the discrete solution $U\in\mathbb{V}(\mathscr{T})$ that changes with the mesh. We account for this change in the next lemma, which is a direct consequence of Lemma 8.4.

Lemma 8.5 (Estimator Reduction). *Let $\mathscr{T}\in\mathbb{T}$ be given with a subset $\mathscr{M}\subset\mathscr{T}$ of marked elements and let $\mathscr{T}_* = \mathsf{REFINE}(\mathscr{T},\mathscr{M})$.*

There exists a constant $\Lambda > 0$, such that all $V \in \mathbb{V}(\mathscr{T})$, $V_* \in \mathbb{V}_*(\mathscr{T}_*)$ and any $\delta > 0$ we have

$$\mathscr{E}^2_{\mathscr{T}_*}(V_*,\mathscr{T}_*) \le (1+\delta)\big(\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}) - \lambda\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M})\big) + (1+\delta^{-1})\,\Lambda\,\eta^2_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})\,|||V_*-V|||^2_\Omega\,,$$

where $\lambda = 1-2^{-b/d}$ and

$$\eta_{\mathscr{T}}(\boldsymbol{A},\mathscr{T}) := \max_{T\in\mathscr{T}} \eta_{\mathscr{T}}(\boldsymbol{A},T).$$

Proof. We proceed in several steps.

1 *Global Estimate.* We first observe that $V \in \mathbb{V}(\mathscr{T}_*)$ since the spaces are nested. We next invoke Lemma 8.4 for $T \in \mathscr{T}_*$ and $V, V_* \in \mathbb{V}(\mathscr{T}_*)$ to get

$$\mathscr{E}_{\mathscr{T}_*}(V_*,T) \le \mathscr{E}_{\mathscr{T}_*}(V,T) + C\,\eta_{\mathscr{T}_*}(\boldsymbol{A},T)\|V_*-V\|_{H^1(\omega_T)}.$$

Given $\delta > 0$, we apply Young's inequality $(a+b)^2 \le (1+\delta)a^2 + (1+\delta^{-1})b^2$ and add over $T \in \mathscr{T}_*$ to arrive at

$$\mathscr{E}^2_{\mathscr{T}_*}(V_*,\mathscr{T}_*) \le (1+\delta)\,\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*) + \Lambda\,(1+\delta^{-1})\eta^2_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})\,|||V_*-V|||^2_\Omega\,. \quad (131)$$

Here, $\Lambda = (d+1)C/\alpha_1$ results from the finite overlapping property of sets ω_T and the relation between norms

$$\alpha_1\,\|\nabla v\|^2_{L^2(\Omega)} \le |||v|||^2_\Omega \quad \text{for all } v \in \mathbb{V}.$$

In addition we have used the monotonicity property $\eta_{\mathscr{T}_*}(\boldsymbol{A},\mathscr{T}_*) \le \eta_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})$.

2 *Accounting for $\mathscr{M}$.* We next decompose $\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*)$ over elements $T \in \mathscr{T}$, and distinguish whether or not $T \in \mathscr{M}$. If $T \in \mathscr{M}$, then T is bisected at least b times and so T can be written as the union of elements $T' \in \mathscr{T}_*$ We denote this set of elements $\mathscr{T}_*(T)$ and observe $h_{T'} \le 2^{-b/d}\,h_T$ for all $T' \in \mathscr{T}_*(T)$. Therefore

$$\sum_{T'\in\mathscr{T}_*(T)} h^2_{T'}\|r(V)\|^2_{L^2(T')} \le 2^{-(2b)/d}\,h^2_T\|r(V)\|^2_{L^2(T)}$$

and

$$\sum_{T'\in\mathscr{T}_*(T)} h_{T'}\,\|j(V)\|^2_{L^2(\partial T'\cap\Omega)} \le 2^{-b/d}\,h_T\,\|j(V)\|^2_{L^2(\partial T\cap\Omega)}.$$

This implies

$$\mathscr{E}^2_{\mathscr{T}_*}(V,T) \le 2^{-b/d}\,\mathscr{E}^2_{\mathscr{T}}(V,T) \qquad \text{for all } T \in \mathscr{M}.$$

For the remaining elements $T \in \mathscr{T}\setminus\mathscr{M}$ we only know that mesh-size does not increased because $\mathscr{T} \le \mathscr{T}_*$, whence

$$\mathscr{E}^2_{\mathscr{T}_*}(V,T) \le \mathscr{E}^2_{\mathscr{T}}(V,T) \qquad \text{for all } T \in \mathscr{T}\setminus\mathscr{M}.$$

3 *Assembling.* Combining the two estimates we see that

$$\begin{aligned}\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*) &\le 2^{-b/d}\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M}) + \mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}\setminus\mathscr{M}) \\ &= \mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}) - \big(1-2^{-b/d}\big)\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M}).\end{aligned}$$

Recalling the definition of $\lambda = 1-2^{-b/d}$ and replacing $\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*)$ in (131) by the right hand side of this estimate yields the assertion. □

8.3 Contraction property of AFEM

Recall that AFEM stands for the iteration loop (104) for the model problem. A key question to ask is what is (are) the quantity(ies) that AFEM may contract. In view of (129), an obvious candidate is the energy error $|||u-U_k|||_\Omega$. We show next that this may not be the case unless REFINE enforces several levels of refinement.

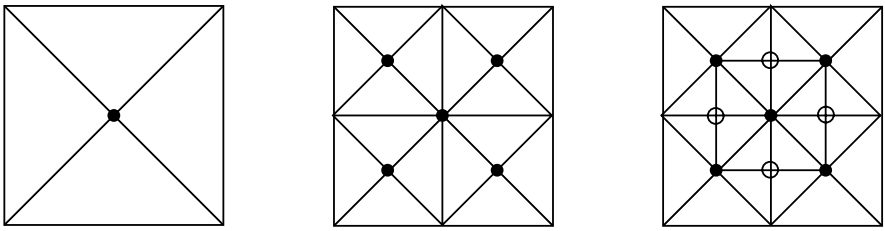


Fig. 15 Grids $\mathscr{T}_0$, $\mathscr{T}_1$, and $\mathscr{T}_2$ of the interior node example

Example 8.1 (Interior Node). Let $\Omega = (0,1)^2$, $\boldsymbol{A} = I$, $f = 1$, and consider the sequence of meshes depicted in Fig. 15. If ϕ_0 denotes the basis function associated with the only interior node of $\mathscr{T}_0$, then

$$U_0 = U_1 = \frac{1}{12}\,\phi_0, \qquad U_2 \neq U_1.$$

The mesh $\mathscr{T}_1 \ge \mathscr{T}_0$ is produced by a standard 2-step bisection ($b=2$) in $2d$. Since $U_0 = U_1$ we conclude that the energy error does not change

$$|||u-U_0|||_\Omega = |||u-U_1|||_\Omega$$

between consecutive steps of AFEM. This is no longer the case provided an interior node in each marked element is created, because then $U_2 \neq U_1$ and so $|||u-U_2|||_\Omega < |||u-U_1|||_\Omega$ (see (128)).

This example appeared first in [52, 53], and was used to justify the *interior node property: $\mathscr{T}_*$ must have one node in each side and interior of every $T \in \mathscr{M}$*. This property entails a minimal number of bisections that increases with the dimension

d. The following heuristics explains why this property, closely related to a local discrete lower bound (see Problem (6.7)), is no longer needed in the present approach.

Heuristics. According to (128), the energy error is monotone, but the previous example shows that strict inequality may fail. However, in case $U_{k+1} = U_k$, the estimator reduction in Lemma 8.5 for $V_* = U_{k+1}$ and $V = U_k$ reveals a strict estimator reduction. We could thus expect that a suitable combination of them, the so-called *quasi error*

$$|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2(U_k, \mathscr{T}_k),$$

may be contractive. This heuristics illustrates a distinct aspect of AFEM theory, the interplay between continuous quantities such the energy error $|||u - U_k|||_{\Omega}$ and discrete ones such as the estimator $\mathscr{E}_k(U_k, \mathscr{T}_k)$: no one alone has the requisite properties to yield a contraction between consecutive adaptive steps.

Theorem 8.1 (Contraction Property). *Let $\theta \in (0,1]$ be the Dörfler Marking parameter, and $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ be a sequence of conforming meshes, finite element spaces and discrete solutions created by AFEM for the model problem (127).*

Then there exist constants $\gamma > 0$ and $0 < \alpha < 1$, additionally depending on the number b of bisections and θ, such that for all $k \geq 0$

$$|||u - U_{k+1}|||_{\Omega}^2 + \gamma \mathscr{E}_{k+1}^2(U_{k+1}, \mathscr{T}_{k+1}) \leq \alpha^2 \Big(|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2(U_k, \mathscr{T}_k) \Big).$$

Proof. We split the proof into four steps. For convenience, we use the notation

$$e_k = |||u - U_k|||_{\Omega}, \; E_k = |||U_{k+1} - U_k|||_{\Omega}, \; \mathscr{E}_k = \mathscr{E}_k(U_k, \mathscr{T}_k), \; \mathscr{E}_k(\mathscr{M}_k) = \mathscr{E}_k(U_k, \mathscr{M}_k).$$

1 The error orthogonality (128) reads

$$e_{k+1}^2 = e_k^2 - E_k^2. \tag{132}$$

Employing Lemma 8.5 with $\mathscr{T} = \mathscr{T}_k$, $\mathscr{T}_* = \mathscr{T}_{k+1}$, $V = U_k$ and $V_* = U_{k+1}$ gives

$$\mathscr{E}_{k+1}^2 \leq (1+\delta)\big(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\big) + (1+\delta^{-1})\,\Lambda_0\, E_k^2, \tag{133}$$

where $\Lambda_0 = \Lambda\, \eta_{\mathscr{T}_0}^2(\boldsymbol{A}, \mathscr{T}_0) \geq \Lambda\, \eta_{\mathscr{T}_k}^2(\boldsymbol{A}, \mathscr{T}_k)$. After multiplying (133) by $\gamma > 0$, to be determined later, we add (132) and (133) to obtain

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \big(\gamma(1+\delta^{-1})\Lambda_0 - 1\big) E_k^2 + \gamma(1+\delta)\big(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\big).$$

2 We now choose the parameters δ, γ, the former so that

$$(1+\delta)\big(1 - \lambda\theta^2\big) = 1 - \frac{\lambda\theta^2}{2},$$

and the latter to verify

$$\gamma(1+\delta^{-1})\Lambda_0 = 1.$$

Note that this choice of γ yields

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \gamma(1+\delta)\left(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\right).$$

3 We next employ Dörfler Marking, namely $\mathscr{E}_k(\mathscr{M}_k) \geq \theta \mathscr{E}_k$, to deduce

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \gamma(1+\delta)(1-\lambda\theta^2)\mathscr{E}_k^2$$

which, in conjunction with the choice of δ, gives

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \gamma\left(1 - \frac{\lambda\theta^2}{2}\right)\mathscr{E}_k^2 = e_k^2 - \frac{\gamma\lambda\theta^2}{4}\mathscr{E}_k^2 + \gamma\left(1 - \frac{\lambda\theta^2}{4}\right)\mathscr{E}_k^2.$$

4 Finally, the upper bound (130a), namely $e_k^2 \leq C_1\, \mathscr{E}_k^2$, implies that

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq \left(1 - \frac{\gamma\lambda\theta^2}{4C_1}\right)e_k^2 + \gamma\left(1 - \frac{\lambda\theta^2}{4}\right)\mathscr{E}_k^2.$$

This in turn leads to

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq \alpha^2\left(e_k^2 + \gamma \mathscr{E}_k^2\right),$$

with

$$\alpha^2 := \max\left\{1 - \frac{\gamma\lambda\theta^2}{4C_1}, 1 - \frac{\lambda\theta^2}{4}\right\},$$

and proves the theorem because $\alpha^2 < 1$. □

Remark 8.1 (Ingredients). This proof hinges on the following basic ingredients: Dörfler marking; symmetry of $\mathscr{B}$ and nesting of spaces, which imply the Pythagoras identity (Lemma 8.2); the a posteriori upper bound (Lemma 8.3); and the estimator reduction property (Lemma 8.5). It does not use the lower bound (130b) and does not require marking by oscillation, as previous proofs do [24, 48, 52, 53, 54]. The marking is driven by $\mathscr{E}_k$ exclusively, as it happens in all practical AFEM.

8.4 Example: Discontinuous coefficients

We invoke the formulas derived by Kellogg [43] to construct an exact solution of an elliptic problem with piecewise constant coefficients and vanishing right-hand side f. We now write these formulas in the particular case $\Omega = (-1,1)^2$, $\boldsymbol{A} = a_1 \boldsymbol{I}$ in the first and third quadrants, and $\boldsymbol{A} = a_2 \boldsymbol{I}$ in the second and fourth quadrants. An exact weak solution u for $f \equiv 0$ is given in polar coordinates by $u(r,\theta) = r^\gamma \mu(\theta)$, where

$$\mu(\theta) = \begin{cases} \cos((\pi/2-\sigma)\gamma)\cdot\cos((\theta-\pi/2+\rho)\gamma) & \text{if } 0 \leq \theta \leq \pi/2, \\ \cos(\rho\gamma)\cdot\cos((\theta-\pi+\sigma)\gamma) & \text{if } \pi/2 \leq \theta \leq \pi, \\ \cos(\sigma\gamma)\cdot\cos((\theta-\pi-\rho)\gamma) & \text{if } \pi \leq \theta < 3\pi/2, \\ \cos((\pi/2-\rho)\gamma)\cdot\cos((\theta-3\pi/2-\sigma)\gamma) & \text{if } 3\pi/2 \leq \theta \leq 2\pi, \end{cases}$$

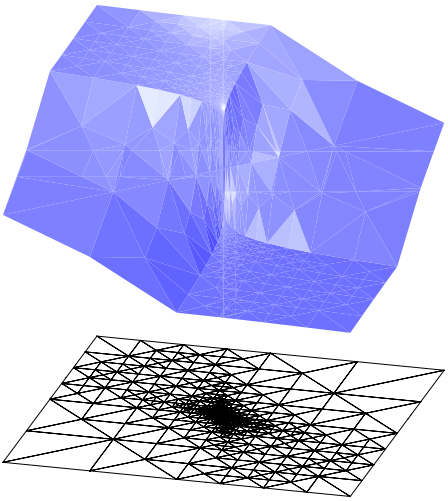


Fig. 16 Discontinuous coefficients in checkerboard pattern: Graph of the discrete solution, which is $u \approx r^{0.1}$, and underlying strongly graded grid. Notice the steep gradient of u at the origin

and the numbers γ, ρ, σ satisfy the nonlinear relations

$$\begin{cases} R := a_1/a_2 = -\tan((\pi/2-\sigma)\gamma)\cdot\cot(\rho\gamma), \\ 1/R = -\tan(\rho\gamma)\cdot\cot(\sigma\gamma), \\ R = -\tan(\sigma\gamma)\cdot\cot((\pi/2-\rho)\gamma), \\ 0 < \gamma < 2, \\ \max\{0, \pi\gamma - \pi\} < 2\gamma\rho < \min\{\pi\gamma, \pi\}, \\ \max\{0, \pi - \pi\gamma\} < -2\gamma\sigma < \min\{\pi, 2\pi - \pi\gamma\}. \end{cases} \tag{134}$$

Since we want to test the algorithm AFEM in a worst case scenario, we choose $\gamma = 0.1$, which produces a very singular solution u that is barely in H^1; in fact $u \in H^s(\Omega)$ for $s < 1.1$ but still piecewise in $W^2_p(\Omega)$ for some $1 < p < \frac{20}{19}$ (see Figure 16). We then solve (134) for R, ρ, and σ using Newton's method to obtain within computing precision

$$R = a_1/a_2 \cong 161.4476387975881, \quad \rho = \pi/4, \quad \sigma \cong -14.92256510455152,$$

and finally choose $a_1 = R$ and $a_2 = 1$. A smaller γ would lead to a larger ratio R, but in principle γ may be as close to 0 as desired.

We realize from Fig. 17 that AFEM attains optimal decay rate for the energy norm. As we have seen in Sect. 5.4, this is consistent with adaptive approximation for functions piecewise in $W^2_p(\Omega)$, but nonobvious for AFEM which does not have direct access to u. We also notice from Fig. 18 that a graded mesh with mesh-size of order 10^{-10} at the origin is achieved with about 2×10^3 elements. To reach a similar resolution with a uniform mesh we would need $N \approx 10^{20}$ elements! This example

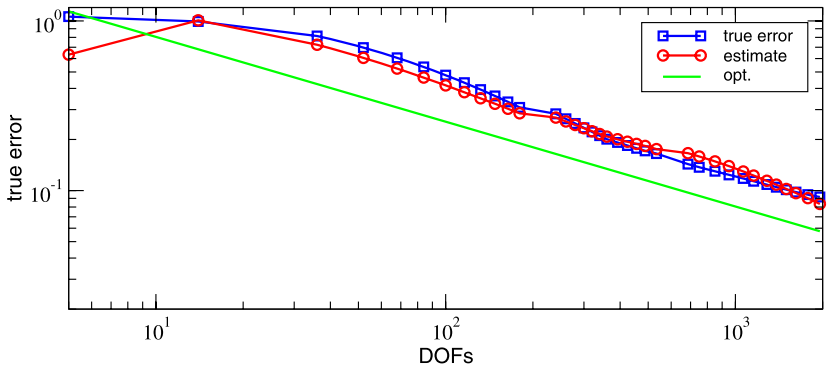


Fig. 17 Quasi-optimality of AFEM for discontinuous coefficients: estimate and true error. The optimal decay for piecewise linear elements in 2d is indicated by the green line with slope $-1/2$

clearly reveals the advantages and potentials of adaptivity within the FEM even with modest computational resources.

What is missing is an explanation of the recovery of optimal error decay $N^{-1/2}$ through mesh grading. This is the subject of Chap. 9, where we have to deal with the interplay between continuous and discrete quantities as already alluded to in the heuristics.

8.5 Problems

Problem 8.1 (Nesting of Spaces). If $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ satisfy $\mathscr{T}_1 \leq \mathscr{T}_2$, that is $\mathscr{T}_2$ is a refinement by bisection of $\mathscr{T}_1$, then the corresponding (Lagrange) finite element spaces are nested, i. e., $\mathbb{V}(\mathscr{T}_1) \subset \mathbb{V}(\mathscr{T}_2)$.

Problem 8.2 (Pythagoras). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V} = H^1_0(\Omega)$ be nested, conforming and closed subspaces. Let $u \in \mathbb{V}$ be the weak solution to (127), $U_1 \in \mathbb{V}_1$ and $U_2 \in \mathbb{V}_2$ the respective Ritz-Galerkin approximations to u. Prove the orthogonality property

$$|||u - U_1|||^2_\Omega = |||u - U_2|||^2_\Omega + |||U_2 - U_1|||^2_\Omega \,. \tag{135}$$

Problem 8.3 (Error in Energy). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V}$ and U_1, U_2, u be as in Problem 8.2. Recalling Problem 2.7, we know that u, U_1, U_2 are the unique minimizer of the quadratic energy

$$I[v] := \tfrac{1}{2}\mathscr{B}[v,v] - \langle f, v\rangle$$

in $\mathbb{V}, \mathbb{V}_1, \mathbb{V}_2$ respectively. Show that (135) is equivalent to the identity

$$I[U_1] - I[u] = (I[U_2] - I[u]) + (I[U_1] - I[U_2]).$$

To this end prove

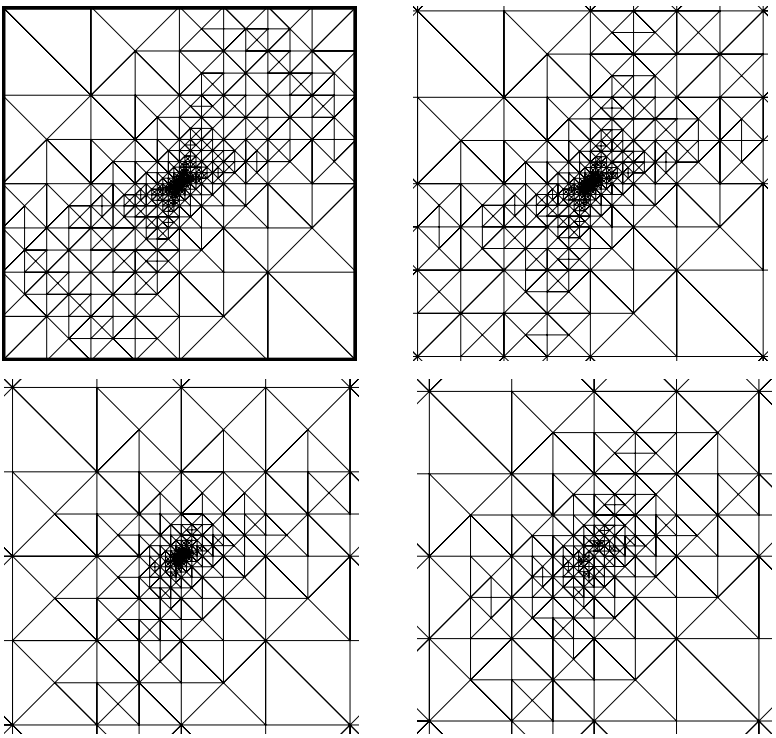


Fig. 18 Discontinuous coefficients in checkerboard pattern: Final grid (full grid with < 2000 nodes) (top left), zooms to $(-10^{-3},10^{-3})^2$ (top right), $(-10^{-6},10^{-6})^2$ (bottom left), and $(-10^{-9},10^{-9})^2$ (bottom right). The grid is highly graded towards the origin. For a similar resolution, a uniform grid would require $N \approx 10^{20}$ elements

$$I[U_i] - I[u] = \tfrac{1}{2}\,|||U_i - u|||_\Omega^2 \qquad \text{and} \qquad I[U_1] - I[U_2] = \tfrac{1}{2}\,|||U_1 - U_2|||_\Omega^2\,.$$

Problem 8.4. Let $S \in \mathring{\mathscr{S}}$ be a side of $T \in \mathscr{T}$, and let $\boldsymbol{A} \in W^1_\infty(T)$. Prove the following inverse estimate by a scaling argument to the reference element

$$\|\boldsymbol{A}\nabla V\|_S \lesssim h_S^{-1/2}\|\nabla V\|_T \quad \text{for all } V \in \mathbb{V}(\mathscr{T}),$$

where the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and $\|\boldsymbol{A}\|_{L^\infty(S)}$.

Problem 8.5. Let K be either a d or a $(d-1)$-simplex. For $\ell \in \mathbb{N}$ denote by $P_m^p\colon L^p(K,\mathbb{R}^\ell) \to \mathbb{P}_m(K,\mathbb{R}^\ell)$ the operator of best L^p-approximation in K. Then for all $v \in L^\infty(K,\mathbb{R}^\ell)$, $V \in \mathbb{P}_n(K,\mathbb{R}^\ell)$ and $m \geq n$, there holds

$$\|vV - P_m^2(vV)\|_{L^2(K)} \leq \|v - P_{m-n}^\infty v\|_{L^\infty(K)}\|V\|_{L^2(K)}.$$

Problem 8.6. Let $\boldsymbol{A} \in W^1_\infty(T)$ for all $T \in \mathscr{T}$. Prove the quasi-local Lipschitz property

$$|\operatorname{osc}_{\mathscr{T}}(V,T) - \operatorname{osc}_{\mathscr{T}}(W,T)| \lesssim \operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)\,\|V-W\|_{H^1(\omega_T)} \qquad \text{for all } V,W \in \mathbb{V},$$

where $\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T) = h_T\,\|\operatorname{div}\boldsymbol{A} - P^\infty_{n-1}(\operatorname{div}\boldsymbol{A})\|_{L^\infty(T)} + \|\boldsymbol{A} - P^\infty_n \boldsymbol{A}\|_{L^\infty(\omega_T)}$. Proceed as in the proof of Lemma 8.4 and use Problem 8.5.

Problem 8.7. Let $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$, with $\mathscr{T} \leq \mathscr{T}_*$. Use Problem 8.6 to prove that, for all $V \in \mathbb{V}(\mathscr{T})$ and $V_* \in \mathbb{V}(\mathscr{T}_*)$, there is a constant $\Lambda_1 > 0$ such that

$$\operatorname{osc}^2_{\mathscr{T}}(V, \mathscr{T} \cap \mathscr{T}_*) \leq 2\operatorname{osc}^2_{\mathscr{T}_*}(V_*, \mathscr{T} \cap \mathscr{T}_*) + \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A}, \mathscr{T}_0)^2\, |||V - V_*|||^2_\Omega\,.$$

9 Adaptivity: Convergence rates

We have already realized in Chap. 5 that we can a priori accommodate the degrees of freedom in such a way that the finite element approximation retains optimal energy error decay for a class of singular functions. This presumes knowledge of the exact solution u. At the same time, we have seen numerical evidence in Sect. 8.4 that the AFEM of Chap. 8, achieves such a performance without direct access to the regularity of u. Practical experience strongly suggests that this is even true for a much larger class of problems and adaptive methods. The challenge ahead is to reconcile these two distinct aspects of AFEM. In doing this we have to restrict ourselves to the setting of Chap. 8. The mathematical foundation to justify the observed optimal error decay of adaptive methods in case of non-symmetric or non-coercive bilinear forms and other marking strategies is completely open.

One key to connect the two worlds for the simplest scenario, the Laplacian and f piecewise constant, is due to Stevenson [69]: *any marking strategy that reduces the energy error relative to the current value must contain a substantial bulk of* $\mathscr{E}_{\mathscr{T}}(U,\mathscr{T})$*, and so it can be related to Dörfler Marking*. This allows us to compare AFEM with an optimal mesh choice and to conclude optimal error decay.

The objective of this section is to study the model problem (127) for general data f and $\boldsymbol{A}$ and the AFEM from Chap. 8. In what follows it is important to use an error notion that is strictly reduced by the adaptive method. In this section we closely follow Cascón et al. [21] by utilizing the quasi-error as contracting quantity. This approach allows us to include variable data f and $\boldsymbol{A}$ and thus improves upon and extends Stevenson [69].

As in Chap. 8, u will always be the weak solution of (127) and, except when stated otherwise, any constant explicit or hidden constant in $\lesssim$ may depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, the (global) eigenvalues of $\boldsymbol{A}$, and the oscillation $\operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)$ of $\boldsymbol{A}$ on the initial mesh $\mathscr{T}_0$, but not on a specific grid $\mathscr{T} \in \mathbb{T}$.

9.1 Approximation class

Since AFEM selects elements for refinement based on information provided exclusively by the error indicators $\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}}$, it is natural that the measure of regularity and ensuing decay rate is closely related to the indicators. We explore this connection now.

The Total Error. We first introduce the concept of *total error* [48]

$$|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}),$$

and next assert that it is equivalent to the quasi error, for the Galerkin function $U\in\mathbb{V}(\mathscr{T})$. In fact, in view of the upper and lower a posteriori error bounds (130a) and (130b), and

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\le\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})$$

we have

$$\begin{aligned}C_2\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})&\le|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\\&\le|||u-U|||_{\Omega}^2+\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})\le(1+C_1)\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}),\end{aligned}$$

whence

$$\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})\approx|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2\,. \tag{136}$$

We thus realize that the decay rate of AFEM must be characterized by the total error. Moreover, on invoking the upper bound once again, we also see that the total error is equivalent to the quasi error

$$|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\approx|||u-U|||_{\Omega}^2+\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}).$$

This is the quantity being strictly reduced by AFEM (Theorem 8.1). Finally, the total error satisfies the following Cea's type-lemma. In fact, if $\boldsymbol{A}$ is piecewise constant, then this is Cea's Lemma stated in Problem 3.1.

Lemma 9.1 (Quasi-Optimality of Total Error). *There exists a constant Λ_2, such that for any $\mathscr{T}\in\mathbb{T}$ and the corresponding Ritz–Galerkin solution $U\in\mathbb{V}(\mathscr{T})$ holds*

$$|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\le\Lambda_2\inf_{V\in\mathbb{V}(\mathscr{T})}\Big(|||u-V|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(V,\mathscr{T})\Big).$$

Proof. For $\varepsilon>0$ choose $V_\varepsilon\in\mathbb{V}(\mathscr{T})$, with

$$|||u-V_\varepsilon|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon,\mathscr{T})\le(1+\varepsilon)\inf_{V\in\mathbb{V}(\mathscr{T})}\Big(|||u-V|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(V,\mathscr{T})\Big).$$

Applying Problem 8.7 with $\mathscr{T}_*=\mathscr{T}$, $V=U$, and $V_*=V_\varepsilon$ yields

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\le2\operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon,\mathscr{T})+C_3\,|||U-V_\varepsilon|||_{\Omega}^2\,,$$

with

$$C_3 := \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A}, \mathscr{T}_0)^2.$$

Since $U \in \mathbb{V}(\mathscr{T})$ is the Galerkin solution, $U - V_\varepsilon \in \mathbb{V}(\mathscr{T})$ is orthogonal to $u - U$ in the energy norm, whence $|||u-U|||_\Omega^2 + |||U-V_\varepsilon|||_\Omega^2 = |||u-V_\varepsilon|||_\Omega^2$ and

$$\begin{aligned} |||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}) &\le \big(1+C_3\big)\, |||u-V_\varepsilon|||_\Omega^2 + 2\operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon,\mathscr{T}) \\ &\le (1+\varepsilon)\,\Lambda_2 \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(V,\mathscr{T}) \Big), \end{aligned}$$

with $\Lambda_2 = \max\big\{2, 1+C_3\big\}$, and the assertion follows from $\varepsilon \to 0$. □

We next give a definition of an appropriate approximation class $\mathbb{A}_s$ that hinges on the concept of total error. We first let $\mathbb{T}_N \subset \mathbb{T}$ be the set of all possible conforming refinements of $\mathscr{T}_0$ with at most N elements more than $\mathscr{T}_0$, i. e.,

$$\mathbb{T}_N = \{\mathscr{T} \in \mathbb{T} \mid \#\mathscr{T} - \#\mathscr{T}_0 \le N\}.$$

The quality of the best approximation in $\mathbb{T}_N$ with respect to the total error is characterized by

$$\sigma(N;u,f,\boldsymbol{A}) := \inf_{\mathscr{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathscr{T})} \big(|||u-V|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(V,\mathscr{T}) \big)^{1/2},$$

and the approximation class $\mathbb{A}_s$ for $s > 0$ is defined by

$$\mathbb{A}_s := \Big\{ (v,f,\boldsymbol{A}) \mid |v,f,\boldsymbol{A}|_s := \sup_{N>0} \big(N^s\, \sigma(N;v,f,\boldsymbol{A}) \big) < \infty \Big\}.$$

Thanks to Lemma 9.1, the solution u with data $(f,\boldsymbol{A})$ satisfies

$$\sigma(N;u,f,\boldsymbol{A}) \approx \inf_{\mathscr{T}\in\mathbb{T}_N} \big\{ \mathscr{E}_{\mathscr{T}}(U,\mathscr{T}) \mid U = \mathsf{SOLVE}(\mathbb{V}(\mathscr{T})) \big\}. \tag{137}$$

We point out the upper bound $s \le n/d$ for polynomial degree $n \ge 1$; this can be seen with full regularity and uniform refinement (see (69)). Note that if $(v,f,\boldsymbol{A}) \in \mathbb{A}_s$ then for all $\varepsilon > 0$ there exist $\mathscr{T}_\varepsilon \ge \mathscr{T}_0$ conforming and $V_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that (see Problem 9.1)

$$|||v-V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2 \le \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \le |v,f,\boldsymbol{A}|_s^{1/s}\, \varepsilon^{-1/s}. \tag{138}$$

For the subsequent discussion we recall Lemma 4.3: the overlay $\mathscr{T}_1 \oplus \mathscr{T}_2 \in \mathbb{T}$ of two meshes $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ is the smallest common refinement of $\mathscr{T}_1$ and $\mathscr{T}_2$ and

$$\#\mathscr{T}_1 \oplus \mathscr{T}_2 \le \#\mathscr{T}_1 + \#\mathscr{T}_2 - \#\mathscr{T}_0. \tag{139}$$

We first investigate the class $\mathbb{A}_s$ for piecewise constant coefficient matrix $\boldsymbol{A}$ with respect to $\mathscr{T}_0$. In this simplified scenario, the oscillation $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ reduces to *data oscillation* (see Remark 6.3):

$$\operatorname{osc}_{\mathscr{T}} = \|h_{\mathscr{T}}(f - P_{2n-2} f)\|_{L^2(\Omega)}.$$

We then have the following characterization of $\mathbb{A}_s$ in terms of the standard approximation classes [13, 14, 69]:

$$\mathscr{A}_s := \Big\{ v \in \mathbb{V} \mid |v|_{\mathscr{A}_s} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathscr{T})} |||v - V|||_\Omega \big) < \infty \Big\},$$
$$\bar{\mathscr{A}_s} := \Big\{ g \in L^2(\Omega) \mid |g|_{\bar{\mathscr{A}_s}} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \|h_{\mathscr{T}}(g - P_{2n-2} g)\|_{L^2(\Omega)} \big) < \infty \Big\}.$$

Lemma 9.2 (Equivalence of Classes). *Let $\boldsymbol{A}$ be piecewise constant over $\mathscr{T}_0$. Then $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ if and only if $(u, f) \in \mathscr{A}_s \times \bar{\mathscr{A}_s}$ and*

$$|u, f, \boldsymbol{A}|_s \approx |u|_{\mathscr{A}_s} + |f|_{\bar{\mathscr{A}_s}}. \tag{140}$$

Proof. It is obvious that $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ implies $(u, f) \in \mathscr{A}_s \times \bar{\mathscr{A}_s}$ as well as the bound $|u|_{\mathscr{A}_s} + |f|_{\bar{\mathscr{A}_s}} \lesssim |u, f, \boldsymbol{A}|_s$.

In order to prove the reverse inequality, let $(u, f) \in \mathscr{A}_s \times \bar{\mathscr{A}_s}$. Then there exist $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}_N$ so that $|||u - U|||_\Omega \le |u|_{\mathscr{A}_s} N^{-s}$ where $U \in \mathbb{V}(\mathscr{T}_1)$ is the best approximation and $\|h_{\mathscr{T}_2}(f - P^2_{2n-2} f)\|_{L^2(\Omega)} \le |f|_{\bar{\mathscr{A}_s}} N^{-s}$.

The overlay $\mathscr{T} = \mathscr{T}_1 \oplus \mathscr{T}_2 \in \mathbb{T}_{2N}$ according to (139), and

$$|||u - U|||^2_\Omega + \operatorname{osc}^2_{\mathscr{T}} \le |||u - U|||^2_\Omega + \operatorname{osc}^2_{\mathscr{T}_2} \le 2^s \big(|u|^2_{\mathscr{A}_s} + |f|^2_{\bar{\mathscr{A}_s}} \big) (2N)^{-s}.$$

This yields $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ together with the bound $|u, f, \boldsymbol{A}|_s \lesssim |u|_{\mathscr{A}_s} + |f|_{\bar{\mathscr{A}_s}}$. □

We next turn to the special case of linear finite elements.

Corollary 9.1 (Membership in $\mathbb{A}_{1/2}$). *Let $d = 2$, polynomial degree $n = 1$, $f \in L^2(\Omega)$, and $\boldsymbol{A}$ piecewise constant with respect to $\mathscr{T}_0$. If $u \in W^2_p(\Omega; \mathscr{T}_0)$ for some $p > 1$, then $(u, f, \boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u, f, \boldsymbol{A}|_{1/2} \lesssim \|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}$$

Proof. We start with the data oscillation $\operatorname{osc}_{\mathscr{T}}$, and realize that

$$\operatorname{osc}_{\mathscr{T}} = \|h_{\mathscr{T}}(f - P_0 f)\|_{L^2(\Omega)} \le h_{\max}(\mathscr{T}) \|f\|_{L^2(\Omega)} \lesssim (\#\mathscr{T})^{-1/2} \|f\|_{L^2(\Omega)},$$

for any uniform refinement $\mathscr{T} \in \mathbb{T}$. This implies $f \in \bar{\mathscr{A}_{1/2}}$ with $|f|_{\bar{\mathscr{A}_{1/2}}} \lesssim \|f\|_{L^2(\Omega)}$.

For $u \in W^2_p(\Omega; \mathscr{T}_0)$ we learn from Corollary 5.2 and Remark 5.6 that $u \in \mathscr{A}_{1/2}$ and $|u|_{\mathscr{A}_{1/2}} \lesssim \|D^2 u\|_{L^2(\Omega;\mathscr{T}_0)}$. The assertion then follows from Lemma 9.2. □

Example 9.1 (Pre-asymptotics). Corollary 9.1 shows that oscillation decays at least with rate $1/2$ for $f \in L^2(\Omega)$. Since the decay rate of the total error is $s \le 1/2$, oscillation can be ignored asymptotically. However, Remark 6.4 shows that oscillation may dominate the total error, or equivalently the class $\mathbb{A}_s$ may fail to describe the behavior of $|||u - U_k|||_\Omega$, in the early stages of adaptivity. In fact, we

recall that $\operatorname{osc}_k(U_k,\mathscr{T}_k) = \|h_k(f-P_0f)\|_{L^2(\Omega)}$, the discrete solution $U_k = 0$, and $|||u-U_k|||_\Omega \approx 2^{-K}$ is constant for as many steps $k \le K$ as desired. In contrast, $\mathscr{E}_k(U_k,\mathscr{T}_k) = \operatorname{osc}_k(U_k,\mathscr{T}_k) = \|h_kf\|_{L^2(\Omega)}$ reduces strictly for $k \le K$ but overestimates $|||u-U_k|||_\Omega$. The fact that the preasymptotic regime $k \le K$ for the energy error could be made arbitrarily long would be problematic if we focus exclusively on $|||u-U_k|||_\Omega$. In practice, this effect is typically less dramatic because f is not orthog-

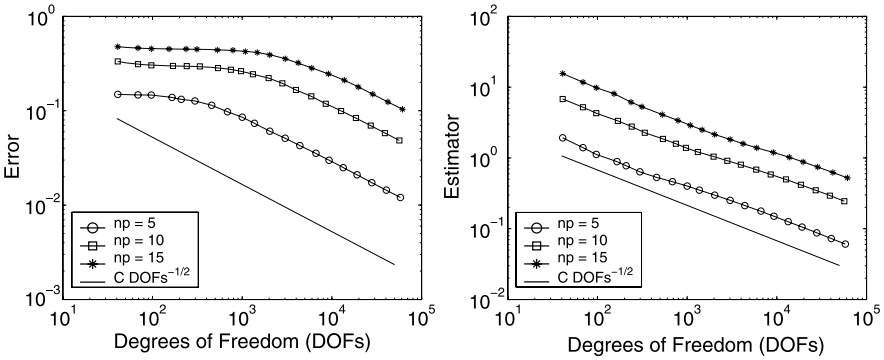


Fig. 19 Decay of the energy error (left) and the estimator (right) for the smooth solution u_S of (141) with frequencies $\kappa = 5, 10$, and 15. The energy error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. This behavior is allowed by $\mathbb{A}_s$

onal to $\mathbb{V}(\mathscr{T}_k)$. Figure 19 displays the behavior of AFEM for the smooth solution $u = u_S$ given by

$$u_S(x,y) = 10^{-2}a_i^{-1}(x^2+y^2)\sin^2(\kappa\pi x)\sin^2(\kappa\pi y), \qquad 1 \le i \le 4. \tag{141}$$

of the problem in Sect. 8.4 with discontinuous coefficients $\{a_i\}_{i=1}^4$ in checkerboard pattern and frequencies $\kappa = 5, 10$, and 15. We can see that the error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. In contrast, the estimator decays always with the optimal rate. Since all decisions of the AFEM are based on the estimator, this behavior has to be expected and is consistent with our notion of approximation class $\mathbb{A}_s$, which can be characterized just by the estimator according to (137).

We next turn to the nonlinear interaction encoded in $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ via the product $\boldsymbol{A}\nabla U$. It is this interaction which makes the class $\mathbb{A}_s$ a non-standard object in approximation theory that deserves further scrutiny.

Lemma 9.3 (Decay Rate of Oscillation). *Let $\boldsymbol{A}$ be piecewise Lipschitz with respect to $\mathscr{T}_0$, $f \in L^2(\Omega)$, and polynomial degree $n = 1$. If $U \in \mathbb{V}(\mathscr{T})$ is the Ritz-Galerkin solution, then oscillation $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ has at least a decay rate of order $-1/d$*

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \operatorname{osc}_{\mathscr{T}}(U,\mathscr{T}) \lesssim \Big(\|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)}\Big)N^{-1/d}.$$

Proof. Let $\mathscr{T} \in \mathbb{T}_N$ be a uniform refinement of $\mathscr{T}_0$ with $\#\mathscr{T} \approx N$. By applying Problem 8.6 with $V = U$ and $W = 0$, we obtain

$$\operatorname{osc}_{\mathscr{T}}(U,T) \lesssim h_T \|f - P_0^2 f\|_{L^2(T)} + \operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)\|U\|_{H^1(\omega_T)}$$

with $h_T\|f - P_0^2 f\|_{L^2(T)} \le h_T\|f\|_{L^2(T)}$ and

$$\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T) = h\|\operatorname{div}\boldsymbol{A} - P_0^\infty(\operatorname{div}\boldsymbol{A})\|_{L^\infty(T)} + \|\boldsymbol{A} - P_1^\infty \boldsymbol{A}\|_{L^\infty(\omega_T)} \lesssim h_T\|\boldsymbol{A}\|_{W^1_\infty(\omega_T;\mathscr{T}_0)}.$$

Uniform refinement yields the relation $h_T \approx N^{-1/d}$ for all $T \in \mathscr{T}$, whence

$$\operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) = \sum_{T\in\mathscr{T}} \operatorname{osc}^2_{\mathscr{T}}(U,T) \lesssim \Big(\|f\|^2_{L^2(\Omega)} + \|\boldsymbol{A}\|^2_{W^1_\infty(\Omega;\mathscr{T}_0)}\Big)N^{-2/d},$$

because $\|U\|_{H^1(\Omega)} \le \alpha_1^{-1}\|f\|_{L^2(\Omega)}$ according to (40). □

Remark 9.1 (Asymptotic Order of Oscillation). Let's assume the following piecewise regularity of data $(f,\boldsymbol{A})$ with respect to a conforming refinement $\mathscr{T}_*$ of $\mathscr{T}_0$:

$$f \in H^1(\Omega;\mathscr{T}_*), \quad \boldsymbol{A} \in W^2_\infty(\Omega;\mathscr{T}_*).$$

The proof of Lemma 9.3, in conjunction with Proposition 5.1(a), shows that for $n = 1$

$$\inf_{\mathscr{T}\in\mathbb{T}_N:\mathscr{T}\ge\mathscr{T}_*} \operatorname{osc}_{\mathscr{T}}(U,\mathscr{T}) \lesssim \Big(\|f\|_{H^1(\Omega;\mathscr{T}_*)} + \|\boldsymbol{A}\|_{W^2_\infty(\Omega;\mathscr{T}_*)}\Big)N^{2/d},$$

and the rate in Lemma 9.3 can be improved. Since the energy error decay is never better than $N^{-1/d}$, according to (69), we realize that oscillation is of higher order than the energy error asymptotically as $N \uparrow \infty$; compare with Remark 6.1.

Corollary 9.2 (Membership in $\mathbb{A}_{1/2}$). *Let $d = 2$, polynomial degree $n = 1$, $\boldsymbol{A} \in W^1_\infty(\Omega;\mathscr{T}_0)$, and $f \in L^2(\Omega)$. If $u \in W^2_p(\Omega;\mathscr{T}_0)$ for some $p > 1$, then $(u,f,\boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u,f,\boldsymbol{A}|_{1/2} \lesssim \|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}.$$

Proof. Repeat the proof of Corollary 9.1 with the help of Lemma 9.3. □

A complete characterization of $\mathbb{A}_s$ for general d and n is still missing. It is important to realize that the nonlinear interaction between data $\boldsymbol{A}$ and U must be accounted for, thereby leading to a new concept of approximation class $\mathbb{A}_s$, which generalizes those in [13, 14, 69]. It is worth mentioning that a near characterization of the standard approximation class $\mathscr{A}_s$ in terms of Besov spaces for $d = 2$ can be found in [13, 14, 37]: $u \in \mathscr{A}_s$ implies that $u \in B^{2s+1}_p(L^p(\Omega))$ for $p = \frac{2}{2s+1}$ [13, Theorem 9.3]; $u \in B^{2s+1}_p(L^p(\Omega))$ for $p > \frac{2}{2s+1}$ implies that $u \in \mathscr{A}_s$ [13, Theorem 9.1]. Note that $p < 1$ for $s > 1/2$; see Remark 5.7.

9.2 Cardinality of $\mathscr{M}_k$

To assess the performance of AFEM in terms of degrees of freedom $\#\mathscr{T}_k$, we need to impose further restrictions on the modules of AFEM beyond those of Sect. 8.1. We recall that $C_2 \le C_1$ are the constants in (130a) and (130b) and $C_3 = \Lambda_1 \operatorname{osc}^2_{\mathscr{T}_0}(\boldsymbol{A}, \mathscr{T}_0)$ is the constant in Problem 8.7 and Lemma 9.1.

Assumption 11.2 (Assumptions for Optimal Decay Rate). *We assume the following additional properties of the marking procedure* MARK *and the initial grid* $\mathscr{T}_0$*:*

(a) *The marking parameter* θ *of Dörfler Marking satisfies* $\theta \in (0, \theta^*)$ *with*

$$\theta_*^2 = \frac{C_2}{1 + C_1(1 + C_3)};$$

(b) MARK *outputs a set* $\mathscr{M}$ *with minimal cardinality;*

(c) *The initial triangulation* $\mathscr{T}_0$ *satisfies Assumption 11.1.*

A few comments are now in order.

- *Threshold* $\theta_* < 1$: We first point out that, according to (130a) and (130b), the ratio $C_2/C_1 \le 1$ is a quality measure of the estimator $\mathscr{E}_{\mathscr{T}}(U, \mathscr{T})$: the closer to 1 the better! It is thus natural to be cautious in marking if the reliability constant C_1 and efficiency constant C_2 are very disparate. The additional factor C_3 accounts for the effect of a function dependent oscillation (see Problem 8.7), and is zero if the oscillation just depends on data f because then $\operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A}, \mathscr{T}_0) = 0$.
- *Minimal* $\mathscr{M}_k$: According to Remark 6.2 about the significance of the local lower a posteriori error estimate for relatively small oscillation, it is natural to mark elements with largest error indicators. This leads to a minimal set $\mathscr{M}_k$ and turns out to be crucial to link AFEM with optimal meshes and approximation classes.
- *Initial Triangulation*: The initial labeling of the element's vertices on $\mathscr{T}_0$ stated in Assumption 11.1 of Sect.4.2 is rather restrictive for dimension $d > 2$ but guarantees the complexity estimate of Theorem 4.3 for our module REFINE. Any other refinement ensuing the same complexity estimate can replace REFINE together with the assumption on $\mathscr{T}_0$.

We stress that we cannot expect local upper bounds between the continuous solution u and discrete solution U due to the global nature of the underlying PDE: the error in a region may be dictated by pollution effects arising somewhere else. The following crucial result shows, however, that this is a matter of scale: if $\mathscr{T}_* \ge \mathscr{T}$, then what determines the error between Galerkin solutions $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ is the refined set $\mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$, namely the region of Ω where the scale of resolution differs from $\mathscr{T}$ to $\mathscr{T}_*$. This is not, of course, in contradiction with the previous statement because one needs an infinitely fine scale to reach the exact solution u.

Lemma 9.4 (Localized Upper Bound). *Let* $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ *satisfy* $\mathscr{T}_* \ge \mathscr{T}$ *and define* $\mathscr{R} := \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ *to be the set of refined elements in* $\mathscr{T}$*. If* $U \in \mathbb{V}(\mathscr{T})$ *and* $U_* \in \mathbb{V}(\mathscr{T}_*)$ *are the corresponding Galerkin solutions, then*

$$|||U_* - U|||_\Omega^2 \le C_1\, \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

where $C_1 > 0$ is the same constant as in (130a).

Proof. Problem 9.2. □

The following result reveals the importance of Dörfler's marking in the present context. The original result, established by Stevenson [69], referred to the energy error alone. We follow [21] in this analysis.

Lemma 9.5 (Optimal Marking). *Let the marking parameter θ satisfy Assumption 11.2(a) and set $\mu := \frac{1}{2}(1-\frac{\theta^2}{\theta_*^2}) > 0$. For $\mathscr{T}_* \ge \mathscr{T}$ let the corresponding Galerkin solution $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfy*

$$|||u - U_*|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*) \le \mu\big(|||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\big). \tag{142}$$

Then the set $\mathscr{R} = \mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ of refined elements of $\mathscr{T}$ satisfies the Dörfler property

$$\mathscr{E}_{\mathscr{T}}(U,\mathscr{R}) \ge \theta\, \mathscr{E}_{\mathscr{T}}(U,\mathscr{T}). \tag{143}$$

Proof. We split the proof into four steps.
1 In view of the global lower bound (130b)

$$C_2\, \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}) \le |||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})$$

and (142), we can write

$$\begin{aligned}(1-2\mu)C_2\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}) &\le (1-2\mu)\big(|||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\big)\\ &\le \big(|||u-U|||_\Omega^2 - 2\,|||u-U_*|||_\Omega^2\big) + \big(\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}) - 2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*)\big).\end{aligned}$$

2 Combining the orthogonality relation (128)

$$|||u-U|||_\Omega^2 - |||u-U_*|||_\Omega^2 = |||U-U_*|||_\Omega^2.$$

with the localized upper bound Lemma 9.4 yields

$$|||u-U|||_\Omega^2 - 2\,|||u-U_*|||_\Omega^2 \le C_1\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

3 To deal with oscillation we decompose the elements of $\mathscr{T}$ into two disjoint sets: $\mathscr{R}$ and $\mathscr{T}\setminus\mathscr{R}$. In the former case, we have

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{R}) - 2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{R}) \le \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{R}) \le \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}),$$

because $\operatorname{osc}_{\mathscr{T}}(U,T) \le \mathscr{E}_{\mathscr{T}}(U,T)$ for all $T \in \mathscr{T}$. On the other hand, we use that $\mathscr{T}\setminus\mathscr{R} = \mathscr{T}\cap\mathscr{T}_*$ and apply Problem 8.7 in conjunction with Lemma 9.4 to arrive at

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}\setminus\mathscr{R}) - 2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}\setminus\mathscr{R}) \le C_3\,|||U-U_*|||_\Omega^2 \le C_1C_3\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

Adding these two estimates gives

$$\operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) - 2\operatorname{osc}^2_{\mathscr{T}_*}(U_*,\mathscr{T}_*) \le (1+C_1C_3)\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{R}).$$

4 Returning to 1 we realize that

$$(1-2\mu)C_2\,\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}) \le \big(1+C_1(1+C_3)\big)\,\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{R}),$$

which is (143) in disguise. In fact, recalling that $\theta_*^2 = C_2/\big(1+C_1(1+C_3)\big)$ then $\theta^2 = (1-2\mu)\theta_*^2 < \theta_*^2$ as asserted. □

We are now ready to explore the cardinality of $\mathscr{M}_k$. To this end, we must relate AFEM with the approximation class $\mathbb{A}_s$. This might appear like an undoable task. However, the key to unravel this connection is given by Lemma 9.5.

Lemma 9.6 (Cardinality of $\mathscr{M}_k$). *Let Assumptions 11.2(a) and 11.2(b) be satisfied. If $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ then*

$$\#\mathscr{M}_k \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,\big(|||u-U_k|||_\Omega + \operatorname{osc}_k(U_k,\mathscr{T}_k)\big)^{-1/s} \quad \text{for all } k \ge 0. \tag{144}$$

Proof. We split the proof into three steps.
1 We set $\varepsilon^2 := \mu\Lambda_2^{-1}\big(|||u-U_k|||^2_\Omega + \operatorname{osc}^2_k(U_k,\mathscr{T}_k)\big)$ with $\mu = \frac{1}{2}\big(1-\frac{\theta^2}{\theta_*^2}\big) > 0$ as in Lemma 9.5 and Λ_2 given Lemma 9.1. Since $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$, in view of (138) there exists $\mathscr{T}_\varepsilon \in \mathbb{T}$ and $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||u-U_\varepsilon|||^2_\Omega + \operatorname{osc}^2_\varepsilon(U_\varepsilon,\mathscr{T}_\varepsilon) \le \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/2}\,\varepsilon^{-1/s}.$$

Since $\mathscr{T}_\varepsilon$ may be totally unrelated to $\mathscr{T}_k$ we introduce the overlay

$$\mathscr{T}_* = \mathscr{T}_k \oplus \mathscr{T}_\varepsilon.$$

2 We claim that the total error over $\mathscr{T}_*$ reduces by a factor μ relative to that one over $\mathscr{T}_k$. In fact, since $\mathscr{T}_* \ge \mathscr{T}_\varepsilon$ and so $\mathbb{V}(\mathscr{T}_*) \supset \mathbb{V}(\mathscr{T}_\varepsilon)$, we use Lemma 9.1 to obtain

$$\begin{aligned}|||u-U_*|||^2_\Omega + \operatorname{osc}^2_{\mathscr{T}_*}(U_*,\mathscr{T}_*) &\le \Lambda_2\Big(|||u-U_\varepsilon|||^2_\Omega + \operatorname{osc}^2_\varepsilon(U_\varepsilon,\mathscr{T}_\varepsilon)\Big)\\ &\le \Lambda_2\varepsilon^2 = \mu\,\big(|||u-U_k|||^2_\Omega + \operatorname{osc}^2_k(U_k,\mathscr{T}_k)\big).\end{aligned}$$

Upon applying Lemma 9.5 we conclude that the set $\mathscr{R} = \mathscr{R}_{\mathscr{T}_k\to\mathscr{T}_*}$ of refined elements satisfies a Dörfler marking (143) with parameter $\theta < \theta_*$.
3 According to Assumption 11.2(b) MARK selects a minimal set $\mathscr{M}_k$ satisfying this property. Therefore, we deduce

$$\#\mathscr{M}_k \le \#\mathscr{R} \le \#\mathscr{T}_* - \#\mathscr{T}_k \le \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,\varepsilon^{-1/s},$$

where we have employed Lemma 4.3 for the overlay. Now recalling the definition of ε we end up with the asserted estimate (144). □

Remark 9.2 (Blow-up). The constant hidden in (144) blows up as $\theta \uparrow \theta_*$ because $\mu \downarrow 0$.

9.3 Quasi-optimal convergence rates

We are ready to prove the main result of this section, which combines Theorem 8.1 and Lemma 9.6.

Theorem 9.1 (Quasi-Optimality). *Let Assumption 11.2 be satisfied. If $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ then AFEM gives rise to a sequence $\{\mathscr{T}_k,\mathbb{V}_k,U_k\}_{k=0}^{\infty}$ such that*

$$|||u-U_k|||_\Omega + \operatorname{osc}_k(U_k,\mathscr{T}_k) \lesssim |u,f,\boldsymbol{A}|_s\,(\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s} \quad \text{for all } k \ge 1.$$

Proof. 1 Since no confusion arises, we use the notation $\operatorname{osc}_j = \operatorname{osc}_j(U_j,\mathscr{T}_j)$ and $\mathscr{E}_j = \mathscr{E}_j(U_j,\mathscr{T}_j)$. In light of Assumption 11.2(c) and (144) we have

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim \sum_{j=0}^{k-1} \#\mathscr{M}_j \lesssim |u,f,\boldsymbol{A}|_s^{1/s} \sum_{j=0}^{k-1} \big(|||u-U_j|||_\Omega^2 + \operatorname{osc}_j^2\big)^{-1/(2s)}.$$

2 Let $\gamma>0$ be the scaling factor in the (contraction) Theorem 8.1. The lower bound (130b) along with $\operatorname{osc}_j \le \mathscr{E}_j$ implies

$$|||u-U_j|||_\Omega^2 + \gamma \operatorname{osc}_j^2 \le |||u-U_j|||_\Omega^2 + \gamma\mathscr{E}_j^2 \le \Big(1+\frac{\gamma}{C_2}\Big)\big(|||u-U_j|||_\Omega^2 + \operatorname{osc}_j^2\big).$$

3 Theorem 8.1 yields for $0 \le j < k$

$$|||u-U_k|||_\Omega^2 + \gamma\mathscr{E}_k^2 \le \alpha^{2(k-j)}\big(|||u-U_j|||_\Omega^2 + \gamma\mathscr{E}_j^2\big),$$

whence

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\big(|||u-U_k|||_\Omega^2 + \gamma\mathscr{E}_k^2\big)^{-1/(2s)} \sum_{j=0}^{k-1} \alpha^{(k-j)/s}.$$

Since $\sum_{j=0}^{k-1}\alpha^{(k-j)/s} = \sum_{j=1}^{k}\alpha^{j/s} < \sum_{j=1}^{\infty}\alpha^{j/s} < \infty$ because $\alpha<1$, the assertion follows immediately. □

Corollary 9.3 (Estimator Decay). *Let Assumption 11.2 be satisfied. If $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ then the estimator $\mathscr{E}_k(U_k,\mathscr{T}_k)$ satisfies*

$$\mathscr{E}_k(U_k,\mathscr{T}_k) \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,(\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s}.$$

Proof. Use (136) and Theorem 9.1. □

Corollary 9.4 (W^2_p-Regularity). *Let $d = 2$, the polynomial degree $n = 1$, $f \in L^2(\Omega)$, and let $\boldsymbol{A}$ be piecewise constant over $\mathscr{T}_0$. If $u \in W^2_p(\Omega; \mathscr{T}_0)$ for $p > 1$, then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying $\operatorname{osc}_k = \|h_k(f - P_0 f)\|_{L^2(\Omega)}$ and*

$$|||u - U_k|||_{\Omega} + \operatorname{osc}_k \lesssim \left(\|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} \right) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}$$

for all $k \geq 1$.

Proof. Combine Corollary 9.1 with Theorem 9.1. □

Corollary 9.5 (W^2_p-Regularity). *Besides the assumptions of Corollary 9.4, let $\boldsymbol{A}$ be piecewise Lipschitz over the initial grid $\mathscr{T}_0$. Then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying for all $k \geq 1$*

$$|||u - U_k|||_{\Omega} + \operatorname{osc}_k(U_k, \mathscr{T}_k) \lesssim \left(\|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} \right) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}.$$

Proof. Combine Corollary 9.2 with Theorem 9.1. □

So far we have assumed that the module SOLVE gives the *exact* Galerkin solution U_k and in doing this we have ignored the effects of numerical integration and inexact solution of the linear system; recall Remarks 3.6 and 3.7. The two issues above are important for AFEM to be fully practical. If one could control a posteriori the errors due to inexactness of SOLVE, then it would still be possible to prove a contraction property, as in Chap. 8, and examine the number of operations of AFEM in terms of $\#\mathscr{T}_k$ for a desired accuracy, following the steps of Sect. 9.2 and Sect. 9.3. We refer to Stevenson [69], who explores this endeavor for problem (127) with $\boldsymbol{A}$ piecewise constant.

9.4 Marking vs optimality

We conclude with a brief discussion of processes that optimize more than one quantity at once and the critical role of marking, i. e., we consider adaptive algorithms that mark in each iteration for different error contributions separately. For instance, in earlier work on adaptivity error indicator and oscillation are treated independently [52, 53, 48]. Furthermore, when dealing with systems one is easily tempted to mark separately for the different components; compare for instance with [40]. It is worth observing that Binev et al. [13], Stevenson [69] and also Cascón et al. [21] avoided to use separate marking in their algorithms when proving optimal error decay. When dealing with the Poisson problem, oscillation becomes data oscillation and allows one to first approximate data sufficiently well and then reduce the energy error. This is done in different ways in [13] and [69]. However, for variable $\boldsymbol{A}$ the oscillation depends on the discrete solution, as discussed in Sect. 9.1, and the above splitting

does not apply. Nonetheless marking solely for the estimator gives an optimal decay rate according to Sect. 9.3.

The design of adaptive algorithms that rely on separate marking is extremely delicate when aiming for optimal decay rates. To shed light on this issue we first present some numerical experiments based on separate marking, and next analyze the effect of separate marking in a simplified setting.

9.4.1 Separate Marking

The procedure ESTIMATE of Morin, Nochetto and Siebert, used in previous convergence proofs [52, 53, 48], calculates both the error and oscillation indicators $\{\mathscr{E}_k(U_k,T), \operatorname{osc}_k(U_k,T)\}_{T\in\mathscr{T}_k}$ (see Remarks 6.1 and 9.1), and the procedure MARK uses Dörfler marking for both the estimator and oscillation. More precisely, the routine MARK is of the form: *given parameters* $0<\theta_{\rm est},\theta_{\rm osc}<1$,

$$\text{mark any subset } \mathscr{M}_k\subset\mathscr{T}_k \text{ such that } \mathscr{E}_k(U_k,\mathscr{M}_k)\geq\theta_{\rm est}\mathscr{E}_k(U_k,\mathscr{T}_k); \tag{145a}$$

$$\text{if necessary enlarge } \mathscr{M}_k \text{ to satisfy } \operatorname{osc}_k(U_k,\mathscr{M}_k)\geq\theta_{\rm osc}\operatorname{osc}_k(U_k,\mathscr{T}_k). \tag{145b}$$

Since oscillation is generically of higher order than the estimator, the issue at stake is whether elements added by oscillation, even though immaterial relative to the error, could ruin the optimal cardinality observed in experiments. If $\mathscr{E}_k(U_k,\mathscr{T}_k)$ has large indicators in a small area, then Dörfler marking for the estimator (145a) could select a set $\mathscr{M}_k$ with a small number of elements relative to $\mathscr{T}_k$. However, if $\operatorname{osc}_k(U_k,\mathscr{T}_k)$ were globally distributed in $\mathscr{T}_k$, then separate marking would require additional marking of a large percentage of all elements to satisfy (145b); i.e., $\#\mathscr{M}_k$ could be large relative to $\#\mathscr{T}_k$.

To explore this idea computationally, we consider a simple modification of the Example of Sect. 8.4 with exact solution that we denote hereafter by u_R. Let u_S be the smooth solution of (141), which is of comparable magnitude with u_R, while the corresponding $f=-\operatorname{div}\boldsymbol{A}\nabla u_S$ exhibits an increasing amount of data oscillation away from the origin. Let $u=u_R+u_S$ be the modified exact solution and let f be the corresponding forcing function. Procedure MARK takes the usual value of $\theta_{\rm est}=0.5$ [32, 52, 53, 63], and procedure REFINE subdivides all elements in $\mathscr{M}_k$ by using two bisections.

The behavior of separate marking for several values of $\theta_{\rm osc}$ is depicted in Figure 20. We can visualize its sensitivity with respect to parameter $\theta_{\rm osc}$. For values of $\theta_{\rm osc}\leq 0.4$ the rate of convergence appears to be quasi-optimal. However, beyond this threshold the curves for both the error and the estimator flatten out, thereby indicating a lack of optimality. The threshold value $\theta_{\rm osc}=0.4$, even though consistent with practice, is tricky to find in general since it is problem-dependent. Therefore, marking by oscillation (145b) is questionable.

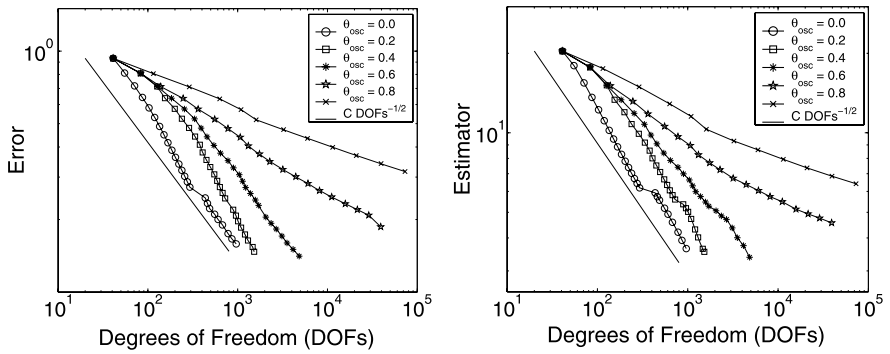


Fig. 20 Decay of the error (left) and the estimator (right) vs. degrees of freedom for $\theta_{\mathrm{est}} = 0.5$ and values $\theta_{\mathrm{osc}} = 0.0, 0.2, 0.4, 0.6$, and 0.8. For values of $\theta_{\mathrm{osc}} \leq 0.4$ the rate of convergence is quasi-optimal, but for $\theta_{\mathrm{osc}} > 0.4$ the curves flatten out, thereby indicating lack of optimality

9.4.2 Analysis of Separate Marking

In order to gain mathematical insight on the key issues related to separate marking, we examine the adaptive approximation of two given functions in an idealized scenario. We show that separate marking, similar to (145), may lead to suboptimal meshes in general. However, a suitable choice of marking parameters may restore optimality. The numerical experiments of Sect. 9.4.1 confirm this theoretical insight in a realistic environment.

For the discussion, we assume that we have two functions u_i, $i = 1, 2$, and have access to their local approximation error

$$e_{\mathscr{T}}(u_i; T) = |u_i - I_{\mathscr{T}} u_i|_{i;T} \qquad \forall T \in \mathscr{T}$$

and global error $e^2_{\mathscr{T}}(u_i) = \sum_{T \in \mathscr{T}} e^2_{\mathscr{T}}(u_i; T)$; hereafter $|\cdot|_i$ are unspecified norms, and $I_{\mathscr{T}}$ is a local interpolation operator over $\mathscr{T} \in \mathbb{T}$. We define the *total error* to be

$$e^2_{\mathscr{T}} := e^2_{\mathscr{T}}(u_1) + e^2_{\mathscr{T}}(u_2)$$

and are interested in its asymptotic decay. If $\mathscr{T} = \mathscr{T}_k$, then we denote $e_k = e_{\mathscr{T}_k}$.

To explore the use of (145), we examine the effect of separate marking for $e_k(u_i)$ on a sequence of meshes $\mathscr{T}^i$ for $i = 1, 2$. We put ourselves in an idealized, but plausible, situation governed by the following three simplifying assumptions:

Independence: $\mathscr{T}^1_k$ and $\mathscr{T}^2_k$ are generated from $\mathscr{T}_0$ and are independent of each other; (146a)

Marking: Separate Dörfler marking with parameters $\theta_i \in (0,1)$ implies that $e_k(u_i) \approx \alpha_i^k$ on $\mathscr{T}^i_k$, with $\alpha_i \in (0,1)$; (146b)

Approximability: $e_k(u_i) \approx (\#\mathscr{T}^i_k - \#\mathscr{T}_0)^{-s_i}$, with $s_1 \leq s_2$ maximal. (146c)

We are interested in the decay of the total error e_k on the overlay $\mathscr{T}_k := \mathscr{T}_k^1 \oplus \mathscr{T}_k^2$. This scenario is a simplification of the more realistic approximation of u_1 and u_2 with separate Dörfler marking on the same sequence of grids $\mathscr{T}_k$ but avoids the complicated interaction of the two marking procedures.

Lemma 9.7 (Separate Marking). *Let assumptions (146) be satisfied. Then the decay of the total error e_k on the overlay $\mathscr{T}_k = \mathscr{T}_k^1 \oplus \mathscr{T}_k^2$ for separate marking is always suboptimal except when α_1 and α_2 satisfy*

$$\alpha_2 \le \alpha_1 \le \alpha_2^{s_1/s_2}.$$

Proof. 1 Assumption (146b) on the average reduction rate implies for the total error that

$$e_k \approx e_k(u_1) + e_k(u_2) \approx \max\{e_k(u_1), e_k(u_2)\} \approx \max\{\alpha_1^k, \alpha_2^k\}. \tag{147}$$

Combining (146b) and (146c) yields $\alpha_i^k \approx (\#\mathscr{T}_k^i - \#\mathscr{T}_0)^{-s_i}$, whence

$$\#\mathscr{T}_k^1 - \#\mathscr{T}_0 \approx \alpha_1^{-k/s_1} = \beta^k \alpha_2^{-k/s_2} \approx \beta^k (\#\mathscr{T}_k^2 - \#\mathscr{T}_0), \tag{148}$$

with $\beta = \alpha_1^{-1/s_1} \alpha_2^{1/s_2}$. In view of Lemma 4.3, this gives for the overlay $\mathscr{T}_k$

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \approx \begin{cases} \#\mathscr{T}_k^1 - \#\mathscr{T}_0, & \beta \ge 1, \\ \#\mathscr{T}_k^2 - \#\mathscr{T}_0, & \beta < 1. \end{cases} \tag{149}$$

The optimal decay of total error e_k corresponds to $e_k \approx (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}$ because $s_1 \le s_2$. In analyzing the relation of e_k to the number of elements $\#\mathscr{T}_k$ in the overlay $\mathscr{T}_k$, we distinguish three cases and employ (147), (148), and (149).

2 *Case*: $\alpha_1 < \alpha_2$. We note that $\alpha_1 < \alpha_2$ and $s_1 \le s_2$ yields $\beta \ge 1$. We thus deduce

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_2^k = (\alpha_2/\alpha_1)^k \alpha_1^k \\ &\approx (\alpha_2/\alpha_1)^k (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \approx (\alpha_2/\alpha_1)^k (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}. \end{aligned}$$

Since $\alpha_2/\alpha_1 > 1$, the approximation of e_k on $\mathscr{T}_k$ is suboptimal.

3 *Case*: $\alpha_1 \ge \alpha_2$ and $\beta < 1$. We obtain

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \\ &\approx \beta^{-ks_1} (\#\mathscr{T}_k^2 - \#\mathscr{T}_0)^{-s_1} \approx \beta^{-ks_1} (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}, \end{aligned}$$

whence the approximation of the total error on $\mathscr{T}_k$ is again suboptimal.

4 *Case*: $\alpha_1 \ge \alpha_2$ and $\beta \ge 1$. We infer that

$$e_k \approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \approx (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}$$

and that $\mathscr{T}_k$ exhibits optimal cardinality. This exceptional case corresponds to the assertion and concludes the proof. □

We learn from Lemma 9.7 that separate marking requires a critical choice of parameters θ_i to retain optimal error decay with respect to the total error e_k. In light of Lemma 9.7, we could identify the AFEM estimator $\mathscr{E}_k$ with the error $e_k(u_1)$ and the AFEM oscillation osc_k with the error $e_k(u_2)$. We observe that $\operatorname{osc}_k \leq \mathscr{E}_k$ combined with (146b) implies that $\alpha_2 \leq \alpha_1$ and that osc_k is generically of higher order than $\mathscr{E}_k$, thereby yielding $s_1 < s_2$.

We wonder whether or not the optimality condition $\alpha_1 \leq \alpha_2^{s_1/s_2}$ is valid. Note that $\alpha_2^{s_1/s_2}$ increases as the gap between s_1 and s_2 increases. Since the oscillation reduction estimate of [52] reveals that α_2 increases as θ_{osc} decreases, we see that separate marking may be optimal for a wide range of marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$; this is confirmed by the numerical experiments in Sect. 9.4.1 even though it is unclear whether $\mathscr{E}_k$ and osc_k satisfy (146). However, choosing marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ is rather tricky in practice because neither the explicit dependence of average reduction rates α_1, α_2 on $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ nor the optimal exponents s_1, s_2 are known. In contrast to [52, 53, 48], the standard AFEM of Chap. 8 marks solely according to the estimator $\mathscr{E}_k(U_k, \mathscr{T}_k)$ and thus avoids separate marking.

9.5 Problems

Problem 9.1. Show that $(v, f, \boldsymbol{A}) \in \mathbb{A}_s$ if and only there exists a constant $\Lambda > 0$ such that for all $\varepsilon > 0$ there exist $\mathscr{T}_\varepsilon \geq \mathscr{T}_0$ conforming and $V_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2 \leq \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \leq \Lambda^{1/s}\, \varepsilon^{-1/s};$$

in this case $|v, f, \boldsymbol{A}|_s \leq \Lambda$. Hint: Let $\mathscr{T}_\varepsilon$ be minimal for $|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2 \leq \varepsilon^2$. This means that for all $\mathscr{T} \in \mathbb{T}$ such that $\#\mathscr{T} = \#\mathscr{T}_\varepsilon - 1$ we have $|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2 > \varepsilon$.

Problem 9.2. Prove Lemma 9.4: if $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ satisfy $\mathscr{T}_* \geq \mathscr{T}$, $\mathscr{R} := \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ is the refined set to go from $\mathscr{T}$ to $\mathscr{T}_*$, and $U \in \mathbb{V}$, $U_* \in \mathbb{V}_*$ are the corresponding Galerkin solutions, then

$$|||U_* - U|||_\Omega^2 \leq C_1\, \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{R}).$$

To this end, write the equation fulfilled by $U_* - U \in \mathbb{V}_*$ and use as a test function the local quasi-interpolant $I_{\mathscr{T}}(U_* - U)$ of $U_* - U$ introduced in Proposition 5.1(b) and compare with Remark 5.5.

Problem 9.3. Trace the dependence as $\theta \to \theta_*$ and $s \to 0$ in Lemma 9.6 and Theorem 9.1.

Problem 9.4. Let $d = 2$ and $n = 1$. Let f be piecewise W_1^1 over the initial mesh $\mathscr{T}_0$, namely $f \in W_1^1(\Omega; \mathscr{T}_0)$. Show that

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \|h_{\mathscr{T}}(f-P_0 f)\|_{L^2(\Omega)} \lesssim \|f\|_{W_1^1(\Omega;\mathscr{T}_0)} N^{-1}.$$

This shows the same decay rate of data oscillation as in Remark 9.1 but with weaker regularity.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Wiley (2000)
2. Arnold, D.N., Mukherjee, A., Pouly, L.: Locally adapted tetrahedral meshes using bisection. SIAM J. Sci. Comput. **22**(2), 431–448 (2000)
3. Atalay, F.B., Mount, D.M.: The cost of compatible refinement of simplex decomposition trees. In: Proc. International Meshing Roundtable 2006 (IMR 2006), pp. 57–69 (2006). Birmingham, AL
4. Babuška, I., Kellogg, R.B., Pitkäranta, J.: Direct and inverse error estimates for finite elements with mesh refinements. Numer. Math. **33**(4), 447–471 (1979)
5. Babuška, I., Rheinboldt, W.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**, 736–754 (1978)
6. Babuška, I., Strouboulis, T.: The finite element method and its reliability. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York (2001)
7. Babuška, I., Vogelius, M.: Feedback and adaptive finite element solution of one-dimensional boundary value problems. Numer. Math. **44**, 75–102 (1984)
8. Babuška, I.: Error-bounds for finite element method. Numer. Math. **16**, 322–333 (1971)
9. Babuška, I., Aziz, A.K.: Survey lectures on the mathematical foundations of the finite element method. with the collaboration of G. Fix and R.B. Kellogg. Math. Found. Finite Elem. Method Appl. Part. Differ. Equations, Sympos. Univ. Maryland, Baltimore 1972, 1-359 (1972). (1972)
10. Bänsch, E.: Local mesh refinement in 2 and 3 dimensions. IMPACT Comput. Sci. Engrg. **3**, 181–191 (1991)
11. Bebendorf, M.: A note on the Poincaré inequality for convex domains. Z. Anal. Anwendungen **22**(4), 751–756 (2003)
12. Beck, R., Hiptmair, R., Hoppe, R.H., Wohlmuth, B.: Residual based a posteriori error estimators for eddy current computation. Math. Model. Numer. Anal. **34**(1), 159–182 (2000)
13. Binev, P., Dahmen, W., DeVore, R.: Adaptive finite element methods with convergence rates. Numer. Math. **97**, 219–268 (2004)
14. Binev, P., Dahmen, W., DeVore, R., Petrushev, P.: Approximation classes for adaptive methods. Serdica Math. J. **28**(4), 391–416 (2002). Dedicated to the memory of Vassil Popov on the occasion of his 60th birthday
15. Braess, D.: Finite Elements. Theory, fast solvers, and applications in solid mechanics, 2nd edition. Cambridge University Press (2001)
16. Brenner, S., Scott, R.: The Mathematical Theory of Finite Element Methods. Springer Texts in Applied Mathematics 15 (2008)
17. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from lagrange multipliers. R.A.I.R.O. Anal. Numer. **R2**, T 129–151 (1974)
18. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer Series in Computational Mathematics 15 (1991)
19. Carroll, R., Duff, G., Friberg, J., Gobert, J., Grisvard, P., Nečas, J., Seeley, R.: Equations aux dérivées partielles. No. 19 in Seminaire de mathematiques superieures. Les Presses de l'Université de Montréal (1966)
20. Carstensen, C., Funken, S.A.: Fully reliable localized error control in the FEM. SIAM J. Sci. Comput. **21**(4), 1465–1484 (electronic) (1999/00)

21. Cascón, J.M., Kreuzer, C., Nochetto, R.H., Siebert, K.G.: Quasi-optimal convergence rate for an adaptive finite element method. Preprint 009/2007, Universität Augsburg (2007)
22. Cea, J.: Approximation variationnelle des problèmes aux limites. Ann. Inst. Fourier **14**(2), 345–444 (1964)
23. Chen, L., Nochetto, R.H., Xu, J.: Adaptive multilevel methods on graded bisection grids. to appear (2009)
24. Chen, Z., Feng, J.: An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems. Math. Comp. **73**, 1167–1042 (2006)
25. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, 40, SIAM (2002)
26. Clément, P.: Approximation by finite element functions using local regularization. R.A.I.R.O. **9**, 77–84 (1975)
27. Dahlke, S., DeVore, R.A.: Besov regularity for elliptic boundary value problems. Commun. Partial Differ. Equations **22**(1-2), 1–16 (1997)
28. DeVore, R.A.: Nonlinear approximation. In: A. Iserles (ed.) Acta Numerica, vol. 7, pp. 51–150. Cambridge University Press (1998)
29. DeVore, R.A., Lorentz, G.G. Constructive approximation, *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. 303. Springer-Verlag, Berlin (1993)
30. DeVore, R.A., Popov, V.A.: Interpolation of Besov spaces. Trans. Amer. Math. Soc. **305**(1), 397–414 (1988)
31. Diening, L., Kreuzer, C.: Convergence of an adaptive finite element method for the p-Laplacian equation. SIAM J. Numer. Anal. **46**(2), 614–638 (2008)
32. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
33. Dupont, T., Scott, R.: Polynomial approximation of functions in Sobolev spaces. Math. Comp. **34**(150), 441–463 (1980)
34. Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems I: A linear model problem. SIAM J. Numer. Anal. **28**, 43–77 (1991)
35. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics. 19. Providence, AMS (1998)
36. Galdi, G.P.: An introduction to the mathematical theory of the Navier-Stokes equations. Vol. 1: Linearized steady problems. Springer Tracts in Natural Philosophy, 38 (1994)
37. Gaspoz, F., Morin, P.: Approximation classes for adaptive higher order finite element approximation. (in preparation) (2009)
38. Gilbarg, D., Trudinger, N.S.: Elliptic partial differential equations of second order. Classics in Mathematics, Springer (2001)
39. Grisvard, P.: Elliptic problems in nonsmooth domains, *Monographs and Studies in Mathematics*, vol. 24. Pitman (Advanced Publishing Program), Boston, MA (1985)
40. Hintermüller, M., Hoppe, R.H., Iliash, Y., Kieweg, M.: An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. ESAIM, Control Optim. Calc. Var. **14**(3), 540–560 (2008)
41. Jarausch, H.: On an adaptive grid refining technique for finite element approximations. SIAM J. Sci. Stat. Comput. **7**, 1105–1120 (1986)
42. Kato, T.: Estimation of iterated matrices, with application to the von Neumann condition. Numer. Math. **2**, 22–29 (1960)
43. Kellogg, R.B.: On the Poisson equation with intersecting interfaces. Applicable Anal. **4**, 101–129 (1974/75). Collection of articles dedicated to Nikolai Ivanovich Muskhelishvili
44. Kossaczký, I.: A recursive approach to local mesh refinement in two and three dimensions. J. Comput. Appl. Math. **55**, 275–288 (1994)
45. Lax, P., Milgram, A.: Parabolic equations. Ann. Math. Stud. **33**, 167–190 (1954)
46. Liu, A., Joe, B.: Quality local refinement of tetrahedral meshes based on bisection. SIAM J. Sci. Comput. **16**, 1269–1291 (1995)
47. Maubach, J.M.: Local bisection refinement for n-simplicial grids generated by reflection. SIAM J. Sci. Comput. **16**, 210–227 (1995)

48. Mekchay, K., Nochetto, R.H.: Convergence of adaptive finite element methods for general second order linear elliptic PDEs. SIAM J. Numer. Anal. **43**(5), 1803–1827 (2005)
49. Mitchell, W.F.: Unified multilevel adaptive finite element methods for elliptic problems. Ph.D. thesis, Department of Computer Science, University of Illinois, Urbana (1988)
50. Mitchell, W.F.: A comparison of adaptive refinement techniques for elliptic problems. ACM Trans. Math. Softw. **15**, 326–347 (1989)
51. Monk, P.: Finite element methods for Maxwell's equations. Numerical Mathematics and Scientific Computation. Oxford University Press. xiv, 450 p. (2003)
52. Morin, P., Nochetto, R.H., Siebert, K.G.: Data oscillation and convergence of adaptive FEM. SIAM J. Numer. Anal. **38**, 466–488 (2000)
53. Morin, P., Nochetto, R.H., Siebert, K.G.: Convergence of adaptive finite element methods. SIAM Review **44**, 631–658 (2002)
54. Morin, P., Nochetto, R.H., Siebert, K.G.: Local problems on stars: A posteriori error estimators, convergence, and performance. Math. Comp. **72**, 1067–1097 (2003)
55. Morin, P., Siebert, K.G., Veeser, A.: A basic convergence result for conforming adaptive finite elements. Math. Models Methods Appl. 18 (2008) **18**, 707–737 (2008)
56. Necas, J.: Sur une méthode pour resoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. Ann. Sc. Norm. Super. Pisa, Sci. Fis. Mat., III. Ser. **16**, 305–326 (1962)
57. Nochetto, R.H., Paolini, M., Verdi, C.: An adaptive finite element method for two-phase Stefan problems in two space dimensions. I. Stability and error estimates. Math. Comp. **57**(195), 73–108 (1991), S1–S11
58. Oswald, P.: On function spaces related to finite element approximation theory. Z. Anal. Anwendungen **9**(1), 43–64 (1990)
59. Otto, F.: On the Babuška–Brezzi condition for the Taylor–Hood element. Diploma thesis Universität Bonn (1990). In German
60. Payne, L.E., Weinberger, H.F.: An optimal Poincaré-inequality for convex domains. Archive Rat. Mech. Anal. **5**, 286–292 (1960)
61. Rivara, M.C.: Mesh refinement processes based on the generalized bisection of simplices. SIAM J. Numer. Anal. **21**(3), 604–613 (1984)
62. Sacchi, R., Veeser, A.: Locally efficient and reliable a posteriori error estimators for Dirichlet problems. Math. Models Methods Appl. Sci. **16**(3), 319–346 (2006)
63. Schmidt, A., Siebert, K.G.: Design of adaptive finite element software. The finite element toolbox ALBERTA. Lecture Notes in Computational Science and Engineering 42, Springer (2005)
64. Schöberl, J.: A posteriori error estimates for Maxwell equations. Math. Comp. **77**(262), 633–649 (2008)
65. Scott, L.R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Mathematics of Computation **54**(190), 483–493 (1990)
66. Sewell, E.G.: Automatic generation of triangulations for piecewise polynomial approximation. Ph.D. dissertation, Purdue Univ., West Lafayette, Ind., 1972
67. Siebert, K.G.: A convergence proof for adaptive finite elements without lower bound (2009). Preprint Universität Duisburg-Essen and Universität Freiburg No. 1/2009
68. Siebert, K.G., Veeser, A.: A unilaterally constrained quadratic minimization with adaptive finite elements. SIAM J. Optim. **18**(1), 260–289 (2007)
69. Stevenson, R.: Optimality of a standard adaptive finite element method. Found. Comput. Math. **7**(2), 245–269 (2007)
70. Stevenson, R.: The completion of locally refined simplicial partitions created by bisection. Math. Comput. **77**(261), 227–241 (2008)
71. Storoženko, È.A., Oswald, P.: Jackson's theorem in the spaces $L^p(\mathbf{R}^k)$, $0 < p < 1$. Sibirsk. Mat. Ž. **19**(4), 888–901, 956 (1978)
72. Traxler, C.T.: An algorithm for adaptive mesh refinement in n dimensions. Computing **59**(2), 115–137 (1997)
73. Veeser, A.: Convergent adaptive finite elements for the nonlinear Laplacian. Numer. Math. **92**(4), 743–770 (2002)

74. Veeser, A., Verfürth, R.: Explicit upper bounds for dual norms of residuals. SIAM J. Numer. Anal. (to appear)
75. Verfürth, R.: A posteriori error estimators for the Stokes equations. Numer. Math. **55**, 309–325 (1989)
76. Verfürth, R.: A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Adv. Numer. Math. John Wiley, Chichester, UK (1996)
77. Wu, H., Chen, Z.: Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems. Sci. China Ser. A **49**(10), 1405–1429 (2006)
78. Xu, J., Chen, L., Nochetto, R.H.: Optimal multilevel methods for $H(\mathrm{grad})$, $H(\mathrm{curl})$ and $H(\mathrm{div})$ systems on graded and unstructured grids, R.A. DeVore and A. Kunoth eds., *Multiscale, Nonlinear and Adaptive Approximation*, pp. 599–659. Springer 2009.
79. Xu, J., Zikatanov, L.: Some observations on Babuška and Brezzi theories. Numer. Math. **94**(1), 195–202 (2003)

Adaptive wavelet methods for solving operator equations: An overview

Rob Stevenson

Abstract In [*Math. Comp*, 70 (2001), 27–75] and [*Found. Comput. Math.*, 2(3) (2002), 203–245], Cohen, Dahmen and DeVore introduced adaptive wavelet methods for solving operator equations. These papers meant a break-through in the field, because their adaptive methods were not only proven to converge, but also with a rate better than that of their non-adaptive counterparts in cases where the latter methods converge with a reduced rate due a lacking regularity of the solution. Until then, adaptive methods were usually assumed to converge via a saturation assumption. An exception was given by the work of Dörfler in [*SIAM J. Numer. Anal.*, 33 (1996), 1106–1124], where an adaptive finite element method was proven to converge, with no rate though.

This work contains a complete analysis of the methods from the aforementioned two papers of Cohen, Dahmen and DeVore. Furthermore, we give an overview over the subsequent developments in the field of adaptive wavelet methods. This includes a precise analysis of the near-sparsity of an operator in wavelet coordinates needed to obtain optimal computational complexity; the avoidance of coarsening; quantitative improvements of the algorithms; their generalization to frames; and their application with tensor product wavelet bases which give dimension independent rates.

1 Introduction

1.1 Non-adaptive methods

In this survey, we discuss optimally converging adaptive wavelet methods for solving well-posed linear operator equations

Rob Stevenson
Korteweg - de Vries (KdV) Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands, e-mail: R.P.Stevenson@uva.nl

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_13,

$$Bu = f.$$

Such methods were introduced by Cohen, Dahmen and DeVore in[CDD01, CDD02]. For wavelet methods in general for solving operator equations, we refer to [Dah97, Coh03, Urb09].

We assume that B is a *boundedly invertible linear operator* between $\mathscr{X}$ and $\mathscr{Y}'$, where $\mathscr{X}$ and $\mathscr{Y}$ are Hilbert spaces. As typical examples, we have in mind linear *partial differential* or *singular integral* equations, in which case $\mathscr{X}$ and $\mathscr{Y}$ are Sobolev spaces or, for non-scalar equations, products of such spaces. We assume that we have Riesz bases $\Psi^{\mathscr{X}} = \{\psi_\lambda^{\mathscr{X}} : \lambda \in \nabla\}$ and $\Psi^{\mathscr{Y}} = \{\psi_\lambda^{\mathscr{Y}} : \lambda \in \nabla\}$ for $\mathscr{X}$ and $\mathscr{Y}$ available, which are of *wavelet type*. In most applications, $\mathscr{X}$ and $\mathscr{Y}$ and $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{Y}}$ will be equal.

The adaptive wavelet methodology has been extended to *non-linear* problems, see [CDD03a, DSX00a, CDD03b, CU05, BDS07, BU08]. Such problems, however, will not be discussed in this paper.

A standard *non-adaptive* numerical (wavelet) method for solving $Bu = f$ consists of selecting a Λ from a fixed sequence $\Lambda_0 \subset \Lambda_1 \subset \cdots$ with $\cup_i \Lambda_i = \nabla$, and computing a (quasi-) best approximation u_Λ to u from $\operatorname{span}\{\psi_\lambda^{\mathscr{X}} : \lambda \in \Lambda\}$. The standard choice for Λ_i is the set of all wavelet indices λ with "level" up to i, so that $\operatorname{span}\{\psi_\lambda^{\mathscr{X}} : \lambda \in \Lambda_i\}$ is equal to the span of all "scaling functions" on level i. The counterpart of this wavelet method in a *finite element setting* is the computation of the finite element approximation with respect to an i times *uniformly refined initial mesh*.

Associated to $\mathscr{X}$ and $\Psi^{\mathscr{X}}$, there exists a parameter

$$s_{\max} > 0$$

such that for a suitable choice of $(\Lambda_i)_i$, for *all* $u \in \mathscr{X}$ that are *sufficiently smooth*

$$\|u - u_{\Lambda_i}\|_{\mathscr{X}} \lesssim (\#\Lambda_i)^{-s_{\max}},$$

where this rate $s_{\max}$ *cannot be improved by imposing additional smoothness conditions or by another selection of* $(\Lambda_i)_i$.

For completeness, here and in the remainder of this work, with $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters on which C and D may depend. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

Remark 1.1. There exist $u \in \mathscr{X}$ for which a rate better than $s_{\max}$ can be realized. Indeed, if u happens to have a finite representation in $\Psi^{\mathscr{X}}$, or if it is exceptionally close to such a function, then with a suitable choice of $(\Lambda_i)_i$ any rate can be realized. Since such cases are exceptional, we may ignore them in the further considerations.

Typically, the parameter $s_{\max}$ is a function of the *order* of the wavelet basis $\Psi^{\mathscr{X}}$, the order of smoothness that is measured in the (Sobolev) space $\mathscr{X}$, and the dimension of the underlying domain.

Example 1.1. Let $\mathcal{X} = H^m(\Omega)$ where Ω is a bounded domain in $\mathbb{R}^n$, and let $\Psi^{\mathcal{X}}$ be a standard wavelet basis of order $d > m$. Then

$$s_{\max} = \frac{d-m}{n},$$

and with Λ_i being the set of all wavelet indices λ with levels up to i, this rate $s_{\max}$ is realized for $u \in H^d(\Omega)$. More generally, for $s \in (0, s_{\max}]$ and $u \in H^{sn+m}(\Omega)$, a rate s is realized. This result is sharp in the sense that for $\varepsilon > 0$, there exists no choice $(\Lambda_i)_i$ such that the rate s is realized for all $u \in H^{sn+m-\varepsilon}(\Omega)$.

1.2 Adaptive methods

Even for smooth right-hand sides f, in many applications the smoothness conditions on u to realize the optimal rate $s_{\max}$ with the standard choice of $(\Lambda_i)_i$ are not fulfilled. Typical examples are boundary value problems on non-smooth domains, where corners, edges etc. induce singularities in the solution. For simple model examples, the precise knowledge of these singularities enables one to select a sequence $(\Lambda_i)_i$ such that the optimal rate $s_{\max}$ is retrieved, assuming f is sufficiently smooth. Such a sequence $(\Lambda_i)_i$ involves local refinements towards the boundary, i.e., the addition of extra wavelets with supports near the boundary. For more general problems, however, such an a priori selection of $(\Lambda_i)_i$ is not feasible.

The topic of this work are *adaptive* (wavelet) methods. With these methods, the expansion of Λ_i to Λ_{i+1} is made based on information provided by u_{Λ_i}. In this way, the sequences $(\Lambda_i)_i$ and $(u_{\Lambda_i})_i$ depend (non-linearly) on u.

The method from [CDD01] is similar to an adaptive finite element method in the sense that information from an a posteriori error estimator is used to guide the expansion of Λ_i to Λ_{i+1} such that the error is reduced, at the expense of a (quasi-) minimal increase in the cardinality.

The idea behind the method from [CDD02] is the application of *some* iterative method to construct $(\Lambda_i)_i$ such that $(u_{\Lambda_i})_i$ converges (linearly) to u. Here a (quasi-) optimal balance between support sizes and accuracy is realized by, after each fixed number of iterations, removing small coefficients from the current approximation, a process known as *coarsening*.

The key to the development of adaptive wavelet methods is the fact that for a large class of operators B, its bi-infinite stiffness or system matrix with respect to suitable wavelet bases is close to a sparse matrix. Here suitable means that the wavelets are *sufficiently smooth* and have sufficiently many *vanishing moments*. Thanks to this near-sparsity, given an approximation $\tilde{\mathbf{u}} \in \ell_0$ to $\mathbf{u}$, its generally infinitely supported residual can be accurately approximated at relatively low cost. This fact allows to run an iterative scheme to the bi-infinite matrix vector equation, in which residuals are computed approximately, essentially being the scheme from [CDD02], or to use the approximate residual as an a posteriori error estimator as in the scheme proposed in [CDD01].

1.3 Best N-term approximation and approximation classes

As a benchmark for these adaptive methods, we consider a (quasi-) *best possible choice* of $(\Lambda_i)_i$ *depending on* u, where we assume to have full knowledge of this function, and thus of its expansion in the wavelet basis $\Psi^{\mathscr{X}}$. Given $u = \mathbf{u}^\top \Psi^{\mathscr{X}} := \sum_{\lambda\in\nabla} \mathbf{u}_\lambda \psi_\lambda^{\mathscr{X}}$ and an approximation $v = \mathbf{v}^\top \Psi^{\mathscr{X}}$, because $\Psi^{\mathscr{X}}$ is a Riesz basis, it holds that

$$\|u-v\|_{\mathscr{X}} \eqsim \|\mathbf{u}-\mathbf{v}\|, \tag{1}$$

where $\|\cdot\|$ is the norm on $\ell_2 = \ell_2(\nabla) := \{\mathbf{v} : \nabla \to \mathbb{R} : \sum_{\lambda\in\nabla} |v_\lambda|^2 < \infty\}$. The subspace of finitely supported $\mathbf{v} \in \ell_2$ will be denoted as ℓ_0. As a consequence of (1), given a budget $N \in \mathbb{N}$, a (quasi-) best choice for an approximation $v = \mathbf{v}^\top \Psi^{\mathscr{X}} \in \mathscr{X}$ with $\#\operatorname{supp}\mathbf{v} \le N$ is to take $\mathbf{v}$ to be a *best N-term approximation* to $\mathbf{u}$, i.e., a vector with at most N non-zero coefficients that has ℓ_2-distance to $\mathbf{u}$ not larger than any vector with support length $\le N$. Obviously, such a best N-term approximation to $\mathbf{u}$, denoted as $\mathbf{u}_N$, coincides with $\mathbf{u}$ on those N positions where $\mathbf{u}$ has its N largest coefficients in modulus, and is zero elsewhere. Note that $\mathbf{u}_N$ is not necessarily unique.

All $\mathbf{u}$ whose best N-term approximations converge with rate $s > 0$ are collected in the approximation class

$$\mathscr{A}^s(=\mathscr{A}^s_\infty) := \{\mathbf{u} \in \ell_2 : \|\mathbf{u}\|_{\mathscr{A}^s} := \sup_{\varepsilon>0} \varepsilon \times [\min\{N \in \mathbb{N}_0 : \|\mathbf{u}-\mathbf{u}_N\|_{\ell_2} \le \varepsilon\}]^s < \infty\}. \tag{2}$$

Indeed, one may verify that $\|\mathbf{u}\|_{\mathscr{A}^s} \eqsim \sup_{N\in\mathbb{N}_0} (N+1)^s \|\mathbf{u}-\mathbf{u}_N\|$, being the commonly used definition of the (quasi-) norm on $\mathscr{A}^s$. Given $\mathbf{u} \in \mathscr{A}^s$ and $\varepsilon > 0$, the smallest N such that $\|\mathbf{u}-\mathbf{u}_N\| \le \varepsilon$ satisfies

$$N \le \varepsilon^{-1/s} \|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}, \tag{3}$$

which bound is generally sharp. Since for $\varepsilon < \|\mathbf{u}\|$, the value of N in the definition of $\|\mathbf{u}\|_{\mathscr{A}^s}$ is positive, furthermore note that

$$\|\mathbf{u}\|_{\mathscr{A}^s} \ge \sup_{0<\varepsilon<\|\mathbf{u}\|} \varepsilon = \|\mathbf{u}\|.$$

As discussed in Remark 1.1, for $s > s_{\max}$, the class $\mathscr{A}^s$, although not empty, is not relevant. For any $s \in (0, s_{\max}]$, the class $\mathscr{A}^s$ is much larger than the class of (representations of) functions that can be approximated with rate s for any fixed choice of $(\Lambda_i)_i$.

Example 1.2. In the situation of Example 1.1, with wavelets that are sufficiently smooth, for $s \in (0, \frac{d-m}{n})$ and with $\tau := (\frac{1}{2}+s)^{-1}$, (representations of) all functions in the Besov space $B^{sn+t}_\tau(L_\tau(\Omega))$ are contained in $\mathscr{A}^s$. Coarsely speaking, $B^{sn+t}_\tau(L_\tau(\Omega))$ is the space of all functions having $sn+t$ orders of smoothness in $L_\tau(\Omega)$, which space, since $\tau < 2$, is thus (much) larger than $H^{sn+t}(\Omega)$ with an increasing difference with growing s. For details about the relation between approximation classes and Besov spaces we refer to [DeV98, Coh03]. For several boundary

value problems, assuming a sufficiently smooth right-hand side f, it has been proved that the solution u has a much higher regularity in this scale of Besov spaces than in the scale of Sobolev spaces $(H^{sn+m}(\Omega))_s$, see [DD97, Dah99].

In view of the definition of $\mathscr{A}^s$, in particular (3), we will call an *adaptive wavelet method to be (quasi-) optimal* if

> *whenever u has a representation $u = \mathbf{u}^\top \Psi^{\mathscr{X}}$ with $\mathbf{u} \in \mathscr{A}^s$ for some $s \in (0, s_{\max}]$, then given a tolerance $\varepsilon > 0$, it produces an approximation $\mathbf{v} \in \ell_0$ with $\|\mathbf{u} - \mathbf{v}\| \le \varepsilon$ and $\#\operatorname{supp}\mathbf{v} \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, at the cost of a number of arithmetic operations that can be bounded by some absolute multiple of the same expression.*

1.4 Structure of the paper

The remainder of this work is organized as follows: In Sect. 2, we reformulate well-posed operator equations as bi-infinite matrix vector equations, and give some typical examples of such operator equations.

In Sect. 3 and 4, we define the adaptive wavelet schemes from [CDD02] and [CDD01], respectively, and prove their (quasi-) optimality. Note that we reverse the order in which these schemes were proposed.

The analysis from Sect. 3 and 4 applies under the assumption that the operator B in wavelet coordinates can be sufficiently well approximated by sparse matrices that are computable in linear complexity. In Sect. 5, we verify this assumption for a class of partial differential operators.

In Sect. 6, we discuss the generalization of the adaptive wavelet approach to the case that instead of Riesz bases we only have frames available. Our motivation will be that on general non-product domains, the construction of (wavelet) frames is easier than that of (wavelet) Riesz bases.

Finally, in Sect. 7, the application of tensor product wavelet bases is discussed. Approximation using such bases does not suffer from the so-called curse of dimensionality. An application of those bases is given by the (quasi-) optimal simultaneously space-time adaptive solution of parabolic initial boundary value problems.

1.5 Some properties of the (quasi-) norms $\|\cdot\|_{\mathscr{A}^s}$

We end this section by recalling two known properties of the $\|\cdot\|_{\mathscr{A}^s}$-norm (cf. e.g. [DeV98]) that will be used in Sect. 3 and 4. In order to keep the presentation in these sections self-contained, we include their proofs.

Lemma 1.1. *For $\mathbf{v} \in \mathscr{A}^s$ and $\mathbf{w} \in \ell_0$,*

$$\|\mathbf{w}\|_{\mathscr{A}^s} \le 2\max(\|\mathbf{v}\|_{\mathscr{A}^s}, (\#\operatorname{supp}\mathbf{w})^s\|\mathbf{v} - \mathbf{w}\|).$$

Proof. For $\varepsilon \in (0, 2\|\mathbf{v}-\mathbf{w}\|]$, the approximation of $\mathbf{w}$ by itself shows that the expression εN^s in the definition of $\|\mathbf{w}\|_{\mathscr{A}^s}$ is bounded by $2\|\mathbf{v}-\mathbf{w}\|(\#\operatorname{supp}\mathbf{w})^s$.

For $\varepsilon \geq 2\|\mathbf{v}-\mathbf{w}\|$, let N be the smallest integer such that $\|\mathbf{v}-\mathbf{v}_N\| \leq \frac{\varepsilon}{2}$. Then $\|\mathbf{w}-\mathbf{v}_N\| \leq \varepsilon$ and $\varepsilon N^s = 2\frac{\varepsilon}{2}N^s \leq 2\|\mathbf{v}\|_{\mathscr{A}^s}$. □

Lemma 1.2. *For $s>0$ and with $\tau := (\frac{1}{2}+s)^{-1}$,*

$$\|\mathbf{v}\|_{\mathscr{A}^s} \eqsim \sup_{\eta>0} \eta \times \#\{\lambda \in \nabla : |v_\lambda| > \eta\}^{1/\tau}, \quad (\mathbf{v} \in \mathscr{A}^s). \tag{4}$$

Proof. Let us denote the expression at the right-hand side of (4) as $|||\mathbf{v}|||_{\mathscr{A}^s}$. Let N be the smallest integer such that the entries of $\mathbf{v}-\mathbf{v}_N$ are in modulus not larger than η. Then $N \leq (|||\mathbf{v}|||_{\mathscr{A}^s}\eta^{-1})^\tau$, and

$$\begin{aligned}\|\mathbf{v}-\mathbf{v}_N\| &\leq \sum_{k=0}^{\infty} 2^{-k}\eta\sqrt{\#\{\lambda \in \nabla : |v_\lambda| \in (2^{-(k+1)}\eta, 2^{-k}\eta]\}} \\ &\leq \sum_{k=0}^{\infty} 2^{-k}\eta(|||\mathbf{v}|||_{\mathscr{A}^s}2^{k+1}\eta^{-1})^{\tau/2} \lesssim \eta^{1-\tau/2}|||\mathbf{v}|||_{\mathscr{A}^s}^{\tau/2},\end{aligned}$$

and so $\|\mathbf{v}\|_{\mathscr{A}^s} \lesssim \sup_{\eta>0}\eta^{1-\tau/2}|||\mathbf{v}|||_{\mathscr{A}^s}^{\tau/2} \times ((|||\mathbf{v}|||_{\mathscr{A}^s}\eta^{-1})^\tau)^s = |||\mathbf{v}|||_{\mathscr{A}^s}$.

To show the other direction, first we note that

$$\|\mathbf{v}-\mathbf{v}_N\| \leq N^{-s}\|\mathbf{v}\|_{\mathscr{A}^s} \quad (\mathbf{v} \in \mathscr{A}^s, N \geq 1). \tag{5}$$

Indeed, if $\|\mathbf{v}-\mathbf{v}_N\| = \|\mathbf{v}-\mathbf{v}_{N-1}\|$, then $\mathbf{v}_N = \mathbf{v}_{N-1} = \mathbf{v}$ and (5) is valid. Otherwise, i.e., when $\|\mathbf{v}-\mathbf{v}_N\| < \|\mathbf{v}-\mathbf{v}_{N-1}\|$, by putting $\varepsilon := \|\mathbf{v}-\mathbf{v}_N\|$, the definition of $\|\cdot\|_{\mathscr{A}^s}$ shows (5).

With $(\gamma_N(\mathbf{v}))_{N\in\mathbb{N}}$ denoting a non-decreasing re-arrangement of $\mathbf{v}$ in modulus, secondly we note that

$$\sup_{N\in\mathbb{N}} N^{1/\tau}|\gamma_N(\mathbf{v})| \lesssim \|\mathbf{v}\|_{\mathscr{A}^s} \quad (\mathbf{v} \in \mathscr{A}^s). \tag{6}$$

Indeed, $|\gamma_1(\mathbf{v})| \leq \|\mathbf{v}\| \leq \|\mathbf{v}\|_{\mathscr{A}^s}$. For $1 \leq k < N$,

$$(N-k)|\gamma_N(\mathbf{v})|^2 \leq \sum_{k<j\leq N}|\gamma_j(\mathbf{v})|^2 \leq \|\mathbf{v}-\mathbf{v}_k\|^2 \leq k^{-2s}\|\mathbf{v}\|_{\mathscr{A}^s}^2,$$

or $|\gamma_N(\mathbf{v})| \leq \min_{1\leq k<N}\frac{k^{-s}}{(N-k)^{\frac{1}{2}}}\|\mathbf{v}\|_{\mathscr{A}^s} \eqsim N^{-1/\tau}\|\mathbf{v}\|_{\mathscr{A}^s}$.

Now given $N \in \mathbb{N}_0$, let $\eta := \gamma_{N+1}(\mathbf{v})$, then $\#\{\lambda \in \nabla : |v_\lambda| > \eta\} = N$. From $\eta \lesssim (N+1)^{-1/\tau}\|\mathbf{v}\|_{\mathscr{A}^s}$, we arrive at $|||\mathbf{v}|||_{\mathscr{A}^s} \lesssim \sup_N (N+1)^{-1/\tau}\|\mathbf{v}\|_{\mathscr{A}^s} \times N^{1/\tau} \leq \|\mathbf{v}\|_{\mathscr{A}^s}$. □

2 Well-posed linear operator equations

2.1 Reformulation as a bi-infinite matrix vector equation

Let $\mathscr{X},\mathscr{Y}$ be separable (infinite dimensional) Hilbert spaces over $\mathbb{R}$ (the complex case does not impose additional difficulties apart from requiring somewhat more complicated notations). Let us assume that we have available a *Riesz basis* $\Psi^{\mathscr{X}} = \{\psi_\lambda^{\mathscr{X}} : \lambda \in \nabla\}$ for $\mathscr{X}$, meaning that the *analysis operator*

$$\mathscr{F}_{\mathscr{X}} : \mathscr{X}' \to \ell_2 : g \mapsto [g(\psi_\lambda^{\mathscr{X}})]_{\lambda\in\nabla},$$

is boundedly invertible. By identifying ℓ_2 with its dual, its adjoint $\mathscr{F}'_{\mathscr{X}}$, known as the *synthesis operator*, and defined by $g(\mathscr{F}'_{\mathscr{X}}\mathbf{c}) = \langle \mathscr{F}_{\mathscr{X}} g, \mathbf{c}\rangle_{\ell_2\times\ell_2}$ ($g \in \mathscr{X}', \mathbf{c} \in \ell_2$), reads as

$$\mathscr{F}'_{\mathscr{X}} : \ell_2 \to \mathscr{X} : \mathbf{c} \mapsto \mathbf{c}^\top \Psi^{\mathscr{X}} := \sum_{\lambda\in\nabla} c_\lambda \psi_\lambda^{\mathscr{X}}.$$

Similarly, let $\Psi^{\mathscr{Y}} = \{\psi_\lambda^{\mathscr{Y}} : \lambda \in \nabla\}$ be a Riesz basis for $\mathscr{Y}$, with analysis operator $\mathscr{F}_{\mathscr{Y}}$ and adjoint $\mathscr{F}'_{\mathscr{Y}}$. For both $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{Y}}$, we have suitable *wavelet* bases in mind. Note that w.l.o.g. we could assume that the index set ∇ is the same for $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{Y}}$.

Now given an $f \in \mathscr{Y}'$, and a boundedly invertible $B \in \mathscr{L}(\mathscr{X},\mathscr{Y}')$, we are interested in solving the *linear operator equation* of finding $u \in \mathscr{X}$ such that

$$Bu = f.$$

Writing $u = s_{\mathscr{X}}\mathbf{u}$, and applying $\mathscr{F}_{\mathscr{Y}}$ to both sides of the equation, we infer that the problem can equivalently be written as the bi-infinite matrix vector problem

$$\mathbf{Bu} = \mathbf{f}, \tag{7}$$

where $\mathbf{f} := \mathscr{F}_{\mathscr{Y}} f = [f(\psi_\lambda^{\mathscr{Y}})]_{\lambda\in\nabla} \in \ell_2$, and the *"stiffness"* or *system matrix*

$$\mathbf{B} := \mathscr{F}_{\mathscr{Y}} B \mathscr{F}'_{\mathscr{X}} = [(B\psi_\mu^{\mathscr{X}})(\psi_\lambda^{\mathscr{Y}})]_{\lambda,\mu\in\nabla} \in \mathscr{L}(\ell_2,\ell_2)$$

is boundedly invertible. With $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{\ell_2\times\ell_2}$, for any $\mathbf{v}, \mathbf{w} \in \ell_2$,

$$\langle \mathbf{Bv}, \mathbf{w}\rangle = \langle \mathscr{F}_{\mathscr{Y}} B \mathscr{F}'_{\mathscr{X}} \mathbf{v}, \mathbf{w}\rangle = (Bv)(w), \tag{8}$$

where $v = \mathscr{F}'_{\mathscr{X}}\mathbf{v}$ and $w = \mathscr{F}'_{\mathscr{Y}}\mathbf{w}$.

With the Riesz constants

$$\Lambda_{\Psi^{\mathscr{X}}} := \|\mathscr{F}_{\mathscr{X}}\|_{\mathscr{X}'\to\ell_2} = \sup_{0\neq g\in\mathscr{X}} \frac{\|\mathscr{F}_{\mathscr{X}} g\|_{\ell_2}}{\|g\|_{\mathscr{X}'}},$$

$$\lambda_{\Psi^{\mathscr{X}}} := \|(\mathscr{F}_{\mathscr{X}})^{-1}\|^{-1}_{\ell_2\to\mathscr{X}'} = \inf_{0\neq g\in\mathscr{X}} \frac{\|\mathscr{F}_{\mathscr{X}} g\|_{\ell_2}}{\|g\|_{\mathscr{X}'}},$$

and $\Lambda_{\mathscr{Y}}$ and $\lambda_{\mathscr{Y}}$ defined analogously, and with $\|\cdot\| := \|\cdot\|_{\ell_2\to\ell_2}$, obviously it holds that

$$\|\mathbf{B}\| \le \|B\|_{\mathscr{X}\to\mathscr{Y}'}\Lambda_{\Psi^{\mathscr{X}}}\Lambda_{\mathscr{Y}}, \tag{9}$$

$$\|\mathbf{B}^{-1}\| \le \frac{\|B^{-1}\|_{\mathscr{Y}'\to\mathscr{X}}}{\lambda_{\Psi^{\mathscr{X}}}\lambda_{\mathscr{Y}}}. \tag{10}$$

Remark 2.1. Although not strictly necessary for the remainder of this paper, we make a few comments about *dual bases*. The collection

$$\Psi^{\mathscr{X}'} = (\mathscr{F}'_{\mathscr{X}}\mathscr{F}_{\mathscr{X}})^{-1}\Psi^{\mathscr{X}}$$

is the Riesz basis for $\mathscr{X}'$ that is dual to $\Psi^{\mathscr{X}}$. Indeed, since the corresponding analysis operator $\mathscr{F}_{\mathscr{X}'}$ reads as $(\mathscr{F}'_{\mathscr{X}})^{-1}$, it is boundedly invertible, and so $\Psi^{\mathscr{X}'}$ is a Riesz basis for $\mathscr{X}'$. It holds that $\mathscr{F}'_{\mathscr{X}'}\mathscr{F}_{\mathscr{X}} = I$, which, since $\Psi^{\mathscr{X}'}$ is a basis, implies that $\psi_\lambda^{\mathscr{X}'}(\psi_\mu^{\mathscr{X}}) = \delta_{\lambda\mu}$.

Given a $\tilde{\nabla}\subset\nabla$, $Q_{\tilde{\nabla}} : \mathscr{X}\to\mathscr{X} : v\mapsto\sum_{\lambda\in\tilde{\nabla}}\psi_\lambda^{\mathscr{X}'}(v)\psi_\lambda^{\mathscr{X}}$ is the *biorthogonal projector* onto $\operatorname{span}\{\psi_\lambda^{\mathscr{X}} : \lambda\in\tilde{\nabla}\}$, i.e., $Q_{\tilde{\nabla}}^2 = Q_{\tilde{\nabla}}$ and $\psi_\lambda^{\mathscr{X}'}$ vanishes on $\operatorname{Im}(\mathrm{Id}-Q_{\tilde{\nabla}})$ for all $\lambda\in\tilde{\nabla}$. We have $\|Q_{\tilde{\nabla}}\|_{\mathscr{X}\to\mathscr{X}} \le \Lambda_{\Psi^{\mathscr{X}}}/\lambda_{\Psi^{\mathscr{X}}}$, and so

$$\|v - Q_{\tilde{\nabla}}v\|_{\mathscr{X}} \le (1+\Lambda_{\Psi^{\mathscr{X}}}/\lambda_{\Psi^{\mathscr{X}}}) \inf_{w\in\operatorname{span}\{\psi_\lambda^{\mathscr{X}}:\lambda\in\tilde{\nabla}\}} \|v-w\|_{\mathscr{X}}.$$

The dual projector $Q'_{\tilde{\nabla}} : \mathscr{X}'\to\mathscr{X}'$, that reads as $Q'_{\tilde{\nabla}}(g) = \sum_{\lambda\in\tilde{\nabla}} g(\psi_\lambda^{\mathscr{X}})\psi_\lambda^{\mathscr{X}'}$, has analogous properties.

If we identify $\mathscr{X}'$ with $\mathscr{X}$ using the Riesz map, then if, using this identification, $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{X}'}$ are equal, then $Q_{\tilde{\nabla}}$, being equal to its adjoint, is the orthogonal projector onto $\operatorname{span}\{\psi_\lambda^{\mathscr{X}} : \lambda\in\tilde{\nabla}\}$.

Obviously, similar observations can be made for the collections $\Psi^{\mathscr{Y}}$ and its dual $\Psi^{\mathscr{Y}'}$.

2.2 Some model examples

We give some examples of partial differential equations or singular integral equations that are of the form $Bu = f$ with $B\in\mathscr{L}(\mathscr{X},\mathscr{Y}')$ boundedly invertible. More examples can be found in [CDD02, DK05].

2.2.1 Second order elliptic boundary value problems

The variational formulation of a second order elliptic boundary value problem on a domain $\Omega\subset\mathbb{R}^n$ with homogeneous Dirichlet boundary conditions reads as $Bu = f$,

where

$$(Bu)(v) := \int_\Omega \mathbf{A}\nabla u \cdot \nabla v + (\mathbf{b}\cdot\nabla u)v + cuv\,dx.$$

If $\mathbf{A} \in L_\infty(\Omega)^{n\times n}$, $\mathbf{b} \in L_\infty(\Omega)^n$, $c \in L_\infty(\Omega)$, $c \geq 0$ (*a.e.*), $\nabla\cdot\mathbf{b} = 0$ (*a.e.*) and, for some $\delta > 0$, $\mathbf{A} \geq \delta > 0$ (*a.e.*), then $(Bv)(v) \geq \delta|v|^2_{H^1(\Omega)} \gtrsim \|v\|^2_{H^1(\Omega)}$ $(v \in H^1_0(\Omega))$, i.e., B is *coercive*. The Lax-Milgram lemma now shows that with $\mathscr{X} := H^1_0(\Omega)$, $B : \mathscr{X} \to \mathscr{X}'$ is boundedly invertible.

Remark 2.2. If $\partial\Omega \in C^2$ or Ω is convex, and the coefficients of the differential operator satisfy some mild smoothness conditions, then $B : H^2(\Omega)\cap H^1_0(\Omega) \to L_2(\Omega)$ is boundedly invertible, e.g., see [Hac92] + references cited there. Since the same is valid for the adjoint B', defined by $(B'v)(u) = (Bu)(v)$, we also have that $B : L_2(\Omega) \to (H^2(\Omega)\cap H^1_0(\Omega))'$ is boundedly invertible. In view of the possibility to take $\mathscr{X} \neq \mathscr{Y}$, we infer that the adaptive wavelet method can be used also to realize the best possible convergence rate in $L_2(\Omega)$.

2.2.2 Boundary integral equations

For Ω being some domain in $\mathbb{R}^3$, let $\Gamma := \partial\Omega$. The Laplace equation on Ω or on $\mathbb{R}^3\backslash\Omega$, with either Dirichlet or Neumann boundary conditions can be reformulated as a boundary integral equation of type $(Bu)(v) := \int_\Gamma Lu(x)v(x)ds_x = f(v)$ $(v \in \mathscr{X})$, where either

$$Lu(x) := \int_\Gamma \frac{u(y)}{4\pi|x-y|}ds_y, \quad \mathscr{X} := H^{-\frac{1}{2}}(\Gamma), \tag{11}$$

or

$$Lu(x) := \pm\frac{1}{2}u(x) + \int_\Gamma \frac{(x-y)^\top \mathbf{n}_y v(y)}{4\pi|x-y|^3}ds_y, \quad \mathscr{X} := L_2(\Gamma), \tag{12}$$

or

$$Lu(x) := -\partial_{\mathbf{n}_x}\int_\Gamma \frac{(x-y)^\top \mathbf{n}_y v(y)}{4\pi|x-y|^3}ds_y, \quad \mathscr{X} := H^{\frac{1}{2}}(\Gamma)/\mathbb{R}. \tag{13}$$

In all three cases, $B : \mathscr{X} \to \mathscr{X}'$ is known to be boundedly invertible.

2.2.3 Stokes equations

The variational formulation of the Stokes equations on a domain $\Omega \subset \mathbb{R}^n$ with homogeneous Dirichlet boundary conditions reads as

$$(B(\vec{u},p))(\vec{v},q) := \int_\Omega \nabla\vec{u} : \nabla\vec{v}\,dx + \int_\Omega p\,\mathrm{div}\,\vec{v}\,dx + \int_\Omega q\,\mathrm{div}\,\vec{u}\,dx = \vec{f}(\vec{v})$$

$(\vec{v} \in H^1_0(\Omega)^n, q \in L_{2,0}(\Omega))$. With $\mathscr{X} := H^1_0(\Omega)^n \times L_{2,0}(\Omega)$, it is well-known that $B : \mathscr{X} \to \mathscr{X}'$ is boundedly invertible.

2.2.4 Parabolic evolution equations

For some domain $\Omega \subset \mathbb{R}^n$ and $T > 0$, we consider the parabolic problem

$$\begin{cases} (\partial_t u + \nabla_x \cdot \mathbf{A}\nabla_x u + \mathbf{b} \cdot \nabla_x u + cu)(t,x) = g(t,x) & (t \in (0,T), x \in \Omega), \\ u(t,x) = 0 & (t \in (0,T), x \in \partial\Omega), \\ u(0,x) = h(x) & (x \in \Omega), \end{cases}$$

where $\mathbf{A} \in L_\infty((0,T) \times \Omega)^{n \times n}$, $\mathbf{b} \in L_\infty((0,T) \times \Omega)^n$, $c \in L_\infty((0,T) \times \Omega)$, and, for some $\delta > 0$, $\mathbf{A} \geq \delta > 0$ (a.e.). With

$$\mathscr{X} := L_2(0,T) \otimes H^1_0(\Omega) \cap H^1(0,T) \otimes H^{-1}(\Omega)$$

i.e., $\mathscr{X}$ is an intersection of Bochner spaces, and

$$\mathscr{Y} := (L_2(0,T) \otimes H^1_0(\Omega)) \times L_2(\Omega),$$

and assuming that $g \in L_2((0,T); H^1_0(\Omega))'$ and $h \in L_2(\Omega)$, a variational formulation of this problem reads as: Find $u \in \mathscr{X}$ such that

$$\begin{aligned} (Bu)(v_1,v_2) &:= \int_0^T \int_\Omega (\partial_t u)v_1 + \mathbf{A}\nabla_x u \cdot \nabla_x v_1 + (\mathbf{b} \cdot \nabla_x u)v_1 + cuv_1 \, dx\, dt + \int_\Omega u(0,\cdot)v_2 \, dx \\ &= \int_0^T \int_\Omega g v_1 dx dt + \int_\Omega h v_2 \, dx \quad ((v_1,v_2) \in \mathscr{Y}). \end{aligned}$$

The operator $B : \mathscr{X} \to \mathscr{Y}'$ is boundedly invertible (cf. [SS09], [DL92, Ch.XVIII, §3], [Wlo82, Ch.IV,§26]).

3 Adaptive wavelet schemes I: Inexact Richardson iteration

3.1 Richardson iteration

Throughout this section, until Sect. 3.4, we will *assume* that there exists an $\alpha \in \mathbb{R}$ such that

$$\|\mathrm{Id} - \alpha\mathbf{B}\| < 1, \tag{14}$$

i.e., we will assume that a properly damped Richardson iteration

$$\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \alpha(\mathbf{f} - \mathbf{B}\mathbf{u}^{(i)})$$

applied to (7) converges linearly.

Lemma 3.1. *In addition to being boundedly invertible, let* $\mathbf{B}$ *satisfy* $\mathbf{B} = \mathbf{B}^\top > 0$. *Then for* $\alpha \in (0, 2/\|\mathbf{B}\|)$,

$$\|\mathrm{Id} - \alpha\mathbf{B}\| = \max(\alpha\|\mathbf{B}\| - 1, 1 - \alpha\|\mathbf{B}^{-1}\|^{-1}) < 1,$$

with minimum $\frac{\kappa(\mathbf{B})-1}{\kappa(\mathbf{B})+1}$ *when* $\alpha = 2/(\|\mathbf{B}\| + \|\mathbf{B}^{-1}\|^{-1})$, *where* $\kappa(\mathbf{B}) := \|\mathbf{B}\|\|\mathbf{B}^{-1}\|$.

Proof. Since $\mathbf{B} = \mathbf{B}^\top$, $\|\mathrm{Id} - \alpha\mathbf{B}\| = \max_{\lambda\in\sigma(\mathrm{Id}-\alpha\mathbf{B})} |\lambda| = \max_{\mu\in\sigma(\mathbf{B})} |1 - \alpha\mu|$, and from $\mathbf{B} > 0$, we have $\sigma(\mathbf{B}) \subset [\|\mathbf{B}^{-1}\|^{-1}, \|\mathbf{B}\|]$. Elementary calculations now complete the proof. □

If, apart from being boundedly invertible between $\mathscr{X}$ and $\mathscr{Y}'$, B is *symmetric*, i.e., $\mathscr{X} = \mathscr{Y}$ and $(Bv)(w) = (Bw)(v)$ $(v, w \in \mathscr{X})$, and *positive definite*, i.e., $(Bv)(v) > 0$ $(v \in \mathscr{X})$, then, because of (8), so is $\mathbf{B}$ and Lemma 3.1 applies. The example from Sect.2.2.1 when $\mathbf{b} = 0$, as well as the example from Sect. 2.2.2 in the cases (11) and (13) fall into this category.

If $\mathscr{X} = \mathscr{Y}$ and B is bounded and *coercive*, i.e., for some $\delta > 0$, $(Bv)(v) \geq \delta\|v\|^2_{\mathscr{X}}$ $(v \in \mathscr{X})$, then (8) and the next lemma show that the properly damped Richardson iteration is again convergent. An application is given by the example from Sect.2.2.1 for general $\mathbf{b} \in L_\infty(\Omega)^n$ with $\nabla \cdot \mathbf{b} = 0$ (*a.e.*).

Lemma 3.2. *If, in addition to* $\mathbf{B}$ *being bounded,* $\mathbf{B}_S := \frac{1}{2}(\mathbf{B} + \mathbf{B}^\top) > 0$ *and has a bounded inverse, then for* $\alpha \in (0, 1/(\|\mathbf{B}_S\| + \|\mathbf{B}_S^{-1}\|^{-1})]$ *with* $\alpha < 2/(\|\mathbf{B}_S^{-1}\|\|\mathbf{B}\|)$,

$$\|\mathrm{Id} - \alpha\mathbf{B}\| \leq \sqrt{1 - 2\alpha\|\mathbf{B}_S^{-1}\|^{-1} + \alpha^2\|\mathbf{B}\|^2} < 1.$$

Proof. As shown in Lemma 3.1, for $\alpha \in (0, 1/(\|\mathbf{B}_S\| + \|\mathbf{B}_S^{-1}\|^{-1})]$, $\|\mathrm{Id} - 2\alpha\mathbf{B}_s\| \leq 1 - 2\alpha\|\mathbf{B}_S^{-1}\|^{-1}$. This shows that

$$\begin{aligned}\|\mathrm{Id} - \alpha\mathbf{B}\|^2 &= \|(\mathrm{Id} - \alpha\mathbf{B})(\mathrm{Id} - \alpha\mathbf{B}^\top)\| \\ &= \|\mathrm{Id} - 2\alpha\mathbf{B}_S + \alpha^2\mathbf{B}\mathbf{B}^\top\| \leq 1 - 2\alpha\|\mathbf{B}_S^{-1}\|^{-1} + \alpha^2\|\mathbf{B}\|^2 < 1,\end{aligned}$$

when $\alpha < 2/(\|\mathbf{B}_S^{-1}\|\|\mathbf{B}\|)$.

□

3.2 Practical scheme

Of course the Richardson iteration cannot performed exactly. Generally the right-hand side $\mathbf{f}$ is infinitely supported, and although $\mathbf{B}$ is close to being sparse, generally so is any column of $\mathbf{B}$. The idea proposed in [CDD02] is to apply Richardson iteration with inexact evaluations of the matrix-vector product and of the right-hand

side **f**. It is easily seen that with a proper decay of the tolerances for these inexact evaluations as the iteration proceeds, the perturbed iteration is still linearly convergent. The issues at stake are whether the support lengths of the iterands are, up to a constant multiple, equal to the generally best possible bounds on the lengths of the best N-term approximations that give rise to the same error, and whether the computational costs to produce such iterands are bounded by the same expressions. To ensure these properties, i.e., to ensure (quasi-) *optimality* of the algorithm, assumptions are needed on the cost of the inexact application of **B** and that of the inexact evaluation of the right-hand side as a function of the prescribed tolerance.

Definition 3.1. For $\bar{s} > 0$, **B** will be called to be $\bar{s}$-*admissible* when we have available an approximate matrix times vector routine

$$\mathbf{APPLY}[\mathbf{w}, \varepsilon] \to \mathbf{z}_\varepsilon$$

that, for any $\varepsilon > 0$ and $\mathbf{w} \in \ell_0$, yields a $\mathbf{z}_\varepsilon \in \ell_0$ with

$$\|\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \le \varepsilon,$$

and, for any $s \in (0, \bar{s}]$,

$$\#\operatorname{supp} \mathbf{z}_\varepsilon \lesssim \varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}, \tag{15}$$

where the number of operations used by the call $\mathbf{APPLY}[\mathbf{w}, \varepsilon]$ is bounded by some absolute multiple of

$$\varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s} + \#\operatorname{supp} \mathbf{w} + 1. \tag{16}$$

As we will see, in order to guarantee optimality of the inexact Richardson iteration, as well as of the alternative Adaptive Wavelet-Galerkin Method discussed in Sect. 4, it will be needed that $\bar{s}$ not less than that s for which the solution **u** happens to be in $\mathscr{A}^s$. That is, with the best possible rate $s_{\max}$ as introduced in Sect. 1.1, *it will be sufficient, and generally necessary,* when

$$\bar{s} \ge s_{\max}, \tag{17}$$

an issue that was somewhat ignored in the early publications on adaptive wavelet methods. In Sect. 5, we will see that for partial differential operators with sufficiently smooth coefficients and for wavelets that are sufficiently smooth and have sufficiently many vanishing moments (or, more generally, cancellation properties) indeed (17) is valid. We include pointers to the literature where it is shown that the same holds for classes of singular integral operators,

In view of the definition of $\mathscr{A}^s$, a consequence of (15) is that **B**, restricted to ℓ_0, is a bounded mapping from $\mathscr{A}^s$ to $\mathscr{A}^s$ for $s \in (0, \bar{s}]$. As shown in [CDD01, Prop. 3.8], we even have:

Proposition 3.1. *Let* **B** *be* $\bar{s}$*-admissible. Then for* $s \in (0, \bar{s}]$, $\mathbf{B} : \mathscr{A}^s \to \mathscr{A}^s$ *is bounded, and for* $\mathbf{z}_\varepsilon := \mathbf{APPLY}[\mathbf{w}, \varepsilon]$, *we have* $\|\mathbf{z}_\varepsilon\|_{\mathscr{A}^s} \lesssim \|\mathbf{w}\|_{\mathscr{A}^s}$, *uniformly in* ε.

Proof. For $s \in (0, \bar{s}]$, $\mathbf{w} \in \mathcal{A}^s$ and $\varepsilon > 0$, let $N \in \mathbb{N}$ be such that $\|\mathbf{B}\|\|\mathbf{w} - \mathbf{w}_N\| \le \varepsilon/2$, and let $\mathbf{z}_{\varepsilon/2} := \mathbf{APPLY}[\mathbf{w}_N, \varepsilon/2]$. Then $\#\operatorname{supp} \mathbf{z}_{\varepsilon/2} \lesssim \varepsilon^{-1/s}\|\mathbf{w}_N\|_{\mathcal{A}^s}^{1/s} \le \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathcal{A}^s}^{1/s}$, and $\|\mathbf{B}\mathbf{w} - \mathbf{z}_{\varepsilon/2}\| \le \varepsilon$, showing the first statement.

Lemma 1.1 and (15) show that $\|\mathbf{z}_\varepsilon\|_{\mathcal{A}^s} \lesssim \max(\|\mathbf{B}\mathbf{w}\|_{\mathcal{A}^s}, \|\mathbf{w}\|_{\mathcal{A}^s}) \lesssim \|\mathbf{w}\|_{\mathcal{A}^s}$. □

The requirement (16) basically means that the cost of producing $\mathbf{z}_\varepsilon$ is proportional to its length plus that of $\mathbf{w}$.

Concerning the inexact evaluation of the right-hand side, *throughout this paper we assume availability of the following routine:*

RHS$[\varepsilon] \to \mathbf{f}_\varepsilon$:
% Input: $\varepsilon > 0$.
% Output: $\mathbf{f}_\varepsilon \in \ell_0$ *with*

$$\|\mathbf{f} - \mathbf{f}_\varepsilon\| \le \varepsilon \quad \textit{and} \quad \#\operatorname{supp} \mathbf{f}_\varepsilon \lesssim \min\{N : \|\mathbf{f} - \mathbf{f}_N\| \le \varepsilon\},$$

% taking a number of operations that is bounded by some absolute multiple of $\#\operatorname{supp} \mathbf{f}_\varepsilon + 1$.

A realization of **RHS** generally has to depend on the right-hand side f at hand, that, however, in contrast to the solution u, is known to the user. Noting that for $\tilde{\nabla} \subset \nabla$,

$$\|\mathbf{f} - \mathbf{f}|_{\tilde{\nabla}}\| \eqsim \|f - \sum_{\lambda \in \tilde{\nabla}} f(\psi_\lambda^{\mathscr{Y}})\psi_\lambda^{\mathscr{Y}'}\|_{\mathscr{Y}'} \eqsim \inf_{\tilde{f} \in \operatorname{span}\{\psi_\lambda^{\mathscr{Y}'} : \lambda \in \tilde{\nabla}\}} \|f - \tilde{f}\|_{\mathscr{Y}'}$$

(cf. Remark 2.1), we see that for sufficiently smooth f, **RHS** is realized by collecting, or more precisely, by approximating using suitable quadrature, the wavelet coefficients of f up to some suitable level.

Corollary 3.1. *Let* $\mathbf{B}$ *be* $\bar{s}$*-admissible. If, for some* $s \in (0, \bar{s}]$, $\mathbf{u} \in \mathcal{A}^s$, *then* $\mathbf{f}_\varepsilon :=$ **RHS**$[\varepsilon]$ *satisfies* $\#\operatorname{supp} \mathbf{f}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}$, *where the number of operations used by the call* **RHS**$[\varepsilon]$ *is bounded by some absolute multiple of*

$$\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} + 1.$$

Proof. By the assumptions and Proposition 3.1, we have $\mathbf{f} \in \mathcal{A}^s$, with $\|\mathbf{f}\|_{\mathcal{A}^s} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}$. Now the proof is completed by the definition of $\mathcal{A}^s$ and the assumptions made on **RHS**. □

Remark 3.1. Recalling that $s_{\max}$ is the approximation order of $\Psi^{\mathscr{X}}$ in $\mathscr{X}$, let $\tilde{s}_{\max}$ denote the approximation order of $\Psi^{\mathscr{Y}'}$ in $\mathscr{Y}'$.

The property, shown in Corollary 3.1, that for any $\mathbf{u} \in \mathcal{A}^s$ with $s \le \bar{s}$, it holds that $\mathbf{f} \in \mathcal{A}^s$ can only be expected when $\tilde{s}_{\max} \ge \min(\bar{s}, s_{\max})$. This means that $\bar{s}$-admissibility of $\mathbf{B}$ with $\bar{s} \ge s_{\max}$ requires that $\tilde{s}_{\max} \ge s_{\max}$.

In the scalar model situation of $\mathscr{X} = \mathscr{Y} = H^m(\Omega)$ for some domain $\Omega \subset \mathbb{R}^n$, and $\Psi^{\mathscr{X}}$, $\Psi^{\mathscr{X}'}$ being wavelet collections of order d, $\tilde{d}$, normalized in $H^m(\Omega)$ or $(H^m(\Omega))'$, respectively, it holds that $s_{\max} = \frac{d-m}{n}$ and $\tilde{s}_{\max} = \frac{\tilde{d}+m}{n}$. In this case,

$\tilde{s}_{\max} \geq s_{\max}$ means that $\tilde{d} \geq d - 2m$. For differential- and integral operators, in Sect. 5 we will see that the condition $\tilde{d} > d - 2m$ suffices to demonstrate $\bar{s}$-admissibility of $\mathbf{B}$ for $\bar{s} \geq s_{\max}$.

Remark 3.2. The properties that $\|\mathbf{f} - \mathbf{f}_\varepsilon\| \leq \varepsilon$ and, when $\mathbf{u} \in \mathscr{A}^s$, that $\#\operatorname{supp} \mathbf{f}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, with the cost of producing it being bounded by some absolute multiple of $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s} + 1$ is all that will be needed about $\mathbf{f}_\varepsilon := \mathbf{RHS}[\varepsilon]$. Our assumptions on **RHS** together with Corollary 3.1 show that these properties hold when $\mathbf{B}$ is $\bar{s}$-admissible for some $\bar{s} \geq s_{\max}$. The assumption, formulated in the description of the **RHS** routine, that we can realize quasi-best N-term approximations for $\mathbf{f}$ in linear complexity is actually stronger than what is needed when $\tilde{s}_{\max} > s_{\max}$.

Besides **APPLY** and **RHS**, the inexact Richardson iteration requires another subroutine:

COARSE$[\mathbf{w}, \varepsilon] \to \mathbf{w}_\varepsilon$*:*
% Input: $\mathbf{w} \in \ell_0$ *and* $\varepsilon > 0$.
% Output: $\mathbf{w}_\varepsilon \in \ell_0$ *with*

$$\|\mathbf{w} - \mathbf{w}_\varepsilon\| \leq \varepsilon \quad \textit{and} \quad \#\operatorname{supp} \mathbf{w}_\varepsilon \lesssim \min\{N : \|\mathbf{w} - \mathbf{w}_N\| \leq \varepsilon\}, \tag{18}$$

% taking a number of operations that is bounded by an absolute multiple of

$$\#\operatorname{supp} \mathbf{w} + \max\big(\log(\varepsilon^{-1}\|\mathbf{w}\|), 1\big).$$

An implementation of a routine **COARSE** with these properties will be given in Sect. 3.3.

The routine **COARSE** will be applied after each fixed number of (inexact) Richardson steps. The idea is to remove small coefficients from the iterands, that, because they are small, little contribute to the approximation, but, because their possibly large number, may hamper an optimal balance between accuracy and support length. Although obviously an application of **COARSE** generally increases the error, the following proposition ([Coh03, Th. 4.9.1]) shows that indeed it creates the aforementioned optimal balance.

Proposition 3.2. *Let* $\zeta > 1$ *and* $s > 0$. *Then for any* $\varepsilon > 0$, $\mathbf{v} \in \mathscr{A}^s$ *and* $\mathbf{w} \in \ell_0$ *with*

$$\|\mathbf{v} - \mathbf{w}\| \leq \varepsilon,$$

for $\mathbf{w}_{\zeta\varepsilon} := \mathbf{COARSE}[\zeta\varepsilon, \mathbf{w}]$ *we have that*

$$\#\operatorname{supp} \mathbf{w}_{\zeta\varepsilon} \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}, \quad \|\mathbf{w}_{\zeta\varepsilon}\|_{\mathscr{A}^s} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s},$$

and $\|\mathbf{v} - \mathbf{w}_{\zeta\varepsilon}\| \leq (1 + \zeta)\varepsilon$.

Proof. The smallest $N \in \mathbb{N}_0$ with $\|\mathbf{v} - \mathbf{v}_N\| \leq (\zeta - 1)\varepsilon$ satisfies

$$N \le ((\zeta-1)\varepsilon)^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.$$

From $\|\mathbf{w}-\mathbf{v}_N\| \le \|\mathbf{w}-\mathbf{v}\| + \|\mathbf{v}-\mathbf{v}_N\| \le \varepsilon + (\zeta-1)\varepsilon = \zeta\varepsilon$ and (18), it follows that $\#\operatorname{supp}\mathbf{w}_{\zeta\varepsilon} \lesssim N \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}$.

The second and last statement follow from Lemma 1.1 and an application of the triangle inequality, respectively. □

We are ready to give the inexact Richardson iteration:

Rich$[\varepsilon,\varepsilon_0] \to \mathbf{u}_\varepsilon$:
% Input: $\varepsilon > 0$ *and* $\varepsilon_0 \ge \|\mathbf{u}\|$.
% Parameters: $\theta < 1/2$, $K \in \mathbb{N}$ *and* $\rho < 1$ *such that* $\|\mathrm{Id}-\alpha\mathbf{B}\| \le \rho$ *and* $2\rho^K < \theta$.

$i := 0$, $\mathbf{u}^{(0)} := 0$
`while` $\varepsilon_i > \varepsilon$ `do`
  $i := i+1$
  $\varepsilon_i := 2\rho^K\varepsilon_{i-1}/\theta$
  $\mathbf{v}^{(i,0)} := \mathbf{u}^{(i-1)}$
  `for` $j = 1,\dots,K$ `do`
    $\mathbf{v}^{(i,j)} := \mathbf{v}^{(i,j-1)} + \alpha(\mathbf{RHS}[\frac{\rho^j\varepsilon_{i-1}}{2\alpha K}] - \mathbf{APPLY}[\mathbf{v}^{(i,j-1)}, \frac{\rho^j\varepsilon_{i-1}}{2\alpha K}])$
  `enddo`
  $\mathbf{u}^{(i)} := \mathbf{COARSE}[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$
`enddo`
$\mathbf{u}_\varepsilon := \mathbf{u}^{(i)}$

Theorem 3.1 ([CDD02]). *Let* $\varepsilon_0 \ge \|\mathbf{u}\|$, *and* $\varepsilon > 0$, *then* $\mathbf{u}_\varepsilon := \mathbf{Rich}[\varepsilon,\varepsilon_0]$ *satisfies* $\|\mathbf{u}-\mathbf{u}_\varepsilon\| \le \varepsilon$. *If for some* $s > 0$, $\mathbf{u} \in \mathscr{A}^s$, *then* $\#\operatorname{supp}\mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *If, additionally,* $\mathbf{B}$ *is* $\bar{s}$*-admissible,* $s \le \bar{s}$ *and* $\varepsilon < \varepsilon_0 \lesssim \|\mathbf{u}\|$, *then the number of operations used by the call* $\mathbf{Rich}[\varepsilon,\varepsilon_0]$ *is bounded by an absolute multiple of* $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *In other words, if* $\bar{s} \ge s_{\max}$, *then the inexact Richardson iteration is (quasi-) optimal.*

Proof. For the first statement, it suffices to show that $\|\mathbf{u}-\mathbf{u}^{(i)}\| \le \varepsilon_i$. For $i = 0$, this is clearly valid. Now for some $i \ge 1$, let $\|\mathbf{u}-\mathbf{u}^{(i-1)}\| \le \varepsilon_{i-1}$. For $1 \le j \le K$, for some δ_j with $\|\delta_j\| \le \frac{\rho^j\varepsilon_{i-1}}{K}$, we have

$$\mathbf{u}-\mathbf{v}^{(i,j)} = (\mathrm{Id}-\alpha\mathbf{B})(\mathbf{u}-\mathbf{v}^{(i,j-1)}) + \delta_j,$$

and so

$$\mathbf{u}-\mathbf{v}^{(i,K)} = (\mathrm{Id}-\alpha\mathbf{B})^K(\mathbf{u}-\mathbf{u}^{(i-1)}) + \sum_{j=1}^{K}(\mathrm{Id}-\alpha\mathbf{B})^{K-j}\delta_j.$$

From $\|\mathrm{Id}-\alpha\mathbf{B}\| \le \rho$, we infer that

$$\|\mathbf{u}-\mathbf{v}^{(i,K)}\| \le \rho^K\varepsilon_{i-1} + \sum_{j=1}^{K}\rho^{K-j}\frac{\rho^j\varepsilon_{i-1}}{K} = 2\rho^K\varepsilon_{i-1} = \theta\varepsilon_i, \tag{19}$$

and conclude that

$$\|\mathbf{u}-\mathbf{u}^{(i)}\| \le \theta\varepsilon_i + (1-\theta)\varepsilon_i = \varepsilon_i$$

as required.

Now for some $s>0$, let $\mathbf{u}\in\mathscr{A}^s$. From (19) and the definition of $\mathbf{u}^{(i)}$, Proposition 3.2 shows that

$$\#\operatorname{supp}\mathbf{u}^{(i)} \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}, \quad \|\mathbf{u}^{(i)}\|_{\mathscr{A}^s} \lesssim \|\mathbf{u}\|_{\mathscr{A}^s},$$

which bounds, as we emphasize here, hold uniformly in i. Since $\varepsilon_i \gtrsim \varepsilon_{i-1}$, the first bound shows the second statement of the theorem.

Now let $\mathbf{B}$ is $\bar{s}$-admissible for some $\bar{s}\ge s$. Since K is fixed, Proposition 3.1 shows that $\|\mathbf{v}^{(i,j)}\|_{\mathscr{A}^s} \lesssim \|\mathbf{u}\|_{\mathscr{A}^s}$, uniformly in i and j. The properties from Definition 3.1, together with Corollary 3.1 show that $\#\operatorname{supp}\mathbf{v}^{(i,j)} \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$ and that the cost of computing it from the previous iterand is bounded by an absolute multiple of $\varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. For the latter, we have used that by assumption on ε_0, $1 \lesssim \varepsilon_0^{-1/s}\|\mathbf{u}\|^{1/s} \le \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. Since the cost of the call $\mathbf{COARSE}[(1-\theta)\varepsilon_i, \mathbf{v}^{(i,K)}]$ is bounded by an absolute multiple of $\#\operatorname{supp}\mathbf{v}^{(i,K)} + \max(\log(((1-\theta)\varepsilon_i)^{-1}\|\mathbf{v}^{(i,K)}\|, 1) \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, the proof is completed by using the linear decrease of ε_i as function of i. □

Remark 3.3. Although for any $s\in(0,\bar{s}]$, $\mathbf{APPLY}[\cdot,\varepsilon]:\mathscr{A}^s\to\mathscr{A}^s$ is bounded, even uniformly in ε, there is no guarantee that by a repeated application the $\|\cdot\|_{\mathscr{A}^s}$ (quasi-) norm of the iterands does not grow beyond any bound. This was the reason to add coarsening to this inexact Richardson iteration. Numerical experiments have shown that indeed generally **COARSE** is needed to ensure optimality of the inexact Richardson iteration.

3.3 *The routines* COARSE *and* APPLY

The obvious implementation of $\mathbf{COARSE}[\mathbf{w},\varepsilon]\to\mathbf{w}_\varepsilon$ would be to order the elements of $\mathbf{w}$ by non-increasing modulus, and then to define $\mathbf{w}_\varepsilon$ as the smallest possible head of $\mathbf{w}$ such that the discarded tail has norm not larger than ε. Unfortunately, with $M := \#\operatorname{supp}\mathbf{w}$, this ordering requires $\mathscr{O}(M\log M)$ operations, so that linear complexity cannot be realized. This is the reason that in [CDD01, CDD02] on many places the suboptimal complexity of the sorting was taken into account separately. Later, this problem was solved independently by Barinka and Metselaar in [Bar05, Met02], who proposed to apply an approximate "bucket" sorting:

$\mathbf{BUCKETSORT}[\mathbf{w},\varepsilon]\to(\mathbf{w}_{[p]})_{1\le p\le P}$:
% Input: $\mathbf{w}\in\ell_0$, $\varepsilon>0$.
% Output: A distribution of the (largest) elements of $\mathbf{w}$ *over* P *"buckets".*

- *Let P be the smallest positive integer with $2^{-P/2}\|\mathbf{w}\|_\infty\sqrt{\#\operatorname{supp}\mathbf{w}} \le \varepsilon$.*
- *Store the indices of $\mathbf{w}$ in one of the P buckets, depending on the modulus of the corresponding coefficient to be in $\left(\frac{1}{\sqrt{2}}\|\mathbf{w}\|_\infty, \|\mathbf{w}\|_\infty\right]$ (first bucket), $\left(\frac{1}{2}\|\mathbf{w}\|_\infty, \frac{1}{\sqrt{2}}\|\mathbf{w}\|_\infty\right]$, …, or $\left(2^{-P/2}\|\mathbf{w}\|_\infty, 2^{-(P-1)/2}\|\mathbf{w}\|_\infty\right]$, and discard them otherwise.*
  Let $\mathbf{w}_{[p]}$ denote the restriction of $\mathbf{w}$ to indices in bucket p.

The number of buckets P is $\max(1, \lceil 2\log_2(\|\mathbf{w}\|_\infty\sqrt{\#\operatorname{supp}\mathbf{w}}/\varepsilon)\rceil)$. This number is chosen so that $\|\mathbf{w} - \sum_{p=1}^{P}\mathbf{w}_{[p]}\| \le \varepsilon$. This means that for the task of finding a (quasi) minimal Λ such that $\|\mathbf{w} - \mathbf{w}|_\Lambda\| \le \varepsilon$, these coefficients can be discarded anyway. This suggest the following coarsening routine:

COARSE$[\mathbf{w}, \varepsilon] \to \mathbf{w}_\varepsilon$:
% Input: $\mathbf{w} \in \ell_0$, $\varepsilon > 0$.

$(\mathbf{w}_{[p]})_{1\le p\le P} :=$ **BUCKETSORT**$[\mathbf{w}, \varepsilon]$
Build $\mathbf{w}_\varepsilon$ by extracting indices from the buckets, starting with the first bucket and when it got empty continuing with the second one and so on, and within each bucket in arbitrary order, until $\|\mathbf{w} - \mathbf{w}_\varepsilon\| \le \varepsilon$.

Note that for small ε, the number of buckets can be larger than $\#\operatorname{supp}\mathbf{w}$. Although then necessarily some buckets are empty, the computational cost of the call cannot be bounded on some absolute multiple of $\#\operatorname{supp}\mathbf{w}$ alone. This cost, however, can be bounded on some absolute multiple of $\#\operatorname{supp}\mathbf{w}$ plus the number of buckets. Further, since squared coefficients within one bucket differ at most a factor 2, $\#\operatorname{supp}\mathbf{w}_\varepsilon$ is at most twice as large as the length of the shortest approximation to $\mathbf{w}$ within tolerance ε. We conclude that the above implementation realizes all properties of **COARSE** that were mentioned in its description in the previous section (if necessary, consult [GHS07, Remark 2.3]).

To define a valid **APPLY** routine, we have to assume that $\mathbf{B}$ can be sufficiently well approximated by computable sparse matrices. We will assume to have available sequences $(e_j)_{j\in\mathbb{N}_0}, (c_j)_{j\in\mathbb{N}_0} \subset \mathbb{R}$, $(\mathbf{B}^{(j)})_{j\in\mathbb{N}_0} \subset \mathscr{L}(\ell_2, \ell_2)$, such that

- $\|\mathbf{B} - \mathbf{B}^{(j)}\| \le e_j$, $\lim_{j\to\infty} e_j = 0$,
- the number of non-zeros in each column of $\mathbf{B}^{(j)}$, as well as the number of operations needed to compute them, is bounded by c_j,
- $\mathbf{B}^{(0)} = 0$ (and thus $\|\mathbf{B}\| \le e_0$), $c_0 = 0$ and $\sup_{j\in\mathbb{N}_0} c_{j+1}/c_j < \infty$.

So the faster e_j decays as function of c_j, the closer is $\mathbf{B}$ to a computable sparse matrix. This motivates the following definition:

Definition 3.2. For $s^* > 0$, $\mathbf{B}$ will be called to be s^*-*computable* when for any $s < s^*$, $\sup_j e_j c_j^s < \infty$.

By specifying an approximate matrix-vector multiplication routine **APPLY**, next we will show that an s^*-computable matrix $\mathbf{B}$ is $\bar{s}$-admissible for any $\bar{s} < s^*$.

In the **APPLY** routine proposed in [DSS08] and recalled below, for some P sufficiently large $\mathbf{w}$ is split into $\sum_{p=1}^{P}\mathbf{w}_{[p]}$ plus its tail $\mathbf{w}-\sum_{p=1}^{P}\mathbf{w}_{[p]}$, after which for $1\le p\le P$, $\mathbf{B}\mathbf{w}_{[p]}$ is approximated by $\mathbf{B}^{(j_p)}\mathbf{w}_{[p]}$, where (usually) j_p grows with decreasing p. On the tail $\mathbf{w}-\sum_{p=1}^{P}\mathbf{w}_{[p]}$, and possibly also on some $\mathbf{w}_{[p]}$ with p close to P, $\mathbf{B}$ is simply approximated by the zero operator. So the basic idea is to approximate columns of $\mathbf{B}$ that correspond to large entries in the input vector $\mathbf{w}$ more accurately that those that correspond to entries that are small. This means that **APPLY** is an adaptive routine, which depends non-linearly on the input $\mathbf{w}$.

A difference with the corresponding original routine proposed in [CDD01] is that instead of the splitting of $\mathbf{w}$ into buckets, each of them containing all entries of $\mathbf{w}$ with modulus in a certain range, there $\mathbf{w}$ was chopped into parts with prescribed lengths. Secondly, and more importantly, instead of taking as in [CDD01] an a priori distribution of the accuracies of the approximations of $\mathbf{B}$ over the parts, which distribution was chosen to yield an error below the prescribed tolerance in a worst case scenario, to enhance its quantitative performance, the current implementation is based on a minimization of the cost for yielding an error below the tolerance using a posteriori information.

APPLY$[\mathbf{w},\varepsilon]\to\mathbf{z}_\varepsilon$:
% Input: $\mathbf{w}\in\ell_0$ *and* $\varepsilon>0$.

1. $[(\mathbf{w}_{[p]})_p]:=$ **BUCKETSORT**$[\mathbf{w},\varepsilon/(2e_0)]$

2. *Compute the smallest* $\ell\in\mathbb{N}_0$ *with*

$$\delta:=e_0\|\mathbf{w}-\sum_{p=1}^{\ell}\mathbf{w}_{[p]}\|\le\varepsilon/2.$$

3. *Determine* $\mathbf{j}\in\mathbb{N}_0^\ell$ *such that* $\sum_{p=1}^{\ell}e_{\mathbf{j}_p}\|\mathbf{w}_{[p]}\|\le\varepsilon-\delta$ *and* $c_{\mathbf{j}_p}\lesssim c_{\tilde{\mathbf{j}}_p}$ $(p=1,\dots,\ell)$, *where* $\tilde{\mathbf{j}}\in\mathbb{N}_0^\ell$ *is the solution of*

$$\sum_{p=1}^{\ell}c_{\tilde{\mathbf{j}}_p}\#\operatorname{supp}\mathbf{w}_{[p]}\to\min!,\quad\sum_{p=1}^{\ell}e_{\tilde{\mathbf{j}}_p}\|\mathbf{w}_{[p]}\|\le\varepsilon-\delta. \tag{20}$$

4. *Compute*

$$z_\varepsilon:=\sum_{p=1}^{\ell}\mathbf{B}^{(\mathbf{j}_p)}\mathbf{w}_{[p]}.$$

In practice, the cost of solving the exact solution (i.e., $\mathbf{j}=\tilde{\mathbf{j}}$) of the small optimization problem in 3 is neglectable. By using that $\ell=\mathcal{O}(|\log\varepsilon|)$ (see the proof of Theorem 3.2) below), and by deriving some a priori bounds for $\|\tilde{\mathbf{j}}\|_\infty$, we expect it to be possible to prove that these cost are indeed always neglectable compared to the other cost of the algorithm. Instead of doing so, however, we show how to find analytically a near optimum in 2 common situations: If for some constants C and D, $c_j=C2^{j/s^*}$ and $e_j=D2^{-j}$, so that $\mathbf{B}$ is s^*-computable, then

$$\tilde{\mathbf{j}}_p = \log_2 \left(\Big(\frac{\|\mathbf{w}_{[p]}\|}{\#\operatorname{supp} \mathbf{w}_{[p]}} \Big)^{\frac{s^*}{s^*+1}} \frac{\sum_{q=1}^{\ell} \|\mathbf{w}_{[q]}\|^{\frac{1}{s^*+1}} (\#\operatorname{supp} \mathbf{w}_{[q]})^{\frac{s^*}{1+s^*}}}{(\varepsilon-\delta)/D} \right)$$

is the solution of (20) when minimization is performed over $\mathbb{R}^\ell$. If for some constants C, D and $\omega > 0$, $c_j = Cj/\omega$ and $e_j = D2^{-j}$, so that $\mathbf{B}$ is even ∞-computable, then

$$\tilde{\mathbf{j}}_p = \log_2 \Big(\frac{\|\mathbf{w}_{[p]}\| \sum_{q=1}^{\ell} \#\operatorname{supp} \mathbf{w}_{[q]}}{\#\operatorname{supp} \mathbf{w}_{[p]} (\varepsilon-\delta)/D} \Big)$$

is the solution of (20) when minimization is performed over $\mathbb{R}^\ell$. Assuming these $\tilde{\mathbf{j}}_p$ are non-negative, by rounding them up to the nearest value in $\mathbb{N}_0$ one obtains a valid $\mathbf{j}$.

Theorem 3.2. $\mathbf{z}_\varepsilon := \mathbf{APPLY}[\mathbf{w}, \varepsilon]$ *satisfies* $\|\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \le \varepsilon$. *If* $\mathbf{B}$ *is* s^**-computable, then for any* $s < s^*$,

$$\#\operatorname{supp} \mathbf{z}_\varepsilon \lesssim \varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}, \tag{21}$$

where the number of operations required by the call is bounded by some absolute multiple of

$$\varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s} + \#\operatorname{supp} \mathbf{w} + 1. \tag{22}$$

In other words, $\mathbf{B}$ *is* $\bar{s}$*-admissible for any* $\bar{s} < s^*$.

Proof. The estimates $\|\mathbf{B}\| \|\mathbf{w} - \sum_{p=1}^{\ell} \mathbf{w}_{[p]}\| \le \delta$ and $\sum_{p=1}^{\ell} \|\mathbf{B} - \mathbf{B}^{(\mathbf{j}_p)}\| \|\mathbf{w}_{[p]}\| \le \varepsilon - \delta$ show the first statement.

Let $s \in (0, s^*)$ and select $s < s_1 < s_2 < s^*$.

As we have seen, the cost of the call $\mathbf{BUCKETSORT}[\mathbf{w}, \varepsilon/(2e_0)]$ is bounded by an absolute multiple of $\#\operatorname{supp} \mathbf{w}$ plus the number of buckets, the latter being not larger than $\max(1, \lceil 2\log_2(\|\mathbf{w}\|_\infty \sqrt{\#\operatorname{supp} \mathbf{w}}/(\varepsilon/(2e_0))) \rceil)$, so that the cost of the call is bounded by an absolute multiple of $\#\operatorname{supp} \mathbf{w} + 1 + \varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}$.

With $\tau := (\frac{1}{2} + s)^{-1}$, Lemma 1.2 shows that

$$\#\operatorname{supp} \mathbf{w}_{[p]} \le \#\{\lambda \in \nabla : |\mathbf{w}_\lambda| > 2^{-p/2} \|\mathbf{w}\|_\infty\} \lesssim 2^{p\tau/2} \|\mathbf{w}\|_\infty^{-\tau} \|\mathbf{w}\|_{\mathscr{A}^s}^{\tau},$$

so that

$$\|\mathbf{w}_{[p]}\| \lesssim 2^{-p/2} \|\mathbf{w}\|_\infty \sqrt{\#\operatorname{supp} \mathbf{w}_{[p]}} \lesssim 2^{-ps\tau/2} \|\mathbf{w}\|_\infty^{1-\tau/2} \|\mathbf{w}\|_{\mathscr{A}^s}^{\tau/2}.$$

The proof will be completed once we have shown that there exists *some* $\mathbf{j} \in \mathbb{N}_0^\ell$ with $\sum_{p=1}^{\ell} e_{\mathbf{j}_p} \|\mathbf{w}_{[p]}\| \le \varepsilon - \delta$ and $\sum_{p=1}^{\ell} c_{\mathbf{j}_p} \#\operatorname{supp} \mathbf{w}_{[p]} \lesssim \varepsilon^{-1/s} \|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}$. For $\ell = 0$ there is nothing to prove, so we assume that $\ell > 0$.

First, we derive an upper bound for ℓ determined in step 2 of **APPLY**. By definition of ℓ, we have

$$\varepsilon/2 < e_0 \|\mathbf{w} - \sum_{p=1}^{\ell-1} \mathbf{w}_{[p]}\| = e_0 \sqrt{\sum_{p=\ell}^{\infty} \|\mathbf{w}_{[p]}\|^2} \lesssim e_0 2^{-\ell s\tau/2} \|\mathbf{w}\|_\infty^{1-\tau/2} \|\mathbf{w}\|_{\mathscr{A}^s}^{\tau/2},$$

or

$$2^{\ell\tau/2}\|\mathbf{w}\|_\infty^{-\tau}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau} \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}. \tag{23}$$

Here we used the notation $\mathbf{w}_{[p]}$ also to denote the restriction of $\mathbf{w}$ to indices in buckets beyond those that were generated by the call $\mathbf{BUCKETSORT}[\mathbf{w}, \varepsilon/(2e_0)]$.

Next, let $J \geq \ell$ be defined as the smallest integer such that

$$\sum_{p=1}^{\ell} 2^{-(J-p)s_1\tau/2}\|\mathbf{w}_{[p]}\| \leq \varepsilon - \delta. \tag{24}$$

In case that $J > \ell$, from $s_1 > s$ we have

$$\begin{aligned}
\varepsilon/2 \leq \varepsilon - \delta &< \sum_{p=1}^{\ell} 2^{-(J-1-p)s_1\tau/2}\|\mathbf{w}_{[p]}\| \\
&< \sum_{p=1}^{\ell} 2^{-(J-1-p)s_1\tau/2}2^{-ps\tau/2}\|\mathbf{w}\|_\infty^{1-\tau/2}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau/2} \\
\textit{synthesis} &\lesssim 2^{-(J-1-\ell)s_1\tau/2}2^{-\ell s\tau/2}\|\mathbf{w}\|_\infty^{1-\tau/2}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau/2} \\
&\leq 2^{-(J-1)s\tau/2}\|\mathbf{w}\|_\infty^{1-\tau/2}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau/2},
\end{aligned}$$

or

$$2^{J\tau/2}\|\mathbf{w}\|_\infty^{-\tau}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau} \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathscr{A}^s}^{1/s}. \tag{25}$$

From (23) we see that the upper bound on J given by (25) is also valid when $J = \ell$.

Now we select $\mathbf{j}_p$ as to be the smallest integer such that $e_{\mathbf{j}_p} \leq 2^{-(J-p)s_1\tau/2}$. Then (24) shows that indeed $\sum_{p=1}^{\ell} e_{\mathbf{j}_p}\|\mathbf{w}_{[p]}\| \leq \varepsilon - \delta$. Because of $\sup_j e_j c_j^{s_2} < \infty$ and $\sup_j c_{j+1}/c_j < \infty$, we have $c_{\mathbf{j}_p} \lesssim c_{\mathbf{j}_p-1} \lesssim e_{\mathbf{j}_p-1}^{-1/s_2} < 2^{(J-p)(s_1/s_2)(\tau/2)}$. From (25), we conclude that

$$\begin{aligned}
\sum_{p=1}^{\ell} c_{\mathbf{j}_p}\#\operatorname{supp}\mathbf{w}_{[p]} &\lesssim \sum_{p=1}^{\ell} 2^{(J-p)(s_1/s_2)\tau/2}2^{p\tau/2}\|\mathbf{w}\|_\infty^{-\tau}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau} \\
&\lesssim 2^{(J-\ell)(s_1/s_2)\tau/2}2^{\ell\tau/2}\|\mathbf{w}\|_\infty^{-\tau}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau} \\
&\lesssim 2^{J\tau/2}\|\mathbf{w}\|_\infty^{-\tau}\|\mathbf{w}\|_{\mathscr{A}^s}^{\tau} \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathscr{A}^s}^{1/s},
\end{aligned}$$

which completes the proof. □

3.4 Non-coercive **B**

If **B** is non-coercive, then the Richardson iteration may not converge, and so the inexact Richardson iteration does not apply. A general applicable remedy is to apply the inexact Richardson iteration to the *normal equations*

$$\mathbf{B}^\top \mathbf{B}\mathbf{u} = \mathbf{B}^\top \mathbf{f}.$$

Clearly, the operator $\mathbf{B}^\top\mathbf{B}$ is symmetric, positive definite, and boundedly invertible with $\|\mathbf{B}^\top\mathbf{B}\| = \|\mathbf{B}\|^2$ and $\|(\mathbf{B}^\top\mathbf{B})^{-1}\| = \|\mathbf{B}^{-1}\|^2$. In order to conclude that the inexact Richardson iteration applied to the normal equations is (quasi-) optimal, what is left to show is that for some $\bar{s} \geq s_{\max}$, $\mathbf{B}^\top\mathbf{B}$ is $\bar{s}$-admissible, and that we have a valid routine for approximating the right-hand side $\mathbf{B}^\top\mathbf{f}$ in the sense of Remark 3.2. Proposition 3.3 from [CDD02, Sect. 7] given below shows that both conditions are fulfilled when $\mathbf{B}$ and $\mathbf{B}^\top$ are $\bar{s}$-admissible for some $\bar{s} \geq s_{\max}$.

Concerning the latter, from Theorem 3.2, recall that $\mathbf{B}$ is $\bar{s}$-admissible for some $\bar{s} \geq s_{\max}$ when it is s^*-computable for some $s^* > s_{\max}$. The results demonstrating s^*-computability of $\mathbf{B}$, that will be given in Sect. 5, are symmetric in the sense that they also show s^*-computability of $\mathbf{B}^\top$ for the same value of s^*.

Proposition 3.3. (a). *If* $\mathbf{B}$ *and* $\mathbf{B}^\top$ *are* $\bar{s}$*-admissible, then so is* $\mathbf{B}^\top\mathbf{B}$. *With the* **APPLY** *routines for* $\mathbf{B}$ *and* $\mathbf{B}^\top$ *denoted as* $\mathbf{APPLY}_{\mathbf{B}}$ *and* $\mathbf{APPLY}_{\mathbf{B}^T}$, *respectively, and with* e_0 *being an upper bound for* $\|\mathbf{B}\|$, *a valid* **APPLY** *for* $\mathbf{B}^\top\mathbf{B}$ *is given by*

$$[\mathbf{w}, \varepsilon] \mapsto \mathbf{z}_\varepsilon := \mathbf{APPLY}_{\mathbf{B}^\top}[\mathbf{APPLY}_{\mathbf{B}}[\mathbf{w}, \varepsilon/(2e_0)], \varepsilon/2].$$

(b). *For* $\varepsilon > 0$, $\mathbf{g}_\varepsilon := \mathbf{APPLY}_{\mathbf{B}^\top}[\mathbf{RHS}[\varepsilon/(2e_0)], \varepsilon/2]$ *satisfies* $\|\mathbf{B}^\top\mathbf{f} - \mathbf{g}_\varepsilon\| \leq \varepsilon$. *If* $\mathbf{B}$ *and* $\mathbf{B}^\top$ *are* $\bar{s}$*-admissible, then whenever for some* $s \in (0, \bar{s}]$, $\mathbf{u} \in \mathcal{A}^s$, *it holds that* $\#\operatorname{supp}\mathbf{g}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}$, *with the cost of producing it being bounded by some absolute multiple of* $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} + 1$.

Proof. (a).

$$\|\mathbf{B}^\top\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \leq \|\mathbf{B}^\top(\mathbf{B}\mathbf{w} - \mathbf{APPLY}_{\mathbf{B}}[\mathbf{w}, \varepsilon/(2e_0)])\| + \varepsilon/2 \leq \|\mathbf{B}\|\varepsilon/(2e_0) + \varepsilon/2 \leq \varepsilon.$$

Let $s \in (0, \bar{s}]$. Putting $\mathbf{t}_\varepsilon := \mathbf{APPLY}_{\mathbf{B}}[\mathbf{w}, \varepsilon/(2e_0)]$, from $\mathbf{B}$ being $\bar{s}$-admissible, we know that

$$\#\operatorname{supp}\mathbf{t}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathcal{A}^s}^{1/s},$$

and that the cost of producing it is bounded by some absolute multiple of

$$\varepsilon^{-1/s}\|\mathbf{w}\|_{\mathcal{A}^s}^{1/s} + \#\operatorname{supp}\mathbf{w} + 1.$$

Moreover, Proposition 3.1 shows that $\|\mathbf{t}_\varepsilon\|_{\mathcal{A}^s} \lesssim \|\mathbf{w}\|_{\mathcal{A}^s}$, uniformly in ε (and in $\mathbf{w}$).

From $\mathbf{B}^\top$ being $\bar{s}$-admissible, we know that

$$\#\operatorname{supp}\mathbf{z}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{t}_\varepsilon\|_{\mathcal{A}^s}^{1/s} \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathcal{A}^s}^{1/s},$$

and that the cost of producing it from $\mathbf{t}_\varepsilon$ is bounded by a constant multiple of $\varepsilon^{-1/s}\|\mathbf{t}_\varepsilon\|_{\mathcal{A}^s}^{1/s} + \#\operatorname{supp}\mathbf{t}_\varepsilon + 1 \lesssim \varepsilon^{-1/s}\|\mathbf{w}\|_{\mathcal{A}^s}^{1/s} + 1$. We conclude that indeed $\mathbf{B}^\top\mathbf{B}$ is $\bar{s}$-admissible.

The proof of (b) is similar to that of (a). □

3.5 Alternatives for the Richardson iteration

As already appears from Lemma 3.1, the quantitative performance of the approximate Richardson scheme will depend on the spectral condition number of the matrix being inverted. In this respect, the approach, for non-coercive $\mathbf{B}$, of applying the inexact Richardson iteration to the normal equations, which gives rise to a squared condition number, might not always be the best possible choice.

For (symmetric) saddle point problems, as the Stokes equations from Sect. 2.2.3, as alternatives, in [CDD02] it was proposed to apply the inexact Richardson iteration to the reformulation introduced in [BP88] of the saddle point problem as a symmetric positive definite system, or to the Schur complement system (if necessary after first switching to the augmented Lagrangian formulation). In the latter case, each iteration requires the application of the Schur complement operator, and so in particular, the solution of an elliptic system. Necessarily, these systems can only be solved approximately, in which case the resulting scheme is known as the inexact Uzawa iteration. With the inner elliptic problems being solved with an adaptive wavelet method, (quasi-) optimality of the overall scheme in the sense of Theorem 3.1 was demonstrated in [DDU02].

Also in cases where the Richardson scheme applies directly to $\mathbf{Bu} = \mathbf{f}$, one may think of applying a more advanced iterative method. For symmetric and positive definite $\mathbf{B}$, in [CU05, DFR$^+$07b] it was shown that an approximate *Steepest Descent* method, with appropriate tolerances for the inexact matrix-vector and right-hand side evaluations, is (quasi-) optimal. Since the asymptotic convergence rate of the optimally damped Richardson iteration is equal to that of the Steepest Descent method, the main advantage of the latter scheme lies in the fact that it frees the user of the task of providing accurate estimates of the extremal eigenvalues of $\mathbf{B}$.

For $\mathbf{B}$ being only coercive, instead of the Steepest Descent method, the *Minimal Residual* method (see e.g. [Saa03]) might be applied. We envisage that (quasi-) optimality of an approximate Minimal Residual method can be proven along the same lines as for the Steepest Descent method. Since for $\mathbf{B}$ being only coercive it is even less obvious how to choose the damping parameter in the Richardson scheme, the advantage of the approximate Minimal Residual method is likely even bigger.

Even more advanced schemes than the Steepest Descent or Minimal Residual method are *Krylov subspace* methods, like the Conjugate Gradient method for symmetric positive definite systems. Clearly, in the infinite dimensional setting, these schemes can only be applied with inexact evaluations of the residuals. Numerical results are reported in [BK08]. With a suitable choice of the tolerances for these inexact evaluations, the approximate Conjugate Gradient method has been shown to converge ([vS04]). Yet, as far as we know, in the infinite dimensional setting it has not been proven that there exists a choice of the tolerances such that the resulting scheme is not only convergent but also (quasi-) optimal. Indeed, recall that the tolerances determine the support lengths of the iterands (except immediately after coarsening), and with that the cost of the algorithm. So in view of this observation, it is not necessarily true that a faster converging iteration gives rise, when applied approximately, to a quantitatively better performing adaptive wavelet scheme.

In the next section, we will study the Adaptive Wavelet-Galerkin Method proposed in [CDD01] and later modified in [GHS07]. As we will see, unlike the methods we discussed so far, this scheme can *not* be viewed as an inexact evaluation of some convergent iterative scheme applied to the bi-infinite matrix vector problem.

4 Adaptive wavelet schemes II: The Adaptive wavelet-Galerkin method

Throughout this section we will assume that $\mathbf{B}$ is symmetric and positive definite, i.e., $\mathbf{B} = \mathbf{B}^\top > 0$. On $\ell_2(\nabla)$, we define

$$|||\cdot||| := \langle \mathbf{B}\cdot,\cdot\rangle^{\frac{1}{2}}.$$

Remark 4.1. If $\mathbf{B}$ is not symmetric and positive definite, then the scheme presented here can be applied to the normal equations $\mathbf{B}^\top\mathbf{B}\mathbf{u} = \mathbf{B}^\top\mathbf{f}$, meaning that in the following everywhere $\mathbf{B}$ should be read as $\mathbf{B}^\top\mathbf{B}$ and $\mathbf{f}$ as $\mathbf{B}^\top\mathbf{f}$.

For any $\Lambda \subset \nabla$, with $\ell_2(\Lambda)$ we will mean the subspace of $\mathbf{v} \in \ell_2(\nabla)$ with supports in Λ. The trivial embedding of $\ell_2(\Lambda)$ into $\ell_2(\nabla)$ will be denoted by $\mathbf{I}_\Lambda$, and its adjoint with respect to $\langle\cdot,\cdot\rangle$, i.e., the operator that replaces coefficients outside Λ by zeros, will be denoted by $\mathbf{P}_\Lambda$. We set

$$\mathbf{B}_\Lambda := \mathbf{P}_\Lambda\mathbf{B}\mathbf{I}_\Lambda.$$

Using that $\mathbf{B}$ is symmetric and positive definite, one verifies that for any $\Lambda \subseteq \nabla$,

$$\begin{aligned}\|\mathbf{B}_\Lambda^{-1}\|^{-\frac{1}{2}}\|\cdot\| \le\ |||\cdot||| \ &\le \|\mathbf{B}_\Lambda\|^{\frac{1}{2}}\|\cdot\| \quad \text{on } \ell_2(\Lambda),\\ \|\mathbf{B}_\Lambda^{-1}\|^{-\frac{1}{2}}|||\cdot||| \le \|\mathbf{B}_\Lambda\cdot\| &\le \|\mathbf{B}_\Lambda\|^{\frac{1}{2}}|||\cdot||| \quad \text{on } \ell_2(\Lambda),\end{aligned}$$

as well as $\|\mathbf{B}_\Lambda\| \le \|\mathbf{B}\|$ and $\|\mathbf{B}_\Lambda^{-1}\| \le \|\mathbf{B}^{-1}\|$, which properties will be used often in the following.

4.1 The adaptive wavelet-Galerkin method (AWGM) in a idealized setting

The solution $\mathbf{u}_\Lambda \in \ell_2(\Lambda)$ of $\mathbf{B}_\Lambda\mathbf{u}_\Lambda = \mathbf{P}_\Lambda\mathbf{f}$ is known as the *Galerkin approximation* to $\mathbf{u}$ from $\ell_2(\Lambda)$. With respect to $|||\cdot|||$, it is the best approximation to $\mathbf{u}$ from this subspace.

The idea of the AWGM is to loop over the following 2 steps: Given $\Lambda \subset \nabla$, compute the Galerkin approximation $\mathbf{u}_\Lambda$. Enlarge Λ to a set $\tilde{\Lambda} \subset \nabla$ such that for some constant $\rho < 1$, $|||\mathbf{u} - \mathbf{u}_{\tilde{\Lambda}}||| \le \rho|||\mathbf{u} - \mathbf{u}_\Lambda|||$. This loop is similar to the one that

underlies the Adaptive Finite Element Method (AFEM), where the enlargement of Λ corresponds to mesh-refinement. The AFEM is discussed by R. Nochetto in another chapter of this book.

In the AFEM, a refinement that guarantees error reduction is obtained by computing an a posteriori error estimator, being the square root of the sum of local error indicators associated to the elements, and by refining those elements that carry the largest error indicators and whose joint sum can be bounded from below on a constant multiple of the total squared a posteriori error estimator (this is known as the so-called *bulk criterion*). The AWGM works according to the same principle, with the role of the a posteriori error estimator being played by the residual $\mathbf{f}-\mathbf{B}\mathbf{u}_\Lambda$, where for the moment we ignore the fact that this residual cannot be computed exactly.

The next lemma, being [CDD01, Lemma 4.1], shows convergence of the AWGM. Although in this lemma $\mathbf{w}$ can be a general function in $\ell_2(\Lambda)$, we have in mind it to be (an approximation to) the Galerkin approximation $\mathbf{u}_\Lambda$.

Lemma 4.1. *Let $\mu \in (0,1]$, $\mathbf{w} \in \ell_2(\Lambda)$ and $\Lambda \subset \tilde{\Lambda} \subset \nabla$ such that*

$$\|\mathbf{P}_{\tilde{\Lambda}}(\mathbf{f}-\mathbf{B}\mathbf{w})\| \geq \mu \|\mathbf{f}-\mathbf{B}\mathbf{w}\|. \tag{26}$$

Then, for $\mathbf{u}_{\tilde{\Lambda}} \in \ell_2(\tilde{\Lambda})$ being the solution of $\mathbf{B}_{\tilde{\Lambda}}\mathbf{u}_{\tilde{\Lambda}} = \mathbf{P}_{\tilde{\Lambda}}\mathbf{f}$, we have

$$|||\mathbf{u}-\mathbf{u}_{\tilde{\Lambda}}||| \leq \left[1-\mu^2\kappa(\mathbf{B})^{-1}\right]^{\frac{1}{2}} |||\mathbf{u}-\mathbf{w}|||.$$

Proof. We have

$$\begin{aligned} |||\mathbf{u}_{\tilde{\Lambda}}-\mathbf{w}||| &\geq \|\mathbf{B}\|^{-\frac{1}{2}} \|\mathbf{B}(\mathbf{u}_{\tilde{\Lambda}}-\mathbf{w})\| \geq \|\mathbf{B}\|^{-\frac{1}{2}} \|\mathbf{P}_{\tilde{\Lambda}}(\mathbf{f}-\mathbf{B}\mathbf{w})\| \\ &\geq \|\mathbf{B}\|^{-\frac{1}{2}} \mu \|\mathbf{f}-\mathbf{B}\mathbf{w}\| \geq \mu\,\kappa(\mathbf{B})^{-\frac{1}{2}} |||\mathbf{u}-\mathbf{w}|||. \end{aligned}$$

The proof of is completed by using the Galerkin orthogonality

$$|||\mathbf{u}-\mathbf{w}|||^2 = |||\mathbf{u}-\mathbf{u}_{\tilde{\Lambda}}|||^2 + |||\mathbf{u}_{\tilde{\Lambda}}-\mathbf{w}|||^2. \qquad \square$$

In Lemma 4.1, $\tilde{\Lambda}$ is some enlargement of Λ such that the bulk criterion (26) is satisfied. The natural approach is to construct $\tilde{\Lambda}$ by gathering the indices of the *largest* elements in modulus of the residual. In [CDD01], the corresponding practical algorithm –i.e., with the inexact solution of the arising Galerkin systems and the inexact evaluation of the residuals using the **APPLY** and **RHS** routines– was shown to be (quasi-) optimal by the addition of a recurrent application of **COARSE**, similar to the inexact Richardson iteration from Sect. 3.

In the next lemma, being [GHS07, Lemma 2.1], it is shown that when μ is taken to be sufficiently small, then the cardinality of the expansion $\tilde{\Lambda}\setminus\Lambda$ can be controlled. This lemma will be the key to show that the algorithm from [CDD01] *without* a recurrent coarsening of the iterands is already (quasi-) optimal (coarsening will still be used to find the large entries from the approximate residuals). Later, basically the

same technique was used to show that the standard adaptive finite element method, so without coarsening, is (quasi-) optimal, see [Ste07].

Lemma 4.2. *If, in the situation of Lemma 4.1,* $\mu < \kappa(\mathbf{B})^{-\frac{1}{2}}$ *and* $\tilde{\Lambda} \supset \Lambda$ *is the* smallest *set satisfying* (26)*, then*

$$\#(\tilde{\Lambda}\setminus\Lambda) \le \min\{N : |||\mathbf{u}-\mathbf{u}_N||| \le [1-\mu^2\kappa(\mathbf{B})]^{\frac{1}{2}}|||\mathbf{u}-\mathbf{w}|||\}. \tag{27}$$

Proof. For an N as in the right-hand side of (27), let $\breve{\Lambda} := \Lambda \bigcup \operatorname{supp}\mathbf{u}_N$. Then, for the solution of $\mathbf{B}_{\breve{\Lambda}}\mathbf{u}_{\breve{\Lambda}} = \mathbf{P}_{\breve{\Lambda}}\mathbf{f}$, we have $|||\mathbf{u}-\mathbf{u}_{\breve{\Lambda}}||| \le |||\mathbf{u}-\mathbf{u}_N|||$, and so by Galerkin orthogonality

$$|||\mathbf{u}_{\breve{\Lambda}}-\mathbf{w}||| \ge \mu\kappa(\mathbf{B})^{\frac{1}{2}}|||\mathbf{u}-\mathbf{w}|||,$$

giving

$$\begin{aligned}\|\mathbf{P}_{\breve{\Lambda}}(\mathbf{f}-\mathbf{B}\mathbf{w})\| &= \|\mathbf{B}_{\breve{\Lambda}}(\mathbf{u}_{\breve{\Lambda}}-\mathbf{w})\| \ge \|\mathbf{B}^{-1}\|^{-\frac{1}{2}}|||\mathbf{u}_{\breve{\Lambda}}-\mathbf{w}|||\\ &\ge \|\mathbf{B}^{-1}\|^{-\frac{1}{2}}\mu\kappa(\mathbf{B})^{\frac{1}{2}}|||\mathbf{u}-\mathbf{w}||| \ge \mu\|\mathbf{f}-\mathbf{B}\mathbf{w}\|.\end{aligned}$$

By our assumption on $\tilde{\Lambda}$, we conclude that $\#(\tilde{\Lambda}\setminus\Lambda) \le \#(\breve{\Lambda}\setminus\Lambda) \le N$. □

Lemmas 4.1 and 4.2 suggest the following routine:

exact-AWGM:
% Parameter: $\mu \in (0, \kappa(\mathbf{B})^{-\frac{1}{2}})$.

$\Lambda_0 := \emptyset$, $\mathbf{u}_{\Lambda_0} := 0$,
`for` $i = 1,2,\ldots$ `do`
  find the smallest $\Lambda_{i+1} \supset \Lambda_i$ *with* $\|\mathbf{P}_{\Lambda_{i+1}}(\mathbf{f}-\mathbf{B}\mathbf{u}_{\Lambda_i})\| \ge \mu\|\mathbf{f}-\mathbf{B}\mathbf{u}_{\Lambda_i}\|$
  solve $\mathbf{B}_{\Lambda_{i+1}}\mathbf{u}_{\Lambda_{i+1}} = \mathbf{P}_{\Lambda_{i+1}}\mathbf{f}$
`enddo`

Proposition 4.1. *For* $(\mathbf{u}_{\Lambda_i})_i$ *produced by* **exact-AWGM***, we have*

$$|||\mathbf{u}-\mathbf{u}_{\Lambda_i}||| \le \left[1-\mu^2\kappa(\mathbf{B})^{-1}\right]^{i/2}|||\mathbf{u}|||,$$

and if for some $s > 0$, $\mathbf{u} \in \mathscr{A}^s$*, then*

$$\#\operatorname{supp}\mathbf{u}_{\Lambda_i} \lesssim \|\mathbf{u}-\mathbf{u}_{\Lambda_{i-1}}\|^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}.$$

Proof. For $0 \le k \le i$, Lemma 4.1 shows that $|||\mathbf{u}-\mathbf{u}_{\Lambda_i}||| \le \rho^{i-k}|||\mathbf{u}-\mathbf{u}_{\Lambda_k}|||$ where $\rho := \left[1-\mu^2\kappa(\mathbf{B})^{-1}\right]^{\frac{1}{2}}$, which in particular shows the first statement.

Assuming that $\mathbf{u} \in \mathscr{A}^s$ for some $s > 0$, with $\sigma := [1-\mu^2\kappa(\mathbf{B})]^{\frac{1}{2}}$, Lemma 4.2 shows that

$$\begin{aligned}\#(\Lambda_k\backslash\Lambda_{k-1}) &\leq \min\{N : |||\mathbf{u}-\mathbf{u}_N||| \leq \sigma|||\mathbf{u}-\mathbf{u}_{\Lambda_{k-1}}|||\}\\ &\leq \min\{N : \|\mathbf{u}-\mathbf{u}_N\| \leq \|\mathbf{B}\|^{-\frac{1}{2}}\sigma|||\mathbf{u}-\mathbf{u}_{\Lambda_{k-1}}|||\}\\ &\leq [\|\mathbf{B}\|^{-\frac{1}{2}}\sigma|||\mathbf{u}-\mathbf{u}_{\Lambda_{k-1}}|||]^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s},\end{aligned}$$

by $|||\cdot||| \leq \|\mathbf{B}\|^{\frac{1}{2}}\|\cdot\|$ and the definition of $\|\cdot\|_{\mathscr{A}^s}$.

By combining both estimates, for $i \in \mathbb{N}$ we have

$$\begin{aligned}\#\operatorname{supp}\mathbf{u}_{\Lambda_i} \leq \#\Lambda_i = \sum_{k=1}^{i}\#(\Lambda_k\backslash\Lambda_{k-1}) &\leq \|\mathbf{B}\|^{1/2s}\sigma^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}\sum_{k=1}^{i}|||\mathbf{u}-\mathbf{u}_{\Lambda_{k-1}}|||^{-1/s}\\ &\leq \|\mathbf{B}\|^{1/2s}\sigma^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}|||\mathbf{u}-\mathbf{u}_{\Lambda_{i-1}}|||^{-1/s}\sum_{k=1}^{i}(\rho^{i-k})^{1/s}\\ &\leq \kappa(\mathbf{B})^{1/2s}\frac{\sigma^{-1/s}}{1-\rho^{1/s}}\|\mathbf{u}-\mathbf{u}_{\Lambda_{i-1}}\|^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s},\end{aligned} \tag{28}$$

by $\|\cdot\| \leq \|\mathbf{B}^{-1}\|^{\frac{1}{2}}|||\cdot|||$. □

In view of the definition of $\mathscr{A}^s$, note that the bound on $\#\operatorname{supp}\mathbf{u}_{\Lambda_i}$ derived in Proposition 4.1 is, up to some constant multiple, the generally best possible one. That is, not taking into account the computational cost, the routine **exact-AWGM** is *(quasi-) optimal*.

4.2 Practical scheme

In this subsection, we turn **exact-AWGM** into a practical scheme by

- computing residuals only approximately,
- allowing that for the enlargement Λ_{i+1} of Λ_i, which satisfies the "bulk criterion", $\#(\Lambda_{i+1}\backslash\Lambda_i)$ is only minimal up to some constant multiple,
- solving the arising Galerkin problems only approximately.

The following proposition extends Lemmas 4.1 and 4.2 to this setting.

Proposition 4.2. *Let $\delta \in (0,\alpha)$, $\gamma > 0$ be constants such that $\mu := \frac{\alpha+\delta}{1-\delta} < \kappa(\mathbf{B})^{-\frac{1}{2}}$ and $\gamma < \frac{(1-\delta)(\alpha-\delta)}{1+\delta}\kappa(\mathbf{B})^{-1}$. Given $\Lambda \subset \nabla$ and $\mathbf{w} \in \ell_2(\Lambda)$, let $\mathbf{r} \in \ell_2(\nabla)$ be such that*

$$\|\mathbf{f}-\mathbf{B}\mathbf{w}-\mathbf{r}\| \leq \delta\|\mathbf{r}\|. \tag{29}$$

Let $\Lambda \subset \tilde{\Lambda} \subset \nabla$ be such that

$$\|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}\| \geq \alpha\|\mathbf{r}\| \tag{30}$$

and such that, up to some absolute multiple, $\#(\tilde{\Lambda}\backslash\Lambda)$ is minimal among all such $\tilde{\Lambda}$. Let $\tilde{\mathbf{w}} \in \ell_2(\tilde{\Lambda})$ be an approximation to $\mathbf{u}_{\tilde{\Lambda}}$ such that

$$\|\mathbf{P}_{\tilde\Lambda}\mathbf{f}-\mathbf{B}_{\tilde\Lambda}\tilde{\mathbf{w}}\|\le\gamma\|\mathbf{r}\|. \tag{31}$$

Then it holds that[1]

$$|||\mathbf{u}-\tilde{\mathbf{w}}|||\le\rho|||\mathbf{u}-\mathbf{w}|||, \tag{32}$$

where $\rho:=\big[1-\big(\frac{\alpha-\delta}{1+\delta}\big)^2\kappa(\mathbf{B})^{-1}+\frac{\gamma^2}{(1-\delta)^2}\kappa(\mathbf{B})\big]^{\frac12}<1$, *and*

$$\#(\tilde\Lambda\setminus\Lambda)\lesssim\min\{N:|||\mathbf{u}-\mathbf{u}_N|||\le[1-\mu^2\kappa(\mathbf{B})]^{\frac12}|||\mathbf{u}-\mathbf{w}|||\}.$$

Proof. From $\|\mathbf{f}-\mathbf{B}\mathbf{w}\|\le(1+\delta)\|\mathbf{r}\|$ and $\|\mathbf{P}_{\tilde\Lambda}\mathbf{r}\|\le\|\mathbf{P}_{\tilde\Lambda}(\mathbf{f}-\mathbf{B}\mathbf{w})\|+\delta\|\mathbf{r}\|$, we have $\|\mathbf{P}_{\tilde\Lambda}(\mathbf{f}-\mathbf{B}\mathbf{w})\|\ge(\alpha-\delta)\|\mathbf{r}\|\ge\frac{\alpha-\delta}{1+\delta}\|\mathbf{f}-\mathbf{B}\mathbf{w}\|$, so that Lemma 4.1 shows that

$$|||\mathbf{u}-\mathbf{u}_{\tilde\Lambda}|||\le[1-(\tfrac{\alpha-\delta}{1+\delta})^2\kappa(\mathbf{B})^{-1}]^{\frac12}|||\mathbf{u}-\mathbf{w}|||. \tag{33}$$

We have

$$\begin{aligned}|||\mathbf{u}_{\tilde\Lambda}-\tilde{\mathbf{w}}|||&\le\|\mathbf{B}^{-1}\|^{\frac12}\|\mathbf{P}_{\tilde\Lambda}\mathbf{f}-\mathbf{B}_{\tilde\Lambda}\tilde{\mathbf{w}}\|\le\|\mathbf{B}^{-1}\|^{\frac12}\gamma\|\mathbf{r}\|\\&\le\|\mathbf{B}^{-1}\|^{\frac12}\tfrac{\gamma}{1-\delta}\|\mathbf{f}-\mathbf{B}\mathbf{w}\|\le\tfrac{\gamma}{1-\delta}\kappa(\mathbf{B})^{\frac12}|||\mathbf{u}-\mathbf{w}|||.\end{aligned}$$

The last two displayed formulas together with $|||\mathbf{u}-\tilde{\mathbf{w}}|||^2=|||\mathbf{u}-\mathbf{u}_{\tilde\Lambda}|||^2+|||\mathbf{u}_{\tilde\Lambda}-\tilde{\mathbf{w}}|||^2$ show (32). The condition on γ shows that $\rho<1$.

Let $\Lambda\subset\hat\Lambda\subset\nabla$ be the smallest set with

$$\|\mathbf{P}_{\hat\Lambda}(\mathbf{f}-\mathbf{B}\mathbf{w})\|\ge\mu\|\mathbf{f}-\mathbf{B}\mathbf{w}\|.$$

Then

$$\mu\|\mathbf{r}\|\le\mu\|\mathbf{f}-\mathbf{B}\mathbf{w}\|+\mu\delta\|\mathbf{r}\|\le\|\mathbf{P}_{\hat\Lambda}(\mathbf{f}-\mathbf{B}\mathbf{w})\|+\mu\delta\|\mathbf{r}\|\le\|\mathbf{P}_{\hat\Lambda}\mathbf{r}\|+(1+\mu)\delta\|\mathbf{r}\|$$

or $\|\mathbf{P}_{\hat\Lambda}\mathbf{r}\|\ge(\mu-(1+\mu)\delta)\|\mathbf{r}\|=\alpha\|\mathbf{r}\|$. We conclude that

$$\#(\tilde\Lambda\setminus\Lambda)\lesssim\#(\hat\Lambda\setminus\Lambda)\le\min\{N:|||\mathbf{u}-\mathbf{u}_N|||\le[1-\mu^2\kappa(\mathbf{B})]^{\frac12}|||\mathbf{u}-\mathbf{w}|||\},$$

where the last inequality follows from Lemma 4.2 using that $\mu<\kappa(\mathbf{B})^{-\frac12}$. □

The selection of a $\tilde\Lambda$ as in (30) will be performed by a call of the following routine.

EXPAND$[\Lambda,\mathbf{r},\alpha]\to\tilde\Lambda$:
% Input: $\Lambda\subset\nabla$, $\#\Lambda<\infty$, $\mathbf{r}\in\ell_0$, $\alpha\in[0,1]$.

$\bar{\mathbf{r}}:=$ **COARSE**$[\mathbf{r}|_{\nabla\setminus\Lambda},\sqrt{1-\alpha^2}\|\mathbf{r}\|]$
$\tilde\Lambda:=\Lambda\cup\operatorname{supp}\bar{\mathbf{r}}$

[1] Under the milder condition $\gamma<\frac13(\alpha-\delta)\kappa(\mathbf{B})^{-\frac12}$, a more complicated proof ([Gan06, Proposition 3.2.2] or [GHS07, Theorem 2.7]) shows (32) for another $\rho<1$.

Proposition 4.3. *$\tilde\Lambda := \mathbf{EXPAND}[\Lambda,\mathbf{r},\alpha]$ satisfies $\tilde\Lambda \supset \Lambda$, $\|\mathbf{P}_{\tilde\Lambda}\mathbf{r}\| \geq \alpha\|\mathbf{r}\|$, and*

$$\#(\tilde\Lambda\setminus\Lambda) \lesssim \min\{\#(\check\Lambda\setminus\Lambda) : \|\mathbf{P}_{\check\Lambda}\mathbf{r}\| \geq \alpha\|\mathbf{r}\|, \Lambda\subset\check\Lambda\subset\nabla\}.$$

The number of operations used by the call $\mathbf{EXPAND}[\Lambda,\mathbf{r},\alpha]$ is bounded by some absolute multiple of $\#\Lambda + \#\operatorname{supp}\mathbf{r} + 1$.

Proof. We have $\|\mathbf{r}-\mathbf{P}_{\tilde\Lambda}\mathbf{r}\| = \|\mathbf{r}|_{\nabla\setminus\Lambda} - \bar{\mathbf{r}}\| \leq \sqrt{1-\alpha^2}\|\mathbf{r}\|$, which is equivalent to $\|\mathbf{P}_{\tilde\Lambda}\mathbf{r}\| \geq \alpha\|\mathbf{r}\|$. The properties of **COARSE** imply the statement about the work as well as that

$$\begin{aligned}\#(\tilde\Lambda\setminus\Lambda) = \#\operatorname{supp}\bar{\mathbf{r}} &\lesssim \min\{\#\bar\Lambda : \bar\Lambda\subset\nabla\setminus\Lambda, \|\mathbf{r}|_{\nabla\setminus\Lambda} - \mathbf{P}_{\bar\Lambda}(\mathbf{r}|_{\nabla\setminus\Lambda})\| \leq \sqrt{1-\alpha^2}\|\mathbf{r}\|\}\\ &= \min\{\#\bar\Lambda : \bar\Lambda\subset\nabla\setminus\Lambda, \|\mathbf{P}_{\Lambda\cup\bar\Lambda}\mathbf{r}\| \geq \alpha\|\mathbf{r}\|\},\end{aligned}$$

which completes the proof. □

The arising Galerkin systems will be solved approximately by the application of an iterative scheme. Since the (approximate) solution of the previous Galerkin system will be used as the starting vector, a uniformly bounded number of iterations will suffice. Each iteration requires the application of $\mathbf{B}_\Lambda$. Although this matrix is close to being sparse, generally its number of non-zero entries is not of the order of $\#\Lambda$. Therefore, the iterative scheme will be executed only approximately. Below we consider the simplest option of applying an inexact Richardson iteration.

GALERKIN$[\Lambda,\bar{\mathbf{w}}_\Lambda,\delta,\varepsilon] \to \mathbf{w}_\Lambda$:
% Input: $\delta,\varepsilon>0$, $\Lambda\subset\nabla$, $\#\nabla<\infty$, $\bar{\mathbf{w}}_\Lambda\in\ell_2(\Lambda)$ with $\|\mathbf{P}_\Lambda\mathbf{f}-\mathbf{B}_\Lambda\bar{\mathbf{w}}_\Lambda\|\leq\delta$.
% Parameters: $\rho,\alpha,e_0\in\mathbb{R}$, $K\in\mathbb{N}$ such that $\|\mathrm{Id}-\alpha\mathbf{B}\|\leq\rho<1$, $2\rho^K\leq\varepsilon/\delta$,
% and $\|\mathbf{B}\|\leq e_0$.

$\mathbf{v}^{(0)} := \bar{\mathbf{w}}_\Lambda$
`for` $i=1,\dots,K$, `do`
$\quad\mathbf{v}^{(i)} := \mathbf{v}^{(i-1)} + \alpha\mathbf{P}_\Lambda(\mathbf{RHS}[\frac{\rho^i\delta}{2\alpha K e_0}] - \mathbf{APPLY}[\mathbf{v}^{(i-1)},\frac{\rho^i\delta}{2\alpha K e_0}])$
`enddo`
$\mathbf{w}_\Lambda := \mathbf{v}^{(K)}$

The following proposition is essentially [CDD01, Prop. 6.7].

Proposition 4.4. *$\mathbf{w}_\Lambda := \mathbf{GALERKIN}[\Lambda,\bar{\mathbf{w}}_\Lambda,\delta,\varepsilon]$ satisfies $\|\mathbf{P}_\Lambda\mathbf{f}-\mathbf{B}_\Lambda\mathbf{w}_\Lambda\|\leq\varepsilon$. Let $\mathbf{B}$ be $\bar s$-admissible, and for some $s\in(0,\bar s]$, $\mathbf{u}\in\mathscr{A}^s$. Then the cost of the call can be bounded on some absolute multiple of*

$$\eta(\delta/\varepsilon)(\delta^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s} + \delta^{-1/s}\|\bar{\mathbf{w}}_\Lambda\|_{\mathscr{A}^s}^{1/s} + \#\Lambda + 1),$$

where $\eta:(0,\infty)\to[1,\infty)$ is some non-decreasing function.

Proof. For some $\delta_1,\dots,\delta_K\in\ell_2(\Lambda)$ with $\|\delta_i\|\leq\frac{\rho^i\delta}{Ke_0}$,

$$\|\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \mathbf{v}^{(K)}\| = \|(\mathrm{Id} - \alpha \mathbf{B}_\Lambda)^K (\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \mathbf{v}^{(0)}) + \mathbf{B}_\Lambda \sum_{i=1}^{K} (\mathrm{Id} - \alpha \mathbf{B}_\Lambda)^{K-i} \delta_i\|$$

$$\leq \rho^K \delta + e_0 \sum_{i=1}^{K} \rho^{K-i} \frac{\rho^i \delta}{K e_0} \leq \varepsilon$$

(cf. proof of Theorem 3.1). The statement about the cost follows from the $\bar{s}$-admissibility of $\mathbf{B}$ and the assumptions on **RHS**, in particular (16), Corollary 3.1 and Proposition 3.1, as well as from the fact that $K < \infty$ depending on δ/ε. □

Remark 4.2. The above implementation of **GALERKIN** can be improved. Instead of computing $\mathbf{P}_\Lambda \mathbf{RHS}[\eta]$ for a decreasing sequence of η's, it is better to compute once an approximation $\bar{\mathbf{f}}_\Lambda \in \ell_2(\Lambda)$ with $\|\mathbf{P}_\Lambda \mathbf{f} - \bar{\mathbf{f}}_\Lambda\| \leq \eta$ for the final accuracy η (actually, then an even less accurate approximation suffices). Further, instead of approximating the application of $\mathbf{B}_\Lambda$ by using the **APPLY** routine and by afterwards restricting the result to Λ, obviously it is better not to compute any entry with index outside Λ. Also with these improvements, the routine remains quantitatively demanding because of the relatively expensive adaptive approximate matrix vector applications.

A more efficient Galerkin routine can be constructed using a *defect correction* principle. Let $\tilde{\mathbf{B}}_\Lambda$ be a *fixed* sparse matrix with $\|\mathrm{Id} - \mathbf{B}_\Lambda \tilde{\mathbf{B}}_\Lambda^{-1}\| \leq \varepsilon/\delta$. Existence of such a matrix follows by assuming s^*-computability of $\mathbf{B}$. Then

$$\mathbf{w}_\Lambda := \bar{\mathbf{w}}_\Lambda + \tilde{\mathbf{B}}_\Lambda^{-1} (\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \bar{\mathbf{w}}_\Lambda)$$

satisfies

$$\|\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \mathbf{w}_\Lambda\| = \|(\mathrm{Id} - \mathbf{B}_\Lambda \tilde{\mathbf{B}}_\Lambda^{-1})(\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \bar{\mathbf{w}}_\Lambda)\| \leq \tfrac{\varepsilon}{\delta} \delta = \varepsilon.$$

By taking $\tilde{\mathbf{B}}_\Lambda$ to be somewhat more accurate, say with $\|\mathrm{Id} - \mathbf{B}_\Lambda \tilde{\mathbf{B}}_\Lambda^{-1}\| \leq \varepsilon/(2\delta)$, room is left to compute the initial *defect* $\mathbf{P}_\Lambda \mathbf{f} - \mathbf{B}_\Lambda \bar{\mathbf{w}}_\Lambda$ approximately, and to approximate the application of $\tilde{\mathbf{B}}_\Lambda^{-1}$. The first task requires single calls of **RHS** and **APPLY**, and for the second task a few iterations of a fast iterative method can be applied, e.g., the conjugate residual method. Details can be found in [GHS07].

We are ready to formulate the practical AWGM. It works according to the principles outlined in Proposition 4.2. The tasks (30) and (31) are realized by calls of the routines **EXPAND** and **GALERKIN**, respectively. The first task (29) amounts to finding an approximation of the residual of the current iterand with a *relative* error not larger than δ. This will be implemented by initially approximating this residual with an *absolute* tolerance equal to some multiple θ of the norm of the previous residual. If this tolerance turns out to be too large, in the sense that it is not less than δ times the norm of the so computed residual, in an inner loop it is halved until the criterion is met.

In view of obtaining a quantitatively efficient implementation, one would like to choose this θ not too small, but sufficiently small such that "usually" one residual computation suffices. It can, however, never be excluded that by sheer chance at an

early stage the (favourable) situation is encountered that the current iterand $\mathbf{u}^{(i)}$ is equal or exceptionally close to the solution $\mathbf{u}$. Then the algorithm will continue halving the tolerance until the norm of the computed residual plus tolerance is not larger than the target ε, showing that the true residual is not larger than ε. Since in this case there is no point in expanding the index set or computing the Galerkin solution more accurately, this behaviour of the algorithm is desired.

AWGM$[\varepsilon,\varepsilon_{-1}]\to\mathbf{u}_\varepsilon$:
% Input: $\varepsilon,\varepsilon_{-1}>0$.
% Parameters: $\alpha,\delta,\gamma,\theta$ *such that* $\delta\in(0,\alpha)$, $\frac{\alpha+\delta}{1-\delta}<\kappa(\mathbf{B})^{-\frac{1}{2}}$, $\theta>0$ *and*
% $\gamma\in\big(0,\frac{(1-\delta)(\alpha-\delta)}{1+\delta}\kappa(\mathbf{B})^{-1}\big)$.

$i:=0$, $\mathbf{u}^{(i)}:=0$, $\Lambda_i:=\emptyset$
`do` $\zeta:=\theta\varepsilon_{i-1}$
`do` $\zeta:=\zeta/2$, $\mathbf{r}^{(i)}:=\mathbf{RHS}[\zeta/2]-\mathbf{APPLY}[\mathbf{u}^{(i)},\zeta/2]$
`if` $\varepsilon_i:=\|\mathbf{r}^{(i)}\|+\zeta\le\varepsilon$ `then` $\mathbf{u}_\varepsilon:=\mathbf{u}^{(i)}$ `stop endif`
`until` $\zeta\le\delta\|\mathbf{r}^{(i)}\|$
$\Lambda_{i+1}:=\mathbf{EXPAND}[\Lambda_i,\mathbf{r}^{(i)},\alpha]$
$\mathbf{u}^{(i+1)}:=\mathbf{GALERKIN}[\Lambda_{i+1},\mathbf{u}^{(i)},\varepsilon_i,\gamma\|\mathbf{r}^{(i)}\|]$
$i:=i+1$
`enddo`

Theorem 4.1 ([GHS07]). *Let* $\varepsilon_{-1},\varepsilon>0$, *then for* $\mathbf{u}_\varepsilon:=\mathbf{AWGM}[\varepsilon,\varepsilon_{-1}]$ *we have that* $\|\mathbf{f}-\mathbf{B}\mathbf{u}_\varepsilon\|\le\varepsilon$. *If for some* $s>0$, $\mathbf{u}\in\mathscr{A}^s$, *then* $\#\operatorname{supp}\mathbf{u}_\varepsilon\lesssim\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *If, additionally,* $\mathbf{B}$ *is* $\bar{s}$*-admissible,* $s\le\bar{s}$ *and* $\varepsilon\lesssim\varepsilon_{-1}\eqsim\|\mathbf{f}\|$, *then the number of operations used by the call* $\mathbf{AWGM}[\varepsilon,\varepsilon_0]$ *is bounded by an absolute multiple of* $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *In other words, if* $\bar{s}\ge s_{\max}$, *then this AWGM is (quasi-) optimal.*

Proof. By definition of ε_i, we have

$$\|\mathbf{f}-\mathbf{B}\mathbf{u}^{(i)}\|\le\varepsilon_i, \tag{34}$$

so that ε_i is a valid parameter for the later call $\mathbf{GALERKIN}[\Lambda_{i+1},\mathbf{u}^{(i)},\varepsilon_i,\gamma\|\mathbf{r}^{(i)}\|]$.

Since ζ is halved in each iteration, if the inner loop does not terminate because of $\zeta\le\delta\|\mathbf{r}^{(i)}\|$, then at some point it will terminate because of $\varepsilon_i\le\varepsilon$.

If the inner loop terminates because of $\zeta\le\delta\|\mathbf{r}^{(i)}\|$, then, because of $\delta<1$,

$$\varepsilon_i\eqsim\|\mathbf{r}^{(i)}\|\eqsim\|\mathbf{f}-\mathbf{B}\mathbf{u}^{(i)}\| \tag{35}$$

and $\|\mathbf{f}-\mathbf{B}\mathbf{u}^{(i)}-\mathbf{r}^{(i)}\|\le\zeta\le\delta\|\mathbf{r}^{(i)}\|$. Since after the subsequent calls of **EXPAND** and **GALERKIN**, $\|\mathbf{P}_{\Lambda_{i+1}}\mathbf{r}^{(i)}\|\ge\alpha\|\mathbf{r}^{(i)}\|$ and $\|\mathbf{P}_{\Lambda_{i+1}}\mathbf{f}^{(i)}-\mathbf{B}_{\Lambda_{i+1}}\mathbf{u}^{(i+1)}\|\le\gamma\|\mathbf{r}^{(i)}\|$, an application of Proposition 4.2 shows that, with $\rho<1$ from that proposition,

$$|||\mathbf{u}-\mathbf{u}^{(i+1)}|||\le\rho|||\mathbf{u}-\mathbf{u}^{(i)}||| \tag{36}$$

and

$$\#(\Lambda_{i+1}\backslash\Lambda_i)\lesssim\min\{N:|||\mathbf{u}-\mathbf{u}_N|||\le[1-\mu^2\kappa(\mathbf{B})]^{\frac{1}{2}}|||\mathbf{u}-\mathbf{u}^{(i)}|||\}. \tag{37}$$

Because $\varepsilon_i = \|\mathbf{r}^{(i)}\| + \zeta \le \|\mathbf{f} - \mathbf{B}\mathbf{u}^{(i)}\| + 2\zeta \le \|\mathbf{f} - \mathbf{B}\mathbf{u}^{(i)}\| + 2\theta\varepsilon_{i-1}$, from (35) and (36), we conclude that eventually the inner loop, and thus the algorithm, will terminate because of $\varepsilon_i \le \varepsilon$. By (34), this proves the first statement of the theorem.

Fully analogous to the proof of Proposition 4.1, from (36) and (37) we conclude that if for some $s > 0$, $\mathbf{u} \in \mathscr{A}^s$, then

$$\#\operatorname{supp}\mathbf{u}^{(i+1)} \le \#\Lambda_{i+1} \lesssim \|\mathbf{u} - \mathbf{u}^{(i)}\|^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}. \tag{38}$$

With K denoting the value of i at termination, i.e., $\mathbf{u}_\varepsilon = \mathbf{u}^{(K)}$, if $K = 0$ then $\#\operatorname{supp}\mathbf{u}_\varepsilon = 0$, and the second statement of the theorem is obviously true. If $K > 0$, then this second statement follows from $\varepsilon < \varepsilon_{K-1} \eqsim \|\mathbf{u} - \mathbf{u}^{(K-1)}\|$ and (38). Together with Lemma 1.1, the same arguments also show that

$$\|\mathbf{u}^{(i)}\|_{\mathscr{A}^s} \lesssim \|\mathbf{u}\|_{\mathscr{A}^s}. \tag{39}$$

Now let $\mathbf{B}$ be $\bar{s}$-admissible for some $\bar{s} \ge s$, and let $\varepsilon \lesssim \varepsilon_{-1} \eqsim \|\mathbf{f}\|$. By definition of $\bar{s}$-admissibility and Corollary 3.1, with C_i denoting the cost of the evaluation of $\mathbf{r}^{(i)} := \mathbf{RHS}[\zeta/2] - \mathbf{APPLY}[\mathbf{u}^{(i)}, \zeta/2]$, we have

$$\begin{aligned}\#\operatorname{supp}\mathbf{r}^{(i)} \lesssim C_i &\lesssim (\zeta/2)^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s} + 1 + (\zeta/2)^{-1/s}\|\mathbf{u}^{(i)}\|_{\mathscr{A}^s}^{1/s} + \#\operatorname{supp}\mathbf{u}^{(i)} + 1\\ &\lesssim \zeta^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s} + \varepsilon_{i-1}^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s},\end{aligned} \tag{40}$$

by (39) and, for $i > 1$, by (38), (35) and $\varepsilon_{i-1} \lesssim \varepsilon_0 \eqsim \|\mathbf{f}\| \lesssim \|\mathbf{u}\|_{\mathscr{A}^s}$ (and thus $1 \lesssim \varepsilon_{i-1}^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$), and, for $i = 0$, by $\#\operatorname{supp}\mathbf{u}^{(0)} = 0$ and $\varepsilon_{-1} \lesssim \|\mathbf{f}\| \lesssim \|\mathbf{u}\|_{\mathscr{A}^s}$.

To proceed, we claim that for $0 \le i < K$, at termination of the inner loop, $\zeta \gtrsim \varepsilon_i$. Indeed, if the inner loop terminates at the first evaluation of the `until`-clause, then $\zeta = \theta\varepsilon_{i-1} \gtrsim \varepsilon_i$, the latter for $i = 0$ being valid by assumption. Otherwise, at the previous evaluation of the `until`-clause, we had $\|\mathbf{f} - \mathbf{B}\mathbf{u}^{(i)}\| \le \|\mathbf{r}^{(i)}\| + \zeta < (\delta^{-1} + 1)\zeta$. Since this ζ is twice the final one, (35) shows that the latter satisfies $\zeta \gtrsim \varepsilon_i$.

From the above claim, (40) and the successive halvings of ζ starting from $\zeta = \theta\varepsilon_{i-1}$, we conclude that for $0 \le i < K$, at termination of the inner loop

$$\#\operatorname{supp}\mathbf{r}^{(i)} \lesssim \bar{C}_i \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s},$$

where $\bar{C}_i$ denotes the *total* cost of the inner loop that produced this $\mathbf{r}^{(i)}$.

Propositions 4.3 and 4.4 show that the cost of subsequent calls of **EXPAND** and **GALERKIN** is bounded by an absolute multiple of $\#\Lambda_i + \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s} \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$ and, since $\varepsilon_i \lesssim \gamma\|\mathbf{r}^{(i)}\|$, of $\#\Lambda_{i+1} \lesssim \varepsilon_i^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, respectively.

From $\varepsilon_i \lesssim \rho^{i-j}\varepsilon_j$ $(i \le j)$, being a consequence of (36) and (35), and, when $K > 0$, $\varepsilon_{K-1} > \varepsilon$, we may conclude that the total cost of the call $\mathbf{AWGM}[\varepsilon, \varepsilon_0]$ is bounded by an absolute multiple of $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, once we have shown that the cost of the final run of the inner loop can be bounded an absolute multiple of this expression. For this goal, it suffices to show that at termination of this last inner loop, $\zeta \gtrsim \varepsilon$.

If this inner loop terminates by the first evaluation of the `if`-clause, then $\zeta = \theta\varepsilon_{K-1} \gtrsim \varepsilon$, for $K = 0$ by assumption. Otherwise, the previous value of ζ, being twice the final one, satisfies both $\|\mathbf{r}^{(i)}\| + \zeta > \varepsilon$ and $\zeta \geq \delta\|\mathbf{r}^{(i)}\|$, and so $(1+\delta^{-1})\zeta > \varepsilon$, with which the proof is completed. □

4.3 Discussion

As we have seen, both the adaptive inexact Richardson scheme **Rich** from Sect. 3 and the Adaptive Wavelet Galerkin Method **AWGM** discussed in the present section are (quasi-) optimal. Practical experiments, see [GHS07] and [DHS07, Sect. 4], show that the **AWGM** is quantitatively more efficient. One reason could be the need for coarsening in **Rich**. Indeed, without coarsening generally this algorithm turns out not to be (quasi-) optimal. This means that in between two coarsening steps, the error as function of the support size does not decay with the optimal rate. As a consequence, in each coarsening step many previously computed coefficients are thrown away. Another possible explanation is that in both algorithms, the expansion of the current wavelet index set via an approximate residual computation is the most costly part. In view of this, given such an index set, it seems most efficient to compute a (near) best approximation from the span of the corresponding wavelets, being the Galerkin approach.

Apart from the aforementioned references, practical experiments with (variants of) the **AWGM** can be found in [BBC+01, Bar01, BK06, BK08]. Numerical results with (variants of) the adaptive inexact Richardson scheme applied to the Schur complement of the Stokes equations (Uzawa scheme) can be found in [DUV02].

5 The approximation of operators in wavelet coordinates by computable sparse matrices

From the main theorems 3.1 and 4.1, recall that the inexact Richardson iteration **Rich** and the Adaptive Wavelet Galerkin Method **AWGM** applied to $\mathbf{Bu} = \mathbf{f}$ are (quasi-) optimal under the condition that $\mathbf{B}$ is $\bar{s}$-admissible (cf. Definition 3.1) for some $\bar{s} \geq s_{\max}$. Consequently, if either of the adaptive wavelet schemes is applied to the normal equations, both $\mathbf{B}$ and $\mathbf{B}^\top$ have to be $\bar{s}$-admissible for some $\bar{s} \geq s_{\max}$. With B being a boundedly invertible operator between $\mathcal{X}$ and $\mathcal{Y}'$, recall that $s_{\max}$ is the generally best possible approximation rate from $\operatorname{span}\Psi^{\mathcal{X}}$ of a function in $\mathcal{X}$. Furthermore, from Theorem 3.2, recall that if $\mathbf{B}$ is s^*-computable (cf. Definition 3.2), then it is $\bar{s}$-admissible for any $\bar{s} < s^*$. In view of these results, our task is therefore to show s^*-computability of $\mathbf{B}$ and possibly $\mathbf{B}^\top$ for some $s^* > s_{\max}$.

The question whether $\mathbf{B}$ (and $\mathbf{B}^\top$) is s^*-computable for some $s^* > s_{\max}$ depends on the operator B and the wavelets at hand. So far, apart from the boundedly invertibility of B, we only assumed that $\Psi^{\mathcal{X}}$ and $\Psi^{\mathcal{Y}}$ are Riesz bases for $\mathcal{X}$ and $\mathcal{Y}$,

respectively. In this section, we study the issue of s^*-computability for B resulting from a *scalar PDE* or a *system of PDE's* on a domain $\Omega \subset \mathbb{R}^n$, and for $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{Y}}$ being collections of commonly applied, locally supported, piecewise smooth wavelets. In Section 7, we comment on the case of $\Psi^{\mathscr{X}}$ and $\Psi^{\mathscr{Y}}$ being collections of tensor product wavelets.

Also for classes for *singular integral operators* and suitable wavelets s^*-computability with $s^* > s_{\max}$ is valid. We refer to [Ste04, GS06b, DHS07] and on the chapter "Rapid Solution of Boundary Integral Equations" by H. Harbrecht and R. Schneider in this book.

5.1 Near-sparsity of partial differential operators in wavelet coordinates

This subsection is devoted to the question how well the representation $\mathbf{B}$ of a partial differential operator with respect to wavelet bases can be approximated by sparse matrices. We will not be concerned with the question how to compute, or more generally, how to approximate the entries of these sparse matrices, and at which cost. These issues will be postponed to the next subsections. Our current task motivates the following definition.

Definition 5.1. For $s^* > 0$, $\mathbf{B} \in \mathscr{L}(\ell_2, \ell_2)$ will be called to be s^*-*compressible* when we have available sequences $(e_j)_{j\in\mathbb{N}_0}, (c_j)_{j\in\mathbb{N}_0} \subset \mathbb{R}$, $(\mathbf{B}^{(j)})_{j\in\mathbb{N}_0} \subset \mathscr{L}(\ell_2, \ell_2)$, such that

- $\|\mathbf{B} - \mathbf{B}^{(j)}\| \le e_j$, $\lim_{j\to\infty} e_j = 0$,
- the number of non-zero entries in each column of $\mathbf{B}^{(j)}$ is bounded by c_j,
- $\mathbf{B}^{(0)} = 0$ (and thus $\|\mathbf{B}\| \le e_0$), $c_0 = 0$ and $\sup_{j\in\mathbb{N}_0} c_{j+1}/c_j < \infty$.

and such that for any $s < s^*$, $\sup_j e_j c_j^s < \infty$.

So compared to the definition of s^*-computability (Definition 3.2), the only difference is that we do not require that number of operations needed to *compute* the non-zero entries in each column of $\mathbf{B}^{(j)}$ is bounded by c_j.

For some $\alpha_l \in \mathbb{N}_0^n$ ($l \in \{1,2\}$), we consider the representation as a bi-infinite matrix of a bounded linear operator $E : H_0^{|\alpha_1|}(\Omega) \to (H_0^{|\alpha_2|}(\Omega))'$ defined by

$$(Eu_1)(u_2) = \int_\Omega g \partial^{\alpha_1} u_1 \partial^{\alpha_2} u_2 \quad (u_l \in H_0^{|\alpha_l|}(\Omega)),$$

with respect to wavelet collections

$$\Psi^{(l)} = \{\psi_\lambda^{(l)} : \lambda \in \nabla\} \subset H_0^{|\alpha_l|}(\Omega).$$

We will assume that the coefficient g is *sufficiently smooth*.

In this paper, we do not discuss the *construction* of wavelet bases on domains, but refer to the numerous papers written on that topic. Some references are included at the end of this subsection. Following standard conventions, $|\lambda| \in \mathbb{N}_0$ will denote the *level* of the wavelet $\psi_\lambda^{(l)}$. Here thinking of the wavelets being *normalized in* $L_2(\Omega)$ and constructed using *dyadic* dilations, for $s \geq 0$ up to some upper bound determined by the smoothness of the wavelets, it holds that $\|\psi_\lambda^{(l)}\|_{H^s(\Omega)} \eqsim 2^{s|\lambda|}$. In view of this, we investigate the approximation of

$$\mathbf{E} := \big[2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}(A\psi_\mu^{(1)})(\psi_\lambda^{(2)})\big]_{\lambda,\mu\in\nabla}$$

by sparse matrices.

The representation of a scalar PDE will result into a sum of such matrices (where, because of the eventual normalization of the wavelets in higher order Sobolev norms, matrices corresponding to lower order terms will be multiplied from left and right by $\mathrm{diag}[2^{-|\lambda|s}]_{\lambda\in\nabla}$ or $\mathrm{diag}[2^{-|\lambda|t}]_{\lambda\in\nabla}$ for some $s,t \geq 0$ with $s+t>0$) and the representation of a system of PDE's will consist of blocks, each of them being a sum of such matrices.

We will assume that the wavelets are *local*, *locally finite* and *piecewise smooth*, where for an easy treatment of the quadrature issue, we assume that the wavelets from both collections are piecewise smooth with respect to the *same* partitions, moreover which are *nested* as function of the level (for the general case, see [SW08]): We assume that for all $k \in \mathbb{N}_0$, there exists a collection $\{\Omega_k^{(\nu)} : \nu \in \mathscr{O}_k\}$ of disjoint, uniformly shape regular, open subdomains, with $\overline{\Omega} = \cup_{\nu\in\mathscr{O}_k}\overline{\Omega_k^{(\nu)}}$, $\mathrm{diam}(\Omega_k^{(\nu)}) \eqsim 2^{-k}$ and $\Omega_k^{(\nu)}$ being the union of some $\Omega_{k+1}^{(\tilde{\nu})}$. These subdomains will be such that $\mathrm{supp}\,\psi_\lambda^{(l)}$ ($l \in \{1,2\}$), which is assumed to be connected, is the union of a uniformly bounded number of $\Omega_{|\lambda|}^{(\nu)}$ (*locality*), and such that each $\Omega_k^{(\nu)}$ has non-empty intersection with the supports of a uniformly bounded number of $\psi_\lambda^{(l)}$ with $|\lambda| = k$ (*locally finiteness*). Typical examples of the $\Omega_k^{(\nu)}$ are n-cubes or n-simplices, or smooth images of such. We assume that $\psi_\lambda^{(l)}|_{\Omega_{|\lambda|}^{(\nu)}}$ is smooth with, for any $\gamma \in \mathbb{N}_0^n$,

$$\sup_{x\in\Omega_{|\lambda|}^{(\nu)}} |\partial^\gamma \psi_\lambda^{(l)}(x)| \lesssim 2^{|\lambda|(\frac{n}{2}+|\gamma|)} \tag{41}$$

(*piecewise smoothness*), the latter being a consequence of the smoothness of the function $\psi_\lambda^{(l)}|_{\Omega_{|\lambda|}^{(\nu)}}$, the normalization of the wavelets in $L_2(\Omega)$ and their construction using dyadic dilations. Note that the singular support of $\psi_\lambda^{(l)}$ is part of the skeleton $\cup_{\nu\in\mathscr{O}_k}\partial\Omega_{|\lambda|}^{(\nu)}$.

We will also need that the wavelets satisfy some *global smoothness* conditions: For some

$$\mathbb{N}_0 \cup \{-1\} \ni r_l \geq |\alpha_l| - 1,$$

we assume that

$$\|\psi_\lambda^{(l)}\|_{W_\infty^t(\Omega)} \lesssim 2^{|\lambda|(\frac{n}{2}+t)} \quad (t \in [0, r_l+1]). \tag{42}$$

For $r_l > -1$, this estimate follows from (41) when $\psi_\lambda^{(l)} \in C^{r_l}(\Omega)$.

We assume that the wavelets have *cancellation properties of order* $\tilde{d}_l \in \mathbb{N}_0$, meaning that

$$\Big|\int_\Omega u\psi_\lambda^{(l)}\Big| \lesssim 2^{-|\lambda|t}\|u\|_{W_\infty^t(\operatorname{supp}\psi_\lambda^{(l)})}\|\psi_\lambda^{(l)}\|_{L_1(\Omega)} \quad (t \in [0,\tilde{d}_l], u \in W_\infty^t(\Omega)). \tag{43}$$

Actually, with some constructions, here $\operatorname{supp}\psi_\lambda^{(l)}$ should read as a neighbourhood of $\operatorname{supp}\psi_\lambda^{(l)}$ with diameter $2^{-|\lambda|}$. For convenience we ignore this fact, but our results extend trivially to this situation.

Finally, for any $\gamma \le \alpha_l$, $\gamma \ne \alpha_l$, we assume the homogeneous Dirichlet boundary conditions

$$\partial^\gamma \psi_\lambda^{(l)} = 0 \quad \text{at } \partial\Omega, \tag{44}$$

actually being a consequence of our earlier assumption that $\Psi^{(l)} \subset H_0^{|\alpha_l|}(\Omega)$.

We split

$$\mathbf{E} = \mathbf{E}^{(r)} + \mathbf{E}^{(s)},$$

where $\mathbf{E}^{(r)}_{\lambda,\mu} = \mathbf{E}_{\lambda,\mu}$ when either $|\lambda| > |\mu|$ and $\operatorname{supp}\psi_\lambda^{(2)} \subset \bar{\Omega}_{|\mu|}^{(\nu)}$ for some $\nu \in \mathscr{O}_{|\mu|}$ or $|\lambda| < |\mu|$ and $\operatorname{supp}\psi_\mu^{(1)} \subset \bar{\Omega}_{|\lambda|}^{(\nu)}$ for some $\nu \in \mathscr{O}_{|\lambda|}$, and $\mathbf{E}^{(r)}_{\lambda,\mu}$ is zero otherwise. So $\mathbf{E}^{(r)}$ contains the *regular entries* of $\mathbf{E}$, i.e., the non-zero entries for which the the interior of the support of the wavelet on the higher level does not intersect the singular support of the wavelet on the lower level. The remaining *singular entries* are gathered in $\mathbf{E}^{(s)}$. As we will see, the size of the singular entries decays less fast as function of the difference in the levels of the indices than with the regular entries, but this will be compensated by a smaller increase of their number.

We write $\mathbf{E}^{(r)} = (\mathbf{E}^{(r)}_{\ell,k})_{\ell,k\in\mathbb{N}_0}$, where $\mathbf{E}^{(r)}_{\ell,k} = (\mathbf{E}^{(r)}_{\lambda,\mu})_{|\lambda|=\ell,|\mu|=k}$ and similarly $\mathbf{E}^{(s)} = (\mathbf{E}^{(s)}_{\ell,k})_{\ell,k\in\mathbb{N}_0}$.

Proposition 5.1. *The number of non-zero entries in each row of* $\mathbf{E}^{(r)}_{\ell,k}$ $(\mathbf{E}^{(s)}_{\ell,k})$ *or each column of* $\mathbf{E}^{(r)}_{k,\ell}$ $(\mathbf{E}^{(s)}_{k,\ell})$ *is bounded by an absolute multiple of* $2^{\max(k-\ell,0)n}$ $\big(2^{\max(k-\ell,0)(n-1)}\big)$.

With

$$\rho_r := \tilde{d}_2 + |\alpha_2|, \quad \rho_s := \tfrac{1}{2} + \min(\tilde{d}_2 + |\alpha_2|, r_1 + 1 - |\alpha_1|),$$

for $|\lambda| > |\mu|$, *we have*

$$|\mathbf{E}^{(s)}_{\lambda,\mu}| \lesssim \|g\|_{W_\infty^{\rho_s-1/2}(\Omega)} 2^{-\left(\left||\lambda|-|\mu|\right|\right)(\frac{n-1}{2}+\rho_s)}, \quad |\mathbf{E}^{(r)}_{\lambda,\mu}| \lesssim \|g\|_{W_\infty^{\rho_r}(\Omega)} 2^{-\left(\left||\lambda|-|\mu|\right|\right)(\frac{n}{2}+\rho_r)}$$

The same statement is valid for $|\lambda| < |\mu|$ *when* $(\alpha_1, \alpha_2, r_1, \tilde{d}_2)$ *is replaced by* $(\alpha_2, \alpha_1, r_2, \tilde{d}_1)$ *in the definitions of* ρ_r *and* ρ_s.

Proof. The first statement follows by the localness and locally finiteness of both wavelet collections, and concerning $\mathbf{E}^{(s)}$, by their piecewise smoothness.

When $r_1+1 \le |\alpha_1|+|\alpha_2|$, select a $\gamma \le \alpha_2$ with $|\alpha_1+\gamma| = r_1+1$ and so $|\alpha_2-\gamma| = |\alpha_1|+|\alpha_2|-(r_1+1)$. Using (44) for the case that $\operatorname{supp}\psi_\lambda^{(2)} \cap \operatorname{supp}\psi_\mu^{(1)} \cap \partial\Omega \neq \emptyset$, integration by parts, $\operatorname{vol}(\operatorname{supp}\psi_\lambda^{(2)}) \lesssim 2^{-|\lambda|n}$ and (42) show that

$$\begin{aligned}
|\mathbf{E}_{\lambda,\mu}| &= 2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}\Big|\int_{\operatorname{supp}\psi_\lambda^{(2)}} (-1)^{|\gamma|}\partial^\gamma(g\partial^{\alpha_1}\psi_\mu^{(1)})\partial^{\alpha_2-\gamma}\psi_\lambda^{(2)}\Big| \\
&\lesssim 2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}\|g\|_{W_\infty^{r_1+1-|\alpha_1|}(\Omega)} 2^{-|\lambda|n} 2^{|\mu|(\frac{n}{2}+r_1+1)} 2^{|\lambda|(\frac{n}{2}+|\alpha_1|+|\alpha_2|-(r_1+1))} \\
&= \|g\|_{W_\infty^{r_1+1-|\alpha_1|}(\Omega)} 2^{-(|\lambda|-|\mu|)(\frac{n}{2}+r_1+1-|\alpha_1|)}.
\end{aligned}$$

For $r_1+1 > |\alpha_1|+|\alpha_2|$ by additionally using that the $\psi_\lambda^{(2)}$ have $\tilde{d}_2$ vanishing moments ((43)) and taking into account that $\psi_\mu^{(1)} \in W_\infty^{r_1+1}(\Omega)$ ((42)), we have

$$\begin{aligned}
|\mathbf{E}_{\lambda,\mu}| &= 2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}\Big|\int_{\operatorname{supp}\psi_\lambda^{(2)}} (-1)^{|\alpha_2|}\partial^{\alpha_2}(g\partial^{\alpha_1}\psi_\mu^{(1)})\psi_\lambda^{(2)}\Big| \\
&\lesssim 2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}2^{-|\lambda|\min(\tilde{d}_2, r_1+1-|\alpha_1|-|\alpha_2|)} \\
&\qquad \times \|\partial^{\alpha_2}(g\partial^{\alpha_1}\psi_\mu^{(1)})\|_{W_\infty^{\min(\tilde{d}_2, r_1+1-|\alpha_1|-|\alpha_2|)}(\operatorname{supp}\psi_\lambda^{(2)})}\|\psi_\lambda^{(2)}\|_{L_1(\Omega)} \\
&\lesssim 2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|}2^{-|\lambda|\min(\tilde{d}_2, r_1+1-|\alpha_1|-|\alpha_2|)} \\
&\qquad \times \|g\|_{W_\infty^{\min(\tilde{d}_2+|\alpha_2|, r_1+1-|\alpha_1|)}(\Omega)} 2^{|\mu|(\frac{n}{2}+\min(\tilde{d}_2+|\alpha_1|+|\alpha_2|, r_1+1))} 2^{-|\lambda|\frac{n}{2}} \\
&\eqsim \|g\|_{W_\infty^{\min(\tilde{d}_2+|\alpha_2|, r_1+1-|\alpha_1|)}(\Omega)} 2^{-(|\lambda|-|\mu|)(\frac{n}{2}+\min(\tilde{d}_2+|\alpha_2|, r_1+1-|\alpha_1|))},
\end{aligned}$$

which completes the proof of the first estimate.

Finally, when $\operatorname{supp}\psi_\lambda^{(2)} \subset \bar{\Omega}_{|\mu|}^{(\nu)}$ for some $\nu \in \mathscr{O}_{|\mu|}$, by estimating $\mathbf{E}_{\lambda,\mu}^{(r)}$ as above, but now applying (41) for sufficiently large γ instead of (42), we obtain the second estimate. □

In the next proposition, we construct sparse approximation for matrices like $\mathbf{E}^{(r)}$ or $\mathbf{E}^{(s)}$.

Proposition 5.2. *Let* $\mathbf{C} = (\mathbf{C}_{\ell,k})_{\ell,k\in\mathbb{N}_0}$ *with* $\mathbf{C}_{\ell,k} = (\mathbf{C}_{\lambda,\mu})_{|\lambda|=\ell,|\mu|=k}$ *be such that for some* $q \in \mathbb{N}_0$ *and* $\rho > 0$, *the number of non-zero entries in each row of* $\mathbf{C}_{\ell,k}$ *or column of* $\mathbf{C}_{k,\ell}$ *is bounded by an absolute multiple of* $2^{\max(k-\ell,0)q}$ *and*

$$|\mathbf{C}_{\lambda,\mu}| \lesssim 2^{-\big||\lambda|-|\mu|\big|(\frac{q}{2}+\rho)}.$$

Then with $\mathbf{C}^{(j)}$ *constructed from* $\mathbf{C}$ *by dropping* $\mathbf{C}_{\lambda,\mu}$ *when*

$$\big||\lambda|-|\mu|\big| > j/\rho,$$

we have

$$\|\mathbf{C}-\mathbf{C}^{(j)}\| \lesssim 2^{-j},$$

where the number of non-zero entries per row and column of $\mathbf{C}^{(j)}$ *is bounded by some absolute multiple of*

$$\max(2^{qj/\rho}, j/\rho).$$

Proof. By two applications of the Schur lemma, we have

$$\|\mathbf{C}_{\ell,k}\|^2 \le \max_{|\lambda|=\ell} \sum_{|\mu|=k} |\mathbf{C}_{\lambda,\mu}| \cdot \max_{|\mu|=k} \sum_{|\lambda|=\ell} |\mathbf{C}_{\lambda,\mu}| \lesssim 4^{-(\ell-m)\rho},$$

$$\|\mathbf{C}-\mathbf{C}^{(j)}\|^2 \le \max_{\ell} \sum_{\{k:|\ell-k|>j/\rho\}} \|\mathbf{C}_{\ell,k}\| \cdot \max_{k} \sum_{\{\ell:|\ell-k|>j/\rho\}} \|\mathbf{C}_{\ell,k}\| \lesssim 4^{-j}. \qquad \square$$

So the result of the last proposition shows that $\mathbf{C}$ and $\mathbf{C}^\top$ are s^**-compressible* with

$$s^* = \rho/q$$

(or $s^* = \infty$ when $q = 0$). We exemplify our findings concerning s^*-compressibility in the model case of an (elliptic) scalar PDE of order $2m$:

Example 5.1. For some bounded domain $\Omega \subset \mathbb{R}^n$, with $n \ge 2$, and $m \in \mathbb{N}$, let $B : H_0^m(\Omega) \to H_0^m(\Omega)'$ be defined as

$$(Bu)(v) = \int_\Omega \sum_{|\alpha|,|\beta|\le m} a_{\alpha,\beta} \partial^\alpha u \partial^\beta v,$$

with coefficients such that B is boundedly invertible and that are sufficiently smooth.

Let $\Psi^{\mathscr{X}} = \Psi^{\mathscr{Y}} = \{\psi_\lambda : \lambda \in \nabla\} \subset H_0^m(\Omega)$ be a dyadic wavelet collection, normalized in $L_2(\Omega)$, such that for some $\mathbb{N} \ni d > m$, $\tilde{d} \in \mathbb{N}_0$, $\mathbb{N}_0 \cup \{-1\} \ni r \ge m-1$,

a). $\inf_{v_i \in \mathrm{span}\{\psi_\lambda : |\lambda| \le i\}} \|u - v_i\|_{H^m(\Omega)} \lesssim 2^{-(d-m)i} \|u\|_{H^d(\Omega)} \quad (u \in H^d(\Omega) \cap H_0^m(\Omega))$,
b). the wavelets are local, locally finite and piecewise smooth (and thus satisfy (41)),
c). the wavelets are in $C^r(\Omega)$ (and thus satisfy (42) with $r_l = r$),
d). the wavelets have cancellation properties of order $\tilde{d}$ ((43)),
e). $\{2^{-|\lambda|m}\psi_\lambda : \lambda \in \nabla\}$ is a Riesz basis for $H_0^m(\Omega)$.

The representation of B with respect to the wavelet basis from e) reads as

$$\mathbf{B} := \sum_{|\alpha|,|\beta|\le m} \big[2^{-(|\lambda|+|\mu|)m} \int_\Omega a_{\alpha,\beta} \partial^\alpha \psi_\mu \partial^\beta \psi_\lambda\big]_{\lambda,\mu\in\nabla}.$$

Due to the scaling factor $2^{-(|\lambda|+|\mu|)m}$, one may verify that it suffices to analyze the s^*-compressibilty of the highest order terms. By applying Propositions 5.1 and 5.2 to those terms, we infer that $\mathbf{B}$ and $\mathbf{B}^\top$ are s^*-compressible with

$$s^* = \min\Big(\frac{\tilde{d}+m}{n}, \frac{\frac{1}{2}+\min(\tilde{d}+m, r+1-m)}{n-1}\Big).$$

As a consequence of the dyadic construction, we have that $\#\{\lambda \in \nabla : |\lambda| \le i\} \eqsim 2^{ni}$, which together with a) shows that

$$s_{\max} = \frac{d-m}{n}.$$

We conclude that $s^* > s_{\max}$ when $\tilde{d} > d-2m$ and $\frac{r+\frac{3}{2}-m}{n-1} > \frac{d-m}{n}$ (the third condition $\tilde{d}+m \ge r+1-m$ follows already from the first one using that always $r \le d-2$).

On $(0,1)^n$, or on a smooth image of that, *biorthogonal spline* wavelets can be constructed that satisfy a)-e) for arbitrary $\tilde{d} \ge d$ with $d+\tilde{d}$ even and $r = d-2$ ([DS98]). Because of $r = d-2$, the conditions for $s^* > s_{\max}$ read as $\tilde{d} > d-2m$ and $\frac{d-m}{n} > \frac{1}{2}$.

For general domains, these wavelets can be applied in combination with non-overlapping *domain decomposition* techniques. The existing techniques fall into 2 categories: With the technique based on *extension operators* proposed in [DS99b], all above conditions can be satisfied. The condition number of the resulting basis, however, turns out to increase rapidly with d. The other technique amounts to a *continuous* gluing of multiresolution analyses over the interfaces between patches, see [DS99a, CTU99]. As a result, wavelets with supports that extend to more than one patches are only continuous, and thus for $d > 2$ not in C^{d-2}, resulting in a reduced value of s^*. For problems of order $2m = 2$, this limitation can be overcome with a construction of wavelets that have *patchwise vanishing moments*, see [HS06].

5.2 The approximate computation of the significant entries

For a non-constant coefficient g, generally the entries of $\mathbf{E}^{(r)}$ and $\mathbf{E}^{(s)}$ have to be approximated by suitable quadrature. In this subsection, we show that such approximations can be made that keep the error on the same level, while taking in each row and column *on average* $\mathcal{O}(1)$ operations per entry. This means these matrices are s^*-computable for the same value of s^* as they were shown to be s^*-compressible. The key observation is that this restriction on the work does allow to spend quite some operations, up to the number of entries in the row or column, to the approximation of the few largest entries with indices that have equal level, as long as the work per entry decays sufficiently fast as function of the difference in the levels of the indices. For simplicity, we exclude the special, although easy case that $q = 0$ in Proposition 5.2. Since with $\mathbf{E}^{(s)}$ the role of q is played by $n-1$, we thus assume that $n > 1$.

Proposition 5.3. *Let* $\mathbf{C}$ *and* $\mathbf{C}^{(j)}$ *be as in Proposition 5.2 assuming that* $q > 0$. *Suppose that for some constants* $\xi, \omega > 0$, $\xi \ne \omega$, *for any* $\lambda, \mu \in \nabla$ *one can compute an approximation* $\tilde{\mathbf{C}}_{\lambda,\mu}$ *to* $\mathbf{C}_{\lambda,\mu}$ *in* $\mathcal{O}(N)$ *operations with*

$$|\mathbf{C}_{\lambda,\mu} - \tilde{\mathbf{C}}_{\lambda,\mu}| \lesssim N^{-\omega} 2^{-\left||\lambda|-|\mu|\right|(\frac{q}{2}+\xi q)}. \tag{45}$$

Now for some $\sigma \in (1, \xi/\omega)$ *when* $\xi > \omega$, *and* $\sigma \in (\xi/\omega, 1)$ *when* $\xi < \omega$, *and* $\theta \le \min(1,\sigma)$, *build* $\tilde{\mathbf{C}}^{(j)}$ *by approximating each non-zero entry of* $\mathbf{C}^{(j)}$ *as above by taking*

$$N = N_{j,\lambda,\mu} \eqsim \max\left(1, 2^{qj\theta/\rho - \left||\lambda|-|\mu|\right|\sigma q}\right)$$

operations. Then the work for computing each row or column of $\tilde{\mathbf{C}}^{(j)}$ *is bounded by some absolute multiple of* $2^{qj/\rho}$, *and*

$$\|\mathbf{C}^{(j)} - \tilde{\mathbf{C}}^{(j)}\| \lesssim \begin{cases} 2^{-qj\omega\theta/\rho} & \text{when } \xi > \omega, \\ 2^{-qj(\xi+(\theta-\sigma)\omega)/\rho} & \text{when } \xi < \omega. \end{cases} \tag{46}$$

In particular, taking $\theta = \min(1,\sigma)$, *we have* $\|\mathbf{C}^{(j)} - \tilde{\mathbf{C}}^{(j)}\| \lesssim 2^{-qj\min(\omega,\xi)/\rho}$.

Proof. The work per row or column is bounded by an absolute multiple of

$$\begin{aligned} \sum_{i=0}^{j/\rho} 2^{iq} \max\left(1, 2^{qj\theta/\rho - i\sigma q}\right) &\eqsim 2^{qj/\rho} + 2^{qj\theta/\rho} \sum_{i=0}^{j/\rho} 2^{iq(1-\sigma)} \\ &\eqsim 2^{qj/\rho} + 2^{qj\theta/\rho} \max(1, 2^{qj(1-\sigma)/\rho}) \eqsim 2^{qj/\rho}, \end{aligned}$$

because of $\theta \le \min(1,\sigma)$.

Taking into account the selection of $N_{j,\lambda,\mu}$, two applications of the Schur lemma show that

$$\begin{aligned} \|\mathbf{C}^{(j)}_{\ell,m} - (\tilde{\mathbf{C}}^{(j)}_{\lambda,\mu})_{|\lambda|=\ell,|\mu|=m}\|^2 &\lesssim 2^{|\ell-m|q} \big(2^{qj\theta/\rho - |\ell-m|\sigma q}\big)^{-2\omega} 2^{-|\ell-m|(q+2\xi q)} \\ &= 2^{-2qj\theta\omega/\rho} 2^{-|\ell-m|2q(\xi-\sigma\omega)}, \\ \|\mathbf{C}^{(j)} - \tilde{\mathbf{C}}^{(j)}\| &\lesssim \sum_{0\le i\le j/\rho} 2^{-qj\theta\omega/\rho} 2^{-iq(\xi-\sigma\omega)}, \end{aligned}$$

which shows (46). □

Comparing Propositions (5.2) and (5.3), we see that in order to prove our earlier claim that $\mathbf{C} = \mathbf{E}^{(r)}$ or $\mathbf{C} = \mathbf{E}^{(s)}$ are s^*-computable for the same value of s^* as they were shown to be s^*-computable, it suffices to have available a family of quadrature formulas satisfying (45) with

$$\min(\omega,\xi) \ge \rho \quad \text{and} \quad \max(\omega,\xi) > \rho.$$

Below, under some mild additional assumption ((48)), we verify this by showing that for any $a, b > 0$, we can construct a family of approximations $(\tilde{\mathbf{E}}_{\lambda,\mu,N})_{N\in\mathbb{N}}$, where $\tilde{\mathbf{E}}_{\lambda,\mu,N}$ requires $\mathscr{O}(N)$ evaluations of $g\partial^{\alpha_1}\psi^{(1)}_{\mu}\partial^{\alpha_2}\psi^{(2)}_{\lambda}$, such that for some $t \in \mathbb{N}$,

$$|\mathbf{E}_{\lambda,\mu} - \tilde{\mathbf{E}}_{\lambda,\mu,N}| \lesssim N^{-a} 2^{-\left||\lambda|-|\mu|\right|(\frac{n}{2}+b)} \|g\|_{W^t_\infty(\Omega)}. \tag{47}$$

This means that (45) is valid with $\omega = a$ and $\xi = b/n$ or $\xi = (b+\frac{1}{2})/(n-1)$ for $q = n$ or $q = n-1$, respectively.

Without loss of generality let us assume that

$$|\lambda| \geq |\mu|.$$

Suppose that for any $k \in \mathbb{N}_0$ and $\nu \in \mathscr{O}_k^{(\nu)}$, there exists a sufficiently smooth transformation of coordinates κ, with derivatives bounded uniformly in k and ν, such that for some $e \in \mathbb{N}$, and all $|\lambda| = k$,

$$\psi_\lambda^{(2)} \circ \kappa|_{\kappa^{-1}(\Omega_k^{(\nu)})} \in P_{e-1}. \tag{48}$$

In the following, for notational convenience, without loss of generality we take $\kappa = \mathrm{id}$.

To approximate an integral $\int_{\Omega_k^{(\nu)}} f$, for any $p \in \mathbb{N}$ we consider internal, uniformly stable, composite quadrature rules $Q_{\Omega_k^{(\nu)},N}(f)$ of *fixed order* (i.e, the degree of polynomial exactness plus one) p, and *variable rank* N. The rank N of a composite quadrature formula denotes the number of subdomains on which the elementary quadrature formula is applied. Since the order p of $Q_{\Omega_k^{(\nu)},N}$ is fixed, the number of abscissae in the composite rule $Q_{\Omega_k^{(\nu)},N}$ is $\mathscr{O}(N)$. For such rules, the following error estimate is valid

$$|\int_{\Omega_k^{(\nu)}} f - Q_{\Omega_k^{(\nu)},N}(f)| \lesssim \mathrm{vol}(\Omega_k^{(\nu)}) N^{-p/n} \mathrm{diam}(\Omega_k^{(\nu)})^p \|f\|_{W_\infty^p(\Omega_k^{(\nu)})} \tag{49}$$

(e.g., see [GS06a, §2]).

To find an upper bound for the quadrature error when these rules are applied with integrand $2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|} g \partial^{\alpha_1} \psi_\mu^{(1)} \partial^{\alpha_2} \psi_\mu^{(2)}$, we have to bound the expression $(\partial^\rho g)(\partial^\sigma \partial^{\alpha_1} \psi_\mu^{(1)})(\partial^\tau \partial^{\alpha_2} \psi_\lambda^{(2)})$ for all multi-indices with $|\rho+\sigma+\tau| \leq p$. Since g is assumed to be sufficiently smooth, $|\lambda| \geq |\mu|$ and $\partial^\tau \partial^{\alpha_2} \psi_\lambda^{(2)}$ vanishes when $|\tau+\alpha_2| \geq e$, by invoking (41) we see that the worst case occurs when $\rho = 0$ and $|\tau+\alpha_2| = z := \min(e-1, p+|\alpha_2|)$, and thus when $|\sigma| = p-z+|\alpha_2|$, yielding

$$2^{-|\mu||\alpha_1|-|\lambda||\alpha_2|} \|g \partial^{\alpha_1} \psi_\mu^{(1)} \partial^{\alpha_2} \psi_\lambda^{(2)}\|_{W_\infty^p(\Omega_k^{(\nu)})} \lesssim 2^{(|\mu|+|\lambda|)\frac{n}{2}} 2^{|\mu|(p-z+|\alpha_2|)} 2^{|\lambda|(z-|\alpha_2|)} \|g\|_{W_\infty^p(\Omega_k^{(\nu)})}.$$

By substituting this result into (49), using that $\mathrm{diam}(\Omega_k^{(\nu)}) \eqsim 2^{-|\lambda|}$ and $\mathrm{vol}(\Omega_k^{(\nu)}) \eqsim 2^{-|\lambda|n}$, by taking p satisfying

$$p \geq \max(na, b-|\alpha_2|+e-1)$$

and by summing over the uniformly bounded number of $\overline{\Omega}_k^{(\nu)}$ that make up supp $\psi_\lambda^{(2)}$ we end up with (47).

This completes the proof of our claim made at the beginning of this subsection that that $\mathbf{C} = \mathbf{E}^{(r)}$ or $\mathbf{C} = \mathbf{E}^{(s)}$ are s^*-computable for the same value of s^* as they were shown to be s^*-computable.

Remark 5.1. The estimate for the quadrature error obtained by summing the error estimates for the quadrature errors over those ν with $\Omega_\ell^{(\nu)} \subset \mathrm{supp}\psi_\lambda^{(2)}$ can be orders of magnitude too pessimistic. The point is that it has not been used that $\psi_\lambda^{(2)}$ is a wavelet and thus is oscillating, which causes cancellation of errors, in particular when $\psi_\mu^{(1)}$ is smooth on the interior of $\mathrm{supp}\psi_\lambda^{(2)}$, i.e, when it concerns a *regular* entry. For that case, much sharper estimates can be found in [SW08], see also [BBD+02].

5.3 Trees

Although, as we demonstrated, it can be done whilst retaining optimal computational complexity, the approximate computation using quadrature of the required entries of the stiffness matrix that may involve wavelets on largely different levels is a rather delicate process. Such computations can be avoided by restricting to wavelet approximations where the underlying index sets form a tree. In this subsection, we briefly indicate the main ingredients of this approach.

We restrict ourselves to the case that $\Psi = \Psi^{\mathscr{X}} = \Psi^{\mathscr{Y}} = \{\psi_\lambda : \lambda \in \nabla\}$ is a Riesz basis for $\mathscr{X} = \mathscr{Y}$. Apart from wavelets, here we will need scaling functions. A set $\Phi_k \subset \mathscr{X}$ is called a collection of scaling functions on level k when $\mathrm{span}\{\psi_\lambda : |\lambda| \leq k\} = \mathrm{span}\Phi_k$. We assume that the Φ_k are (uniformly) local and locally finite (cf. definitions in Subsec. 5.1), and that each wavelet ψ_λ is a linear combination of a uniformly bounded number of scaling functions on level $|\lambda|$ (and that $\mathrm{supp}\,\psi_\lambda$ is connected).

We equip the index set ∇ with a tree structure by assigning to each $\lambda \in \nabla$ with $|\lambda| > 0$ a *parent* μ with $|\mu| = |\lambda| - 1$ and $\mathrm{supp}\psi_\lambda \cap \mathrm{supp}\psi_\mu \neq \emptyset$. By our assumptions, the number of children of any parent is uniformly bounded. We call $\Lambda \subset \nabla$ a *tree*, when all $\lambda \in \nabla$ with $|\lambda| = 0$ are in Λ (the "roots"), and when whenever $\lambda \in \nabla$ with $|\lambda| > 0$ is in Λ then so is its parent.

Analogously to (2), we define approximation classes $\mathscr{A}^s$, and corresponding (quasi-) norms $\|\cdot\|_{\mathscr{A}^s}$, where we now consider only best N-term approximations $\mathbf{u}_N$ to $\mathbf{u} \in \ell_2$ whose supports, apart from having a length not larger than N, form a tree. For $\mathscr{X}$ being a Sobolev space, it has been shown that the resulting classes are only slightly smaller than those one obtains with unconstrained best N-term approximation, see [CDDD01] for details.

The reason to consider tree approximation is that any $\mathbf{w} \in \ell_0$ whose support forms a tree, can be expressed as a linear combination of K scaling functions, where

$K \lesssim \#\operatorname{supp}\mathbf{w}$ and where the supports of any two scaling functions in this expansion can only intersect when their difference in levels is not larger than 1. Moreover, this scaling function representation can be found in $\mathcal{O}(\#\operatorname{supp}\mathbf{w})$ operations, see [DSX00b].

As an application, now let $\mathbf{B} \in \mathscr{L}(\ell_2,\ell_2)$ be s^*-compressible, let the support of $\mathbf{w} \in \ell_0$ form a tree, and let $\varepsilon > 0$ be given. Then as shown in [DHS07], using the near best N-term tree approximation algorithm from [BD04], trees $\Lambda_j \subset \cdots \subset \Lambda_2 \subset \Lambda_1 \subset \operatorname{supp}\mathbf{w}$ can be found such that, with $\mathbf{w}_{[p]} := \mathbf{w}|_{\Lambda_p\setminus\Lambda_{p+1}}$ ($\Lambda_{j+1} := \emptyset$) and suitable $j_p \in \mathbb{N}_0$, $\mathbf{z}_\varepsilon := \sum_{p=1}^{j} \mathbf{B}^{(j_p)}\mathbf{w}_{[p]}$ satisfies $\|\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \le \varepsilon$ and, for any $s < s^*$, $\#\operatorname{supp}\mathbf{z}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{z}_\varepsilon\|_{\mathscr{A}^s}^{-1/s}$, where the cost of determining $\operatorname{supp}\mathbf{z}_\varepsilon$ is bounded by some absolute multiple of $\varepsilon^{-1/s}\|\mathbf{z}_\varepsilon\|_{\mathscr{A}^s}^{-1/s} + \#\operatorname{supp}\mathbf{w} + 1$. What is more, taking the construction of the sparse matrices $\mathbf{B}^{(j)}$ into account, for both partial differential and singular integral operators, $\operatorname{supp}\mathbf{z}_\varepsilon$ forms a tree.

Instead of approximating the required entries of the involved matrices $\mathbf{B}^{(j_p)}$, this opens another possibility to approximate $\mathbf{B}\mathbf{w}$. Since $\|\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \le \varepsilon$ is shown by estimating $\|\mathbf{B}\mathbf{w} - \mathbf{z}_\varepsilon\| \le \sum_{p=1}^{j} \|\mathbf{B} - \mathbf{B}^{(j_p)}\|\|\mathbf{w}_{[p]}\|$, and by bounding $\|\mathbf{B} - \mathbf{B}^{(j_p)}\|$ by summing over upper bounds for the entries of $\mathbf{B}$ that were dropped in the definition of $\mathbf{B}^{(j_p)}$, one infers that also $\|\mathbf{B}\mathbf{w} - (\mathbf{B}\mathbf{w})|_{\operatorname{supp}\mathbf{z}_\varepsilon}\| \le \varepsilon$ as well as that $\|\mathbf{B}\mathbf{w} - (\mathbf{B}\mathbf{w})|_{\bar\Lambda}\| \le \varepsilon$, where $\bar\Lambda := \operatorname{supp}\mathbf{w} \cup \operatorname{supp}\mathbf{z}_\varepsilon$ is a tree.

Now with $\bar{\boldsymbol{\Phi}}$ denoting the collection of the single scale functions with $\operatorname{span}\bar{\boldsymbol{\Phi}} = \{\psi_\lambda : \lambda \in \bar\Lambda\}$ and $T_{\bar\Lambda}$ the corresponding basis transformation from multiscale to single scale representation, we have $\mathbf{B}|_{\bar\Lambda\times\bar\Lambda} = T_{\bar\Lambda}^\top \mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})T_{\bar\Lambda}$, thus with $\mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})$ being the single-scale representation of $\mathbf{B}|_{\bar\Lambda\times\bar\Lambda}$. Since $(\mathbf{B}\mathbf{w})|_{\bar\Lambda} = T_{\bar\Lambda}^\top \mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})T_{\bar\Lambda}\mathbf{w}$, in order to construct a valid **APPLY**, what is left is to approximate the multiplication with $\mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})$ in $\mathcal{O}(\varepsilon^{-1/s}\|\mathbf{z}_\varepsilon\|_{\mathscr{A}^s}^{-1/s} + \#\operatorname{supp}\mathbf{w} + 1)$ operations, while keeping the error on the level of a multiple of ε. For partial differential operators, the advantage is that non-zeros entries of $\mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})$ only involve pairs of scaling functions on equal or consecutive levels. For singular integral operators, to approximate the multiplication with $\mathbf{B}(\bar{\boldsymbol{\Phi}},\bar{\boldsymbol{\Phi}})$ one may think of the application of *panel clustering* ([HN89]) or *multipole expansions* ([GR87]).

Finally, whereas for the optimal adaptive solution of *linear* operator equations, the restriction to tree approximations is not really necessary, for such a solution of *nonlinear* operator equations it seems indispensable (see [CDD03a]). Indeed, note that for a nonlinear operator of the form $f(v)(x) = g(v(x))$, the evaluation of $f(\mathbf{w}^\top\boldsymbol{\Psi})(x)$ already requires a number of operations of the order of the number of wavelets in the expansion that are non-zero in x. If $\operatorname{supp}\mathbf{w}$ is a tree, however, then after transformation to the locally finite single scale representation, any of such a point evaluations can be done in $\mathcal{O}(1)$ operations.

6 Adaptive frame methods

6.1 Introduction

A drawback of wavelet methods for solving operator equations is the rather complicated construction of wavelet bases on non-product domains. As was already mentioned at the end of Sect. 5.1, the usual construction is via a non-overlapping decomposition of the n-dimensional domain or manifold into subdomains, each of them being a smooth parametric image of the n-dimensional unit cube. Loosely speaking, wavelets or scaling functions constructed on this n-cube are lifted to the subdomains, after which those functions that do not vanish at an interface between subdomains are either continuously connected to functions from neighbouring subdomains or are smoothly extended into these subdomains. Apart from the fact that these constructions are not that easy to implement, another disadvantage is that the condition numbers of the resulting bases are quite somewhat larger than that of the corresponding bases on the n-cube.

As an alternative, for $\mathscr{X}$ being a Sobolev space, in [Ste03] it was suggested to use an *overlapping* domain decomposition, and to define $\Psi^{\mathscr{X}}$ simply as the union of the wavelet bases on the subdomains. By a proper choice of the bases on these subdomains, the span of $\Psi^{\mathscr{X}}$ will be dense in $\mathscr{X}$, but due to the overlap regions, it cannot be a basis for $\mathscr{X}$. Instead it will be a *frame* for $\mathscr{X}$. In [DFR07a], such a frame was called an *aggregated wavelet frame*.

6.2 Frames

Let $\mathscr{X}$ be a separable Hilbert space. A collection $\Psi = \{\psi_\lambda : \lambda \in \nabla\} \subset \mathscr{X}$ is called a *frame* for $\mathscr{X}$ when the *analysis operator*

$$\mathscr{F} : \mathscr{X}' \to \ell_2 : g \mapsto [g(\psi_\lambda)]_{\lambda\in\nabla},$$

is a boundedly invertible mapping between $\mathscr{X}'$ and *its range* $\operatorname{ran}\mathscr{F}$. From Sect. 2, recall that its adjoint, known as the *synthesis operator*, reads as

$$\mathscr{F}' : \ell_2 \to \mathscr{X} : \mathbf{c} \mapsto \mathbf{c}^\top \Psi.$$

We set the *frame constants*

$$\Lambda_\Psi := \|\mathscr{F}\|_{\mathscr{X}'\to\ell_2}, \quad \lambda_\Psi := \inf_{0\neq g\in\mathscr{X}'} \frac{\|\mathscr{F}g\|_{\ell_2}}{\|g\|_{\mathscr{X}}}.$$

The composition $\mathscr{F}'\mathscr{F} : \mathscr{X}' \to \mathscr{X}$ is boundedly invertible with $\|\mathscr{F}'\mathscr{F}\|_{\mathscr{X}'\to\mathscr{X}} = \Lambda_\Psi^2$ and $\|(\mathscr{F}'\mathscr{F})^{-1}\|_{\mathscr{X}\to\mathscr{X}'} = \lambda_\Psi^{-2}$.

The collection $\tilde{\Psi} := (\mathscr{F}'\mathscr{F})^{-1}\Psi$ is a frame for $\mathscr{X}'$, known as the canonical dual frame, with analysis operator $\tilde{\mathscr{F}} := \mathscr{F}(\mathscr{F}'\mathscr{F})^{-1}$ and frame constants λ_Ψ^{-1} and Λ_Ψ^{-1}. From $\mathscr{F}'\tilde{\mathscr{F}} = I$, one infers that any $v \in \mathscr{X}$ has a representation $v = \mathbf{v}^\top\Psi$ with $\Lambda_\Psi^{-1} \leq \|\mathbf{v}\|_{\ell_2}/\|v\|_{\mathscr{X}} \leq \lambda_\Psi^{-1}$, actually a property that is equivalent to Ψ being a frame with frame constants Λ_Ψ and λ_Ψ. Note that generally a representation of $v \in \mathscr{X}$ in frame coordinates is not unique (unless Ψ is a Riesz basis).

We have $\ell_2 = \operatorname{ran}\mathscr{F} \oplus^\perp \ker\mathscr{F}'$ and $\mathbf{Q} := \tilde{\mathscr{F}}\mathscr{F}'$ is the orthogonal projector onto $\operatorname{ran}\mathscr{F}$. The frame Ψ is a Riesz basis for $\mathscr{X}$ if and only if $\ker\mathscr{F}' = 0$ or equivalently $\operatorname{ran}\mathscr{F} = \ell_2$.

Many examples of frames can be given. Besides aggregated wavelet frames, here we only mention curvelets ([CD04]) and shearlets ([LLKW05]).

For a given $f \in \mathscr{X}'$ and a boundedly invertible $B \in \mathscr{L}(\mathscr{X}, \mathscr{X}')$, let us consider the problem of finding $u \in \mathscr{X}$ such that

$$Bu = f. \tag{50}$$

Writing $u = \mathscr{F}'\mathbf{u}$ for *some* $\mathbf{u} \in \ell_2$, this $\mathbf{u}$ solves

$$\mathbf{Bu} = \mathbf{f}, \tag{51}$$

where

$$\mathbf{B} := \mathscr{F}B\mathscr{F}', \quad \mathbf{f} := \mathscr{F}f.$$

Obviously, we have $\|\mathbf{B}\| \leq \Lambda_\Psi^2\|B\|_{\mathscr{X}\to\mathscr{X}'}$. With respect to the decomposition $\ell_2 = \operatorname{ran}\mathscr{F} \oplus^\perp \ker\mathscr{F}'$, $\mathbf{B}$ is of the form $\left[\begin{smallmatrix}\mathbf{B}_0 & 0\\ 0 & 0\end{smallmatrix}\right]$. From $\tilde{\mathscr{F}}B^{-1}\tilde{\mathscr{F}}'\mathbf{B} = \mathbf{B}\tilde{\mathscr{F}}B^{-1}\tilde{\mathscr{F}}' = \mathbf{Q}$, we conclude that $\mathbf{B}_0 = \mathbf{B}|_{\operatorname{ran}\mathscr{F}} : \operatorname{ran}\mathscr{F} \to \operatorname{ran}\mathscr{F}$ is boundedly invertible with $\|\mathbf{B}_0^{-1}\| \leq \lambda_\Psi^{-2}\|B^{-1}\|_{\mathscr{X}'\to\mathscr{X}}$. Finally, we note that for $\mathbf{v}, \mathbf{w} \in \operatorname{ran}\mathscr{F}$,

$$\langle \mathbf{B}_0\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{Bv}, \mathbf{w}\rangle = \langle \mathscr{F}B\mathscr{F}'\mathbf{v}, \mathbf{w}\rangle = (Bv)(w), \tag{52}$$

where $v = \mathscr{F}'\mathbf{v}$ and $w = \mathscr{F}'\mathbf{w}$, or equivalently because $\mathbf{v}, \mathbf{w} \in \operatorname{ran}\mathscr{F}$, $\mathbf{v} = \tilde{\mathscr{F}}v$ and $\mathbf{w} = \tilde{\mathscr{F}}w$.

6.3 *The adaptive solution of an operator equation in frame coordinates*

In case the operator B in (50) is symmetric and positive definite, one may think of applying the adaptive wavelet Galerkin approach discussed in Sect. 4 onto $\mathbf{Bu} = \mathbf{f}$ from (51). Since, however, for a "true" frame, $\mathbf{B}$ has a non-trivial kernel, for $\Lambda \subsetneq \nabla$ the generalized condition number of $\mathbf{B}|_{\Lambda\times\Lambda}$, i.e., the quotient of its largest and its smallest non-negative eigenvalue, can be arbitrarily large. This makes this approach unfeasible.

Therefore, we return to the damped Richardson iteration discussed in Sect. 3.1. Denoting its ith iterand as $\mathbf{u}^{(i)}$, and with $\mathbf{u}$ *some* solution of $\mathbf{Bu} = \mathbf{f}$, we have

$$\mathbf{u} - \mathbf{u}^{(i)} = (\mathbf{I} - \alpha\mathbf{B})(\mathbf{u} - \mathbf{u}^{(i-1)}),$$

which, due to the non-trivial kernel of $\mathbf{B}$, shows no convergence. By applying $\mathbf{Q}$, however, we obtain

$$\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)}) = (\mathbf{I} - \alpha\mathbf{B}_0)\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i-1)}).$$

If B is symmetric and positive definite or only coercive, then in view of (52), analogously to the analysis from Sect. 4, we infer that by a proper choice of α, $\|\mathbf{I} - \alpha\mathbf{B}_0\| < 1$. Since $u - \mathscr{F}'\mathbf{u}^{(i)} = \mathscr{F}'\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})$, we conclude linear convergence of $\mathscr{F}'\mathbf{u}^{(i)}$ to u in $\mathscr{X}$. For non-coercive B, the iteration can be applied to the normal equations.

When applying the damped Richardson iteration with an inexact evaluation of the matrix-vector multiplication and that of the right-hand side $\mathbf{f}$, then, with a proper choice of decaying tolerances, for the resulting iteration a linear decrease of the projected error $\mathbf{Q}(\mathbf{u} - \mathbf{u}^{(i)})$ can still be shown. These inexact evaluations, however, generally produce error components that are in $\ker \mathscr{F}'$. Since $\ker \mathscr{F}' = \ker \mathbf{B}$, these error components will not be changed by subsequent Richardson steps. Although these error components do not affect the projected error, generally they do affect the $\mathscr{A}^s$-norms of the iterands, and with that, the cost of the applications of the **APPLY** routine.

In spite of this, in [Ste03] it was proved that the algorithm **Rich**, as given in Sect. 3.2 but with a modified choice of the tolerances (see [Ste03] for details), is again (quasi-) optimal in the sense of Theorem 3.1: *Given an* $\varepsilon > 0$, *it produces an* $\mathbf{u}_\varepsilon$ *with* $\|\mathbf{Q}(\mathbf{u} - \mathbf{u}_\varepsilon)\| \leq \varepsilon$. *If for some* $s > 0$, $\mathbf{Bu} = \mathbf{f}$ *has* some *solution* $\mathbf{u} \in \mathscr{A}^s$, *then* $\#\operatorname{supp} \mathbf{u}_\varepsilon \lesssim \varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *If, additionally, for some* $\bar{s} > s$, $\mathbf{B}$ *is* $\bar{s}$*-admissible and* $\mathbf{Q} : \mathscr{A}^{\bar{s}} \to \mathscr{A}^{\bar{s}}$ *is bounded, then the number of operations used by the call is bounded by an absolute multiple of* $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$. *In order words, if* $\bar{s} > s_{\max}$, *with* $s_{\max}$ *defined similarly as in the basis case* (*see Sect. 1.1*), *then this inexact Richardson iteration is* (quasi-) optimal.

The additional condition that $\mathbf{Q} : \mathscr{A}^{\bar{s}} \to \mathscr{A}^{\bar{s}}$ is bounded is satisfied when $\mathbf{Q}$ is $\bar{s}$-admissable (cf. Definition 3.1 and Proposition 3.1), which in turn is satisfied when, for some $s^* > \bar{s}$, $\mathbf{Q}$ is s^*-compressible (cf. Definitions 3.2, 5.1 and Theorem 3.2, and realize that the question about cost of computing entries of $\mathbf{Q}$ is not relevant, since $\mathbf{Q}$ does not enter the algorithm, but its boundedness in $\mathscr{A}^s$ is only needed for the proof of optimality).

Unfortunately, although we expect it to hold more generally, in the aggregated wavelet frame case so far the s^*-compressibility of $\mathbf{Q}$ was proved (in [Ste03, §4.3]) only in the case that the wavelets on each subdomain are L_2-orthogonal and that, before aggregation, they were multiplied by a smooth function that is positive on the subdomain and that vanishes outside the subdomain. Numerical results reported

in [DFR+07b] indicate (quasi-) optimality in other cases. In [DFR07a], the boundedness of $\mathbf{Q} : \mathscr{A}^{\bar{s}} \to \mathscr{A}^{\bar{s}}$ was shown for time-frequency localized Gabor frames.

Sufficient for $\bar{s}$-admissability of $\mathbf{B}$ is that it is s^*-computable for some $s^* > \bar{s}$. For aggregated wavelet frames, a proof of s^*-compressiblity of $\mathbf{B}$ can follows the same lines as in Sect. 5.1 for the basis case. In the aggregated wavelet frame case, the approximate computation using quadrature of the significant entries of $\mathbf{B}$ is a harder task. Indeed, wavelets from different subdomains whose supports overlap will be piecewise smooth with respect to different underlying partitions. Nevertheless, in [SW08], for partial differential operators with smooth coefficients, s^*-computability for an $s^* > s_{\max}$ was demonstrated.

Thinking of a symmetric and positive definite B, the selection of a suitable damping parameter α for the Richardson iteration requires estimating the smallest nonnegative eigenvalue of $\mathbf{B}$. Other than in the Riesz basis case where $\mathbf{B}$ has no zero eigenvalues, in the true frame case it is difficult to estimate this eigenvalue numerically. In [DFR+07b], it was shown that an approximate steepest descent iteration, which does not require information about the spectrum of $\mathbf{B}$, is (quasi-) optimal under the same conditions as the approximate Richardson iteration.

6.4 An adaptive Schwarz method for aggregated wavelet frames

Let $B \in \mathscr{L}(\mathscr{X}, \mathscr{X}')$ be symmetric and positive definite, where $\mathscr{X}$ is a Sobolev space with positive smoothness index on a domain Ω. Let Ψ be an aggregated wavelet frame being the union of wavelet bases $\Psi_1, \dots, \Psi_m$ on overlapping subdomains $\Omega_1, \dots, \Omega_m$, respectively. Each of these bases is a Riesz basis of the corresponding Sobolev space on the subdomain, with homogeneous Dirichlet boundary conditions on the internal boundary.

The partition of the domain into overlapping subdomains, or that of the frame into the different Riesz systems, suggest the application of a Schwarz method to solve $\mathbf{Bu} = \mathbf{f}$, being the representation of the operator equation $Bu = f$ in frame coordinates. An multiplicative adaptive Schwarz method was studied in [SW09].

Let $\mathbf{B} = (\mathbf{B}_{k\ell})_{1\le k,\ell\le m}$ and $\mathbf{v} = (\mathbf{v}_k)_{1\le k\le m}$ denote the corresponding partitions of the system matrix $\mathbf{B}$ and any vector of frame coordinates, respectively. Then the (exact) multiplicative Schwarz algorithm reads as follows:

```
for i = 1,2,... do
    for k = 1 to m do
        solve B_kk u_k^(i) = f_k − Σ_{ℓ=1}^{k−1} B_kℓ u_ℓ^(i) − Σ_{ℓ=k+1}^{m} B_kℓ u_ℓ^(i−1)
    enddo
enddo
```

Using the general theory of Schwarz methods (e.g. see [Xu92]), one shows that $\mathscr{F}'\mathbf{u}^{(i)} = \mathbf{u}^{(i)\top}\Psi$ converges linearly to u in $\mathscr{X}$.

The idea behind an inexact, adaptive variant is to find an approximation to $\mathbf{u}_k^{(i)}$ by the application of an adaptive *wavelet* method on subdomain Ω_k (either of inexact

Richardson type or an adaptive wavelet Galerkin method). By a suitable choice of decaying tolerances, the resulting method will still be linearly convergent.

For each k, the sequence $(\mathbf{u}_k^{(i)})_i$ of approximate solutions of the subdomain problems on Ω_k converges to some $\mathbf{u}_k$, that depends on the choice of the initial vectors $(\mathbf{u}_\ell^{(0)})_{1\le\ell\le m}$. With $\mathbf{u}$ being *some* representation of u, i.e., $\mathbf{u}^\top\Psi = u$, it is not clear that the splitting $\mathbf{u} = \mathbf{u}_k + (\mathbf{u} - \mathbf{u}_k)$ is smoothness preserving, in the sense that if $\mathbf{u} \in \mathscr{A}^s$, then $\mathbf{u}_k \in \mathscr{A}^s$ with $\|\mathbf{u}_k\|_{\mathscr{A}^s} \lesssim \|\mathbf{u}\|_{\mathscr{A}^s}$. From our considerations about the cost of the **APPLY** routine that is part of the adaptive wavelet method, it is however clear that such a smoothness preservation would be needed to conclude (quasi-) optimality of the resulting method. Actually, numerical experiments indicated that generally this splitting is not smoothness preserving.

In order to solve this problem, again consider the system

$$\mathbf{B}_{kk}\mathbf{u}_k^{(i)} = \mathbf{f}_k - \sum_{\ell=1}^{k-1} \mathbf{B}_{k\ell}\mathbf{u}_\ell^{(i)} - \sum_{\ell=k+1}^{m} \mathbf{B}_{k\ell}\mathbf{u}_\ell^{(i-1)}.$$

Note that if, before solving, coefficients from $(\mathbf{u}_\ell^{(i)})_{1\le\ell\le k-1}$ and $(\mathbf{u}_\ell^{(i-1)})_{k+1\le\ell\le m}$ that correspond to wavelets that are fully supported in Ω_k are modified, in particular, are *deleted*, then this will not change the approximation $\mathbf{u}^{(i)\top}\Psi = \sum_{\ell=1}^k \mathbf{u}_\ell^{(i)\top}\Psi_\ell + \sum_{\ell=k+1}^m \mathbf{u}_\ell^{(i-1)\top}\Psi_\ell$ after this solution process, although the vectors $(\mathbf{u}_\ell^{(i)})_{1\le\ell\le k}$ and $(\mathbf{u}_\ell^{(i-1)})_{k+1\le\ell\le m}$ generally do change. For this process, but then with an inexact adaptive solving, it was shown that if the sizes of the overlap regions are sufficiently large compared to the maximal diameter of the support of any wavelet, then the aforementioned splitting *is* smoothness preserving. Using this result, the overall method was shown to be (quasi-) optimal assuming that $\mathbf{B}$ is $\bar{s}$-admissable for some $\bar{s} \ge s_{\max}$ (cf. the discussion in Sect. 6.3). The boundedness of $\mathbf{Q} : \mathscr{A}^{\bar{s}} \to \mathscr{A}^{\bar{s}}$ is *not* required.

Note that the method with the deletion of the coefficients that correspond to wavelets associated to other subdomains, but that are fully supported in the current subdomain is actually closer to the original Schwarz method from [Sch90] than the method we described first. Indeed, what is left after this deletion process is essentially only boundary data for the problem on the current subdomain. The method with deletion is also cheaper to implement since it requires the computation of less entries in the system matrix corresponding to pairs of wavelets associated to different subdomains. Recall that the quadrature problem to approximate those entries is more demanding.

Numerical results reported in [SW09] show that quantitatively this multiplicative adaptive Schwarz method is much more efficient that the adaptive steepest descent method described in Sect. 6.3.

7 Adaptive methods based on tensor product wavelet bases

7.1 Tensor product wavelets

Let Ω be a product domain, i.e., $\Omega = \Omega_1 \times \cdots \times \Omega_n$, then for $t \geq 0$,

$$H^t(\Omega) = H^t(\Omega_1) \otimes L_2(\Omega_2) \otimes \cdots \otimes L_2(\Omega_n) \cap \cdots \cap L_2(\Omega_1) \otimes \cdots \otimes L_2(\Omega_{n-1}) \otimes H^t(\Omega_n)$$

For $t \notin \mathbb{N}_0 + \frac{1}{2}$, the same holds true with $H^t(\Omega)$ reading as $H^t_0(\Omega)$ and $H^t(\Omega_i)$ as $H^t_0(\Omega)$. Similar statements involving boundary conditions of lower order, or with boundary conditions on a part of the boundary (of product type) are also valid ([DS09a]).

Now for $1 \leq i \leq n$, let $\Psi^{(i)} = \{\psi^{(i)}_\lambda : \lambda \in \nabla_i\} \subset H^t(\Omega_i)$ be a Riesz basis for $L_2(\Omega_i)$ that, when normalized in $H^t(\Omega_i)$, is a Riesz basis for $H^t(\Omega_i)$. Wavelet bases are known to have this property for a range of t. Then using above characterization of $H^t(\Omega)$, it can be shown (cf. [GO95]) that the tensor product wavelet basis

$$\boldsymbol{\Psi} := \Psi^{(1)} \otimes \cdots \otimes \Psi^{(n)} = \{\boldsymbol{\psi}_{\boldsymbol{\lambda}} := \psi^{(1)}_{\lambda_1} \otimes \cdots \psi^{(n)}_{\lambda_n} : \boldsymbol{\lambda} \in \boldsymbol{\nabla} := \nabla^{(1)} \times \cdots \times \nabla^{(n)}\}$$

is a Riesz basis for $H^t(\Omega)$.

Note that the widths of the support of a tensor product wavelet measured in the coordinate directions can differ to an arbitrarily large extend. Furthermore, other than with a (standard) wavelet basis, there exists no multiresolution analysis on Ω such that (biorthogonal) complement spaces are spanned by a subset of $\boldsymbol{\Psi}$.

In spite of these differences, tensor product wavelet bases can be applied in adaptive wavelet algorithms. In order to show that these algorithms give (quasi-) optimal results, what is needed to verify is that the representation of the operator under consideration in tensor product wavelet coordinates can be sufficiently well approximated by computable sparse matrices in relation to the best possible convergence rate that can be expected. That is, what is needed to check is whether $s^* > s_{\max}$, with $s_{\max}$ being defined in Sect. 1.1 and s^* from Definition 3.2 in Sect. 3.3.

7.2 Non-adaptive approximation

Let Ω_i be a domain of dimension n_i and $\Psi^{(i)}$ be a wavelet basis of order $d_i > t$, cf. Example 1.1. Then it is well-known that a sufficiently smooth function on Ω can be approximated in $H^t(\Omega)$ from the sequence of spaces $(\mathrm{span}\{\boldsymbol{\psi}_{\boldsymbol{\lambda}} : \sum_{i=1}^n |\lambda_i| \leq \ell\})_\ell$ with rate $s_{\max} = \max_i \frac{d_i - t}{n_i}$, up to some log-factors (the error bound reads as $N^{-\max_i \frac{d_i - t}{n_i}} (\log N)^q$ for some $q > 0$ with N being the number of unknowns). This type of approximation is known as *sparse-grid* or *hyperbolic cross* approximation (see [Zen91, DKT98, BG04]). For $t > 0$, the aforementioned log-factors can even be removed by considering slightly modified approximation spaces, known

as *optimized sparse-grid* spaces ([GK00]). In particular, from now on thinking of $n_1 = \ldots = n_n = 1$ and $d_1 = \ldots = d_n =: d > t$, a sufficiently smooth function on an n-rectangle is approximated in H^t for $t > 0$ by optimized sparse grids with rate $s_{\max} = d - t$. That is, the so-called "curse of dimensionality" – the fact that with standard wavelet (or finite element) approximation the rate is inversely proportional with the space dimension – is completely removed.

7.3 *Best N-term approximation and regularity*

Sparse grid approximation is *non-adaptive*, and the aforementioned high convergence rate requires a smoothness of the function being approximated that the solution of an operator equation may not possess. Indeed, in [DS09a] it was shown that for the Poisson problem on the n-rectangle with homogeneous Dirichlet boundary conditions and a smooth right-hand side, the optimized sparse grid convergence rate in H^1 is $\frac{1}{2} + \frac{1}{n}$, instead of $s_{\max} = d - 1$ that would be obtained when the solution was sufficiently smooth. Only if the right-hand side vanishes to a sufficiently high order at the non-smooth parts of the boundary, the best possible rate is obtained.

The requirements to approximate a function on the n-rectangle with a certain rate $s \leq s_{\max} = d - t$ with *best N-term approximation* from the tensor product basis, i.e., the requirements for the function to be in $\mathscr{A}^s$, are (much) milder than the requirements to obtain this rate with (optimized) sparse grid approximation. For $s < s_{\max}$, a characterization of $\mathscr{A}^s$ in terms of intersections of tensor products of Besov spaces was given in [Nit06]. Following earlier work in [Nit05], for $t \in \mathbb{N}$ in [DS09a] it was shown that if a function u on the n-rectangle has partial derivatives up to order nd in certain weighted L_2 spaces, with weights that vanish at the boundary, then $u \in \mathscr{A}^{d-t}$. What is more, additionally it was shown that the solution of an elliptic boundary value problem of order $2t$ on the n-rectangle with smooth coefficients, homogeneous Dirichlet boundary conditions and a smooth right-hand side satisfies these regularity conditions.

Here we emphasize that for sufficiently large n and d, a rate $d - t$ cannot be realized with best N-term standard wavelet approximation. Indeed, with wavelets of order $\hat{d}$, in n space dimensions the best possible rate is $\frac{\hat{d}-t}{n}$. A (near) characterization of $\mathscr{A}^{\frac{\hat{d}-t}{n}}$ can be given in terms of certain Besov spaces. It is known, however, that for $n \geq 3$, the solution of an elliptic boundary value problem has limited smoothness in this scale of Besov spaces. In other words, one cannot simply choose the rate at one's convenience by increasing the order $\hat{d}$. In any case in three dimensions, with finite elements of order $\hat{d}$ one can realize the best possible rate $\frac{\hat{d}-t}{3}$ by including *anisotropic* refinements towards the boundary ([Ape99]). The tensor product wavelet approach has the unique additional feature that the rate $s_{\max}$ does not deteriorate with an increasing space dimension.

7.4 s^*-computability

In order to conclude that the adaptive tensor product wavelet method converges at the same rate as the sequence of best N-term approximations with respect to the tensor product basis in linear complexity, it is needed that $s^* > s_{\max} = d - t$. For boundary value problems with homogeneous Dirichlet boundary conditions and smooth coefficients and piecewise smooth, sufficiently globally smooth univariate wavelets with sufficiently many vanishing moments, this has been verified in [SS08]. Thinking of the arbitrarily stretched supports of the tensor product wavelets, one might consider it as counterintuitive that an operator is better compressible in a tensor product wavelet basis than it is in a standard wavelet basis. The key is that the sizes of the entries decay exponentially as function of the *sum* of the absolute differences in levels of the tensor product wavelets involved. Compressiblity of integrodifferential operators has been investigated in [Rei08].

7.5 Truly sparse stiffness matrices

Recently, in [DS09b] a univariate wavelet basis of cubic Hermite splines was constructed that has the property that any second order boundary value problem with constant coefficients and homogeneous Dirichlet boundary conditions on the n-cube with respect to the n-fold tensor product basis is *truly sparse*. As a consequence, the application of an adaptive wavelet method simplifies enormously. Indeed, the application of the stiffness matrix to any finitely supported vector can be performed exactly in linear complexity. Also with non-constant, smooth coefficients, the application of this basis in the adaptive wavelet Galerkin method is advantageous. For the approximate residual computation, being the most time consuming part of the algorithm, entries outside the nonzero pattern of a constant coefficient operator, except those that correspond to wavelets on a few coarsest levels, are an order of magnitude smaller than those inside this pattern, and so can be discarded.

7.6 Problems in space high dimension

We have seen that the sequence of approximations produced by an adaptive tensor product wavelet method converges with the same *rate* as the sequence of best N-term approximations with respect to the tensor product basis. This does not exclude the possibility that the quotient of the error produced by the adaptive method and that of the best N-term approximation of the same length grows with increasing n. Actually, generally in any case any available upper bound for this quotient will grow exponentially as function of n. A reason is that various estimates to bound the error for the adaptive method depend critically on the condition number of the

n-fold tensor product basis. If and only if the univariate wavelets are chosen to be L_2-orthogonal, this condition number is bounded uniformly in n, whereas it grows exponentially in n otherwise.

In [DSS08], the n-fold tensor product of the univariate piecewise polynomial L_2-orthogonal wavelet basis from [DGH96] is applied to solve constant coefficient elliptic boundary value problems on the n-rectangle. For this case, it was shown that even the *factor* that the adaptive method might lose compared to the best N-term approximations is bounded by an absolute constant. Experiments for the Poisson problem on the n-cube with right-hand side 1 show, however, that the best N-term approximations themselves still suffer from another, although much milder curse of dimensionality. Although for any dimension n, the rate of approximation in H^1 is $d-1$, the number of unknowns needed to achieve a relative error below some given tolerance grows exponentially with n. Apparently, the constant C in the error bound CN^{d-1} grows exponentially with n. In view of the result from [NW08] saying that the approximation of a general infinitely differentiable multivariate function is *intractable*, this exponential growth of the constant is not surprising.

Likely, to approximate a function in high space dimensions, with the current hardware think of dimensions higher than say 8-10, one should exploit more information about the function than only that it is the solution of a boundary value problem with some *general* smooth right-hand side. As demonstrated in [Gra04, BM02, HK07], a class of functions that can be accurately approximated in high space dimensions are the solutions of boundary value problems with right-hand sides that can be well approximated by a small number of separable functions.

7.7 Non-product domains

The application of tensor product wavelet bases is not restricted to Sobolev spaces $H^t(\Omega)$ with $t \geq 0$ where Ω is product domain. Indeed, recall that the commonly applied approaches to construct wavelet bases on a non-product domain start with writing this domain as a non-overlapping union of subdomains, each of them being a smooth parametric image of the n-cube. With the approach based on extension operators, wavelet bases on the n-cube are lifted to the subdomains, after which those that do not vanish at an interface between subdomains are smoothly extended into neighbouring subdomains. This approach can be applied verbatim to tensor product wavelet bases on the n-cube.

Using a *non-overlapping* domain decomposition, one may also think of constructing an aggregated frame based on tensor product wavelet bases on the subdomains. In the general case, however, where the underlying partitions in the overlap regions are not aligned, the compressibility of the resulting system matrix will be too low.

7.8 Other, non-elliptic problems

We considered well-posed linear operator equations of the form $B : \mathscr{X} \to \mathscr{X}'$, where $\mathscr{X} = H^t(\Omega)$ or $H^t_0(\Omega)$ and Ω is a product domain. In this case, $H^t(\Omega)$ is an intersection of tensor product of Sobolev spaces. Well-posed operator equations $B : \mathscr{X} \to \mathscr{Y}'$, where $\mathscr{X}$ and $\mathscr{Y}$ are of this type arise more generally. We mention here the "unfolding" of elliptic n-scale homogenization problems (cf. [AB96, HS05]) as well as the higher dimensional partial differential equations for the mean field, two-point correlation and possibly higher order moments of the random solution of an elliptic PDE with stochastic input data (cf. e.g. [ST03, HSS08, vPS06]).

Another example is given by the space-time variational formulation of the parabolic initial boundary value problem presented in Sect. 2.2.4. In this case $\mathscr{X} = L_2(0,T) \otimes H^1_0(\Omega) \cap H^1(0,T) \otimes H^{-1}(\Omega)$ and $\mathscr{Y} = (L_2(0,T) \otimes H^1_0(\Omega)) \times L_2(\Omega)$.

A classical approach to the numerical solution of the parabolic initial boundary value problem is the *Method of Lines*, which reduces the problem by spatial semidiscretization to a system of coupled ordinary differential equations to be solved numerically in $(0,T)$. Conversely, in Rothe's Method the problem is reduced by time semidiscretization to a sequence of coupled spatial, elliptic problems to be solved. Both these approaches, and the more recently proposed discontinuous Galerkin method are essentially time marching methods. The ultimate aim of adaptive methods is to achieve an approximate solution with an error below a prescribed tolerance at the expense of, up to an absolute multiple, minimal amount of computer time and storage. Due to the character of time stepping this seems hard to realize and, unlike for elliptic problems, so far no optimality results seem to be known.

In [SS09], the aforementioned spaces $\mathscr{X}$ and $\mathscr{Y}$ were equipped with tensor product wavelet bases. The resulting system matrix was proven to be sufficiently compressible and so the adaptive wavelet method applied to the simultaneously space-time variational formulation converges with the rate as that of the best N-term approximations. While keeping discrete solutions on all time levels is prohibitive for time marching methods, thanks to the use of tensor product bases, with the method in [SS09] there is no penalty in complexity because of the additional time dimension.

References

[AB96] G. Allaire and M. Briane. Multiscale convergence and reiterated homogenisation. *Proc. Roy. Soc. Edinburgh Sect. A*, 126(2):297–342, 1996.

[Ape99] T. Apel. *Anisotropic finite elements: local estimates and applications*. Advances in Numerical Mathematics. B.G. Teubner, Stuttgart, 1999.

[Bar01] T. Barsch. *Adaptive Multiskalenverfahren für elliptische partielle Differentialgleigungen – Realisierung, Umsetzung and numerische Ergebnisse*. PhD thesis, RTWH Aachen, 2001.

[Bar05] A. Barinka. *Fast Evaluation Tools for Adaptive Wavelet Schemes*. PhD thesis, RTWH Aachen, March 2005.

[BBC+01] A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban. Adaptive wavelet schemes for elliptic problems - Implementation and numerical experiments. *SISC*, 23(3):910–939, 2001.

[BBD+02] A. Barinka, T. Barsch, S. Dahlke, M. Mommer, and M. Konik. Quadrature formulas for refinable functions and wavelets. II. Error analysis. *J. Comput. Anal. Appl.*, 4(4):339–361, 2002.

[BD04] P. Binev and R. DeVore. Fast computation in adaptive tree approximation. *Numer. Math.*, 97(2):193–217, 2004.

[BDS07] A. Barinka, W. Dahmen, and R. Schneider. Fast computation of adaptive wavelet expansions. *Numer. Math.*, 105(4):549–589, 2007.

[BG04] H.J. Bungartz and M. Griebel. Sparse grids. *Acta Numer.*, 13:147–269, 2004.

[BK06] S. Berrone and T. Kozubek. An adaptive WEM algorithm for solving elliptic boundary value problems in fairly general domains. *SIAM J. Sci. Comput.*, 28(6):2114–2138 (electronic), 2006.

[BK08] C. Burstedde and A. Kunoth. A wavelet-based nested iteration-inexact conjugate gradient algorithm for adaptively solving elliptic PDEs. *Numer. Algorithms*, 48(1-3):161–188, 2008.

[BM02] G. Beylkin and M.J. Mohlenkamp. Numerical operator calculus in higher dimensions. *Proc. Natl. Acad. Sci. USA*, 99(16):10246–10251 (electronic), 2002.

[BP88] J.H. Bramble and J.E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181):1–17, 1988.

[BU08] K. Bittner and K. Urban. Adaptive wavelet methods using semiorthogonal spline wavelets: sparse evaluation of nonlinear functions. *Appl. Comput. Harmon. Anal.*, 24(1):94–119, 2008.

[CD04] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.

[CDD01] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations – Convergence rates. *Math. Comp*, 70:27–75, 2001.

[CDD02] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods II - Beyond the elliptic case. *Found. Comput. Math.*, 2(3):203–245, 2002.

[CDD03a] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet schemes for nonlinear variational problems. *SIAM J. Numer. Anal.*, 41:1785–1823, 2003.

[CDD03b] A. Cohen, W. Dahmen, and R. DeVore. Sparse evaluation of compositions of functions using multiscale expansions. *SIAM J. Math. Anal.*, 35(2):279–303 (electronic), 2003.

[CDDD01] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.

[Coh03] A. Cohen. *Numerical Analysis of Wavelet Methods*. Elsevier, Amsterdam, 2003.

[CTU99] C. Canuto, A. Tabacco, and K. Urban. The wavelet element method part I: Construction and analysis. *Appl. Comput. Harmon. Anal.*, 6:1–52, 1999.

[CU05] C. Canuto and K. Urban. Adaptive optimization of convex functionals in Banach spaces. *SIAM J. Numer. Anal.*, 42(5):2043–2075, 2005 (electronic).

[Dah97] W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numer.*, 6:55–228, 1997.

[Dah99] S. Dahlke. Besov regularity for the Stokes problem. In W. Haussmann, K. Jetter, and M. Reimer, editors, *Advances in Multivariate Approximation*, Math. Res. 107, pages 129–138, Berlin, 1999. Wiley-VCH.

[DD97] S. Dahlke and R. DeVore. Besov regularity for elliptic boundary value problems. *Comm. Partial Differential Equations*, 22(12):1–16, 1997. &

[DDU02] S. Dahlke, W. Dahmen, and K. Urban. Adaptive wavelet methods for saddle point problems - Optimal convergence rates. *SIAM J. Numer. Anal.*, 40:1230–1262, 2002.

[DeV98] R. DeVore. Nonlinear approximation. *Acta Numer.*, 7:51–150, 1998.

[DFR07a] S. Dahlke, M. Fornasier, and T. Raasch. Adaptive frame methods for elliptic operator equations. *Adv. Comput. Math.*, 27(1):27–63, 2007.

[DFR+07b] S. Dahlke, M. Fornasier, T. Raasch, R.P. Stevenson, and M. Werner. Adaptive frame methods for elliptic operator equations: The steepest descent approach. *IMA J. Numer. Math.*, 27(4):717–740, 2007.

[DGH96] G.C. Donovan, J.S. Geronimo, and D.P. Hardin. Intertwining multiresolution analyses and the construction of piecewise-polynomial wavelets. *SIAM J. Math. Anal.*, 27(6):1791–1815, 1996.

[DHS07] W. Dahmen, H. Harbrecht, and R. Schneider. Adaptive methods for boundary integral equations - complexity and convergence estimates. *Math. Comp.*, 76:1243–1274, 2007.

[DK05] W. Dahmen and A. Kunoth. Adaptive wavelet methods for linear-quadratic elliptic control problems: convergence rates. *SIAM J. Control Optim.*, 43(5):1640–1675, 2005 (electronic).

[DKT98] R.A. DeVore, S.V. Konyagin, and V.N. Temlyakov. Hyperbolic wavelet approximation. *Constr. Approx.*, 14(1):1–26, 1998.

[DL92] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 5*. Springer-Verlag, Berlin, 1992. Evolution problems I.

[DS98] W. Dahmen and R. Schneider. Wavelets with complementary boundary conditions—function spaces on the cube. *Results Math.*, 34(34):255–293, 1998.-

[DS99a] W. Dahmen and R. Schneider. Composite wavelet bases for operator equations. *Math. Comp.*, 68:1533–1567, 1999.

[DS99b] W. Dahmen and R. Schneider. Wavelets on manifolds I: Construction and domain decomposition. *SIAM J. Math. Anal.*, 31:184–230, 1999.

[DS09a] M. Dauge and R.P. Stevenson. Sparse tensor product wavelet approximation of singular functions. Technical report, 2009. Submitted.

[DS09b] T.J. Dijkema and R.P. Stevenson. A sparse Laplacian in tensor product wavelet coordinates. Technical report, 2009. Submitted.

[DSS08] T.J. Dijkema, Ch. Schwab, and R.P. Stevenson. An adaptive wavelet method for solving high-dimensional elliptic PDEs. Technical report, 2008. To appear in *Constr. Approx.*

[DSX00a] W. Dahmen, R. Schneider, and Y. Xu. Nonlinear functionals of wavelet expansions—adaptive reconstruction and fast evaluation. *Numer. Math.*, 86(1):49–101, 2000.

[DSX00b] W. Dahmen, R. Schneider, and Y. Xu. Nonlinear functionals of wavelet expansions—adaptive reconstruction and fast evaluation. *Numer. Math.*, 86(1):49–101, 2000.

[DUV02] W. Dahmen, K. Urban, and J. Vorloeper. Adaptive wavelet methods - Basic concepts and applications to the Stokes problem. In D.-X. Zhou, editor, *Wavelet Analysis*, New Jersey, 2002. World Scientific.

[Gan06] T. Gantumur. *Adaptive Wavelet Algorithms for solving operator equations*. PhD thesis, Department of Mathematics, Utrecht University, 2006.

[GHS07] T. Gantumur, H. Harbrecht, and R.P. Stevenson. An optimal adaptive wavelet method without coarsening of the iterands. *Math. Comp.*, 76:615–629, 2007.

[GK00] M. Griebel and S. Knapek. Optimized tensor-product approximation spaces. *Constr. Approx.*, 16(4):525–540, 2000.

[GO95] M. Griebel and P. Oswald. Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems. *Adv. Comput. Math.*, 4(12):171–206, 1995.–

[GR87] L. Greengard and V. Rokhlin. A fast algorithm for particle simulation. *J. Comput. Phys.*, 73:325–348, 1987.

[Gra04] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(34):247–265, 2004.-

[GS06a] T. Gantumur and R.P. Stevenson. Computation of differential operators in wavelet coordinates. *Math. Comp.*, 75:697–709, 2006.

[GS06b] T. Gantumur and R.P. Stevenson. Computation of singular integral operators in wavelet coordinates. *Computing*, 76:77–107, 2006.

[Hac92] W. Hackbusch. *Elliptic Differential Equations. Theory and Numerical Treatment*. Springer-Verlag, Berlin, 1992.

[HK07] W. Hackbusch and B.N. Khoromskij. Tensor-product approximation to operators and functions in high dimensions. *J. Complexity*, 23(46):697–714, 2007.-

[HN89] W. Hackbusch and Z.P. Nowak. On the fast matrix multiplication in the boundary element by panel clustering. *Numer. Math.*, 54:463–491, 1989.

[HS05] V.H. Hoang and Ch. Schwab. High-dimensional finite elements for elliptic problems with multiple scales. *SIAM J. Multiscale Model. Simul.*, 3(1):168–194, 2005.

[HS06] H. Harbrecht and R. Stevenson. Wavelets with patchwise cancellation properties. *Math. Comp.*, 75(256):1871–1889, 2006.

[HSS08] H. Harbrecht, R. Schneider, and Ch. Schwab. Sparse second moment analysis for elliptic problems in stochastic domains. *Numer. Math.*, 109(3):385–414, 2008.

[LLKW05] D. Labate, W-Q. Lim, G. Kutyniok, and G. Weiss. Sparse multidimensional representation using shearlets. In M. Papadakis, A. Laine, and M. Unser, editors, *Wavelets XI*, SPIE, Proc. 5914, pages 254–262, 2005.

[Met02] A. Metselaar. *Handling Wavelet Expansions in Numerical Methods*. PhD thesis, University of Twente, 2002.

[Nit05] P.-A. Nitsche. Sparse approximation of singularity functions. *Constr. Approx.*, 21(1):63–81, 2005.

[Nit06] P.-A. Nitsche. Best N-term approximation spaces for tensor product wavelet bases. *Constr. Approx.*, 24(1):49–70, 2006.

[NW08] E. Novak and H. Woźniakovski. Approximation of infinitely differentiable multivariate functions is intractable. Technical report, 2008. To appear in *Journal of Complexity*.

[Rei08] N.C. Reich. *Wavelet compression of anisotropic integrodifferential operators on sparse tensor product spaces*. PhD thesis, ETH, Zürich, 2008.

[Saa03] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[Sch90] H.A. Schwarz. *Gesammelte Mathematische Abhandlungen*. volume 2, pages 133–143, Berlin, 1890. Springer.

[SS08] Ch. Schwab and R.P. Stevenson. Adaptive wavelet algorithms for elliptic PDEs on product domains. *Math. Comp.*, 77:71–92, 2008.

[SS09] Ch. Schwab and R.P. Stevenson. A space-time adaptive wavelet method for parabolic evolution problems. Technical report, 2009. To appear in *Math. Comp.*

[ST03] Ch. Schwab and R.A. Todor. Sparse finite elements for stochastic elliptic problems—higher order moments. *Computing*, 71(1):43–63, 2003.

[Ste03] R.P. Stevenson. Adaptive solution of operator equations using wavelet frames. *SIAM J. Numer. Anal.*, 41(3):1074–1100, 2003.

[Ste04] R.P. Stevenson. On the compressibility of operators in wavelet coordinates. *SIAM J. Math. Anal.*, 35(5):1110–1132, 2004.

[Ste07] R.P. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[SW08] R.P. Stevenson and M. Werner. Computation of differential operators in aggregated wavelet frame coordinates. *IMA J. Numer. Anal.*, 28(2):354–381, 2008.

[SW09] R.P. Stevenson and M. Werner. A multiplicative schwarz adaptive wavelet method for elliptic boundary value problems. *Math. Comp.*, 78:619–644, 2009.

[Urb09] K. Urban. *Wavelet Methods for Elliptic Partial Differential Equations*. Oxford University Press, 2009.

[vPS06] T. von Petersdorff and Ch. Schwab. Sparse finite element methods for operator equations with stochastic data. *Appl. Math.*, 51(2):145–180, 2006.

[vS04] J. van den Eshof and G.L.G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 26(1):125–153, 2004.

[Wlo82] J. Wloka. *Partielle Differentialgleichungen*. B.G. Teubner, Stuttgart, 1982. Sobolevräume und Randwertaufgaben.

[Xu92] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.

[Zen91] Ch. Zenger. Sparse grids. In *Parallel algorithms for partial differential equations (Kiel, 1990)*, volume 31of *Notes Numer. Fluid Mech.*, pages 241–251. Vieweg, Braunschweig, 1991.

Optimal multilevel methods for $H(\mathrm{grad})$, $H(\mathrm{curl})$, and $H(\mathrm{div})$ systems on graded and unstructured grids

Jinchao Xu, Long Chen, and Ricardo H. Nochetto

We give an overview of multilevel methods, such as V-cycle multigrid and BPX preconditioner, for solving various partial differential equations (including $H(\mathrm{grad})$, $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems) on quasi-uniform meshes and extend them to graded meshes and completely unstructured grids. We first discuss the classical multigrid theory on the basis of the method of subspace correction of Xu and a key identity of Xu and Zikatanov. We next extend the classical multilevel methods in $H(\mathrm{grad})$ to graded bisection grids upon employing the decomposition of bisection grids of Chen, Nochetto, and Xu. We finally discuss a class of multilevel preconditioners developed by Hiptmair and Xu for problems discretized on unstructured grids and extend them to $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems over graded bisection grids.

1 Introduction

How to effectively solve the large scale algebraic systems arising from the discretization of partial differential equations is a fundamental problem in scientific and engineering computing. In this paper, we give an overview of a special class of methods for solving such systems: multilevel iterative methods based on the method of subspace corrections [18, 91] and the method of auxiliary spaces [92, 52].

Jinchao Xu
Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA and LMAM, The School of Mathematical Sciences, Peking University, China.
e-mail: xu@math.psu.edu.

Long Chen
Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA.
e-mail: chenlong@math.uci.edu.

Ricardo H. Nochetto
Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA.
e-mail: rhn@math.umd.edu.

R.A. DeVore, A. Kunoth (eds.), *Multiscale, Nonlinear and Adaptive Approximation*,
DOI 10.1007/978-3-642-03413-8_14,

The method of subspace corrections proves to be a very useful general framework for the design and analysis of various iterative methods. We give a rather detailed description of this method in Section §2 and apply it to additive and multiplicative multilevel methods. Of special interest is the sharp convergence identity of Xu and Zikatanov [94], which we also prove.

Most of the multilevel methods are dictated by the underlying mesh structure. In this paper, roughly speaking, we consider the following three types of grids:

- Quasi-uniform (and structured) grids with a hierarchy of nested sub-grids.
- Graded grids obtained by bisection with a hierarchy of nested sub-grids.
- Unstructured grids without a hierarchy of sub-grids.

Multilevel methods on quasi-uniform grids

The theoretical and algorithmic development of most traditional multilevel methods are devoted to quasi-uniform structured grids; see Brandt [21], Hackbusch [44], Xu [91, 16], and Yserentant [96]. In Section §3, using the method of subspace correction framework [18, 91], we discuss the classical V-cycle multigrid method and the BPX preconditioner. We also include a recent result by Xu and Zhu [93] that demonstrates that the conjugate gradient method with classical V-cycle multigrid or BPX-preconditioner as preconditioners provides a robust method with respect to jump discontinuities of coefficients.

Multilevel methods on graded bisection grids

Multilevel algorithms for graded grids generated by adaptive finite element methods (AFEM) is one main topic to be discussed in this paper. AFEM are now widely used in scientific and engineering computation to optimize the relation between accuracy and computational labor (degrees of freedom). We refer to the survey to [63] for an introduction to the theory of AFEM.

Of all possible refinement strategies, we are interested in *bisection*, the most popular and effective procedure for refinement in any dimension; see [63] and the references therein. Our goal is to design optimal multilevel solvers and analyze them within the framework of highly graded meshes created by bisection, from now on called *bisection meshes*.

In Section §4, we present multilevel methods and analysis for $H(\mathrm{grad})$ based on the novel decomposition of bisection grids of Chen, Nochetto, and Xu [27], which is conceptually simple and dimension and polynomial degree independent. Roughly speaking, for any triangulation $\mathscr{T}_N$ constructed from $\mathscr{T}_0$ by bisection, we can write

$$\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}, \quad \mathscr{B} = (b_1, b_2, \cdots, b_N),$$

where $\mathscr{B}$ denotes a sequence of N elementary bisections b_i. Each such b_i is restricted to a local region and the corresponding local grid is quasi-uniform. This decom-

position serves as a general bridge to transfer results from quasi-uniform grids to graded bisection grids. We exploit this flexibility to design and analyze local multigrid methods for the $H(\text{curl})$ and $H(\text{div})$ systems in three dimensions in Section §5; we explicitly follow Chen, Nochetto, and Xu [28], which in turn build on Hiptmair and Xu [52].

Multilevel methods on unstructured grids

In practical applications, finite element grids are often unstructured, namely, they have no natural geometric hierarchy that can be extracted from the mesh data structure and used for designing optimal multilevel algorithms. For such problems we turn to algebraic multigrid methods (AMG). What makes AMG attractive in practice is that they generate coarse-level equations without using any (or much) geometric information or re-discretization on the coarse levels. Despite the lack of rigorous theoretical justification, AMG methods are very successful in practice for various Poisson-like equations; see [73, 81] and reference therein.

Even though we do not describe AMG in any detail, in Section §6 we present a technique developed by Hiptmair and Xu [52] for quasi-uniform meshes that converts the solution of both $H(\text{curl})$ and $H(\text{div})$ systems into that of a number of Poisson-like equations, which can be efficiently solved by AMG.

2 The method of subspace corrections

Most partial differential equations, after discretization, are reduced to solve some linear algebraic equations in the form

$$Au = f, \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ is a sparse matrix and $f \in \mathbb{R}^N$. How to solve (1) efficiently remains a basic question in numerical PDEs (and in all scientific computing). The Gaussian elimination still remains the most commonly used method in practice. It is a black-box as it can be applied to any problem in principle. But it is expensive: for a general $N \times N$ matrix, it required $\mathcal{O}(N^3)$ operations. For a sparse matrix, it may require less operations but still too expensive for large scale problems. Multigrid methods, on the other hand, are examples of problem-oriented algorithms, which, for some problems, only require $\mathcal{O}(N|\log N|^\sigma), \sigma > 0$, operations. In this section, we will give some general and basic results that will be used in later sections to construct efficient iterative methods (such as multigrid methods) for discretized partial differential equations.

Following [91], we shall use notation $x \lesssim y$ to stand for $x \le Cy$. We also use $x \eqsim y$ to mean $x \lesssim y$ and $y \lesssim x$.

2.1 Iterative methods

2.1.1 Basic iterative method

In general, a basic linear iterative method for $Au = f$ can be written in the following form:

$$u^{k+1} = u^k + B(f - Au^k),$$

starting from an initial guess $u^0 \in \mathscr{V}$. It can be interpreted as a result of the following three steps:

1. form the residual $r = f - Au^k$;
2. solve the residual equation $Ae = r$ approximately by $\hat{e} = Br$ with $B \approx A^{-1}$;
3. correct the solution $u^{k+1} = u^k + \hat{e}$.

Here B is called *iterator*. As simple examples, if $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ and $A = D + L + U$, we may take $B = D^{-1}$ to obtain the Jacobi method and $B = (D+L)^{-1}$ to obtain the Gauss-Seidel method.

The art of constructing *efficient* iterative methods lies on the design of B which captures the essential information of A^{-1} and its action is easily computable. In this context the notion of "efficient" implies two essential requirements:

1. One iteration requires a computational effort proportional to the number of unknowns.
2. The rate of convergence is well below 1 and independent with the number of unknowns.

2.1.2 Preconditioned Krylov space methods

The approximate inverse B, when it is SPD, can be used as a preconditioner for Conjugate Gradient (CG) method. The resulting method, known as preconditioned conjugate gradient method (PCG), admits the following error estimate:

$$\frac{\|u-u^k\|_A}{\|u-u^0\|_A} \le 2\left(\frac{\sqrt{\kappa(BA)}-1}{\sqrt{\kappa(BA)}+1}\right)^k (k \ge 1), \quad \left(\kappa(BA) = \frac{\lambda_{\max}(BA)}{\lambda_{\min}(BA)}\right).$$

Here B is called *preconditioner*. A good preconditioner should have the properties that the action of B is easy to compute and that $\kappa(BA)$ is significantly smaller than $\kappa(A)$.

An interesting fact is that the linear iterative method using iterator B may not be convergent at all whereas B can always be a preconditioner. For example, the Jacobi method is not convergent for all SPD systems, but $B = D^{-1}$ can always be used as a preconditioner which is often known as the diagonal preconditioner.

2.1.3 Convergence analysis

Let $e^k = u - u^k$. The error equation of the basic iterative method is

$$e^{k+1} = (I - BA)e^k = (I - BA)^k e^0.$$

Thus the basic iterative method converges if and only if the spectral radius of the error operator $I - BA$ is less than one, i.e., $\rho(I - BA) < 1$.

Given an iterator B, we define the iteration operator $\Phi_B u = u + B(f - Au)$ and introduce a symmetric scheme $\Phi_{\overline{B}} = \Phi_{B^t}\Phi_B$. The convergence of the iteration scheme Φ_B and its symmetrization $\Phi_{\overline{B}}$ is connected by the following inequality:

$$\rho(I - BA) \le \sqrt{\rho(I - \overline{B}A)},$$

and the equality holds if $B = B^t$. Hence we shall focus on the analysis of the symmetric scheme.

By definition, we have the following formula for the error operator $I - \overline{B}A$

$$I - \overline{B}A = (I - B^t A)(I - BA), \text{ and thus } \overline{B} = B^t(B^{-t} + B^{-1} - A)B. \tag{2}$$

Since $\overline{B}$ is symmetric, $I - \overline{B}A$ is symmetric with respect to the inner product $(u, v)_A := (Au, v)$. Indeed, let $(\cdot)^*$ be the adjoint operator with respect to $(\cdot,\cdot)_A$, it is easy to show

$$I - \overline{B}A = (I - BA)^*(I - BA). \tag{3}$$

Consequently, $I - \overline{B}A$ is SPD with respect to $(\cdot,\cdot)_A$ and $\lambda_{\max}(\overline{B}A) < 1$. Therefore

$$\rho(I - \overline{B}A) = \max\{|1 - \lambda_{\min}(\overline{B}A)|, |1 - \lambda_{\max}(\overline{B}A)|\} = 1 - \lambda_{\min}(\overline{B}A). \tag{4}$$

A more quantitative information on $\lambda_{\min}(\overline{B}A)$ is given in the following lemma.

Lemma 2.1 (Least Eigenvalue). *When B is symmetric and nonsingular,*

$$\lambda_{\min}(BA) = \inf_{u \in \mathscr{V}\setminus\{0\}} \frac{(ABAu, u)}{(Au, u)} = \inf_{u \in \mathscr{V}\setminus\{0\}} \frac{(Au, u)}{(B^{-1}u, u)} = \left(\sup_{u \in \mathscr{V}\setminus\{0\}} \frac{(B^{-1}u, u)}{(Au, u)} \right)^{-1}.$$

Proof. The first two identities comes from the fact BA is symmetric with respect to $(\cdot,\cdot)_A$ and $(\cdot,\cdot)_{B^{-1}}$. The third identity comes from

$$\lambda_{\min}^{-1}(BA) = \lambda_{\max}((BA)^{-1}) = \sup_{u \in \mathscr{V}\setminus\{0\}} \frac{((BA)^{-1}u, u)_A}{(u, u)_A} = \sup_{u \in \mathscr{V}\setminus\{0\}} \frac{(B^{-1}u, u)}{(Au, u)}.$$

This completes the proof. □

2.2 Space decomposition and method of subspace correction

In the spirit of dividing and conquering, we shall decompose the space $\mathscr{V}$ as the summation of subspaces. Then the original problem (1) can be split into sub-problems with smaller sizes which are relatively easier to solve.

Let $\mathscr{V}_i \subset \mathscr{V}$, $i = 0,\dots,J$, be subspaces of $\mathscr{V}$. If $\mathscr{V} = \sum_{i=0}^{J} \mathscr{V}_i$, then $\{\mathscr{V}_i\}_{i=0}^{J}$ is called a *space decomposition* of $\mathscr{V}$, and we can write $u = \sum_{i=0}^{J} u_i$. Since $\sum_{i=0}^{J} \mathscr{V}_i$ is not necessarily a direct sum, decompositions of u are in general not unique.

Throughout this paper, we use the following operators, for $i = 0, 1, \dots, J$:

- $Q_i : \mathscr{V} \mapsto \mathscr{V}_i$ the projection with the inner product $(\cdot,\cdot)$;
- $I_i : \mathscr{V}_i \mapsto \mathscr{V}$ the natural inclusion which is often called prolongation;
- $P_i : \mathscr{V} \mapsto \mathscr{V}_i$ the projection with the inner product $(\cdot,\cdot)_A$;
- $A_i : \mathscr{V}_i \mapsto \mathscr{V}_i$ the restriction of A to the subspace $\mathscr{V}_i$;
- $R_i : \mathscr{V}_i \mapsto \mathscr{V}_i$ an approximation of A_i^{-1} (often known as smoother).

It is easy to verify the relation $Q_i A = A_i P_i$ and $Q_i = I_i^t$. The operator I_i^t is often called restriction. If $R_i = A_I^{-1}$, then we have an exact local solver and $R_i Q_i A = P_i$.

For a given residual $r \in \mathscr{V}$, we let $r_i = I_i^t r$ denote the restriction of the residual to the subspace and solve the residual equation in the subspaces

$$A_i e_i = r_i \quad \text{by} \quad \hat{e}_i = R_i r_i.$$

Subspace corrections $\hat{e}_i$ are assembled to give a correction in the space $\mathscr{V}$ and therefore is called the method of subspace correction. There are two basic ways to assemble subspace corrections.

Parallel Subspace Correction (PSC)

This method performs the correction on each subspace in parallel. In operator form, it reads

$$u^{k+1} = u^k + B(f - Au^k), \tag{5}$$

where

$$B = \sum_{i=0}^{J} I_i R_i I_i^t. \tag{6}$$

The subspace correction is $\hat{e}_i = I_i R_i I_i^t (f - Au^k)$, and the correction in $\mathscr{V}$ is $\hat{e} = \sum_{i=0}^{J} \hat{e}_i$.

Successive subspace correction (SSC)

This method performs the correction in a successive way. In operator form, it reads

$$v^0 = u^k, \quad v^{i+1} = v^i + I_i R_i I_i^t (f - A v^i), i = 0, \ldots, J, \quad u^{k+1} = v^{J+1}. \tag{7}$$

We have the following error formulae for PSC and SSC:

- Parallel Subspace Correction (PSC):

$$u - u^{k+1} = \left[I - \Big(\sum_{i=0}^{J} I_i R_i I_i^t \Big) A \right] (u - u^k);$$

- Successive Subspace Correction (SSC):

$$u - u^{k+1} = \left[\prod_{i=0}^{J} (I - I_i R_i I_i^t A) \right] (u - u^k).$$

Thus PSC is also called additive method while SSC is called multiplicative method. In the notation $\prod_{i=0}^{J} a_i$, we assume there is a build-in ordering from $i = 0$ to J, i.e., $\prod_{i=0}^{J} a_i = a_0 a_1 \ldots a_J$.

As a trivial example, we consider the space decomposition $\mathbb{R}^J = \sum_{i=1}^{J} \text{span}\{e_i\}$. In this case, if we use exact (one dimensional) subspace solvers, the resulting SSC is just the Gauss-Seidel method and the PSC is just the Jacobi method. More complicated examples, including multigrid methods and multilevel preconditioners, will be discussed later on.

PSC or SSC can be also understood as Jacobi or Gauss-Seidel methods for a bigger equation in the product space [43, 94], respectively. The analysis of classical iterative methods can then be applied to more advanced PSC or SSC methods.

Given a decomposition $\mathcal{V} = \sum_{i=0}^{J} \mathcal{V}_i$, we can construct a product space $\tilde{\mathcal{V}} = \mathcal{V}_0 \times \mathcal{V}_1 \times \ldots \times \mathcal{V}_J$, with an inner product $(\tilde{u}, \tilde{v})_{\tilde{\mathbb{V}}} = \sum_{i=0}^{J} (u_i, v_i)$. We will reformulate the linear operator equation $Au = f$ to an equation posed on $\tilde{\mathcal{V}} : \tilde{A}\tilde{u} = \tilde{f}$.

Let us introduce the operator $\mathcal{R} : \tilde{\mathcal{V}} \to \mathcal{V}$ by $\mathcal{R}\tilde{u} = \sum_{i=0}^{J} u_i$. Because of the decomposition $\mathcal{V} = \sum_{i=0}^{J} \mathcal{V}_i$, $\mathcal{R}$ is surjective. In general$\mathcal{R}$ is not injective but it will be in the quotient space $\overline{\mathcal{V}} = \tilde{\mathcal{V}} / \ker(\mathcal{R})$. We define $\mathcal{R}^* : \mathcal{V} \mapsto \tilde{\mathcal{V}}$, the adjoint of $\mathcal{R}$ with respect to $(\cdot, \cdot)_A$, to be

$$(\mathcal{R}^* u, \tilde{v})_{\tilde{\mathcal{V}}} := (u, \mathcal{R}\tilde{v})_A = \sum_{i=0}^{J} (u, v_i)_A = \sum_{i=0}^{J} (Q_i A u, v_i), \quad \text{for all } \tilde{v} = (v_i)_{i=0}^{J} \in \tilde{\mathcal{V}}.$$

Therefore

$$\mathcal{R}^* = (Q_0 A, Q_1 A, \cdots, Q_J A)^t.$$

Similarly, the transpose $\mathcal{R}^t : \mathcal{V} \mapsto \tilde{\mathcal{V}}$ of $\mathcal{R}$ with respect to $(\cdot, \cdot)$ is

$$\mathcal{R}^t = (Q_0, Q_1, \cdots, Q_J)^t.$$

Since $\mathcal{R}$ is surjective, we conclude that $\mathcal{R}^t$ is injective. Let $\tilde{A} = \mathcal{R}^* \mathcal{R}$ and $\tilde{f} = \mathcal{R}^t f$. If $\tilde{u}$ is a solution of $\tilde{A}\tilde{u} = \tilde{f}$, it is straightforward to verify that then $u = \mathcal{R}\tilde{u}$ is the solution of $Au = f$.

SSC as Gauss-Seidel Method

The new formulation of the problem is used to characterize SSC for solving $Au = f$ as a Gauss-Seidel method for $\tilde{A}\tilde{u} = \tilde{f}$. In the sequel, we consider the SSC applied to the space decomposition $\mathscr{V} = \sum_{k=0}^{J} \mathscr{V}_j$ with $R_i = A_i^{-1}$, namely we solve the problem posed on the subspaces exactly.

Let $\tilde{A} = \tilde{D} + \tilde{L} + \tilde{U}$ and $\tilde{B} = (\tilde{D} + \tilde{L})^{-1}$. Then SSC for $Au = f$ with exact local solvers $R_i = A_i^{-1}$ is equivalent to the Gauss-Seidel method for solving $\tilde{A}\tilde{u} = \tilde{f}$:

$$\tilde{u}^{k+1} = \tilde{u}^k + \tilde{B}(\tilde{f} - \tilde{A}\tilde{u}^k). \tag{8}$$

The verification of the equivalence is as follows. We first compute the entries for $\tilde{A} = (\tilde{a}_{ij})_{(J+1)\times(J+1)}$. By definition,

$$\tilde{a}_{ij} = Q_i A I_j = A_i P_i I_j : \mathscr{V}_j \mapsto \mathscr{V}_i.$$

In particular $\tilde{a}_{ii} = A_i : \mathscr{V}_i \mapsto \mathscr{V}_i$ is SPD on $\mathscr{V}_i$.

We can write the standard Gauss-Seidel method using iterator $\tilde{B} = (\tilde{D} + \tilde{L})^{-1}$ as

$$\tilde{u}^{k+1} = \tilde{u}^k + \tilde{D}^{-1}(\tilde{f} - \tilde{L}\tilde{u}^{k+1} - (\tilde{D} + \tilde{U})\tilde{u}^k).$$

The component-wise formula is

$$\begin{aligned} u_i^{k+1} &= u_i^k + A_i^{-1}(f_i - \sum_{j=0}^{i-1} \tilde{a}_{ij} u_j^{k+1} - \sum_{j=i}^{J} \tilde{a}_{ij} u_j^k) \\ &= u_i^k + A_i^{-1} Q_i (f - A \sum_{j=0}^{i-1} u_j^{k+1} - A \sum_{j=i}^{J} u_j^k). \end{aligned}$$

Let

$$v^i = \sum_{j=0}^{i-1} u_j^{k+1} + \sum_{j=i}^{J} u_j^k.$$

Noting that $v^i - v^{i-1} = u_i^{k+1} - u_i^k$, we then get

$$v^i = v^{i-1} + A_i^{-1} Q_i (f - A v^{i-1}),$$

which is the correction on $\mathscr{V}_i$.

Similarly one can easily verify that PSC using exact local solvers $R_i = A_i^{-1}$ is equivalent to the Jacobi method for solving the large system $\tilde{A}\tilde{u} = \tilde{f}$.

2.3 Sharp convergence identities

The analysis of additive multilevel operator relies on the following identity which is well known in the literature [87, 91, 42, 94]. For completeness, we include a concise proof taken from [94].

Theorem 2.1 (Identity for PSC). *If R_i is SPD on $\mathscr{V}_i$ for $i = 0, \ldots, J$, then B defined by (6) is also SPD on $\mathscr{V}$. Furthermore*

$$(B^{-1}v, v) = \inf_{\sum_{i=0}^{J} v_i = v} \sum_{i=0}^{J} (R_i^{-1} v_i, v_i), \tag{9}$$

and

$$\lambda_{\min}(BA)^{-1} = \sup_{\|v\|_A = 1} \inf_{\sum_{i=0}^{J} v_i = v} (R_i^{-1} v_i, v_i). \tag{10}$$

Proof. Note that B is symmetric, and

$$(Bv, v) = (\sum_{i=0}^{J} I_i R_i I_i^t v, v) = \sum_{i=0}^{J} (R_i Q_i v, Q_i v),$$

whence B is invertible and thus SPD. We now prove (9) by constructing a decomposition achieving the infimum. Let $v_i^* = R_i Q_i B^{-1} v$, $i = 0, \ldots, J$. By definition of B, we get a special decomposition $\sum_i v_i^* = v$, and

$$\begin{aligned}\inf_{\sum v_i = v} \sum_{i=0}^{J} (R_i^{-1} v_i, v_i) &= \inf_{\sum w_i = 0} \sum_{i=0}^{J} (R_i^{-1}(v_i^* + w_i), v_i^* + w_i) \\ &= \sum_{i=0}^{J} (R_i^{-1} v_i^*, v_i^*) + \inf_{\sum w_i = 0} \Big[\sum_{i=0}^{J} 2(R_i^{-1} v_i^*, w_i) + \sum_{i=0}^{J} (R_i^{-1} w_i, w_i) \Big]\end{aligned}$$

Since

$$\sum_{i=0}^{J} (R_i^{-1} v_i^*, u_i) = \sum_{i=0}^{J} (B^{-1} v, u_i) = (B^{-1} v, \sum_{i=0}^{J} u_i)$$

for all $(u_i)_{i=0}^{J} \in \mathscr{V}$, we deduce

$$\begin{aligned}\inf_{\sum v_i = v} \sum_{i=0}^{J} (R_i^{-1} v_i, v_i) &= (B^{-1} v, \sum_{i=0}^{J} v_i^*) \\ &\quad + \inf_{\sum w_i = 0} \Big[2(B^{-1} v, \sum_{i=0}^{J} w_i) + \sum_{i=0}^{J} (R_i^{-1} w_i, w_i) \Big] = (B^{-1} v, v).\end{aligned}$$

The proof of the equality (10) is a simple consequence of Lemma 2.1. □

As for additive methods, we now present an identity developed by Xu and Zikatanov [94] for multiplicative methods. For simplicity, we focus on the case $R_i = A_i^{-1}$, $i = 0, \ldots, J$, i.e., the subspace solvers are exact. In this case $I - I_i R_i I_i^t A = I - P_i$.

Theorem 2.2 (X-Z Identity for SSC). *The following identity is valid*

$$\Big\|\prod_{i=0}^{J}(I-P_i)\Big\|_A^2 = 1-\frac{1}{1+c_0}, \tag{11}$$

with

$$c_0 = \sup_{\|v\|_A=1}\inf_{\sum_{i=0}^J v_i=v}\sum_{i=0}^{J}\Big\|P_i\sum_{j=i+1}^{J}v_j\Big\|_A^2. \tag{12}$$

Proof. Recall that SSC for solving $Au=f$ with exact local solvers $R_i=A_i^{-1}$ is equivalent to the Gauss-Seidel method for solving $\tilde{A}\tilde{u}=\tilde{f}$ using iterator $\tilde{B}=(\tilde{D}+\tilde{L})^{-1}$. Let $\overline{B}$ be the symmetrization of $\tilde{B}$ from (2). Direct computation yields

$$\overline{B}^{-1}=\tilde{A}+\tilde{L}\tilde{D}^{-1}\tilde{U}. \tag{13}$$

On the quotient space $\overline{\mathscr{V}}=\tilde{V}/\ker(\mathscr{R})$, $\tilde{A}$ is SPD and thus defines an inner produce $(\cdot,\cdot)_{\tilde{A}}$. Using Lemma 2.1 and (13), we have

$$\begin{aligned}\|\tilde{I}-\tilde{B}\tilde{A}\|_{\tilde{A}}^2 &= \|\tilde{I}-\overline{B}\tilde{A}\|_{\tilde{A}} = 1-\left[\sup_{\tilde{v}\in\overline{\mathscr{V}}\setminus\{0\}}\frac{(\overline{B}^{-1}\tilde{v},\tilde{v})_{\bar{\mathscr{V}}}}{(\tilde{A}\tilde{v},\tilde{v})_{\bar{\mathscr{V}}}}\right]^{-1}\\ &= 1-\left[1+\sup_{\tilde{v}\in\overline{\mathscr{V}}\setminus\{0\}}\frac{(\tilde{D}^{-1}\tilde{U}\tilde{v},\tilde{U}\tilde{v})}{(\tilde{A}\tilde{v},\tilde{v})}\right]^{-1}.\end{aligned}$$

To finish the proof, we verify that

$$\sup_{\tilde{v}\in\overline{\mathscr{V}},\tilde{v}\neq 0}\frac{(\tilde{D}^{-1}\tilde{U}\tilde{v},\tilde{U}\tilde{v})}{(\tilde{A}\tilde{v},\tilde{v})} = \sup_{v\in\mathscr{V},\|v\|_A=1}\inf_{\sum v_i=v}\sum_{i=0}^{J}\|P_i\sum_{j=i+1}^{J}v_j\|_A^2.$$

For any $\tilde{v}\in\overline{\mathscr{V}}$, and corresponding $v=\mathscr{R}\tilde{v}$, we have

$$(\tilde{A}\tilde{v},\tilde{v})_{\bar{\mathscr{V}}} = (\mathscr{R}^*\mathscr{R}\tilde{v},\tilde{v})_{\bar{\mathscr{V}}} = (\mathscr{R}\tilde{v},\mathscr{R}\tilde{v})_A = (v,v)_A,$$

and

$$(\tilde{D}^{-1}\tilde{U}\tilde{v},\tilde{U}\tilde{v})_{\bar{\mathscr{V}}} = \sum_{i=0}^{J}(A_i^{-1}\sum_{j=i+1}^{J}A_iP_iv_j,\sum_{j=i+1}^{J}A_iP_iv_j)$$

because $Q_iA=A_iP_i$ and $\sum_{j=i+1}^{J}Q_jAv_j=A_iP_i\sum_{j=i+1}^{J}v_j$. Consequently,

$$(\tilde{D}^{-1}\tilde{U}\tilde{v},\tilde{U}\tilde{v})_{\bar{\mathscr{V}}} = \sum_{i=0}^{J}(\sum_{j=i+1}^{J}P_iv_j,A_i\sum_{j=i+1}^{J}P_iv_j) = \sum_{i=0}^{J}\|\sum_{j=i+1}^{J}P_iv_j\|_A^2.$$

Since $\tilde{v}\in\overline{\mathscr{V}}$, we should use the quotient norm (which gives the inf) to finish the proof. □

For SSC method with general smoothers, we present the following sharp estimate of Xu and Zikatanov [94] (see also [30]). We refer to [94, 30] for a proof.

Theorem 2.3 (X-Z General Identity for SSC). *The SSC is convergent if each subspace solver $T_i = R_i Q_i A$ is convergent. Furthermore*

$$\left\| \prod_{i=1}^{J} (I - T_i) \right\|_A^2 = 1 - \frac{1}{K}, \quad K = 1 + \sup_{\|v\|=1} \inf_{\sum_i v_i = v} \sum_{i=1}^{J} \|T_i^* w_i\|^2_{\overline{T}_i^{-1}} \tag{14}$$

where $w_i = \sum_{j=i}^{J} v_i - T_i^{-1} v_i$ and $\overline{T}_i := T_i^* + T_i - T_i^* T_i$.

3 Multilevel methods on quasi-uniform grids

In this section, we apply PSC and SSC to the finite element discretization of second order elliptic equations. We use theory developed in the previous section to give a convergence analysis of multilevel iteration methods.

3.1 Finite element methods

For simplicity we illustrate the technique by considering the linear finite element method for the Poisson equation.

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega, \tag{15}$$

where $\Omega \subset \mathbb{R}^d$ is a polyhedral domain.

3.1.1 Weak formulation

The weak formulation of (15) reads: given an $f \in H^{-1}(\Omega)$ find $u \in H_0^1(\Omega)$ so that

$$a(u, v) = \langle f, v \rangle \quad \text{for all } v \in H_0^1(\Omega), \tag{16}$$

where

$$a(u, v) = (\nabla u, \nabla v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x,$$

and $\langle \cdot, \cdot \rangle$ is the duality pair between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

By the Poincaré inequality, $a(\cdot, \cdot)$ defines an inner product on $H_0^1(\Omega)$. Thus by the Riesz representation theorem, for any $f \in H^{-1}(\Omega)$, there exists a unique $u \in H_0^1(\Omega)$ such that (16) holds. Furthermore, we have the following regularity result. There exists $\alpha \in (0, 1]$ which depends on the smoothness of $\partial\Omega$ such that

$$\|u\|_{1+\alpha} \lesssim \|f\|_{\alpha-1}. \tag{17}$$

This inequality is valid if Ω is convex or $\partial\Omega$ is $C^{1,1}$.

3.1.2 Triangulation and properties

Let Ω be a polyhedral domain in $\mathbb{R}^d$. A triangulation $\mathscr{T}$ (also called mesh or grid) of Ω is a partition of $\overline{\Omega}$ into a set of d-simplexes.

We impose two conditions on a triangulation $\mathscr{T}$ which are important in finite element construction. First, a triangulation $\mathscr{T}$ is called *conforming* or *compatible* if the intersection of any two simplexes τ and τ' in $\mathscr{T}$ is either empty or a common lower dimensional simplex.

The second important condition is shape regularity. A set of triangulations $\mathbb{T}$ is called *shape regular* if there exists a constant σ_1 such that

$$\max_{\tau\in\mathscr{T}} \frac{\operatorname{diam}(\tau)^d}{|\tau|} \leq \sigma_1, \quad \text{for all } \mathscr{T} \in \mathbb{T}, \tag{18}$$

where $\operatorname{diam}(\tau)$ is the diameter of τ and $|\tau|$ is the measure of τ in $\mathbb{R}^d$. For shape regular triangulations, $\operatorname{diam}(\tau) \eqsim h_\tau := |\tau|^{1/d}$ which will be used to represent the size of τ.

Furthermore, a shape regular class of triangulations $\mathbb{T}$ is called *quasi-uniform* if there exists a constant σ_2 such that

$$\frac{\max_{\tau\in\mathscr{T}} h_\tau}{\min_{\tau\in\mathscr{T}} h_\tau} \leq \sigma_2, \quad \text{for all } \mathscr{T} \in \mathbb{T}.$$

For a quasi-uniform triangulation $\mathscr{T}$, we simply call $h = \max_{\tau\in\mathscr{T}} h_\tau$ the mesh size and denote $\mathscr{T}$ by $\mathscr{T}_h$.

3.1.3 Finite element approximation

The standard finite element method is to solve problem (16) in a piecewise polynomial finite dimensional subspace. For simplicity we consider the piecewise linear finite element space $\mathscr{V}_h \subset H_0^1(\Omega)$ on quasi-uniform triangulations $\mathscr{T}_h$ of Ω:

$$\mathscr{V}_h := \{v \in H_0^1(\Omega) : v|_\tau \in \mathscr{P}_1(\tau) \text{ for all } \tau \in \mathscr{T}\}.$$

We now solve (16) in the finite element space $\mathscr{V}_h$: find $u_h \in \mathscr{V}_h$ such that

$$a(u_h, v_h) = \langle f, v_h\rangle, \quad \text{for all } v_h \in \mathscr{V}_h. \tag{19}$$

The existence and uniqueness of the solution to (19) follows again from the Riesz representation theorem since $\mathscr{V}_h \subset H_0^1(\Omega)$. By approximation and regularity theory,

we can easily get an error estimate on quasi-uniform grids

$$\|u - u_h\|_1 \lesssim h^\alpha \|u\|_{1+\alpha} \lesssim h^\alpha \|f\|_{\alpha-1},$$

where $\alpha > 0$ is determined by the regularity result (17). Thus u_h converges to u when $h \to 0$. When the solution u is rough, e.g., $\alpha \ll 1$, the convergence rate can be improved using adaptive grids [12, 79, 23, 63]. We will assume $\mathcal{V}_h$ is given, and the main objective of this paper is to discuss how to compute u_h efficiently. We focus on quasi-uniform grids in this section and on graded grids in the next section.

In this application, the SPD operator A is $(Au, v) = (\nabla u, \nabla v)$ and $\|\cdot\|_A$ is $|\cdot|_1$. For quasi-uniform mesh $\mathcal{T}_h$, let A_h be the restriction of A on the finite element space $\mathcal{V}_h$ over $\mathcal{T}_h$. We then end up with a linear operator equation $A_h : \mathcal{V}_h \mapsto \mathcal{V}_h$ that is

$$A_h u_h = f_h. \tag{20}$$

It is easy to see A_h is a self-adjoint operator in the Hilbert space $\mathcal{V}_h$ using L^2 inner product. To simplify notation in the sequel, we remove the subscript h when it is clear from the context and leads to no confusion.

It can be easily shown that $\kappa(A_h) \eqsim h^{-2}$ and the convergence rate of classical iteration methods, including Richardson, Jacobi, and Gauss-Seidel methods, for solving (19) is like

$$\rho \leq 1 - Ch^2.$$

Thus when $h \to 0$, we observe slow convergence of those classical iterative methods. We will construct efficient iterative methods using multilevel space decompositions.

3.2 Multilevel space decomposition and multigrid method

We first present a multilevel space decomposition. Let us assume that we have an initial quasi-uniform triangulation $\mathcal{T}_0$ and a nested sequence of triangulations $\{\mathcal{T}_k\}_{k=0}^J$ where $\mathcal{T}_k$ is obtained by the uniform refinement of $\mathcal{T}_{k-1}$ for $k > 0$. We then get a nested sequence (in the sense of trees [63]) of quasi-uniform triangulations

$$\mathcal{T}_0 \leq \mathcal{T}_1 \leq \cdots \leq \mathcal{T}_J = \mathcal{T}_h.$$

Note that h_k, the mesh size of $\mathcal{T}_k$, satisfies $h_k \eqsim \gamma^{2k}$ for some $\gamma \in (0,1)$, and thus $J \eqsim |\log h|$. Let $\mathcal{V}_k$ denote the corresponding linear finite element space of $H_0^1(\Omega)$ based on $\mathcal{T}_k$. We thus get a sequence of multilevel nested spaces

$$\mathcal{V}_0 \subset \mathcal{V}_1 \ldots \subset \mathcal{V}_J = \mathcal{V},$$

and a macro space decomposition

$$\mathcal{V} = \sum_{k=0}^{J} \mathcal{V}_k. \tag{21}$$

There is redundant overlapping in this multilevel decomposition, so the sum is not direct. The subspace solvers need only to take care of the "non-overlapping" components of the error (high frequencies in $\mathscr{V}_k$). For each subspace problem $A_k e_k = r_k$ posed on $\mathscr{V}_k$, we use a simple Richardson method

$$R_k = h_k^2 I_k,$$

where $I_k : \mathscr{V}_k \to \mathscr{V}_k$ is the identity and $h_k \approx \lambda_{\max}(A_k)$.

Let N_k be the dimension of $\mathscr{V}_k$, i.e., the number of interior vertices of $\mathscr{T}_k$. The standard nodal basis in $\mathscr{V}_k$ will be denoted by $\phi_{(k,i)}, i = 1, \cdots, N_k$. By our characterization of Richardson method, it is PSC method on the micro decomposition $\mathscr{V}_k = \sum_{i=1}^{N_k} \mathscr{V}_{(k,i)}$ with $\mathscr{V}_{(k,i)} = \operatorname{span}\{\phi_{(k,i)}\}$. In summary we choose the space decomposition:

$$\mathscr{V} = \sum_{k=0}^{J} \mathscr{V}_k = \sum_{k=0}^{J} \sum_{i=1}^{N_k} \mathscr{V}_{(k,i)}. \tag{22}$$

If we apply PSC to the decomposition (22) with $R_{(k,i)} = h_k^2 I_{(K,i)}$, we obtain $I_{(k,i)} R_{(k,i)} I_{(k,i)}^t u = h^{2-d}(u, \phi_{(k,i)})\phi_{(k,i)}$. The resulting operator B, according to (6), is the so-called BPX preconditioner [19]

$$Bu = \sum_{k=0}^{J} \sum_{i=1}^{N_k} h_k^{2-d}(u, \phi_{(k,i)})\phi_{(k,i)}. \tag{23}$$

If we apply SSC to the decomposition (22) with exact subspace solvers $R_i = A_i^{-1}$, we obtain a V-cycle multigrid method with Gauss-Seidel smoothers.

3.3 Stable decomposition and optimality of BPX preconditioner

For the optimality of the BPX preconditioner, we are to prove that the condition number $\kappa(BA)$ is uniformly bounded and thus PCG using BPX preconditioner converges in a fixed number of steps for a given tolerance regardless of the mesh size.

The estimate $\lambda_{\min}(BA) \gtrsim 1$ follows from the stability of the subspace decomposition. The first result is on the macro decomposition $\mathscr{V} = \sum_{k=0}^{J} \mathscr{V}_k$.

Lemma 3.1 (Stability of Macro Decomposition). *For any $v \in \mathscr{V}$, there exists a decomposition $v = \sum_{k=0}^{J} v_k$ with $v_k \in \mathscr{V}_k, k = 0, \ldots, J$ such that*

$$\sum_{k=0}^{J} h_k^{-2} \|v_k\|^2 \lesssim |v|_1^2. \tag{24}$$

Proof. Following the chronological development, we present two proofs. The first one uses full regularity and the second one minimal regularity.

1 *Full regularity* H^2: We assume $\alpha = 1$ in (17), which holds for convex polygons or polyhedrons. Recall that $P_k : \mathscr{V} \to \mathscr{V}_k$ is the projection onto $\mathscr{V}_k$ with the inner product $(u, v)_A = (\nabla u, \nabla v)$, and let $P_{-1} = 0$. We prove that the following decomposition

$$v = \sum_{k=0}^{J} (P_k - P_{k-1})v \tag{25}$$

satisfies (24). The full regularity assumption leads to the L^2 error estimate of P_k via a standard duality argument:

$$\|(I - P_k)v\| \lesssim h_k |(I - P_k)v|_1, \quad \text{for all } v \in H_0^1(\Omega). \tag{26}$$

Since $\mathscr{V}_{k-1} \subset \mathscr{V}_k$, we have $P_{k-1}P_k = P_{k-1}$ and

$$P_k - P_{k-1} = (I - P_{k-1})(P_k - P_{k-1}). \tag{27}$$

In view of (26) and (27), we have

$$\begin{aligned} \sum_{k=0}^{J} h_k^{-2} \|(P_k - P_{k-1})v\|^2 &= \sum_{k=0}^{J} h_k^{-2} \|(I - P_{k-1})(P_k - P_{k-1})v\|^2 \\ &\lesssim \sum_{k=0}^{J} |(P_k - P_{k-1})v|_{1,\Omega}^2 = |v|_{1,\Omega}^2. \end{aligned}$$

In the last step, we have used the fact $(P_k - P_{k-1})v$ is the orthogonal decomposition in the A-inner product.

2 *Minimal regularity* H^1: We relax the H^2-regularity upon using the decomposition

$$v = \sum_{k=0}^{J} (Q_k - Q_{k-1})v, \tag{28}$$

where $Q_k : \mathscr{V} \to \mathscr{V}_k$ is the L^2-projection onto $\mathscr{V}_k$. A simple proof of nearly optimal stability of (28) proceeds as follows. Invoking approximability and H^1-stability of the L^2-projection Q_k on quasi-uniform grids, we infer that

$$\|(Q_k - Q_{k-1})u\| = \|(I - Q_{k-1})Q_k u\| \lesssim h_k |Q_k u|_1 \lesssim h_k |u|_1.$$

Therefore

$$\sum_{k=0}^{J} h_k^2 \|(Q_k - Q_{k-1})u\|^2 \lesssim J|u|_1^2 \lesssim |\log h||u|_1^2.$$

The factor $|\log h|$ in the estimate can be removed by a more careful analysis based on the theory of Besov spaces and interpolation spaces. The following crucial inequality can be found, for example, in [91, 31, 64, 15, 65]:

$$\sum_{k=0}^{J} h_k^2 \|(Q_k - Q_{k-1})u\|^2 \lesssim |u|_1^2. \tag{29}$$

This completes the proof. □

We next state the stability of the micro decomposition. For a finite element space $\mathscr{V}$ with nodal basis $\{\phi_i\}_{i=1}^N$, let Q_{ϕ_i} be the L^2-projection to the one dimensional subspace spanned by ϕ_i. We have the following norm equivalence which says the nodal decomposition is stable in L^2. The proof is classical in the finite element analysis and thus omitted here.

Lemma 3.2 (Stability of Micro Decomposition). *For any $u \in \mathscr{V}$ over a quasi-uniform mesh $\mathscr{T}$, we have the norm equivalence*

$$\|u\|^2 \eqsim \sum_{i=1}^{N} \|Q_{\phi_i} u\|^2.$$

Theorem 3.1 (Stable Space Decomposition). *For any $v \in \mathscr{V}$, there exists a decomposition of v of the form*

$$v = \sum_{k=0}^{J} \sum_{i=1}^{N_k} v_{(k,i)}, \quad v_{(k,i)} \in \mathscr{V}_{(k,i)}, i = 1, \ldots, N_k, k = 0, \ldots, J,$$

such that

$$\sum_{k=0}^{J} \sum_{i=1}^{N_k} h_k^{-2} \|v_{(k,i)}\|^2 \lesssim |v|_1^2.$$

Consequently $\lambda_{\min}(BA) \gtrsim 1$ for the BPX preconditioner B defined in (23).

Proof. In light of Lemma 2.1, it suffices to combine Lemmas 3.1 and 3.2, and use (23). □

To estimate $\lambda_{\max}(BA)$, we first present a *strengthened Cauchy-Schwarz* (SCS) inequality for the macro decomposition.

Lemma 3.3 (Strengthened Cauchy-Schwarz Inequality (SCS)). *For any $u_i \in \mathscr{V}_i, v_j \in \mathscr{V}_j, j \geq i$, we have*

$$(u_i, v_j)_A \lesssim \gamma^{j-i} |u_i|_1 h_j^{-1} \|v_j\|_0,$$

where $\gamma < 1$ is a constant such that $h_i \eqsim \gamma^{2i}$.

Proof. Let us first prove the inequality on one element $\tau \in \mathscr{T}_i$. Using integration by parts, Cauchy-Schwarz inequality, trace theorem, and inverse inequality, we have

$$\begin{aligned}\int_\tau \nabla u_i \cdot \nabla v_j \, \mathrm{d}x = \int_{\partial\tau} \frac{\partial u_i}{\partial n} v_j \, ds &\lesssim \|\nabla u_i\|_{0,\partial\tau} \|v_j\|_{0,\partial\tau} \lesssim h_i^{-1/2} \|\nabla u_i\|_{0,\tau} h_j^{-1/2} \|v_j\|_{0,\tau} \\ &\lesssim (\frac{h_j}{h_i})^{1/2} |u_i|_{1,\tau} h_j^{-1} \|v_j\|_{0,\tau} \eqsim \gamma^{j-i} |u_i|_{1,\tau} h_j^{-1} \|v_j\|_{0,\tau}.\end{aligned}$$

Adding over $\tau \in \mathscr{T}_i$, and using Cauchy-Schwarz again, yields

$$(\nabla u_i, \nabla v_j) = \sum_{\tau\in\mathscr{T}_i} (\nabla u_i, \nabla v_j)_\tau \lesssim \gamma^{j-i} h_j^{-1} \sum_{\tau\in\mathscr{T}_i} |u_i|_{1,\tau} \|v_j\|_{0,\tau}$$
$$\lesssim \gamma^{j-i} h_j^{-1} \Big(\sum_{\tau\in\mathscr{T}_i} |u_i|_{1,\tau}^2 \Big)^{1/2} \Big(\sum_{\tau\in\mathscr{T}_i} \|v_j\|_{0,\tau}^2 \Big)^{1/2} = \gamma^{j-i} |u_i|_1 h_j^{-1} \|v_j\|_0,$$

which is the asserted estimate. □

Before we prove the main consequence of SCS, we need an elementary estimate.

Lemma 3.4 (Auxiliary Estimate). *Given $\gamma < 1$, we have*

$$\sum_{i,j=1}^n \gamma^{|j-i|} x_i y_j \le \frac{2}{1-\gamma} \Big(\sum_{i=1}^n x_i^2 \Big)^{1/2} \Big(\sum_{i=1}^n y_i^2 \Big)^{1/2} \quad \forall (x_i)_{i=1}^n, (y_i)_{i=1}^n \in \mathbb{R}^n.$$

Proof. Let $\Gamma \in \mathbb{R}^{n\times n}$ be the matrix $\Gamma = (\gamma^{|j-i|})_{i,j=1}^n$. The spectral radius $\rho(\Gamma)$ of Γ satisfies

$$\rho(\Gamma) \le \|\Gamma\|_1 = \max_{1\le j\le n} \sum_{i=1}^n \gamma^{|j-i|} \le \frac{2}{1-\gamma}.$$

Consequently, utilizing the Cauchy-Schwarz inequality yields

$$\sum_{i,j=1}^n \gamma^{|j-i|} x_i y_j = (\Gamma \mathbf{x}, \mathbf{y}) \le \rho(\Gamma) \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \quad \forall \mathbf{x} = (x_i)_{i=1}^n, \mathbf{y} = (y_i)_{i=1}^n \in \mathbb{R}^n,$$

which is the desired estimate. □

Theorem 3.2 (Largest Eigenvalue of BA). *For any $v \in \mathscr{V}$, we have*

$$(Av, v) \lesssim \inf_{\sum_{k=0}^J v_k = v} \sum_{k=0}^J h_k^{-2} \|v_k\|^2.$$

Consequently $\lambda_{\max}(BA) \lesssim 1$ for the BPX preconditioner B defined in (23).

Proof. For $v \in \mathscr{V}$, let $v = \sum_{k=0}^J v_k$, $v_k \in \mathscr{V}_k$, $k = 0, \ldots, J$, be an arbitrary decomposition. By the SCS inequality of Lemma 3.3, we have

$$(\nabla v, \nabla v) = 2 \sum_{k=0}^J \sum_{j=k+1}^J (\nabla v_k, \nabla v_j) + \sum_{k=0}^J (\nabla v_k, \nabla v_k) \lesssim \sum_{k=0}^J \sum_{j=k}^J \gamma^{j-k} |v_k|_1 h_j^{-1} \|v_j\|.$$

Combining Lemma 3.4 with the inverse estimate $|v_k|_1 \lesssim h_k^{-1} \|v_k\|$, we obtain

$$(\nabla v, \nabla v) \lesssim \Big(\sum_{k=0}^J |v_k|_1^2 \Big)^{1/2} \Big(\sum_{k=0}^J h_k^{-2} \|v_k\|^2 \Big)^{1/2} \lesssim \sum_{k=0}^J h_k^{-2} \|v_k\|^2.$$

which is the assertion. □

We finally prove the optimality of the BPX preconditioner.

Corollary 3.1 (Optimality of BPX Preconditioner). *For the preconditioner B defined in (23), we have*

$$\kappa(BA) \lesssim 1$$

Proof. Simply combine Theorems 3.1 and 3.2. □

3.4 Uniform convergence of V-cycle multigrid

In this section, we prove the uniform convergence of V-cycle multigrid, namely SSC applied to the decomposition (22) with exact subspace solvers.

Lemma 3.5 (Nodal Decomposition). *Let $\mathscr{T}$ be a quasi-uniform triangulation with N nodal basis ϕ_i. For the nodal decomposition*

$$v = \sum_{i=1}^{N} v_i, \quad v_i = v(x_i)\phi_i,$$

we have

$$\sum_{i=1}^{N} \Big| P_i \sum_{j>i}^{N} v_j \Big|_1^2 \lesssim h^{-2}\|v\|^2.$$

Proof. For every $1 \le i \le N$, we define the index set $L_i := \{j \in \mathbb{N} : i < j \le N, \operatorname{supp}\phi_j \cap \operatorname{supp}\phi_i \neq \emptyset\}$ and $\Omega_i = \cup_{j\in L_i} \operatorname{supp}\phi_j$. Since $\mathscr{T}$ is shape-regular, the numbers of integers in each L_i is uniformly bounded. So we have

$$\sum_{i=1}^{N} \Big| P_i \sum_{j>i}^{N} v_j \Big|_{1,\Omega}^2 = \sum_{i=1}^{N} \Big| P_i \sum_{j\in L_i} v_j \Big|_{1,\Omega}^2 \lesssim \sum_{i=1}^{N} \sum_{j\in L_i} |v_j|_{1,\Omega_i}^2 \lesssim \sum_{i=1}^{N} |v_i|_{1,\Omega_i}^2 \lesssim \sum_{i=1}^{N} h_i^{-2}\|v_i\|_{0,\Omega_i}^2,$$

where we have used an inverse inequality in the last step. Since $\mathscr{T}$ is quasi-uniform, and the nodal basis decomposition is stable in the L^2 inner product (Lemma 3.2), i.e. $\sum_{i=1}^{N} \|v_i\|_{0,\Omega_i}^2 \approx \|v\|_{0,\Omega}^2$, we deduce

$$\sum_{i=1}^{N} \Big| P_i \sum_{j>i}^{N} v_j \Big|_{1,\Omega}^2 \lesssim h^{-2}\|v\|_{0,\Omega}^2,$$

which is the desired estimate. □.

Lemma 3.6 (H^1 vs L^2 Stability). *The following inequality holds for all $v \in \mathscr{V}$*

$$\sum_{k=0}^{J} |(P_k - Q_k)v|_{1,\Omega}^2 \lesssim \sum_{k=0}^{J} h_k^{-2}\|(Q_k - Q_{k-1})v\|^2. \tag{30}$$

Proof. We first use the definition of P_k, together with $(I-Q_k)v = \sum_{j=k+1}^{J}(Q_j - Q_{j-1})v$, to write

$$\sum_{k=0}^{J}|(P_k-Q_k)v|_1^2 = \sum_{k=0}^{J}((P_k-Q_k)v,(I-Q_k)v)_A$$
$$= \sum_{k=0}^{J}\sum_{j=k+1}^{J}((P_k-Q_k)v,(Q_j-Q_{j-1})v)_A.$$

Applying now Lemma 3.3 yields

$$\sum_{k=0}^{J}|(P_k-Q_k)v|_1^2 \lesssim \left(\sum_{k=1}^{J}|(P_k-Q_k)v|_1^2\right)^{1/2}\left(\sum_{k=0}^{J}h_k^{-2}\|(Q_k-Q_{k-1})v\|^2\right)^{1/2}.$$

The desired result then follows. □

Theorem 3.3 (Optimality of V-cycle Multigrid). *The V-cycle multigrid method, using SSC applied to the decomposition (22) with exact subspace solvers $R_i = A_i^{-1}$, converges uniformly.*

Proof. We use the telescopic multilevel decomposition

$$v = \sum_{k=0}^{J} v_k, \quad v_k = (Q_k - Q_{k-1})v,$$

along with the nodal decomposition

$$v_k = \sum_{i=1}^{N_k} v_{(k,i)}, \quad v_{(k,i)} = v_k(x_i)\phi_{(k,i)},$$

for each level k. By the X-Z identity of Theorem 2.2, it suffices to prove the inequality

$$\sum_{k=0}^{J}\sum_{i=1}^{N_k}|P_{(k,i)}\sum_{(l,j)>(k,i)}v_{(l,j)}|_1^2 \lesssim |v|_1^2, \tag{31}$$

where the inner sum is understood in lexicographical order. We first simplify the left hand side of (31) upon writing

$$\sum_{(l,j)>(k,i)}v_{(l,j)} = \sum_{j>i}^{N_k}v_{(k,j)} + \sum_{l>k}v_l = \sum_{j>i}^{N_k}v_{(k,j)} + (v-Q_kv).$$

We apply Lemma 3.5 and the stable decomposition (29) to get

$$\sum_{k=0}^{J}\sum_{i=1}^{N_k}|P_{(k,i)}\sum_{j>i}v_{(k,j)}|_1^2 \lesssim \sum_{k=0}^{J}h_k^{-2}\|v_k\|^2 \lesssim |v|_1^2.$$

We now estimate the remaining terms $|P_0(v-Q_0)v|^2$ and $\sum_{k=1}^J |P_{(k,i)}(v-Q_kv)|^2_{1,\Omega}$. For any function $u \in \mathscr{V}$,

$$\sum_{i=1}^{N_k} |P_{(k,i)}u|_1^2 = \sum_{i=1}^{N_k} |P_{(k,i)}P_k u|^2_{1,\Omega_{(k,i)}} \leq \sum_{i=1}^{N_k} |P_k u|^2_{1,\Omega_{(k,i)}} \lesssim |P_k u|_1^2.$$

Thus, by (30) and (29), we get

$$\begin{aligned} &|P_0(v-Q_0)v|^2 + \sum_{k=1}^J |P_k(v-Q_kv)|_1^2 \\ &\qquad \lesssim \sum_{k=0}^J |(P_k - Q_k)v|_1^2 \lesssim \sum_{k=0}^J h_k^{-2} \|Q_k - Q_{k-1})v\|^2 \lesssim |v|_1^2. \end{aligned}$$

This completes the proof. □

The proof of Theorem 3.3 hinges on Theorem 2.2 (X-Z identity), which in turn requires exact solvers $R_i = A_i^{-1}$ and makes $P_i = A_i^{-1}Q_iA$ the key operator to appear in (11). If the smoothers R_i are not exact, namely $R_i \neq A_i^{-1}$, then the key operator becomes $T_i = R_iQ_iA$ and Theorem 2.2 must be replaced by Theorem 2.3. We refer to [30] for details.

3.5 Systems with strongly discontinuous coefficients

Elliptic problems with strongly discontinuous coefficients arise often in practical applications and are notoriously difficult to solve for iterative methods such as multigrid and domain decomposition. We are interested in the performance of these methods with respect to jumps. Consider the following model problem

$$\begin{cases} -\nabla\cdot(\omega\nabla u) = f & \text{in } \Omega, \\ u = g_D & \text{on } \Gamma_D, \\ -\omega\frac{\partial u}{\partial n} = g_N & \text{on } \Gamma_N \end{cases} \tag{32}$$

where $\Omega \in \mathbb{R}^d (d = 1,\ 2 \text{ or } 3)$ is a polygonal or polyhedral domain with Dirichlet boundary Γ_D and Neumann boundary Γ_N. We assume that the coefficient function $\omega = \omega(x)$ is positive and piecewise constant with respect to given subdomains Ω_m $(m = 1, \cdots, M)$ with $\overline{\Omega} = \cup_{m=1}^M \overline{\Omega}_m$, i.e., $\omega|_{\Omega_m} = \omega_m$ and

$$\mathscr{J}(\omega) \equiv \frac{\omega_{\max}}{\omega_{\min}} \gg 1.$$

These subdomains Ω_m are matched by the initial grid $\mathscr{T}_0$.

The question is how to make multigrid and domain decomposition methods converge (nearly) uniformly, not only with respect to the mesh size, but also with respect to the jump $\mathscr{J}(\omega)$. There has been a lot of interest in the development of iterative

methods with robust convergence rates with respect to the size of both jumps and mesh; see [17, 25, 77, 85, 86, 95] and the references cited therein. Domain decomposition (DD) methods have been developed for this purpose with special coarse spaces [95]. We refer to the monograph [82] and the survey [24] for a summary on DD methods. However, in general, the convergence rates of multigrid and domain decomposition methods are known to deteriorate with respect to $\mathscr{J}(\omega)$, especially in three dimensions.

The BPX and overlapping domain decomposition preconditioners are proven to be robust for some special cases: interface has no cross points [20, 66]; every subdomain touches part of the Dirichlet boundary [93]; and quasi-monotone coefficients [33, 34]. If the number of levels is fixed, multigrid converges uniformly with the convergence rate $\rho_k \leq 1-\delta^k$ where $\delta \in (0,1)$ is a constant and k is the number of levels. In general, the worst convergence rate is $1-Ch$ and, for BPX preconditioned system, $\sup_\omega \kappa(BA) \geq Ch^{-1}$ (see [66, 90]).

An interesting open problem is how to make multigrid method work uniformly with respect to jumps without introducing "expensive" coarse spaces. Recently, Xu and Zhu [93] proved that BPX and multigrid V-cycle lead to a nearly uniform convergent preconditioned conjugate gradient method (see [97] for a similar result on DD preconditioners). We now report this result.

Theorem 3.4 (Nearly Optimal PCG). *For BPX and multigrid V-cycle preconditioners (without using any special coarse spaces), PCG converges uniformly with respect to jumps in the sense that there exist c_0, c_1 and m_0 so that*

$$\|u-u_k\|_A \leq 2(c_0/h-1)^{m_0}(1-c_1/|\log h|)^{k-m_0}\|u-u_0\|_A \quad (k \geq m_0), \tag{33}$$

where m_0 is a fixed number depending only on the distribution of the coefficients.

This result is motivated by [41, 84] where PCG with diagonal scaling or overlapping DD is considered, and the following convergence result is proved by using pure algebraic methods:

$$\|u-u_k\|_A \leq C(h, \mathscr{J}(\omega))(1-ch)^{k-m_0}\|u-u_0\|_A.$$

Unfortunately, this estimate deteriorates severely with respect to mesh size. The improved estimate (33) implies that after m_0 steps, the convergent rate of the PCG is nearly uniform with respect to the mesh size and uniform with respect to jumps. The first m_0 steps are necessary for PCG to deal with small eigenvalues created by the jumps. To account for the effect of a finite cluster of eigenvalues in the convergence rate of PCG, the following estimate from [45] will be instrumental. Suppose that we can split the spectrum $\sigma(BA)$ of BA into two sets $\sigma_0(BA)$ and $\sigma_1(BA)$, where σ_0 consists of all "bad" eigenvalues and the remaining eigenvalues in σ_1 are bounded above and below.

Theorem 3.5 (CG for Clusters of Eigenvalues). *If $\sigma(BA) = \sigma_0(BA) \cup \sigma_1(BA)$ is such that $\sigma_0(BA)$ contains m eigenvalues and $\lambda \in [a,b]$ for each $\lambda \in \sigma_1(BA)$, then*

$$\|u-u_k\|_A \le 2(\kappa(BA)-1)^m \left(\frac{\sqrt{b/a}-1}{\sqrt{b/a}+1}\right)^{k-m} \|u-u_0\|_A.$$

Proof of Theorem 3.4. We introduce the weighted L^2 and H^1 inner products and corresponding norms

$$(u,v)_{0,\omega} = \int_\Omega uv\omega\,\mathrm{d}x = \sum_{m=1}^M \omega_m (u,v)_{\Omega_m},\ \|u\|_{0,\omega} = (u,u)_\omega^{1/2},$$
$$(u,v)_{1,\omega} = \int_\Omega \nabla u\cdot\nabla v\omega\,\mathrm{d}x = \sum_{m=1}^M \omega_m (u,v)_{1,\Omega_m},\ \|u\|_{1,\omega} = (u,u)_{1,\omega}^{1/2}.$$

The SPD operator A and corresponding inner product of finite element discretization of (32) is $(Au,v) = (u,v)_{1,\omega}$. Let $\mathscr{V}_h$ be the linear finite element space based on a shape regular triangulation $\mathscr{T}_h$. The weighted L^2-projection to $\mathscr{V}_h$ with respect to $(\cdot,\cdot)_{0,\omega}$ will be denoted by Q_h^ω.

We now introduce the following auxiliary subspace:

$$\widetilde{\mathscr{V}_h} = \left\{ v\in\mathscr{V}_h : \int_{\Omega_m} v\,\mathrm{d}x = 0,\ |\partial\Omega_m\cap\Gamma_D| = 0 \right\}.$$

Note that this subspace satisfies $\dim(\widetilde{\mathscr{V}_h}) = n - m_0$ where $m_0 < M$ is a fixed number, and more importantly,

$$\|v\|_{0,\omega} \lesssim |v|_{1,\omega} \quad \text{for all } v\in\widetilde{\mathscr{V}_h}.$$

As a consequence, we obtain the approximation and stability of the weighted L^2-projection Q_h^ω (see [20, 93, 97]),

$$\|(I-Q_h^\omega)v\|_{0,\omega} \lesssim h|\log h|^{\frac{1}{2}}|v|_{1,\omega},\quad |Q_h^\omega v|_{1,\omega} \lesssim |\log h|^{\frac{1}{2}}|v|_{1,\omega},\quad \text{for all } v\in\widetilde{\mathscr{V}_h}.$$

Using the arguments in Lemma 3.1-step 2, we can prove that the decomposition using weighted L^2 projection is almost stable, i.e.,

$$\sum_{k=0}^J h_k^{-2}\|(Q_k^\omega - Q_{k-1}^\omega)u\|^2 \lesssim |\log h|^2|u|_{1,\omega}^2. \tag{34}$$

Repeating the argument of Theorem 3.1, we obtain the estimate $\lambda_{\min}(BA)\gtrsim|\log h|^{-2}$.

On the other hand, the strengthened Cauchy Schwarz inequality (SCS) of Lemma 3.3 is valid for weighted inner products because its proof can be carried out element-wise when ω is piecewise constant. Consequently Theorem 3.2 holds for weighted L^2-norm and implies $\lambda_{\max}(BA)\lesssim 1$. We thus infer that the condition number of BA restricted to $\widetilde{\mathscr{V}_h}$ is nearly uniformly bounded, namely $\kappa(BA)\lesssim|\log h|^2$.

To estimate the convergent rate of PCG in the space $\mathscr{V}_h$, we introduce the mth effective condition number by $\kappa_{m+1}(A) = \lambda_{\max}(A)/\lambda_{m+1}(A)$, where $\lambda_{m+1}(A)$ is

the $(m+1)$th minimal eigenvalue of A. By the Courant "minmax" principle (see e.g., [40])

$$\lambda_{m+1}(A) = \max_{S,\dim(S)=m} \min_{0\neq v\in S^\perp} \frac{(Av,v)_{0,\omega}}{(v,v)_{0,\omega}}.$$

In particular, the fact $\dim(\widetilde{\mathscr{V}_h}) = n - m_0$, together with the nearly stable decomposition (34), implies that $\lambda_{m_0+1}(BA) \geq |\log h|^{-2}$.

The asserted estimate finally follows from Theorem 3.5 □.

Results such as Theorem 3.5 provide convincing evidence of a general rule of thumb: an iterative method, whenever possible, should be used together with certain preconditioned Krylov space (such as conjugate gradient) method.

4 Multilevel methods on graded grids

Adaptive methods are now widely used in scientific and engineering computation to optimize the relation between accuracy and computational labor (degrees of freedom). Let $\mathscr{V}_0 \subseteq \mathscr{V}_1 \subseteq \cdots \subseteq \mathscr{V}_J = \mathscr{V}$ be nested finite element spaces obtained by local mesh refinement. A standard multilevel method contains a smoothing step on the spaces $\mathscr{V}_j, j = 0,\ldots,J$. For graded grids obtained by adaptive procedure, it is possible that $\mathscr{V}_j$ results from $\mathscr{V}_{j-1}$ by just adding few, say one, basis function. Thus smoothing on both $\mathscr{V}_j$ and $\mathscr{V}_{j-1}$ leads to a lot of redundancy. If we let N be the number of unknowns in the finest space $\mathscr{V}$, then the complexity of smoothing can be as bad as $\mathscr{O}(N^2)$ [62]. To achieve optimal complexity $\mathscr{O}(N)$, the smoothing in each space $\mathscr{V}_j$ must be restricted to the new unknowns and their neighbors. Such methods are referred to as *adaptive multilevel methods* or *local multilevel methods*.

Of all possible refinement strategies, we are interested in *bisection*, the most popular and effective procedure for refinement in any dimension [6, 9, 56, 59, 68, 69, 70, 71, 72, 74, 80, 83]. We refer to [31] for the optimality of BPX preconditioner for regular refinement (one triangle is divided into four similar triangles) in 2-D and [1] for similar results in 3-D (one tetrahedron is divided into eight tetrahedrons).

We still consider the finite element approximation of Poisson equation (15); see Section §3.1 for the problem setting. The additional difficulty is that the mesh is no longer quasi-uniform. We present a decomposition of bisection grids and transfer results from quasi-uniform grids to bisection grids. As an example, we present a stable decomposition of finite element spaces and SCS inequality. The optimality of BPX preconditioner and uniform convergence of multigrid can then be established upon applying the general theory of Section §3; we refer to [27].

4.1 Bisection methods

In this section, we introduce bisection methods for simplicial grids and present a novel decomposition of conforming triangulations obtained by bisection methods.

Given a simplex τ, we assign one of its edges as the *refinement edge* of τ. Starting from an initial triangulation $\mathscr{T}_0$, a bisection method consists of the following rules:

R1. assign refinement edges for each element $\tau \in \mathscr{T}_0$;
R2. divide a simplex with a refinement edge into two simplexes;
R3. assign refinement edges to the two children of a bisected simplex.

We now give a mathematical description. Let τ be a simplex that bisects into simplexes τ_1 and τ_2. R2 can be described by a mapping $b_\tau : \{\tau\} \to \{\tau_1, \tau_2\}$. If we denote a simplex τ with a refinement edge e by a pair (τ, e), then R2 and R3 can be described by a mapping $\{(\tau,e)\} \to \{(\tau_1,e_1),(\tau_2,e_2)\}$. The pair (τ,e) is called a *labeled simplex* and the set $(\mathscr{T},L) := \{(\tau,e) : \tau \in \mathscr{T}\}$ is called a *labeled triangulation*. Then R1 can be described by a mapping $\mathscr{T}_0 \to (\mathscr{T}_0,L)$ and called *initial labeling*. The first rule is an essential ingredient of bisection methods. Once the initial labeling is done, the subsequent grids inherit labels according to R2-R3 such that the bisection process can proceed. We refer to [63, Section 4] for details.

For a labeled triangulation $(\mathscr{T},L)$, and a bisection

$$b_\tau : \{(\tau,e)\} \to \{(\tau_1,e_1),(\tau_2,e_2)\}$$

for $\tau \in \mathscr{T}$, we define a formal addition

$$\mathscr{T} + b_\tau := (\mathscr{T},L)\backslash\{(\tau,e)\} \cup \{(\tau_1,e_1),(\tau_2,e_2)\}.$$

For a sequence of bisections $\mathscr{B} = (b_{\tau_1}, b_{\tau_2}, \cdots, b_{\tau_N})$, we define

$$\mathscr{T} + \mathscr{B} := ((\mathscr{T} + b_{\tau_1}) + b_{\tau_2}) + \cdots + b_{\tau_N},$$

whenever the addition is well defined (i.e. τ_i should exists in the previous labeled triangulation). These additions are a convenient mathematical description of bisection on triangulations.

Given a labeled initial grid $\mathscr{T}_0$ of Ω and a bisection method, we define

$$\mathbb{F}(\mathscr{T}_0) = \{\mathscr{T} : \text{ there exists a bisection sequence } \mathscr{B} \text{ such that } \mathscr{T} = \mathscr{T}_0 + \mathscr{B}\},$$
$$\mathbb{T}(\mathscr{T}_0) = \{\mathscr{T} \in \mathbb{F}(\mathscr{T}_0) : \mathscr{T} \text{ is conforming}\}.$$

Therefore $\mathbb{F}(\mathscr{T}_0)$ contains all triangulations obtained from $\mathscr{T}_0$ using the bisection method, and is unique once the rules R1-3 have been set. But a triangulation $\mathscr{T} \in \mathbb{F}(\mathscr{T}_0)$ could be non-conforming and thus we define $\mathbb{T}(\mathscr{T}_0)$ as a subset of $\mathbb{F}(\mathscr{T}_0)$ containing only conforming triangulations.

We also define the sequence of uniformly refined meshes $\{\overline{\mathscr{T}}_k\}_{k=0}^{\infty}$ by:

$$\overline{\mathscr{T}}_0 = \mathscr{T}_0, \text{ and } \overline{\mathscr{T}}_k = \overline{\mathscr{T}}_{k-1} + \{b_\tau : \tau \in \overline{\mathscr{T}}_{k-1}\}, \text{ for } k \geq 1.$$

This means that $\overline{\mathscr{T}}_k$ is obtained by bisecting all elements in $\overline{\mathscr{T}}_{k-1}$ only once. Note that $\overline{\mathscr{T}}_k \in \mathbb{F}(\mathscr{T}_0)$ but not necessarily in the set $\mathbb{T}(\mathscr{T}_0)$.

We consider bisection methods which satisfy the following two assumptions:

(B1) Shape Regularity: $\mathbb{F}(\mathscr{T}_0)$ is shape regular.

(B2) Conformity of Uniform Refinement: $\overline{\mathscr{T}}_k \in \mathbb{T}(\mathscr{T}_0)$, i.e., $\overline{\mathscr{T}}_k$ is conforming for all $k \geq 0$.

All existing bisection methods share the same rule R2 described now. Given a simplex τ with refinement edge e, the two children of τ are defined by bisecting e and connecting the midpoint of e to the other vertices of τ. More precisely, let $\{x_1, x_2, \cdots, x_{d+1}\}$ be vertices of τ and let $e = \overline{x_1 x_2}$ be the refinement edge. Let x_m denote the midpoint of e. The children of τ are two simplexes τ_1 with vertices $\{x_1, x_m, x_3, \cdots, x_{d+1}\}$ and τ_2 with $\{x_2, x_m, x_3, \cdots, x_{d+1}\}$; we refer to [63, Section 4] for a through discussion of the notion of vertex type order and type. There is another class of refinement method, called regular refinement, which divide one simplex into 2^d children; see [8, 58].

All existing bisection methods differ in R1 and R3. For the so-called *longest edge bisection* [68, 70, 71, 72, 69], the refinement edge of a simplex is always assigned as the longest edge of this simplex. It is also used in R1 to assign the longest edge for each element in the initial triangulation. This method is simple, but (B1) is only proved for two dimensional triangulations [72] and (B2) only holds for special cases.

Regarding R3, the *newest vertex bisection* method for two dimensional triangulations assigns the edge opposite to the newest vertex of each child as their refinement edge. Sewell [76] showed that all the descendants of a triangle in $\mathscr{T}_0$ fall into four similarity classes and hence (B1) holds. Note that (B2) may not hold for an arbitrary rule R1, namely the refinement edge for elements in the initial triangulation cannot be selected freely. Mitchell [60] came up with a rule R1 for which (B2) holds. He proved the existence of such initial labeling scheme (so-called *compatible initial labeling*), and Biedl, Bose, Demaine, and Lubiw [11] gave an optimal $\mathscr{O}(N)$ algorithm to find a compatible initial labeling for a triangulation with N elements. In summary, in two dimensions, newest vertex bisection with compatible initial labeling is a bisection method which satisfies (B1) and (B2).

There are several bisection methods proposed in three and higher dimensions which generalize the newest vertex bisection in two dimensions [9, 56, 67, 6, 59, 80]. We shall not give detailed description of these bisection methods since the description of rules R1 and R3 is very technical for three and higher dimensions; we refer to [63, Section 4]. In these methods, (B1) is relatively easy to prove by showing all descendants of a simplex in $\mathscr{T}_0$ fall into similarity classes. As in the two dimensional case, (B2) requires special initial labeling, i.e., R1. We refer to Kossaczký [56] for the discussion of such rule in three dimensions and Stevenson [80] for the generalization to d-dimensions. However the algorithms proposed in [56, 80] to enforce such initial labeling consist of modifying the initial triangulation by further refinement of each element, which deteriorates the shape regularity. Although (B2) imposes a severe restriction on the initial labeling, we emphasize that it is also used to prove the optimal complexity of adaptive finite element methods [23, 63].

4.2 Compatible bisections

The set of vertices of the triangulation $\mathscr{T}$ will be denoted by $\mathscr{N}(\mathscr{T})$ and the set of all edges will be denoted by $\mathscr{E}(\mathscr{T})$. For a vertex $x \in \mathscr{N}(\mathscr{T})$ or an edge $e \in \mathscr{E}(\mathscr{T})$, we define the *first ring* of x or e to be

$$\mathscr{R}_x = \{\tau \in \mathscr{T} \,|\, x \in \tau\}, \quad \mathscr{R}_e = \{\tau \in \mathscr{T} \,|\, e \subset \tau\},$$

and the local patch of x or e as $\omega_x = \cup_{\tau \in \mathscr{R}_x} \tau$, and $\omega_e = \cup_{\tau \in \mathscr{R}_e} \tau$. Note that ω_x and ω_e are subsets of Ω, while $\mathscr{R}_x$ and $\mathscr{R}_e$ are subsets of $\mathscr{T}$ which can be thought of as triangulations of ω_x and ω_e, respectively. The cardinality of a set S will be denoted by $\#S$.

Given a labeled triangulation $(\mathscr{T}, L)$, an edge $e \in \mathscr{E}(\mathscr{T})$ is called a *compatible edge* if e is the refinement edge of τ for all $\tau \in \mathscr{R}_e$. For a compatible edge, the ring $\mathscr{R}_e$ is called a *compatible ring*, and the patch ω_e is called a *compatible patch*. Let x be the midpoint of e and $\mathscr{R}_x$ be the ring of x in $\mathscr{T} + \{b_\tau : \tau \in \mathscr{R}_e\}$. A *compatible bisection* is a mapping $b_e : \mathscr{R}_e \to \mathscr{R}_x$. We then define the addition

$$\mathscr{T} + b_e := \mathscr{T} + \{b_\tau : \tau \in \mathscr{R}_e\} = \mathscr{T} \backslash \mathscr{R}_e \cup \mathscr{R}_x.$$

For a compatible bisection sequence $\mathscr{B}$, the addition $\mathscr{T} + \mathscr{B}$ is defined as before.

Note that if $\mathscr{T}$ is conforming, then $\mathscr{T} + b_e$ is conforming for a compatible bisection b_e, whence compatible bisections preserve the conformity of triangulations. Hence, compatible bisection is a fundamental concept both in theory and practice.

In two dimensions, a compatible bisection b_e has only two possible configurations; see Fig. 1. One is bisecting an interior compatible edge, in which case the patch ω_e is a quadrilateral. Another case is bisecting a boundary edge, which is always compatible, and ω_e is a triangle. In three dimensions, the configuration of compatible bisections depends on the initial labeling; see Fig. 2 for a simple case.

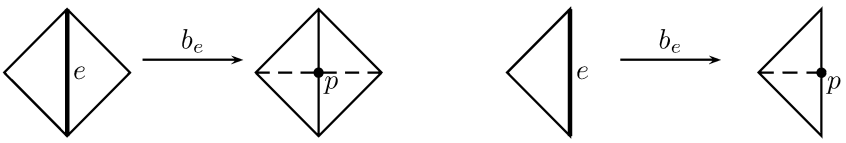
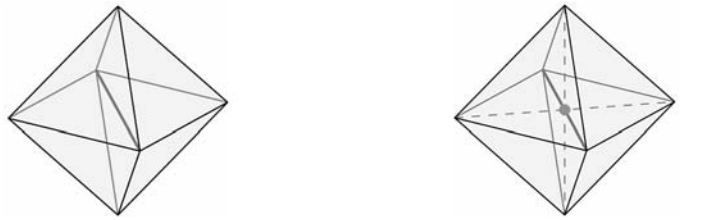


Fig. 1 Two compatible bisections for $d = 2$. Left: interior edge; right: boundary edge. The edge with boldface is the compatible refinement edge, and the dash-line represents the bisection

The bisection of paired triangles was first introduced by Mitchell [60, 61]. The idea was generalized by Kossaczký [56] to three dimensions, and Maubach [59] and Stevenson [80] to any dimension. In the aforementioned references, efficient recursive completion procedures of bisection methods are introduced based on compatible bisections. We use them to characterize the conforming mesh obtained by bisection methods.



(a) A compatible patch ω_e (b) After a compatible bisection ω_p

Fig. 2 A compatible bisection for $d = 3$: the edge e (in bold) is the refinement edge all elements in the patch ω_e. Connecting e to the other vertices bisects each element of the compatible ring $\mathscr{R}_e$ and keeps the mesh conforming without spreading refinement outside ω_e. This is an atomic operation

4.3 Decomposition of bisection grids

We now present a decomposition of meshes in $\mathbb{T}(\mathscr{T}_0)$ using compatible bisections. This is due to Chen, Nochetto, and Xu [27] and will be instrumental later.

Theorem 4.1 (Decomposition of Bisection Grids). *Let $\mathscr{T}_0$ be a conforming triangulation. Suppose the bisection method satisfies assumptions (B2), i.e., for all $k \geq 0$ all uniform refinements $\overline{\mathscr{T}}_k$ of $\mathscr{T}_0$ are conforming. Then for any $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$, there exists a compatible bisection sequence $\mathscr{B} = (b_1, b_2, \cdots, b_N)$ with $N = \#\mathscr{N}(\mathscr{T}) - \#\mathscr{N}(\mathscr{T}_0)$ such that*

$$\mathscr{T} = \mathscr{T}_0 + \mathscr{B}. \tag{35}$$

We use the example in Figure 3 to illustrate the decomposition of a bisection grid. In Figure 3 (a), we display the initial triangulation $\mathscr{T}_0$ which uses the longest edge as the refinement edge for each triangle. We display the fine grid $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$ in Figure 3 (f). In Figure 3 (b)-(e), we give several intermediate triangulations during the refinement process: each triangulation is obtained by performing several compatible bisections on the previous one. Each compatible patch is indicated by a gray region and the new vertices introduced by bisections are marked by black dots. In these figures, we denoted by $\mathscr{T}_i := \mathscr{T}_0 + (b_1, b_2, \cdots, b_i)$ for $1 \leq i \leq 19$.

To prove Theorem 4.1, we introduce the *generation* of elements and vertices. The generation of each element in the initial grid $\mathscr{T}_0$ is defined to be 0, and the generation of a child is 1 plus that of the father. The generation of an element $\tau \in \mathscr{T} \in \mathbb{F}(\mathscr{T}_0)$ is denoted by g_τ and coincides with the number of bisections needed to create τ from $\mathscr{T}_0$. Consequently, the uniformly refined mesh $\overline{\mathscr{T}}_k$ can be characterized as the triangulation in $\mathbb{F}(\mathscr{T}_0)$ with all elements of $\overline{\mathscr{T}}_k$ of the same generation k. Vice versa, an element τ with generation k can only exist in $\overline{\mathscr{T}}_k$.

Let $\mathbb{N}(\mathscr{T}_0) = \cup\{\mathscr{N}(\mathscr{T}) : \mathscr{T} \in \mathbb{F}(\mathscr{T}_0)\}$ denote the set of all possible vertices. For any vertex $p \in \mathbb{N}(\mathscr{T}_0)$, the generation of p is defined as the minimal integer k such that $p \in \mathscr{N}(\overline{\mathscr{T}}_k)$ and is denoted by g_p. For convenience of notation, we regard g as

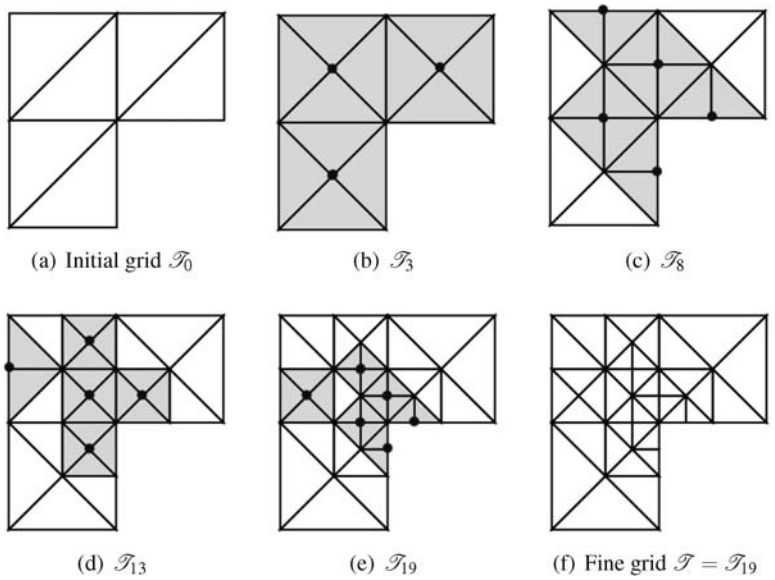


Fig. 3 Decomposition of a bisection grid for $d = 2$: Each frame displays a mesh $\mathscr{T}_{i+k} = \mathscr{T}_i + \{b_{i+1}, \cdots, b_{i+k}\}$ obtained from $\mathscr{T}_i$ by a sequence of compatible bisections $\{b_j\}_{j=i+1}^{i+k}$ using the longest edge. The order of bisections is irrelevant within each frame, but matters otherwise

either a piecewise linear function on $\mathscr{T}$ defined as $g(p) = g_p$ for $p \in \mathscr{N}(\mathscr{T})$ or a piecewise constant defined as $g(\tau) = g_\tau$ for $\tau \in \mathscr{T}$.

The following properties about the generation of elements or vertices for uniformly refined mesh $\overline{\mathscr{T}}_k$ are a consequence of the definition above:

$$\tau \in \overline{\mathscr{T}}_k \text{ if and only if } g_\tau = k; \tag{36}$$

$$p \in \mathscr{N}(\overline{\mathscr{T}}_k) \text{ if and only if } g_p \le k; \tag{37}$$

$$\text{for } \tau \in \overline{\mathscr{T}}_k, \ \max_{q \in \mathscr{N}(\tau)} g_q = k = g_\tau. \tag{38}$$

Lemma 4.1. *Let $\mathscr{T}_0$ be a conforming triangulation. Let the bisection method satisfy assumption (B2). For any $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$, let $p \in \mathscr{N}(\mathscr{T})$ be a vertex with maximal generation in the sense that $g_p = \max_{q \in \mathscr{N}(\mathscr{T})} g_q$. Then*

$$g_\tau = g_p \quad \textit{for all } \tau \in \mathscr{R}_p \tag{39}$$

and

$$\mathscr{R}_p = \overline{\mathscr{R}}_{k,p}, \tag{40}$$

where $k = g_p$ and $\overline{\mathscr{R}}_{k,p}$ is the first ring of p in the uniformly refined mesh $\overline{\mathscr{T}}_k$. Equivalently, all elements in $\mathscr{R}_p$ have the same generation g_p.

Proof. We prove (39) by showing $g_p \le g_\tau$ and $g_\tau \le g_p$. Since $\mathscr{T}$ is conforming, p is a vertex of each element $\tau \in \mathscr{R}_p$. This implies that $p \in \mathscr{N}(\overline{\mathscr{T}}_{g_\tau})$ and thus $g_\tau \ge g_p$ by (37). On the other hand, from (38), we have

$$g_\tau = \max_{q\in\mathscr{N}(\tau)} g_q \le \max_{q\in\mathscr{N}(\mathscr{T})} g_q = g_p, \quad \text{for all } \tau \in \mathscr{R}_p.$$

Now we prove (40). By (36), $\overline{\mathscr{R}}_{k,p}$ is made of all elements with generation k containing p. By (39), we conclude $\mathscr{R}_p \subseteq \overline{\mathscr{R}}_{k,p}$. On the other hand, p cannot belong to the domain of $\Omega\backslash\omega_p$, because of the topology of ω_p, whence $\overline{\mathscr{R}}_{k,p}\backslash\mathscr{R}_p = \varnothing$. This proves (40). □

Now we are in the position to prove Theorem 4.1.

Proof of Theorem 4.1. We prove the result by the induction on $N = \#\mathscr{N}(\mathscr{T}) - \#\mathscr{N}(\mathscr{T}_0)$. Nothing needs to be proved for $N = 0$. Assume that (35) holds for N.

Let $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$ with $\#\mathscr{N}(\mathscr{T}) - \#\mathscr{N}(\mathscr{T}_0) = N+1$. Let $p \in \mathscr{N}(\mathscr{T})$ be a vertex with maximal generation, i.e., $g_p = \max_{q\in\mathscr{N}(\mathscr{T})} g_q$. Then by Lemma 4.1, we know that $\mathscr{R}_p = \overline{\mathscr{R}}_{k,p}$ for $k = g_p$. Now by assumption (B2), $\overline{\mathscr{R}}_{k,p}$ is created by a compatible bisection, say

$$b_e : \overline{\mathscr{R}}_e \to \overline{\mathscr{R}}_{k,p},$$

with $e \in \mathscr{E}(\overline{\mathscr{T}_{k-1}})$. Since the compatible bisection giving rise to p is unique within $\mathbb{F}(\mathscr{T}_0)$, it must thus be b_e. This means that if we undo the bisection operation, then we still have a conforming mesh $\mathscr{T}'$, or equivalently $\mathscr{T} = \mathscr{T}' + b_e$. We can now apply the induction assumption to $\mathscr{T}' \in \mathbb{T}(\mathscr{T}_0)$ with $\#\mathscr{N}(\mathscr{T}') - \#\mathscr{N}(\mathscr{T}_0) = N$ to finish the proof. □

4.4 Generation of compatible bisections

For a compatible bisection $b_i \in \mathscr{B}$, we use the same subscript i to denote related quantities such as:

- e_i: the refinement edge;
- p_i: the midpoint of e_i;
- $\widetilde{\omega}_i = \omega_{p_i} \cup \omega_{p_{l_i}} \cup \omega_{p_{r_i}}$;
- $\mathscr{T}_i = \mathscr{T}_0 + \{b_1, \cdots, b_i\}$;
- ω_i: the patch of p_i i.e. ω_{p_i};
- p_{l_i}, p_{r_i}: two end points of e_i;
- h_i: the local mesh size of ω_i;
- $\mathscr{R}_i$: the first ring of p_i in $\mathscr{T}_i$.

We understand $h \in L^\infty(\Omega)$ as a piecewise constant mesh-size function, i.e., $h_\tau = \mathrm{diam}(\tau)$ in each simplex $\tau \in \mathscr{T}$.

Lemma 4.2. *If $b_i \in \mathscr{B}$ is a compatible bisection, then all elements of $\mathscr{R}_i$ have the same generation g_i.*

Proof. Let $p_i \in \mathscr{N}(\mathscr{T}_0)$ be the vertex associated with b_i. Let $\overline{\mathscr{T}_k}$ be the coarsest uniformly refined mesh containing p_i, so $k = g_{p_i}$. In view of assumption (B2), p_i

arises from uniform refinement of $\overline{\mathscr{T}}_{k-1}$. Since the bisection giving rise to p_i is unique within $\mathbb{F}(\mathscr{T}_0)$, we realize that all elements in $\mathscr{R}_{e_i}$ are bisected and have generation $k-1$ because they belong to $\overline{\mathscr{T}}_{k-1}$. This implies that all elements of $\mathscr{R}_{p_i}$ have generation k, as asserted. □

This lemma allows us to introduce the concept of generation of compatible bisections. For a compatible bisection $b_i : \mathscr{R}_{e_i} \to \mathscr{R}_{p_i}$, we define $g_i = g(\tau), \tau \in \mathscr{R}_{p_i}$. Throughout this paper we always assume $h(\tau) \eqsim 1$ for $\tau \in \mathscr{T}_0$. We have the following important relation between generation and mesh size

$$h_i \eqsim \gamma^{g_i}, \quad \text{with } \gamma = \left(\frac{1}{2}\right)^{1/d} \in (0,1). \tag{41}$$

Beside this relation, we give now two more important properties on the generation of compatible bisections. The first property says that different bisections with the same generation have weakly disjoint local patches.

Lemma 4.3. *Let $\mathscr{T}_N \in \mathbb{T}(\mathscr{T}_0)$ be $\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}$, where $\mathscr{B}$ is a compatible bisection sequence $\mathscr{B} = (b_1, \cdots, b_N)$. For any $i \neq j$ and $g_j = g_i$, we have*

$$\mathring{\omega}_i \cap \mathring{\omega}_j = \varnothing. \tag{42}$$

Proof. Since $g_i = g_j = g$, both bisection patches $\mathscr{R}_i$ and $\mathscr{R}_j$ belong to the uniformly refined mesh $\overline{\mathscr{T}}_q$. If (42) were not true, then there would exists $\tau \in \mathscr{R}_i \cap \mathscr{R}_j \subset \overline{\mathscr{T}}_q$ containing distinct refinement edges e_i and e_j because $i \neq j$. This contradicts rules R2 and R3 which assign a unique refinement edge to each element. □

A simple consequence of (42) is that, for all $u \in L^2(\Omega)$ and $k \geq 1$,

$$\sum_{g_i=k} \|u\|^2_{\omega_i} \leq \|u\|^2_{\Omega}, \tag{43}$$

$$\sum_{g_i=k} \|u\|^2_{\tilde{\omega}_i} \lesssim \|u\|^2_{\Omega}. \tag{44}$$

The second property is on the ordering of generations. For a given bisection sequence $\mathscr{B}$, we define $b_i < b_j$ if $i < j$, which means bisection b_i is performed before b_j. The generation sequence $(g_1, \cdots, g_N)$, however, is not necessary monotone increasing; there could exist $b_i < b_j$ but $g_i > g_j$. This happens for bisections driven by *a posteriori* error estimators in practice. Adaptive algorithms usually refine elements around a singularity region first, thereby creating many elements with large generations, and later they refine coarse elements away from the singularity. This mixture of generations is the main difficulty for the analysis of multilevel methods on adaptive grids. We now prove the following quasi-monotonicity property of generations restricted to a fixed bisection patch.

Lemma 4.4. *Let $\mathscr{T}_N \in \mathbb{T}(\mathscr{T}_0)$ be $\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}$, where $\mathscr{B}$ is a compatible bisection sequence $\mathscr{B} = (b_1, \cdots, b_N)$. For any $j > i$ and $\mathring{\tilde{\omega}}_j \cap \mathring{\tilde{\omega}}_i \neq \varnothing$, we have*

$$g_j \geq g_i - g_0, \tag{45}$$

where $g_0 > 0$ is an integer depending only the shape regularity of $\mathscr{T}_0$.

Proof. Since $\overset{\circ}{\widetilde{\omega}}_j \cap \overset{\circ}{\widetilde{\omega}}_i \neq \varnothing$, there must be elements $\tau_j \in \mathscr{R}_{p_j} \cup \mathscr{R}_{p_{l_j}} \cup \mathscr{R}_{p_{r_j}}$ and $\tau_i \in \mathscr{R}_{p_i} \cup \mathscr{R}_{p_{l_i}} \cup \mathscr{R}_{p_{r_i}}$ such that $\overset{\circ}{\tau}_j \cap \overset{\circ}{\tau}_i \neq \varnothing$. Since we consider triangulations in $\mathbb{T}(\mathscr{T}_0)$, the intersection $\tau_j \cap \tau_i$ is still a simplex. When b_j is performed, only τ_j exists in the current mesh. Thus $\tau_j = \tau_j \cap \tau_i \subseteq \tau_i$ and $g_{\tau_j} \geq g_{\tau_i}$.

Shape regularity implies the existence of a constant g_0 only depending on $\mathscr{T}_0$ such that

$$g_j + g_0/2 \geq g_{\tau_j} \geq g_{\tau_i} \geq g_i - g_0/2,$$

and (45) follows. □

4.5 Node-oriented coarsening algorithm

A key practical issue is to find a decomposition of a bisection grid. We present a node-oriented coarsening algorithm recently developed by Chen and Zhang [29].

A crucial observation is that the inverse of a compatible bisection can be thought as a coarsening process. It is restricted to a compatible star and thus no conformity issue arises; See Figure 1. For a triangulation $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$, a vertex p is called a *good-for-coarsening vertex*, or a *good vertex* in short, if there exist a compatible bisection b_e such that p is the middle point of e. The set of all good vertices in the grid $\mathscr{T}$ will be denoted by $G(\mathscr{T})$. By the decomposition of bisection grids (Theorem 4.1), the existence of good vertices is evident. Moreover, for bisection grids in 2-D, we have the following characterization of good vertices due to Chen and Zhang [29].

Theorem 4.2 (Coarsening). *Let $\mathscr{T}_0$ be a conforming triangulation. Suppose the bisection method satisfies assumptions (B2), i.e., for all $k \geq 0$ all uniform refinements $\overline{\mathscr{T}}_k$ of $\mathscr{T}_0$ are conforming. Then for any $\mathscr{T} \in \mathbb{T}(\mathscr{T}_0)$ and $\mathscr{T} \neq \mathscr{T}_0$, the set of good vertices $G(\mathscr{T})$ is not empty. Furthermore $x \in G(\mathscr{T})$ if and only if*

1. *it is not a vertex of the initial grid $\mathscr{T}_0$;*
2. *it is the newest vertex of all elements in the ring of $\mathscr{R}_p$.*
3. *$\#\mathscr{R}_p = 4$ for an interior vertex x or $\#\mathscr{R}_p = 2$ for a boundary vertex p.*

Remark 4.1. The assumption that $\mathscr{T}_0$ is compatible labeled could be further relaxed by using the longest edge of each triangle as its refinement edge for the initial triangulation $\mathscr{T}_0$; see Kossaczký [56].

The coarsening algorithm is simply read as the following:

```
ALGORITHM COARSEN (𝒯)
  Find all good nodes G(𝒯) of 𝒯.
  For each good node p ∈ G(𝒯)
```

```
    Replace the star R_p by b_e^{-1}(R_p).
END
```

Chen and Zhang [29] prove that one can finally obtain the initial grid back by applying the coarsening algorithm `coarsen` repeatedly. It is possible that `coarsen(T)` applied to the current grid $\mathscr{T}$ gives a coarse grid which is not in the adaptive history. Indeed our coarsening algorithm may remove vertices added in several different stages of the adaptive procedure.

For details on the implementation of this coarsening algorithm and the application to multilevel preconditioners and multigrid methods, we refer to [29] and [26].

4.6 Space decomposition on bisection grids

We give a space decomposition for Lagrange finite element spaces on bisection grids. Given a conforming triangulation $\mathscr{T}$ of the domain $\Omega \subset \mathbb{R}^d$ and an integer $m \geq 1$, the mth order finite element space on $\mathscr{T}$ is defined as follows:

$$\mathscr{V}(\mathscr{P}_m, \mathscr{T}) := \{v \in H^1(\Omega) : v|_\tau \in \mathscr{P}_m(\tau) \text{ for all } \tau \in \mathscr{T}\}.$$

We restrict ourselves to bisection grids in $\mathbb{T}(\mathscr{T}_0)$ satisfying (B1) and (B2). Therefore by Theorem 4.1, for any $\mathscr{T}_N \in \mathbb{T}(\mathscr{T}_0)$, there exists a compatible bisection sequence $\mathscr{B} = (b_1, \cdots, b_N)$ such that

$$\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}.$$

We give a decomposition of the finite element space $\mathscr{V} := \mathscr{V}(\mathscr{P}_m, \mathscr{T}_N)$ using this decomposition of $\mathscr{T}_N$. If $\mathscr{T}_i$ is the triangulation $\mathscr{T}_0 + (b_1, \cdots, b_i)$, let $\phi_{i,p} \in \mathscr{V}(\mathscr{P}_1, \mathscr{T}_i)$ denote the linear nodal basis at a vertex $p \in \mathscr{N}(\mathscr{T}_i)$. Motivated by the stable three-point wavelet constructed by Stevenson [78], we define the sub-spaces

$$\mathscr{V}_0 = \mathscr{V}(\mathscr{P}_1, \mathscr{T}_0), \text{ and } \mathscr{V}_i = \operatorname{span}\{\phi_{i,p_i}, \phi_{i,p_{l_i}}, \phi_{i,p_{r_i}}\}. \tag{46}$$

Since the basis functions of $\mathscr{V}_i, i = 0, \ldots, N$, are piecewise linear polynomials on $\mathscr{T}_N$, we know $\mathscr{V}_i \subseteq \mathscr{V}$. Let $\{\phi_p, p \in \Lambda\}$ be a basis of $\mathscr{V}(\mathscr{P}_m, \mathscr{T}_N)$ such that $v = \sum_{p\in\Lambda} v(p)\phi_p$ for all $v \in \mathscr{V}(\mathscr{P}_m, \mathscr{T}_N)$, where Λ is the index set of basis. For example, for quadratic element spaces, Λ consists of vertices and middle points of edges. We define $\mathscr{V}_p = \operatorname{span}\{\phi_p\}$ and end up with the following space decomposition:

$$\mathscr{V} = \sum_{p\in\Lambda} \mathscr{V}_p + \sum_{i=0}^{N} \mathscr{V}_i. \tag{47}$$

Since $\dim \mathscr{V}_i = 3$, we have a three-point local smoother and the total computational cost for subspace correction methods based on (47) is CN. This is optimal and the constant in front of N is relatively small. In addition, the three-point local smoother simplifies the implementation of multilevel methods especially in dimensions higher

than 3. For example, we only need to maintain an ordered vertex array with two parent vertices and do not need tree structure to maintain a hierarchical structure of meshes. The following result is due to Chen, Nochetto, and Xu [27].

Theorem 4.3 (Space Decomposition over Graded Meshes). *For any $v \in \mathscr{V}$, there exist $v_p, p \in \Lambda, v_i \in \mathscr{V}_i, i = 0, \cdots, N$ such that $v = \sum_{p\in\Lambda} v_p + \sum_{i=0}^{N} v_i$ and*

$$\sum_{p\in\Lambda} h_p^{-2}\|v_p\|^2 + \sum_{i=0}^{N} h_i^{-2}\|v_i\|^2 \lesssim \|v\|_A^2. \tag{48}$$

The idea of the proof is to use Scott-Zhang quasi-interpolation operator [75]

$$\mathscr{I}_{\mathscr{T}} : H^1(\Omega) \mapsto \mathscr{V}(\mathscr{P}_1, \mathscr{T})$$

for a conforming triangulation $\mathscr{T}$; see also Oswald [65]. For any $p \in \mathscr{N}(\mathscr{T})$ and p is an interior point, we choose a $\tau_p \subset \mathscr{R}_p$. Let $\{\lambda_{\tau_p,i}, i = 1, \cdots, d+1\}$ be the barycentric coordinates of τ which span $\mathscr{P}_1(\tau_p)$. We construct the L^2-dual basis $\Theta(\tau_p) = \{\theta_{\tau_p,i} : i = 1, \cdots, d+1\}$ of $\{\lambda_{\tau_p,i} : i = 1, \cdots, d+1\}$. Suppose $\theta_p \in \Theta(\tau_p)$ is the dual basis such that $\int_{\tau_p} \theta_p v \,\mathrm{d}x = v(p)$, for all $v \in \mathscr{P}_1(\tau_p)$. We then define

$$\mathscr{I}_{\mathscr{T}} v = \sum_{p\in\mathscr{N}(\mathscr{T})} \Big(\int_{\tau_p} \theta_p v \,\mathrm{d}x\Big)\phi_p.$$

For boundary vertex p, we simply define $\mathscr{I}_{\mathscr{T}} v(p) = 0$ to reflect the vanishing boundary condition of v. By definition, $\mathscr{I}_{\mathscr{T}}$ preserves piecewise linear functions and satisfies the following estimate and stability [75, 65]

$$|\mathscr{I}_{\mathscr{T}} v|_1 + \|h^{-1}(v - \mathscr{I}_{\mathscr{T}} v)\| \lesssim |v|_1, \tag{49}$$

$$h_i^{d-2}|\mathscr{I}_{\mathscr{T}} v(p_i)|^2 \lesssim h_i^{-2}\|v\|_{\tau_{p_i}}, \tag{50}$$

where h_i is the size of τ_{p_i}.

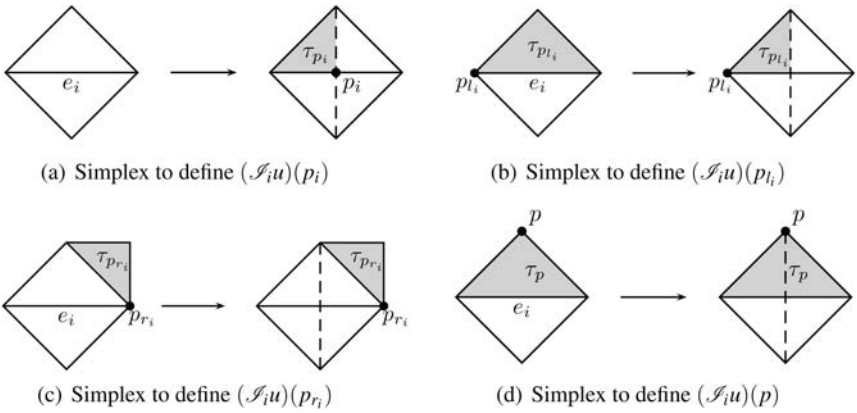
Given $v \in \mathscr{V}(\mathscr{P}_m, \mathscr{T})$, we define $u = \mathscr{I}_{\mathscr{T}} v$ and a decomposition $v = u + (v - u)$, where $\mathscr{I}_{\mathscr{T}} : \mathscr{V}(\mathscr{P}_m, \mathscr{T}) \to \mathscr{V}(\mathscr{P}_1, \mathscr{T})$. We first give a multilevel decomposition of u using quasi-interpolation. For a vertex p, we denote by τ_p the simplex used to define the nodal value at p. The following construction of a sequence of quasi-interpolations will update τ_p carefully.

Let $\mathscr{I}_0$ be a quasi-interpolation operator defined $\mathscr{V}(\mathscr{P}_1, \mathscr{T}) \to \mathscr{V}_0$. Suppose $\mathscr{I}_{i-1}$ is defined on $\mathscr{V}(\mathscr{P}_1, \mathscr{T}_{i-1})$. After the compatible bisection b_i, we define the nodal values at the new added vertex p_i using a simplex introduced by the bisection, i.e. $\tau_{p_i} \subset \omega_i$. For other vertices p, let $\tau_p \in \mathscr{T}_{i-1}$ be the simplex used to define $(\mathscr{I}_{i-1}u)(p)$, we define $(\mathscr{I}_i u)(p)$ according to the following two cases:

1. if $\tau_p \subset \omega_p(\mathscr{T}_i)$ we keep the nodal value, i.e., $(\mathscr{I}_i u)(p) = (\mathscr{I}_{i-1}u)(p)$;
2. otherwise we choose a new $\tau_p \subset \omega_p(\mathscr{T}_i) \cap \omega_p(\mathscr{T}_{i-1})$ to define $(\mathscr{I}_i u)(p)$.

In either case, we ensure that the simplex $\tau_p \subset \omega_p(\mathscr{T}_i)$.

An important property of the bisection is that b_i only changes the local patches of two end points of the refinement edge e_i going from $\mathscr{T}_{i-1}$ to $\mathscr{T}_i$. The construction in the second case is thus well defined. By construction $(\mathscr{I}_i - \mathscr{I}_{i-1})u(p) = 0$ for $p \in \mathscr{N}(\mathscr{T}_i), p \neq p_i, p_{l_i}$ or p_{r_i}, which implies $(\mathscr{I}_i - \mathscr{I}_{i-1})u \in \mathscr{V}_i$. Furthermore a close look reveals that if $(\mathscr{I}_i - \mathscr{I}_{i-1})u(p) \neq 0$, then the elements τ_p used to define $\mathscr{I}_i(p)$ or $\mathscr{I}_{i-1}(p)$ are inside the patch ω_i; see Figure 4.



(a) Simplex to define $(\mathscr{I}_i u)(p_i)$ (b) Simplex to define $(\mathscr{I}_i u)(p_{l_i})$

(c) Simplex to define $(\mathscr{I}_i u)(p_{r_i})$ (d) Simplex to define $(\mathscr{I}_i u)(p)$

Fig. 4 Update of nodal values $\mathscr{I}_i u$ to yield $\mathscr{I}_{i-1}u$: the element τ chosen to perform the averaging that gives $(\mathscr{I}_i u)(p)$ must belong to $\omega_p(\mathscr{T}_i)$. This implies $(\mathscr{I}_i - \mathscr{I}_{i-1})u(p) \neq 0$ possibly for $p = p_i, p_{l_i}, p_{r_i}$ and $= 0$ otherwise

In this way, we obtain a sequence of quasi-interpolation operators

$$\mathscr{I}_i : \mathscr{V}(\mathscr{P}_1, \mathscr{T}_N) \to \mathscr{V}(\mathscr{P}_1, \mathscr{T}_i), i = 0 : N.$$

We define $v_i = (\mathscr{I}_i - \mathscr{I}_{i-1})u \in \mathbb{V}_i$ for $i = 1 : N$. In general $\mathscr{I}_N u \neq u$ since the simplex used to define nodal values of $\mathscr{I}_N u$ may not be in the finest mesh $\mathscr{T}_N$ but in $\mathscr{T}_{N-1}$. Nevertheless, the difference $v - \mathscr{I}_N u$ is of high frequency in the finest mesh.

Let $v - \mathscr{I}_N u = \sum_{p \in \Lambda} v_p$ be the basis decomposition. We then obtain a decomposition

$$v = \sum_{p \in \Lambda} v_p + \sum_{i=0}^{N} v_i, \quad v_i \in \mathbb{V}_i, \tag{51}$$

where for convenience we define $\mathscr{I}_{-1} := 0$.

To prove that the decomposition (51) is stable we first study $\sum_{p \in \Lambda} v_p$. Let τ_p be the simplex used to define $\mathscr{I}_N u(p)$ for $p \in \mathscr{N}(\mathscr{T}_N)$. By construction, although τ_p may not be a simplex in the triangulation $\mathscr{T}_N$, it is still in the patch $\omega_p(\mathscr{T}_N)$. Then by (49)

$$\sum_{p \in \Lambda} h_p^{-2} \|v_p\|^2 \lesssim \|h^{-1}(v - Q_N v)\|^2 \lesssim |v|_1^2. \tag{52}$$

We next prove that the decomposition $\mathscr{I}_N u = \sum_{i=0}^{N}(\mathscr{I}_i - \mathscr{I}_{i-1})u$ is stable. For this purpose, we need the auxiliary decomposition on the uniform refinement. We choose minimal L such that $\mathscr{V} \subseteq \overline{\mathscr{V}}_L$. By Lemma 3.1, we have a stable decomposition $u = \sum_{k=0}^{L} \bar{v}_k$, with $\bar{v}_k = (\overline{Q}_k - \overline{Q}_{k-1})u, k = 0, \cdots, L$.

We apply the slicing operator $\mathscr{I}_i - \mathscr{I}_{i-1}$ to this decomposition. When $k \leq g_i - 1$, $\bar{v}_k$ is piecewise linear in ω_{e_i}, $(\mathscr{I}_i - \mathscr{I}_{i-1})\bar{v}_k = 0$ since $\mathscr{I}_i$ preserves piecewise linear functions. So the slicing operator detects frequencies higher than or equal to the generation of bisection, namely

$$v_i = (\mathscr{I}_i - \mathscr{I}_{i-1}) \sum_{l=g_i}^{L} \bar{v}_l. \tag{53}$$

By construction of v_i and the stability of quasi-interpolation, we conclude

$$\|v_i\|_{\tilde{\omega}_i}^2 \lesssim h_i^{2+d}\left[v_i(p_i)^2 + v_i(p_{l_i})^2 + v_i(p_{r_i})^2\right] \lesssim \Big\| \sum_{l=g_i}^{L} \bar{v}_l \Big\|_{\omega_i}^2.$$

In the last step, the domain is changed to ω_i since the simplexes used to define nonzero values of $v_i(p_i), v_i(p_{l_i r})$ or $v_i(p_{l_i r})$ are inside ω_i.

Note that for different bisections with the same generation, their local patches are weakly disjoint (Lemma 4.3): for any $i \neq j$ and $g_j = g_i$, we have

$$\overset{\circ}{\omega}_i \cap \overset{\circ}{\omega}_j = \varnothing. \tag{54}$$

Consequently

$$\sum_{g_i=k} \|v_i\|^2 = \sum_{g_i=k} \|v_i\|_{\tilde{\omega}_i}^2 \lesssim \sum_{g_i=k} \Big\| \sum_{l=g_i}^{L} \bar{v}_l \Big\|_{\omega_i}^2 \lesssim \Big\| \sum_{l=g_i}^{L} \bar{v}_l \Big\|_{\Omega}^2 = \sum_{l=k}^{L} \|\bar{v}_l\|^2.$$

In the last step, we use the fact $\bar{v}_k$ are L^2-orthogonal decomposition.

The following elementary result will be useful and can be found in [32].

Lemma 4.5 (Discrete Hardy Inequality). *If the sequences* $\{a_k\}_{k=0}^{L}, \{b_k\}_{k=0}^{L}$ *satisfy*

$$b_k \leq \sum_{l=k}^{L} a_l, \quad \textit{for all } k \geq 0$$

and are non-negative, then for any $s \in (0,1)$*, we have*

$$\sum_{k=0}^{L} s^{-k} b_k \leq \frac{1}{1-s} \sum_{k=0}^{L} s^{-k} a_k.$$

Proof. Since

$$\sum_{k=0}^{L} s^{-k} b_k \leq \sum_{k=0}^{L}\sum_{l=k}^{L} s^{-k} a_l = \sum_{l=0}^{L}\sum_{k=0}^{l} s^{-k} a_l = \sum_{l=0}^{L} s^{-l} a_l \sum_{k=0}^{l} s^{l-k},$$

and $s < 1$, the geometric series is bounded by $1/(1-s)$ and concludes the proof. □

Applying Lemma 4.5 to $a_k = \|\bar{v}_k\|^2$ and $b_k = \sum_{g_i=k} \|v_i\|^2$, we obtain

$$\sum_{k=0}^{L} \bar{h}_k^{-2} \sum_{g_i=k} \|v_i\|^2 \lesssim \sum_{k=0}^{L} \bar{h}_k^{-2} \|\bar{v}_k\|^2,$$

and thus from the stable decomposition corresponding to uniform refinement, we conclude

$$\sum_{i=0}^{N} h_i^{-2} \|v_i\|^2 = \sum_{k=0}^{L} \bar{h}_k^{-2} \sum_{g_i=k} \|v_i\|^2 \lesssim \sum_{k=0}^{L} \bar{h}_k^{-2} \|\bar{v}_k\|^2 \lesssim |\mathscr{I}_{\mathscr{T}} v|_1^2 \lesssim |v|_1^2. \tag{55}$$

4.7 Strengthened Cauchy-Schwarz inequality

In this section we establish the SCS inequality for the space decomposition $\sum_{i=0}^{N} \mathscr{V}_i$.

Theorem 4.4. *For any $u_i, v_i \in \mathscr{V}_i, i = 0, \cdots, N$, we have*

$$\Big| \sum_{i=0}^{N} \sum_{j=i+1}^{N} (u_i, v_j)_A \Big| \lesssim \left(\sum_{i=0}^{N} \|u_i\|_A^2 \right)^{1/2} \left(\sum_{i=0}^{N} h_i^{-2} \|v_i\|^2 \right)^{1/2}. \tag{56}$$

Proof. The proof consists of several careful summations using the concept of generation to relate with uniform refinements. The proof is divided into four steps.

1 For a fixed index $i \in [1, N]$, we denote by

$$n(i) = \{ j > i : \widetilde{\omega}_j \cap \widetilde{\omega}_i \neq \varnothing \} \quad \text{and} \quad w_k^i = \sum_{j \in n(i), g_j = k} v_j.$$

Shape regularity implies that $w_k^i \in \overline{\mathscr{V}}_{k+g_0}$ and $k = g_j \geq g_i - g_0$ (Lemma 4.4). For any $\tau \in \tilde{\omega}_i$, we apply the SCS inequality of Lemma 3.3 over τ to u_i and w_k^i and obtain

$$(u_i, w_k^i)_{A,\tau} \lesssim \gamma^{k+g_0-g_i} \|u_i\|_{A,\tau} \bar{h}_{k+g_0}^{-1} \|w_k^i\|_\tau \lesssim \gamma^{k-g_i} \|u_i\|_{A,\tau} \bar{h}_k^{-1} \|w_k^i\|_\tau.$$

Then

$$
\begin{aligned}
(u_i, w_k^i)_{A,\tilde{\omega}_i} &= \sum_{\tau\subset\tilde{\omega}_i} (u_i, w_k^i)_{A,\tau} \\
&\lesssim \gamma^{k-g_i} \sum_{\tau\subset\tilde{\omega}_i} \|u_i\|_{A,\tau} \bar{h}_k^{-1} \|w_k^i\|_\tau \\
&\lesssim \gamma^{k-g_i} \|u_i\|_{A,\tilde{\omega}_i} \bar{h}_k^{-1} \Big(\sum_{\tau\subset\tilde{\omega}_i} \|w_k^i\|_\tau^2 \Big)^{1/2}.
\end{aligned}
$$

Since v_j's with the same generation $g_j = k$ have supports with finite overlap, we infer that $\|w_k^i\|_\tau^2 \lesssim \sum_{j\in n(i), g_j=k} \|v_j\|_\tau^2 \le \sum_{g_j=k} \|v_j\|_\tau^2$ and

$$
(u_i, w_k^i)_{A,\tilde{\omega}_i} \lesssim \gamma^{k-g_i} \|u_i\|_{A,\tilde{\omega}_i} \bar{h}_k^{-1} \Big(\sum_{g_j=k} \|v_j\|_{0,\tilde{\omega}_i}^2 \Big)^{1/2}.
$$

2 We fix u_i and consider

$$
|(u_i, \sum_{j=i+1}^{N} v_j)_A| = |(u_i, \sum_{j\in n(i)} v_j)_{A,\tilde{\omega}_i}| = |(u_i, \sum_{k=(g_i-g_0)^+}^{L} \sum_{j\in n(i), g_j=k} v_j)_{A,\tilde{\omega}_i}|,
$$

because $w_k^j = 0$ for $k < g_i - g_0$ (Lemma 4.4). Since $k \ge 0$, this is equivalent to $k \ge (g_i - g_0)^+ := \max\{g_i - g_0, 0\}$, whence

$$
\begin{aligned}
|(u_i, \sum_{j=i+1}^{N} v_j)_A| &\lesssim \sum_{k=(g_i-g_0)^+}^{L} |(u_i, w_k^i)_{A,\tilde{\omega}_i}| \\
&\lesssim \sum_{k=(g_i-g_0)^+}^{L} \gamma^{k-g_i} \|u_i\|_{A,\tilde{\omega}_i} \bar{h}_k^{-1} \Big(\sum_{g_j=k} \|v_j\|_{0,\tilde{\omega}_i}^2 \Big)^{1/2}.
\end{aligned}
$$

3 We now sum over i but keeping the generation $g_i = l \ge 0$ fixed:

$$
\begin{aligned}
\sum_{g_i=l} |(u_i, \sum_{j=i+1}^{N} v_j)_A| &\lesssim \sum_{k=(l-g_0)^+}^{L} \gamma^{k-l} \left\{ \sum_{g_i=l} \left[\|u_i\|_{A,\tilde{\omega}_i} \Big(\bar{h}_k^{-2} \sum_{g_j=k} \|v_j\|_{\tilde{\omega}_i}^2 \Big)^{1/2} \right] \right\} \\
&\lesssim \sum_{k=(l-g_0)^+}^{L} \gamma^{k-l} \Big(\sum_{g_i=l} \|u_i\|_{A,\tilde{\omega}_i}^2 \Big)^{1/2} \Big(\bar{h}_k^{-2} \sum_{g_i=l} \sum_{g_j=k} \|v_j\|_{\tilde{\omega}_i}^2 \Big)^{1/2}.
\end{aligned}
$$

In view of the finite overlap of patches $\tilde{\omega}_i$ for generation $g_i = l$ (see (44)), we deduce

$$
\sum_{g_i=l} |(u_i, \sum_{j=i+1}^{N} v_j)_A| \lesssim \sum_{k=(l-g_0)^+}^{L} \gamma^{k-l} \Big(\sum_{g_i=l} \|u_i\|_{A,\tilde{\omega}_i}^2 \Big)^{1/2} \Big(\bar{h}_k^{-2} \sum_{g_j=k} \|v_j\|^2 \Big)^{1/2}.
$$

4. We finally sum over all generations $0 \le l \le L$ to get

$$\sum_{l=0}^{L}\sum_{g_i=l}|(u_i,\sum_{j=i+1}^{N}v_j)_A| \lesssim \sum_{l=0}^{L}\sum_{k=(l-g_0)^+}^{L}\gamma^{k-l}\left(\sum_{g_i=l}\|u_i\|^2_{A,\tilde{\omega}_i}\right)^{1/2}\left(\bar{h}_k^{-2}\sum_{g_j=k}\|v_j\|^2\right)^{1/2}$$
$$\lesssim \left(\sum_{l=0}^{L}\sum_{g_i=l}\|u_i\|^2_{A,\tilde{\omega}_i}\right)^{1/2}\left(\sum_{k=0}^{L}\bar{h}_k^{-2}\sum_{g_j=k}\|v_j\|^2\right)^{1/2}.$$

where we have applied Lemma 3.4. Therefore, since $\sum_{i=0}^{N}=\sum_{l=0}^{L}\sum_{g_i=l}$ and $\bar{h}_k=h_j$ for $k=g_j$, we end up with the desired estimate (56). □

4.8 BPX preconditioner and multigrid on graded bisection grids

Proceeding as in Section §3, with quasi-uniform grids created by uniform refinement, we can obtain the optimality of BPX preconditioner and optimal convergent rate of V-cycle multigrid. We state the results below and refer to [27] for proofs.

Theorem 4.5 (Optimality of BPX on Graded Bisection Grids). *For the BPX preconditioner based on the space decomposition (47)*

$$Bu=\sum_{p\in\Lambda}h_p^{2-d}(u,\phi_p)\phi_p+\sum_{i=1}^{N}h_i^{2-d}[(u,\phi_{p_i})\phi_{p_i}+(u,\phi_{p_{l_i}})\phi_{p_{l_i}}+(u,\phi_{p_{r_i}})\phi_{p_{r_i}}],$$

we have

$$\kappa(BA)\lesssim 1.$$

A V-cycle type multigrid method can be obtained by applying SSC to the space decomposition (47). A symmetric V-cycle loop is like

1. pre-smoothing (forward Gauss-Seidel) in the finest space $\mathscr{V}(\mathscr{P}_m,\mathscr{T}_N)$;
2. multilevel smoothing in *linear* finite element spaces $\mathscr{V}_i$ for $i=N$ to 1;
3. exact solver in the coarsest linear finite element spaces $\mathscr{V}_0$;
4. multilevel smoothing in *linear* finite element spaces $\mathscr{V}_i$ for $i=1$ to N;
5. post-smoothing (backward Gauss-Seidel) in the finest space $\mathscr{V}(\mathscr{P}_m,\mathscr{T}_N)$.

Theorem 4.6 (Uniform Convergence of V-cycle Multigrid on Graded Bisection Grids). *The above V-cycle multigrid, namely SSC based on the space decomposition (47), is uniformly convergent.*

5 Multilevel methods for H(curl) and H(div) systems

In this section, we design and analyze multigrid methods for solving finite element discretization of $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems

$$\mathrm{curl} \times \mathrm{curl} \times \boldsymbol{u} + \boldsymbol{u} = \boldsymbol{f}, \qquad \text{in } \Omega, \tag{57}$$

$$-\,\mathrm{grad}\,\mathrm{div}\,\boldsymbol{u} + \boldsymbol{u} = \boldsymbol{f}, \qquad \text{in } \Omega, \tag{58}$$

with homogeneous Neumann boundary condition. Here $\Omega \subset \mathbb{R}^3$ is a simply connected and bounded polyhedron. We study edge elements for (57) and face elements for (58) over shape regular tetrahedra triangulations $\mathscr{T}$ of Ω.

Standard multigrid methods developed for H^1 problem, i.e.,

$$-\Delta u + u = -\,\mathrm{div}\,\mathrm{grad}\,u + u = f$$

cannot be transferred to the $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems directly. The reason is that for vector fields, the operators $\mathrm{curl} \times \mathrm{curl}$ and $-\,\mathrm{grad}\,\mathrm{div}$ are only part of the Laplace operator because

$$-\boldsymbol{\Delta} := \mathrm{curl} \times \mathrm{curl} - \mathrm{grad}\,\mathrm{div}\,.$$

Therefore in the divergence free space, the operator $\mathrm{curl} \times \mathrm{curl} + \boldsymbol{I}$ behaves like $-\boldsymbol{\Delta} + \boldsymbol{I}$, while in the kernel space of the curl operator, the space of gradients, it is like $\boldsymbol{I}$. Similarly, the operator $-\,\mathrm{grad}\,\mathrm{div} + \boldsymbol{I}$ behaves like $-\Delta + I$ on gradients and I on curls. Efficient solvers should account for the different behavior of curl and div in their kernel and orthogonal complement. In particular, the smoother in the kernel space is critical. We note that for the grad operator, the kernel space is a one dimensional (constant) space, while for the curl and div operators, the kernel space is infinite dimensional. The decomposition of spaces used in multigrid methods should satisfy certain properties (see [57] and [98]). One approach is to perform a smoothing in the kernel space which can be expressed explicitly using properties of exact sequences between finite element spaces of $H^1, \boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems. This is used by Hiptmair to obtain the first results for multigrid of $\boldsymbol{H}(\mathrm{div})$ [47] and $\boldsymbol{H}(\mathrm{curl})$ [49] systems in three dimensions. See also Hiptmair and Toselli [51] for a unified and simplified treatment. Another important approach taken by Arnold, Falk and Winther in [3, 4] is to perform the smoothing on patches of vertices which contain a basis of the kernel space of curl and div operator. In [3, 4], the analysis hinges on the following two assumptions:

- Ω is a bounded and *convex* polyhedron in $\mathbb{R}^3$;
- $\mathscr{T}$ is a shape regular and quasi-uniform mesh of Ω.

The first assumption is used in duality arguments which require full regularity of the elliptic equations, whereas the second one is used to prove certain approximation properties. We regard both items as regularity assumptions, first on the solutions of the elliptic equation and second on the underlying mesh.

In practice, most problems are posed on non-convex domains and thus solutions exhibit singularities. Finite element approximations based on quasi-uniform grids cannot deliver optimal rates due to lack of regularity. Mesh refinements restore optimal convergence rates in terms of degree of freedoms, but make the above regularity assumptions inadequate for studying adaptive finite element methods for $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems.

We will design multilevel methods for these systems on graded grids obtained by bisection. In the analysis, we relax the regularity assumptions used in the previous work [47, 49, 3, 4] by using two new techniques developed recently in [52] and [27]. More precisely, we employ

- Discrete regular decompositions of finite element spaces [52] to relax the regularity assumption on the solution;
- Decomposition of bisection grids and corresponding space decompositions [27], already discussed in section §4, to relax the regularity assumption on the grids.

We should mention that a local multigrid method similar to ours for $\boldsymbol{H}(\mathrm{curl})$ system on adaptive grids has been independently developed by Hiptmair and Zheng [53]. We follow closely our recent work [28] to present a unified treatment for both $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems.

To focus on the two aforementioned issues, we consider the simplest scenario, that is we do not include Dirichlet type boundary conditions for (57) or (58) nor variable coefficients. We note that results in [4] hold uniformly for variable coefficients and results in this paper extend to this case as well.

5.1 Preliminaries

5.1.1 Sobolev spaces and finite element spaces

Let $\Omega \subset \mathbb{R}^3$ be a bounded domain which is homeomorphic to a ball. We define the following Sobolev spaces

$$\begin{aligned} H(\mathrm{grad};\Omega) &= \{v \in L^2(\Omega) : \mathrm{grad}\, v \in \boldsymbol{L}^2(\Omega)\} = H^1(\Omega), \\ \boldsymbol{H}(\mathrm{curl};\Omega) &= \{\boldsymbol{v} \in (L^2(\Omega))^3 : \mathrm{curl}\, \boldsymbol{v} \in (L^2(\Omega))^3\}, \\ \boldsymbol{H}(\mathrm{div};\Omega) &= \{\boldsymbol{v} \in (L^2(\Omega))^3 : \mathrm{div}\, \boldsymbol{v} \in (L^2(\Omega))^3\}. \end{aligned}$$

We use a generic notation $\boldsymbol{H}(\mathscr{D},\Omega)$ to refer to $H(\mathrm{grad};\Omega)$, $\boldsymbol{H}(\mathrm{curl};\Omega)$ or $\boldsymbol{H}(\mathrm{div};\Omega)$, where $\mathscr{D} = \mathrm{grad}, \mathrm{curl}$ or div represents differential operators according to the context. Since $\mathrm{curl}\, \boldsymbol{v}$ and $\mathrm{div}\, \boldsymbol{v}$ are special combinations of components of $\mathrm{grad}\, \boldsymbol{v}$, in general $\boldsymbol{H}^1(\Omega) \subset \boldsymbol{H}(\mathscr{D},\Omega)$.

Let $(\cdot,\cdot)$ denote the inner product for $L^2(\Omega)$ or $[L^2(\Omega)]^3$. As subspaces of $[L^2(\Omega)]^3$, $H(\mathrm{grad};\Omega)$, $\boldsymbol{H}(\mathrm{curl};\Omega)$, and $\boldsymbol{H}(\mathrm{div};\Omega)$ are endowed with $(\cdot,\cdot)$ as their default inner product. We assign new inner products using differential operator $\mathscr{D}$ to these spaces:

$$\begin{aligned} H(\mathrm{grad};\Omega): &\quad (u,v)_{A^g} := (u,v) + (\mathrm{grad}\, u, \mathrm{grad}\, v), \\ \boldsymbol{H}(\mathrm{curl};\Omega): &\quad (\boldsymbol{u},\boldsymbol{v})_{A^c} := (\boldsymbol{u},\boldsymbol{v}) + (\mathrm{curl}\, \boldsymbol{u}, \mathrm{curl}\, \boldsymbol{v}), \\ \boldsymbol{H}(\mathrm{div};\Omega): &\quad (\boldsymbol{u},\boldsymbol{v})_{A^d} := (\boldsymbol{u},\boldsymbol{v}) + (\mathrm{div}\, \boldsymbol{u}, \mathrm{div}\, \boldsymbol{v}). \end{aligned}$$

The corresponding norm are denoted by $\|\cdot\|_{A^g}$, $\|\cdot\|_{A^c}$ and $\|\cdot\|_{A^d}$, respectively.

These inner products introduce corresponding symmetric positive definite operators (with respect to the default $(\cdot,\cdot)$ inner product).

$$\begin{aligned} A^g &: H(\mathrm{grad};\Omega) \to H(\mathrm{grad};\Omega)^* \qquad (A^g u, v) := (u,v)_{A^g}, \\ A^c &: \boldsymbol{H}(\mathrm{curl};\Omega) \to H(\mathrm{curl};\Omega)^* \qquad (A^c u, v) := (u,v)_{A^c}, \\ A^d &: \boldsymbol{H}(\mathrm{div};\Omega) \to H(\mathrm{div};\Omega)^* \qquad (A^d u, v) := (u,v)_{A^d}. \end{aligned}$$

We focus on the $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems, namely,

$$A^c \boldsymbol{u} = \mathrm{curl} \times \mathrm{curl}\, u + u = \boldsymbol{f}, \tag{59}$$

$$A^d \boldsymbol{u} = -\,\mathrm{grad}\,\mathrm{div}\, u + u = \boldsymbol{f}, \tag{60}$$

with homogeneous Neumann boundary condition. We build on the study of the H^1 problem, $A^g u = f$, in previous sections.

Given a shape regular triangulation $\mathscr{T}$ of Ω and integer $k \geq 1$, we define the following finite element spaces:

$$\begin{aligned} \mathscr{V}(\mathrm{grad}, \mathscr{P}_k, \mathscr{T}) &:= \{v \in H(\mathrm{grad};\Omega) : v|_\tau \in \mathscr{P}_k(\tau),\ \forall \tau \in \mathscr{T}\}, \\ \mathscr{V}(\mathrm{curl}, \mathscr{P}_k^-, \mathscr{T}) &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};\Omega) : \boldsymbol{v}|_\tau \in \mathscr{P}_{k-1}^3(\tau) + \mathscr{P}_{k-1}^3(\tau) \times \boldsymbol{x},\ \forall \tau \in \mathscr{T}\}, \\ \mathscr{V}(\mathrm{curl}, \mathscr{P}_k, \mathscr{T}) &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};\Omega) : \boldsymbol{v}|_\tau \in \mathscr{P}_k^3(\tau),\ \forall \tau \in \mathscr{T}\}, \\ \mathscr{V}(\mathrm{div}, \mathscr{P}_k^-, \mathscr{T}) &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega) : \boldsymbol{v}|_\tau \in \mathscr{P}_{k-1}^3(\tau) + \mathscr{P}_{k-1}(\tau)\boldsymbol{x},\ \forall \tau \in \mathscr{T}\} \\ \mathscr{V}(\mathrm{div}, \mathscr{P}_k, \mathscr{T}) &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega) : \boldsymbol{v}|_\tau \in \mathscr{P}_k^3(\tau),\ \forall \tau \in \mathscr{T}\} \\ \mathscr{V}(L^2, \mathscr{P}_{k-1}, \mathscr{T}) &:= \{v \in L^2(\Omega) : v|_\tau \in \mathscr{P}_{k-1}(\tau),\ \forall \tau \in \mathscr{T}\}. \end{aligned}$$

As in [5], the notation $\mathscr{P}_k^-$ indicates that the polynomial space is a proper subspace of $\mathscr{P}_k$. When we do not refer to a specific finite element space, we use the generic notation $\mathscr{V}(\mathscr{D}, \mathscr{T})$. In particular, we simply denote by $\mathscr{V} = \mathscr{V}(\mathrm{grad}, \mathscr{P}_1, \mathscr{T})$ the continuous piecewise linear finite element space.

The degrees of freedom of these finite element spaces, and their unisolvency, are not easy to sketch here. We refer to [2, 5, 48, 50] for a unified presentation using differential forms.

Since $\mathscr{V}(\mathscr{D}, \mathscr{T}) \subset \boldsymbol{H}(\mathscr{D};\Omega)$, the operator equations (59) or (60) can be restricted to the finite element spaces $\mathscr{V}(\mathrm{curl}, \mathscr{T})$ or $\mathscr{V}(\mathrm{div}, \mathscr{T})$. Existence and uniqueness of the ensuing discrete problems follow from the Riesz representation theorem. Our task is to develop fast solvers for these linear algebraic systems over graded bisection grids as well as unstructured grids $\mathscr{T}$.

5.1.2 Exact sequences and commutative diagram

The following exact sequence, called de Rham differential complex, plays an important role in the error analysis of finite element approximations as well as the iteration methods for solving the algebraic systems:

$$\mathbb{R} \hookrightarrow H^1(\Omega) \xrightarrow{\text{grad}} \boldsymbol{H}(\text{curl};\Omega) \xrightarrow{\text{curl}} \boldsymbol{H}(\text{div};\Omega) \xrightarrow{\text{div}} L^2(\Omega). \tag{61}$$

For a differential operator $\mathscr{D}$, we denote by $\mathscr{D}^-$ the previous one in the exact sequence: if $\mathscr{D} = \text{curl}$, then $\mathscr{D}^- = \text{grad}$, and if $\mathscr{D} = \text{div}$, then $\mathscr{D}^- = \text{curl}$. The following crucial properties of (61) are valid:

$$\ker(\text{grad}) = \mathbb{R}, \quad \ker(\text{curl}) = \text{img}(\text{grad}), \quad \ker(\text{div}) = \text{img}(\text{curl}). \tag{62}$$

We now state two results, Theorem 5.1 for $\mathscr{D} = \text{curl}$ and Theorem 5.2 for $\mathscr{D} = \text{div}$, which make this precise. We refer to Girault-Raviart [39] for Theorem 5.1.

Theorem 5.1 (Irrotational Fields). *Let Ω be a bounded, simply connected Lipschitz domain in $\mathbb{R}^3$ and suppose $\boldsymbol{u} \in [L^2(\Omega)]^3$. Then $\text{curl}\,\boldsymbol{u} = 0$ in Ω if and only if there exists a scalar potential $\phi \in H^1(\Omega)$ such that $\boldsymbol{u} = \text{grad}\,\phi$ and*

$$\|\phi\|_1 \lesssim \|u\|. \tag{63}$$

To verify that $\ker(\text{div}) = \text{img}(\text{curl})$, we first present a result in $\mathbb{R}^3$.

Lemma 5.1. *Let $N(\text{div};\mathbb{R}^3) = \{\boldsymbol{v} \in H(\text{div};\mathbb{R}^3) : \text{div}\, v = 0\}$ be the kernel of operator* div. *Then for any $\boldsymbol{u} \in N(\text{div};\mathbb{R}^3)$ there exists $\boldsymbol{\phi} \in [H^1_{\text{loc}}(\mathbb{R}^3)]^3$ such that*

$$\text{curl}\,\boldsymbol{\phi} = \boldsymbol{u}, \quad \text{div}\,\boldsymbol{\phi} = 0, \quad \|\boldsymbol{\phi}\|_{1,\text{loc},\mathbb{R}^3} \lesssim \|\boldsymbol{u}\|_{0,\mathbb{R}^3}. \tag{64}$$

Proof. In terms of Fourier transform, the conditions $u = \text{curl}\,\phi$ and $\text{div}\,\phi = 0$ become

$$\hat{u} = 2\pi i\, \xi \times \hat{\phi} = 2\pi i(\xi_2\hat{\phi}_3 - \xi_3\hat{\phi}_2, \xi_3\hat{\phi}_1 - \xi_1\hat{\phi}_3, \xi_1\hat{\phi}_2 - \xi_2\hat{\phi}_1),$$
$$\xi \cdot \hat{\phi} = \sum_{j=1}^{3} \xi_j \hat{\phi}_j = 0,$$

respectively. We observe that the first relation implies

$$\xi \cdot \hat{u} = \sum_{j=1}^{3} \xi_j \hat{u}_j = 0,$$

or equivalently $\text{div}\, u = 0$. Computing $\hat{u} \times \xi$ and using the first two relations gives $\hat{\phi}$ uniquely as follows:

$$\hat{\phi} = \frac{1}{2\pi i|\xi|^2}\hat{u} \times \xi = \frac{1}{2\pi i|\xi|^2}(\xi_3\hat{u}_2 - \xi_2\hat{u}_3, \xi_1\hat{u}_3 - \xi_3\hat{u}_1, \xi_2\hat{u}_1 - \xi_1\hat{u}_2).$$

The desirable ϕ is the inverse Fourier transform of $\hat{\phi}$. In addition, we have

$$|\xi_j\phi_i| \le \sum_{i=1}^{3} |\hat{u}_i|.$$

Parseval's identity shows that $\boldsymbol{\phi} \in \boldsymbol{H}^1_{\text{loc}}(\mathbb{R}^3)$. □

Theorem 5.2 (Solenoidal Fields). *Let Ω be a simply connected bounded domain. For any function $\boldsymbol{u} \in \boldsymbol{H}(\mathrm{div};\Omega)$ such that $\mathrm{div}\,\boldsymbol{u} = 0$, there exists a vector field $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ such that $\boldsymbol{u} = \mathrm{curl}\,\phi$ and $\mathrm{div}\,\phi = 0$ in Ω and*

$$\|\boldsymbol{\phi}\|_{H^1(\Omega)} \lesssim \|\boldsymbol{u}\|_{L^2(\Omega)}.$$

Proof. We first construct an extension of $\boldsymbol{u}$ to $N(\mathrm{div};\mathbb{R}^3)$. Let $\mathscr{O}$ be a smooth domain containing Ω. We let $p \in H^1(\mathscr{O}\backslash\Omega)/\mathbb{R}$ satisfy

$$-\Delta p = 0 \text{ in } \mathscr{O}\backslash\Omega,$$
$$\frac{\partial p}{\partial n} = \boldsymbol{u}\cdot\boldsymbol{n} \text{ on } \partial\Omega, \quad \frac{\partial p}{\partial n} = 0 \text{ on } \partial\mathscr{O}.$$

This solution exists since $\langle \boldsymbol{u}\cdot\boldsymbol{n}, 1\rangle_{\partial\Omega} = \int_\Omega \mathrm{div}\,\boldsymbol{u}\,\mathrm{dx} = 0$. We define $\tilde{\boldsymbol{u}} \in \boldsymbol{L}^2(\mathbb{R}^3)$ by

$$\tilde{u} = \begin{cases} \boldsymbol{u} & \text{in } \Omega, \\ \mathrm{grad}\,p & \text{in } \mathscr{O}\backslash\Omega, \\ 0 & \text{in } \mathbb{R}^3\backslash\mathscr{O}. \end{cases}$$

Since $\mathrm{div}\,\tilde{\boldsymbol{u}} = 0$ in Ω and $\mathscr{O}\backslash\Omega$ and the normal component of $\tilde{\boldsymbol{u}}$ is continuous across the common boundary $\partial\Omega$, we conclude $\tilde{u} \in \boldsymbol{H}(\mathrm{div};\mathbb{R}^3)$ and $\mathrm{div}\,\tilde{u} = 0$.

We then apply Lemma 5.1 to get a $\boldsymbol{\phi}$ satisfying (64). Since

$$\|\tilde{u}\|_{L^2(\mathbb{R}^3)} = \|\tilde{u}\|_{H(\mathrm{div};\mathbb{R}^3)} \lesssim \|u\|_{H(\mathrm{div};\Omega)} = \|u\|_{L^2(\Omega)},$$

restricting $\boldsymbol{\phi}$ to Ω leads to a desirable $\boldsymbol{\phi}$. □

Exact Sequences (ES). The discrete counterpart of the de Rham differential complex (61) is also valid for the finite element spaces $\mathscr{V}(\mathscr{D},\mathscr{T})$:

$$\mathbb{R} \hookrightarrow \mathscr{V}(\mathrm{grad},\mathscr{T}) \xrightarrow{\mathrm{grad}} \mathscr{V}(\mathrm{curl},\mathscr{T}) \xrightarrow{\mathrm{curl}} \mathscr{V}(\mathrm{div},\mathscr{T}) \xrightarrow{\mathrm{div}} \mathscr{V}(L^2,\mathscr{T}). \tag{65}$$

The starting finite element space $\mathscr{V}(\mathrm{grad},\mathscr{T})$ and the ending space $\mathscr{V}(L^2,\mathscr{T})$ are continuous and discontinuous complete polynomial spaces, respectively. For the two spaces in the middle, each one has two types. Therefore we have 4 exact sequences in $\mathbb{R}^3$ and these are all possible exact sequences in $\mathbb{R}^3$ [5]. For completeness we list these exact sequences below:

$$\mathbb{R} \hookrightarrow \mathscr{V}(\mathrm{grad},\mathscr{P}_k,\mathscr{T}) \xrightarrow{\mathrm{grad}} \mathscr{V}(\mathrm{curl},\mathscr{P}_{k-1},\mathscr{T}) \xrightarrow{\mathrm{curl}} \mathscr{V}(\mathrm{div},\mathscr{P}_{k-2},\mathscr{T}) \xrightarrow{\mathrm{div}} \mathscr{V}(L^2,\mathscr{P}_{k-3},\mathscr{T})$$
$$\mathbb{R} \hookrightarrow \mathscr{V}(\mathrm{grad},\mathscr{P}_k,\mathscr{T}) \xrightarrow{\mathrm{grad}} \mathscr{V}(\mathrm{curl},\mathscr{P}_{k-1},\mathscr{T}) \xrightarrow{\mathrm{curl}} \mathscr{V}(\mathrm{div},\mathscr{P}^-_{k-1},\mathscr{T}) \xrightarrow{\mathrm{div}} \mathscr{V}(L^2,\mathscr{P}_{k-2},\mathscr{T})$$
$$\mathbb{R} \hookrightarrow \mathscr{V}(\mathrm{grad},\mathscr{P}_k,\mathscr{T}) \xrightarrow{\mathrm{grad}} \mathscr{V}(\mathrm{curl},\mathscr{P}^-_k,\mathscr{T}) \xrightarrow{\mathrm{curl}} \mathscr{V}(\mathrm{div},\mathscr{P}_{k-1},\mathscr{T}) \xrightarrow{\mathrm{div}} \mathscr{V}(L^2,\mathscr{P}_{k-2},\mathscr{T})$$
$$\mathbb{R} \hookrightarrow \mathscr{V}(\mathrm{grad},\mathscr{P}_k,\mathscr{T}) \xrightarrow{\mathrm{grad}} \mathscr{V}(\mathrm{curl},\mathscr{P}^-_k,\mathscr{T}) \xrightarrow{\mathrm{curl}} \mathscr{V}(\mathrm{div},\mathscr{P}^-_k,\mathscr{T}) \xrightarrow{\mathrm{div}} \mathscr{V}(L^2,\mathscr{P}_{k-1},\mathscr{T}).$$

There exist a sequence of interpolation operators

$$\Pi^{\mathscr{D}} : \boldsymbol{H}(\mathscr{D},\Omega)\cap \mathrm{dom}(\Pi^{\mathscr{D}}) \to \mathscr{V}(\mathscr{D},\mathscr{T})$$

to connect the Sobolev spaces $\boldsymbol{H}(\mathscr{D},\Omega)$ with corresponding finite element spaces $\mathscr{V}(\mathscr{D},\mathscr{T})$. These operators enjoy the following commutative diagram:

$$\begin{array}{ccccccccc}
\mathbb{R} & \longrightarrow & C^\infty(\Omega) & \xrightarrow{\mathrm{grad}} & C^\infty(\Omega) & \xrightarrow{\mathrm{curl}} & C^\infty(\Omega) & \xrightarrow{\mathrm{div}} & C^\infty(\Omega) \\
\downarrow & & \downarrow{\scriptstyle \Pi^{\mathrm{grad}}} & & \downarrow{\scriptstyle \Pi^{\mathrm{curl}}} & & \downarrow{\scriptstyle \Pi^{\mathrm{div}}} & & \downarrow{\scriptstyle \Pi^{L^2}} \\
\mathbb{R} & \longrightarrow & \mathscr{V}(\mathrm{grad},\mathscr{T}) & \xrightarrow{\mathrm{grad}} & \mathscr{V}(\mathrm{curl},\mathscr{T}) & \xrightarrow{\mathrm{curl}} & \mathscr{V}(\mathrm{div},\mathscr{T}) & \xrightarrow{\mathrm{div}} & \mathscr{V}(L^2,\mathscr{T}),
\end{array}$$

where for simplicity, we replace $\boldsymbol{H}(\mathscr{D},\Omega)\cap \mathrm{dom}(\Pi^{\mathscr{D}})$ by its subspace $C^\infty(\Omega)$.

The sequence in the bottom should be one of the 4 exact sequences in (**ES**). The operator $\Pi^{\mathscr{D}}$, of course, also depends on the specific choice of $\mathscr{V}(\mathscr{D},\mathscr{T})$. Operator $\Pi^{\mathscr{D}}$ is the identity restricted to $\mathscr{V}(\mathscr{D},\mathscr{T})$, namely

$$\Pi^{\mathscr{D}}\boldsymbol{v} = \boldsymbol{v}, \quad \text{for all } \boldsymbol{v} \in \mathscr{V}(\mathscr{D},\mathscr{T}). \tag{66}$$

We refer to [5, 48, 50] for the construction of such canonical interpolation operators. Here we list properties used later and refer to [50, Section 3.6 and Lemma 4.6] for proofs.

Lemma 5.2 (Operator Π^{curl}). *The interpolation operator Π^{curl} is bounded on $V = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) : \mathrm{curl}\,\boldsymbol{v} \in \mathscr{V}(\mathrm{div},\mathscr{T})\}$ and, with constants only depending on the shape regularity of $\mathscr{T}$, it satisfies*

$$\|h^{-1}(I-\Pi^{\mathrm{curl}})\boldsymbol{v}\| \lesssim \|\boldsymbol{v}\|_1, \quad \textit{for all } \boldsymbol{v} \in V. \tag{67}$$

Lemma 5.3 (Operator Π^{div}). *The interpolation operator Π^{div} is bounded on $\boldsymbol{H}^1(\Omega)$ and, with constants only depending on the shape regularity of $\mathscr{T}$, it satisfies*

$$\|h^{-1}(I-\Pi^{\mathrm{div}})\boldsymbol{v}\| \lesssim \|\boldsymbol{v}\|_1, \quad \textit{for all } \boldsymbol{v} \in \boldsymbol{H}^1(\Omega). \tag{68}$$

5.1.3 Regular decomposition

The Helmholtz (or Hodge) decomposition states that a vector field can be written as the sum of a gradient plus a curl. This decomposition is orthogonal in $L^2(\Omega)$ but requires regularity of Ω to be useful to us. Upon sacrificing L^2 orthogonality, we can decompose the space $\boldsymbol{H}(\mathscr{D},\Omega)$ into a regular part $\boldsymbol{H}^1(\Omega)$ plus the kernel of $\mathscr{D}$.

Theorem 5.3 (Regular Decomposition of $\boldsymbol{H}(\mathrm{curl};\Omega)$). *For any $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};\Omega)$, there exists $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ and $u \in H^1(\Omega)$ such that*

$$\boldsymbol{v} = \boldsymbol{\phi} + \mathrm{grad}\, u.$$

This decomposition is stable in the sense that

$$\|\boldsymbol{\phi}\|_1 + \|u\|_1 \lesssim \|\boldsymbol{v}\|_{A^c}.$$

Proof. For $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};\Omega)$, let $\boldsymbol{u} = \mathrm{curl}\,\boldsymbol{v} \in [L^2(\Omega)]^3$. Since $\mathrm{div}\,\mathrm{curl}\,\boldsymbol{v} = 0$, we can apply Theorem 5.2 to obtain $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ such that

$$\mathrm{curl}\,\boldsymbol{\phi} = \boldsymbol{u} = \mathrm{curl}\,\boldsymbol{v}, \text{ in } \Omega,$$

and

$$\|\boldsymbol{\phi}\|_1 \lesssim \|\boldsymbol{u}\| \leq \|v\|_{A^c}.$$

Since $\mathrm{curl}(\boldsymbol{v} - \boldsymbol{\phi}) = 0$, by Theorem 5.1, there exists $u \in H(\mathrm{grad};\Omega)$ such that

$$\mathrm{grad}\,u = \boldsymbol{v} - \boldsymbol{\phi},$$

and

$$\|u\|_1 \lesssim \|\boldsymbol{v}\| + \|\boldsymbol{\phi}\| \lesssim \|\boldsymbol{v}\|_{A^c}.$$

This completes the proof. □

The following lemma concerns the regular inversion of div operator.

Lemma 5.4 (Regular Inverse of div**).** *For any* $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega)$*, there exists* $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ *such that*

$$\mathrm{div}\,\boldsymbol{\phi} = \mathrm{div}\,\boldsymbol{v}, \quad \|\boldsymbol{\phi}\|_1 \lesssim \|\mathrm{div}\,\boldsymbol{v}\|.$$

Proof. Given $v \in \boldsymbol{H}(\mathrm{div};\Omega)$, let f be the zero extension of $\mathrm{div}\,\boldsymbol{v}$ to a smooth domain $\mathscr{O} \subset \mathbb{R}^3$ containing Ω; obviously $f \in L^2(\mathscr{O})$. We then solve the Poisson equation

$$-\Delta u = f \text{ in } \mathscr{O}, \quad u|_{\partial\mathscr{O}} = 0.$$

If $\boldsymbol{\phi} = -\mathrm{grad}\,u$, then $\mathrm{div}\,\boldsymbol{\phi} = -\Delta u = \mathrm{div}\,\boldsymbol{v}$ in $L^2(\mathscr{O})$. Since $u \in H^2(\mathscr{O})$ and $\|u\|_{2,\mathscr{O}} \lesssim \|f\|_{0,\mathscr{O}}$ because $\mathscr{O}$ is smooth, we deduce that $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ and

$$\|\boldsymbol{\phi}\|_{1,\Omega} \leq \|\boldsymbol{\phi}\|_{1,\mathscr{O}} \leq \|\mathrm{grad}\,u\|_{2,\mathscr{O}} \lesssim \|f\|_{0,\mathscr{O}} = \|\mathrm{div}\,\boldsymbol{v}\|_{0,\Omega},$$

which proves the assertion. □

Similar results can even be established for functions with appropriate traces on the boundary $\partial\Omega$. We refer to [35, 7] for specific constructions.

Theorem 5.4 (Regular Decomposition of $\boldsymbol{H}(\mathrm{div};\Omega)$**).** *For any* $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega)$*, there exist* $\boldsymbol{\phi}, \boldsymbol{u} \in [H^1(\Omega)]^3$ *such that*

$$\boldsymbol{v} = \boldsymbol{\phi} + \mathrm{curl}\,\boldsymbol{u}.$$

This decomposition is stable in the sense that

$$\|\boldsymbol{\phi}\|_1 + \|\boldsymbol{u}\|_1 \lesssim \|\boldsymbol{v}\|_{A^d}.$$

Proof. We first apply Lemma 5.4 to $\boldsymbol{v}$ to find $\boldsymbol{\phi} \in [H^1(\Omega)]^3$ such that

$$\mathrm{div}\,\boldsymbol{\phi} = \mathrm{div}\,\boldsymbol{v}, \quad \|\boldsymbol{\phi}\|_1 \lesssim \|\mathrm{div}\,\boldsymbol{v}\|.$$

Now since $\operatorname{div}(\boldsymbol{v}-\boldsymbol{\phi})=0$, we apply Theorem 5.2 to find $\boldsymbol{u}\in[H^1(\Omega)]^3$ such that

$$\operatorname{curl}\boldsymbol{u}=\boldsymbol{v}-\boldsymbol{\phi},\quad \|\boldsymbol{u}\|_1\lesssim\|\boldsymbol{v}-\boldsymbol{\phi}\|\le\|\boldsymbol{v}\|+\|\boldsymbol{\phi}\|\lesssim\|\boldsymbol{v}\|_{A^d}.$$

This is the asserted estimate. □

5.1.4 Discrete regular decomposition

We now present discrete regular decompositions for finite element spaces $\mathscr{V}(\operatorname{curl},\mathscr{T})$ and $\mathscr{V}(\operatorname{div},\mathscr{T})$, Theorem 5.5 and 5.6, following Hiptmair and Xu [52].

Theorem 5.5 (Discrete Regular Decomposition of $\mathscr{V}(\operatorname{curl},\mathscr{T})$). *Let $\mathscr{V}(\operatorname{grad},\mathscr{T}_h)$ and $\mathscr{V}(\operatorname{curl},\mathscr{T}_h)$ be a pair in the four exact sequences. For any $\boldsymbol{v}\in\mathscr{V}(\operatorname{curl},\mathscr{T}_h)$, there exist $\tilde{\boldsymbol{v}}\in\mathscr{V}(\operatorname{curl},\mathscr{T}_h),\boldsymbol{\phi}\in\mathscr{V}^3$, and $u\in\mathscr{V}(\operatorname{grad},\mathscr{T}_h)$ such that*

$$\boldsymbol{v}=\tilde{\boldsymbol{v}}+\Pi^{\operatorname{curl}}\boldsymbol{\phi}+\operatorname{grad}u, \tag{69}$$

$$\|h^{-1}\tilde{\boldsymbol{v}}\|+\|\boldsymbol{\phi}\|_1+\|u\|_1\lesssim\|\boldsymbol{v}\|_{A^c}. \tag{70}$$

Proof. For $\boldsymbol{v}\in\boldsymbol{H}(\operatorname{curl};\Omega)$, we apply the regular decomposition of Theorem 5.3 to obtain $\boldsymbol{v}=\boldsymbol{\Psi}+\operatorname{grad}U$ with

$$\boldsymbol{\Psi}\in[H^1(\Omega)]^3, U\in H^1(\Omega),\quad \|\boldsymbol{\Psi}\|_1+\|U\|_1\lesssim\|\boldsymbol{v}\|_{A^c}.$$

We then split $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi}=(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}+\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi}$, where $\mathscr{I}_{\mathscr{T}}:[H^1(\Omega)]^3\to\mathscr{V}^3$ is the vector version of the Scott-Zhang quasi-interpolation operator.

Since $\operatorname{curl}\boldsymbol{\Psi}=\operatorname{curl}\boldsymbol{v}\in\mathscr{V}(\operatorname{div},\mathscr{T}_h)$, by Lemma 5.2, $\Pi^{\operatorname{curl}}\boldsymbol{\Psi}$ is well defined. We apply the interpolation operator $\Pi^{\operatorname{curl}}$ to the decomposition

$$\boldsymbol{v}=(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}+\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi}+\operatorname{grad}U,$$

and use (66) to obtain the discrete decomposition

$$\boldsymbol{v}=\Pi^{\operatorname{curl}}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}+\Pi^{\operatorname{curl}}\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi}+\operatorname{grad}\Pi^{\operatorname{grad}}U.$$

This implies (69) with

$$\begin{aligned}
&\tilde{\boldsymbol{v}}=\Pi^{\operatorname{curl}}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}\in\mathscr{V}(\operatorname{curl},\mathscr{T}_h),\\
&\boldsymbol{\phi}=\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi}\in\mathscr{V}^3,\ \text{and}\\
&u=\Pi^{\operatorname{grad}}U-\frac{1}{\Omega}\int_\Omega\Pi^{\operatorname{grad}}U\,\mathrm{d}x\in\mathscr{V}(\operatorname{grad},\mathscr{T}_h).
\end{aligned}$$

We then prove this decomposition satisfies (70). First, by (67) and (49), we get

$$\begin{aligned}
\|h^{-1}\tilde{\boldsymbol{v}}\|&\le\|h^{-1}(I-\Pi^{\operatorname{curl}})(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}\|+\|h^{-1}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}\|\\
&\lesssim\|(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi}\|_1+\|\boldsymbol{\Psi}\|_1\lesssim\|\boldsymbol{\Psi}\|_1\lesssim\|\boldsymbol{v}\|_{A^c}.
\end{aligned}$$

Second, by the stability of $\mathscr{I}_T$ we obtain

$$\|\boldsymbol{\phi}\|_1 = \|\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi}\|_1 \lesssim \|\boldsymbol{\Psi}\|_1 \lesssim \|\boldsymbol{v}\|_{A^c},$$

and by that of Π^{grad} we have

$$\|u\|_1 \lesssim \|U\|_1 \lesssim \|\boldsymbol{v}\|_{A^c}.$$

This finishes the proof. □

The following regular decomposition is taken from Hiptmair and Xu [52]; see also Cascón, Nochetto, and Siebert [22].

Theorem 5.6 (Discrete Regular Decomposition of $\mathscr{V}(\mathrm{div},\mathscr{T})$). *Let $\mathscr{V}(\mathrm{curl},\mathscr{T}_h)$ and $\mathscr{V}(\mathrm{div},\mathscr{T}_h)$ be a pair in the four exact sequences. For any $\boldsymbol{v}\in\mathscr{V}(\mathrm{div},\mathscr{T}_h)$, there exist $\tilde{\boldsymbol{v}}\in\mathscr{V}(\mathrm{div},\mathscr{T}_h)$, $\boldsymbol{\phi}\in\mathscr{V}^3$, and $\boldsymbol{u}\in\mathscr{V}(\mathrm{curl},\mathscr{T}_h)$ such that*

$$\boldsymbol{v} = \tilde{\boldsymbol{v}} + \Pi^{\mathrm{div}}\boldsymbol{\phi} + \mathrm{curl}\,\boldsymbol{u}, \tag{71}$$

$$\|h^{-1}\tilde{\boldsymbol{v}}\| + \|\boldsymbol{\phi}\|_1 + \|\boldsymbol{u}\|_{A^c} \lesssim \|\boldsymbol{v}\|_{A^d}. \tag{72}$$

Proof. The proof is similar to that of Theorem 5.5 but a bit trickier. We first obtain

$$\boldsymbol{v} = \boldsymbol{\Psi} + \mathrm{curl}\,\boldsymbol{U}, \quad \|\boldsymbol{\Psi}\|_1 + \|\boldsymbol{U}\|_1 \lesssim \|\boldsymbol{v}\|_{A^d}.$$

But we cannot apply the interpolation operator Π^{div} directly and use the commutative diagram relation $\Pi^{\mathrm{div}}\,\mathrm{curl}\,\boldsymbol{U} = \mathrm{curl}\,\Pi^{\mathrm{curl}}\boldsymbol{U}$ because $\boldsymbol{U}\in[H^1(\Omega)]^3$ only and the interpolation Π^{curl} is not well defined on $[H^1(\Omega)]^3$.

To overcome this difficulty, we further split v as follows:

$$\boldsymbol{v} = (I-\Pi^{\mathrm{div}})\boldsymbol{\Psi} + \Pi^{\mathrm{div}}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi} + \Pi^{\mathrm{div}}\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi} + \mathrm{curl}\,\boldsymbol{U}. \tag{73}$$

Invoking the commutative diagram property $\mathrm{div}\,\Pi^{\mathrm{div}}\Psi = \Pi^{L^2}\,\mathrm{div}\,\Psi$, and the fact $\mathrm{div}\,\boldsymbol{\Psi} = \mathrm{div}\,\boldsymbol{v}\in\mathscr{V}(L^2,\mathscr{T})$, we have $\mathrm{div}(I-\Pi^{\mathrm{div}})\boldsymbol{\Psi} = 0$. Applying the regular inversion of curl operator (Lemma 5.3), there exists $\boldsymbol{Q}\in\boldsymbol{H}^1(\Omega)$ such that $\mathrm{curl}\,\boldsymbol{Q} = (I-\Pi^{\mathrm{div}})\boldsymbol{\Psi}$.

If $\tilde{\boldsymbol{U}} = \boldsymbol{U}+\boldsymbol{Q}$, then $\tilde{\boldsymbol{U}}\in\boldsymbol{H}^1(\Omega)$ and $\mathrm{curl}\,\tilde{\boldsymbol{U}}\in\mathscr{V}(\mathrm{div},\mathscr{T}_h)$. By Lemma 5.2, $\Pi^{\mathrm{curl}}\tilde{\boldsymbol{U}}$ is well defined. The decomposition (73) thus becomes

$$\boldsymbol{v} = \Pi^{\mathrm{div}}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi} + \Pi^{\mathrm{div}}\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi} + \mathrm{curl}\,\tilde{\boldsymbol{U}}.$$

We then apply Π^{div} operator to both sides and use property $\Pi^{\mathrm{div}}\,\mathrm{curl}\,\tilde{U} = \mathrm{curl}\,\Pi^{\mathrm{curl}}\tilde{U}$ to obtain

$$\boldsymbol{v} = \Pi^{\mathrm{div}}(I-\mathscr{I}_{\mathscr{T}})\boldsymbol{\Psi} + \Pi^{\mathrm{div}}\mathscr{I}_{\mathscr{T}}\boldsymbol{\Psi} + \mathrm{curl}\,\Pi^{\mathrm{curl}}\tilde{\boldsymbol{U}},$$

which implies (71).

The stability (72) of this decomposition is similar to that of Theorem 5.5. □

5.2 Space decomposition and multigrid methods

In this section, we first recall the space decomposition of $V(\mathrm{grad}, \mathscr{T})$ discussed in Section §4, following [27], and then present space decompositions for $V(\mathrm{curl}, \mathscr{T})$ and $V(\mathrm{div}, \mathscr{T})$. On the basis of these space decompositions, we develop multigrid methods for solving $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ systems. We consider bisection grids $\mathscr{T}_N$ which admits a decomposition $\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}$.

Let $\{\phi_p : p \in \mathscr{P}\}$, $\{\phi_e : e \in \mathscr{E}\}$, and $\{\phi_f : f \in \mathscr{F}\}$ be "nodal" basis functions. Namely $V(\mathrm{grad}, \mathscr{T}) = \mathrm{span}\{\phi_p : p \in \mathscr{P}\}, V(\mathrm{curl}, \mathscr{T}) = \mathrm{span}\{\phi_e : e \in \mathscr{E}\}$, and $V(\mathrm{div}, \mathscr{T}) = \mathrm{span}\{\phi_f : f \in \mathscr{F}\}$, where $\mathscr{P}$ (nodes), $\mathscr{E}$ (edges), and $\mathscr{F}$ (faces) are the degrees of freedom of the three spaces under consideration.

If $V_p = \mathrm{span}\{\phi_p\}$, $V_e = \mathrm{span}\{\phi_e\}$, and $V_f = \mathrm{span}\{\phi_f\}$ denote one dimensional subspaces, we then have the standard basis decompositions:

$$V(\mathrm{grad}, \mathscr{T}) = \sum_{p\in\mathscr{P}} V_p, \quad V(\mathrm{curl}, \mathscr{T}) = \sum_{e\in\mathscr{E}} V_e, \quad V(\mathrm{div}, \mathscr{T}) = \sum_{f\in\mathscr{F}} V_f.$$

Moreover, if $v = \sum_{p\in\mathscr{P}} v_p$, $\boldsymbol{v} = \sum_{e\in\mathscr{E}} \boldsymbol{v}_e$ and $\boldsymbol{v} = \sum_{f\in\mathscr{F}} \boldsymbol{v}_f$, then mesh shape regularity implies

$$\begin{aligned} \sum_{p\in\mathscr{P}} \|h^{-1} v_p\|^2 &\lesssim \|h^{-1} v\|^2, \\ \sum_{e\in\mathscr{E}} \|h^{-1} \boldsymbol{v}_e\|^2 &\lesssim \|h^{-1} \boldsymbol{v}\|^2, \\ \sum_{f\in\mathscr{F}} \|h^{-1} \boldsymbol{v}_f\|^2 &\lesssim \|h^{-1} \boldsymbol{v}\|^2. \end{aligned} \tag{74}$$

Let $\mathscr{T}_i = \mathscr{T}_0 + (b_1, \cdots, b_i)$ be the i-th mesh and $\phi_{i,p_i} \in V(\mathscr{T}_i; \mathscr{P}_1)$ denote the linear nodal basis associated with vertex $p_i \in \mathscr{N}(\mathscr{T}_i)$. We define the sub-spaces

$$V_0 = V(\mathscr{T}_0; \mathscr{P}_1), \quad V_i = \mathrm{span}\{\phi_{i,p_i}, \phi_{i,p_{l_i}}, \phi_{i,p_{r_i}}\},\ p_i \in \mathscr{N}(\mathscr{T}_i), \tag{75}$$

where recall that p_{l_i} and p_{r_i} are two end points of the edge and p_i is the middle point of that edge.

Space decompositions. We now present space decompositions of $V(\mathrm{curl}, \mathscr{T})$ and $V(\mathrm{div}, \mathscr{T})$ in the same vein of that for $V(\mathrm{grad}, \mathscr{T})$ of Section §4.6:

$$V(\mathrm{grad}, \mathscr{T}) = \sum_{p\in\mathscr{P}} V_p + \sum_{i=1}^{N} V_i. \tag{76}$$

If $\mathscr{R}_i$ is the ring of vertex p_i, which consists of all simplexes of $\mathscr{T}_i$ containing the vertex p_i, we define $V_i(\mathscr{D}, \mathscr{R}_i)$ as follows:

$$V_i(\mathrm{curl}, \mathscr{R}_i) = \Pi_i^{\mathrm{curl}} V_i^3 + \mathrm{grad}\, V_i. \tag{77}$$

$$V_i(\mathrm{div}, \mathscr{R}_i) = \Pi_i^{\mathrm{div}} V_i^3 + \mathrm{curl}\, V_i. \tag{78}$$

If $\mathscr{V}_i^3 \subset V(\mathscr{D},\mathscr{T})$, then the interpolation operator $\Pi_i^{\mathscr{D}}$ is the identity and we can ignore it. The macro space decompositions of $\mathscr{V}(\mathscr{D},\mathscr{T})$ are as follows:

$$\mathscr{V}(\mathrm{curl},\mathscr{T}) = \sum_{e\in\mathscr{E}} \mathscr{V}_e + \sum_{p\in\mathscr{P}} \mathrm{grad}\,\mathscr{V}_p + \sum_{i=0}^{N} \mathscr{V}_i(\mathrm{curl},\mathscr{R}_i), \tag{79}$$

$$\mathscr{V}(\mathrm{div},\mathscr{T}) = \sum_{F\in\mathscr{F}} \mathscr{V}_f + \sum_{e\in\mathscr{E}} \mathrm{curl}\,\mathscr{V}_e + \sum_{i=0}^{N} \mathscr{V}_i(\mathrm{div},\mathscr{R}_i). \tag{80}$$

Here for the convenience of notation, we include the coarsest space by defining $\mathscr{R}_0 = \mathscr{T}_0$ and $\mathscr{V}_0(\mathscr{D},\mathscr{R}_0) = \mathscr{V}(\mathscr{D},\mathscr{T}_0)$.

We will apply the Successive Subspace Correction (SSC) method to the space decompositions (79) and (80). The common feature is to apply smoothing in the finest space first and then the multilevel iteration to $\mathscr{V}_i(\mathscr{D},\mathscr{R}_i)$. For completeness, we also list the algorithm for $H(\mathrm{grad})$ problem.

$H(\mathrm{grad})$ **Problem**

$$u \leftarrow u + B^{\mathrm{grad}}(f - A^g u).$$

The operation of B^{grad} consists of two steps:

1. Smoothing in the finest space: $u \leftarrow u + S^{\mathrm{grad}}(f - A^g u)$
2. SSC for $H(\mathrm{grad})$ system on $\sum_i \mathscr{V}_i$:

$$u \leftarrow u + R_i Q_i (f - A^g u), \quad i = 0 : N.$$

$\boldsymbol{H}(\mathrm{curl})$ **System**

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{B}^{\mathrm{curl}}(\boldsymbol{f} - \boldsymbol{A}^c \boldsymbol{u}).$$

The operation of $\boldsymbol{B}^{\mathrm{curl}}$ consists of three steps:

1. Smoothing in the finest space: $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{S}^{\mathrm{curl}}(\boldsymbol{f} - \boldsymbol{A}^c \boldsymbol{u})$
2. Smoothing in the kernel space for the finest space $\sum_p \mathscr{V}_p$:

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \mathrm{grad}\, S^{\mathrm{grad}}(\boldsymbol{f} - \boldsymbol{A}^c \boldsymbol{u}),$$

3. SSC for $\boldsymbol{H}(\mathrm{curl})$ system on $\sum_i \mathscr{V}_i(\mathrm{curl},\mathscr{R}_i)$:

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{R}_i \boldsymbol{Q}_i (\boldsymbol{f} - \boldsymbol{A}^c \boldsymbol{u}), \quad i = 0 : N.$$

$\boldsymbol{H}(\mathrm{div})$ **System**

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{B}^{\mathrm{div}}(\boldsymbol{f} - \boldsymbol{A}^d \boldsymbol{u}).$$

The operation of $\boldsymbol{B}^{\mathrm{div}}$ consists of three steps:

1. Smoothing in the finest space: $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{S}^{\mathrm{div}}(\boldsymbol{f} - \boldsymbol{A}^d \boldsymbol{u})$
2. Smoothing in the kernel space for the finest space $\sum_e \mathscr{V}_e$:

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \mathrm{curl}\, \boldsymbol{S}^{\mathrm{curl}}(\boldsymbol{f} - \boldsymbol{A}^d \boldsymbol{u}).$$

3. SSC for $\boldsymbol{H}(\mathrm{div})$ system on $\sum_i \mathscr{V}_i(\mathrm{curl}, \mathscr{R}_i)$:

$$\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{R}_i \boldsymbol{Q}_i (\boldsymbol{f} - \boldsymbol{A}^d \boldsymbol{u}), \quad i = 0 : N.$$

5.3 Stable decomposition

We now prove that the multilevel space decompositions (79) and (80) are stable. Our approach is based on the stable decomposition for $\mathscr{V}(\mathrm{grad}; \mathscr{T})$ discussed in Section §4 (Theorem 4.3): for any $v \in \mathscr{V}(\mathrm{grad}, \mathscr{T})$, there exist $v_p \in \mathscr{V}_p, v_i \in \mathscr{V}_i$ such that

$$v = \sum_{p \in \mathscr{P}} v_p + \sum_{i=0}^{N} v_i, \tag{81}$$

$$\sum_{p \in \mathscr{P}} \|h^{-1} v_p\|^2 + \sum_{i=0}^{N} \|h_i^{-1} v_i\|^2 \lesssim \|v\|_{A^g}^2. \tag{82}$$

We first use the stable decomposition (81) and the discrete regular decomposition to give a space decomposition for $\mathscr{V}(\mathrm{curl}, \mathscr{T})$.We next employ the results of $\mathscr{V}(\mathrm{curl}, \mathscr{T})$ to give a stable decomposition of $\mathscr{V}(\mathrm{div}, \mathscr{T})$.

Theorem 5.7 (Stable Decomposition of $\mathscr{V}(\mathrm{curl}; \mathscr{T})$). *Let $\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}$ be a bisection grid. For every $\boldsymbol{v} \in \mathscr{V}(\mathrm{curl}, \mathscr{T}_N)$, there exist $\tilde{\boldsymbol{v}}_e \in \mathscr{V}_e, \tilde{u}_p \in \mathscr{V}_p$ and $w_i = \Pi_i^{\mathrm{curl}} \boldsymbol{\phi}_i + \mathrm{grad}\, u_i \in \mathscr{V}_i(\mathrm{curl}, \mathscr{R}_i)$ for all $e \in \mathscr{E}, p \in \mathscr{P}, i = 1 : N$, such that*

$$\boldsymbol{v} = \sum_{e \in \mathscr{E}} \tilde{\boldsymbol{v}}_e + \sum_{p \in \mathscr{P}} \mathrm{grad}\, \tilde{u}_p + \sum_{i=0}^{N} w_i, \tag{83}$$

$$\sum_{e \in \mathscr{E}} \|h^{-1} \tilde{\boldsymbol{v}}_e\|^2 + \sum_{p \in \mathscr{P}} \|h^{-1} \tilde{u}_p\|^2 + \sum_{i=0}^{N} \left(\|h^{-1} \boldsymbol{\phi}_i\|^2 + \|h^{-1} u_i\|^2 \right) \lesssim \|\boldsymbol{v}\|_{A^c}^2. \tag{84}$$

Proof. 1 We first consider the case $\mathscr{V}^3 \subset \mathscr{V}(\mathrm{curl}, \mathscr{T}_N)$ which excludes only the lowest order space $\mathscr{V}(\mathrm{curl}, \mathscr{P}_1^-, \mathscr{T}_N)$.

For any $\boldsymbol{v} \in \mathscr{V}(\mathrm{curl}, \mathscr{T}_N)$, we can apply Theorem 5.5 to obtain a discrete regular decomposition $\tilde{\boldsymbol{v}} \in \mathscr{V}(\mathrm{curl}, \mathscr{T}_N), \boldsymbol{\phi} \in \mathscr{V}^3$ and $u \in \mathscr{V}(\mathrm{grad}, \mathscr{T}_N)$ such that

$$\boldsymbol{v} = \tilde{\boldsymbol{v}} + \boldsymbol{\phi} + \mathrm{grad}\, u$$
$$\|h^{-1} \tilde{\boldsymbol{v}}\|^2 + \|\boldsymbol{\phi}\|_1^2 + \|u\|_1^2 \lesssim \|\boldsymbol{v}\|_1^2.$$

For $\mathscr{T}_N$, we can choose $\boldsymbol{\phi}$ so that $\boldsymbol{\phi} = \sum_{i=0}^{N} \boldsymbol{\phi}_i$ using the quasi-interpolation operator $\mathscr{I}_{\mathscr{T}}$ adapted to bisection grids; see Section §4.6 for the construction of $\mathscr{I}_{\mathscr{T}}$.

We apply the basis and multilevel decompositions of H^1 finite element spaces to obtain the desirable decomposition

$$\tilde{\boldsymbol{v}} = \sum_{e\in\mathscr{E}} \tilde{\boldsymbol{v}}_e, \boldsymbol{\phi} = \sum_{i=0}^{N} \boldsymbol{\phi}_i, \quad u = \sum_{p\in\mathscr{P}} \tilde{u}_p + \sum_{i=0}^{N} u_i.$$

The stability (84) of the decomposition results from the following inequalities:

1. $\sum_{e\in\mathscr{E}} \|h^{-1}\tilde{\boldsymbol{v}}_e\|^2 \lesssim \|h^{-1}\tilde{\boldsymbol{v}}\|^2$ by (74);
2. $\sum_{i=0}^{N} \|h^{-1}\boldsymbol{\phi}_i\|^2 \lesssim \|\boldsymbol{\phi}\|_1^2$ by the stable decomposition (55);
3. $\sum_{p\in\mathscr{P}} \|h^{-1}\tilde{u}_p\|^2 + \sum_{i=0}^{N} \|h^{-1}u_i\|^2 \lesssim \|u\|_1^2$ by the stable decomposition (82).

2 Now we consider the case $\mathscr{V}^3 \nsubseteq \mathscr{V}(\mathrm{curl}, \mathscr{T}_N)$, i.e., the space $\mathscr{V}(\mathrm{curl}, \mathscr{P}_1^-, \mathscr{T}_N)$. By Theorem 5.5, we have the discrete regular decomposition

$$\boldsymbol{v} = \tilde{\boldsymbol{v}} + \Pi^{\mathrm{curl}}\boldsymbol{\phi} + \mathrm{grad}\, u. \tag{85}$$

The key is a multilevel decomposition of the middle term. If $\boldsymbol{\phi} = \sum_{i=0}^{N} \boldsymbol{\phi}_i$ is the stable decomposition of $\boldsymbol{\phi}$, then

$$\Pi^{\mathrm{curl}}\boldsymbol{\phi} = \sum_{i=0}^{N} \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i + \Pi^{\mathrm{curl}} \sum_{i=0}^{N} (\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i), \tag{86}$$

because $\mathscr{V}(\mathrm{curl}, \mathscr{R}_i) \subset \mathscr{V}(\mathrm{curl}, \mathscr{T}_N)$ and $\Pi_i^{\mathrm{curl}} = \Pi^{\mathrm{curl}}\Pi_i^{\mathrm{curl}}$. We now show $\mathrm{curl}(\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i) = 0$. For any face $f \in \mathscr{F}(\mathscr{R}_i)$, using integration by parts and the definition of Π_i^{curl}, we conclude

$$\int_f \mathrm{curl}(\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i)\cdot \boldsymbol{n}\, dS = \int_{\partial f} (\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i)\cdot \boldsymbol{t}\, ds = 0.$$

Since $\mathrm{curl}(\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i)$ is piecewise constant, we deduce $\mathrm{curl}(\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i) = 0$.

From the exact sequence

$$\mathscr{V}(\mathrm{grad}, \mathscr{P}_2, \mathscr{R}_i) \to \mathscr{V}(\mathrm{curl}, \mathscr{P}_1, \mathscr{R}_i) \to \mathscr{V}(\mathrm{div}, \mathscr{P}_1^-, \mathscr{R}_i),$$

there exists $q_i \in \mathscr{V}(\mathrm{grad}, \mathscr{P}_2, \mathscr{R}_i)$ such that $\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i = \mathrm{grad}\, q_i$ and $\|q_i\| \lesssim \|\mathrm{grad}\, q_i\|$. Let $q = \sum q_i$ and $\int_\Omega q\, \mathrm{d}x = 0$. Using the commutative diagram, we have

$$\Pi^{\mathrm{curl}} \sum_{i=0}^{N} (\boldsymbol{\phi}_i - \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i) = \Pi^{\mathrm{curl}}\, \mathrm{grad} \sum_{i=0}^{N} q_i = \mathrm{grad}\, \Pi^{\mathrm{grad}} q,$$

where $\Pi^{\mathrm{grad}} : \mathscr{V}(\mathrm{grad}, \mathscr{P}_2, \mathscr{T}_N) \to \mathscr{V}(\mathrm{grad}, \mathscr{P}_1, \mathscr{T}_N)$. Let $\hat{u} = u + \Pi^{\mathrm{grad}} q$. Then $\hat{u} \in \mathscr{V}(\mathrm{grad}, \mathscr{P}_1, \mathscr{T}_N)$ and the decomposition (85) becomes

$$\boldsymbol{v} = \tilde{\boldsymbol{v}} + \sum_i \Pi_i^{\mathrm{curl}}\boldsymbol{\phi}_i + \mathrm{grad}\, \hat{u}. \tag{87}$$

We then apply the decomposition (81) to $\hat{u}$ as in the previous case, i.e.

$$\hat{u} = \sum_{p\in\mathscr{P}} \tilde{u}_p + \sum_{i=0}^{N} u_i,$$

to obtain the desired decomposition (83).

To prove the stability (84) of the decomposition, it suffices to prove

$$\|\operatorname{grad} q\| \lesssim \|\boldsymbol{v}\|_{A^c}, \tag{88}$$

which can be obtained from the Strengthened Cauchy Schwarz inequality

$$\begin{aligned}\|\operatorname{grad} q\|^2 &= (\sum_{i=0}^{N}\operatorname{grad} q_i, \sum_{i=0}^{N}\operatorname{grad} q_j) \le \sum_{i=0}^{N}\|\operatorname{grad} q_i\|^2 + 2\sum_{i=0}^{N}\sum_{j>i}^{N}|(\operatorname{grad} q_i, \operatorname{grad} q_j)| \\ &\lesssim \sum_{i=0}^{N}\|\operatorname{grad} q_i\|^2 = \sum_{i=0}^{N}\|\boldsymbol{\phi}_i - \Pi_i^{\operatorname{curl}}\boldsymbol{\phi}_i\|^2 \lesssim \sum_{i=0}^{N}\|h^{-1}\boldsymbol{\phi}_i\|^2 \lesssim \|\boldsymbol{\phi}\|_1^2 \lesssim \|\boldsymbol{v}\|_{A_c}^2.\end{aligned}$$

This completes the proof. □

We conclude with a similar result for $\mathscr{V}(\operatorname{div},\mathscr{T})$. Its proof follows along the same lines as those of Theorem 5.7. We refer to [28] for details.

Theorem 5.8 (Stable Decomposition of $H(\operatorname{div};\Omega)$). *Let $\mathscr{T}_N = \mathscr{T}_0 + \mathscr{B}$ be a bisection grid. For every $\boldsymbol{v} \in \mathscr{V}(\operatorname{div},\mathscr{T}_N)$ with $\mathscr{V}^3 \subset \mathscr{V}(\operatorname{curl},\mathscr{T}_N)$, there exist $\tilde{\boldsymbol{v}}_f \in \mathscr{V}_f, \tilde{\boldsymbol{u}}_e \in \mathscr{V}_e$ and $\boldsymbol{w}_i \in \mathscr{V}_i(\operatorname{div},\mathscr{R}_i)$ for all $f \in \mathscr{F}, e \in \mathscr{E}, i = 0:N$, such that*

$$\boldsymbol{v} = \sum_{f\in\mathscr{F}} \tilde{\boldsymbol{v}}_f + \sum_{e\in\mathscr{E}} \operatorname{curl}\tilde{\boldsymbol{u}}_e + \sum_{i=0}^{N}\boldsymbol{w}_i, \tag{89}$$

and

$$\sum_{F\in\mathscr{F}}\|h^{-1}\boldsymbol{v}_f\|^2 + \sum_{e\in\mathscr{E}}\|h^{-1}\tilde{\boldsymbol{u}}_e\|^2 + \sum_{i=0}^{N}\left(\|h^{-1}\boldsymbol{\phi}_i\|^2 + \|h^{-1}\boldsymbol{u}_i\|^2\right) \lesssim \|\boldsymbol{v}\|_{A^d}^2. \tag{90}$$

A remaining important ingredient, the SCS inequality for the space decompositions (79) and (80), can be established as well. Consequently, we have uniform convergence of multigrid methods for $\boldsymbol{H}(\operatorname{curl})$ or $\boldsymbol{H}(\operatorname{div})$ systems. We state the result below and refer to [28] for details.

Theorem 5.9. *The multigrid methods (c.f. algorithms in §5.2) for $\boldsymbol{H}(\operatorname{curl})$ or $\boldsymbol{H}(\operatorname{div})$ systems based on the space decompositions (79) or (80)), respectively, are uniformly convergent.*

6 The auxiliary space method and HX preconditioner for unstructured grids

In previous sections, we study multilevel methods formulated over a hierarchy of quasi-uniform or graded meshes. The geometric structure of these meshes is essential for both the design and analysis of such methods. Unfortunately, many grids in practice are not hierarchical.

We use the term *unstructured grids* to refer those grids that do not possess much geometric or topological structure. The design and analysis of efficient multilevel solvers for unstructured grids is a topic of great theoretical and practical interest. In this section, we discuss a special class of optimal preconditioners developed by Hiptmair and Xu [52] that can be effectively applied to unstructured grids. This type of preconditioners have been developed in the theoretical framework of the *auxiliary space method*.

6.1 The auxiliary space method

The method of subspace correction consists of solving a system of equations in a vector space by solving on appropriately chosen *subspaces* of the original space. Such subspaces are, however, not always available. The auxiliary space method (Xu 1996 [92]) is for designing preconditioners using auxiliary spaces which are not necessarily subspaces of the original subspace.

To solve the equation $a(u,v) = (f,v)$ in a Hilbert space $\mathscr{V}$, we consider

$$\overline{\mathscr{V}} = \mathscr{V} \times \mathscr{W}_1 \times \cdots \times \mathscr{W}_J \,, \tag{91}$$

where $\mathscr{W}_1, \ldots, \mathscr{W}_J$, $J \in \mathbb{N}$ are auxiliary (Hilbert) spaces endowed with inner products $\overline{a}_j(\cdot,\cdot)$, $j = 1, \ldots, J$.

A distinctive feature of the auxiliary space method is the presence of $\mathscr{V}$ in (91), but as a component of $\overline{\mathscr{V}}$. The space $\mathscr{V}$ is equipped with an inner product $d(\cdot,\cdot)$ different from $a(\cdot,\cdot)$. The operator $D : \mathscr{V} \mapsto \mathscr{V}$ induced by $d(\cdot,\cdot)$ on $\mathscr{V}$ leads to the smoother $S = D^{-1}$. For each $\mathscr{W}_j$ we need $\Pi_j : \mathscr{W}_j \mapsto \mathscr{V}$ which gives

$$\Pi := Id \times \Pi_1 \times \cdots \times \Pi_J : \overline{\mathscr{V}} \mapsto \mathscr{V}, \tag{92}$$

with properties

$$\|\Pi_j w_j\|_A \le c_j \overline{a}(w_j, w_j)^{1/2} \,, \qquad \text{for all } w_j \in \mathscr{W}_j, j = 1, \cdots, J \,, \tag{93}$$

$$\|v\|_A \le c_s d(v,v)^{1/2} \,, \qquad \text{for all } v \in \mathscr{V} \,, \tag{94}$$

and for every $v \in \mathscr{V}$, there exist $v_0 \in \mathscr{V}$ and $w_j \in \mathscr{W}_j$ such that $v = v_0 + \sum_{j=1}^{J} \Pi_j w_j$ and

$$d(v_0,v_0)^{1/2}+\sum_{j=1}^{J}\overline{a}_j(w_j,w_j)^{1/2}\le c_0\|v\|_A\,. \tag{95}$$

Let $\bar{A}_i$, for $i=1,\cdots,J$, be operators induced by $(\cdot,\cdot)_{A_i}$. Then the auxiliary space preconditioner is given by

$$B=S+\sum_{j=1}^{J}\Pi_j\bar{A}_j^{-1}\Pi_j^*. \tag{96}$$

The estimate of the condition number $\kappa(BA)$ is given below.

Theorem 6.1. *Let $\Pi=Id\times\Pi_1\times\cdots\times\Pi_J:\bar{\mathscr{V}}=\mathscr{V}\times\mathscr{W}_1\times\cdots\times\mathscr{W}_J\mapsto\mathscr{V}$ satisfy properties (93), (94), and (95). Then the auxiliary space preconditioner B given in (96) admits the following estimate:*

$$\kappa(BA)\le c_0^2(c_s^2+c_1^2+\cdots+c_J^2)\,. \tag{97}$$

Proof. 1 We first prove $(BAu,u)_A\le(c_s^2+c_1+\cdots+c_J^2)(u,u)_A$ and consequently $\lambda_{\max}(BA)\le(c_s^2+c_1+\cdots+c_J^2)$. By definition of B, we have:

$$(BAu,u)_A=(SAu,u)_A+\sum_{j=1}^{J}(\Pi_j\bar{A}_j^{-1}\Pi_j^*Au,u)_A.$$

We use Cauchy-Schwarz inequality and (94) to control the first term as

$$\begin{aligned}(SAu,u)_A&\le(SAu,SAu)_A^{1/2}(u,u)_A^{1/2}\le c_s d(SAu,SAu)^{1/2}(u,u)_A^{1/2}\\&=c_s(SAu,Au)^{1/2}(u,u)_A^{1/2}=c_s(SAu,u)_A^{1/2}(u,u)_A^{1/2},\end{aligned}$$

which leads to $(SAu,u)_A\le c_s^2(u,u)_A$.

Similarly we use Cauchy-Schwarz inequality and (93) to control the term as

$$\begin{aligned}(\Pi_j\bar{A}_j^{-1}\Pi_j^*Au,u)_A&\le(\Pi_j\bar{A}_j^{-1}\Pi_j^*Au,\Pi_j\bar{A}_j^{-1}\Pi_j^*Au)_A^{1/2}(u,u)_A^{1/2}\\&\le c_j(\bar{A}_j^{-1}\Pi_j^*Au,\bar{A}_j^{-1}\Pi_j^*Au)_{\bar{A}_j}^{1/2}(u,u)_A^{1/2}\\&=c_j(\bar{A}_j^{-1}\Pi_j^*Au,\Pi_j^*Au)^{1/2}(u,u)_A^{1/2}\\&=c_j(\Pi_j\bar{A}_j^{-1}\Pi_j^*Au,u)_A^{1/2}(u,u)_A^{1/2},\end{aligned}$$

which leads to $(\Pi_j\bar{A}_j^{-1}\Pi_j^*Au,u)_A\le c_j^2(u,u)_A$.

2 We then prove there exists $u\in\mathscr{V}$ such that $(u,u)_A\le c_0^2(BAu,u)_A$ and consequently $\lambda_{\min}(BA)\ge c_0^{-2}$.

We choose $u=v_0+\sum_{j=1}^{J}\Pi_j w_j$ satisfying (95). Then

$$(\Pi_j w_j, u)_A = (\Pi_j w_j, Au) = (w_j, \Pi_j^* Au) = (w_j, \bar{A}_j^{-1}\Pi_j^* Au)_{\bar{A}_j}$$
$$\leq \|w_j\|_{\bar{A}_j}(\bar{A}_j^{-1}\Pi_j^* Au, \bar{A}_j^{-1}\Pi_j^* Au)_{\bar{A}_j}^{1/2} = \|w_j\|_{\bar{A}_j}(BAu, u)_A^{1/2}.$$

Similarly $(v_0, u)_A \leq \|v_0\|_D (BAu, u)_A^{1/2}$. Therefore

$$(u,u)_A = (v_0 + \sum_{j=1}^{J} w_j, u)_A \leq (\|v_0\|_D + \sum_{j=1}^{J} \|w_j\|_{\bar{A}_j})(BAu, u)_A^{1/2}$$
$$\leq c_0 (u,u)_A^{1/2} (BAu, u)_A^{1/2},$$

which leads to the desired result. □

6.2 HX preconditioner

We present an *auxiliary space preconditioner* for $H(\text{curl})$ and $H(\text{div})$ systems developed in Hiptmair and Xu [52] (see also R. Beck [10] for a special case). The basic idea is to apply an auxiliary space preconditioner framework in [92], to the discrete regular decompositions of $\mathscr{V}(\text{curl}, \mathscr{T})$ or $\mathscr{V}(\text{div}, \mathscr{T})$. The resulting preconditioner for the $H(\text{curl})$ systems is

$$B^{\text{curl}} = S^{\text{curl}} + \Pi^{\text{curl}} \boldsymbol{B}^{\text{grad}} (\Pi^{\text{curl}})^t + \text{grad}\ B^{\text{grad}} (\text{grad})^t. \tag{98}$$

The implementation makes use of the input data: the $\boldsymbol{H}(\text{curl})$ stiffness matrix A, the coordinates of the grid points, along with the discrete gradient grad (for the lowest order Nédélec element case, it is simply the "vertex"-to-"edge" mapping with entries 1 or -1). Based on the coordinates, one can easily construct the interpolation operator Π_h^{curl}. Then the "Auxiliary space Maxwell solver" consists of the following three components:

1. The smoother S^{curl} of A (it could be the standard Jacobi or symmetric Gauss-Seidel methods).
2. An algebraic multigrid (AMG) solver B^{grad} for $\text{grad}^t A\, \text{grad}$
3. An (vector) AMG solver $\boldsymbol{B}^{\text{grad}}$ for $\left(\Pi^{\text{curl}}\right)^T A \Pi^{\text{curl}}$.

Similarly

$$B^{\text{div}} = S^{\text{div}} + \Pi^{\text{div}} \boldsymbol{B}^{\text{grad}} (\Pi_h^{\text{div}})^t + \text{curl}\ B^{\text{curl}} (\text{curl})^t$$
$$= S^{\text{div}} + \Pi^{\text{div}} \boldsymbol{B}^{\text{grad}} (\Pi^{\text{div}})^t + \text{curl}\ S^{\text{curl}} (\text{curl})^t + \text{curl}\, \Pi^{\text{curl}} \boldsymbol{B}^{\text{grad}} (\Pi^{\text{curl}})^t (\text{curl})^t.$$

This preconditioner consists of 4 Poisson solvers B^{grad} for $\boldsymbol{H}(\text{curl})$ (and 6 for $\boldsymbol{H}(\text{div})$) as well as 1 simple relaxation method (S^{curl}) such as point Jacobi for $\boldsymbol{H}(\text{curl})$ (and 2 relaxation methods for $\boldsymbol{H}(\text{div})$).

The point here is that we can use well-developed AMG for H^1 systems for the Poisson solver B^{grad} to obtain robust AMG methods for $\boldsymbol{H}(\text{curl})$ and $\boldsymbol{H}(\text{div})$ sys-

tems. These classes of preconditioners are in some way a "grey-box" AMG as it makes use of information on geometric grids (and associated interpolation operators). But the overhead is minimal and it requires very little programming effort. It has been proved in [52] that it is optimal and efficient for problems on unstructured grids.

To interpret B^{curl} as an auxiliary space preconditioner, we choose $\mathscr{V} = \mathscr{V}(\mathrm{curl}, \mathscr{T})$ and $\mathscr{W}_1 = \mathscr{W}_2 = \mathscr{V}(\mathrm{grad}, \mathscr{T})$. The inner product for the smoother is induced by the diagonal matrix of A^{curl} and the inner product $\bar{A}_1, \bar{A}_2$ is induced by $(\boldsymbol{B}^{\mathrm{grad}})^{-1}$. The operator $\Pi_1 : \mathscr{W}_1 \to \mathscr{V}$ is the interpolation Π^{curl} and $\Pi_2 = \mathrm{grad} : \mathscr{W}_1 \to \mathscr{V}$.

Theorem 6.2. *Suppose B^{grad} is an SPD matrix such that $((B^{\mathrm{grad}})^{-1}u, u) \eqsim (u,u)_1$. Then the preconditioner B^{curl} defined by (98) admits the estimate*

$$\kappa(B^{\mathrm{curl}} A^{\mathrm{curl}}) \lesssim 1.$$

Proof. In view of Theorem 97, it suffices to verify properties (93), (94), and (95).

The property (94) is an easy consequence of Cauchy-Schwarz inequality and shape regularity of the mesh. We use the stability of the operator $\Pi_1 = \Pi^{\mathrm{curl}}$ and $\Pi_2 = \mathrm{grad}$ discussed in Section 5.1 and inequality $(u,u)_1 \lesssim ((B^{\mathrm{grad}})^{-1}u, u)$ to get (94). To get (95), we can use the discrete regular decomposition in Section 5.1.4 and the inequality $((B^{\mathrm{grad}})^{-1}u, u) \lesssim (u,u)_1$. This completes the proof.

We state a similar result for B^{div} below and leave the proof to readers.

Theorem 6.3. *Suppose B^{grad} is an SPD matrix such that $((B^{\mathrm{grad}})^{-1}u, u) \eqsim (u,u)_1$. Then the preconditioner*

$$B^{\mathrm{div}} = S^{\mathrm{div}} + \Pi^{\mathrm{div}} \boldsymbol{B}^{\mathrm{grad}} (\Pi^{\mathrm{div}})^t + \mathrm{curl}\, S^{\mathrm{curl}} (\mathrm{curl})^t + \mathrm{curl}\, \Pi^{\mathrm{curl}} \boldsymbol{B}^{\mathrm{grad}} (\Pi^{\mathrm{curl}})^t (\mathrm{curl})^t$$

admits the estimate

$$\kappa(B^{\mathrm{div}} A^{\mathrm{div}}) \lesssim 1.$$

For $\boldsymbol{H}(\mathrm{curl})$ systems, the preconditioners have been included and tested in LLNL's *hypre* package [36, 37, 38] based on its parallel algebraic multigrid solver "BoomerAMG" [46]. It is a parallel implementation, almost a 'black-box" as it requires only discrete gradient matrix plus vertex coordinates, it can handle complicated geometries and coefficient jumps, scales with the problem size and on large parallel machines, supports simplified magnetostatics mode, and can utilize Poisson matrices, when available. Extensive numerical experiments demonstrate that this preconditioner is also efficient and robust for more general equations (see Hiptmair and Xu [52], and Kolev and Vassilevski [54, 55]) such as

$$\mathrm{curl}(\mu(x)\,\mathrm{curl}\, u) + \sigma(x) u = f \tag{99}$$

where μ and σ may be discontinuous, degenerate, and exhibit large variations.

For this type of general equations, we may not expect that the simple Poisson solvers are sufficient to handle possible variations of μ and σ. Let us argue roughly

what the right equations are to replace the Poisson equations. Let us assume our problems has sufficient regularity (e.g., Ω is convex). We then have

$$\|\operatorname{grad} \boldsymbol{u}\|^2 \eqsim \|\operatorname{curl} \boldsymbol{u}\|^2 + \|\operatorname{div} \boldsymbol{u}\|^2.$$

If $\boldsymbol{u}(=\operatorname{curl} w) \in N(\operatorname{curl})^{\perp}$, then $\|\operatorname{grad} \boldsymbol{u}\| = \|\operatorname{curl} \boldsymbol{u}\|$. Roughly, we get the following equivalence:

$$(\mu \operatorname{curl} \boldsymbol{u}, \operatorname{curl} \boldsymbol{u}) + (\sigma \boldsymbol{u}, \boldsymbol{u}) \eqsim (\mu \operatorname{grad} \boldsymbol{u}, \operatorname{grad} \boldsymbol{u}) + (\sigma \boldsymbol{u}, \boldsymbol{u}),$$

which corresponds to the following operator:

$$\boldsymbol{L}_1 \boldsymbol{u} \equiv -\operatorname{div}(\mu(x) \operatorname{grad} \boldsymbol{u}) + \sigma(x) \boldsymbol{u}. \tag{100}$$

On the other hand, if $\boldsymbol{u}, \boldsymbol{v} \in N(\operatorname{curl})$, $\boldsymbol{u} = \operatorname{grad} p$ and $\boldsymbol{v} = \operatorname{grad} q$,

$$(\mu \operatorname{curl} \boldsymbol{u}, \operatorname{curl} \boldsymbol{v}) + (\sigma \boldsymbol{u}, \boldsymbol{v}) = (\sigma \operatorname{grad} p, \operatorname{grad} q)$$

which corresponds to the following operator:

$$L_2 \boldsymbol{u} \equiv -\operatorname{div}(\sigma(x) \operatorname{grad} p). \tag{101}$$

We obtain the following preconditioner for the general equation (99):

$$B^{\mathrm{curl}} = S^{\mathrm{curl}} + \Pi^{\mathrm{curl}} \boldsymbol{B}_1^{\mathrm{grad}} (\Pi^{\mathrm{curl}})^t + \operatorname{grad} B_2^{\mathrm{grad}} (\operatorname{grad})^t$$

where is $\boldsymbol{B}_1^{\mathrm{grad}}$ is a preconditioner for the operator in the equation (100) and B_2^{grad} is a preconditioner for the operator in the equation (101).

The $\boldsymbol{H}(\mathrm{div})$ systems arise naturally from numerous problems of practical importance, such as stabilized mixed formulations of the Stokes problem, least squares methods for $H(\mathrm{grad})$ systems, and mixed methods for $H(\mathrm{grad})$ systems, see [3, 88]. Motivated by [13], we have recently designed a compatible gauge AMG algorithm for $\boldsymbol{H}(\mathrm{div})$ systems in [14], and the numerical experiments demonstrate the efficiency and robustness of this algorithm.

Acknowledgements L. Chen was supported in part by NSF Grant DMS-0811272, and in part by NIH Grant P50GM76516 and R01GM75309. This work is also partially supported by the Beijing International Center for Mathematical Research.

R.H. Nochetto was partially supported by NSF Grant DMS-0807811.

J. Xu was partially by NSF DMS-0609727 and DMS 0749202, and by Alexander von Humboldt Research Award for Senior US Scientists.

References

1. B. Aksoylu and M. Holst. Optimality of multilevel preconditioners for local mesh refinement in three dimensions. *SIAM J. Numer. Anal.*, 44(3):1005–1025, 2006.

2. D.N. Arnold. Differential complexes and numerical stability. *Plenary address delivered at ICM 2002 International Congress of Mathematicians*, 2004.
3. D.N. Arnold, R.S. Falk, and R. Winther. Preconditioning in $H(div)$ and applications. *Math. Comp.*, 66:957–984, 1997.
4. D.N. Arnold, R.S. Falk, and R. Winther. Multigrid in $H(div)$ and $H(curl)$. *Numer. Math.*, 85:197–218, 2000.
5. D.N. Arnold, R.S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, pages 1–155, 2006.
6. D.N. Arnold, A. Mukherjee, and L. Pouly. Locally adapted tetrahedral meshes using bisection. *SIAM J. Sci. Comput.*, 22(2):431–448, 2000.
7. D.N. Arnold, L.R. Scott, and M. Vogelius. Regular inversion of the divergence operator with Dirichlet boundary conditions on a polygon. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 15(2):169–192, (1989), 1988.
8. R.E. Bank, A.H. Sherman, and A. Weiser. Refinement algorithms and data structures for regular local mesh refinement. In *Scientific Computing*, pages 3–17. IMACS/North-Holland Publishing Company, Amsterdam, 1983.
9. E. Bänsch. Local mesh refinement in 2 and 3 dimensions. *Impact of Computing in Science and Engineering*, 3:181–191, 1991.
10. R. Beck. Algebraic multigrid by component splitting for edge elements on simplicial triangulations. Technical Report SC 99-40, ZIB, Berlin, Germany, 1999.
11. T.C. Biedl, P. Bose, E.D. Demaine, and A. Lubiw. Efficient algorithms for Petersen's matching theorem. In *Symposium on Discrete Algorithms*, pages 130–139, 1999.
12. P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219–268, 2004.
13. P.B. Bochev, J.J. Hu, C.M. Siefert, and R.S. Tuminaro. An Algebraic Multigrid Approach Based on a Compatible Gauge Reformulation of Maxwell's Equations. *SIAM J. Sci. Comput.*, 31:557–583, 2008.
14. P.B. Bochev, C.M. Siefert, R.S. Tuminaro, J. Xu, and Y. Zhu. Compatible Gauge Approaches for H(div) Equations. In *SNL-CSRI Proceeding*, 2007.
15. F.A. Bornemann and H. Yserentant. A basic norm equivalence for the theory of multilevel methods. *Numer. Math.*, 64:455–476, 1993.
16. J.H. Bramble. *Multigrid Methods*, volume 294 of *Pitman Research Notes in Mathematical Sciences*. Longman Scientific & Technical, Essex, England, 1993.
17. J.H. Bramble, J.E. Pasciak, and A.H. Schatz. The construction of preconditioners for elliptic problems by substructuring, IV. *Math. Comp.*, 53:1–24, 1989.
18. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57:1–21, 1991.
19. J.H. Bramble, J.E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.
20. J.H. Bramble and J. Xu. Some estimates for a weighted L^2 projection. *Math. Comp.*, 56:463–476, 1991.
21. A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31:333–390, 1977.
22. J.M. Cascón, R.H. Nochetto and K.G. Siebert. Design and Convergence of AFEM in $H(\mathrm{div};\Omega)$. *Math. Models Methods Appl. Sci.*, 17:1849–1881, 2007.
23. J.M. Cascón, C. Kreuzer, R.H. Nochetto, and K.G. Siebert. Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal.*, 46(5):2524–2550, 2008.
24. T.F. Chan and T.P. Mathew. *Domain Decomposition Algorithms*, volume 3 of *Acta Numerica*, pages 61–143. Cambridge University Press, Cambridge, 1994.
25. T.F. Chan and W.L. Wan. Robust multigrid methods for nonsmooth coefficient elliptic linear systems. *J. Comput. Appl. Math.*, 123(12):323–352, 2000. -
26. L. Chen. *i*FEM: an innovative finite element methods package in MATLAB. *Submitted*, 2009.
27. L. Chen, R.H. Nochetto, and J. Xu. Local multilevel methods on graded bisection grids: H^1 problem. Technical report, University of Maryland at College Park, 2007.

28. L. Chen, R.H. Nochetto, and J. Xu. Local multilevel methods on graded bisection grids: $H(curl$ and $H(div)$ systems. Technical report, University of Maryland at College Park, 2008.
29. L. Chen and C.-S. Zhang. A coarsening algorithm and multilevel methods on adaptive grids by newest vertex bisection. Technical report, Department of Mathematics, University of Maryland at College Park, 2007.
30. D. Cho, J. Xu, and L. Zikatanov. New estimates for the rate of convergence of the method of subspace corrections. *Numer. Math. Theor. Meth. Appl.*, 1:44–56, 2008.
31. W. Dahmen and A. Kunoth. Multilevel preconditioning. *Numer. Math.*, 63:315–344, 1992.
32. R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Spring-Verlag, New York, NY, 1993.
33. M. Dryja, M.V. Sarkis, and O.B. Widlund. Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. *Numer. Math.*, 72(3):313–348, 1996.
34. M. Dryja, B.F. Smith, and O.B. Widlund. Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. *SIAM J. Numer. Anal.*, 31, 1994.
35. R.G. Durán and M.A. Muschietti. An explicit right inverse of the divergence operator which is continuous in weighted norms. *Studia Math.*, 148(3):207–219, 2001.
36. R.D. Falgout, J.E. Jones, and U.M. Yang. Pursuing scalability for hypre's conceptual interfaces. *ACM Trans. Math. Software*, 31(3):326–350, 2005.
37. R.D. Falgout, J.E. Jones, and U.M. Yang. The design and implementation of hypre, a library of parallel high performance preconditioners. In *Numerical solution of partial differential equations on parallel computers*, volume 51 of *Lect. Notes Comput. Sci. Eng.*, pages 267–294. Springer, Berlin, 2006.
38. R.D. Falgout and U.M. Yang. Hypre: A Library of High Performance Preconditioners. In *International Conference on Computational Science (3)*, pages 632–641, 2002.
39. V. Girault and P.A. Raviart. *Finite Element Methods for Navier–Stokes Equations*. Springer-Verlag, Berlin, 1986.
40. G. Golub and C. van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
41. I.G. Graham and M.J. Hagger. Unstructured additive Schwarz-conjugate gradient method for elliptic problems with highly discontinuous coefficients. *SIAM J. Sci. Comput.*, 20:2041–2066, 1999.
42. M. Griebel and P. Oswald. On additive Schwarz preconditioners for sparse grid discretization. *Numer. Math.*, 66:449–464, 1994.
43. M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz methods. *Numer. Math.*, 70:163–180, 1995.
44. W. Hackbusch. *Multigrid Methods and Applications*, volume 4 of *Computational Mathematics*. Springer–Verlag, Berlin, 1985.
45. W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*, volume 95 of *Applied Mathematical Sciences*. Springer-Verlag, New York, translated and revised from the 1991 german original edition, 1994.
46. V.E. Henson and U.M. Yang. Boomeramg: a parallel algebraic multigrid solver and preconditioner. *Appl. Numer. Math.*, 41(1):155–177, 2002.
47. R. Hiptmair. Multigrid method for $H(div)$ in three dimensions. *Electron. Trans. Numer. Anal.*, 6:133–152, 1997.
48. R. Hiptmair. Canonical construction of finite elements. *Math. Comp.*, 68:1325–1346, 1999.
49. R. Hiptmair. Multigrid method for Maxwell's equations. *SIAM J. Numer. Anal.*, 36(1):204–225, 1999. unreadable.
50. R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, pages 237–339, 2002.
51. R. Hiptmair and A. Toselli. Overlapping and multilevel schwarz methods for vector valued elliptic problems in three dimensions. *IMA Volumes in Mathematics and its Applications*, 2000.
52. R. Hiptmair and J. Xu. Nodal auxiliary space preconditioning in H(curl) and H(div) spaces. *SIAM J. Numer. Anal.*, 45(6):2483–2509, 2007.
53. R. Hiptmair and W. Zheng. Local Multigrid in $H(curl)$. *J. Comput. Math.*, 27:573–603, 2009.
54. T. Kolev and P.S. Vassilevski. Some experience with a H1-based auxiliary space AMG for H(curl) problems. Technical Report 221841, LLNL, 2006.

55. T.V. Kolev and P.S. Vassilevski. Parallel Auxiliary Space AMG For H(curl) Problems. *J. Comput. Math.*, 27:604–623, 2009.
56. I. Kossaczký. A recursive approach to local mesh refinement in two and three dimensions. *J. Comput. Appl. Math.*, 55:275–288, 1994.
57. Y. Lee, J. Wu, J. Xu, and L. Zikatanov. Robust subspace correction methods for nearly singular systems. *M3AS*, 17:1937–1963, 2007.
58. A. Liu and B. Joe. Quality local refinement of tetrahedral meshes based on bisection. *SIAM J. Sci. Comput.*, 16(6):1269–1291, 1995.
59. J. Maubach. Local bisection refinement for *n*-simplicial grids generated by reflection. *SIAM J. Sci. Comput.*, 16(1):210–227, 1995.
60. W.F. Mitchell. *Unified Multilevel Adaptive Finite Element Methods for Elliptic Problems*. PhD thesis, University of Illinois at Urbana-Champaign, 1988.
61. W.F. Mitchell. A comparison of adaptive refinement techniques for elliptic problems. *ACM Trans. Math. Software (TOMS) archive*, 15(4):326–347, 1989.
62. W.F. Mitchell. Optimal multilevel iterative methods for adaptive grids. *SIAM J. Sci. Stat. Comp.*, 13:146–167, 1992.
63. R. Nochetto, K. Siebert, and A. Veeser. Theory of adaptive finite element methods: an introduction. In R.A. DeVore and A. Kunoth, editors, *Multiscale, Nonlinear and Adaptive Approximation*. Springer, 2009.
64. P. Oswald. On discrete norm estimates related to multilevel preconditioners in the finite element method. In *Constructive Theory of Functions, Proc. Int. Conf. Varna 1991*, pages 203–214, Sofia, 1992. Bulg. Acad. Sci.
65. P. Oswald. *Multilevel Finite Element Approximation, Theory and Applications*. Teubner Skripten zur Numerik. Teubner Verlag, Stuttgart, 1994.
66. P. Oswald. On the robustness of the BPX-preconditioner with respect to jumps in the coefficients. *Math. Comp.*, 68:633–650, 1999.
67. A. Plaza and G.F. Carey. Local refinement of simplicial grids based on the skeleton. *Appl. Numer. Math.*, 32(2):195–218, 2000.
68. A. Plaza and M.-C. Rivara. Mesh refinement based on the 8-tetrahedra longest-edge partition. *12th meshing roundtable*, pages 67–78, 2003.
69. M.C. Rivara. Mesh refinement processes based on the generalized bisection of simplexes. *SIAM J. Numer. Anal.*, 21:604–613, 1984.
70. M.-C. Rivara and P. Inostroza. Using longest-side bisection techniques for the automatic refinement fo Delaunay triangulations. *Int. J. Numer. Methods. Eng.*, 40:581–597, 1997.
71. M.C. Rivara and G. Iribarren. The 4-triangles longest-side partition of triangles and linear refinement algorithms. *Math. Comp.*, 65(216):1485–1501, 1996.
72. M.C. Rivara and M. Venere. Cost analysis of the longest-side (triangle bisection) refinement algorithms for triangulations. *Engineering with Computers*, 12:224–234, 1996.
73. J.W. Ruge and K. Stüben. Algebraic multigrid (AMG). In S.F. McCormick, editor, *Multigrid Methods*, volume 3 of *Frontiers in Applied Mathematics*, pages 73–130. SIAM, Philadelphia, PA, 1987.
74. A. Schmidt and K.G. Siebert. *Design of Adaptive Finite Element Software*, volume 42 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2005. The finite element toolbox ALBERTA, With 1 CD-ROM (Unix/Linux).
75. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54:483–493, 1990.
76. E.G. Sewell. Automatic generation of triangulations for piecewise polynomial approximation. In *Ph. D. dissertation*. Purdue Univ., West Lafayette, Ind., 1972.
77. B.F. Smith. A domain decomposition algorithm for elliptic problems in three dimensions. *Numer. Math.*, 60:219–234, 1991.
78. R. Stevenson. Stable three-point wavelet bases on general meshes. *Numer. Math.*, 80(1):131–158, 1998. V
79. R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

80. R. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77:227–241, 2008.
81. K. Stüben. An introduction to algebraic multigrid. In U. Trottenberg, C. Oosterlee, and A. Schüller, editors, *Multigrid*, pages 413–528. Academic Press, San Diego, CA, 2001.
82. A. Toselli and O. Widlund. *Domain Decomposition Methods: Algorithms and Theory*. Springer Series in Computational Mathematics, 2005.
83. C.T. Traxler. An algorithm for adaptive mesh refinement in n dimensions. *Computing*, 59(2):115–137, 1997.
84. C. Vuik, A. Segal, and J.A. Meijerink. An efficient preconditioned cg method for the solution of a class of layered problems with extreme contrasts in the coefficients. *J. Comput. Phys.*, 152(1):385–403, June 1999.
85. J. Wang. New convergence estimates for multilevel algorithms for finite-element approximations. *J. Comput. Appl. Math.*, 50:593–604, 1994.
86. J. Wang and R. Xie. Domain decomposition for elliptic problems with large jumps in coefficients. In *the Proceedings of Conference on Scientific and Engineering Computing*, pages 74–86. National Defense Industry Press, 1994.
87. O.B. Widlund. Some Schwarz methods for symmetric and nonsymmetric elliptic problems. In D.E. Keyes, T.F. Chan, G.A. Meurant, J.S. Scroggs, and R.G. Voigt, editors, *Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 19–36, Philadelphia, 1992. SIAM.
88. B.I. Wohlmuth, A. Toselli, and O.B. Widlund. An iterative substructuring method for Raviart–Thomas vector fields in three dimensions. *SIAM J. Numer. Anal.*, 37(5):1657–1676, 2000.
89. H. Wu and Z. Chen. Uniform convergence of multigrid v-cycle on adaptively refined finite element meshes for second order elliptic problems. *Science in China: Series A Mathematics*, 49(1):1–28, 2006.
90. J. Xu. Counter examples concerning a weighted L^2 projection. *Math. Comp.*, 57:563–568, 1991.
91. J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.
92. J. Xu. The auxiliary space method and optimal multigrid preconditioning techniques for unstructured meshes. *Computing*, 56:215–235, 1996.
93. J. Xu and Y. Zhu. Uniform convergent multigrid methods for elliptic problems with strongly discontinuous coefficients. *M3AS*, 2007.
94. J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.*, 15:573–597, 2002.
95. J. Xu and J. Zou. Some nonoverlapping domain decomposition methods. *SIAM Rev.*, 40(4):857–914, 1998.
96. H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numer.*, pages 285–326, 1993.
97. Y. Zhu. Domain decomposition preconditioners for elliptic equations with jump coefficients. *Numer. Linear Algebra Appl.*, 15:271–289, 2008.
98. L. Zikatanov. Two-sided bounds on the convergence rate of two-level methods. *Numer. Linear Algebra Appl.*, 15(5):439–454, 2008.