LECTURE NOTES IN COMPUTATIONAL SCIENCE AND ENGINEERING 128

Thomas Apel · Ulrich Langer
Arnd Meyer · Olaf Steinbach *Editors*

Advanced Finite Element Methods with Applications

Selected Papers from the 30th Chemnitz Finite Element Symposium 2017

Springer

Lecture Notes in Computational Science and Engineering 128

Editors:

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

Thomas Apel • Ulrich Langer • Arnd Meyer •
Olaf Steinbach
Editors

Advanced Finite Element Methods with Applications

Selected Papers from the 30th Chemnitz Finite Element Symposium 2017

Springer

Editors
Thomas Apel
Institut für Mathematik &
Computergestützte Simulation
Universität der Bundeswehr München
Neubiberg, Germany

Ulrich Langer
Institute for Computational Mathematics
Johannes Kepler University Linz
Linz, Austria

Arnd Meyer
Fakultät für Mathematik
TU Chemnitz
Chemnitz, Germany

Olaf Steinbach
Institut für Angewandte Mathematik
Technische Universität Graz
Graz, Austria

ISSN 1439-7358 ISSN 2197-7100 (electronic)
Lecture Notes in Computational Science and Engineering
ISBN 978-3-030-14243-8 ISBN 978-3-030-14244-5 (eBook)
https://doi.org/10.1007/978-3-030-14244-5

Mathematics Subject Classification (2010): 65-Fxx, 65-Mxx, 65-Nxx, 65-Yxx, 74-Sxx, 76-Mxx

Cover illustration: Image reprinted with kind permission from Ulrich Langer und Andreas Schafelner.

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

Preface

The annual Chemnitz Finite Element Symposium is a successful series of meetings on discretization methods for partial differential equations which is usually held in or nearby Chemnitz, Germany. The 30th Chemnitz Finite Element Symposium was on tour in Austria and held at the Federal Institute for Adult Education (BIfEB, Bundesinstitut für Erwachsenenbildung) in St. Wolfgang/Strobl, Austria, from September 25 to 27, 2017. It was jointly organized by the Technical University of Chemnitz, the Johannes Kepler University Linz, and the Johann Radon Institute for Computational and Applied Mathematics (RICAM) at the Austrian Academy of Sciences. About 80 participants from more than 10 countries attended the Symposium; see Fig. 1.

The scientific committee invited four experts to give keynote presentations on the theory and the application of finite element and related methods. Our keynote speakers and the titles of their talks were (in alphabetical order):

Mark Ainsworth (Brown University, Providence, USA), Fractional Cahn–Hilliard Equation(s): Analysis, Properties and Approximation
Volker John (WIAS and FU Berlin, Germany), Finite Elements for Scalar Convection-Dominated Equations and Incompressible Flow Problems—a Never Ending Story
Ricardo H. Nochetto (University of Maryland, College Park, USA), Numerical Methods for Fractional Diffusion
Gabriel Wittum (Goethe University Frankfurt, Germany, and KAUST, Saudi Arabia), Extreme Scale Solvers for Coupled Systems

Three invited speakers (M. Ainsworth, V. John, and R. Nochetto) contributed to a special issue of the Springer journal *Computing and Visualization in Science (CVS)*, volume 19, issue 5–6, 2018, which is dedicated to Ulrich Langer and Arnd Meyer on the occasion of their 65th birthdays in 2017. Both were the driving forces in the organization of the Finite Element Symposia in different periods. After Reinhard Lehmann organized the first Finite Element Symposium in 1978, the next six symposia were organized by U. Langer who moved to the Johannes Kepler University of Linz in October 1993. Since then, Arnd Meyer has served as

Fig. 1 Participants of the 30th Chemnitz Finite Element Symposium

the main organizer of Chemnitz Finite Element Symposia. The 30th Symposium was organized by both. In a special session, former PhD students of U. Langer and A. Meyer presented their results. Two of these contributions are also included in the CVS special issue, namely, the paper by D. Ganellari, G. Haase, and G. Zumbusch and the paper by C. Pechstein and S. Reitzinger.

The 60 contributed talks were given in parallel sessions. These sessions reflect the main topics of the conference including:

- High-order finite element methods (FEM) and isogeometric analysis
- The numerical treatment of partial differential equations (PDEs) with fractional derivatives
- Numerical methods for time-dependent problems
- Boundary element methods
- Computational structural mechanics
- A priori error estimates
- A posteriori error estimates and adaptivity
- Fast solution methods and parallel computing

The collection of contributions in this book contains original papers that are arranged in 19 chapters in alphabetical order. Here we give short discussions of these contributions in a systematic way reflecting the topics mentioned above.

There were several talks on *high-order finite elements methods (HOFEM) and isogeometric analysis (IGA)* that are closely related to each other. Two papers on HOFEM and two papers on IGA are included in this book. In Chap. 2, L. Banz, J. Petsche, and A. Schröder consider dual mixed *hp*-finite element methods for the Poisson equation with mixed boundary conditions and discuss different a posteriori error estimates and adaptive schemes, which was another topic of the symposium that is discussed below. M. Bernkopf and J. M. Melenk (Chap. 4) give an analysis of the *hp*-version of a first-order system least squares (FOSLS) method for the Helmholtz equation, especially for high wave numbers k. They provide a

quantitative analysis in terms of k, the mesh size h, and the polynomial degree p. In Chap. 11, C. Hofer and S. Takacs consider multi-patch IGA for elliptic boundary value problems and focus on the efficient solution of IGA equations on parallel computers; see also the corresponding topic that is discussed below in detail. The contribution by F. Scholz, A. Mantzaflaris, and B. Jüttler (Chap. 15) considers trimmed domains in IGA. They propose and analyze new trimmed quadrature rules that are very important for the correct generation of the IGA equations. Several numerical tests confirm the error estimates derived.

A second scientific topic was *PDEs with fractional derivatives* with two invited and eight contributed talks. While the invited speakers submitted their talks to the CVS special issue, see above, one contributed talk is reflected in the volume at hand. S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and J. Pasciak (Chap. 9) compare several methods for the solution of algebraic problems of the form $A^\alpha x = f$ where A is a symmetric and positive definite matrix and $\alpha \in (0, 1)$. Hence, the method is related to a discretization of the fractional Laplacian $(-\Delta)^\alpha$.

Chapters 6, 13, 16, 17, and 18 deal with *numerical methods for time-dependent problems*. H. Egger and Th. Kugler (Chap. 6) consider a time-stepping scheme in combination with some mixed finite element approximation in space in order to obtain fully discrete approximations to a linear damped wave system modeling the propagation of pressure waves in a pipeline. In Chap. 13, U. Langer, M. Neumüller, and A. Schafelner present a conforming, locally stabilized space–time finite element method for the numerical solution of parabolic initial boundary value problems with variable in space and time, possibly discontinuous coefficients on completely unstructured simplicial space–time meshes. In Chap. 16, O. Steinbach and H. Yang construct and analyze a Galerkin–Petrov space–time finite element method for solving a simplified version of the nonlinear bidomain equations. The contribution by O. Steinbach and M. Zank (Chap. 17) provides a stabilized space–time finite element method for the wave equation. In Chap. 18, I. Voulis presents a space–time discretization of parabolic initial boundary value problems with time-dependent Dirichlet boundary conditions that uses a continuous Galerkin discretization in space and a discontinuous Galerkin discretization in time on tensor product space–time meshes.

There are two contributions to efficient realizations of *boundary element methods*. H. Harbrecht and M. Moor (Chap. 8) consider wavelet boundary element methods for boundary integral equations that typically arise from boundary value problems for elliptic PDEs. In particular, they propose a goal-oriented refinement strategy that is based on an a posteriori error estimate for the linear goal functional. Several numerical examples show the efficiency of this approach. This chapter also contributes to the topic *a posteriori error estimates and adaptivity*. In Chap. 14, S. Rjasanow and S. Weißer propose a modification of the clustering procedure for the adaptive cross approximation (ACA) to boundary element matrices in the practically important case where edges appear in the surface on which the boundary integral operator is given. The numerical results show the efficiency of the proposed modification of the surface segmentation.

Three contributions are devoted to *computational structural mechanics*. L. Banz, J. Petsche, and A. Schröder introduce in Chap. 3 two stabilized three-field formulations for the biharmonic problem and derive a priori error estimates which are explicit in the mesh size and the polynomial degree. M. Frolov and O. Chistiakova (Chap. 7) justify an adaptive mesh refinement algorithm based on functional-type a posteriori local error indicator for Reissner–Mindlin plates. In Chap. 19, M. Weise introduces new hierarchic plate and shell elements and compares them with existing ones with respect to convergence, locking phenomena, and the performance of iterative solvers.

A classical subject of numerical analysis is the derivation of *a priori error estimates*. In Chap. 1, Th. Apel, M. Mateos, J. Pfefferer, and A. Rösch review results for graded and for superconvergent meshes and discuss novel meshes which are both graded and superconvergent, with a focus on Dirichlet control problems. As already mentioned, L. Banz, J. Petsche, and A. Schröder analyze two stabilized three-field formulations for the biharmonic problem in Chap. 3. I. Voulis shows convergence with optimal order for a discretization of a parabolic PDE with inhomogeneous Dirichlet boundary condition in Chap. 18.

Modern methods are *adaptive* which includes *a posteriori error estimation* and the definition of an improved finite element space by modifying the mesh and/or the distribution of the polynomial degrees. L. Banz, J. Petsche, and A. Schröder describe in Chap. 2 two a posteriori error estimates for the dual mixed finite element method on quadrilateral grids applied to the Poisson model problem and use adaptive refinement simultaneously in the mesh and in the polynomial degree. M. Bruchhäuser, K. Schwegler, and M. Bause investigate in Chap. 5 the dual weighted residual (DWR) method for the adaptive solution of the stationary diffusion–advection–reaction problem with stabilized finite elements. M. Frolov and O. Chistiakova (Chap. 7) recall a functional-type a posteriori local error indicator for Reissner–Mindlin plates and apply it within a new adaptive strategy. H. Harbrecht and M. Moor (Chap. 8) review the quasi-optimal convergence behavior of the adaptive wavelet boundary element method for boundary integral equations and extend it to goal-oriented adaptive refinement. V. Korneev studies in Chap. 12 guaranteed, robust, and consistent a posteriori error estimators for reaction–diffusion equations with a focus on flux recovery techniques.

Two chapters are devoted to *fast solution methods and parallel computing*. In Chap. 10, A. Heinlein, A. Klawonn, O. Rheinbach, and F. Röver extend the generalized Dryja–Smith–Widlund (GDSW) overlapping Schwarz domain decomposition preconditioner to a three-level version by recursively applying the GDSW preconditioner to the coarse problem. The authors present numerical results on the supercomputer JUQUEEN using up to 90,000 cores. The new three-level GDSW preconditioner shows excellent parallel scalability. In Chap. 11, C. Hofer and S. Takacs propose a new parallel multigrid solver for large-scale systems arising from the multi-patch isogeometric analysis of elliptic boundary value problems. The multigrid solver is robust with respect to both the mesh size and the spline degree. Moreover, the solver scales well in a parallel environment.

We very much hope that this collection of papers will be of interest to many applied mathematicians and engineers working at universities and research institutions as well as in industry. We would like to thank all the participants for their valuable contributions to the Chemnitz Finite Element Symposium 2017. In particular, we are grateful to the authors of the papers published in this collection of selected papers. Furthermore, we would like to thank the anonymous referees. In a short time, they did an excellent work that helped the authors to improve the quality of their contributions.

The editors are grateful to Christof Haubner (Universität der Bundeswehr München) who did a great job in combining and unifying the LaTeX contributions of all the authors. Finally, we express our thanks to Anne Comment and Jan Holland from Springer for continuing support and patience while preparing this volume.

Neubiberg, Germany — Thomas Apel
Linz, Austria — Ulrich Langer
Chemnitz, Germany — Arnd Meyer
Graz, Austria — Olaf Steinbach
December 2018

Contents

Contributors

Thomas Apel Institut für Mathematik und Computergestützte Simulation, Universität der Bundeswehr München, Neubiberg, Germany

Lothar Banz Department of Mathematics, University of Salzburg, Salzburg, Austria

Markus Bause Helmut Schmidt University, Hamburg, Germany

Maximilian Bernkopf Institute for Analysis and Scientific Computing, Technische Universität Wien, Vienna, Austria

Marius Paul Bruchhäuser Helmut Schmidt University, Hamburg, Germany

Olga Chistiakova Peter the Great St Petersburg Polytechnic University, St. Petersburg, Russia

Herbert Egger AG Numerics and Scientific Computing, TU Darmstadt, Darmstadt, Germany

Maxim Frolov Peter the Great St Petersburg Polytechnic University, St. Petersburg, Russia

Helmut Harbrecht Department Mathematik und Informatik, Universität Basel, Basel, Switzerland

Stanislav Harizanov Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria

Alexander Heinlein Mathematisches Institut, Universität zu Köln, Köln, Germany

Christoph Hofer Doctoral Program Computational Mathematics, University Linz, Linz, Austria

Bert Jüttler Institute of Applied Geometry, Johannes Kepler University Linz, Linz, Austria

Axel Klawonn Mathematisches Institut, Universität zu Köln, Köln, Germany

Vadim G. Korneev St. Petersburg State University, St. Petersburg, Russia

Thomas Kugler TU Darmstadt, AG Numerics and Scientific Computing, Darmstadt, Germany

Ulrich Langer Institute of Computational Mathematics, Johannes Kepler University Linz, Linz, Austria

Raytcho Lazarov Department of Mathematics, Texas A&M University, College Station, TX, USA

Angelos Mantzaflaris Institute of Applied Geometry, Johannes Kepler University Linz, Linz, Austria

Svetozar Margenov Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria

Pencho Marinov Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria

Mariano Mateos Departamento de Matemáticas, E.P.I. de Gijón, Universidad de Oviedo, Gijón, Spain

Jens Markus Melenk Institute for Analysis and Scientific Computing, Technische Universität Wien, Vienna, Austria

Manuela Moor Universitätsspital Basel, Basel, Switzerland

Martin Neumüller Institute of Computational Mathematics, Johannes Kepler University Linz, Linz, Austria

Joseph Pasciak Department of Mathematics, Texas A&M University, College Station, TX, USA

Jan Petsche Department of Mathematics, University of Salzburg, Salzburg, Austria

Johannes Pfefferer Lehrstuhl für Optimalsteuerung, Technische Universität München, Garching bei München, Germany

Oliver Rheinbach Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Freiberg, Germany

Sergej Rjasanow Department of Mathematics, Saarland University, Saarbrücken, Germany

Arnd Rösch Fakultät für Mathematik, Universtät Duisburg-Essen, Essen, Germany

Friederike Röver Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Freiberg, Germany

Andreas Schafelner Doctoral Program Computational Mathematics, Johannes Kepler University Linz, Linz, Austria

Felix Scholz Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria

Andreas Schröder Department of Mathematics, University of Salzburg, Salzburg, Austria

Kristina Schwegler Helmut Schmidt University, Hamburg, Germany

Olaf Steinbach Institut für Angewandte Mathematik, TU Graz, Graz, Austria

Stefan Takacs Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria

Igor Voulis Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Aachen, Germany

Michael Weise TU Chemnitz, Chemnitz, Germany

Steffen Weißer Department of Mathematics, Saarland University, Saarbrücken, Germany

Huidong Yang Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

Marco Zank Institut für Angewandte Mathematik, TU Graz, Graz, Austria

Acronyms

ACA	Adaptive Cross Approximation
AMG	Algebraic Multigrid
BDDC	Balancing Domain Decomposition by Constraints
BDM	Brezzi–Douglas–Marini
BEM	Boundary Element Method
BFS	Bogner–Fox–Schmit
B-rep	Boundary representation
BURA	Best Uniform Rational Approximation
CAD	Computer-Aided Design
CFL	Courant–Friedrichs–Lewy
CG	Conjugate Gradient Method
CG, cG	continuous Galerkin
cGP	continuous Galerkin–Petrov
CLT	Corrected Linearized Trimmed
DAE(s)	Differential Algebraic Equation(s)
DG, dG	discontinuous Galerkin
dGP	discontinuous Galerkin–Petrov
DOF(s), Dof(s), dof(s)	Degree(s) Of Freedom
DWR	Dual Weighted Residual
EAFE	Edge Average Finite Element
EKd	Endo–Kimura Plate Decoupled
EKh	Endo–Kimura Plate in Hierarchic Formulation
EKs	Plate Formulation of Endo and Kimura
EOC	Experimental Order of Convergence
FCT	Flux-Corrected Transport
FDTS	First Dualize and Then Stabilize
FE	Finite Element
FEM	Finite Element Method
FHN	FitzHugh–Nagumo
FOSLS	First-Order System Least Squares
FROSch	Fast and Robust Overlapping Schwarz

FSTD	First Stabilize and Then Dualize
GDSW	Generalized Dryja–Smith–Widlund
GMRES	Generalized Minimal RESidual
IgA	Isogeometric Analysis
ILU	Incomplete LU factorization
LPS	Local Projection Stabilization
LT	Linearized Trimmed
LU	Lower-Upper
MITC	Mixed Interpolation of Tensorial Components
MREK	Hierarchic Mindlin–Reissner Based on Endo–Kimura
MRh	Mindlin–Reissner Plate in Hierarchic Formulation
MRs	Standard Mindlin–Reissner Plate Formulation
MPI	Message Passing Interface
NURBS	Non-uniform Rational B-Splines
ORB	Mindlin–Reissner Plate in a Rotation Free Formulation by Oesterle, Ramm, and Bischoff
PCG	Preconditioned Conjugate Gradient
PDE	Partial Differential Equation
rHCT	reduced Hsieh–Clough–Tocher
RT	Raviart–Thomas
SPD	Symmetric Positive Definite
SPR	Superconvergent Path Recovery
SUPG	Streamline Upwind Petrov–Galerkin

Chapter 1
Superconvergent Graded Meshes for an Elliptic Dirichlet Control Problem

Thomas Apel, Mariano Mateos, Johannes Pfefferer, and Arnd Rösch

Abstract Superconvergent discretization error estimates can be obtained when the solution is smooth enough and the finite element meshes enjoy some structural properties. The simplest one is that any two adjacent triangles form a parallelogram. Existing results on finite element estimates on superconvergent meshes are reviewed, which can be used for numerical analysis of Dirichlet control problems. Moreover, an error estimate is given for a variational normal derivative which is of higher order on superconvergence meshes. Graded meshes can be used as a remedy of the reduced convergence order in the case of quasi-uniform meshes when elliptic boundary value problems with singularities in the vicinity of corners are treated. Discretization error estimates on graded meshes are reviewed. Depending on the construction, graded meshes may or may not have superconvergence properties. The discretization error of an elliptic Dirichlet control problem is discussed in the case of superconvergent graded meshes. Results of a paper in preparation are announced, where error estimates for Dirichlet optimal control problems on superconvergent graded meshes will be shown.

T. Apel (✉)
Institut für Mathematik und Computergestützte Simulation, Universität der Bundeswehr München, Neubiberg, Germany
e-mail: Thomas.Apel@unibw.de

M. Mateos
Departamento de Matemáticas, E.P.I. de Gijón, Universidad de Oviedo, Gijón, Spain
e-mail: mmateos@uniovi.es

J. Pfefferer
Lehrstuhl für Optimalsteuerung, Technische Universität München, Garching bei München, Germany
e-mail: pfefferer@ma.tum.de

A. Rösch
Fakultät für Mathematik, Universtät Duisburg-Essen, Essen, Germany
e-mail: Arnd.Roesch@uni-due.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_1

1.1 Introduction

Our motivation for the investigation of superconvergence properties of graded meshes comes from the investigation of an elliptic Dirichlet control problem. To introduce it, let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain with boundary Γ. The state variable y satisfies the Laplace equation with non-homogeneous Dirichlet boundary condition which is the control $u \in U_{\rm ad} := \{u \in L^2(\Gamma) : a \le u(x) \le b \text{ for a.a. } x \in \Gamma\}$, with $-\infty \le a < b \le +\infty$. The aim is to minimize the target functional

$$J(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Gamma)}$$

for some parameter $\nu > 0$ and given desired state $y_d \in H^s(\Omega)$, $s \ge 0$ as needed. Since the control is considered in $L^2(\Gamma)$ the state equation is understood in very weak sense, see [5]. This optimal control problem has a unique solution $\bar u \in U_{ad}$; see [3, Lemma 3.1]. The optimal control $\bar u \in U_{\rm ad}$, the corresponding state $\bar y \in Y$, and the corresponding adjoint state $\bar\varphi \in V$, satisfy the first order optimality conditions

$$\begin{aligned}
-\Delta \bar y &= 0 \text{ in } \Omega, \qquad \bar y = \bar u \text{ on } \partial\Omega \quad \text{in very weak sense,}\\
-\Delta \bar\varphi &= \bar y - y_d \text{ in } \Omega, \quad \bar\varphi = 0 \text{ on } \partial\Omega \quad \text{in weak sense,}\\
\bar u &= \Pi_{[a,b]}\left(\tfrac{1}{\nu}\partial_n \bar\varphi\right) \text{ on } \partial\Omega,
\end{aligned}$$

where the projection operator $\Pi_{[a,b]}$ is defined pointwise by $c \mapsto \min\{b, \max\{a,c\}\}$.

For the numerical solution of the optimal control problem consider a family of conforming finite element meshes $\mathscr{T}_h$. Define the finite element spaces

$$\begin{aligned}
Y_h &= \{v_h \in H^1(\Omega) : v_h|_T \in \mathscr{P}_1 \ \forall T \in \mathscr{T}_h\},\\
Y_{0h} &= Y_h \cap H^1_0(\Omega),\\
U_h &= Y_h|_\Gamma,\\
U^h_{ad} &= U_h \cap U_{ad},
\end{aligned}$$

where $\mathscr{P}_1$ is the space of first order polynomials. The discrete Dirichlet control problem is to minimize $J(y_h,u_h)$ for $(y_h,u_h) \in Y_h \times U^h_{ad}$ subject to

$$(\nabla y_h, \nabla v_h)_{L^2(\Omega)} = 0 \ \forall v_h \in Y_{0h}, \quad y_h|_\Gamma = u_h.$$

The corresponding discrete optimality system reads [14]

$$\begin{aligned}(\nabla \bar{y}_h, \nabla v_h)_{L^2(\Omega)} &= 0 \; \forall v_h \in Y_{0h}, \quad \bar{y}_h|_\Gamma = \bar{u}_h,\\ (\nabla \bar{\varphi}_h, \nabla v_h)_{L^2(\Omega)} &= (\bar{y}_h - y_d, v_h)_{L^2(\Omega)} \; \forall v_h \in Y_{0h},\\ (\nu \bar{u}_h - \partial_n^h \bar{\varphi}_h, u_h - \bar{u}_h)_{L^2(\Gamma)} &\geq 0 \; \forall u_h \in U_{ad}^h,\end{aligned}$$

where the variational discrete normal derivative $\partial_n^h \bar{\varphi}_h \in Y_h|_\Gamma$ is defined by [14]

$$(\partial_n^h \bar{\varphi}_h, v_h)_{L^2(\Gamma)} = -(\bar{y}_h - y_d, v_h)_{L^2(\Omega)} + (\nabla \bar{\varphi}_h, \nabla v_h)_{L^2(\Omega)} \; \forall v_h \in Y_h \backslash Y_{0h}. \tag{1.1}$$

Various approximation results are contained in [7, 14, 16, 22]. To state a general error estimate, we introduce the discrete harmonic extension operator $S_h : U_h \to Y_h$ and the $L^2(\Gamma)$-projection $Q_h : L^2(\Gamma) \to U_h$ by

$$\begin{aligned}(\nabla S_h u_h, \nabla v_h)_{L^2(\Omega)} &= 0 \quad \forall v_h \in Y_{0h}, \quad (S_h u_h)|_\Gamma = u_h,\\ (u - Q_h u, v_h)_{L^2(\Gamma)} &= 0 \quad \forall v_h \in U_h.\end{aligned}$$

Then the estimate

$$\begin{aligned}&\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} + \|\bar{y} - \bar{y}_h\|_{L^2(\Omega)}\\ &\leq c \left(\|\bar{u} - u_h^*\|_{L^2(\Gamma)} + \|\bar{y} - S_h Q_h \bar{u}\|_{L^2(\Omega)} + \sup_{v_h \in U_h} \frac{\left|(\nabla \bar{\varphi}, \nabla S_h v_h)_{L^2(\Omega)}\right|}{\|v_h\|_{L^2(\Gamma)}} \right)\end{aligned} \tag{1.2}$$

holds where $u_h^* \in U_{ad}^h$ is to be chosen such that

$$(\nu \bar{u} - \partial_n \bar{\varphi}, u_h^* - \bar{u})_{L^2(\Gamma)} = 0.$$

The first term is a special quasi-interpolation error, the second term contains the approximation of a non-smooth boundary condition and is non-trivial since $y \notin H^1(\Omega)$ in general. The third term corresponds to an error estimate of the normal derivative, see (1.9) below, and determines the overall convergence order. By the definition of S_h, the numerator is the absolute value of

$$(\nabla(\bar{\varphi} - I_h \bar{\varphi}), \nabla S_h v_h)_{L^2(\Omega)} \tag{1.3}$$

where we use the Lagrange interpolant $I_h : C(\bar{\Omega}) \to Y_h$. On a first look one can expect at best first order convergence for the gradient of the interpolation error and, together with $\|\nabla S_h v_h\|_{L^2(\Omega)} \leq c h^{-1/2} \|v_h\|_{L^2(\Gamma)}$, see (1.5) below, convergence with order $\frac{1}{2}$ for the third term in (1.2). However, we obtained at least first order for $\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)}$ in numerical tests, see [7], (which can be proved with

sharper arguments than above, see [7]), and sometimes even order 1.5 depending on the sequence of meshes used. This observation led us to the consideration of superconvergent meshes.

Many authors have contributed to the investigation of superconvergence effects, see for example [12, 13, 20, 21, 28, 29] and the references therein. All these authors consider families of quasi-uniform meshes with special structure. We found the paper by Bank and Xu [12] particularly useful for our investigations. Therefore we review it in Sect. 1.2 and add some further applications.

There is a second important point in the choice of the sequence of finite element meshes. The proof of a convergence order is always based on regularity assumptions on the functions to be approximated. It is well known that large interior angles in the corners of the domain Ω lead to reduced regularity of the solution of boundary value problems. In order to avoid a corresponding reduction of the convergence order one can use mesh grading, meshes with smaller and smaller element sizes towards the corners. For contributions in this direction we mention the papers [10, 11, 17, 23, 25, 26] and the overview article [2] and note that superconvergence effects are not considered there. We give a short overview in Sect. 1.3 with the focus on techniques to construct such families of graded meshes.

Coming back to our observation of superconvergence effects with certain families of graded meshes, we review few results from the literature and our own ones in Sect. 1.4. They are applied in Sect. 1.5 to the third term of estimate (1.2).

Unique features of the paper are the discussion of graded meshes with and without superconvergence properties and the connection to the discretization of Dirichlet control problems. Some estimates in the paper are new or at least less well known, see the estimate of the normal derivative in (1.10) and the estimate (1.15) for the trace of discrete harmonic functions. The paper gives a preview on some of the results in our upcoming paper [8].

1.2 Superconvergent Meshes

Bank and Xu [12] investigate quasi-uniform meshes with an $O(h^2)$ approximate parallelogram property. This means that the two elements which share an interior edge of the mesh, form an approximate parallelogram: the lengths of any two opposite edges differ by $O(h^2)$, see Fig. 1.1 for an illustration. For situations in which this property holds everywhere except in a region of size $O(h^{2\sigma})$, the quantity σ is traced. For some applications like Neumann boundary conditions there is another condition for boundary edges. The result in [12] which is important here, is the estimate (we denote the variable by φ in view of (1.3))

$$(\nabla(\varphi - I_h\varphi), \nabla v_h)_{L^2(\Omega)} \le ch^{1+\min\{1,\sigma\}} |\log h|^{1/2} \|\varphi\|_{W^{3,\infty}(\Omega)} |v_h|_{H^1(\Omega)} \tag{1.4}$$

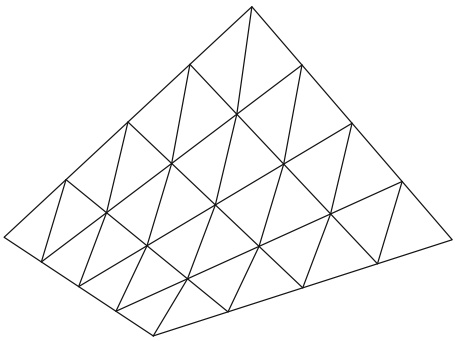


Fig. 1.1 Mesh with $O(h^2)$ approximate parallelogram property

for $\varphi \in W^{3,\infty}(\Omega)$ and any $v_h \in Y_h$ (piecewise linears). In the case that $\Omega = \bigcup_{j=1}^{n} \Omega_j$ and the meshes possess the $O(h^2)$ approximate parallelogram property in each polygon Ω_j, we speak about a *piecewise* $O(h^2)$ approximate parallelogram property. Note that this one is sufficient for (1.4) since we can apply the estimate (1.4) in subdomains and sum up.

Formula (1.4) is not immediately applicable to bound the third term in (1.2) since we would need $\|v_h\|_{L^2(\Gamma)}$ instead of $|v_h|_{H^1(\Omega)}$ on the right hand side of (1.4). The estimate $|S_h\psi_h|_{H^1(\Omega)} \le ch^{-1/2}\|\psi_h\|_{L^2(\Gamma)}$ for $\psi_h \in U_h$ helps. This estimate can be derived from $S_h\psi_h = \operatorname*{argmin}_{v_h \in \{w_h \in Y_h : w_h|_\Gamma = \psi_h\}} |v_h|_{H^1(\Omega)}$ since then

$$|S_h\psi_h|_{H^1(\Omega)} \le |\tilde{S}_h\psi_h|_{H^1(\Omega)} \le h^{-1/2}\|\psi_h\|_{L^2(\Gamma)} \tag{1.5}$$

with $\tilde{S}_h : U_h \to Y_h$ being the extension by zero operator.

As further applications of formula (1.4) we find supercloseness results for the interpolant and estimates for the recovered gradient in [12]. For the review here, denote by $\varphi \in W^{3,\infty}(\Omega)$ and $\varphi_h \in Y_h$ the solution of the homogenous Dirichlet problem for the Poisson equation and its finite element approximation, respectively. For the Lagrange interpolant $I_h : C(\bar{\Omega}) \to Y_h$ we have the supercloseness result

$$\begin{aligned} |\varphi_h - I_h\varphi|_{H^1(\Omega)} &= \sup_{v_h \in Y_{0h}} \frac{(\nabla(\varphi_h - I_h\varphi), \nabla v_h)}{|v_h|_{H^1(\Omega)}} = \sup_{v_h \in Y_{0h}} \frac{(\nabla(\varphi - I_h\varphi), \nabla v_h)}{|v_h|_{H^1(\Omega)}} \\ &\le ch^{1+\min\{1,\sigma\}}|\log h|^{1/2}\|\varphi\|_{W^{3,\infty}(\Omega)}, \end{aligned} \tag{1.6}$$

compare with

$$|\varphi - \varphi_h|_{H^1(\Omega)} \le |\varphi - I_h\varphi|_{H^1(\Omega)} \le ch|\varphi|_{H^2(\Omega)}$$

for the discretization error. As a corollary, Bank and Xu show in [12, Thm. 4.2] the superconvergence property of a recovered gradient,

$$\|\nabla\varphi - Q_h\nabla\varphi_h\|_{L^2(\Omega)} \le ch^{1+\min\{1,\sigma\}}|\log h|^{1/2}\|\varphi\|_{W^{3,\infty}(\Omega)}, \tag{1.7}$$

where $Q_h : L^2(\Omega)\times L^2(\Omega) \to Y_h \times Y_h$ is here the $L^2(\Omega)$-projection operator for each component.

Deckelnick et al. [16] considered the approximation of smooth domains, modified the estimate (1.4) to

$$\begin{aligned}&(\nabla(\varphi - I_h\varphi), \nabla v_h)_{L^2(\Omega_h)}\\ &\le c\|\varphi\|_{W^{3,r}(\Omega_h)}\left(h^{1+\min\{1,\sigma\}}\|v_h\|_{H^1(\Omega_h)} + h^{3/2}\|v_h\|_{L^2(\Gamma_h)}\right)\end{aligned} \tag{1.8}$$

with $r > 2$, and used it for the analysis of a Dirichlet control problem with L^2-regularization, as it was stated in the Introduction. They focused on smooth domains in order to avoid corner singularities such that quasi-uniform meshes are appropriate. With formula (1.5), they obtain approximation order $\frac{3}{2}$ for control and state which is optimal due to regularity issues. We would like to underline that these authors used meshes with a global $O(h^2)$ approximate parallelogram property (up to a region of size $O(h^{2\sigma})$). It is not obvious whether this estimate holds for meshes with only a piecewise $O(h^2)$ approximate parallelogram property; we discuss this in the Appendix. The result in [16] stimulated our treatment of superconvergence meshes within the investigation of Dirichlet control problems in non-smooth domains.

An application of formula (1.8) concerns the approximation of normal derivatives.

Lemma 1.1 *Let $\partial_n^h\varphi_h$ be the variational normal derivative of the Ritz projection $\varphi_h \in Y_h$ of $\varphi \in W^{3,r}(\Omega)$ with some $r > 2$, which can be defined by*

$$(\partial_n\varphi - \partial_n^h\varphi_h, z_h)_\Gamma = (\nabla(\varphi - \varphi_h), \nabla z_h)_\Omega \quad \forall z_h \in Y_h. \tag{1.9}$$

Then the estimate

$$\|\partial_n\varphi - \partial_n^h\varphi_h\|^2_{L^2(\Gamma)} \le ch^{1/2+\min\{1,\sigma\}}|\log h|^{1/2}\|\varphi\|_{W^{3,r}(\Omega)}, \ r > 2, \tag{1.10}$$

holds on superconvergent meshes.

Proof We can write

$$\begin{aligned}&\|\partial_n\varphi - \partial_n^h\varphi_h\|^2_{L^2(\Gamma)} = (\partial_n\varphi - \partial_n^h\varphi_h, \partial_n\varphi - \partial_n^h\varphi_h)_\Gamma\\ &= (\partial_n\varphi - \partial_n^h\varphi_h, \partial_n\varphi - Q_h\partial_n\varphi)_\Gamma + (\partial_n\varphi - \partial_n^h\varphi_h, Q_h\partial_n\varphi - \partial_n^h\varphi_h)_\Gamma =: I + II,\end{aligned}$$

where $Q_h : L^2(\Gamma) \to U_h$ is here the $L^2(\Gamma)$-projection. By using the Cauchy–Schwarz inequality, the approximation error estimate for Q_h, and

$$\|\partial_n \varphi\|_{W^{3/2,2}_{pw}(\Gamma)} := \sum_i \|\partial_n \varphi\|_{W^{3/2,2}(\Gamma_i)} \le c\|\varphi\|_{W^{3,r}(\Omega)}$$

for $r \ge 2$, where we denote by Γ_i the smooth parts of Γ, we obtain

$$\begin{aligned} I &\le \|\partial_n \varphi - \partial_n^h \varphi_h\|_{L^2(\Gamma)} \|\partial_n \varphi - Q_h \partial_n \varphi\|_{L^2(\Gamma)} \\ &\le c h^{3/2} \|\varphi\|_{W^{3,r}(\Omega)} \|\partial_n \varphi - \partial_n^h \varphi_h\|_{L^2(\Gamma)}. \end{aligned}$$

With $e_h := Q_h \partial_n \varphi - \partial_n^h \varphi_h = Q_h(\partial_n \varphi - \partial_n^h \varphi_h)$ and the discrete harmonic extension operator S_h we get with the help of the formula (1.8)

$$\begin{aligned} II &= (\partial_n \varphi - \partial_n^h \varphi_h, e_h)_\Gamma = (\nabla(\varphi - \varphi_h), \nabla S_h e_h)_\Omega = (\nabla(\varphi - I_h \varphi), \nabla S_h e_h)_\Omega \\ &\le c\|\varphi\|_{W^{3,r}(\Omega)} \left(h^{1+\min\{1,\sigma\}} |\log h|^{1/2} |S_h e_h|_{H^1(\Omega)} + h^{3/2} \|e_h\|_{L^2(\Gamma)} \right) \\ &\le c h^{1/2+\min\{1,\sigma\}} |\log h|^{1/2} \|\varphi\|_{W^{3,r}(\Omega)} \|e_h\|_{L^2(\Gamma)} \\ &\le c h^{1/2+\min\{1,\sigma\}} |\log h|^{1/2} \|\varphi\|_{W^{3,r}(\Omega)} \|\partial_n \varphi - \partial_n^h \varphi_h\|_{L^2(\Gamma)} \end{aligned}$$

where we have used the estimate $|S_h e_h|_{H^1(\Omega)} \le c h^{-1/2} \|e_h\|_{L^2(\Gamma)}$, the definition of e_h, and the stability of Q_h. Collecting all terms and dividing by $\|\partial_n \varphi - \partial_n^h \varphi_h\|_{L^2(\Gamma)}$ we find that (1.10) holds on superconvergent meshes. □

1.3 Graded Meshes

Graded meshes are widely used in the literature to cope with corner singularities in the solution of boundary value problems. To get an idea, consider a polygonal domain Ω with boundary Γ and the partial differential equation

$$-\Delta u + u = f \quad \text{in } \Omega$$

with Dirichlet or Neumann boundary conditions. The solution of this boundary value problem behaves in the vicinity of corners with interior angle ω like

$$u = u_r + u_s, \quad u_s = k\xi(r)\, r^\lambda\, \Phi(\theta)$$

with a regular part u_r, a coefficient k, a cut-off function ξ, polar coordinates (r, θ) centered in the corner, a smooth function Φ, and the singularity exponent $\lambda = \pi/\omega$. Simple calculations reveal how the regularity of u (resp. u_s) depends on the interior

angles of the domain: $u_s \in H^2(\Omega)$ for $\omega < \pi$ and $u_s \in W^{2,\infty}(\Omega)$ for $\omega < \pi/2$. To keep the explanation less technical, we assume here that there is only one critical corner and the solution is as smooth as needed near the other corners of the domain. Since corner singularities and their treatment are local phenomena, this is not a loss of generality.

Let us consider graded meshes of the following type. With the global mesh parameter h and the grading parameter $\mu \in (0, 1]$, let the element size $h_T := \operatorname{diam} T$ be related with the distance r_T to the critical corner by

$$h_T \sim \begin{cases} h^{1/\mu} & \text{for } r_T = 0, \\ h r_T^{1-\mu} & \text{for } R \geq r_T > 0, \\ h & \text{for } r_T > R, \end{cases} \tag{1.11}$$

where $a \sim b$ means the existence of constants c_1 and c_2 such that $c_1 b \leq a \leq c_2 b$.

The purpose of the mesh grading becomes clear when we consider approximation error estimates. For the piecewise linear finite element approximation u_h we have [10, 23–25]

$$\|u - u_h\|_{H^1(\Omega)} \leq c h^{\min\{1,\lambda/\mu-\varepsilon\}} \|f\|_{L^2(\Omega)},$$
$$\|u - u_h\|_{L^2(\Omega)} \leq c h^{2\min\{1,\lambda/\mu-\varepsilon\}} \|f\|_{L^2(\Omega)},$$

this means a grading condition $\mu < \lambda$ for optimal convergence, see also [2] for further references. $L^\infty(\Omega)$ error estimates for graded meshes,

$$\|u - u_h\|_{L^\infty(\Omega)} \leq c h^{\min\{2,\lambda/\mu\}-\varepsilon} \|f\|_X,$$

were originally proved by Schatz and Wahlbin [26] for smooth right hand sides f and $c = c(f)$ and later improved by Sirch [27] (Dirichlet problem, $X = C^{0,\sigma}(\Omega)$, $h^{-\varepsilon}$ being specified to $|\log h|^{3/2}$) and Rogovs et al. [9] (Neumann and Dirichlet problems, $X = C^{0,\sigma}(\Omega)$, $h^{-\varepsilon}$ specified to $|\log h|$). Estimates on the boundary like

$$\|u - u_h\|_{L^2(\Gamma)} \leq c h^{\min\{2,(\frac{1}{2}+\lambda)/\mu\}-\varepsilon} \|f\|_{C^{0,\sigma}(\Omega)}$$
$$\|\partial_n u - Q_h \partial_n u\|_{L^2(\Gamma)} \leq c h^{\min\{1/2,(\lambda-\frac{1}{2})/\mu\}-\varepsilon} \|f\|_{L^2(\Omega)}$$

for the Neumann and the Dirichlet problem, respectively, were proven by Pfefferer et al. [4, 6], with $h^{-\varepsilon}$ specified to $|\log h|^{1+\delta}$, $\delta = \delta(\lambda, \mu)$, in the first estimate, and Q_h being the $L^2(\Gamma)$-projection in the second estimate.

Let us now introduce three methods to generate graded meshes. The first method can be called *bisection method* and is described in [17]. The algorithm is initialized with a coarse start mesh. Afterwards, every element $T \in \mathcal{T}_h$ is bisected which satisfies

$$h_T > h \quad \text{or} \quad h_T > h \left(\frac{r_T}{R}\right)^{1-\mu}$$

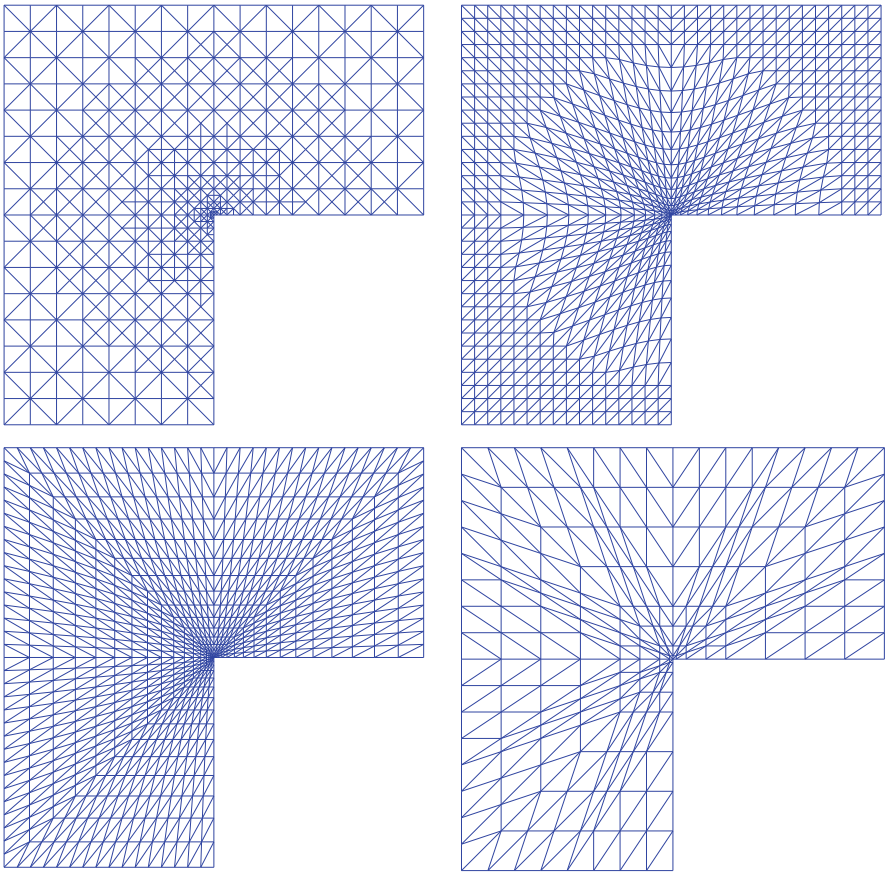


Fig. 1.2 Graded meshes with $\mu = 0.5$; top left: bisection, $R = 0.8$; top right: relocation with Euclidean metric, $R = 0.8$; bottom left: relocation with Manhattan metric, $R = 1$; bottom right: hierarchic with relocation

where R denotes the radius of the refinement zone. The splitting step is repeated until all elements have the desired mesh size. An illustration is provided in Fig. 1.2, top left. Advantages of this approach are the simple construction of a sequence of hierarchic meshes and the smooth transition of the element size from one element to the next. However, the approximate parallelogram property introduced in Sect. 1.2 is in general violated.

The second method can be called *relocation* and goes back to ideas by Oganesyan and Rukhovets [23, 24]. The idea is first to refine a coarse start mesh uniformly until $h_T \sim h$ for all $T \in \mathscr{T}_h$ with desired mesh size h, and then to transform the nodes $X^{(i)} \in \Omega$ with $r(X^{(i)}) < R$ according to

$$X^{(i)}_{new} = X^{(i)} \left(\frac{r(X^{(i)})}{R} \right)^{1/\mu - 1},$$

see the illustration in Fig. 1.2, top right. Raugel [25] used the same idea with the Manhattan metric instead of the Euclidean metric, see Fig. 1.2, bottom left. The relocation method does not lead to hierarchic meshes but the approximate parallelogram property can be achieved. We discuss this further in Sect. 1.4 but mention already that the approximate parallelogram property is not satisfied in general near the boundaries of the elements of the coarse mesh and near the boundary of the refinement zone.

The third method will be called *hierarchic with relocation* and was invented independently by Apel and Schöberl, [1], and Băcuţă, Nistor, and Zikatanov, [11]. The algorithm starts with a coarse mesh and splits the triangles recursively into four subtriangles by introducing new nodes at each edge. The new nodes are usually the midpoints of the edges except if an edge is adjacent to a singular corner. In this case the edge is split in the ratio $2^{-1/\mu} : 1$, see Fig. 1.2, bottom right, for an illustration. One can see clearly that there are subdomains with uniform meshes, meaning that all elements are congruent and hence a parallelogram property is satisfied within the subdomain. However, the number of such patches depends on $\log h$ for $\mu < 1$.

1.4 Superconvergent Graded Meshes

Numerical tests revealed that superconvergence effects occur also with certain graded meshes. The study of the literature led us only to the publications [15, 19]. Huang [19] investigated Raugel-type meshes and proved a supercloseness result as in (1.6) for $\mu < \frac{1}{2}\lambda$. Chen and Li [15] used hierarchic meshes with relocation (third method above) and proved estimates like (1.6) and (1.7) with an exponent $\frac{5}{4} - \varepsilon$ for $\mu < \lambda$. The proofs of all these estimates are based on results like (1.4) which can be used in subdomains and summed up. For the estimate of the third term on the right hand side of (1.2), which is of main interest for us, this did not help. In particular, we are using a different grading parameter μ.

In order to summarize our results of [8] we have to adapt the approximate parallelogram property to graded meshes with grading parameter μ: For an interior edge e, let T_e and T_e' denote the two elements of $\mathscr{T}_h$ which share this edge e, and denote by r_e the distance of e to the set of vertices of Ω. Then the lengths of any two opposite edges of the quadrilateral $T_e \cup T_e'$ differ only by $O(h_e^2 r_e^{-1})$. Such a property can be achieved in special situations, see for example Fig. 1.3 where a smoothened relocation was applied to a uniform initial mesh. In [8] we prove for any $v_h \in Y_h$ the estimate

$$\left| \int_\Omega \nabla(\varphi - I_h\varphi) \cdot \nabla v_h \right| \le c \left(\|r^{3(1-\mu)/2} \nabla^3 \varphi\|_{L^2(\Omega)} + \|r^{(1-3\mu)/2} \nabla^2 \varphi\|_{L^2(\Omega)} \right) \cdot \left(h^2 \|r^{(1-\mu)/2} \nabla v_h\|_{L^2(\Omega)} + h^{3/2} \|v_h\|_{L^2(\Gamma)} \right)$$

provided that the norms are finite. It is related to (1.4) but using weighted norms.

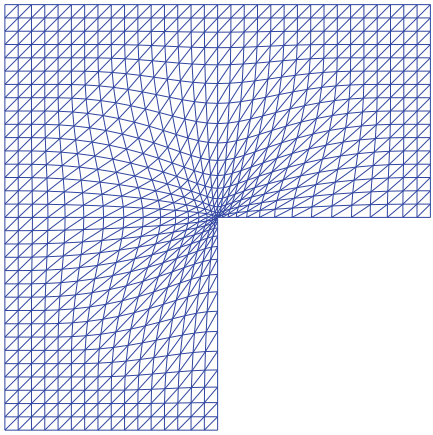


Fig. 1.3 Graded mesh with $\mu = 0.5$, smoothened relocation with Euclidean metric, $R = 1$

Let us discuss this result. First, if the function φ is the solution of a homogenous Dirichlet problem with sufficiently smooth right hand side, then the assumptions $r^{3(1-\mu)/2}\nabla^3\varphi \in L^2(\Omega)$ and $r^{(1-3\mu)/2}\nabla^2\varphi \in L^2(\Omega)$ are satisfied for $\mu < \frac{2}{3}(\lambda - \frac{1}{2})$ since $\nabla^m\varphi$ has leading singularity $r^{\lambda-m}$ and $r^\beta r^{\lambda-m} \in L^2(\Omega)$ for $\beta+\lambda-m > -1$. Second, in improvement of the proofs in [12] and [16] we avoided (weighted) L^r-norms with $r > 2$ of second and third derivatives of φ. Third, if the function v_h is discrete harmonic then

$$\|r^{(1-\mu)/2}\nabla v_h\|_{L^2(\Omega)} \le ch^{-1/2}\|v_h\|_{L^2(\Gamma)}$$

and the second factor on the right hand side is just $h^{3/2}\|v_h\|_{L^2(\Gamma)}$. Therefore we use $\|r^{(1-\mu)/2}\nabla v_h\|_{L^2(\Omega)}$ and not just $\|\nabla v_h\|_{L^2(\Omega)}$.

In the investigation of a Dirichlet control problem we use the estimate for the optimal adjoint state $\bar\varphi$, compare (1.2). Consequently, we get

$$\left|\int_\Omega \nabla(\bar\varphi - I_h\bar\varphi)\cdot\nabla v_h\right| \le ch^{3/2}\|v_h\|_{L^2(\Gamma)} \tag{1.12}$$

for $\mu < \frac{2}{3}(\lambda - \frac{1}{2})$ and discrete harmonic v_h.

Finally, we were able to relax the assumption of globally superconvergent graded meshes to piecewise superconvergent graded meshes. Note that if $\Omega = \bigcup_{j=1}^n \Omega_j$ and the meshes are superconvergent in each polygon Ω_j, we get from (1.12)

$$\left|\int_\Omega \nabla(\bar\varphi - I_h\bar\varphi)\cdot\nabla v_h\right| \le ch^{3/2}\sum_j \|v_h\|_{L^2(\partial\Omega_j)}$$

But we were able to show for discrete harmonic v_h and $\mu < 2\lambda - 1$

$$\sum_{j=1}^{n} \|v_h\|_{L^2(\partial\Omega_j)} \le c_n \|v_h\|_{L^2(\Gamma)} \tag{1.13}$$

such that (1.12) holds also for piecewise superconvergent graded meshes.

1.5 Application to Dirichlet Control Problems

We finally are able to obtain the following results for the error in the solution of the Dirichlet optimal control problem, whose proofs can be found in [8].

Our first result is for unconstrained problems, $a = -\infty$, $b = +\infty$. To prove the estimate

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} + \|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \le ch^{3/2-\varepsilon} \tag{1.14}$$

for arbitrary $\varepsilon > 0$, we assume $y_\Omega \in W^{1,p}(\Omega)$, $p > 2$, and use a family of superconvergent graded meshes with grading parameter $\mu < 2(\lambda - 1/2)/3$. Numerical experiments show that this quite strong grading is really necessary. Note for example that $\omega = 7\pi/4$ leads to $\lambda = 4/7$ and $2(\lambda - 1/2)/3 = 1/21$. Note also that grading is necessary near convex corners with obtuse angle, for example $\omega = 3\pi/4$ leads to $\lambda = 4/3$ and $2(\lambda - 1/2)/3 = 5/9$.

For constrained problems, $-\infty < a < b < \infty$, a technical assumption must be made. As we elaborate in [3], the normal derivative of the adjoint state will normally exhibit a singularity of type $r^{\lambda-1}$ at a critical corner, which tends to infinity for $\lambda = \pi/\omega < 1$, i.e., at a non-convex corner. We can deduce from this fact and the projection formula that the optimal control $\bar{u}$ is a constant function in a neighborhood of such corners. We will assume that the discrete optimal controls also satisfies this property in a neighborhood independent of the discretization parameter h, see (H), which has been observed in all our numerical experiments. Note that even a weaker assumption could be made, see the discussion in [7, Sect. 5].

(H) If $\partial_n\bar{\varphi}(x) \sim r^{\lambda-1}$ with $\lambda < 1$ at a critical corner, then there exists a neighborhood N with fixed radius such that $\bar{u}_h = \bar{u}$ in N for all $h < h_0$.

At nonconvex corners, it is also possible, though rare, that $\partial_n\bar{\varphi}$ behaves like $r^{2\lambda-1}$ which does not tend to infinity. This situation is analyzed in [8] as well, but here we assume for simplicity that this case does not happen.

To get the error estimate (1.14) in the constrained case, we may still need mesh grading: To cope with the polluting effects of the state singularity near concave corners we need $\mu \le 4\lambda/3$. Furthermore, grading with $\mu < 2(\pi/\omega - 1/2)/3$ near convex corners with interior angle $\omega \in [\pi/2, \pi)$ is necessary.

The computation of the optimal μ can be easily done for specific cases. For instance, suppose a polygonal domain with nonconvex angle $\omega = 7\pi/4$, so that the rest of interior angles are less than $\pi/2$. We have $\lambda = 4/7$, hence the optimal grading parameter is $\mu^* = 16/21 \approx 0.762$. If $\mu \leq \mu^*$, we obtain the order of convergence 1.5. Otherwise, the polluting effects of the state singularity influence the control approximation, despite the fact of being constant at the nonconvex corner, exhibiting a lower order of convergence $8/(7\mu)$. This behaviour is indeed observed in numerical experiments done with the different strategies of mesh grading described above; see [8] for the details.

Acknowledgements The project was supported by DFG through the International Research Training Group IGDK 1754 *Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures*. The second author was partially supported by the Spanish Ministerio Español de Economía y Competitividad under research projects MTM2014-57531-P and MTM2017-83185-P.

Appendix

We show here a result which was announced at the end of Sect. 1.2. Let Ω be split into subdomains Ω_i in which the quasi-uniform finite element mesh possesses locally the $O(h^2)$ approximate parallelogram property with $\sigma \geq 1$. Then the result (1.8) (for simplicity without approximating curved boundaries) can be written as

$$\begin{aligned}(\nabla(\varphi - I_h\varphi), \nabla v_h)_{L^2(\Omega)} &= \sum_i (\nabla(\varphi - I_h\varphi), \nabla v_h)_{L^2(\Omega_i)} \\ &\leq c \sum_i \|\varphi\|_{W^{3,r}(\Omega_i)} \left(h^2 \|v_h\|_{H^1(\Omega_i)} + h^{3/2} \|v_h\|_{L^2(\partial\Omega_i)} \right) \\ &\leq c \|\varphi\|_{W^{3,r}(\Omega)} \left(h^2 \|v_h\|_{H^1(\Omega)} + h^{3/2} \sum_i \|v_h\|_{L^2(\partial\Omega_i)} \right)\end{aligned}$$

where the constant depends on the number of subdomains which is assumed to be $O(1)$. In view of (1.2) and (1.3) we would like to bound both terms in the parentheses by $h^{3/2}\|v_h\|_{L^2(\Gamma)}$, and we can assume that v_h is discrete harmonic. We discussed the first term already in (1.5) such that $\|v_h\|_{H^1(\Omega)} \leq h^{-1/2}\|v_h\|_{L^2(\Gamma)}$. The challenge in the estimate of the second term is not to lose a power of h. We give here a proof for the case that Ω is convex and the mesh is quasi-uniform, and discuss extensions at the end of the section.

Lemma 1.2 *Let Ω be a convex domain which is split into subdomains Ω_i. Assume that in each Ω_i the quasi-uniform finite element meshes possess locally the $O(h^2)$*

approximate parallelogram property with $\sigma \geq 1$. Then the estimate

$$\sum_i \|v_h\|_{L^2(\partial\Omega_i)} \leq \|v_h\|_{L^2(\Gamma)}. \tag{1.15}$$

holds for discrete harmonic functions $v_h \in Y_h$.

Proof Let $S : U_h \to H^1(\Omega)$ and $S_h : U_h \to Y_h$ be defined via

$$\begin{aligned} (\nabla S v_h, \nabla v) &= 0 \quad \forall v \in H_0^1(\Omega), \quad (Sv_h)|_\Gamma = v_h, \\ (\nabla S_h v_h, \nabla v) &= 0 \quad \forall v \in Y_{0h}, \qquad (S_h v_h)|_\Gamma = v_h. \end{aligned}$$

By using the triangle inequality we have

$$\|S_h v_h\|_{L^2(\partial\Omega_i)} \leq \|Sv_h\|_{L^2(\partial\Omega_i)} + \|(S - S_h)v_h\|_{L^2(\partial\Omega_i)}.$$

The first term can be estimated by using the trace theorem for harmonic functions from [18, Lemma 2.3] and the regularity result from [5, Lemma 2.4]. We get

$$\|Sv_h\|_{L^2(\partial\Omega_i)} \leq c\|Sv_h\|_{H^{1/2}(\Omega)} \leq c\|v_h\|_{L^2(\Gamma)}.$$

For the second term we use a more standard trace theorem and get

$$\|(S - S_h)v_h\|_{L^2(\partial\Omega_i)} \leq c\,\|(S - S_h)v_h\|_{L^2(\Omega)}^{1/2} \|\nabla(S - S_h)v_h\|_{L^2(\Omega)}^{1/2}. \tag{1.16}$$

The first term was bounded in [5, Corollary 3.3] by

$$\|(S - S_h)v_h\|_{L^2(\Omega)} \leq ch^{1/2}\|v_h\|_{L^2(\Gamma)}$$

for convex domains. The second term is estimated by using the triangle inequality,

$$\|\nabla(S - S_h)v_h\|_{L^2(\Omega)} \leq \|\nabla S v_h\|_{L^2(\Omega)} + \|\nabla S_h v_h\|_{L^2(\Omega)}.$$

Both terms can be bounded by $ch^{-1/2}\|v_h\|_{L^2(\Gamma)}$; the second one is proved in (1.5), the first one can be proved with the same arguments. All theses estimates together show the desired result (1.15). Note that we used in several places that v_h is discrete harmonic and Ω is convex. □

The proof does not work for graded meshes although we would need a similar result, see (1.13). To cope with the nonuniform mesh size we suggest to replace the trace estimate (1.16) by a weighted one,

$$\|v\|_{L^2(\partial\Omega_i)} \leq c\,\|\sigma^{-(1-\mu)/2}v\|_{L^2(\Omega)}^{1/2} \|\sigma^{(1-\mu)/2}\nabla v\|_{L^2(\Omega)}^{1/2}$$

with $\sigma(x) = r(x) + c_I h^{1/\mu}$ and c_I being a large enough constant. It turns out that these weighted norms can also be bounded by $ch^{1/2}\|v_h\|_{L^2(\Gamma)}$ and $ch^{-1/2}\|v_h\|_{L^2(\Gamma)}$, respectively, see our upcoming article [8].

References

1. Apel, T., Schöberl, J.: Multigrid methods for anisotropic edge refinement. SIAM J. Numer. Anal. **40**(5), 1993–2006 (2002). https://doi.org/10.1137/S0036142900375414
2. Apel, T., Pfefferer, J., Rösch, A.: Locally refined meshes in optimal control for elliptic partial differential equations – an overview. In: Leugering, G., Benner, P., Engell, S., Griewank, A., Harbrecht, H., Hinze, M., Rannacher, R., Ulbrich, S. (eds.) Trends in PDE Constrained Optimization. International Series of Numerical Mathematics, vol. 165, pp. 285–302. Birkhäuser (Springer), Cham (2014)
3. Apel, T., Mateos, M., Pfefferer, J., Rösch, A.: On the regularity of the solutions of Dirichlet optimal control problems in polygonal domains. SIAM J. Control Optim. **53**(6), 3620–3641 (2015). https://doi.org/10.1137/140994186
4. Apel, T., Pfefferer, J., Rösch, A.: Finite element error estimates on the boundary with application to optimal control. Math. Comput. **84**(291), 33–70 (2015)
5. Apel, T., Nicaise, S., Pfefferer, J.: Discretization of the Poisson equation with non-smooth data and emphasis on non-convex domains. Numer. Methods Partial Differ. Equ. **32**, 1433–1454 (2016)
6. Apel, T., Nicaise, S., Pfefferer, J.: Adapted numerical methods for the numerical solution of the Poisson equation with L^2 boundary data in nonconvex domains. SIAM J. Numer. Anal. **55**, 1937–1957 (2017)
7. Apel, T., Mateos, M., Pfefferer, J., Rösch, A.: Error estimates for Dirichlet control problems in polygonal domains: quasi-uniform meshes. Math. Control Relat. Fields **8**, 217–245 (2018)
8. Apel, T., Mateos, M., Pfefferer, J., Rösch, A.: Error estimates for Dirichlet control problems in polygonal domains: superconvergent graded meshes. In preparation (2018)
9. Apel, T., Pfefferer, J., Rogovs, S., Winkler, M.: L^∞-error estimates for Neumann boundary value problems on graded meshes. IMA J. Numer. Anal. 1–24 (2018). https://doi.org/10.1093/imanum/dry076
10. Babuška, I.: Finite element method for domains with corners. Computing **6**, 264–273 (1970)
11. Băcuţă, C., Nistor, V., Zikatanov, L.: Improving the rate of convergence of 'high order finite elements' on polygons and domains with cusps. Numer. Math. **100**(2), 165–184 (2005). https://doi.org/10.1007/s00211-005-0588-3
12. Bank, R., Xu, J.: Asymptotically exact a posteriori error estimators, part I: grids with superconvergence. SIAM J. Numer. Anal. **41**(6), 2294–2312 (2003)
13. Brandts, J., Křížek, M.: History and future of superconvergence in three-dimensional finite element methods. In: Finite Element Methods (Jyväskylä, 2000). GAKUTO International Series. Mathematical Sciences and Applications, vol. 15, pp. 22–33. Gakkōtosho, Tokyo (2001)
14. Casas, E., Raymond, J.P.: Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations. SIAM J. Control Optim. **45**(5), 1586–1611 (2006) (electronic). https://doi.org/10.1137/050626600
15. Chen, L., Li, H.: Superconvergence of gradient recovery schemes on graded meshes for corner singularities. J. Comput. Math. **28**(1), 11–31 (2010). https://doi.org/10.4208/jcm.2009.09-m1002
16. Deckelnick, K., Günther, A., Hinze, M.: Finite element approximation of Dirichlet boundary control for elliptic PDEs on two- and three-dimensional curved domains. SIAM J. Control Optim. **48**(4), 2798–2819 (2009). https://doi.org/10.1137/080735369

17. Fritzsch, R., Oswald, P.: Zur optimalen Gitterwahl bei Finite-Elemente-Approximationen. Wiss. Z. TU Dresd. **37**(3), 155–158 (1988)
18. Gesztesy, F., Mitrea, M.: Generalized Robin boundary conditions, Robin-to-Dirichlet maps, and Krein-type resolvent formulas for Schrödinger operators on bounded Lipschitz domains. In: Perspectives in Partial Differential Equations, Harmonic Analysis and Applications. Proceedings of Symposia in Pure Mathematics, vol. 79, pp. 105–173. American Mathematical Society, Providence (2008). https://doi.org/10.1090/pspum/079/2500491
19. Huang, Y.Q.: The superconvergence of finite element methods on domains with reentrant corners. In: Finite Element Methods (Jyväskylä, 1997). Lecture Notes in Pure and Applied Mathematics, vol. 196, pp. 169–182. Dekker, New York (1998)
20. Křížek, M., Neittaanmäki, P.: On superconvergence techniques. Acta Appl. Math. **9**(3), 175–198 (1987). https://doi.org/10.1007/BF00047538
21. Lakhany, A., Marek, I., Whiteman, J.: Superconvergence results on mildly structured triangulations. Comput. Methods Appl. Mech. Eng. **189**(1), 1–75 (2000). https://doi.org/10.1016/S0045-7825(99)00281-9
22. May, S., Rannacher, R., Vexler, B.: Error analysis for a finite element approximation of elliptic Dirichlet boundary control problems. SIAM J. Control Optim. **51**(3), 2585–2611 (2013). https://doi.org/10.1137/080735734
23. Oganesyan, L., Rukhovets, L.: Variational-difference schemes for linear second-order elliptic equations in a two-dimensional region with piecewise smooth boundary. Zh. Vychisl. Mat. Mat. Fiz. **8**, 97–114 (1968). In Russian. English translation in USSR Comput. Math. Math. Phys. 8, 129–152 (1968)
24. Oganesyan, L., Rukhovets, L.: Variational-difference methods for the solution of elliptic equations. Izd. Akad. Nauk Armyanskoi SSR, Jerevan (1979). In Russian
25. Raugel, G.: Résolution numérique par une méthode d'éléments finis du problème de Dirichlet pour le Laplacien dans un polygone. C. R. Acad. Sci. Paris A **286**(18), A791–A794 (1978)
26. Schatz, A., Wahlbin, L.: Maximum norm estimates in the finite element method on plane polygonal domains. Part 2: Refinements. Math. Comput. **33**(146), 465–492 (1979)
27. Sirch, D.: Finite element error analysis for pde-constrained optimal control problems: the control constrained case under reduced regularity. Ph.D. thesis, TU München (2010). http://mediatum2.ub.tum.de/node?id=977779
28. Wahlbin, L.: Superconvergence in Galerkin Finite Element Methods. Lecture Notes in Mathematics, vol. 1605. Springer, Berlin (1995)
29. Zhu, J., Zienkiewicz, O.: Superconvergence recovery technique and a posteriori error estimators. Int. J. Numer. Methods Eng. **30**, 1321–1339 (1990)

Chapter 2
Explicit and Implicit Reconstructions of the Potential in Dual Mixed *hp*-Finite Element Methods

Lothar Banz, Jan Petsche, and Andreas Schröder

Abstract In this paper, two different reconstruction techniques for the discrete potential in dual mixed *hp*-finite element methods are discussed and compared: a post-processed reconstruction based on an explicit computation of local solutions and a reconstruction technique in which the reconstruction is not explicitly computed. Both approaches enable the derivation of a posteriori error estimates which can be used to drive *h*- as well as *hp*-adaptive schemes. In several numerical experiments the convergence of the adaptive schemes and resulting efficiency indices are studied. Moreover, efficiency indices for *h*-uniform and *hp*-geometric refinements as well as the computational amount required for the error estimations are compared for both approaches.

2.1 Introduction

Dual mixed methods are well-established finite element approaches, which are based on the introduction of a flux field as an additional unknown in $H(\mathrm{div})$ [5, 12]. Raviart-Thomas finite elements are often used for the discretization of the $H(\mathrm{div})$-space providing its continuity requirements in the normal direction of the edges of the underlying mesh. An alternative is to apply dual mixed-hybrid methods to ensure these continuity requirements by defining an additional Lagrange multiplier on the edges. The provision of the flux, in particular, without the need of additional post-processing is an advantage of dual mixed methods as opposed to primal methods. A certain drawback is that the potential is just in L^2 and, thus, not conforming with respect to the primal trial H^1-spaces. Typically, post-processing schemes are used for the reconstruction of the potential in H^1. A commonly used technique for reconstruction is proposed in [36], which is the basis of the reconstructions

L. Banz · J. Petsche · A. Schröder (✉)
University of Salzburg, Salzburg, Austria
e-mail: lothar.banz@sbg.ac.at; jan.petsche@sbg.ac.at; andreas.schroeder@sbg.ac.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_2

introduced in [1] for lowest order finite elements, in [28] for finite elements with uniform polynomial degrees and in [2] for hp-finite elements with non-uniform polynomial degrees, where regular triangular meshes without hanging nodes are assumed. These reconstructions are computed via the solution of local problems and enable the derivation of a posteriori error estimates for mixed methods. The latter can be applied to drive h- and hp-adaptivity, which is also discussed in these publications. Post-processing reconstruction, error control and adaptivity for mixed methods have been well studied and are documented in literature. We refer to [3, 9, 16–18, 20, 21, 29, 30, 37, 38] for further approaches on reconstruction techniques and/or a posteriori error estimation for mixed methods.

In this paper, we discuss and compare two different reconstruction techniques for dual mixed hp-finite element methods: First, the post-processed reconstruction as described in [2] with a small generalization to hp-finite elements on quadrilateral meshes with multilevel hanging nodes and, second, a reconstruction technique in which the reconstruction is not explicitly computed, but enables, nevertheless, the derivation of a posteriori error estimates [7]. Throughout this paper we call the first approach explicit and the second one implicit. We compare both approaches by discussing the a posteriori error estimates based on them because the derivation of error estimates may be their main field of application. For this purpose, we study the convergence of h- and hp-adaptive schemes driven by the error estimates and resulting efficiency indices in several numerical experiments, but we also discuss the efficiency indices (quotient of the exact and the estimated error) resulting from h-uniform as well as hp-geometric refinements. Since an explicit computation of the reconstruction is not needed in the implicit approach, the CPU-times required for the estimation are also compared. We observe that the efficiency indices are constant for both approaches. They are nearly 1 if error estimates based on the implicit reconstruction are applied. However, this advantage has to be considered with some care, since the estimates based on the implicit reconstruction contain an unknown efficiency constant, so that it is actually not fully computable; in contrast to the estimates with the explicit reconstruction, which give guaranteed upper bounds, even though their efficiency indices are significantly smaller than 1. It has still to be determined by what (computable) real number the efficiency constant of the implicit approach can be bounded. The explicit reconstruction can be computed efficiently as it only requires the solution of local problems on each mesh element. This is, in particular, true, if meshes without hanging nodes are used. If hp-finite elements on meshes with (multilevel) hanging nodes are applied, the proposed explicit reconstruction is more involved since it needs explicit knowledge of some connectivity information of the degrees of freedom for the reconstruction in H^1. In the numerical experiments, we observe that it requires noticeable computational time, whereas the implicit reconstruction does not cause an essentially more computational amount for its evaluation in comparison to the case without hanging nodes.

In this paper, we particularly consider the derivation of the implicit reconstruction in more detail, which is defined in three steps: first, the lowest order (vertex) contributions are separated, second, an element-wise lifting of the average of the

smoother remainder from the element boundary to the interior of the element is applied, and, third, a well-known low-order averaging is added. We show that the reconstruction error given as the difference of the discrete potential and its reconstruction can be estimated by the jumps of the discrete potential on the edges. Hence, we can interpret the reconstruction error as a measure for the non-conformity of the discrete potential with respect to the H^1-space. We note that the role of the jumps of the discrete potential is also underlined in [1] in a similar context.

The error estimates derived from the implicit reconstruction rely on measuring the error of the potential in a weighted, mesh-dependent H^1-norm (instead of the L^2-norm), which seems to be more natural in terms of broken Sobolev spaces. In particular, we obtain the same error estimates as in [9], but without any saturation assumption, due to the reconstruction in H^1. Thus, it does not face the problem of (possibly) unavailable regularity assumptions regarding to $H(\mathrm{div})$. We refer to [7] for more details with respect to the error estimation.

The paper is organized as follows: In Sect. 2.2 we introduce the dual mixed and dual mixed-hybrid finite element method for the Poisson problem. The explicit reconstruction and the error estimates based on this reconstruction as well as some numerical experiments are introduced in Sects. 2.3 and 2.4, respectively. The implicit reconstruction and the estimates of the reconstruction error are presented in Sect. 2.5. We discuss the a posteriori error estimates based on the implicit reconstruction and relating numerical experiments in Sect. 2.6. We compare both reconstruction approaches in more detail in Sect. 2.7, where we consider efficiency indices and CPU-times for h-uniform and hp-geometric refinements.

In this paper $A \lesssim B$ abbreviates $A \leq CB$ with a positive constant C which is independent from A, B and all other quantities of interest like mesh size and polynomial degree. The expression $A \lesssim_\epsilon B$ means that the constant C may dependent on an ϵ.

2.2 Mixed and Mixed-Hybrid Formulations of the Poisson Problem and Their Discretizations

In this section we consider the dual mixed formulation of the Poisson problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = u_D \quad \text{on } \Gamma_D, \qquad \partial_n u = u_N \quad \text{on } \Gamma_N \tag{2.1}$$

for a given $f \in L^2(\Omega)$, $u_D \in H^{1/2}(\Gamma_D)$, $u_N \in L^2(\Gamma_N)$, where ∂_n is the derivative in the direction of the outer normal n. Here, $\Omega \subset \mathbb{R}^2$ is a bounded, polygonal domain with boundary $\Gamma := \partial\Omega = \overline{\Gamma_D} \cup \overline{\Gamma_N}$ and $\Gamma_D \cap \Gamma_N = \emptyset$. With

$$H^1_{g;\Gamma_D}(\Omega) := \left\{ v \in H^1(\Omega) \mid v = g \text{ on } \Gamma_D \right\},$$

$$H_{g;\Gamma_N}(\mathrm{div}, \Omega) := \{ \tau \in H(\mathrm{div}, \Omega) \mid \tau \cdot n = g \text{ on } \Gamma_N \}$$

a dual mixed formulation of (2.1) consists in finding $(u, \sigma) \in L^2(\Omega) \times H_{u_N;\Gamma_N}(\mathrm{div}, \Omega)$ such that

$$(\sigma, \tau) + (u, \mathrm{div}\,\tau) = \langle u_D, \tau \cdot n \rangle_{\Gamma_D} \tag{2.2a}$$

$$(\mathrm{div}\,\sigma, v) = (-f, v) \tag{2.2b}$$

for all $v \in L^2(\Omega)$ and for all $\tau \in H_{0;\Gamma_N}(\mathrm{div}, \Omega)$. Here, $(\cdot, \cdot) = (\cdot, \cdot)_\Omega$ is the $L^2(\Omega)$-inner product and $\langle \cdot, \cdot \rangle_{\Gamma_D}$ is the duality pairing between $H^{1/2}(\partial\Omega)$ and

$$H^{-1/2}_{0;\Gamma_N}(\partial\Omega) := \{\mu \in H^{-1/2}(\partial\Omega) \mid \exists \tau \in H_{0;\Gamma_N}(\mathrm{div}, \Omega) : \ \tau \cdot n = \mu \text{ on } \partial\Omega\}.$$

Recall that $u_D \in H^{1/2}(\Gamma_D)$ implies that there exists an extension $\tilde{u}_D \in H^{1/2}(\partial\Omega)$ such that $\tilde{u}_D = u_D$ on Γ_D.

Conforming discretizations of the mixed formulation (2.2) with Raviart-Thomas elements are introduced, for instance, in [5, 12]. We refer to [4], which discusses their relationship to non-conforming methods. Let $\hat{K} = (-1, 1)^k$, $k \in \{1, 2\}$ and $F_K : \hat{K} \to K$ be the (bi-)linear transformation of $\hat{K}$ to an edge or a convex quadrilateral K with diameter h_K. We denote the Jacobian of F_K by D_{F_K} and its determinant by J_{F_K}. Here and in the following, a hatted variable denotes a variable in the reference setting. The finite element mesh of Ω is denoted by $\mathcal{T}_h$, consists of convex quadrilaterals and may have hanging nodes. In particular, multilevel hanging nodes are allowed, i.e. more than one hanging node per edge or hanging nodes on edges with hanging nodes as endpoints, see Fig. 2.1a. The mesh is assumed to be locally quasi-uniform and its boundary edges match with Γ_N and Γ_D. Furthermore, a polynomial degree distribution $(p_T)_{T \in \mathcal{T}_h}$ is defined on $\mathcal{T}_h$. We denote the set of inner constraining edges by $\mathcal{E}_h^i$ (i.e. the shortest edges, of which the start and end

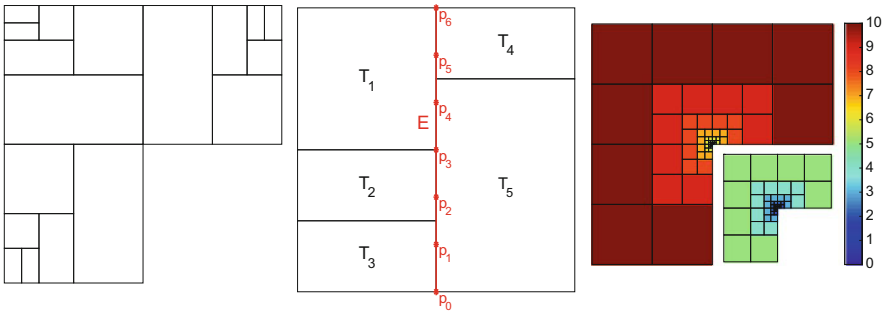


Fig. 2.1 **(a)** Initial mesh with multilevel hanging nodes of order 5. **(b)** Interpolation points $p_0, \ldots, p_6$ on the constraining edge E of degree $p_E = 6$. **(c)** Mesh resulting from hp-geometric refinements with zoom towards the reentrant corner and polynomial degrees marked by some colors

points are common nodes of the quadrilaterals left and right to these edges), the set of Dirichlet boundary edges by $\mathscr{E}_h^D$ and the set of Neumann boundary edges by $\mathscr{E}_h^N$. The polynomial degree on $\mathscr{E}_h := \mathscr{E}_h^i \cup \mathscr{E}_h^D \cup \mathscr{E}_h^N$ is given by the maximum rule and is denoted by $(p_e)_{e\in\mathscr{E}_h}$, i.e. p_e is the maximum of all p_T where T is adjacent to e. To specify local tensor product shape functions, we introduce the standard pullback transformation $\mathscr{F}_{F_K} : H^1(\hat{K}) \to H^1(K)$ as $\mathscr{F}_{F_K}(\hat{v}) = \hat{v} \circ F_K^{-1}$ for an edge or a convex quadrilateral K. The contravariant Piola transformation $\mathscr{P}_{F_T} : H(\mathrm{div}, \hat{T}) \to H(\mathrm{div}, T)$, $T \in \mathscr{T}_h$, is given by $\mathscr{P}_{F_T}(\hat{\tau}) = J_{F_T}^{-1} D_{F_T} \hat{\tau} \circ F_T^{-1}$ and preserves the normal component continuity in $H(\mathrm{div}, \Omega)$. Herewith, we define the conforming finite element spaces

$$RT_{hp} := \left\{\tau_{hp} \in H(\mathrm{div}, \Omega) \mid \forall\, T \in \mathscr{T}_h :\ \tau_{hp}|_T \in \mathscr{P}_{F_T}(P_{p_T+1,p_T} \times P_{p_T,p_T+1})\right\},$$

$$RT_{g;\Gamma_N;hp} := \left\{\tau_{hp} \in RT_{hp} \mid \tau_{hp} \cdot n = g \text{ on } \Gamma_N\right\},$$

$$V_{hp} := \left\{v \in L^2(\Omega) \mid \forall\, T \in \mathscr{T}_h :\ v|_T \in \mathscr{F}_{F_T}(P_{p_T,p_T})\right\}$$

with $P_{k,l} := \mathrm{span}\{x^i\, y^j \mid 0 \le i \le k,\ 0 \le j \le l\}$. With $u_{N,hp} \in W_{hp;N}$ as the piecewise orthogonal L^2 projection of u_N onto

$$W_{hp;N} := \left\{v_{hp} \in L^2(\Gamma_N) \mid \forall e \in \mathscr{E}_h^N :\ v_{hp}|_e \in \mathscr{F}_{F_e}(P_{p_e})\right\}$$

the discrete mixed formulation is to find $(u_{hp}, \sigma_{hp}) \in V_{hp} \times RT_{u_{N,hp};\Gamma_N;hp}$ such that

$$(\sigma_{hp}, \tau_{hp}) + (u_{hp}, \mathrm{div}\, \tau_{hp}) = (u_D, \tau_{hp} \cdot n)_{\Gamma_D} \tag{2.3a}$$

$$(\mathrm{div}\, \sigma_{hp}, v_{hp}) = (-f, v_{hp}) \tag{2.3b}$$

for all $v_{hp} \in V_{hp}$ and for all $\tau_{hp} \in RT_{0;\Gamma_N;hp}$.

In this $H(\mathrm{div}, \Omega)$-conforming Raviart-Thomas discretization the continuity in the normal component is strongly enforced. Using a Lagrange multiplier as an additional variable we can relax this continuity requirement in a weak sense, see, e.g., [5, 12]. The resulting discrete dual mixed-hybrid formulation is to find $(u_{hp}, \sigma_{hp}, \lambda_{hp}) \in V_{hp} \times \widetilde{RT}_{hp} \times W_{hp}$ such that

$$(\sigma_{hp}, \tau_{hp}) + \sum_{T\in\mathscr{T}_h} \int_T u_{hp}\ \mathrm{div}\, \tau_{hp}\, dx + \sum_{e\in\mathscr{E}_h^i\cup\mathscr{E}_h^N} \int_e [\![\tau_{hp} \cdot n]\!] \lambda_{hp}\, ds = (u_D, \tau_{hp} \cdot n)_{\Gamma_D} \tag{2.4a}$$

$$\sum_{T\in\mathscr{T}_h} \int_T \mathrm{div}\, \sigma_{hp}\, v_{hp}\, dx = -(f, v_{hp}) \tag{2.4b}$$

$$\sum_{e\in\mathscr{E}_h^i\cup\mathscr{E}_h^N} \int_e [\![\sigma_{hp} \cdot n]\!] \mu_{hp}\, ds = (u_N, \mu_{hp})_{\Gamma_N} \tag{2.4c}$$

for all $v_{hp} \in V_{hp}$, $\tau_{hp} \in \widetilde{RT}_{hp}$ and $\mu_{hp} \in W_{hp}$. Here,

$$\widetilde{RT}_{hp} := \left\{ \tau_{hp} \in \left[L^2(\Omega) \right]^2 \mid \forall T \in \mathscr{T}_h : \ \tau_{hp}|_T \in \mathscr{P}_{F_T}(P_{p_T+1,p_T} \times P_{p_T,p_T+1}) \right\},$$

$$W_{hp} := \left\{ \mu_{hp} \in L^2(\mathscr{E}_h^i \cup \mathscr{E}_h^N) \mid \forall e \in \mathscr{E}_h^i \cup \mathscr{E}_h^N : \ \mu_{hp}|_e \in \mathscr{F}_{F_e}(P_{p_e}) \right\},$$

and $[\![v]\!]$ denote the jump of v on the interior edges and v on the edges on Γ_N. It is well known, c.f. [5, 12], that the discrete problems (2.3) and (2.4) have unique solutions and are equivalent. Furthermore, λ_{hp} approximates $-u$ on $\mathscr{E}_h^i \cup \mathscr{E}_h^N$. Note that the relaxation of the continuity requirements in the dual-mixed-hybrid formulation essentially facilitates its implementation, in particular, if hanging nodes occur, see [8].

2.3 Explicit Reconstruction of the Potential

In order to find an appropriate reconstruction $\tilde{u}_{hp}$ in $H^1(\Omega)$, it seems to be straight forward to seek it in a higher-order polynomial space $V_{h,p+1} \cap H^1_{u_{D,hp},\Gamma_D}(\Omega)$ such that

$$(\nabla \tilde{u}_{hp}, \nabla v_{hp}) = -(\operatorname{div} \sigma_{hp}, v_{hp}) + \langle u_{N,hp}, v_{hp} \rangle_{\Gamma_N} \tag{2.5}$$

for all $v_{hp} \in V_{h,p+1} \cap H^1_{0,\Gamma_D}(\Omega)$. Here, $u_{D,hp}$ is a suitable approximation of u_D in

$$W_{h,p+1,D} := \left\{ v_{hp} \in H^{1/2}(\Gamma_D) \mid \forall e \in \mathscr{E}_h^D : \ v_{hp}|_e \in \mathscr{F}_{F_e}(P_{p_e+1}) \right\}.$$

Integration by parts in (2.5) yields

$$(\nabla \tilde{u}_{hp} - \sigma_{hp}, \nabla v_{hp}) = 0$$

for all $v_{hp} \in V_{h,p+1} \cap H^1_{u_{D,hp},\Gamma_D}(\Omega)$ and, therefore,

$$\begin{aligned} \|\nabla \tilde{u}_{hp} - \sigma_{hp}\|^2_{L^2(\Omega)} &= (\nabla \tilde{u}_{hp} - \sigma_{hp}, \nabla v_{hp} - \sigma_{hp}) \\ &\leq \|\nabla \tilde{u}_{hp} - \sigma_{hp}\|_{L^2(\Omega)} \|\nabla v_{hp} - \sigma_{hp}\|_{L^2(\Omega)} \end{aligned}$$

yielding the best-approximation property

$$\|\nabla \tilde{u}_{hp} - \sigma_{hp}\|_{L^2(\Omega)} = \min_{v_{hp} \in V_{h,p+1} \cap H^1_{u_{D,hp},\Gamma_D}(\Omega)} \|\nabla v_{hp} - \sigma_{hp}\|_{L^2(\Omega)}.$$

Clearly, the global reconstruction is quite expensive, since it has the same magnitude of degrees of freedom as the discretizations of the mixed or mixed-hybrid formulations. For this reason, we utilize the local reconstruction as proposed in [2]. This reconstruction is done in a three-step procedure consisting of a local Neumann problem, an averaging step and a local Dirichlet problem. In this section, we briefly describe a generalization of this reconstruction approach defined for non-uniform polynomial degrees to quadrilateral meshes with multilevel hanging nodes [32].

The first step of the reconstruction is to find $u^\circ_{hp,T} \in \mathscr{F}_{F_T}(P_{p_T+1,p_T+1})$ such that

$$\begin{aligned}(\nabla u^\circ_{hp,T}, \nabla v_{hp})_T &= (\sigma_{hp}, \nabla v_{hp})_T \\ (u^\circ_{hp,T}, 1)_T &= (u_{hp}, 1)_T\end{aligned} \tag{2.6}$$

for all $v_{hp} \in \mathscr{F}_{F_T}(P_{p_T+1,p_T+1})$ and $T \in \mathscr{T}_h$. Note that the mean value constraints enforce uniqueness of the solution. This first step of the reconstruction is a variant of the post-processing introduced in [36]. Since the solutions in (2.6) are locally discontinuous from element to element, the second averaging step smoothens them, where we allow for (multilevel) hanging nodes (in contrast to [2]). For this purpose, we evaluate the Neumann solutions $u^\circ_{hp,T}$ on each regular (non-hanging) node and on a certain number of interpolation points along each constraining edge. The number of interpolation points are determined by the minimum rule. Figure 2.1b shows an example of a constraining edge E of degree $p_E = 6$ with hanging nodes and a set of interpolation points $p_0, \ldots, p_6$ along E. Then, we apply a weighted averaging to these evaluations, where the weighting is based on the area of each element as in common ZZ-averaging techniques. The regular nodal degrees of freedom of the averaged approximation $\tilde{u}^\circ_{hp}$ are given by the averaged values associated to these degrees of freedom. The degrees of freedom along the constraining edges are given through the solution of the local interpolation problem for each edge as described above. One may replace the averaged data by the approximated Dirichlet data $u_{D,hp}$ if available. In order to specify the degrees of freedom for the local Dirichlet data of the following third reconstruction step from the averaged $\tilde{u}^\circ_{hp}$ we use connectivity matrices as, for instance, introduced in [13, 33]. These matrices manage the local and global numbering of degrees of freedom as well as varying polynomial degrees, edge orientation and contain the constrained coefficients for (multilevel) hanging nodes (as shown in Fig. 2.1a).

The third and final step of the explicit reconstruction is to solve local Dirichlet problems $\tilde{u}_{hp,T} \in \mathscr{F}_{F_T}(P_{p_T+1,p_T+1})$ where the Dirichlet data of each problem results from the previous averaging step: Find $\tilde{u}_{hp} \in V_{h,p+1} \cap H^1(\Omega)$ such that

$$(\nabla \tilde{u}_{hp,T}, \nabla v_{hp})_T = (\sigma_{hp}, \nabla v_{hp})_T \tag{2.7}$$

$$\tilde{u}_{hp,T} = \tilde{u}^\circ_{hp} \text{ on } \partial T \tag{2.8}$$

for all $v_{hp} \in \left\{ v_{hp} \in \mathscr{F}_{F_T}(P_{p_T+1,p_T+1}) \mid v_{hp} = 0 \text{ on } \partial T \right\}$ and $T \in \mathscr{T}_h$. Note that $\tilde{u}_{hp}$ is continuous due to the continuity of the averaging approximation $\tilde{u}^\circ_{hp}$.

2.4 A Posteriori Error Estimates Based on the Explicit Reconstruction

A posteriori error estimates based on explicit reconstructions are well-known in literature (see the references in the introduction). For hp-finite elements a fully computable upper bound for $\|\sigma - \sigma_{hp}\|_{L^2(\Omega)}$ is introduced in [2], where triangular meshes without hanging nodes are assumed. This can easily be extended to quadrilateral meshes (cf. [32]) so that it holds

$$\|\sigma - \sigma_{hp}\|^2_{L^2(\Omega)} \leq \|\sigma_{hp} - \nabla v_{hp}\|^2_{L^2(\Omega)} + \gamma^2 \tag{2.9}$$

for an arbitrary $v_{hp} \in H^1_{u_D;\Gamma_D}(\Omega)$, where $\gamma^2 := \sum_{T\in\mathscr{T}_h} \gamma_T^2$ and

$$\gamma_T := \frac{h_T}{\pi}\|f - \Pi(f)\|_{0,T} + \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^N} C_{e,T}\|u_N - u_{N,hp}\|_{0,e},$$

$$C_{e,T} := \frac{2h_T}{\pi l_e}\left(\frac{h_T}{\pi} + L_e\right).$$

Here, x_e is one of the vertices of T opposite to $e \in \mathscr{E}_T$, $L_e := \max_{x\in e}|x - x_e|$, $l_e := \min_{x\in e}|x - x_e|$ and $\mathscr{E}_T$ denotes the edges of T. Furthermore, $\Pi(f)$ is the L^2-projection of f onto V_{hp}. Note that $v_{hp} \in H^1_{u_D;\Gamma_D}(\Omega)$ cannot be fulfilled for non-polynomial Dirichlet data u_D, which may lead to some additional error term.

In order to evaluate the explicit reconstruction approach of Sect. 2.3 and the resulting a posteriori error estimate, we consider some numerical experiments based on the Poisson equation on the L-shape domain $\Omega := (-1, 1)^2\backslash[0, 1) \times (-1, 0]$, $\Gamma_D := (0, 1) \times \{0\} \cup (0, -1) \times \{0\}$, with the solution $u := r^{2/3}\sin(2\theta/3)$ given in the polar coordinates (r, θ), see [32]. Due to the singularity at the origin, we have $u \in H^{5/3-\epsilon}(\Omega)$, $\epsilon > 0$ arbitrarily small. Thus, the order of convergence of h-uniform refinements is bounded by the suboptimal algebraic order $\mathscr{O}(N^{-1/3})$, where $N := \dim V_{hp} + \dim RT_{u_{N,hp};\Gamma_N;hp}$, see Fig. 2.2a. Here, we set $e_\sigma := \|\sigma - \sigma_{hp}\|_{L^2(\Omega)}$. Indeed, the optimal order of convergence is $\mathscr{O}(N^{-p/2})$ if u is sufficiently regular (which is not the case here). Using (isotropic) h-adaptive refinements with Dörfler-marking we are able to recover the optimal algebraic order $\mathscr{O}(N^{-1/2})$ for $p = 1$, $\mathscr{O}(N^{-1})$ for $p = 2$ and $\mathscr{O}(N^{-3/2})$ for $p = 3$, whereas the use of hp-refinements driven by the a posteriori error estimates based on the explicit reconstruction yields optimal exponential order of convergence $\mathscr{O}(\exp(-b_e N^{1/3}))$ with some slope $b_e > 0$, see Fig. 2.3a. The p-refinement strategy is based on a Legendre polynomial expansion, see, e.g. [24]. We observe the typical refinement patterns with h-refinements and low polynomial degrees near to the reentrant corner ($p_T = 1$ at the reentrant corner) and increasing polynomial degrees away from the reentrant corner, see Fig. 2.3b. Note that the polynomial degree at the reentrant corner can not be seen in the zoom of Fig. 2.3b. Furthermore, in Fig. 2.3a the

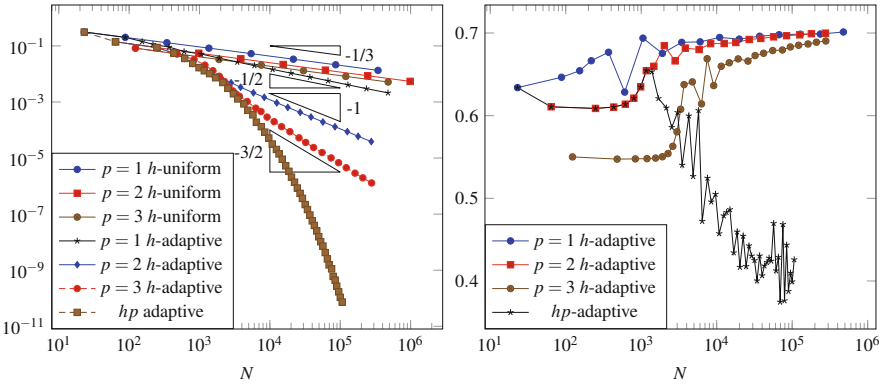


Fig. 2.2 **(a)** e_σ on y-axis in logarithmic scaling, **(b)** efficiency indices e_σ/η_e

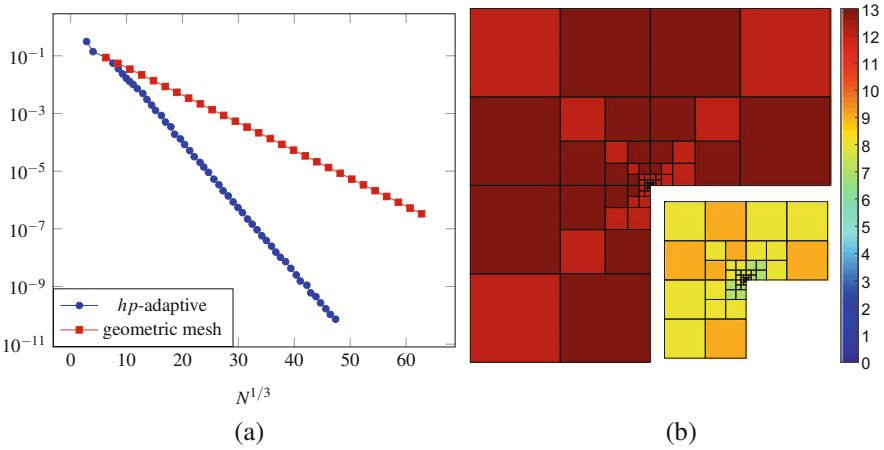


Fig. 2.3 **(a)** e_σ on y-axis in logarithmic scaling, **(b)** hp-adaptive mesh with zoom towards the reentrant corner and polynomial degrees marked by some colors

convergence resulting from a common variant of hp mesh refinements is depicted, which is called hp-geometric refinements below. In this variant the polynomial degree is increased in layers away from the reentrant corner and h-refinements are performed in those mesh elements which contain the reentrant corner, see, e.g., [34] and Fig. 2.1c. As shown in Fig. 2.3a, we also get an optimal exponential order of convergence $\mathcal{O}(\exp(-b_g N^{1/3}))$, but with some slope $b_g < b_e$, which shows that the hp-refinements driven by the a posteriori error estimates are more effective than these hp-geometric refinements. In Table 2.1 the constants of the exponential convergence order $\mathcal{O}(\exp(-bN^{1/3}))$ are tabulated in more detail, where the hidden constants in the $\mathcal{O}$-notation are denoted by C_e, C_i, C_g and $\tilde{C}_g$. The subindices e, i

Table 2.1 Approximatively computed constants of the order of convergence $\mathcal{O}(\exp(-bN^{1/3}))$ with hidden constant C

C_e	b_e	C_i	b_i	C_g	b_g	$\tilde{C}_g$	$\tilde{b}_g$
3.85419	0.52700	6.77538	0.51400	0.80760	0.22133	1.00717	0.22138
3.00665	0.51859	5.83012	0.50763	0.81094	0.22146	1.01036	0.22148
3.66507	0.52332	5.66514	0.50748	0.81421	0.22158	1.01410	0.22159
3.90773	0.52694	6.15360	0.50962	0.81612	0.22164	1.01606	0.22164
3.94131	0.52661	5.85910	0.50767	0.81799	0.22170	1.01828	0.22170
2.69936	0.51325	5.18236	0.50448	0.82018	0.22177	1.02093	0.22177
3.18891	0.51880	5.01065	0.50423	0.82135	0.22180	1.02236	0.22180
2.53258	0.51378	5.22435	0.50532	0.82298	0.22185	1.02431	0.22184
3.94542	0.52747	4.07533	0.49848	0.82378	0.22188	1.02534	0.22187
1.08734	0.50382	3.90671	0.49833	0.82587	0.22193	1.02792	0.22193
4.07108	0.52417	5.21802	0.50476	0.82575	0.22192	1.02773	0.22192

The subindices e, i and g denote the use of the explicit or implicit reconstruction in the a posteriori error estimates or the use of hp-geometric refinements, respectively

and g indicate the use of the explicit or implicit reconstruction in the a posteriori error estimates or the use of the hp-geometric refinements, respectively. Note that the constants C_i, b_i, $\tilde{C}_g$ and $\tilde{b}_g$ included in Table 2.1 are described in Sect. 2.6. The constants are approximately computed by

$$b_k := -\log(e_k/e_{k+l})/(N_k^{1/3} - N_{k+l}^{1/3}), \quad C_k := e_k \exp(b_k N_k^{1/3})$$

for the errors e_k (which is e_σ for the explicit approach), number of degrees of freedom N_k and some appropriately chosen k and l indicating the data sets in Figs. 2.3a and 2.5a. Here, we take $k = 20, \ldots, 30$ and $l := 15$ for the explicit and the implicit approach, and $k = 5, \ldots, 15$ and $l := 10$ for the hp-geometric refinements. In Fig. 2.2b, some efficiency indices (defined as the quotient e_σ/η_e of the exact error e_σ and the estimated error η_e) are presented resulting from the adaptive refinements. As expected, all indices are smaller than 1, which confirms that (2.9) gives a guaranteed upper bound of the error. We observe that the efficiency indices in the case of uniform polynomial degrees converge to a constant value indicating efficiency and reliability as well as a certain independency of the polynomial degree. This seems not to be the case if hp-adaptivity is applied. However, using more than 10^4 degrees of freedom we have efficiency indices which oscillate in the small range between 0.37 and 0.47. These oscillations may go back to the hp-adaptive scheme resulting in (multilevel) hanging nodes and non-uniform polynomial degrees. If hp-geometric refinements (with at most one hanging node per edge and a very structured polynomial degree distribution) are applied, we get constant efficiency indices again, which indicate a certain p-robustness (at least in this case), see Fig. 2.9a. We note that p-robustness of a posteriori error estimation is an important aspect which has been well analyzed in several recent contributions [10, 14, 15, 20, 21].

2.5 Implicit Reconstruction of the Potential

In this section we present a reconstruction $\tilde{u}_{hp}$ in $H^1(\Omega)$, which is not explicitly given, but allows for a computable estimation of $u_{hp} - \tilde{u}_{hp}$ in terms of the jumps of u_{hp} on the edges of $\mathscr{T}_h$ (see Theorems 2.1 and 2.2). In fact, we can interpret $u_{hp} - \tilde{u}_{hp}$ as the non-conformity of $u_{hp} \notin H^1(\Omega)$. The definition of $\tilde{u}_{hp}$ is based on an element-wise lifting of the average of u_{hp} restricted to the element edges. Here, the lifting is the solution of a Poisson equation with some adapted Dirichlet boundary conditions which ensure the continuity of the reconstruction. Although, this lifting is unknown, we can use it to derive a posteriori error estimates. In Sect. 2.6 we present some a posteriori error estimates based on this reconstruction, where an explicit evaluation of the reconstruction is not needed (in contrast to the estimates of Sect. 2.4).

We construct $\tilde{u}_{hp}$ in three steps: First, we separate the lowest order (vertex) contributions and, second, apply an element-wise lifting of the average of the smoother remainder from the element boundary to the interior of the element. Third, we add a well-known low-order averaging. For simplicity, we assume that $u_{D,hp} \in V_{hp}|_{\Gamma_D} \cap C^0(\Gamma_D)$ and assume that $\mathscr{T}_h$ does not contain hanging nodes. We refer to Remark 2.3 for extensions with respect to meshes with hanging nodes.

Step One Let $u_h^n \in V_{h,1} := \left\{ v \in L^2(\Omega) \mid \forall T \in \mathscr{T}_h : v|_T \in \mathscr{F}_{F_T}(P_{1,1}) \right\}$ be defined element by element such that $u_h^n|_T$ is the (bi)-linear nodal interpolant of $u_{hp}|_T$. Furthermore, define $w_{hp} := u_{hp} - u_h^n$ and let $\{\{v\}\} := (v|_{T^+ \to e} + v|_{T^- \to e})/2$ be the average of v on the edge $e \in \mathscr{E}_h^i$ where T^+ and T^- are the adjacent elements of e. Here, it is not important which element $T \in \mathscr{T}_h$ is T^+ or T^-, as it only effects the sign of the jump term in the non-conformity estimation (see below). For $e \in \mathscr{E}_h^D \cup \mathscr{E}_h^N$, we simply define $\{\{v\}\} := v|_e$ and set

$$g_{hp}|_e := \begin{cases} \{\{w_{hp}\}\}, & e \in \mathscr{E}_h^i \cup \mathscr{E}_h^N \\ u_{D,hp} - I_{h,1}(u_{D,hp}), & e \in \mathscr{E}_h^D, \end{cases} \tag{2.10}$$

where $I_{h,1}$ is the globally continuous, piecewise linear interpolation operator defined in the Dirichlet boundary vertices. Since w_{hp} is zero in each vertex, this is also true for g_{hp}, and, hence, $(g_{hp} - w_{hp})|_{\partial T} \in H^{1/2}(\partial T)$.

Step Two On each element $T \in \mathscr{T}_h$, let $w^*|_T \in H^1(T)$ be the unique solution of

$$-\Delta w^* = -\Delta w_{hp} \text{ in } T, \quad w^* = g_{hp} \text{ on } \partial T. \tag{2.11}$$

By construction of the Dirichlet data it holds $w^* \in H^1_{u_{D,hp} - I_{h,1}(u_{D,hp}); \Gamma_D}(\Omega)$. It is well known, that the dependence of the solution on the Dirichlet data is Lipschitz-continuous. Hence, since ∂T is a closed curve, we obtain from the von Petersdorff

inequality, c.f., e.g., [23, Lem. 1]:

$$|w^* - w_{hp}|^2_{H^1(T)} \lesssim \|g_{hp} - w_{hp}\|^2_{H^{1/2}(\partial T)} \lesssim \sum_{e\in\mathscr{E}_T} \|g_{hp} - w_{hp}\|^2_{\widetilde{H}^{1/2}(e)}. \tag{2.12}$$

Here, $\mathscr{E}_T := \{e \in \mathscr{E}_h \mid e \subset \partial T\}$ and

$$\widetilde{H}^{1/2}(e) := \{v = v'|_e \mid \exists\, v' \in H^{1/2}(\partial T),\ \text{supp}\ v' \subset e\}$$

with the norm $\|u\|_{\widetilde{H}^{1/2}(e)} := \|u_0\|_{H^{1/2}(\partial T)}$ where u_0 is the extension of u onto ∂T by zero.

Step Three In [27, Thm. 2.2 and Thm. 2.3] it is proven that there exists a $\tilde{u}^n_h \in V_{h,1} \cap H^1(\Omega)$ with $\tilde{u}^n_h|_{\Gamma_D} = I_{h,1}(u_{D,hp})$ such that

$$\sum_{T\in\mathscr{T}_h} |\tilde{u}^n_h - u^n_h|^2_{H^1(T)} \lesssim \sum_{e\in\mathscr{E}^i_h} \frac{1}{h_e} \left\|[\![u^n_h]\!]\right\|^2_{L^2(e)} + \sum_{e\in\mathscr{E}^D_h} \frac{1}{h_e} \|I_{h,1}(u_{D,hp}) - u^n_h\|^2_{L^2(e)} \tag{2.13}$$

provided that $u_{D,hp} = 0$ and $\Gamma_N = \emptyset$. For a non-vanishing $u_{D,hp}$ and $\Gamma_N \neq \emptyset$ this assertion remains valid since the proof of [27, Thm. 2.2] can be verbatim transferred with the small modification that in [27, Eq. (2.16)] the expansion coefficients of $I_{h,1}(u_{D,hp})$ need to be subtracted from the expansion coefficients of u^n_h which are associated to the Dirichlet nodes. To this end, we have to assume that h is sufficiently small in order to prevent an interior edge having both endpoints on Γ_N. The function $\tilde{u}^n_h$ is uniquely given by its nodal values,

$$\tilde{u}^n_h(q) := \begin{cases} |\mathscr{T}_h(q)|^{-1} \sum_{T\in\mathscr{T}_h(q)} u^n_h|_T(q), & \text{if } q \in \mathscr{N}_h^i \cup \mathscr{N}_h^N \\ u_{D,hp}(q), & \text{if } q \in \mathscr{N}_h^D \end{cases}$$

where $\mathscr{T}_h(q)$ denotes the set of elements in $\mathscr{T}_h$ which contain the vertex q. The number $|\mathscr{T}_h(q)|$ is the cardinality of $\mathscr{T}_h(q)$. Furthermore, $\mathscr{N}_h^D$, $\mathscr{N}_h^N$, $\mathscr{N}_h^i$ are the vertices of $\mathscr{T}_h$ on Γ_D, Γ_N and the interior of Ω. Finally, we set

$$\tilde{u}_{hp} := \tilde{u}^n_h + w^*, \tag{2.14}$$

which is in $H^1(\Omega)$ by construction.

Lemma 2.1 *There holds*

$$\left\|[\![u_h^n]\!]\right\|_{L^2(e)} \lesssim p_e \left\|[\![u_{hp}]\!]\right\|_{L^2(e)},$$
$$\|I_{h,1}(u_{D,hp}) - u_h^n\|_{L^2(e)} \lesssim p_e \left\|u_{D,hp} - u_{hp}\right\|_{L^2(e)}.$$

Proof Let q_1, q_2 be the two endpoints of the edge e and $v(s) := ([\![u_h^n]\!]((1-s)q_1 + sq_2))^2$ with $s \in \mathbb{R}$. Note that $v'' \geq 0$. Applying the trapezoidal rule with some $\xi \in [0,1]$ and an inverse polynomial estimate we have

$$\begin{aligned}
\left\|[\![u_h^n]\!]\right\|_{L^2(e)}^2 &= \frac{h_e}{2}\left([\![u_h^n]\!](q_1)^2 + [\![u_h^n]\!](q_2)^2\right) - \frac{h_e^3}{12}v''(\xi) \\
&\leq \frac{h_e}{2}\left([\![u_{hp}]\!](q_1)^2 + [\![u_{hp}]\!](q_2)^2\right) \\
&\leq h_e \|[\![u_{hp}]\!]\|_{L^\infty(e)}^2 \\
&\lesssim p_e^2 \left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2.
\end{aligned}$$

The second assertion follows analogously. □

Theorem 2.1 *There holds*

$$\sum_{T\in\mathscr{T}_h} |u_{hp} - \tilde{u}_{hp}|_{H^1(T)}^2 \lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e}\left\|u_{D,hp} - u_{hp}\right\|_{L^2(e)}^2. \tag{2.15}$$

Proof Using (2.12), the triangle inequality, Young's inequality and an inverse polynomial estimate (with $\widetilde{H}^{1/2}(e)$ as an interpolation space between $L^2(e)$ and $H_0^1(e)$) and applying Lemma 2.1 we obtain

$$\begin{aligned}
&|w^* - w_{hp}|_{H^1(T)}^2 \\
&\quad\lesssim \sum_{e\in\mathscr{E}_T} \|g_{hp} - w_{hp}\|_{\widetilde{H}^{1/2}(e)}^2 \\
&\quad= \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^i} \frac{1}{4}\left\|[\![u_{hp} - u_h^n]\!]\right\|_{\widetilde{H}^{1/2}(e)}^2 + \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^D} \|u_{D,hp} - u_{hp} - I_{h,1}(u_{D,hp}) + u_h^n\|_{\widetilde{H}^{1/2}(e)}^2 \\
&\quad\leq \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^i} \frac{1}{2}\left\|[\![u_{hp}]\!]\right\|_{\widetilde{H}^{1/2}(e)}^2 + \frac{1}{2}\left\|[\![u_h^n]\!]\right\|_{\widetilde{H}^{1/2}(e)}^2 \\
&\qquad+ \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^D} 2\|u_{D,hp} - u_{hp}\|_{\widetilde{H}^{1/2}(e)}^2 + 2\|I_{h,1}(u_{D,hp}) - u_h^n\|_{\widetilde{H}^{1/2}(e)}^2
\end{aligned}$$

$$\lesssim \sum_{e\in\partial T\cap\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \frac{1}{h_e}\left\|[\![u_h^n]\!]\right\|_{L^2(e)}^2$$

$$+ \sum_{e\in\partial T\cap\mathscr{E}_h^D} \frac{p_e^2}{h_e}\|u_{D,hp}-u_{hp}\|_{L^2(e)}^2 + \frac{1}{h_e}\|I_{h,1}(u_{D,hp})-u_h^n\|_{L^2(e)}^2$$

$$\lesssim \sum_{e\in\partial T\cap\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\partial T\cap\mathscr{E}_h^D} \frac{p_e^2}{h_e}\|u_{D,hp}-u_{hp}\|_{L^2(e)}^2. \tag{2.16}$$

Summing over all elements yields

$$\sum_{T\in\mathscr{T}_h} |w^*-w_{hp}|_{H^1(T)}^2 \lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e}\|u_{D,hp}-u_{hp}\|_{L^2(e)}^2. \tag{2.17}$$

From (2.13) and Lemma 2.1 we conclude

$$\sum_{T\in\mathscr{T}_h} |\tilde{u}_h^n-u_h^n|_{H^1(T)}^2 \lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e}\|u_{D,hp}-u_{hp}\|_{L^2(e)}^2.$$

Exploiting $\tilde{u}_{hp}|_{\Gamma_D} = I_{h,1}(u_{D,hp}) + u_{D,hp} - I_{h,1}(u_{D,hp})$ we eventually have

$$\begin{aligned}\sum_{T\in\mathscr{T}_h} |\tilde{u}_{hp}-u_{hp}|_{H^1(T)}^2 &= \sum_{T\in\mathscr{T}_h} |\tilde{u}_h^n + w^* - u_{hp} - u_h^n + u_h^n|_{H^1(T)}^2 \\ &\le 2\sum_{T\in\mathscr{T}_h} |w^*-w_{hp}|_{H^1(T)}^2 + |\tilde{u}_h^n-u_h^n|_{H^1(T)}^2 \\ &\lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e}\|u_{D,hp}-u_{hp}\|_{L^2(e)}^2.\end{aligned}$$

□

Theorem 2.2 *There holds*

$$\sum_{T\in\mathscr{T}_h} \|u_{hp}-\tilde{u}_{hp}\|_{H^1(T)}^2 \lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e}\left\|[\![u_{hp}]\!]\right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e}\left\|u_{D,hp}-u_{hp}\right\|_{L^2(e)}^2.$$

Proof By the Poincaré-Friedrichs inequality, e.g. [11, p. 135], it holds

$$\begin{aligned}\|u_{hp}-\tilde{u}_{hp}\|^2_{H^1(T)} &\lesssim \left|\int_{\partial T} u_{hp}-\tilde{u}_{hp}\,ds\right|^2 + |u_{hp}-\tilde{u}_{hp}|^2_{H^1(T)}\\ &= \left|\int_{\partial T} w^*-u_{hp}+u_h^n+\tilde{u}_h^n-u_h^n\,ds\right|^2 + |u_{hp}-\tilde{u}_{hp}|^2_{H^1(T)}\\ &\lesssim \left|\int_{\partial T} w^*-u_{hp}+u_h^n\,ds\right|^2 + \left|\int_{\partial T}\tilde{u}_h^n-u_h^n\,ds\right|^2 + |u_{hp}-\tilde{u}_{hp}|^2_{H^1(T)}.\end{aligned} \tag{2.18}$$

Applying Cauchy-Schwarz inequality and Lemma 2.1, we obtain

$$\begin{aligned}&\left|\int_{\partial T} w^*-u_{hp}+u_h^n\,ds\right|\\ &\quad= \left|\sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^i}\int_e [\![u_{hp}]\!]-[\![u_h^n]\!]\,ds+\sum_{e\in\partial T\cap\mathscr{E}_h^D}\int_e u_{D,hp}-u_{hp}-I_{h,1}(u_{D,hp})+u_h^n\,ds\right|\\ &\quad\le \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^i} h_e^{1/2}\left(\|[\![u_{hp}]\!]\|_{L^2(e)}+\|[\![u_h^n]\!]\|_{L^2(e)}\right)\\ &\qquad+\sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^D} h_e^{1/2}\left(\|u_{D,hp}-u_{hp}\|_{L^2(e)}+\|I_{h,1}(u_{D,hp})-u_h^n\|_{L^2(e)}\right)\\ &\quad\lesssim \sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^i} p_e\|[\![u_{hp}]\!]\|_{L^2(e)}+\sum_{e\in\mathscr{E}_T\cap\mathscr{E}_h^D} p_e\|u_{D,hp}-u_{hp}\|_{L^2(e)},\end{aligned}$$

where we exploit that h_e is trivially bounded from above by the diameter of Ω. Let $\mathscr{N}_e$ be the set of the two end points of the edge e. Applying the trapezoidal rule and an inverse polynomial estimate we get

$$\begin{aligned}&\left|\int_{\partial T}\tilde{u}_h^n-u_h^n|_T\,ds\right|\\ &\quad\le \sum_{e\in\mathscr{E}_T}\frac{h_e}{2}\sum_{q\in\mathscr{N}_e}|(\tilde{u}_h^n-u_h^n|_T)(q)|\\ &\quad= \sum_{e\in\mathscr{E}_T}\frac{h_e}{2}\sum_{q\in\mathscr{N}_e}\frac{1}{|\mathscr{T}_h(q)|}\left(\sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^i}|[\![u_{hp}]\!]|_{\tilde{e}}(q)|+\sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^D}|(u_{D,hp}-u_{hp})|_{\tilde{e}}(q)|\right)\end{aligned}$$

$$\leq \sum_{e\in\mathscr{E}_T} \frac{h_e}{2} \sum_{q\in\mathscr{N}_e} \frac{1}{|\mathscr{T}_h(q)|} \left(\sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^i} \|[\![u_{hp}]\!]\|_{L^\infty(\tilde{e})} + \sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^D} \|u_{D,hp} - u_{hp}\|_{L^\infty(\tilde{e})} \right)$$

$$\lesssim \sum_{e\in\mathscr{E}_T} \sum_{q\in\mathscr{N}_e} \left(\sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^i} \frac{p_{\tilde{e}}}{h_{\tilde{e}}^{1/2}} \|[\![u_{hp}]\!]\|_{L^2(\tilde{e})} + \sum_{\tilde{e}\in\mathscr{E}_q\cap\mathscr{E}_h^D} \frac{p_{\tilde{e}}}{h_{\tilde{e}}^{1/2}} \|u_{D,hp} - u_{hp}\|_{L^2(\tilde{e})} \right).$$

Using Young's inequality we have

$$\sum_{T\in\mathscr{T}_h} \left| \int_{\partial T} u_{hp} - \tilde{u}_{hp}\, ds \right|^2 \lesssim \sum_{e\in\mathscr{E}_h^i} \frac{p_e^2}{h_e} \left\| [\![u_{hp}]\!] \right\|_{L^2(e)}^2 + \sum_{e\in\mathscr{E}_h^D} \frac{p_e^2}{h_e} \left\| u_{D,hp} - u_{hp} \right\|_{L^2(e)}^2.$$

Inserting this estimate into (2.18) and applying Theorem 2.1 yield the assertion. □

The existence of a reconstruction with the same estimate as stated in Theorem 2.1 can be found in [25, Prop. 5.2] for $u_{D,hp} = 0$ and $\Gamma_N = \emptyset$. The advantage of this approach consists in the use of (2.11) to approximate the higher order contributions of u_{hp}. This seems to be a general technique: As long as a low order approximation operator (such as given in (2.13)) is known, it is possible to carry over this approach for the reconstruction in higher Sobolev spaces (for instance, $H^2(\Omega)$ as also done in [6]).

Remark 2.1 One may ask for an extension of the reconstruction to three dimensions. However, this may be difficult since an appropriate definition of g_{hp} in (2.10) on edges and faces to define some proper Dirichlet conditions in (2.11) is not straight forward.

Remark 2.2 We emphasize that the reconstruction of the potential as introduced above enables the use of variable polynomial degrees. The application of a classical averaging is not obvious for such polynomial degree distributions. For instance, the averaging proposed in [27] is defined for a fixed uniform polynomial degree. Moreover, the dependency of the polynomial degree on the constant of its error estimation is not explicitly given. We refer to [19] for a further potential reconstruction based on an averaging, which also includes variable polynomial degrees.

Remark 2.3 The reconstruction is also valid for meshes with hanging nodes. In this case one may simply use a virtual mesh $\tilde{\mathscr{T}}_h$ resulting from additional local refinements of all mesh elements responsible for hanging nodes such that all hanging nodes of $\mathscr{T}_h$ are resolved by these refinements. Note that new hanging nodes on former regular edges or in the interior of former elements can occur, but in these points the discrete u_{hp} is continuous in the tangential direction of the new edges by construction. Thus the old hanging nodes are resolved and the new ones do not effect u_{hp}. The virtual mesh defines new sets $\tilde{\mathscr{E}}_h^i$, $\tilde{\mathscr{E}}_h^D$ and $\tilde{\mathscr{E}}_h^N$ of interior edges and of edges on Γ_D and Γ_N. The reconstruction is modified in the following way: The functions u_h^n, w_{hp} and g_{hp} are determined with respect to $\tilde{\mathscr{T}}_h$, $\tilde{\mathscr{E}}_h^i$, $\tilde{\mathscr{E}}_h^D$ and

$\tilde{\mathscr{E}}_h^N$, respectively, whereas w^* in (2.11) is still defined on the original mesh $\mathscr{T}_h$. Applying the average operator introduced in [27, Thm. 2.3] (for hanging nodes) to u_h^n, we again define the reconstruction as in (2.14). The estimations of $|w^* - w_{hp}|_{1,T}^2$ in (2.12) and (2.16) and of the jump $\|[\![u_h^n]\!]\|_{L^2(e)}$ in Lemma 2.1 can be done in the same way as before, but with the edges of the virtual mesh $\tilde{\mathscr{T}}_h$. Summing over the original mesh $\mathscr{T}_h$ in (2.17) gives an estimation with respect to the edges in $\tilde{\mathscr{E}}_h^i$ and $\tilde{\mathscr{E}}_h^D$. Under the additional assumption that the number of hanging nodes per edge is limited the length of theses edges can uniformly be estimated by the length of the edges of $\mathscr{E}_h^i$ and $\mathscr{E}_h^D$, which gives the same estimation as in (2.17). This and the estimation for the average operator finally yields exactly the same error estimation as in Theorem 2.1.

The assumption that the number of hanging nodes has to be limited seems to be unavoidable in the argumentation above. We refer to [20, 22] for the derivation of p-robust a posteriori error estimates based on a flux reconstruction, where a completely arbitrary number of levels of hanging nodes is allowed.

2.6 A Posteriori Error Estimates Based on the Implicit Reconstruction

In the following we present an a posteriori error estimate based on the implicit reconstruction as introduced in Sect. 2.5. Instead of estimating the error of the flux in the L^2-norm (as in Sect. 2.4), we consider the error in the mesh dependent norm

$$|v|_{1,hp}^2 := \sum_{T\in\mathscr{T}_h} |v|_{H^1(T)}^2 + \sum_{e\in\mathscr{E}_h^i\cup\mathscr{E}_h^D} \frac{p_e^2}{h_e} \left\|[\![v]\!]\right\|_{L^2(e)}^2 \tag{2.19}$$

which is well-known in the context of hp-discontinuous Galerkin methods, see, e.g., [25, 26, 31]. Obviously, there holds $|v|_{1,hp} = |v|_{H^1(\Omega)}$ for all $v \in H^1_{0;\Gamma_D}(\Omega)$. The norm (2.19) is well-defined even for $v \in \prod_{T\in\mathscr{T}_h} H^1(T) \supseteq H^1(\Omega)$. We refer to [9, p. 2434] for some a priori error estimates with respect to this norm in the case of the h-version with $u_D = 0$ and $\Gamma_N = \emptyset$.

The residual based a posteriori error estimate for the discrete mixed variational equation (2.3) is stated in (2.21) and (2.23). It consists of the following local error contributions

$$\eta_{T,1}^2 := \frac{h_T^2}{p_T^2} \|f_{hp} + \operatorname{div}\sigma_{hp}\|_{L^2(T)}^2, \quad T \in \mathscr{T}_h, \tag{2.20a}$$

$$\eta_{T,2}^2 := \|\sigma_{hp} - \nabla u_{hp}\|_{L^2(T)}^2, \quad T \in \mathscr{T}_h, \tag{2.20b}$$

$$\eta_{e,1}^2 := \frac{p_e^2}{h_e} \left\| [\![u_{hp}]\!] \right\|_{L^2(e)}^2, \qquad e \in \mathscr{E}_h^i, \tag{2.20c}$$

$$\eta_{e,2}^2 := \frac{p_e^2}{h_e} \left\| u_{D,hp} - u_{hp} \right\|_{L^2(e)}^2, \qquad e \in \mathscr{E}_h^D, \tag{2.20d}$$

where $f_{hp} \in V_{hp}$ is a (suitable) piecewise polynomial approximation of f. The error contributions may be interpreted as internal residuals $\eta_{T,1}$, the error with respect to the flux $\eta_{T,2}$, the lack of $H^1(\Omega)$-conformity $\eta_{e,1}$ and the violation of the (weak) Dirichlet data $\eta_{e,2}$. Volume, Dirichlet and Neumann data oscillation terms are defined as

$$\operatorname{osc}_{\mathscr{T}_h}^2 := \sum_{T \in \mathscr{T}_h} \frac{h_T^2}{p_T^2} \|f - f_{hp}\|_{L^2(T)}^2, \quad \operatorname{osc}_D^2 := \|u_D - u_{D,hp}\|_{H^{1/2}(\Gamma_D)}^2,$$

$$\operatorname{osc}_{\mathscr{E}_h^D}^2 := \sum_{e \in \mathscr{E}_h^D} \frac{p_e^2}{h_e} \|u_D - u_{D,hp}\|_{L^2(e)}^2, \quad \operatorname{osc}_N^2 := \|u_N - u_{N,hp}\|_{H^{-1/2}(\Gamma_N)}^2.$$

The $H^{-1/2}(\Gamma_N)$ and $H^{1/2}(\Gamma_D)$-norms may be realized via the single layer potential and hypersingular operator (in conjunction with the identity operator) for the Poisson equation known from boundary element methods, see, e.g., [35]. Alternatively, assuming $u_D \in H^1(\Gamma_D)$, $u_N \in L^2(\Gamma_N)$ and $u_{N,hp}$ as the $L^2(\Gamma_N)$-projection of u_N onto $V_{hp}|_{\Gamma_N}$, we have

$$\operatorname{osc}_D^2 \lesssim \|u_D - u_{D,hp}\|_{H^1(\Gamma_D)} \|u_D - u_{D,hp}\|_{L^2(\Gamma_D)},$$

$$\operatorname{osc}_N^2 \lesssim \sum_{e \in \mathscr{E}_h^N} \frac{h_e}{p_e} \|u_N - u_{N,hp}\|_{L^2(e)}^2.$$

With

$$\eta^2 := \sum_{T \in \mathscr{T}_h} \left(\eta_{T,1}^2 + \eta_{T,2}^2 \right) + \sum_{e \in \mathscr{E}_h^i} \eta_{e,1}^2 + \sum_{e \in \mathscr{E}_h^D} \eta_{e,2}^2,$$

$$\operatorname{osc}^2 := \operatorname{osc}_{\mathscr{T}_h}^2 + \operatorname{osc}_D^2 + \operatorname{osc}_{\mathscr{E}_h}^2 + \operatorname{osc}_N^2$$

there holds the reliable a posteriori error estimate

$$|u - u_{hp}|_{1,hp}^2 + \|\sigma - \sigma_{hp}\|_{L^2(\Omega)}^2 \lesssim \eta^2 + \operatorname{osc}^2. \tag{2.21}$$

Furthermore, one obtains guaranteed efficiency of the error contributions (2.20) by applying standard arguments with the typical loss in the p-scaling in (2.22a):

$$\eta_{T,1} \lesssim_\varepsilon p_T \|\sigma - \sigma_{hp}\|_{L^2(T)} + p_T^{1/2+\epsilon} \frac{h_T}{p_T} \|f - f_{hp}\|_{L^2(T)}, \tag{2.22a}$$

$$\eta_{T,2} \leq \|\sigma_{hp} - \sigma\|_{L^2(T)} + |u - u_{hp}|_{H^1(T)}, \tag{2.22b}$$

$$\eta_{e,1} = \frac{p_e}{\sqrt{h_e}} \left\| [\![u - u_{hp}]\!] \right\|_{L^2(e)}, \tag{2.22c}$$

$$\eta_{e,2} \leq \frac{p_e}{\sqrt{h_e}} \left\| u - u_{hp} \right\|_{L^2(e)} + \frac{p_e}{\sqrt{h_e}} \left\| u_D - u_{D,hp} \right\|_{L^2(e)} \tag{2.22d}$$

with an arbitrary $\epsilon > 0$. Note that (2.22b)–(2.22d) are h- and p-robust by the definition of the norm in (2.19). Eventually, we have

$$\begin{aligned} &\sum_{T \in \mathscr{T}_h} \left(\eta_{T,1}^2 + \eta_{T,2}^2 \right) + \sum_{e \in \mathscr{E}_h^i} \eta_{e,1}^2 + \sum_{e \in \mathscr{E}_h^D} \eta_{e,2}^2 \\ &\lesssim \sum_{T \in \mathscr{T}_h} p_T^2 \|\sigma - \sigma_{hp}\|_{L^2(T)}^2 + |u - u_{hp}|_{1,hp}^2 + p_T^{1+2\epsilon} \operatorname{osc}_{\mathscr{T}_h}^2 + \operatorname{osc}_{\mathscr{E}_h^D}^2 . \end{aligned} \tag{2.23}$$

The proof of (2.21) and (2.22) as well as the specific application of the implicit reconstruction can be found in [7].

As in Sect. 2.4, we consider the Poisson model problem on the L-shape domain $\Omega := (-1,1)^2 \backslash ([0,1) \times (-1,0])$ with the solution $u := r^{2/3} \sin(2\theta/3)$. As we can see in Fig. 2.4a, the use of h-adaptive refinements (based on Dörfler-marking) recovers the optimal algebraic convergence order $\mathscr{O}(N^{-1/2})$ for $p = 1$, $\mathscr{O}(N^{-1})$ for $p = 2$ and $\mathscr{O}(N^{-3/2})$ for $p = 3$. Here, the error $(e_{u,hp}^2 + e_\sigma^2)^{1/2}$ with $e_{u,hp} := |u - u_{hp}|_{1,hp}$ is plotted. Moreover, hp-adaptive refinements (based on Legendre polynomial expansion) lead to an optimal exponential convergence order $\mathscr{O}(\exp(-b_i N^{1/3}))$ with some slope $b_i > 0$, see Fig. 2.5a and Table 2.1, and are again characterized by the typical hp-refinement patterns with low polynomial degrees near to the reentrant corner ($p_T = 1$ at the reentrant corner), see Fig. 2.5b. As in Sect. 2.4, the use of hp-adaptivity based a posteriori error estimates is superior to

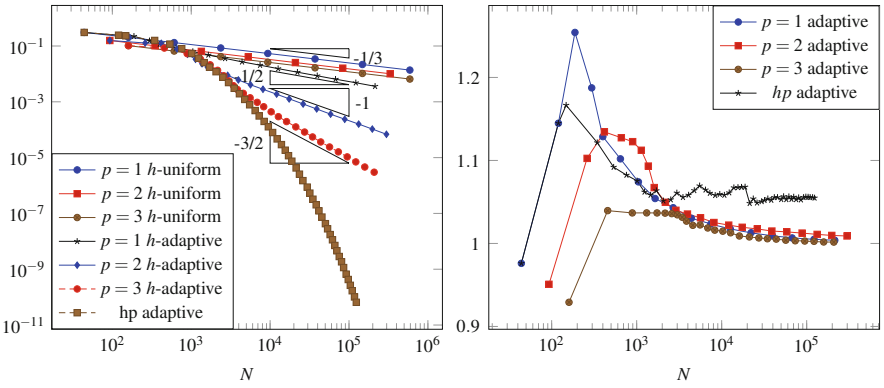


Fig. 2.4 **(a)** $(e_{u,hp}^2 + e_\sigma^2)^{1/2}$ on y-axis in logarithmic scaling. **(b)** Efficiency indices $(e_{u,hp}^2 + e_\sigma^2)^{1/2}/\eta_i$

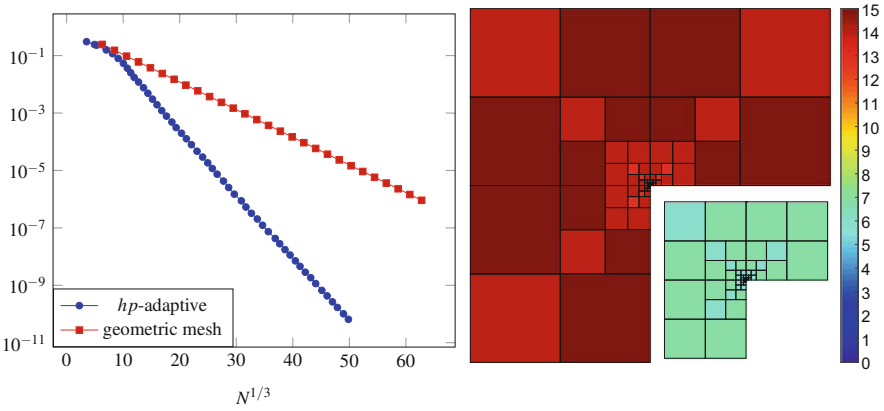


Fig. 2.5 **(a)** $(e_{u,hp}^2 + e_\sigma^2)^{1/2}$ on y-axis in logarithmic scaling, **(b)** hp-adaptive mesh with zoom towards the reentrant corner and polynomial degrees marked by some colors

the use of the hp-geometric refinements, see Fig. 2.5a and Table 2.1, where C_i, b_i, $\tilde{C}_g$and $\tilde{b}_g$ are computed with respect to $e_k := (e_{u,hp}^2 + e_\sigma^2)^{1/2}$. The efficiency indices defined as $(e_{u,hp}^2 + e_\sigma^2)^{1/2}/\eta_i$ with estimated error η_i are nearly constant 1 and, thus, very robust even for the hp-adaptive scheme, see Fig. 2.4b. We note that this observation and, in particular, the gap between the observed and proven efficiency indices could be an interesting follow up research question.

2.7 Comparison of the Explicit and Implicit Reconstruction Approaches

The numerical results of Sects. 2.4 and 2.6 are very similar. We obtain optimal algebraic and exponential convergence, if adaptivity is applied, as well as nearly constant efficiency indices. In the case of the implicit reconstruction the efficiency indices are even nearly 1.The numerical results are particularly confirmed if uniform and hp-geometric refinements are applied (i.e. meshes, which are not generated by refinements based on a posteriori error estimates). In Figs. 2.6a and 2.9a the efficiency indices e_σ/η_e and $(e_{u,hp}^2 + e_\sigma^2)^{1/2}/\eta_i$ for uniform and hp-geometric refinements are plotted. In particular, Fig. 2.6b shows the efficiency indices for h-uniform mesh refinements with multilevel hanging nodes of order 5, where an initial mesh as shown in Fig. 2.1a is used. These results show that both reconstruction approaches can be applied in the presence of hanging nodes. Figure 2.7a and b shows the errors $e_u := \|u - u_{hp}\|_{L^2(\Omega)}$, e_σ and $e_{u,hp}$ for uniform and hp-geometric refinements. We see that e_u is of higher order, which can be expected. Moreover, it seems to be important to determine the efficiency indices with both errors e_σ and

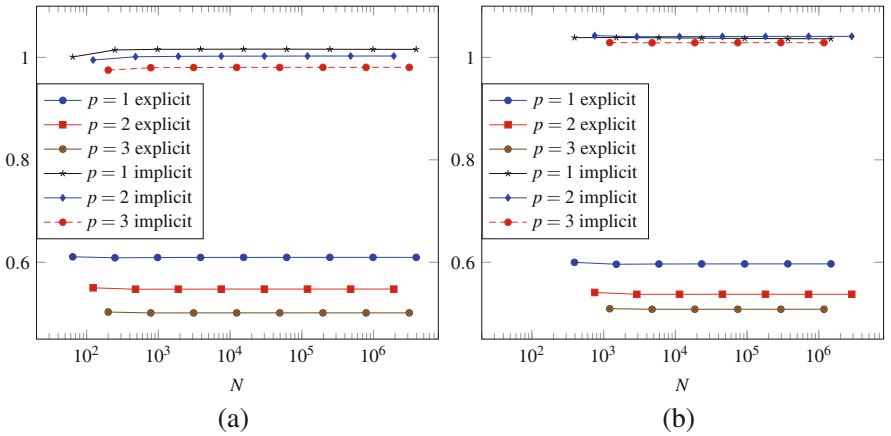


Fig. 2.6 Efficiency indices e_σ/η_e (explicit) and $(e_{u,hp}^2 + e_\sigma^2)^{1/2}/\eta_i$ (implicit) for uniform mesh refinements and polynomial degree $p = 1, 2, 3$: **(a)** without hanging nodes, **(b)** with multilevel hanging nodes of order 5

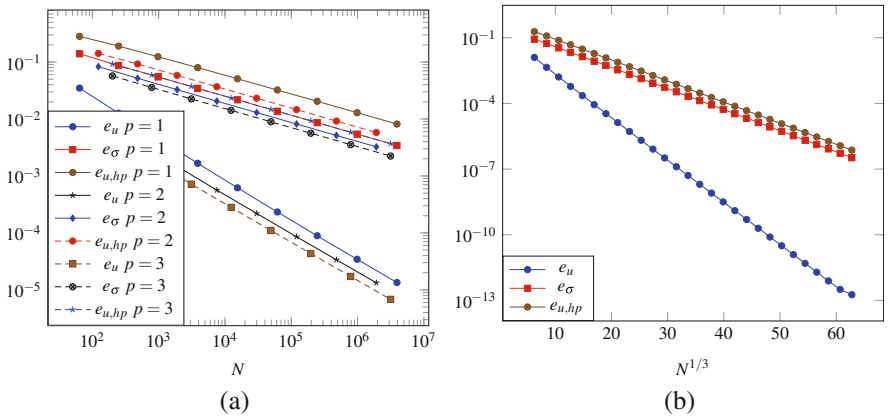


Fig. 2.7 Errors e_u, e_σ and $e_{u,hp}$: **(a)** uniform refinements, **(b)** hp-geometric refinements

$e_{u,hp}$ in the case of the implicit reconstruction as they are of the same magnitude and order. A certain advantage of the use of the implicit reconstruction in error control may be seen in Figs. 2.8a, b and 2.9b, where the CPU times (in seconds) needed for the computations are depicted. We used a matlab implementation on a compute-server with 4x Intel Xeon Gold 6144 8C, 3.50 GHz and 1.5TB memory. As we can see in the figures, the additional computations of local solutions needed for the explicit reconstruction and the evaluation of the estimation based on this approach lead to a certain computational amount. If uniform polynomial degrees and a mesh without hanging nodes are assumed, this amount is comparable to the computational amount which is required for the estimation resulting from the

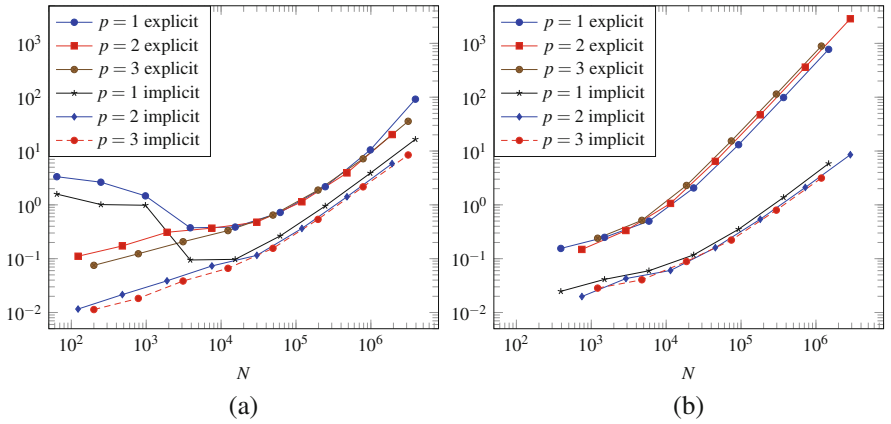


Fig. 2.8 CPU time for uniform mesh refinements: **(a)** without hanging nodes, **(b)** with multilevel hanging nodes of order 5

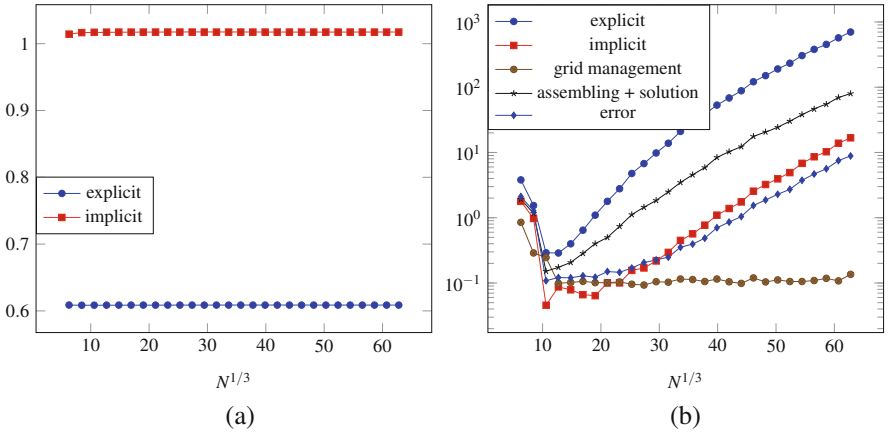


Fig. 2.9 **(a)** Efficiency indices e_σ/η_e (explicit) and $(e_{u,hp}^2 + e_\sigma^2)^{1/2}/\eta_i$ (implicit) for hp-geometric refinements. **(b)** CPU time for hp-geometric refinements

implicit reconstruction, see Fig. 2.8a. In contrast, if multilevel hanging nodes occur (as depicted in Fig. 2.1a), the explicit reconstruction seems to take essentially more computational time due to the complicated connectivity of the degrees of freedom, see Fig. 2.8b. Comparing Fig. 2.8a and b, we observe that the presence of hanging nodes has no effect on the computational time needed for the estimation based on the implicit reconstruction. In Fig. 2.9b several computational times resulting from the assembling and solution process, grid management as well as error evaluation are depicted, where hp-geometric refinements are used. We observe that the computational time for the solution of the local problems of the explicit reconstruction seems to dominate even the finite element computations for

assembling and solution, which may also be traced back to the resolving of the involved hanging nodes in this example. The computational time resulting from the implicit reconstruction does not dominate the assembling and solution process.

Acknowledgements The third author gratefully acknowledges support by the Deutsche Forschungsgemeinschaft in the Priority Programme 1748 "Reliable simulation techniques in solid mechanics. Development of non-standard discretization methods, mechanical and mathematical analysis" under the project "High-order immersed-boundary methods in solid mechanics for structures generated by additive processes", Grant SCHR 1244/4-2.

References

1. Ainsworth, M.: A posteriori error estimation for lowest order Raviart-Thomas mixed finite elements. SIAM J. Sci. Comput. **30**(1), 189–204 (2007/2008)
2. Ainsworth, M., Ma, X.: Non-uniform order mixed FEM approximation: implementation, post-processing, computable error bound and adaptivity. J. Comput. Phys. **231**(2), 436–453 (2012)
3. Alonso, A.: Error estimators for a mixed method. Numer. Math. **74**(4), 385–395 (1996)
4. Arbogast, T., Chen, Z.: On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. Math. Comp. **64**(211), 943–972 (1995)
5. Arnold, D., Brezzi, F.: Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. RAIRO Modél. Math. Anal. Numér. **19**, 7–32 (1985)
6. Banz, L., Petsche, J., Schröder, A.: hp-FEM for a stabilized three-field formulation of the biharmonic problem. Comput. Math. Appl. **77**(9), 2463–2488 (2019)
7. Banz, L., Petsche, J., Schröder, A.: A posteriori error estimates for hp-dual mixed and mixed-hybrid finite elements. Submitted for publication
8. Banz, L., Petsche, J., Schröder, A.: Hybridization and stabilization for hp-finite element methods. Appl. Numer. Math. **136**, 66–102 (2019)
9. Braess, D., Verfürth, R.: A posteriori error estimators for the Raviart-Thomas element. SIAM J. Numer. Anal. **33**(6), 2431–2444 (1996)
10. Braess, D., Pillwein, V., Schöberl, J.: Equilibrated residual error estimates are p-robust. Comput. Methods Appl. Mech. Eng. **198**(13–14), 1189–1197 (2009)
11. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15, 3rd edn. Springer, New York (2008)
12. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer-Verlag, New York (1991)
13. Byfut, A., Schröder, A.: Unsymmetric multi-level hanging nodes and anisotropic polynomial degrees in H^1-conforming higher-order finite element methods. Comput. Math. Appl. **73**(9), 2092–2150 (2017)
14. Canuto, C., Nochetto, R.H., Stevenson, R., Verani, M.: Convergence and optimality of hp-afem. Numer. Math. **135**(4), 1073–1119 (2017)
15. Canuto, C., Nochetto, R.H., Stevenson, R., Verani, M.: On p-robust saturation for hp-AFEM. Comput. Math. Appl. **73**(9), 2004–2022 (2017)
16. Carstensen, C.: A posteriori error estimate for the mixed finite element method. Math. Comput. **66**(218), 465–476 (1997)
17. Carstensen, C., Peterseim, D., Schröder, A.: The norm of a discretized gradient in $H(\mathrm{div})^*$ for a posteriori finite element error analysis. Numer. Math. **132**(3), 519–539 (2016)
18. Cockburn, B., Zhang, W.: An a posteriori error estimate for the variable-degree Raviart-Thomas method. Math. Comput. **83**(287), 1063–1082 (2014)

19. Dolejší, V., Šebestová, I., Vohralík, M.: Algebraic and discretization error estimation by equilibrated fluxes for discontinuous Galerkin methods on nonmatching grids. J. Sci. Comput. **64**(1), 1–34 (2015)
20. Dolejší, V., Ern, A., Vohralík, M.: hp-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems. SIAM J. Sci. Comput. **38**(5), A3220–A3246 (2016)
21. Ern, A., Vohralík, M.: Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. SIAM J. Numer. Anal. **53**(2), 1058–1081 (2015)
22. Ern, A., Smears, I., Vohralík, M.: Discrete p-robust $H(\mathrm{div})$-liftings and a posteriori estimates for elliptic problems with H^{-1} source terms. Calcolo **54**(3), 1009–1025 (2017)
23. Heuer, N.: Additive Schwarz method for the p-version of the boundary element method for the single layer potential operator on a plane screen. Numer. Math. **88**(3), 485–511 (2001)
24. Houston, P., Süli, E.: A note on the design of hp-adaptive finite element methods for elliptic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**(2–5), 229–243 (2005)
25. Houston, P., Schötzau, D., Wihler, T.P.: Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems. Math. Models Methods Appl. Sci. **17**(1), 33–62 (2007)
26. Houston, P., Süli, E., Wihler, T.P.: A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs. IMA J. Numer. Anal. **28**(2), 245–273 (2008)
27. Karakashian, O.A., Pascal, F.: A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. SIAM J. Numer. Anal. **41**(6), 2374–2399 (2003)
28. Kim, K.-Y.: Guaranteed a posteriori error estimator for mixed finite element methods of elliptic problems. Appl. Math. Comput. **218**(24), 11820–11831 (2012)
29. Larson, M.G., Målqvist, A.: A posteriori error estimates for mixed finite element approximations of elliptic problems. Numer. Math. **108**(3), 487–500 (2008)
30. Lovadina, C., Stenberg, R.: Energy norm a posteriori error estimates for mixed finite element methods. Math. Comput. **75**(256), 1659–1674 (2006)
31. Perugia, I., Schötzau, D.: An hp-analysis of the local discontinuous Galerkin method for diffusion problems. In: Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala), vol. 17, pp. 561–571 (2002)
32. Petsche, J., Schröder, A.: A posteriori error control and adaptivity of hp-finite elements for mixed and mixed-hybrid methods. Comput. Math. Appl. **74**(7), 1661–1674 (2017)
33. Schröder, A.: Constrained approximation in hp-FEM: unsymmetric subdivisions and multi-level hanging nodes. In: Spectral and High Order Methods for Partial Differential Equations. Lecture Notes in Computational Science and Engineering, vol. 76, pp. 317–325. Springer, Heidelberg (2011)
34. Schwab, C.: p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics. Numerical Mathematics and Scientific Computation. The Clarendon Press/Oxford University Press, New York (1998)
35. Steinbach, O.: Numerical Approximation Methods for Elliptic Boundary Value Problems. Springer, New York (2008). Finite and boundary elements, Translated from the 2003 German original
36. Stenberg, R.: Postprocessing schemes for some mixed finite elements. RAIRO Modél. Math. Anal. Numér. **25**(1), 151–167 (1991)
37. Vohralík, M.: A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. SIAM J. Numer. Anal. **45**(4), 1570–1599 (2007)
38. Vohralík, M.: Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. Math. Comput. **79**(272), 2001–2032 (2010)

Chapter 3
Two Stabilized Three-Field Formulations for the Biharmonic Problem

Lothar Banz, Jan Petsche, and Andreas Schröder

Abstract We consider two three-field formulations for the biharmonic equation. Both consist of the potential u, the flux $\boldsymbol{\sigma}$ and a Lagrange multiplier to weakly enforce $\boldsymbol{\sigma} = \nabla u$. They differ in their bilinear form and in the regularity requirement of $\boldsymbol{\sigma}$, and thus exhibit different numerical behaviors. We propose a stabilization of the discrete mixed formulations to allow for independent discretizations of the three variables and, thus, even the use of the same mesh and same polynomial degree for all variables. We derive a priori error estimates which are explicitly given in h and p. Several numerical experiments demonstrate the behavior of the methods and underline our theoretical results. In particular, for uniform mesh refinements we obtain optimal algebraic convergence and for uniform p-refinements exponential convergence, provided that the solutions are sufficiently smooth.

3.1 Introduction

Thin beams and plates, strain gradient elasticity [10, 15], the Stokes problem [16] and the phase separation of a binary mixture [23] are all modeled by fourth order problems of which the biharmonic problem is the most prominent model problem. A standard variational formulation of such a fourth order problem requires H^2-test and -trial spaces, and its conforming finite element discretization requires the implementation of globally C^1-continuous basis functions. With Argyris elements and Bogner-Fox-Schmit elements such basis functions are known, but they are far from being easy to implement, in particular in the presence of hanging nodes and varying polynomial degrees. As hanging nodes and varying polynomial degrees appear naturally in adaptive methods, two general strategies to deal with that challenge have attracted a lot of interest. The first strategy is a discontinuous Galerkin method as proposed in [2, 8, 13, 15, 21, 23] in which the non-conformity is

L. Banz (✉) · J. Petsche · A. Schröder
Department of Mathematics, University of Salzburg, Salzburg, Austria
e-mail: lothar.banz@sbg.ac.at; jan.petsche@sbg.ac.at; andreas.schroeder@sbg.ac.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_3

handled, for example, by an interior penalty approach. The second strategy is the use of a mixed method, where typically the flux $\boldsymbol{\sigma} = \nabla u$ of the potential u is introduced as a new variable, which relaxes the regularity requirements of the potential u, see e.g. [1, 4, 9–12, 17–19]. There are many ways to enforce the constraint $\boldsymbol{\sigma} = \nabla u$. One way is to enforce it weakly by a Lagrange multiplier which leads to a three-field formulation of the biharmonic equation [4], a strategy we adopt here as well.

A certain drawback of a mixed method is that an inf-sup condition must be satisfied on the discrete level, i.e. the dual space for the Lagrange multiplier needs to be sufficiently small (while keeping appropriate approximation properties). For h-methods this can be guaranteed by coarsening the mesh size for the dual space, with the obvious disadvantage that one has to work with at least two meshes simultaneously. In the mixed method of [4] the potential u does not appear in the (leading) bilinear form and is only controlled via the constraint $\boldsymbol{\sigma} = \nabla u$. Consequently, the dual space needs to be sufficiently large to enable a good enough approximation of u, and to be sufficiently small for guaranteeing the discrete inf-sup condition at the same time. It might be difficult to find a good balance between these competing requirements, in particular, for hp-methods. In [4] such a balance was found for the lowest order h-version. Alternatively, the mixed problem can be stabilized as in e.g. [3, 5, 6], dating back to [7], to circumvent the discrete inf-sup condition, which allows for the discrete dual space to be arbitrarily large.

In this paper, we propose and compare two stabilized methods, which both enable the use of the same mesh and the same polynomial degree for all three variables. The two corresponding weak three-field mixed formulations rely on two different integration by parts formulas. The first formulation (3.4) has minimal regularity requirements on the flux $\boldsymbol{\sigma}$ whereas the second formulation (3.7) has better numerical properties. As the two weak formulations use the same Lagrange multiplier space and the same bilinear form to couple the three variables, we use the same stabilization technique for both of them. A priori and a posteriori error analysis for the stabilized second formulation were carried out by us in [6]. These results are cited here for the comparison of the stabilized second formulation with the stabilized first formulation. The latter is analyzed in this paper in more detail.

The rest of the paper is structured as follows: In Sect. 3.2 we formally introduce the two three-field mixed formulations, as well as their discretization and stabilization. Section 3.3 is devoted to the a priori error estimates, which concludes with guaranteed convergence rates under some regularity assumptions, see Theorem 3.2. Numerical experiments underline our theoretical findings and enable the comparison of the two formulations. They are discussed in Sect. 3.4.

Notations Beside the standard notations we use the Sobolev spaces $H_0^1(\Omega) := \{v \in H^1(\Omega) \mid v|_{\partial\Omega} = 0\}$, its dual $H^{-1}(\Omega) := \left(H_0^1(\Omega)\right)^*$, $H_0^2(\Omega) := \left\{v \in H^2(\Omega) \mid v|_{\partial\Omega} = 0,\ \nabla v|_{\partial\Omega} = 0\right\}$ and $\mathbf{H}_0(\mathrm{div}, \Omega) := \left\{\boldsymbol{\tau} \in \mathbf{L}^2(\Omega) \mid \mathrm{div}\,\boldsymbol{\tau} \in L^2(\Omega),\ \boldsymbol{\tau} \cdot \mathbf{n} = 0 \text{on } \partial\Omega\right\}$. Here and in the following a quantity in bold font stands for a vector-valued quantity. Often the generic constant $C(d) > 0$, which is independent of the discretization but potentially dependent on d, is expressed by using the inequality sign $\lesssim_d$, i.e. $a \lesssim_d b :\Leftrightarrow a \leq C(d)b$.

3.2 Two Three-Field Formulations and Their Stabilized Discretizations

Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain with boundary $\partial\Omega$ and outer unit normal $\mathbf{n}$. We denote the normal derivative by $\partial_n v := \partial v/\partial \mathbf{n}$ and consider the biharmonic equation (3.1) with homogeneous Dirichlet boundary conditions, i.e. $u = \partial_n u = 0$ on $\partial\Omega$ and, thus, $\nabla u = 0$ on $\partial\Omega$: Find $u \in H_0^2(\Omega)$ such that

$$\Delta^2 u = f \text{ in } \Omega \tag{3.1}$$

for a given $f \in H^{-1}(\Omega)$. The higher smoothness of f is needed for the mixed methods below. There exist two different integration by parts formulas. The first one

$$\int_\Omega \Delta^2 u v \, dx = \int_\Omega \Delta u \Delta v \, dx \quad \forall u, v \in H_0^2(\Omega) \tag{3.2}$$

is obtained by applying the divergence theorem twice. The second one

$$\int_\Omega \Delta^2 u v \, dx = \int_\Omega \nabla\nabla u : \nabla\nabla v \, dx \quad \forall u, v \in H_0^2(\Omega) \, , \tag{3.3}$$

with $\nabla\nabla u : \nabla\nabla v = \sum_{i,j=1}^2 \partial^2 u/(\partial x_i \partial x_j) \cdot \partial^2 v/(\partial x_i \partial x_j)$ is obtained by integration by parts, rearranging the order of derivatives and integration by parts again, see e.g. [6] for the intermediate steps. These two integration by parts formulas lead to two different weak formulations with different bilinear forms. However, both formulations still suffer from the use of $H_0^2(\Omega)$-test and -trial functions. We follow here the strategy recently introduced in [4] to reduce the differentiability requirements by introducing the flux $\boldsymbol{\sigma} = \nabla u$ as a new variable and by weakly enforcing the constraint $\boldsymbol{\sigma} = \nabla u$ with a Lagrange multiplier $\boldsymbol{\phi}$. Using (3.2), this leads to the first mixed formulation: Find $(u, \boldsymbol{\sigma}, \boldsymbol{\phi}) \in H_0^1(\Omega) \times \mathbf{H}_0(\mathrm{div}, \Omega) \times \mathbf{M}$ such that

$$\int_\Omega \mathrm{div}\, \boldsymbol{\sigma} \, \mathrm{div}\, \boldsymbol{\tau} \, dx + b(\boldsymbol{\phi}; v, \boldsymbol{\tau}) = \int_\Omega f v \, dx \tag{3.4a}$$

$$b(\boldsymbol{\psi}; u, \boldsymbol{\sigma}) = 0 \tag{3.4b}$$

for all $(v, \boldsymbol{\tau}, \boldsymbol{\psi}) \in H_0^1(\Omega) \times \mathbf{H}_0(\mathrm{div}, \Omega) \times \mathbf{M}$. Here,

$$b(\boldsymbol{\psi}; v, \boldsymbol{\tau}) := \int_\Omega \boldsymbol{\psi} \cdot (\boldsymbol{\tau} - \nabla v) \, dx \tag{3.5}$$

results from the duality pairing of $\boldsymbol{\psi}$ with $\boldsymbol{\tau}$ and ∇v, respectively. The Lagrange multiplier space $\mathbf{M} := \left\{ \boldsymbol{\psi} \in \mathbf{H}^{-1}(\Omega) \mid \|\boldsymbol{\psi}\|_{\mathbf{M}} < \infty \right\}$ is equipped with its natural

(semi-) norms

$$\|\boldsymbol{\psi}\|_{\mathbf{M}}^2 := \|\boldsymbol{\psi}\|_{\mathbf{H}^{-1}(\Omega)}^2 + |\boldsymbol{\psi}|_{\mathbf{H}^{-1}(\Omega)}^2, \quad |\boldsymbol{\psi}|_{\mathbf{H}^{-1}(\Omega)} := \sup_{v \in H_0^1(\Omega),\, \|v\|_{H^1(\Omega)}=1} \int_\Omega \boldsymbol{\psi} \cdot \nabla v \, dx \,. \tag{3.6}$$

Note that $|\boldsymbol{\psi}|_{\mathbf{H}^{-1}(\Omega)} = \|\operatorname{div} \boldsymbol{\psi}\|_{H^{-1}(\Omega)}$. The use of the second integration by parts formula (3.3) instead of (3.2) yields the second mixed formulation: Find $(u, \boldsymbol{\sigma}, \boldsymbol{\phi}) \in H_0^1(\Omega) \times \mathbf{H}_0^1(\Omega) \times \mathbf{M}$ such that

$$\int_\Omega \nabla \boldsymbol{\sigma} : \nabla \boldsymbol{\tau} \, dx + b(\boldsymbol{\phi}; v, \boldsymbol{\tau}) = \int_\Omega f v \, dx \tag{3.7a}$$

$$b(\boldsymbol{\psi}; u, \boldsymbol{\sigma}) = 0 \tag{3.7b}$$

for all $(v, \boldsymbol{\tau}, \boldsymbol{\psi}) \in H_0^1(\Omega) \times \mathbf{H}_0^1(\Omega) \times \mathbf{M}$. We refer to [4, 6] for a derivation and analysis of these two mixed formulations and summarize the following properties:

Lemma 3.1

1. *There holds* $\mathbf{L}^2(\Omega) \subseteq \mathbf{M}$.
2. *The inf-sup condition is satisfied, i.e. for all* $\boldsymbol{\psi} \in \mathbf{M}$ *there holds*

$$\sup_{0 \neq (v,\boldsymbol{\tau}) \in H_0^1(\Omega) \times \mathbf{H}_0(\operatorname{div},\Omega)} \frac{b(\boldsymbol{\psi}; v, \boldsymbol{\tau})}{\left(\|v\|_{H^1(\Omega)}^2 + \|\boldsymbol{\tau}\|_{\mathbf{H}(\operatorname{div},\Omega)}^2 \right)^{1/2}} \geq 2^{-1} \|\boldsymbol{\psi}\|_{\mathbf{M}} \,, \tag{3.8}$$

$$\sup_{0 \neq (v,\boldsymbol{\tau}) \in H_0^1(\Omega) \times \mathbf{H}_0^1(\Omega)} \frac{b(\boldsymbol{\psi}; v, \boldsymbol{\tau})}{\left(\|v\|_{H^1(\Omega)}^2 + \|\boldsymbol{\tau}\|_{\mathbf{H}^1(\Omega)}^2 \right)^{1/2}} \geq 2^{-1} \|\boldsymbol{\psi}\|_{\mathbf{M}} \,. \tag{3.9}$$

3. *The bilinear forms* $\int_\Omega \operatorname{div} \boldsymbol{\sigma} \operatorname{div} \boldsymbol{\tau} \, dx$ *and* $\int_\Omega \nabla \boldsymbol{\sigma} : \nabla \boldsymbol{\tau} \, dx$ *are coercive on their associated kernel spaces.*
4. *The mixed problems (3.4) and (3.7) have a unique solution with* $\boldsymbol{\sigma} = \nabla u$ *and* $\boldsymbol{\phi} = \nabla(\operatorname{div} \boldsymbol{\sigma}) = \nabla(\Delta u)$.

We note that there is an interesting challenge in discretizing (3.4) and (3.7). On the one hand, the discrete Lagrange multiplier space needs to be sufficiently small (while keeping good approximation properties) to satisfy the discrete inf-sup condition. On the other hand it needs to be sufficiently large as u is only controlled via the constraining equations (3.4b) and (3.7b), respectively. We resolve this conflict by stabilizing the discrete problems, which allows the discrete Lagrange multiplier space to be arbitrarily large. First, we stabilize the bilinear forms to ensure positive definiteness independent of the discrete kernel space, i.e.

$$a_1(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) := \int_\Omega \operatorname{div} \boldsymbol{\sigma} \operatorname{div} \boldsymbol{\tau} \, dx + \int_\Omega (\boldsymbol{\sigma} - \nabla u) \cdot (\boldsymbol{\tau} - \nabla v) \, dx \,, \tag{3.10}$$

$$a_2(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) := \int_\Omega \nabla \boldsymbol{\sigma} : \nabla \boldsymbol{\tau} \, dx + \int_\Omega (\boldsymbol{\sigma} - \nabla u) \cdot (\boldsymbol{\tau} - \nabla v) \, dx \,. \tag{3.11}$$

Second, we introduce stabilizing terms to circumvent the discrete inf-sup condition, i.e.

$$c_\gamma(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}) := \int_\Omega \gamma\, (\boldsymbol{\phi} - \nabla_h \operatorname{div} \boldsymbol{\sigma}) \cdot \nabla_h \operatorname{div} \boldsymbol{\tau}\, dx\ , \tag{3.12}$$

$$d_\gamma(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\psi}) := \int_\Omega \gamma\, \boldsymbol{\psi} \cdot (\boldsymbol{\phi} - \nabla_h \operatorname{div} \boldsymbol{\sigma})\, dx\ , \tag{3.13}$$

where γ is a piecewise constant function with $\gamma|_T := \overline{\gamma}\, h_T^2\, p_T^{-4}$ for $T \in \mathscr{T}_h$ and a sufficiently small $\overline{\gamma} \in \mathbb{R}_{>0}$ (see Lemma 3.2). Here ∇_h is the elementwise gradient operator on $\mathscr{T}_h$. In the following, $\mathscr{T}_h$ and $\mathscr{T}_k$ are two independent, locally quasi-uniform triangulations of Ω into quadrilaterals with diameters $\{h_T\}_{T\in\mathscr{T}_h}$, $\{k_T\}_{T\in\mathscr{T}_k}$, and $\{p_T\}_{T\in\mathscr{T}_h}$, $\{q_T\}_{T\in\mathscr{T}_k}$ are polynomial degree distributions on $\mathscr{T}_h$ and $\mathscr{T}_k$, respectively. The mapping $F_T : \hat{T} \to T$ is the bijective, bilinear coordinate transformation from the reference element $\hat{T} = (-1, 1)^2$ to the element $T \in \mathscr{T}_h$ and $\mathscr{F}_T : H^1(\hat{T}) \to H^1(T)$, $\mathscr{P}_T : H(\operatorname{div}, \hat{T}) \to H(\operatorname{div}, T)$ are the standard pullback, contravariant Piolatransformation of F_T, respectively, i.e.

$$\mathscr{F}_T(\hat{v}) := \hat{v} \circ F_T^{-1}\ , \quad \mathscr{P}_T(\hat{\boldsymbol{\tau}}) := J_{F_T}^{-1}\, D_{F_T}\, \hat{\boldsymbol{\tau}} \circ F_T^{-1} \tag{3.14}$$

with the Jacobian D_{F_T} of F_T and its determinant J_{F_T}. We use the conforming finite element spaces

$$V_{hp} := \left\{ v \in H_0^1(\Omega) \mid \forall\, T \in \mathscr{T}_h :\ v|_T \in \mathscr{F}_T(\mathbb{P}_{p_T+d}) \right\} \tag{3.15}$$

$$\mathbf{W}_{hp}^1 := \left\{ \boldsymbol{\tau} \in \mathbf{H}_0(\operatorname{div}, \Omega) \mid \forall\, T \in \mathscr{T}_h :\ \boldsymbol{\tau}|_T \in \mathscr{P}_T(\mathbf{RT}_{p_T}) \right\} \tag{3.16}$$

$$\mathbf{W}_{hp}^2 := \left\{ \boldsymbol{\tau} \in \mathbf{H}_0^1(\Omega) \mid \forall\, T \in \mathscr{T}_h :\ \boldsymbol{\tau}|_T \in [\mathscr{F}_T(\mathbb{P}_{p_T})]^2 \right\} \tag{3.17}$$

$$\mathbf{M}_{kq} := \left\{ \boldsymbol{\psi} \in \mathbf{L}^2(\Omega) \mid \forall\, T \in \mathscr{T}_k :\ \boldsymbol{\psi}|_T \in [\mathscr{F}_T(\mathbb{P}_{q_T})]^2 \right\} \tag{3.18}$$

with an offset $d \in \mathbb{N}_0$, polynomial space $\mathbb{P}_{k,l} := \operatorname{span}\left\{ x^i y^j \mid 0 \le i \le k,\ 0 \le j \le l \right\}$ and $\mathbf{RT}_{k,l} := \mathbb{P}_{k+1,l} \times \mathbb{P}_{k,l+1}$. Thus, the two stabilized discrete mixed formulations are: Find $(u_{hp}, \boldsymbol{\sigma}_{hp}, \boldsymbol{\phi}_{kq}) \in V_{hp} \times \mathbf{W}_{hp}^i \times \mathbf{M}_{kq}$, $i = 1, 2$, such that

$$a_i(u_{hp}, \boldsymbol{\sigma}_{hp}; v_{hp}, \boldsymbol{\tau}_{hp}) + b(\boldsymbol{\phi}_{kq}; v_{hp}, \boldsymbol{\tau}_{hp}) + c_\gamma(\boldsymbol{\sigma}_{hp}, \boldsymbol{\phi}_{kq}; \boldsymbol{\tau}_{hp}) = \int_\Omega f\, v_{hp}\, dx \tag{3.19a}$$

$$b(\boldsymbol{\psi}_{kq}; u_{hp}, \boldsymbol{\sigma}_{hp}) - d_\gamma(\boldsymbol{\sigma}_{hp}, \boldsymbol{\phi}_{kq}; \boldsymbol{\psi}_{kq}) = 0 \tag{3.19b}$$

for all $(v_{hp}, \boldsymbol{\tau}_{hp}, \boldsymbol{\psi}_{kq}) \in V_{hp} \times \mathbf{W}_{hp}^i \times \mathbf{M}_{kq}$.

Lemma 3.2 *If $\overline{\gamma}$ is sufficiently small, then $a_1(\cdot,\cdot;\cdot,\cdot)+c_\gamma(\cdot,0;\cdot)$ is positive definite and $a_2(\cdot,\cdot;\cdot,\cdot)+c_\gamma(\cdot,0;\cdot)$ is coercive.*

Proof The coercivity of $a_2(\cdot,\cdot;\cdot,\cdot)+c_\gamma(\cdot,0;\cdot)$ is proven in [6]. An elementwise polynomial inverse estimate yields

$$\|\gamma^{1/2}\nabla_h \operatorname{div} \boldsymbol{\tau}_{hp}\|^2_{\mathbf{L}^2(\Omega)} \leq \overline{\gamma}\, C \|\operatorname{div} \boldsymbol{\tau}_{hp}\|^2_{L^2(\Omega)} \,. \tag{3.20}$$

Thus,

$$\begin{aligned} a_1(v_{hp},\boldsymbol{\tau}_{hp};v_{hp},\boldsymbol{\tau}_{hp}) - \int_\Omega \gamma \nabla_h \operatorname{div} \boldsymbol{\tau}_{hp}\cdot\nabla_h \operatorname{div}\boldsymbol{\tau}_{hp}\,dx \\ \geq (1-\overline{\gamma}\,C)\|\operatorname{div}\boldsymbol{\tau}_{hp}\|^2_{L^2(\Omega)} + \|\boldsymbol{\tau}_{hp}-\nabla v_{hp}\|^2_{\mathbf{L}^2(\Omega)} \geq 0. \end{aligned}$$

For $\overline{\gamma}$ sufficiently small the lower bound is non-negative. It is zero if and only if $\operatorname{div}\boldsymbol{\tau}_{hp}=0$ and $\boldsymbol{\tau}_{hp}-\nabla v_{hp}=0$ almost everywhere. Hence,

$$\int_\Omega \nabla v_{hp}\nabla v_{hp}\,dx = \int_\Omega \boldsymbol{\tau}_{hp}\nabla v_{hp}\,dx = -\int_\Omega \operatorname{div}\boldsymbol{\tau}_{hp} v_{hp}\,dx = 0\,. \tag{3.21}$$

From Poincaré-Friedrich's inequality for $v_{hp}\in H^1_0(\Omega)$ we have $v_{hp}=0$ in this case. We conclude $a_1(\cdot,\cdot;\cdot,\cdot)+c_\gamma(\cdot,0;\cdot)=0$ if and only if $(v_{hp},\boldsymbol{\tau}_{hp})=0$. □

Straightforward computations show that (3.19) is equivalent to the saddle point problem

$$\mathscr{L}(u_{hp},\boldsymbol{\sigma}_{hp};\boldsymbol{\psi}_{kq}) \leq \mathscr{L}(u_{hp},\boldsymbol{\sigma}_{hp};\boldsymbol{\phi}_{kq}) \leq \mathscr{L}(v_{hp};\boldsymbol{\tau}_{hp},\boldsymbol{\phi}_{kq})$$

for all $(v_{hp},\boldsymbol{\tau}_{hp},\boldsymbol{\psi}_{kq})\in V_{hp}\times\mathbf{W}^i_{hp}\times\mathbf{M}_{kq}$ with

$$\begin{aligned} \mathscr{L}(v,\boldsymbol{\tau};\boldsymbol{\psi}) := \frac{1}{2}a_i(v,\boldsymbol{\tau};v,\boldsymbol{\tau}) - \int_\Omega f v\,dx + b(\boldsymbol{\psi};v,\boldsymbol{\tau}) \\ -\frac{1}{2}\int_\Omega \gamma(\boldsymbol{\psi}-\nabla_h\operatorname{div}\boldsymbol{\tau})\cdot(\boldsymbol{\psi}-\nabla_h\operatorname{div}\boldsymbol{\tau})\,dx\,. \end{aligned}$$

As $\mathscr{L}(v,\boldsymbol{\tau};\boldsymbol{\psi})$ is strictly convex in $(v_{hp},\boldsymbol{\tau}_{hp})$ (see Lemma 3.2) and strictly concave in $\boldsymbol{\psi}_{kq}$ as $\overline{\gamma}>0$ classical saddle point theory [14] yields:

Theorem 3.1 *The stabilized discrete mixed problem (3.19) has a unique solution $(u_{hp},\boldsymbol{\sigma}_{hp},\boldsymbol{\phi}_{kq})\in V_{hp}\times\mathbf{W}^i_{hp}\times\mathbf{M}_{kq}$ for $i=1,2$.*

3.3 A Priori Error Estimates

The a priori error analysis of this section avoids the use of a discrete inf-sup condition. It relies on an alternative representation of the Lagrange multiplier error in a scaled $\mathbf{L}^2$-norm. Based on that a statement similar to Céa-lemma can be derived, Lemma 3.4, which can be exploited to get guaranteed convergence rates under certain regularity assumptions, see Theorem 3.2.

Lemma 3.3 *If $\boldsymbol{\phi} \in \mathbf{L}^2(\Omega)$, then there holds*

$$\begin{aligned}\|\gamma^{1/2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} &= \int_\Omega (\boldsymbol{\phi}-\boldsymbol{\psi}_{kq})\cdot(\boldsymbol{\sigma}_{hp}-\nabla u_{hp}-\gamma(\boldsymbol{\phi}_{kq}-\nabla_h \operatorname{div}\boldsymbol{\sigma}_{hp}))\,dx\\ &\quad+\int_\Omega \gamma(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\cdot(\boldsymbol{\phi}-\nabla_h\operatorname{div}\boldsymbol{\sigma}_{hp})\,dx\\ &\quad+b(\boldsymbol{\phi}_{kq}-\boldsymbol{\phi};u_{hp}-u,\boldsymbol{\sigma}_{hp}-\boldsymbol{\sigma})\end{aligned}$$

for all $\boldsymbol{\psi}_{kq} \in \mathbf{M}_{kq}$.

Proof The proof is identical for $i = 1$ and $i = 2$ and is given in [6] for $i = 2$. □

Lemma 3.4 *Let $i = 1$ and $\overline{\gamma}$ be sufficiently small. If $\boldsymbol{\phi} \in \mathbf{L}^2(\Omega)$, then there holds*

$$\begin{aligned}&\|\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\|^2_{L^2(\Omega)}+\|\boldsymbol{\sigma}_{hp}-\nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)}+\|hp^{-2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)}\\ &\quad\lesssim_{\overline{\gamma}}\|\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)}+\|h^{-1}p^2(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})\|^2_{\mathbf{L}^2(\Omega)}+\|h^{-1}p^2\nabla(u-v_{hp})\|^2_{\mathbf{L}^2(\Omega)}\\ &\qquad+\|\boldsymbol{\phi}-\boldsymbol{\psi}_{kq}\|^2_{\mathbf{L}^2(\Omega)}\end{aligned}$$

for all $(v_{hp},\boldsymbol{\tau}_{hp},\boldsymbol{\psi}_{kq}) \in V_{hp}\times\mathbf{W}^1_{hp}\times\mathbf{M}_{kq}$.

Proof For the conforming discretization of $i = 1$ it holds that

$$\begin{aligned}&\int_\Omega \operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\operatorname{div}\boldsymbol{\tau}_{hp}-(\boldsymbol{\sigma}_{hp}-\nabla u_{hp})\cdot(\boldsymbol{\tau}_{hp}-\nabla v_{hp})\,dx\\ &\quad+b(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq};v_{hp},\boldsymbol{\tau}_{hp})-c_\gamma(\boldsymbol{\sigma}_{hp},\boldsymbol{\phi}_{kq};\boldsymbol{\tau}_{hp})=0\end{aligned}$$

for all $(v_{hp},\boldsymbol{\tau}_{hp})\in V_{hp}\times\mathbf{W}^1_{hp}$. Choosing the test function as $(u_{hp}-v_{hp},\boldsymbol{\sigma}_{hp}-\boldsymbol{\tau}_{hp})$ in the previous equation and using Lemma 3.3 we obtain

$$\begin{aligned}&\|\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\|^2_{L^2(\Omega)}+\|\boldsymbol{\sigma}_{hp}-\nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)}+\|\gamma^{1/2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)}\\ &\quad=\int_\Omega \operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})+(\boldsymbol{\sigma}_{hp}-\nabla u_{hp})(\boldsymbol{\tau}_{hp}-\nabla v_{hp})\,dx\\ &\qquad+b(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq};u_{hp}-v_{hp},\boldsymbol{\sigma}_{hp}-\boldsymbol{\tau}_{hp})-c_\gamma(\boldsymbol{\sigma}_{hp},\boldsymbol{\phi}_{kq};\boldsymbol{\sigma}_{hp}-\boldsymbol{\tau}_{hp})\end{aligned}$$

$$
\begin{aligned}
&+\|\gamma^{1/2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{L^2(\Omega)}\\
&=\underbrace{\int_\Omega \operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})\,dx}_{\mathrm{I}}+\underbrace{\int_\Omega(\boldsymbol{\sigma}_{hp}-\nabla u_{hp})(\boldsymbol{\tau}_{hp}-\nabla v_{hp})\,dx}_{\mathrm{II}}\\
&+\underbrace{b(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq};u-v_{hp},\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})}_{\mathrm{III}}\\
&\underbrace{-c_\gamma(\boldsymbol{\sigma}_{hp},\boldsymbol{\phi}_{kq};\boldsymbol{\sigma}_{hp}-\boldsymbol{\tau}_{hp})+\int_\Omega\gamma(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})(\boldsymbol{\phi}-\nabla_h\operatorname{div}\boldsymbol{\sigma}_{hp})\,dx}_{\mathrm{IV}}\\
&+\underbrace{\int_\Omega(\boldsymbol{\phi}-\boldsymbol{\psi}_{kq})(\boldsymbol{\sigma}_{hp}-\nabla u_{hp}-\gamma(\boldsymbol{\phi}_{kq}-\nabla_h\operatorname{div}\boldsymbol{\sigma}_{hp}))\,dx}_{\mathrm{V}}\,.
\end{aligned}
$$

For an arbitrarily $\varepsilon>0$, Cauchy-Schwarz inequality and Young's inequality yield

$$
\mathrm{I}\le\varepsilon\|\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{hp})\|^2_{L^2(\Omega)}+\frac{1}{4\varepsilon}\|\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)}
$$

and analogously with $\boldsymbol{\sigma}=\nabla u$ there holds

$$
\begin{aligned}
\mathrm{II}&\le\varepsilon\|\boldsymbol{\sigma}_{hp}-\nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)}+\frac{1}{4\varepsilon}\|\boldsymbol{\tau}_{hp}-\nabla v_{hp}\|^2_{\mathbf{L}^2(\Omega)}\\
&\le\varepsilon\|\boldsymbol{\sigma}_{hp}-\nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)}+\frac{1}{2\varepsilon}\left(\|\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp}\|^2_{\mathbf{L}^2(\Omega)}+\|\nabla(u-v_{hp})\|^2_{\mathbf{L}^2(\Omega)}\right)\,.
\end{aligned}
$$

The estimates of III, IV and V are (almost literally) identical to [6, Proof of Lem. 14], i.e.

$$
\begin{aligned}
\mathrm{III}&\le\varepsilon\|\gamma^{1/2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)}+\frac{1}{2\varepsilon}\|\gamma^{-1/2}(\boldsymbol{\sigma}-\boldsymbol{\tau}_{hp})\|^2_{\mathbf{L}^2(\Omega)}\\
&\quad+\frac{1}{2\varepsilon}\|\gamma^{-1/2}\nabla(u-v_{hp})\|^2_{\mathbf{L}^2(\Omega)}\,,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{IV}&\le\varepsilon\|\gamma^{1/2}(\boldsymbol{\phi}-\nabla_h\operatorname{div}\boldsymbol{\sigma}_{hp})\|^2_{\mathbf{L}^2(\Omega)}+\frac{C\,\overline{\gamma}}{4\varepsilon}\|\operatorname{div}(\boldsymbol{\sigma}_{hp}-\boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)}\\
&\quad+\varepsilon\|\gamma^{1/2}(\boldsymbol{\phi}-\boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)}+\frac{1}{4\varepsilon}\|\gamma^{1/2}(\boldsymbol{\phi}-\nabla_h\operatorname{div}\boldsymbol{\tau}_{hp})\|^2_{\mathbf{L}^2(\Omega)}\,,
\end{aligned}
$$

where C is the constant from the polynomial inverse estimate, and

$$\begin{aligned}\mathrm{V} \le \varepsilon\|\boldsymbol{\sigma}_{hp} - \nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)} + \frac{1}{4\varepsilon}\|\boldsymbol{\phi} - \boldsymbol{\psi}_{kq}\|^2_{\mathbf{L}^2(\Omega)} + \frac{1}{4\varepsilon}\|\gamma^{1/2}(\boldsymbol{\phi} - \boldsymbol{\psi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} \\ +2\varepsilon\|\gamma^{1/2}(\boldsymbol{\phi}_{kq} - \boldsymbol{\phi})\|^2_{\mathbf{L}^2(\Omega)} + 2\varepsilon\|\gamma^{1/2}(\boldsymbol{\phi} - \nabla_h \operatorname{div}\boldsymbol{\sigma}_{hp})\|^2_{\mathbf{L}^2(\Omega)} \,.\end{aligned}$$

It remains to estimate the terms

$$\mathrm{A} := \|\gamma^{1/2}(\boldsymbol{\phi} - \nabla_h \operatorname{div}\boldsymbol{\tau}_{hp})\|^2_{\mathbf{L}^2(\Omega)} \,, \quad \mathrm{B} := \|\gamma^{1/2}(\boldsymbol{\phi} - \nabla_h \operatorname{div}\boldsymbol{\sigma}_{hp})\|^2_{\mathbf{L}^2(\Omega)} \,,$$
$$\mathrm{C} := \|\operatorname{div}(\boldsymbol{\sigma}_{hp} - \boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)} \,.$$

From [6, Proof of Lem. 14] we obtain

$$\begin{aligned}\mathrm{A} &\le 2\|\gamma^{1/2}(\boldsymbol{\phi} - \boldsymbol{\psi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} + 4C\,\overline{\gamma}\sum_{T\in\mathscr{T}_h}\|\boldsymbol{\psi}_{kq} - \boldsymbol{\phi}\|^2_{\mathbf{H}^{-1}(T)} \\ &\quad +8C\,\overline{\gamma}\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)} \,, \\ \mathrm{B} &\le 2\|\gamma^{1/2}(\boldsymbol{\phi} - \boldsymbol{\psi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} + 4C\,\overline{\gamma}\sum_{T\in\mathscr{T}_h}\|\boldsymbol{\psi}_{kq} - \boldsymbol{\phi}\|^2_{\mathbf{H}^{-1}(T)} \\ &\quad +8C\,\overline{\gamma}\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp})\|^2_{L^2(\Omega)}\end{aligned}$$

and

$$\mathrm{C} \le 2\|\operatorname{div}(\boldsymbol{\sigma}_{hp} - \boldsymbol{\sigma})\|^2_{L^2(\Omega)} + 2\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)} \,.$$

Combining all these estimates yields

$$\begin{aligned}&\Big(1 - \varepsilon - C\,\overline{\gamma}(24\varepsilon + 0.5\varepsilon^{-1})\Big)\,\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp})\|^2_{L^2(\Omega)} + (1-2\varepsilon)\|\boldsymbol{\sigma}_{hp} - \nabla u_{hp}\|^2_{\mathbf{L}^2(\Omega)} \\ &\quad +(1-4\varepsilon)\|\gamma^{1/2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} \\ &\le \frac{1 + 10C\,\overline{\gamma}}{4\varepsilon}\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|^2_{L^2(\Omega)} + \frac{1}{2\varepsilon}\|\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp}\|^2_{\mathbf{L}^2(\Omega)} \\ &\quad +\frac{1}{2\varepsilon}\Big(\|\gamma^{-1/2}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|^2_{\mathbf{L}^2(\Omega)} + \|\nabla(u - v_{hp})\|^2_{\mathbf{L}^2(\Omega)} + \|\gamma^{-1/2}\nabla(u - v_{hp})\|^2_{\mathbf{L}^2(\Omega)}\Big) \\ &\quad +\frac{1}{4\varepsilon}\|\boldsymbol{\phi} - \boldsymbol{\psi}_{kq}\|^2_{\mathbf{L}^2(\Omega)} + (\frac{3}{4\varepsilon} + 6\varepsilon)\|\gamma^{1/2}(\boldsymbol{\phi} - \boldsymbol{\psi}_{kq})\|^2_{\mathbf{L}^2(\Omega)} \\ &\quad +C\,\overline{\gamma}(12\varepsilon + \frac{1}{\varepsilon})\sum_{T\in\mathscr{T}_h}\|\boldsymbol{\phi} - \boldsymbol{\psi}_{kq}\|^2_{\mathbf{H}^{-1}(T)} \,.\end{aligned}$$

Finally, choosing first ε and then $\overline{\gamma}$ sufficiently small so that $C\varepsilon^{-1}\overline{\gamma}$ is also sufficiently small and omitting the obviously dominated terms yield the assertion. □

We refer to [6] for a corresponding statement in the case $i = 2$. With well known interpolation operators [20, 22] we obtain the following guaranteed convergence rates under some regularity assumption on u.

Theorem 3.2 *Let $\overline{\gamma}$ be sufficiently small as in Lemma 3.4 and $u \in H^s(\Omega)$ for $s \geq 3$. There holds for $i = 1$*

$$\begin{aligned}&\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp})\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}_{hp} - \nabla u_{hp}\|_{\mathbf{L}^2(\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{kq})\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \lesssim_{\overline{\gamma}} \left(h^{\min\{p,p+d-1,s-2\}}p^{-s+7/2} + k^{\min\{q+1,s-3\}}q^{-s+3}\right)\|u\|_{H^s(\Omega)}\end{aligned}$$

and for $i = 2$

$$\begin{aligned}&\|u - u_{hp}\|_{H^1(\Omega)} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|_{\mathbf{H}^1(\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{kq})\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \lesssim_{\overline{\gamma}} \left(h^{\min\{p,p+d-1,s-2\}}p^{-s+3} + k^{\min\{q+1,s-3\}}q^{-s+3}\right)\|u\|_{H^s(\Omega)} .\end{aligned}$$

Proof For $i = 1$ we estimate the terms on the right hand side of Lemma 3.4 individually. By standard interpolation results for $H_0^1(\Omega)$, $\mathbf{H}_0(\operatorname{div}, \Omega)$ and $\mathbf{L}^2(\Omega)$ functions, see e.g. [20, 22], we obtain with $\boldsymbol{\sigma} = \nabla u \in \mathbf{H}^{s-1}(\Omega)$ and $\boldsymbol{\phi} = \nabla\Delta u \in \mathbf{H}^{s-3}(\Omega)$ that

$$\begin{aligned}\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|_{L^2(\Omega)} &\lesssim h^{\min\{p+1,s-2\}}p^{-s+2}\|u\|_{H^s(\Omega)} \\ \|\gamma^{-1/2}(\boldsymbol{\sigma} - \boldsymbol{\tau}_{hp})\|_{\mathbf{L}^2(\Omega)} &\lesssim h^{\min\{p,s-2\}}p^{-s+7/2}\|u\|_{H^s(\Omega)} \\ \|\gamma^{-1/2}\nabla(u - v_{hp})\|_{\mathbf{L}^2(\Omega)} &\lesssim h^{\min\{p+d-1,s-2\}}p^{-s+3}\|u\|_{H^s(\Omega)} \\ \|\boldsymbol{\phi} - \boldsymbol{\psi}_{kq}\|_{\mathbf{L}^2(\Omega)} &\lesssim k^{\min\{q+1,s-3\}}q^{-s+3}\|u\|_{H^s(\Omega)} .\end{aligned}$$

Omitting the dominated convergence rates yields the assertion for $i = 1$. The case of $i = 2$ is proven in [6]. □

Obviously, it is desirable to use the same mesh for the definition of V_{hp}, $\mathbf{W}_{hp}^i$ and $\mathbf{M}_{kq}$, i.e. $h = k$, and to choose $d = 1$ and $p = q + 1$. For this discretization, which, in fact, does not satisfy the discrete inf-sup condition, we obtain the following guaranteed convergence rates.

Corollary 3.1 *Let $\overline{\gamma}$ be sufficiently small and $u \in H^s(\Omega)$ for $s \geq 3$. If $d = 1$, $h = k$ and $p = q + 1$, then there holds for $i = 1$*

$$\begin{aligned}&\|\operatorname{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp})\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}_{hp} - \nabla u_{hp}\|_{\mathbf{L}^2(\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{kq})\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \lesssim_{\overline{\gamma}} h^{\min\{p,s-3\}}p^{-s+7/2}\|u\|_{H^s(\Omega)}\end{aligned}$$

and for $i = 2$

$$\|u - u_{hp}\|_{H^1(\Omega)} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|_{\mathbf{H}^1(\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{kq})\|_{\mathbf{L}^2(\Omega)} \\ \lesssim_{\overline{\gamma}} h^{\min\{p,s-3\}} p^{-s+3} \|u\|_{H^s(\Omega)} .$$

The convergence rates for the h-versions are optimal if the solution u is sufficiently smooth. For a less smooth solution $u \in H^s(\Omega)$ with $2 \leq s \leq 3$ Theorem 3.2 and Corollary 3.1 can not be applied in order to guarantee convergence rates as $\boldsymbol{\phi}$ might not be in $\mathbf{L}^2(\Omega)$. However, our computations in [6] for a typical L-shaped problem show observable convergence anyway.

3.4 Numerical Results

Let $\Omega := [0, 1]^2$ and the data f be chosen such that

$$u(x, y) := [\exp(x) + (x + 1)\exp(y)]x^2y^2(1 - x)^2(1 - y)^2$$

is the exact solution of (3.1). For all experiments we use the same mesh for $u_{hp}, \boldsymbol{\sigma}_{hp}$ and $\boldsymbol{\phi}_{kq}$, i.e. $h = k$. The polynomial degrees for $(u_{hp}, \boldsymbol{\sigma}_{hp}, \boldsymbol{\phi}_{kq})$ are $(p + 1, p, p)$ or $(p + 1, p, p - 1)$, i.e. $d = 1$ and $q = p$ or $q = p - 1$. The stabilization constant is $\overline{\gamma} = 0.06$ for $i = 1$ and $\overline{\gamma} = 0.2$ for $i = 2$ which is slightly less than the maximal allowed value for this constant. That upper bound can be computed by solving the finite dimensional eigenvalue problem: Find $(u^*_{hp}, \boldsymbol{\sigma}^*_{hp}) \in V_{hp} \times \mathbf{W}^i_{hp}$ and $\lambda^*_{hp} \in \mathbb{R}$ such that

$$a_i(u^*_{hp}, \boldsymbol{\sigma}^*_{hp}; v_{hp}, \boldsymbol{\tau}_{hp}) = \lambda^*_{hp} \sum_{T \in \mathcal{T}_h} \int_T h_T^2 p_T^{-4} \nabla_h \operatorname{div} \boldsymbol{\sigma}^*_{hp} \cdot \nabla_h \operatorname{div} \boldsymbol{\tau}_{hp} \, dx \tag{3.22}$$

for all $(v_{hp}, \boldsymbol{\tau}_{hp}) \in V_{hp} \times \mathbf{W}^i_{hp}$. From Lemma 3.2 it follows $\overline{\gamma} < \min \lambda^*_{hp}$. Our computations suggest that $\min \lambda^*_{hp}$ is independent of h and p, i.e. the maximal allowed amount of stabilization may be computed on the coarsest possible mesh.

Figures 3.1 and 3.2 display the reduction of the error

$$\left(\|u - u_{hp}\|^2_{H^1(\Omega)} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|^2_{\mathbf{H}^1(\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{hq})\|^2_{\mathbf{L}^2(\Omega)}\right)^{1/2} \tag{3.23}$$

for $i = 2$, i.e. the discretization of $H^1_0(\Omega) \times \mathbf{H}^1_0(\Omega) \times \mathbf{M}$. Figures 3.3 and 3.4 display the reduction of the error

$$\left(\|u - u_{hp}\|^2_{H^1(\Omega)} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|^2_{\mathbf{H}(\operatorname{div},\Omega)} + \|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{hq})\|^2_{\mathbf{L}^2(\Omega)}\right)^{1/2} \tag{3.24}$$

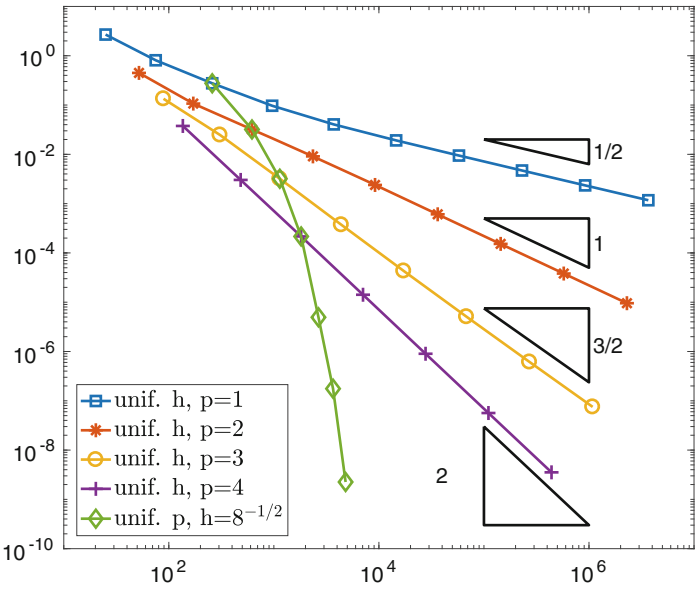


Fig. 3.1 $i = 2$, $(p+1, p, p)$: Error (3.23) vs. degrees of freedom

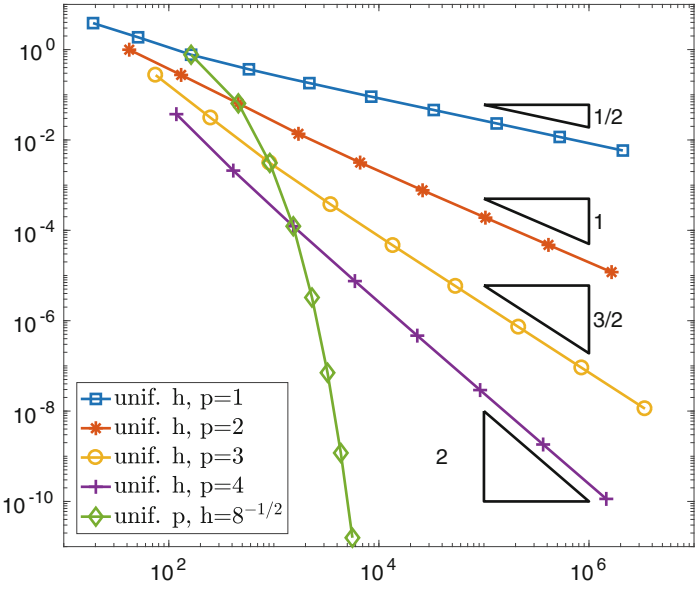


Fig. 3.2 $i = 2$, $(p+1, p, p-1)$: Error (3.23) vs. degrees of freedom

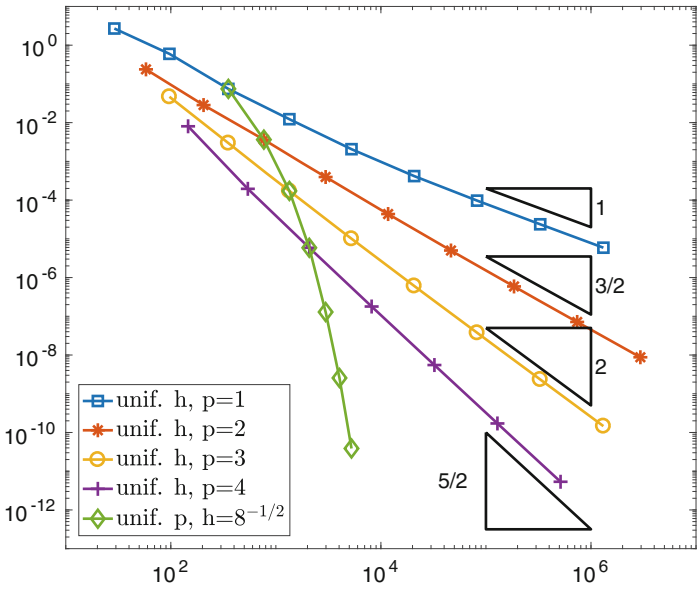


Fig. 3.3 $i = 1$ and $(p+1, p, p)$: Error (3.24) vs. degrees of freedom

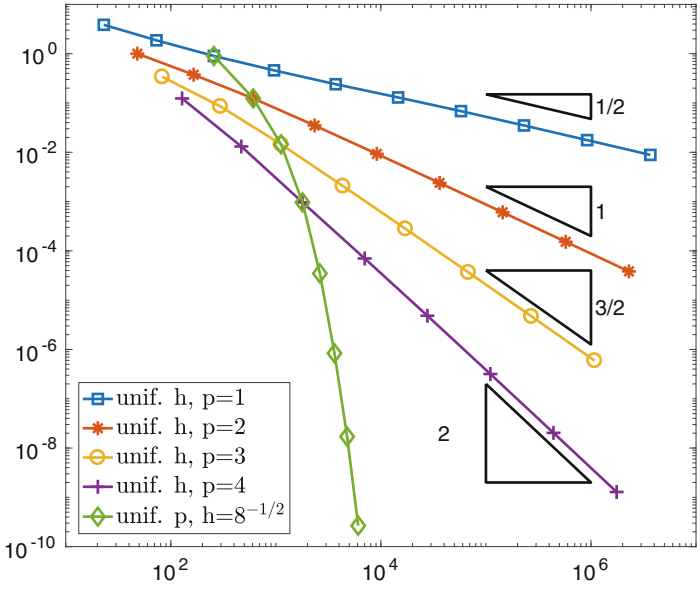


Fig. 3.4 $i = 1$, $(p+1, p, p-1)$: Error (3.24) vs. degrees of freedom

Table 3.1 Experimental order of convergence for uniform h-versions with $p = 1, 2, 3, 4$ for $i = 2$

Type	$\|u - u_{hp}\|_{H^1(\Omega)}$	$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|_{\mathbf{H}^1(\Omega)}$	$\|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{hq})\|_{\mathbf{L}^2(\Omega)}$
$(p+1, p, p)$	1; 1.5; 2; 2.5	0.5; 1; 1.5; 2	0.75; 1; 1.5; 2
$(p+1, p, p-1)$	0.5; 1; 1.5; 2	0.5; 1; 1.5; 2	0.75; 1; 1.5; 2

Table 3.2 Experimental order of convergence for uniform h-versions with $p = 1, 2, 3, 4$ for $i = 1$

Type	$\|u - u_{hp}\|_{H^1(\Omega)}$	$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{hp}\|_{\mathbf{H}(\mathrm{div},\Omega)}$	$\|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{hq})\|_{\mathbf{L}^2(\Omega)}$
$(p+1, p, p)$	1; 1.5; 2; 2.5	1; 1.5; 2; 2.5	1.5; 1.75; 2.4; 2.8
$(p+1, p, p-1)$	0.5; 1; 1.5; 2	0.5; 1; 1.5; 2	1; 1.5; 2; 2.5

for $i = 1$, i.e. the discretization of $H_0^1(\Omega) \times \mathbf{H}_0(\mathrm{div}, \Omega) \times \mathbf{M}$.

In particular, Figs. 3.1 and 3.2 show that for $i = 2$ the uniform h-versions with polynomial degrees $(p+1, p, p)$ and $(p+1, p, p-1)$, $p = 1, 2, 3, 4$, converge with an optimal algebraic rate of $1/2, 1, 3/2, 2$, respectively. Moreover, the uniform p-versions converge exponentially fast as the exact solution is analytic. The convergence rates for the individual variables are stated in Table 3.1. Interestingly, the error in u_{hp} converges at an increased rate for the $(p+1, p, p)$ scheme.

Figures 3.3 and 3.4 show the same quantities as before but for $i = 1$. The uniform h-versions with polynomial degrees $(p+1, p, p)$ converge with the rates $1, 3/2, 2, 5/2$ which is half an order faster than for the $(p+1, p, p-1)$ scheme. Again the uniform p-versions converge exponentially fast. The convergence rates for the individual variables are stated in Table 3.2. Interestingly, the error in $\boldsymbol{\phi}_{hp}$ converges at an increased rate.

Despite the fact that the convergence rates for $\|hp^{-2}(\boldsymbol{\phi} - \boldsymbol{\phi}_{hq})\|_{\mathbf{L}^2(\Omega)}$ are higher than for u_{hp} and $\boldsymbol{\sigma}_{hp}$ in the case of $i = 1$, we observed that this Lagrange multiplier $\boldsymbol{\phi}$ is sensitive to changes in $\overline{\gamma}$. In particular, we noticed that changes in $\overline{\gamma}$, even to 10^{-13}, does not effect the unique solvability, but with smaller $\overline{\gamma}$ it seems that $\boldsymbol{\phi}$ starts to exhibit increasingly checkerboard type oscillations. As the latter is a drawback for the evaluation of an a posteriori error estimate and thus for adaptivity, we do not favor the first method $i = 1$. The second method $i = 2$ is much more robust to changes in $\overline{\gamma}$ and also to the introduction of hanging nodes, and we refer to [6] for an a posteriori error estimate and corresponding adaptive computations.

References

1. Arnold, D.N., Brezzi, F.: Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. ESAIM Math. Model. Numer. Anal. **19**(1), 7–32 (1985)
2. Baker, G.A.: Finite element methods for elliptic equations using nonconforming elements. Math. Comput. **31**(137), 45–59 (1977)

3. Banz, L., Gimperlein, H., Issaoui, A., Stephan, E.P.: Stabilized mixed hp-BEM for frictional contact problems in linear elasticity. Numer. Math. **135**, 217–263 (2017)
4. Banz, L., Lamichhane, B.P., Stephan, E.P.: A new three-field formulation of the biharmonic problem and its finite element discretization. Numer. Methods Partial Differ. Equ. **33**, 199–217 (2017)
5. Banz, L., Milicic, G., Ovcharova, N.: Improved stabilization technique for frictional contact problems solved with hp-BEM. Preprint submitted for publication
6. Banz, L., Petsche, J., Schröder, A.: hp-fem for a stabilized three-field formulation of the biharmonic problem. Comput. Math. Appl. **77**(9), 2463–2488 (2019)
7. Barbosa, H., Hughes, T.: The finite element method with the Lagrange multipliers on the boundary: circumventing the Babuška-Brezzi condition. Comput. Methods Appl. Mech. Eng. **85**, 109–128 (1991)
8. Brenner, S.C., Sung, L.-Y.: C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. J. Sci. Comput. **22–23**, 83–118 (2005)
9. Cheng, X.-L., Han, W., Huang, H.-C.: Some mixed finite element methods for biharmonic equation. J. Comput. Appl. Math. **126**(1–2), 91–109 (2000)
10. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Studies in Mathematics and its Applications. North Holland, Amsterdam (1978)
11. Ciarlet, P., Glowinski, R.: Dual iterative techniques for solving a finite element approximation of the biharmonic equation. Comput. Methods Appl. Mech. Eng. **5**, 277–295 (1975)
12. Comodi, M.I.: The Hellan-Herrmann-Johnson method: some new error estimates and postprocessing. Math. Comput. **52**(185), 17–29 (1989)
13. Da Veiga, L.B., Niiranen, J., Stenberg, R.: A family of C^0 finite elements for Kirchhoff plates I: error analysis. SIAM J. Numer. Anal. **45**(5), 2047–2071 (2007)
14. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems. Society for Industrial and Applied Mathematics, Philadelphia (1999). Unabridged, corrected republication of the 1976 english original edition
15. Engel, G., Garikipati, K., Hughes, T.J.R., Larson, M.G., Mazzei, L., Taylor, R.L.: Continuous/discontinuous finite element approximations of fourth-order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity. Comput. Methods Appl. Mech. Eng. **191**(34), 3669–3750 (2002)
16. Girault, V., Raviart, P.-A.: Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms. Springer, Berlin (1986)
17. Johnson, C., Pitkäranta, J.: Analysis of some mixed finite element methods related to reduced integration. Math. Comput. **38**, 375–400 (1982)
18. Lamichhane, B.P.: A mixed finite element method for the biharmonic problem using biorthogonal or quasi-biorthogonal systems. J. Sci. Comput. **46**(3), 379–396 (2011)
19. Lamichhane, B.P.: A stabilized mixed finite element method for the biharmonic equation based on biorthogonal systems. J. Comput. Appl. Math. **235**(17), 5188–5197 (2011)
20. Melenk, J.: hp-interpolation of nonsmooth functions and an application to hp-a posteriori error estimation. SIAM J. Numer. Anal. **43**(1), 127–155 (2005)
21. Meng, X., Shu, C.-W., Wu, B.: Superconvergence of the local discontinuous Galerkin method for linear fourth-order time-dependent problems in one space dimension. IMA J. Numer. Anal. **32**(4), 1294–1328 (2012)
22. Suri, M.: On the stability and convergence of higher-order mixed finite element methods for second-order elliptic problems. Math. Comput. **54**(189), 1–19 (1990)
23. Wells, G.N., Kuhl, E., Garikipati, K.: A discontinuous Galerkin method for the Cahn-Hilliard equation. J. Comput. Phys. **218**(2), 860–877 (2006)

Chapter 4
Analysis of the *hp*-Version of a First Order System Least Squares Method for the Helmholtz Equation

Maximilian Bernkopf and Jens Markus Melenk

Abstract Extending the wavenumber-explicit analysis of Chen and Qiu (J Comput Appl Math, 309:145–162, 2017), we analyze the L^2-convergence of a least squares method for the Helmholtz equation with wavenumber k. For domains with an analytic boundary, we obtain improved rates in the mesh size h and the polynomial degree p under the scale resolution condition that hk/p is sufficiently small and $p/\log k$ is sufficiently large.

4.1 Introduction

We consider the following Helmholtz problem:

$$\begin{aligned} -\Delta u - k^2 u &= f \quad \text{in } \Omega, \\ \partial_n u - iku &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $k \geq k_0 > 0$ is real. For large k, the numerical solution of (4.1) is challenging due to the requirement to resolve the oscillatory nature of the solution. A second challenge arises in classical, H^1-conforming discretizations of (4.1) from the fact that the Galerkin method is *not* an energy projection, and a meaningful approximation is only obtained under more stringent conditions on the mesh size h and the polynomial degree p than purely approximation theoretical considerations suggest. This shortcoming has been analyzed in the literature. In particular, as discussed in more detail in [6, 20], the analyses [1, 6, 11–13, 19, 20] show that high order methods are much better suited for the high-frequency case of large k than low order methods. Alternatives to the classical Galerkin methods that are still based on high order methods include stabilized methods for Helmholtz [8–10, 28],

M. Bernkopf (✉) · J. M. Melenk
Technische Universität Wien, Institute for Analysis and Scientific Computing, Vienna, Austria
e-mail: maximilian.bernkopf@tuwien.ac.at; melenk@tuwien.ac.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_4

hybridizable methods [4], least-squares type methods [3, 15] and Discontinuous Petrov Galerkin methods, [5, 24]. An attractive feature of least squares type methods is that the resulting linear system is always solvable and that they feature quasi-optimality, albeit in some nonstandard residual norms. In the present paper, we show for the least squares method (4.4) an a priori estimate in the more tractable $L^2(\Omega)$-norm under the scale resolution condition (4.35). For that, we closely follow [3]. Our key refinement over [3] is an improved regularity estimate for the solution of a suitable dual problem (cf. Lemma 4.1 vs. [3, Lemma 5.1]) that allows us to establish the improved p-dependence in the $L^2(\Omega)$-error estimate (cf. Theorem 4.4 vs. [3, Thm. 2.5]). As a tool, which is of independent interest, we develop approximation operators in Raviart-Thomas and Brezzi-Douglas-Marini spaces with optimal (in h and p) approximation rates simultaneously in $L^2(\Omega)$ and $\boldsymbol{H}(\mathrm{div}, \Omega)$.

Throughout this paper, if not otherwise stated, we assume the following:

Assumption 4.1 *In spatial dimension $d = 2, 3$ the bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ has an analytic boundary. The wavenumber k satisfies $k \geq k_0 > 0$. Furthermore $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$.*

Remark 4.1 Under Assumption 4.1 we may apply [2, Thm. 1.8] to conclude that the solution $u \in H^1(\Omega)$ satisfies the a priori bound

$$\|u\|_{H^1(\Omega)} + k\,\|u\|_{L^2(\Omega)} \leq C(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}), \tag{4.2}$$

where $C > 0$ is independent of k. □

Notation and Preliminaries Boldface letters like $\boldsymbol{V}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\Pi}$ will be reserved for quantities having more than one spatial dimensions, while normal letters like W, u and Π will be used for quantities with one spatial dimension. The reference element will be denoted by $\widehat{K}$, whereas the physical one will just be denoted by K. In a similar way, we will distinguish between objects associated with the reference element and the physical one. A function defined on the reference element $\widehat{K}$ will therefore be denoted by $\hat{u}$, while a function defined on the physical element K will be denoted by u. We will follow the same convention when it comes to operators acting on a function space. Therefore operators acting on functions defined on $\widehat{K}$ or K will be denoted by $\widehat{\Pi}$ or Π respectively. Generic constants will either be denoted by C or hidden inside a $\lesssim$ and will be independent of the wavenumber k, the mesh size h and the polynomial degree p, if not otherwise stated.

Outline The outline of this paper is as follows. In Sect. 4.2 we introduce the first order system least squares (FOSLS) method itself, followed by Sect. 4.3, where we prove a refined duality argument (Lemma 4.1), which is later used to derive an a priori estimate (Theorem 4.4) of the method. Key ingredients are the results of [22], where a frequency explicit splitting of the solution to (4.1) is performed when the data has higher order Sobolev regularity. Section 4.4 is concerned with the approximation properties of Raviart-Thomas and Brezzi-Douglas-Marini spaces. We therefore follow the methodology of [19] in order to construct approximation operators, which are not only p-optimal and approximate simultaneous in $L^2(\Omega)$

and $H^1(\Omega)$, but also admit an elementwise construction. Section 4.5 is then devoted to the a priori estimate. Concluding, we give numerical examples which complement the theoretical findings and compare the method to the classical FEM in Sect. 4.6.

4.2 First Order System Least Squares Method and Useful Results

In the present section we introduce the method of [3] and list some useful results which are used later in the paper.

4.2.1 First Order System Least Squares

We employ the complex Hilbert spaces

$$\boldsymbol{V} = \{\boldsymbol{\varphi} \in \boldsymbol{H}(\mathrm{div}, \Omega) \colon \boldsymbol{\varphi} \cdot \boldsymbol{n} \in L^2(\partial\Omega)\} \quad \text{and} \quad W = H^1(\Omega),$$

where $\boldsymbol{V}$ is endowed with the usual graph norm and W with the classical $H^1(\Omega)$-norm. On $\boldsymbol{V} \times W$ we introduce the bilinear form b and the linear functional F by

$$\begin{aligned} b((\boldsymbol{\varphi}, u), (\boldsymbol{\psi}, v)) &\coloneqq (ik\boldsymbol{\varphi} + \nabla u, ik\boldsymbol{\psi} + \nabla v)_\Omega + (iku + \nabla \cdot \boldsymbol{\varphi}, ikv + \nabla \cdot \boldsymbol{\psi})_\Omega + \\ &\qquad k(\boldsymbol{\varphi} \cdot \boldsymbol{n} + u, \boldsymbol{\psi} \cdot \boldsymbol{n} + v)_{\partial\Omega}, \\ F((\boldsymbol{\psi}, v)) &\coloneqq (-ik^{-1} f, ikv + \nabla \cdot \boldsymbol{\psi})_\Omega + (ig, \boldsymbol{\psi} \cdot \boldsymbol{n} + v)_{\partial\Omega}, \end{aligned}$$

where $(u, v)_\Omega = \int_\Omega u\overline{v}\,\mathrm{d}x$. If $u \in H^1(\Omega)$ is the weak solution to (4.1) then the pair $(\boldsymbol{\varphi}, u)$ with $\boldsymbol{\varphi} = ik^{-1}\nabla u$ is in fact in $\boldsymbol{V} \times W$ due to the assumed regularity of the data and the domain and therefore satisfies

$$b((\boldsymbol{\varphi}, u), (\boldsymbol{\psi}, v)) = F((\boldsymbol{\psi}, v)) \quad \forall (\boldsymbol{\psi}, v) \in \boldsymbol{V} \times W. \tag{4.3}$$

For a given regular mesh $\mathscr{T}_h$ we consider the finite element spaces $\boldsymbol{V}_h = \mathbf{RT}_p(\mathscr{T}_h) \subset \boldsymbol{V}$ or $\boldsymbol{V}_h = \mathbf{BDM}_p(\mathscr{T}_h) \subset \boldsymbol{V}$ and $W_h = S_p(\mathscr{T}_h) \subset W$, where $\mathbf{RT}_p(\mathscr{T}_h)$ denotes the Raviart-Thomas space and $\mathbf{BDM}_p(\mathscr{T}_h)$ the Brezzi-Douglas-Marini space; see Sect. 4.4 for further detail and definitions. The FOSLS method is to find $(\boldsymbol{\varphi}_h, u_h) \in \boldsymbol{V}_h \times W_h$ such that

$$b((\boldsymbol{\varphi}_h, u_h), (\boldsymbol{\psi}_h, v_h)) = F((\boldsymbol{\psi}_h, v_h)) \quad \forall (\boldsymbol{\psi}_h, v_h) \in \boldsymbol{V}_h \times W_h. \tag{4.4}$$

Remark 4.2 Based on the a priori estimate (4.2) reference [3, Thm. 2.4] asserts the existence of $C > 0$ independent of k such that

$$\|\boldsymbol{\varphi}\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)} + k\,\|\boldsymbol{\varphi}\cdot\boldsymbol{n} + u\|^2_{L^2(\partial\Omega)} \le Cb((\boldsymbol{\varphi},u),(\boldsymbol{\varphi},u)), \quad \forall (\boldsymbol{\varphi},u)\in \boldsymbol{V}\times W,$$

which immediately gives uniqueness. Together with the fact that the pair $(\boldsymbol{\varphi}, u)$ with $\boldsymbol{\varphi} = ik^{-1}\nabla u$ is a solution, we have unique solvability of (4.3). □

4.2.2 Auxiliary Results

We will need the following decomposition result for the refined duality argument in Lemma 4.1.

Proposition 4.1 ([22, Thm. 4.5] Combined with [2, Thm. 1.8]) *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded Lipschitz domain with an analytic boundary. Fix $s \in \mathbb{N}_0$. Then there exist constants $C, \gamma > 0$ independent of k such that for every $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$ the solution $u = S_k(f,g)$ of (4.1) can be written as $u = u_A + u_{H^{s+2}}$, where, for all $n \in \mathbb{N}_0$, there holds*

$$\|u_A\|_{H^1(\Omega)} + k\,\|u_A\|_{L^2(\Omega)} \le C(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}), \tag{4.5}$$

$$\left\|\nabla^{n+2} u_A\right\|_{L^2(\Omega)} \le C\gamma^n k^{-1} \max\{n,k\}^{n+2}(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}), \tag{4.6}$$

$$\left\|u_{H^{s+2}}\right\|_{H^{s+2}(\Omega)} + k^{s+2}\left\|u_{H^{s+2}}\right\|_{L^2(\Omega)} \le C(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}). \tag{4.7}$$

Remark 4.3 Interpolation between $L^2(\Omega)$ and $H^{s+2}(\Omega)$ in Proposition 4.1 gives estimates for lower order Sobolev norms: Since we have for any $v \in H^m(\Omega)$

$$\|v\|_{H^j(\Omega)} \le C\,\|v\|^{\frac{j}{m}}_{H^m(\Omega)}\,\|v\|^{\frac{m-j}{m}}_{L^2(\Omega)}, \qquad j \in \{0,\dots,m\},$$

Proposition 4.1 implies for $j \in \{0,\dots,s+2\}$

$$k^{s+2-j}\left\|u_{H^{s+2}}\right\|_{H^j(\Omega)} \le C(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}).$$

□

Furthermore we often use the multiplicative trace inequality. We remind the reader of the general form, even though we only need it in the special case $s = 1$.

Proposition 4.2 ([17, Thm. A.2]) *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and $s \in (1/2, 1]$. Then there exists a constant $C > 0$ such that for all $u \in H^s(\Omega)$ there holds*

$$\|u\|_{L^2(\partial\Omega)} \leq C \|u\|_{L^2(\Omega)}^{1-1/(2s)} \|u\|_{H^s(\Omega)}^{1/(2s)},$$

where the left-hand side is understood in the trace sense.

4.3 Duality Argument

We extend the results of [3, Lemma 5.1] by showing that the function $\boldsymbol{\psi}_{H^2} \in \boldsymbol{H}^1(\mathrm{div}, \Omega)$, constructed therein, can actually be modified to satisfy $\boldsymbol{\psi}_{H^2} \in \boldsymbol{H}^2(\Omega)$ and still allow for wavenumber-explicit higher order Sobolev norm estimates.

Lemma 4.1 *For any $(\boldsymbol{\varphi}, w) \in \boldsymbol{V} \times W$ there exists $(\boldsymbol{\psi}, v) \in \boldsymbol{V} \times W$ such that $\|w\|_{L^2(\Omega)}^2 = b((\boldsymbol{\varphi}, w), (\boldsymbol{\psi}, v))$. The pair $(\boldsymbol{\psi}, v)$ admits a decomposition $\boldsymbol{\psi} = \boldsymbol{\psi}_A + \boldsymbol{\psi}_{H^2}$, $v = v_A + v_{H^2}$, where $\boldsymbol{\psi}_A$ and v_A are analytic in Ω, $\boldsymbol{\psi}_{H^2} \in \boldsymbol{H}^2(\Omega)$, and $v_{H^2} \in H^2(\Omega)$. Furthermore there exist constants $C, \gamma > 0$ independent of k such that for all $n \in \mathbb{N}_0$*

$$\|\boldsymbol{\psi}_A\|_{H^1(\Omega)} + k \|\boldsymbol{\psi}_A\|_{L^2(\Omega)} \leq Ck \|w\|_{L^2(\Omega)}, \tag{4.8}$$

$$\|v_A\|_{H^1(\Omega)} + k \|v_A\|_{L^2(\Omega)} \leq Ck \|w\|_{L^2(\Omega)}, \tag{4.9}$$

$$\left\|\nabla^{n+2}\boldsymbol{\psi}_A\right\|_{L^2(\Omega)} + \left\|\nabla^{n+2} v_A\right\|_{L^2(\Omega)} \leq C\gamma^n \max\{n, k\}^{n+2} \|w\|_{L^2(\Omega)}, \tag{4.10}$$

$$\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^2(\Omega)} + k \left\|\boldsymbol{\psi}_{H^2}\right\|_{H^1(\Omega)} + k^2 \left\|\boldsymbol{\psi}_{H^2}\right\|_{L^2(\Omega)} \leq C \|w\|_{L^2(\Omega)}, \tag{4.11}$$

$$\left\|v_{H^2}\right\|_{H^2(\Omega)} + k \left\|v_{H^2}\right\|_{H^1(\Omega)} + k^2 \left\|v_{H^2}\right\|_{L^2(\Omega)} \leq C \|w\|_{L^2(\Omega)}. \tag{4.12}$$

Proof The proof follows the ideas of [3, Lemma 5.1]; for the readers' convenience we recapitulate the important steps of the proof. The novelty over [3] is the ability to choose $\boldsymbol{\psi}_{H^2} \in \boldsymbol{H}^2(\Omega)$ together with $\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^2(\Omega)} \leq C \|w\|_{L^2(\Omega)}$.

Consider the problem

$$\begin{aligned} -\Delta z - k^2 z &= w \qquad \text{in } \Omega, \\ \partial_n z + ikz &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

For any $\boldsymbol{\varphi} \in \boldsymbol{V}$ we have, using the weak formulation and integrating by parts,

$$\begin{aligned}\|w\|^2_{L^2(\Omega)} &= (\nabla w, \nabla z)_\Omega - k^2(w,z)_\Omega - ik(w,z)_{\partial\Omega} \\ &= (ik\boldsymbol{\varphi} + \nabla w, \nabla z)_\Omega - (ik\boldsymbol{\varphi}, \nabla z)_\Omega - k^2(w,z)_\Omega - ik(w,z)_{\partial\Omega} \\ &= (ik\boldsymbol{\varphi} + \nabla w, \nabla z)_\Omega + (\nabla\cdot\boldsymbol{\varphi} + ikw, -ikz)_\Omega + (\boldsymbol{\varphi}\cdot\boldsymbol{n} + w, ikz)_{\partial\Omega}.\end{aligned}$$

Applying Proposition 4.1 together with Remark 4.3 we decompose z into $z = z_A + z_{H^2}$ with z_A analytic and $z_{H^2} \in H^2(\Omega)$. Furthermore we have, for all $n \in \mathbb{N}_0$,

$$\|z_A\|_{H^1(\Omega)} + k\,\|z_A\|_{L^2(\Omega)} \le C\,\|w\|_{L^2(\Omega)}, \tag{4.13}$$

$$\left\|\nabla^{n+2} z_A\right\|_{L^2(\Omega)} \le C\gamma^n k^{-1} \max\{n,k\}^{n+2}\,\|w\|_{L^2(\Omega)}, \tag{4.14}$$

$$\left\|z_{H^2}\right\|_{H^2(\Omega)} + k\left\|z_{H^2}\right\|_{H^1(\Omega)} + k^2\left\|z_{H^2}\right\|_{L^2(\Omega)} \le C\,\|w\|_{L^2(\Omega)}. \tag{4.15}$$

Let $(\boldsymbol{\psi}, v) \in \boldsymbol{V} \times W$ solve

$$\begin{aligned} ik\boldsymbol{\psi} + \nabla v &= \nabla z && \text{in } \Omega, \\ ikv + \nabla\cdot\boldsymbol{\psi} &= -ikz && \text{in } \Omega, \\ k^{1/2}(\boldsymbol{\psi}\cdot\boldsymbol{n} + v) &= ik^{1/2} z && \text{on } \partial\Omega. \end{aligned}$$

Indeed, this system is uniquely solvable by Remark 4.2. This gives the desired representation such that $\|w\|^2_{L^2(\Omega)} = b((\boldsymbol{\varphi}, w), (\boldsymbol{\psi}, v))$. Using the decomposition $z = z_A + z_{H^2}$ we obtain $(\boldsymbol{\psi}, v) = (\tilde{\boldsymbol{\psi}}_A, \tilde{v}_A) + (\tilde{\boldsymbol{\psi}}_{H^2}, \tilde{v}_{H^2})$, where

$$\begin{aligned} ik\tilde{\boldsymbol{\psi}}_A + \nabla\tilde{v}_A &= \nabla z_A && \text{in } \Omega, & \qquad ik\tilde{\boldsymbol{\psi}}_{H^2} + \nabla\tilde{v}_{H^2} &= \nabla z_{H^2} && \text{in } \Omega, \\ ik\tilde{v}_A + \nabla\cdot\tilde{\boldsymbol{\psi}}_A &= -ikz_A && \text{in } \Omega, & \qquad ik\tilde{v}_{H^2} + \nabla\cdot\tilde{\boldsymbol{\psi}}_{H^2} &= -ikz_{H^2} && \text{in } \Omega, \\ k^{1/2}(\tilde{\boldsymbol{\psi}}_A\cdot\boldsymbol{n} + \tilde{v}_A) &= ik^{1/2} z_A && \text{on } \partial\Omega, & \qquad k^{1/2}(\tilde{\boldsymbol{\psi}}_{H^2}\cdot\boldsymbol{n} + \tilde{v}_{H^2}) &= ik^{1/2} z_{H^2} && \text{on } \partial\Omega. \end{aligned}$$

One can immediately verify that

$$\begin{aligned} -\Delta(\tilde{v}_A - z_A) - k^2(\tilde{v}_A - z_A) &= 2k^2 z_A && \text{in } \Omega, \\ \partial_n(\tilde{v}_A - z_A) - ik(\tilde{v}_A - z_A) &= (1+i)kz_A && \text{on } \partial\Omega, \end{aligned} \tag{4.16}$$

as well as

$$\begin{aligned} -\Delta(\tilde{v}_{H^2} - z_{H^2}) - k^2(\tilde{v}_{H^2} - z_{H^2}) &= 2k^2 z_{H^2} && \text{in } \Omega, \\ \partial_n(\tilde{v}_{H^2} - z_{H^2}) - ik(\tilde{v}_{H^2} - z_{H^2}) &= (1+i)kz_{H^2} && \text{on } \partial\Omega. \end{aligned} \tag{4.17}$$

Note that the right-hand sides in Eq. (4.16) are analytic. This fact is used in [3, Lemma 5.1, Lemma 4.4] to prove the following bounds for all $n \in \mathbb{N}_0$:

$$\left\| \nabla^{n+2} \tilde{v}_A \right\|_{L^2(\Omega)} \leq C\gamma^n \max\{n,k\}^{n+2} \|w\|_{L^2(\Omega)}, \tag{4.18}$$

$$\|\tilde{v}_A\|_{H^1(\Omega)} + k \|\tilde{v}_A\|_{L^2(\Omega)} \leq Ck \|w\|_{L^2(\Omega)}, \tag{4.19}$$

$$\left\| \nabla^{n+2} \tilde{\boldsymbol{\psi}}_A \right\|_{L^2(\Omega)} \leq C\gamma^n \max\{n,k\}^{n+2} \|w\|_{L^2(\Omega)}, \tag{4.20}$$

$$\left\| \tilde{\boldsymbol{\psi}}_A \right\|_{H^1(\Omega)} + k \left\| \tilde{\boldsymbol{\psi}}_A \right\|_{L^2(\Omega)} \leq Ck \|w\|_{L^2(\Omega)}. \tag{4.21}$$

Since $\tilde{v}_{H^2} - z_{H^2} = S_k(2k^2 z_{H^2}, (1+i)k z_{H^2})$, where S_k denotes the solution operator for (4.1), we can exploit the regularity of the right-hand sides in Eq. (4.17). Applying Proposition 4.1 with $s = 1$ as well as Remark 4.3 we decompose $\tilde{v}_{H^2} - z_{H^2} = \hat{v}_A + \hat{v}_{H^3}$, where $\hat{v}_A$ is analytic and $\hat{v}_{H^3} \in H^3(\Omega)$. For every $j \in \{0,1,2,3\}$ we have

$$\begin{aligned} k^{3-j} \left\| \hat{v}_{H^3} \right\|_{H^j(\Omega)} &\lesssim \left\| 2k^2 z_{H^2} \right\|_{H^1(\Omega)} + \left\| (1+i)k z_{H^2} \right\|_{H^{3/2}(\partial\Omega)} \\ &\lesssim \underbrace{k^2 \left\| z_{H^2} \right\|_{H^1(\Omega)}}_{\overset{(4.15)}{\lesssim} k\|w\|_{L^2(\Omega)}} + \underbrace{k \left\| z_{H^2} \right\|_{H^{3/2}(\partial\Omega)}}_{\lesssim k\|z_{H^2}\|_{H^2(\Omega)} \overset{(4.15)}{\lesssim} k\|w\|_{L^2(\Omega)}} \\ &\lesssim k \|w\|_{L^2(\Omega)}. \end{aligned}$$

Summarizing the above we have

$$k^{-1} \left\| \hat{v}_{H^3} \right\|_{H^3(\Omega)} + \left\| \hat{v}_{H^3} \right\|_{H^2(\Omega)} + k \left\| \hat{v}_{H^3} \right\|_{H^1(\Omega)} + k^2 \left\| \hat{v}_{H^3} \right\|_{L^2(\Omega)} \leq C \|w\|_{L^2(\Omega)}. \tag{4.22}$$

In order to analyze the behavior of $\hat{v}_A$ we first estimate

$$\left\| 2k^2 z_{H^2} \right\|_{L^2(\Omega)} + \left\| (1+i)k z_{H^2} \right\|_{H^{1/2}(\partial\Omega)} \overset{(4.15)}{\lesssim} \|w\|_{L^2(\Omega)}.$$

We therefore conclude, again with Proposition 4.1, that

$$\left\| \hat{v}_A \right\|_{H^1(\Omega)} + k \left\| \hat{v}_A \right\|_{L^2(\Omega)} \leq C \|w\|_{L^2(\Omega)}, \tag{4.23}$$

$$\left\| \nabla^{n+2} \hat{v}_A \right\|_{L^2(\Omega)} \leq C\gamma^n k^{-1} \max\{n,k\}^{n+2} \|w\|_{L^2(\Omega)}. \tag{4.24}$$

We turn to the final decompositions with associated norm bounds.

Final Decomposition of v

$$v = \tilde{v}_A + \tilde{v}_{H^2} = \tilde{v}_A + \underbrace{\tilde{v}_{H^2} - z_{H^2}}_{=\hat{v}_A + \hat{v}_{H^3}} + z_{H^2} = \underbrace{\tilde{v}_A + \hat{v}_A}_{=:v_A} + \underbrace{\hat{v}_{H^3} + z_{H^2}}_{=:v_{H^2}} .$$

Verification of (4.9)

$$\begin{aligned}\|v_A\|_{H^1(\Omega)} + k\|v_A\|_{L^2(\Omega)} &\le \underbrace{\|\tilde{v}_A\|_{H^1(\Omega)} + k\|\tilde{v}_A\|_{L^2(\Omega)}}_{\overset{(4.19)}{\le} Ck\|w\|_{L^2(\Omega)}} + \underbrace{\|\hat{v}_A\|_{H^1(\Omega)} + k\|\hat{v}_A\|_{L^2(\Omega)}}_{\overset{(4.23)}{\le} C\|w\|_{L^2(\Omega)}} \\ &\le Ck\|w\|_{L^2(\Omega)} .\end{aligned}$$

Verification of (4.12)

$$\begin{aligned}&\|v_{H^2}\|_{H^2(\Omega)} + k\|v_{H^2}\|_{H^1(\Omega)} + k^2\|v_{H^2}\|_{L^2(\Omega)} \\ &\quad\le \underbrace{\|\hat{v}_{H^3}\|_{H^2(\Omega)} + k\|\hat{v}_{H^3}\|_{H^1(\Omega)} + k^2\|\hat{v}_{H^3}\|_{L^2(\Omega)}}_{\overset{(4.22)}{\le} C\|w\|_{L^2(\Omega)}} \\ &\qquad + \underbrace{\|z_{H^2}\|_{H^2(\Omega)} + k\|z_{H^2}\|_{H^1(\Omega)} + k^2\|z_{H^2}\|_{L^2(\Omega)}}_{\overset{(4.15)}{\le} C\|w\|_{L^2(\Omega)}} \\ &\quad\le C\|w\|_{L^2(\Omega)} .\end{aligned}$$

Final Decomposition of $\boldsymbol{\psi}$ Since $-ik\tilde{\boldsymbol{\psi}}_{H^2} = \nabla(\tilde{v}_{H^2} - z_{H^2}) = \nabla\hat{v}_A + \nabla\hat{v}_{H^3}$, we decompose $\tilde{\boldsymbol{\psi}}_{H^2} = \hat{\boldsymbol{\psi}}_A + \hat{\boldsymbol{\psi}}_{H^2}$ accordingly such that $-ik\hat{\boldsymbol{\psi}}_A = \nabla\hat{v}_A$ and consequently $-ik\hat{\boldsymbol{\psi}}_{H^2} = \nabla\hat{v}_{H^3}$. The final decomposition takes the form

$$\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}_A + \tilde{\boldsymbol{\psi}}_{H^2} = \underbrace{\tilde{\boldsymbol{\psi}}_A + \hat{\boldsymbol{\psi}}_A}_{=:\boldsymbol{\psi}_A} + \underbrace{\hat{\boldsymbol{\psi}}_{H^2}}_{=:\boldsymbol{\psi}_{H^2}} .$$

Verification of (4.8)

$$\begin{aligned}&\|\boldsymbol{\psi}_A\|_{H^1(\Omega)} + k\|\boldsymbol{\psi}_A\|_{L^2(\Omega)} \\ &\quad\le \underbrace{\|\tilde{\boldsymbol{\psi}}_A\|_{H^1(\Omega)} + k\|\tilde{\boldsymbol{\psi}}_A\|_{L^2(\Omega)}}_{\overset{(4.21)}{\le} Ck\|w\|_{L^2(\Omega)}} + \|\hat{\boldsymbol{\psi}}_A\|_{H^1(\Omega)} + k\|\hat{\boldsymbol{\psi}}_A\|_{L^2(\Omega)} \\ &\quad\le Ck\|w\|_{L^2(\Omega)} + k^{-1}\|\nabla\hat{v}_A\|_{H^1(\Omega)} + \|\nabla\hat{v}_A\|_{L^2(\Omega)}\end{aligned}$$

$$\leq Ck\|w\|_{L^2(\Omega)} + k^{-1}\underbrace{\|\hat{v}_A\|_{H^1(\Omega)}}_{\overset{(4.23)}{\leq} C\|w\|_{L^2(\Omega)}} + k^{-1}\underbrace{\left\|\nabla^2\hat{v}_A\right\|_{L^2(\Omega)}}_{\overset{(4.24)}{\leq} Ck\|w\|_{L^2(\Omega)}} + \underbrace{\|\hat{v}_A\|_{H^1(\Omega)}}_{\overset{(4.23)}{\leq} C\|w\|_{L^2(\Omega)}}$$

$$\leq Ck\|w\|_{L^2(\Omega)}.$$

Verification of (4.10) This is an immediate consequence of (4.18), (4.20), (4.24), and the fact that $-ik\hat{\boldsymbol{\psi}}_A = \nabla\hat{v}_A$.

Verification of (4.11) Since $-ik\hat{\boldsymbol{\psi}}_{H^2} = \nabla\hat{v}_{H^3}$ we estimate

$$\begin{aligned}
&\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^2(\Omega)} + k\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^1(\Omega)} + k^2\left\|\boldsymbol{\psi}_{H^2}\right\|_{L^2(\Omega)} \\
&\qquad = k^{-1}\left\|\nabla\hat{v}_{H^3}\right\|_{H^2(\Omega)} + \left\|\nabla\hat{v}_{H^3}\right\|_{H^1(\Omega)} + k\left\|\nabla\hat{v}_{H^3}\right\|_{L^2(\Omega)} \\
&\qquad \leq \underbrace{k^{-1}\left\|\hat{v}_{H^3}\right\|_{H^3(\Omega)} + \left\|\hat{v}_{H^3}\right\|_{H^2(\Omega)} + k\left\|\hat{v}_{H^3}\right\|_{H^1(\Omega)}}_{\overset{(4.22)}{\leq} C\|w\|_{L^2(\Omega)}} \\
&\qquad \leq C\|w\|_{L^2(\Omega)},
\end{aligned}$$

which concludes the proof. □

4.4 Approximation Properties of Raviart-Thomas and Brezzi-Douglas-Marini Spaces

In the present section we analyze the approximation properties of Raviart-Thomas and Brezzi-Douglas-Marini spaces. To that end, we first state some standard assumptions on the mesh and recall the relevant function spaces. Next, we will prove the existence of a polynomial approximation operator acting on functions defined on the reference element having certain desirable properties, as outlined below. This operator will then be used to construct a global polynomial approximation operator by means of the Piola transformation.

4.4.1 Preliminaries

We start with assumptions on the triangulation.

Assumption 4.2 (Quasi-Uniform Regular Meshes) *Let $\widehat{K}$ be the reference simplex. Each element map $F_K : \widehat{K} \to K$ can be written as $F_K = R_K \circ A_K$, where A_K*

is an affine map and the maps R_K and A_K satisfy, for constants $C_{\mathrm{affine}}, C_{\mathrm{metric}}, \gamma > 0$ independent of K:

$$\left\|A_K'\right\|_{L^\infty(\widehat{K})} \le C_{\mathrm{affine}} h_K, \qquad \left\|(A_K')^{-1}\right\|_{L^\infty(\widehat{K})} \le C_{\mathrm{affine}} h_K^{-1},$$

$$\left\|(R_K')^{-1}\right\|_{L^\infty(\tilde{K})} \le C_{\mathrm{metric}}, \qquad \left\|\nabla^n R_K\right\|_{L^\infty(\tilde{K})} \le C_{\mathrm{metric}} \gamma^n n! \qquad \forall n \in \mathbb{N}_0.$$

Here, $\tilde{K} = A_K(\widehat{K})$ and $h_K > 0$ denotes the element diameter.

We recall the definition of the Sobolev space $H^{1/2}_{00}(\omega)$. If ω is an edge of a triangle or face of a tetrahedron, then the norm $\|\cdot\|_{H^{1/2}_{00}(\omega)}$ is given by

$$\|u\|^2_{H^{1/2}_{00}(\omega)} := \|u\|^2_{H^{1/2}(\omega)} + \left\| \frac{u}{\sqrt{\operatorname{dist}(\cdot, \partial\omega)}} \right\|^2_{L^2(\omega)},$$

and the space $H^{1/2}_{00}(\omega)$ is the completion of $C_0^\infty(\omega)$ under this norm. Since this norm is induced by a scalar product the space $H^{1/2}_{00}(\omega)$ is a Hilbert space.

On the reference element $\widehat{K}$ we introduce the Raviart-Thomas and Brezzi-Douglas-Marini elements of degree $p \ge 0$ in dimension d:

$$\begin{aligned} \mathscr{P}_p(\widehat{K}) &:= \operatorname{span}\left\{\boldsymbol{x}^{\boldsymbol{\alpha}} : |\boldsymbol{\alpha}| \le p\right\}, \\ \mathbf{BDM}_p(\widehat{K}) &:= \mathscr{P}_p(\widehat{K})^d, \\ \mathbf{RT}_p(\widehat{K}) &:= \left\{\boldsymbol{p} + \boldsymbol{x} q : \boldsymbol{p} \in \mathscr{P}_p(\widehat{K})^d, q \in \mathscr{P}_p(\widehat{K})\right\}. \end{aligned}$$

Note that trivially $\mathbf{BDM}_p(\widehat{K}) \subset \mathbf{RT}_p(\widehat{K})$. We also recall the classical Piola transformation, which is the appropriate change of variables for $\boldsymbol{H}(\operatorname{div}, \Omega)$. For a function $\boldsymbol{\varphi} : K \to \mathbb{R}^d$ and the element map $F_K : \widehat{K} \to K$ its Piola transform $\widehat{\boldsymbol{\varphi}} : \widehat{K} \to \mathbb{R}^d$ is given by

$$\widehat{\boldsymbol{\varphi}} = (\det F_K')(F_K')^{-1} \boldsymbol{\varphi} \circ F_K.$$

Furthermore we introduce the spaces $S_p(\mathscr{T}_h)$, $\mathbf{BDM}_p(\mathscr{T}_h)$, and $\mathbf{RT}_p(\mathscr{T}_h)$ by standard transformation and (contravariant) Piola transformation respectively:

$$\begin{aligned} S_p(\mathscr{T}_h) &:= \left\{u \in H^1(\Omega) : u|_K \circ F_K \in \mathscr{P}_p(\widehat{K}) \text{ for all } K \in \mathscr{T}_h\right\}, \\ \mathbf{BDM}_p(\mathscr{T}_h) &:= \left\{\boldsymbol{\varphi} \in \boldsymbol{H}(\operatorname{div}, \Omega) : (\det F_K')(F_K')^{-1} \boldsymbol{\varphi}|_K \circ F_K \in \mathbf{BDM}_p(\widehat{K}) \text{ for all } K \in \mathscr{T}_h\right\}, \\ \mathbf{RT}_p(\mathscr{T}_h) &:= \left\{\boldsymbol{\varphi} \in \boldsymbol{H}(\operatorname{div}, \Omega) : (\det F_K')(F_K')^{-1} \boldsymbol{\varphi}|_K \circ F_K \in \mathbf{RT}_p(\widehat{K}) \text{ for all } K \in \mathscr{T}_h\right\}. \end{aligned}$$

4.4.2 Polynomial Approximation on the Reference Element

We construct a polynomial approximation operator on the reference element $\widehat{K}$:

Definition 4.1 Let $\widehat{K}$ be the reference simplex in $\mathbb{R}^d$, $s > d/2$ and $p \in \mathbb{N}$. We define the operator $\widehat{\Pi}_p : H^s(\widehat{K}) \to \mathscr{P}_p(\widehat{K})$ by the following consecutive minimization steps:

1. Fix $\widehat{\Pi}_p u$ in the vertices: $(\widehat{\Pi}_p u)(\widehat{V}) = u(\widehat{V})$ for all $d+1$ vertices $\widehat{V}$ of $\widehat{K}$.
2. Fix $\widehat{\Pi}_p u$ on the edges: for every edge $\hat{e}$ of $\widehat{K}$ the restriction $(\widehat{\Pi} u_p)\big|_{\hat{e}}$ is the unique minimizer of

$$\mathscr{P}_p(\hat{e}) \ni \pi \mapsto p \, \|u - \pi\|^2_{L^2(\hat{e})} + \|u - \pi\|^2_{H^{1/2}_{00}(\hat{e})}, \qquad \text{s.t. } \pi \text{ satisfies 1.} \tag{4.25}$$

3. Fix $\widehat{\Pi}_p u$ on the faces (only for $d = 3$): for every face $\hat{f}$ of $\widehat{K}$ the restriction $(\widehat{\Pi} u_p)\big|_{\hat{f}}$ is the unique minimizer of

$$\mathscr{P}_p(\hat{f}) \ni \pi \mapsto p^2 \, \|u - \pi\|^2_{L^2(\hat{f})} + \|u - \pi\|^2_{H^1(\hat{f})}, \qquad \text{s.t. } \pi \text{ satisfies 1, 2.} \tag{4.26}$$

4. Fix $\widehat{\Pi}_p u$ in the volume: $\widehat{\Pi}_p u$ is the unique minimizer of

$$\mathscr{P}_p(\widehat{K}) \ni \pi \mapsto p^2 \, \|u - \pi\|^2_{L^2(\widehat{K})} + \|u - \pi\|^2_{H^1(\widehat{K})}, \quad \text{s.t. } \pi \text{ satisfies 1, 2, 3.} \tag{4.27}$$

It is convenient to construct an approximant Iu of a function u in an elementwise fashion. The drawback is that one has to check if the approximant is in fact in the finite element space. A useful property to achieve this is the following: The restriction of the approximant $Iu\big|_E$ to lower dimensional entities E of the mesh, i.e., edges, faces or vertices, is completely determined by the corresponding restriction of u. To put this rigorously, we employ the following concept:

Definition 4.2 (Restriction Property) Let $\widehat{K}$ be the reference simplex in $\mathbb{R}^d$, $s > d/2$, and $p \in \mathbb{N}$. A polynomial $\pi \in \mathscr{P}_p(\widehat{K})$ is said to satisfy the *restriction property* of polynomial degree p for $u \in H^s(\widehat{K})$, if it satisfies 1, 2, 3 of Definition 4.1.

Remark 4.4 Note that minimizations in the definition of the operator $\widehat{\Pi}_p$ are uniquely solvable. This is due to the fact these minimizations are constrained minimizations of norms induced by Hilbert spaces. These constraints are given by an affine subspace $\mathscr{V}_p^u \leq \mathscr{P}_p(\widehat{K})$, the space of all polynomials satisfying the restriction property for u. Step 4 is therefore the orthogonal projection onto the space $\mathscr{V}_p^u$ with respect to the scalar product inducing the norm

$$|||u|||^2 := p^2 \, \|u\|^2_{L^2(\widehat{K})} + \|u\|^2_{H^1(\widehat{K})}.$$

Furthermore the affine space $\mathscr{V}_p^u$ can be written as $\mathscr{V}_p^u = \pi^u + \mathscr{P}_p^0$ for some $\pi^u \in \mathscr{V}_p^u$, where $\mathscr{P}_p^0(\widehat{K}) \leq \mathscr{P}_p(\widehat{K})$ is the space of polynomials vanishing on $\partial\widehat{K}$. The operator $\widehat{\Pi}_p$ can, apart from being the solution to a minimization problem, also be written as:

$$\widehat{\Pi}_p u = \operatorname{argmin}\{|||u - \pi||| : \pi \in \mathscr{V}_p^u\} = \pi^u + \widehat{\Pi}_{\mathscr{P}_p^0}(u - \pi^u), \tag{4.28}$$

where $\widehat{\Pi}_{\mathscr{P}_p^0}$ denotes the orthogonal projection onto the space $\mathscr{P}_p^0(\widehat{K})$, again with respect to the scalar product inducing $|||\cdot|||$. The operator $\widehat{\Pi}_p : H^s(\widehat{K}) \to \mathscr{P}_p(\widehat{K})$ is furthermore linear. This is easily seen when one explicitly constructs the Steps 1, 2, 3 in Definition 4.1: First, one picks polynomials $\pi_{\widehat{V}}$, which are 1 at the vertex $\widehat{V}$ and zero on all the others. Consider the mapping $\widehat{\Pi}_{\widehat{V}} : u \mapsto \sum_{\widehat{V}} u(\widehat{V})\pi_{\widehat{V}}$. This realizes Step 1. Next one considers the mapping $\tilde{\Pi}_{\hat{e}} : z \mapsto \operatorname{argmin}\{p\,\|u - \pi\|^2_{L^2(\hat{e})} + \|u - \pi\|^2_{H_{00}^{1/2}(\hat{e})} : z(\widehat{V}) = 0 \text{ for all vertices } \widehat{V}\}$ and extending it to the reference element. Step 2 is then realized by the map $\widehat{\Pi}_{\hat{e}} : u \mapsto \widehat{\Pi}_{\widehat{V}} u + \tilde{\Pi}_{\hat{e}}(u - \widehat{\Pi}_{\widehat{V}} u)$. One can easily continue this procedure for Step 3 and 4. As a composition of linear operators $\widehat{\Pi}_p$ is therefore also linear. □

Remark 4.5 Definition 4.2 of the restriction property was introduced in [19, Definition 5.3] under the name *element-by-element construction*. This is due to the fact that, when working in $S_p(\mathscr{T}_h) \leq H^1(\Omega)$, a polynomial, which is constructed in an elementwise fashion on the reference simplex $\widehat{K}$, satisfying the restriction property is already an element of the conforming element space $S_p(\mathscr{T}_h)$. However, when working in $\boldsymbol{H}(\mathrm{div}, \Omega)$ or $\boldsymbol{H}(\mathrm{curl}, \Omega)$ one only needs continuity of the inter element normal or tangential trace. Furthermore it is necessary to use the Piola transformation to go back and forth between the reference element and the physical element to ensure that normal and tangential vectors are mapped appropriately. For the purpose of this paper we therefore use the name restriction property, rather than element-by-element construction. □

In Propositions 4.3, 4.4, and 4.5 we recall certain useful results concerning approximation properties of polynomials satisfying the restriction property. These results can be found in [19].

Proposition 4.3 ([19, Thm. B.4]) *Let $\widehat{K}$ be the reference triangle or reference tetrahedron. Let $s > d/2$. Then there exists $C > 0$ (depending only on s and d) and for every p a linear operator $\widehat{\Pi}_p^{\mathrm{MS}} : H^s(\widehat{K}) \to \mathscr{P}_p(\widehat{K})$, such that $\widehat{\Pi}_p^{\mathrm{MS}} u$ satisfies the restriction property of Definition 4.2 as well as*

$$p \left\| u - \widehat{\Pi}_p^{\mathrm{MS}} u \right\|_{L^2(\widehat{K})} + \left\| u - \widehat{\Pi}_p^{\mathrm{MS}} u \right\|_{H^1(\widehat{K})} \leq C p^{-(s-1)} |u|_{H^s(\widehat{K})} \qquad \forall p \geq s - 1. \tag{4.29}$$

Remark 4.6 The operator $\widehat{\Pi}_p^{\mathrm{MS}}$ does in general not preserve polynomials $q \in \mathscr{P}_p(\widehat{K})$. See also [18] for operators with the projection property. □

Proposition 4.4 ([19, Lemma C.2]) *Let $d \in \{1, 2, 3\}$, and let $\widehat{K} \subset \mathbb{R}^d$ be the reference simplex. Let $\gamma, \tilde{C} > 0$ be given. Then there exist constants $C, \sigma > 0$ that depend solely on γ and $\tilde{C}$ such that the following is true: For any function u that satisfies for some C_u, h, $R > 0$ and $\kappa > 1$ the conditions*

$$\left\| \nabla^n u \right\|_{L^2(\widehat{K})} \leq C_u (\gamma h)^n \max\{n/R, \kappa\}^n \qquad \forall n \in \mathbb{N}_{\geq 2},$$

and for any polynomial degree $p \in \mathbb{N}$ that satisfies

$$\frac{h}{R} + \frac{\kappa h}{p} \leq \tilde{C}$$

there holds

$$\inf_{\pi \in \mathscr{P}_p(\widehat{K})} \|u - \pi\|_{W^{2,\infty}(\widehat{K})} \leq C C_u \left[\left(\frac{h/R}{\sigma + h/R} \right)^{p+1} + \left(\frac{h\kappa}{\sigma p} \right)^{p+1} \right].$$

Proposition 4.5 ([19, Lemma C.3]) *Assume the hypotheses of Proposition 4.4. Then one can find a polynomial $\pi \in \mathscr{P}_p(\widehat{K})$ that satisfies*

$$\|u - \pi\|_{W^{1,\infty}(\widehat{K})} \leq C C_u \left[\left(\frac{h/R}{\sigma + h/R} \right)^{p+1} + \left(\frac{h\kappa}{\sigma p} \right)^{p+1} \right].$$

and additionally satisfies the restriction property of Definition 4.2.

It is not clear whether the polynomial $\widehat{\Pi}_p^{\mathrm{MS}} u$ has the same approximation properties as the polynomial given by Proposition 4.5. However, it is desirable to have both the simultaneous approximation properties in $L^2(\widehat{K})$ and $H^1(\widehat{K})$ as stated in Proposition 4.3 as well as the exponential approximation properties of an analytic function as stated in Proposition 4.5. In the following we will show that the operator $\widehat{\Pi}_p$ constructed in Definition 4.1 has these properties.

Theorem 4.3 (Properties of $\widehat{\Pi}_p$) *Let $\widehat{K}$ be the reference triangle or reference tetrahedron. Let $s > d/2$. Let $\widehat{\Pi}_p \colon H^s(\widehat{K}) \to \mathscr{P}_p(\widehat{K})$ be given by Definition 4.1. Then the following holds:*

(i) The operator $\widehat{\Pi}_p$ is linear and satisfies the restriction property of Definition 4.2.

(ii) The operator $\widehat{\Pi}_p$ preserves $\mathscr{P}_p(\widehat{K})$, i.e., $\widehat{\Pi}_p q = q$ for all $q \in \mathscr{P}_p(\widehat{K})$.

(iii) There exists $C_s > 0$ (depending only on s and d) such that

$$p \left\| u - \widehat{\Pi}_p u \right\|_{L^2(\widehat{K})} + \left\| u - \widehat{\Pi}_p u \right\|_{H^1(\widehat{K})} \leq C_s p^{-(s-1)} |u|_{H^s(\widehat{K})} \qquad \forall p \geq s - 1.$$

(iv) For given γ, $\tilde{C} > 0$, there exist constants C_A, $\sigma > 0$ that depend solely on γ and $\tilde{C}$ such that the following is true: For any function u and polynomial

degree p that satisfy the assumptions of Proposition 4.4 there holds

$$\left\|u-\widehat{\Pi}_p u\right\|_{W^{1,\infty}(\widehat{K})} \leq C_A C_u \left[\left(\frac{h/R}{\sigma+h/R}\right)^{p+1}+\left(\frac{h\kappa}{\sigma p}\right)^{p+1}\right].$$

Idea The crucial points of Theorem 4.3 are items (iii) and (iv). To verify (iii) we will exploit the approximation properties of $\widehat{\Pi}_p^{\mathrm{MS}}$ given by Proposition 4.3 together with the fact that $\widehat{\Pi}_p u$ is the solution to a minimization problem. To prove (iv) we use the affine projection representation (4.28) of $\widehat{\Pi}_p$ together with the approximation properties of polynomials satisfying the restriction property given in Proposition 4.5.

Proof Assertion (i) is trivially satisfied due to the construction in Definition 4.1 and Remark 4.4.
Assertion (ii) is also trivially satisfied, since for a given polynomial $q \in \mathscr{P}_p(\widehat{K})$ the norms in Definition 4.1 are minimized at q.
To prove Assertion (iii) recall that Step 4 in Definition 4.1 is exactly the minimization of the norm in question, constrained to all polynomials satisfying the restriction property for u. Since $\widehat{\Pi}_p^{\mathrm{MS}} u$ given by Proposition 4.3 also satisfies the restriction property we can immediately conclude for $p \geq s-1$ that

$$\begin{aligned} p\left\|u-\widehat{\Pi}_p u\right\|_{L^2(\widehat{K})}+\left\|u-\widehat{\Pi}_p u\right\|_{H^1(\widehat{K})} &\leq p\left\|u-\widehat{\Pi}_p^{\mathrm{MS}} u\right\|_{L^2(\widehat{K})}+\left\|u-\widehat{\Pi}_p^{\mathrm{MS}} u\right\|_{H^1(\widehat{K})} \\ &\leq C_s p^{-(s-1)}|u|_{H^s(\widehat{K})}. \end{aligned}$$

We turn to Assertion (iv). Since polynomials up to degree p are preserved under $\widehat{\Pi}_p$, we immediately have

$$\left\|u-\widehat{\Pi}_p u\right\|_{W^{1,\infty}(\widehat{K})} \leq \|u-q\|_{W^{1,\infty}(\widehat{K})}+\left\|\widehat{\Pi}_p q-\widehat{\Pi}_p u\right\|_{W^{1,\infty}(\widehat{K})}, \tag{4.30}$$

for any $q \in \mathscr{P}_p(\widehat{K})$. We estimate the second term in (4.30). We have seen in (4.28) that the operator $\widehat{\Pi}_p$ can be written as $\widehat{\Pi}_p u=\pi^u+\widehat{\Pi}_{\mathscr{P}_p^0}(u-\pi^u)$ for any $\pi^u \in \mathscr{V}_p^u$ (the affine space of polynomials with restriction property for u), where $\widehat{\Pi}_{\mathscr{P}_p^0}$ is the orthogonal projection onto $\mathscr{P}_p^0(\widehat{K}) \leq \mathscr{P}_p(\widehat{K})$, the space of polynomials vanishing on $\partial\widehat{K}$, with respect to the norm $|||\cdot|||$. Therefore we have

$$\widehat{\Pi}_p q-\widehat{\Pi}_p u=\pi^q-\pi^u+\widehat{\Pi}_{\mathscr{P}_p^0}(q-u+\pi^u-\pi^q)$$

for any $\pi^u \in \mathscr{V}_p^u$ and $\pi^q \in \mathscr{V}_p^q$. Selecting $q \in \mathscr{V}_p^u$ allows us to choose $\pi^u = \pi^q = q$, which immediately gives

$$\widehat{\Pi}_p q - \widehat{\Pi}_p u = \widehat{\Pi}_{\mathscr{P}_p^0}(q-u)$$

for all $q \in \mathscr{V}_p^u$. Using the polynomial inverse estimates $\|\pi\|_{L^\infty(\Omega)} \leq C p^d \|\pi\|_{L^2(\Omega)}$ for all $\pi \in \mathscr{P}_p(\widehat{K})$, (see, e.g., [27, Thm. 4.76] for the case $d = 2$), we find

$$\left\|\widehat{\Pi}_p q - \widehat{\Pi}_p u\right\|_{W^{1,\infty}(\widehat{K})} = \left\|\widehat{\Pi}_{\mathscr{P}_p^0}(q-u)\right\|_{W^{1,\infty}(\widehat{K})} \lesssim p^d \left\|\widehat{\Pi}_{\mathscr{P}_p^0}(q-u)\right\|_{H^1(\widehat{K})}.$$

Since $\widehat{\Pi}_{\mathscr{P}_p^0}$ is the orthogonal projection with respect to the norm $|||\cdot|||$ we obtain

$$p^d \left\|\widehat{\Pi}_{\mathscr{P}_p^0}(q-u)\right\|_{H^1(\widehat{K})} \leq p^d |||q-u||| \lesssim p^{d+1} \|q-u\|_{W^{1,\infty}(\widehat{K})}.$$

We therefore conclude that

$$\left\|u - \widehat{\Pi}_p u\right\|_{W^{1,\infty}(\widehat{K})} \lesssim p^{d+1} \|u-q\|_{W^{1,\infty}(\widehat{K})}$$

for all $q \in \mathscr{V}_p^u$. Proposition 4.5 provides a polynomial $q \in \mathscr{V}_p^u$ with the desired approximation properties. Absorbing the algebraic factor p^{d+1} into the exponential factor then yields the result. □

4.4.3 $\boldsymbol{H}(\mathrm{div}, \Omega)$-Conforming Approximation Operators

In the following we will construct an approximation operator $\boldsymbol{\Pi}_p^{\mathrm{div},s}: \boldsymbol{H}^s(\Omega) \to \mathbf{BDM}_p(\mathscr{T}_h) \subset \mathbf{RT}_p(\mathscr{T}_h)$ that features the optimal convergence rates in p simultaneously in $L^2(\Omega)$ and $\boldsymbol{H}(\mathrm{div}, \Omega)$ for $s > d/2$. The operator will act elementwise. First we consider any operator $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}: \boldsymbol{H}^s(\widehat{K}) \to \mathbf{BDM}_p(\widehat{K}) \subset \mathbf{RT}_p(\widehat{K})$ and define $\boldsymbol{\Pi}_p^{\mathrm{div},s}$ on $\boldsymbol{H}^s(\Omega)$ elementwise using the Piola transformation by

$$\left(\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right)\Big|_K := \left[(\det F_K')^{-1} F_K' \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\left[(\det F_K')(F_K')^{-1}\boldsymbol{\varphi} \circ F_K\right]\right] \circ F_K^{-1}. \tag{4.31}$$

In order for $\boldsymbol{\Pi}_p^{\mathrm{div},s}$ to map into the conforming finite element space one has to select the operator $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}$ correctly. We choose $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}: \boldsymbol{H}^s(\widehat{K}) \to \mathscr{P}_p(\widehat{K})^d = \mathbf{BDM}_p(\widehat{K}) \subset \mathbf{RT}_p(\widehat{K})$ to be the componentwise application of $\widehat{\Pi}_p$ from Definition 4.1 and analyzed in Theorem 4.3:

$$\left(\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right)_i := \widehat{\Pi}_p\boldsymbol{\varphi}_i, \qquad \text{for } i = 1,\dots,d. \tag{4.32}$$

This choice will ensure the desired approximation properties, and will also map into the conforming finite element space due to the restriction property. We will summarize and prove certain properties of the above constructed operators $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}$ and $\boldsymbol{\Pi}_p^{\mathrm{div},s}$. See [21] for a similar construction concerning the space $\boldsymbol{H}(\mathrm{curl},\Omega)$.

Lemma 4.2 *Let $s > d/2$ and let the operators $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}$ and $\boldsymbol{\Pi}_p^{\mathrm{div},s}$ be defined as above. Then there holds:*

(i) The operator $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}: \boldsymbol{H}^s(\widehat{K}) \to \mathbf{BDM}_p(\widehat{K}) \subset \mathbf{RT}_p(\widehat{K})$ satisfies for $p \geq s-1$

$$p\left\|\boldsymbol{\varphi} - \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{L^2(\widehat{K})} + \left\|\boldsymbol{\varphi} - \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{H^1(\widehat{K})} \lesssim p^{-(s-1)}|\boldsymbol{\varphi}|_{H^s(\widehat{K})}. \tag{4.33}$$

(ii) Under the assumptions Theorem 4.3, (iv) *(with $C_{\boldsymbol{\varphi}}$ replacing C_u) there holds for some constants C_A, $\sigma > 0$ independent of p, h, R*

$$\left\|\boldsymbol{\varphi} - \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{W^{1,\infty}(\widehat{K})} \leq C_A C_{\boldsymbol{\varphi}}\left[\left(\frac{h/R}{\sigma + h/R}\right)^{p+1} + \left(\frac{h\kappa}{\sigma p}\right)^{p+1}\right].$$

(iii) The operator $\boldsymbol{\Pi}_p^{\mathrm{div},s}$ defined on $\boldsymbol{H}^s(\Omega)$ maps to the conforming space $\mathbf{BDM}_p(\mathscr{T}_h) \subset \mathbf{RT}_p(\mathscr{T}_h)$.

Proof The first two assertions hold by construction and Theorem 4.3, properties (iii), (iv). To prove the third assertion, note that $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}$ maps to $\mathbf{BDM}_p(\widehat{K})$ so that

$$(\det F_K')(F_K')^{-1}\left(\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right)\Big|_K \circ F_K \in \mathbf{BDM}_p(\widehat{K}) \qquad \text{for all } K \in \mathscr{T}_h, \tag{4.34}$$

by construction. We are therefore left with verifying that $\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi} \in \boldsymbol{H}(\mathrm{div},\Omega)$. Since $\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}$ is piecewise smooth it suffices to show inter element continuity of the normal trace. We will first show that the normal trace of $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}$ in fact only depends on the normal trace of $\boldsymbol{\varphi}$. Consider a face $\hat{f}$ of $\widehat{K}$. Let $\gamma_{\hat{\boldsymbol{n}}_{\hat{f}}}$ denote the

normal trace for the face $\hat{f}$. We calculate

$$\gamma_{\hat{\boldsymbol{n}}_{\hat{f}}}\left(\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right) = \left(\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right)\Big|_{\hat{f}} \cdot \hat{\boldsymbol{n}}_{\hat{f}} = \begin{pmatrix} \widehat{\Pi}_p\boldsymbol{\varphi}_1 \\ \vdots \\ \widehat{\Pi}_p\boldsymbol{\varphi}_d \end{pmatrix}\Bigg|_{\hat{f}} \cdot \hat{\boldsymbol{n}}_{\hat{f}}$$

$$= \begin{pmatrix} \widehat{\Pi}_p(\boldsymbol{\varphi}_1|_{\hat{f}}) \\ \vdots \\ \widehat{\Pi}_p(\boldsymbol{\varphi}_d|_{\hat{f}}) \end{pmatrix} \cdot \hat{\boldsymbol{n}}_{\hat{f}} = \widehat{\Pi}_p(\boldsymbol{\varphi}\cdot\hat{\boldsymbol{n}}_{\hat{f}}) = \widehat{\Pi}_p(\gamma_{\hat{\boldsymbol{n}}_{\hat{f}}}\boldsymbol{\varphi}).$$

Here we used that the operator $\widehat{\Pi}_p$ satisfies the restriction property and the fact that $\hat{\boldsymbol{n}}_{\hat{f}}$ is constant on $\hat{f}$. Furthermore note that we abused notation in that the symbol $\widehat{\Pi}_p$ is used both for the d dimensional as well as the $d-1$ dimensional version. We conclude the proof using the fact that if $\hat{\boldsymbol{n}}$ is the unit outward normal to $\widehat{K}$ the vector $\boldsymbol{n}$ on K given by

$$\boldsymbol{n}\circ F_K = \frac{1}{\|(F_K')^{-T}\hat{\boldsymbol{n}}\|}(F_K')^{-T}\hat{\boldsymbol{n}}$$

is a unit normal to K, see, e.g., [23, Section 3.9 and 5.4]. □

We have p-optimal approximation properties on the reference element $\widehat{K}$ by the operator $\widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}$.

Corollary 4.1 (Approximation of $H^s(\Omega)$ Functions) *For $d = 2, 3$ and $s > d/2$ the operator $\boldsymbol{\Pi}_p^{\mathrm{div},s}\colon \boldsymbol{H}^s(\Omega) \to \mathbf{BDM}_p(\mathscr{T}_h) \subset \mathbf{RT}_p(\mathscr{T}_h)$ satisfies*

$$\frac{p}{h}\left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{L^2(\Omega)} + \left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{H^1(\mathscr{T}_h)} \lesssim \left(\frac{h}{p}\right)^{s-1}\|\boldsymbol{\varphi}\|_{H^s(\Omega)} \quad \forall p \geq s-1,$$

where $\|\cdot\|_{H^1(\mathscr{T}_h)}$ denotes the broken H^1-norm.

Proof The proof follows from Lemma 4.2 together with a scaling argument. □

Corollary 4.2 (Approximation of Analytic Functions) *Let $\boldsymbol{\varphi}$ satisfy, for some $C_{\boldsymbol{\varphi}}$, $\gamma > 0$,*

$$\left\|\nabla^n\boldsymbol{\varphi}\right\|_{L^2(\Omega)} \leq C_{\boldsymbol{\varphi}}\gamma^n \max(n,k)^n \qquad \forall n \in \mathbb{N}_0.$$

There exist C, $\sigma > 0$ independent of h, p, and k such that

$$\left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{H^1(\mathscr{T}_h)} + k\left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{L^2(\Omega)}$$
$$\leq CC_{\boldsymbol{\varphi}}\left[\left(\frac{h}{h+p}\right)^p\left(1+\frac{hk}{h+\sigma}\right) + k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p}+\frac{kh}{\sigma p}\right)\right].$$

Proof We mimic the procedure of [19, Thm. 5.5] and [3, Lemma 4.7]. First consider for each element $K \in \mathscr{T}_h$ the constant C_K given by

$$C_K^2 := \sum_{n\geq 0} \frac{\|\nabla^n \boldsymbol{\varphi}\|_{L^2(K)}^2}{(2\gamma \max(n,k))^{2n}},$$

which is finite by assumption. Note that we immediately have

$$\left\|\nabla^n \boldsymbol{\varphi}\right\|_{L^2(K)} \leq 2^n \gamma^n \max(n,k)^n C_K,$$
$$\sum_{K\in\mathscr{T}_h} C_K^2 \leq \frac{4}{3} C_{\boldsymbol{\varphi}}^2.$$

We write $\widehat{\boldsymbol{\varphi}}$ as

$$\begin{aligned}\widehat{\boldsymbol{\varphi}} &= \det(F_K')(F_K')^{-1}\boldsymbol{\varphi}\circ F_K = \det(R_K'\circ A_K A_K')(R_K'\circ A_K A_K')^{-1}\boldsymbol{\varphi}\circ F_K \\ &= \det(A_K')(A_K')^{-1}\tilde{\boldsymbol{\varphi}}\circ A_K,\end{aligned}$$

with

$$\tilde{\boldsymbol{\varphi}} = \det(R_K')(R_K')^{-1}\boldsymbol{\varphi}\circ R_K.$$

As in [19, Lemma C.1] for simple changes of variables, we apply [16, Lemma 4.3.1] to the function $\tilde{\boldsymbol{\varphi}}$ and obtain the existence of constants $\overline{\gamma}$, $C > 0$ depending additionally on the constants describing the analyticity of the map R_K such that

$$\left\|\nabla^n \tilde{\boldsymbol{\varphi}}\right\|_{L^2(\hat{K})} \leq C\overline{\gamma}^n \max(n,k)^n C_K \qquad \forall n \in \mathbb{N}_0.$$

Since A_K is affine we immediately deduce that

$$\left\|\nabla^n \widehat{\boldsymbol{\varphi}}\right\|_{L^2(\widehat{K})} \lesssim h^{d/2-1}h^n \left\|\nabla^n \tilde{\boldsymbol{\varphi}}\right\|_{L^2(\hat{K})} \leq h^{d/2-1}(\overline{\gamma}h)^n \max(n,k)^n C_K \quad \forall n \in \mathbb{N}_{n\geq 1}.$$

Hence by Lemma 4.2 with $R = 1$ we have

$$\left\|\widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\widehat{\boldsymbol{\varphi}}\right\|_{W^{1,\infty}(\widehat{K})} \lesssim C_K h^{d/2-1}\left[\left(\frac{h}{\sigma+h}\right)^{p+1} + \left(\frac{hk}{\sigma p}\right)^{p+1}\right]$$

for some $\sigma > 0$. By a change of variables there holds for $q = 0, 1$

$$\begin{aligned}\left\|\boldsymbol{\varphi} - \boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{H^q(K)} &\lesssim h^{-d/2+1-q}\left\|\widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\Pi}}_p^{\mathrm{div},s}\widehat{\boldsymbol{\varphi}}\right\|_{H^q(\widehat{K})} \\ &\lesssim h^{-q} C_K\left[\left(\frac{h}{\sigma+h}\right)^{p+1} + \left(\frac{hk}{\sigma p}\right)^{p+1}\right].\end{aligned}$$

Summation over all elements gives

$$\begin{aligned}
&\left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{H^1(\mathscr{T}_h)}+k\left\|\boldsymbol{\varphi}-\boldsymbol{\Pi}_p^{\mathrm{div},s}\boldsymbol{\varphi}\right\|_{L^2(\Omega)}\\
&\lesssim\left[\left(\frac{h}{\sigma+h}\right)^p+k\left(\frac{h}{\sigma+h}\right)^{p+1}+\frac{k}{p}\left(\frac{hk}{\sigma p}\right)^p+k\left(\frac{hk}{\sigma p}\right)^{p+1}\right]\sqrt{\sum_{K\in\mathscr{T}_h}C_K^2}\\
&\lesssim\left[\left(\frac{h}{h+p}\right)^p\left(1+\frac{hk}{h+\sigma}\right)+k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p}+\frac{kh}{\sigma p}\right)\right]C_{\boldsymbol{\varphi}},
\end{aligned}$$

which completes the proof. □

4.5 A Priori Estimate

We now turn to an a priori estimate of the FOSLS method. Again the proof follows the ideas of [3, Lemma 5.1], resting, however, on the refined duality argument given in Lemma 4.1 and the approximation properties derived in Sect. 4.4 to obtain the factor h/p. For the readers' convenience we recapitulate the important steps. As in [19] we show that this can be achieved under the conditions kh/p sufficiently small and p of order $\log k$.

Theorem 4.4 (A Priori Estimate) *Let Assumptions 4.1, 4.2 be valid. Then there exist constants $c_1, c_2 > 0$ that are independent of h, p, and k such that the conditions*

$$\frac{kh}{p}\le c_1 \qquad \text{and} \qquad p\ge c_2(\log k+1) \tag{4.35}$$

imply that the approximation $(\boldsymbol{\varphi}_h, u_h)$ of the FOSLS method satisfies the following: For any $(\boldsymbol{\psi}_h, v_h)\in \boldsymbol{V}_h\times W_h$ there holds

$$\begin{aligned}
\|u-u_h\|_{L^2(\Omega)}\lesssim\frac{h}{p}\Big(&\|\nabla(u-v_h)\|_{L^2(\Omega)}+k\,\|u-v_h\|_{L^2(\Omega)}+\\
&\|\nabla\cdot(\boldsymbol{\varphi}-\boldsymbol{\psi}_h)\|_{L^2(\Omega)}+k\,\|\boldsymbol{\varphi}-\boldsymbol{\psi}_h\|_{L^2(\Omega)}+k^{1/2}\,\|(\boldsymbol{\varphi}-\boldsymbol{\psi}_h)\cdot\boldsymbol{n}\|_{L^2(\partial\Omega)}\Big).
\end{aligned}$$

Proof Let $e^u=u-u_h$ and $\boldsymbol{e}^{\boldsymbol{\varphi}}=\boldsymbol{\varphi}-\boldsymbol{\varphi}_h$ denote the errors of the two components. We apply the duality argument from Lemma 4.1 with $w=e^u$ and also apply the corresponding splitting:

$$\|e^u\|_{L^2(\Omega)}^2=b((\boldsymbol{e}^{\boldsymbol{\varphi}},e^u),(\boldsymbol{\psi},v))=b((\boldsymbol{e}^{\boldsymbol{\varphi}},e^u),(\boldsymbol{\psi}_A,v_A))+b((\boldsymbol{e}^{\boldsymbol{\varphi}},e^u),(\boldsymbol{\psi}_{H^2},v_{H^2})).$$

Exploiting the Galerkin orthogonality we have

$$\left\|e^u\right\|^2_{L^2(\Omega)} = b((\boldsymbol{e}^{\boldsymbol{\varphi}}, e^u), (\boldsymbol{\psi}_A - \tilde{\boldsymbol{\psi}}_A, v_A - \tilde{v}_A)) + b((\boldsymbol{e}^{\boldsymbol{\varphi}}, e^u), (\boldsymbol{\psi}_{H^2} - \tilde{\boldsymbol{\psi}}_{H^2}, v_{H^2} - \tilde{v}_{H^2})).$$

for any $(\tilde{\boldsymbol{\psi}}_A, \tilde{v}_A), (\tilde{\boldsymbol{\psi}}_{H^2}, \tilde{v}_{H^2}) \in \boldsymbol{V}_h \times W_h$. Using Cauchy-Schwarz we arrive at

$$\begin{aligned}
\left\|e^u\right\|^2_{L^2(\Omega)} \lesssim & \left[\left\|ik\boldsymbol{e}^{\boldsymbol{\varphi}} + \nabla e^u\right\|_{L^2(\Omega)} + \left\|ike^u + \nabla\cdot\boldsymbol{e}^{\boldsymbol{\varphi}}\right\|_{L^2(\Omega)} + k^{1/2}\left\|\boldsymbol{e}^{\boldsymbol{\varphi}}\cdot\boldsymbol{n} + e^u\right\|_{L^2(\partial\Omega)}\right]\cdot \\
& \Big(\left\|\nabla\cdot(\boldsymbol{\psi}_A - \tilde{\boldsymbol{\psi}}_A)\right\|_{L^2(\Omega)} + k\left\|\boldsymbol{\psi}_A - \tilde{\boldsymbol{\psi}}_A\right\|_{L^2(\Omega)} + k^{1/2}\left\|(\boldsymbol{\psi}_A - \tilde{\boldsymbol{\psi}}_A)\cdot\boldsymbol{n}\right\|_{L^2(\partial\Omega)} + \\
& \left\|\nabla\cdot(\boldsymbol{\psi}_{H^2} - \tilde{\boldsymbol{\psi}}_{H^2})\right\|_{L^2(\Omega)} + k\left\|\boldsymbol{\psi}_{H^2} - \tilde{\boldsymbol{\psi}}_{H^2}\right\|_{L^2(\Omega)} + k^{1/2}\left\|(\boldsymbol{\psi}_{H^2} - \tilde{\boldsymbol{\psi}}_{H^2})\cdot\boldsymbol{n}\right\|_{L^2(\partial\Omega)} + \\
& \|\nabla(v_A - \tilde{v}_A)\|_{L^2(\Omega)} + k\|v_A - \tilde{v}_A\|_{L^2(\Omega)} + k^{1/2}\|v_A - \tilde{v}_A\|_{L^2(\partial\Omega)} + \\
& \left\|\nabla(v_{H^2} - \tilde{v}_{H^2})\right\|_{L^2(\Omega)} + k\left\|v_{H^2} - \tilde{v}_{H^2}\right\|_{L^2(\Omega)} + k^{1/2}\left\|v_{H^2} - \tilde{v}_{H^2}\right\|_{L^2(\partial\Omega)}\Big).
\end{aligned} \tag{4.36}$$

We are going to exploit the approximation properties in the corresponding norms and spaces.

Approximation of v_A and v_{H^2} For the approximation we may apply [3, Lemma 4.10], which is essentially the procedure of [19, Thm. 5.5] together with a multiplicative trace inequality. Using the estimates (4.9), (4.10), and (4.12) in Lemma 4.1 as well as [19, Thm. B.4] to find appropriate approximations $\tilde{v}_{H^2}$ and $\tilde{v}_A$ we have

$$\begin{aligned}
&\|\nabla(v_A - \tilde{v}_A)\|_{L^2(\Omega)} + k\|v_A - \tilde{v}_A\|_{L^2(\Omega)} + k^{1/2}\|v_A - \tilde{v}_A\|_{L^2(\partial\Omega)} \\
&\quad\lesssim \left[\left(\frac{h}{h+p}\right)^p\left(1 + \frac{hk}{h+\sigma}\right) + k\left(\frac{kh}{\sigma p}\right)^p\left(\frac{1}{p} + \frac{kh}{\sigma p}\right)\right]\left\|e^u\right\|_{L^2(\Omega)} \\
&\quad\lesssim \frac{h}{p}\left\|e^u\right\|_{L^2(\Omega)}
\end{aligned}$$

as well as

$$\begin{aligned}
&\left\|\nabla(v_{H^2} - \tilde{v}_{H^2})\right\|_{L^2(\Omega)} + k\left\|v_{H^2} - \tilde{v}_{H^2}\right\|_{L^2(\Omega)} + k^{1/2}\left\|v_{H^2} - \tilde{v}_{H^2}\right\|_{L^2(\partial\Omega)} \\
&\quad\lesssim \frac{1}{k}\left(\frac{kh}{p} + \left(\frac{kh}{p}\right)^2\right)\left\|e^u\right\|_{L^2(\Omega)} \lesssim \frac{h}{p}\left\|e^u\right\|_{L^2(\Omega)},
\end{aligned}$$

where the latter estimates are due to the boundedness of $\Omega, \sigma > 0$, and choosing c_1 small and c_2 sufficiently large as well as elementary but tedious calculations.

Approximation of $\boldsymbol{\psi}_A$ To approximate $\boldsymbol{\psi}_A$ we choose $\tilde{\boldsymbol{\psi}}_A = \boldsymbol{\Pi}_p^{\mathrm{div},2}\boldsymbol{\psi}_A$ with $\boldsymbol{\Pi}_p^{\mathrm{div},2}$ as in Corollary 4.2 and apply the results therein. Furthermore we apply

the estimates (4.8) and (4.10) of Lemma 4.1. Proceeding as above together with a multiplicative trace inequality, again after tedious calculations, gives

$$\left\|\nabla\cdot(\boldsymbol{\psi}_A-\tilde{\boldsymbol{\psi}}_A)\right\|_{L^2(\Omega)}+k\left\|\boldsymbol{\psi}_A-\tilde{\boldsymbol{\psi}}_A\right\|_{L^2(\Omega)}+k^{1/2}\left\|(\boldsymbol{\psi}_A-\tilde{\boldsymbol{\psi}}_A)\cdot\boldsymbol{n}\right\|_{L^2(\partial\Omega)}\lesssim\frac{h}{p}\left\|e^u\right\|_{L^2(\Omega)}.$$

Approximation of $\boldsymbol{\psi}_{H^2}$ To approximate $\boldsymbol{\psi}_{H^2}$ we choose $\tilde{\boldsymbol{\psi}}_{H^2}=\boldsymbol{\Pi}_p^{\mathrm{div},2}\boldsymbol{\psi}_{H^2}$ with $\boldsymbol{\Pi}_p^{\mathrm{div},2}$ as in Corollary 4.1 and apply the results therein. We apply the estimate (4.11) of Lemma 4.1. Due to the multiplicative trace inequality we also have

$$\left\|(\boldsymbol{\psi}_{H^2}-\tilde{\boldsymbol{\psi}}_{H^2})\cdot\boldsymbol{n}\right\|_{L^2(\partial\Omega)}\leq\left(\frac{h}{p}\right)^{3/2}\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^2(\Omega)}. \tag{4.37}$$

Therefore we arrive at

$$\left\|\nabla\cdot(\boldsymbol{\psi}_{H^2}-\tilde{\boldsymbol{\psi}}_{H^2})\right\|_{L^2(\Omega)}+k\left\|\boldsymbol{\psi}_{H^2}-\tilde{\boldsymbol{\psi}}_{H^2}\right\|_{L^2(\Omega)}+k^{1/2}\left\|(\boldsymbol{\psi}_{H^2}-\tilde{\boldsymbol{\psi}}_{H^2})\cdot\boldsymbol{n}\right\|_{L^2(\partial\Omega)}\lesssim\frac{h}{p}\left\|\boldsymbol{\psi}_{H^2}\right\|_{H^2(\Omega)}\lesssim\frac{h}{p}\left\|e^u\right\|_{L^2(\Omega)},$$

where we used the estimate (4.11) of Lemma 4.1. Putting it all together we have

$$\begin{aligned}\left\|e^u\right\|_{L^2(\Omega)}&\lesssim\frac{h}{p}(\left\|ik\boldsymbol{e}^{\boldsymbol{\varphi}}+\nabla e^u\right\|_{L^2(\Omega)}+\left\|ike^u+\nabla\cdot\boldsymbol{e}^{\boldsymbol{\varphi}}\right\|_{L^2(\Omega)}+k^{1/2}\left\|\boldsymbol{e}^{\boldsymbol{\varphi}}\cdot\boldsymbol{n}+e^u\right\|_{L^2(\partial\Omega)})\\&\lesssim\frac{h}{p}\sqrt{b((\boldsymbol{e}^{\boldsymbol{\varphi}},e^u),(\boldsymbol{e}^{\boldsymbol{\varphi}},e^u))}.\end{aligned}$$

Applying again the Galerkin orthogonality and using the multiplicative trace inequality to absorb the term $k^{1/2}\left\|u-v_h\right\|_{L^2(\partial\Omega)}$ into the L^2 norms of the volume yields the result.

□

We conclude this section with a simple consequence of standard regularity theory and approximation properties of the employed finite element spaces in higher order Sobolev norms.

Corollary 4.3 *For $s\geq 0$, $f\in H^s(\Omega)$ and $g\in H^{s+1/2}(\partial\Omega)$ we have $u\in H^{s+2}(\Omega)$, $u\in H^{s+3/2}(\partial\Omega)$, $\partial_n u\in H^{s+1/2}(\partial\Omega)$, $\boldsymbol{\varphi}\in\boldsymbol{H}^{s+1}(\Omega)$, $\nabla\cdot\boldsymbol{\varphi}\in H^s(\Omega)$ and $\boldsymbol{\varphi}\cdot\boldsymbol{n}\in\boldsymbol{H}^{s+1/2}(\partial\Omega)$. Furthermore there exist constants c_1, $c_2>0$ that are independent of h, p, and k such that the conditions*

$$\frac{kh}{p}\leq c_1\qquad\text{and}\qquad p\geq c_2(\log k+1) \tag{4.38}$$

imply that the solution $(\boldsymbol{\varphi}_h, u_h)$ *satisfies*

$$\|u - u_h\|_{L^2(\Omega)} \lesssim \left(\frac{h}{p}\right)^{s+1} (\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\Omega)}),$$

for $p \geq s$ *with a wavenumber independent constant.*

Proof The first assertion follows immediately from standard regularity theory. Consider the case $s > 0$. Theorem 4.4 together with a multiplicative trace inequality, which is applicable due to the already derived regularity of $\boldsymbol{\varphi}$, gives

$$\|u - u_h\|_{L^2(\Omega)} \lesssim \frac{h}{p}\Big(\|u - v_h\|_{H^1(\Omega)} + k\,\|u - v_h\|_{L^2(\Omega)} + \|\boldsymbol{\varphi} - \boldsymbol{\psi}_h\|_{H^1(\mathcal{T}_h)} + k\,\|\boldsymbol{\varphi} - \boldsymbol{\psi}_h\|_{L^2(\Omega)} \Big).$$

Applying the higher order splitting of Theorem 4.1 and using the fact that $\boldsymbol{\varphi} = ik^{-1}\nabla u$, one can easily estimate, as in the proof of Theorem 4.4 together with the Corollaries 4.1 and 4.2,

$$\|\boldsymbol{\varphi} - \boldsymbol{\psi}_h\|_{H^1(\Omega)} + k\,\|\boldsymbol{\varphi} - \boldsymbol{\psi}_h\|_{L^2(\Omega)} \lesssim \left(\frac{h}{p}\right)^{s} (\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\Omega)}).$$

Note the exponent s, since $\boldsymbol{\varphi}$ is only in $\boldsymbol{H}^{s+1}(\Omega)$. Furthermore, again as in the proof of Theorem 4.4, see also [22, Thm. 4.8], we have

$$\|u - v_h\|_{H^1(\Omega)} + k\,\|u - v_h\|_{L^2(\Omega)} \lesssim \left(\frac{h}{p}\right)^{s+1} (\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\Omega)}),$$

now with the exponent $s+1$ since $u \in H^{s+2}(\Omega)$, which yields the result for $s > 0$. In the case $s = 0$ one simply sets $v_h = 0$ as well as $\boldsymbol{\psi}_h = 0$ and uses the wavenumber-explicit estimates of Theorem 4.1. □

Remark 4.7 Note that although we assume $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$ in Corollary 4.3, we only obtained a convergence rate $s+1$. This seems suboptimal when compared with classical FEM where, given sufficient regularity of the data and the geometry, one can expect a rate of $s+2$ for the convergence in the $L^2(\Omega)$-norm. Especially for $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ one can only expect h/p for the FOSLS method compared to h^2/p^2 for the FEM. The proof of Corollary 4.3 is in that sense sharp since the leading error term in the a priori estimate is

$$\|\nabla \cdot (\boldsymbol{\varphi} - \boldsymbol{\psi}_h)\|_{L^2(\Omega)} = \left\| ik^{-1} f + iku - \nabla \cdot \boldsymbol{\psi}_h \right\|_{L^2(\Omega)},$$

where we used the fact $\boldsymbol{\varphi} = ik^{-1}\nabla u$. The essential part is therefore to approximate an f that is just in $L^2(\Omega)$ and therefore no further powers of h can be gained.

Assuming more regularity on f would resolve this problem, however, the boundary data would restrict a further lifting of $\boldsymbol{\varphi}$ in classical Sobolev spaces, but not in $H(\mathrm{div}, \Omega)$ spaces. This in turn would make it necessary to directly estimate $\|\nabla \cdot (\boldsymbol{\varphi} - \boldsymbol{\psi}_h)\|_{L^2(\Omega)}$ instead of generously bounding it by $\|\boldsymbol{\varphi} - \boldsymbol{\psi}_h\|_{H^1(\mathscr{T}_h)}$. Last but not least there is the boundary term

$$\|(\boldsymbol{\varphi} - \boldsymbol{\psi}_h) \cdot \boldsymbol{n}\|_{L^2(\partial\Omega)} = \left\| ik^{-1} g - u - \boldsymbol{\psi}_h \cdot \boldsymbol{n} \right\|_{L^2(\partial\Omega)}.$$

Again if g is only $H^{1/2}(\partial\Omega)$ one can only expect $\sqrt{h/p}$, but favorable in terms of k. □

4.6 Numerical Examples

All our calculations are performed with the hp-FEM code NETGEN/NGSOLVE by Schöberl, [25, 26]. We plot the error against N_λ, the number of degrees of freedom per wavelength,

$$N_\lambda = \frac{2\pi \sqrt[d]{\mathrm{DOF}}}{k \sqrt[d]{|\Omega|}},$$

where the wavelength λ and the wavenumber k are related via $k = 2\pi/\lambda$ and DOF denotes the size of the linear system to be solved. We compare the results of the classical FEM with the FOSLS method, measured in the relative $L^2(\Omega)$ error. For the classical FEM we use the standard space $S_p(\mathscr{T}_h)$. For the FOSLS method we employ the pairing $\boldsymbol{V}_h \times W_h = \mathbf{BDM}_p(\mathscr{T}_h) \times S_p(\mathscr{T}_h)$.

Example 4.1 Let Ω be the unit circle in $\mathbb{R}^2$ and consider the problem

$$\begin{aligned} -\Delta u - k^2 u &= 0 \qquad \text{in } \Omega, \\ \partial_n u - iku &= g \qquad \text{on } \partial\Omega. \end{aligned}$$

where the data g is such that the exact solution is given by $u(x, y) = \mathrm{e}^{i(k_1 x + k_2 y)}$ with $k_1 = -k_2 = \frac{1}{\sqrt{2}}k$. For the numerical studies, this problem will be solved using h-FEM and h-FOSLS with polynomial degrees $p = 1, 2, 3, 4$. The results are visualized in Fig. 4.1. For both methods we observe the expected convergence $O(h^{p+1})$ in the relative $L^2(\Omega)$ error. Note that for both methods higher order versions are less prone to the pollution effect. At the same number of degrees of freedom per wavelength we also observe that the classical FEM is superior to FOSLS, when measured in achieved accuracy in $L^2(\Omega)$. This is not surprising since, for the same mesh and polynomial degree p, the number of degrees of freedom of the FOSLS is roughly three times as large as for the classical FEM. Note, however, that we do not consider any solver aspects of the employed methods, where FOSLS

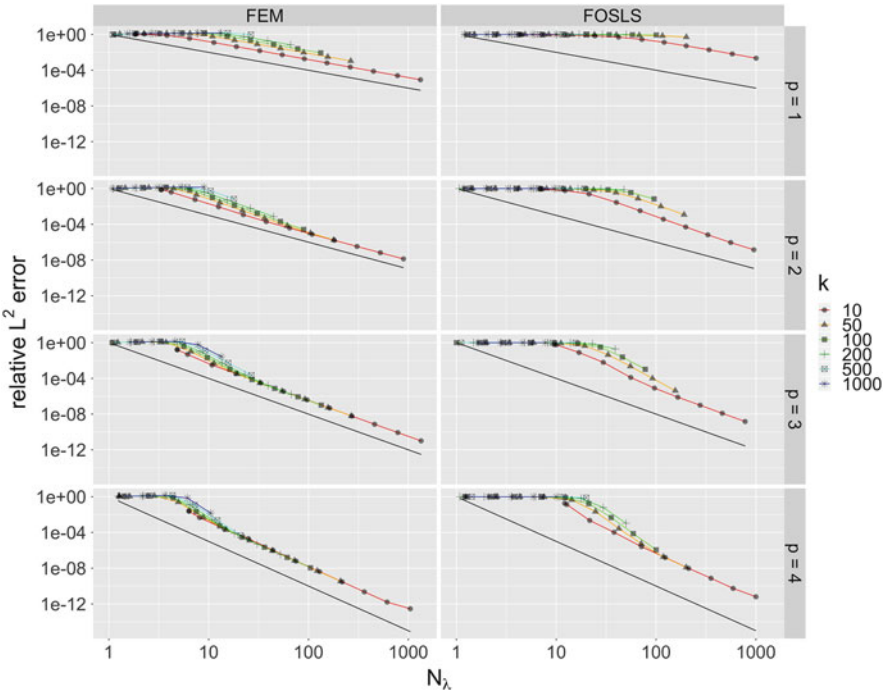


Fig. 4.1 Comparison between the h-FEM (left) and h-FOSLS (right) for $p = 1, 2, 3, 4$ as described in Example 4.1. The reference line in black corresponds to h^{p+1}

might have advantages over the classical FEM since its system matrix is positive definite.

Example 4.2 For $\pi < \omega < 2\pi$ let $\Omega = \{(r\cos\varphi, r\sin\varphi) : r \in (0,1), \varphi \in (0,\omega)\} \subset \mathbb{R}^2$ and consider

$$\begin{aligned} -\Delta u - k^2 u &= 0 \qquad \text{in } \Omega, \\ \partial_n u - iku &= g \qquad \text{on } \partial\Omega. \end{aligned}$$

The data g is such that the exact solution is given by $u(x,y) = J_\alpha(kr)\cos(\alpha\varphi)$, with $\alpha = 3\pi/2$. Standard regularity theory gives $u \in H^{1+\alpha-\varepsilon}(\Omega)$ for every $\varepsilon > 0$. In the numerical experiments we keep $kh = 5$ and perform a p-FEM and a p-FOSLS method up to $p = 46$ and $p = 29$, respectively. The results are visualized in Fig. 4.2. We observe that the FEM has significantly smaller errors than the FOSLS. For a discussion of the expected $L^2(\Omega)$-convergence rates of the p-FEM, we refer the reader to [14, Remark after Thm. 3 and Section 3].

The next example focuses on the Helmholtz equation with right-hand side f with finite Sobolev regularity.

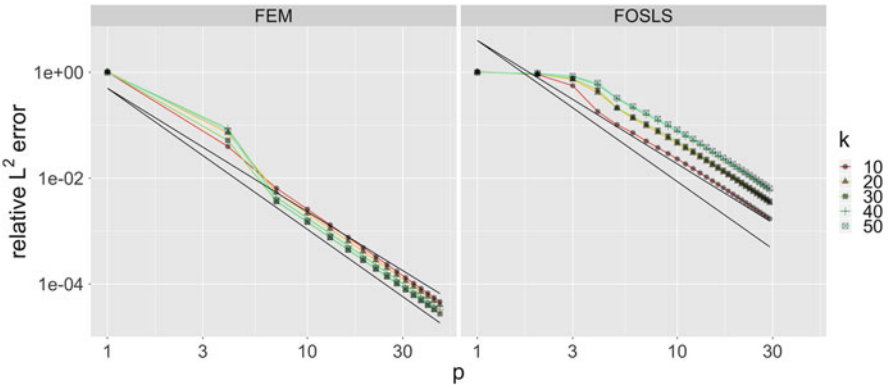


Fig. 4.2 Comparison between the p-FEM (left) and p-FOSLS (right) for $kh = 5$ as described in Example 4.2. We include the reference lines $p^{-4\cdot 2/3} = p^{-8/3}$ and $p^{-(2\cdot 2/3+1)} = p^{-7/3}$

Example 4.3 Let $\Omega = (-1, 1) \subset \mathbb{R}$ and $f = -\chi_{(-1,0]} + \chi_{(0,1)}$, where χ_A denotes the indicator function on $A \subset \mathbb{R}$. The function f is in $H^{1/2-\varepsilon}(\Omega)$ for every $\varepsilon > 0$. We consider uniform meshes $\mathscr{T}_h$ on Ω such that the break point zero is *not* a node, as otherwise the piecewise smooth solution could be approximated very well. We study

$$\begin{aligned} -u'' - k^2 u &= f \qquad \text{in } \Omega, \\ \partial_n u - iku &= g \qquad \text{on } \partial\Omega. \end{aligned}$$

where the data g is such that the exact solution is given by

$$u(x) = \begin{cases} \cos(kx) + \frac{1}{k^2} & x \le 0 \\ (1 + \frac{2}{k^2})\cos(kx) - \frac{1}{k^2} & x > 0 \end{cases}$$

Standard regularity theory gives $u \in H^{2.5-\varepsilon}(\Omega)$ for every $\varepsilon > 0$. For the h-FEM we expect $O(h^{\min(2+0.5,p+1)})$. In fact for $p > 1$ one can show (cf. [7, Cor. 4.6]) that $k \left\| u - u_h^{\mathrm{FEM}} \right\|_{L^2(\Omega)} \lesssim h^{2.5}$ and, by inspection, $\|u\|_{L^2(\Omega)} = O(1)$ (uniformly in k). It is therefore expedient to plot $k^{3.5} \left\| u - u_h^{\mathrm{FEM}} \right\|_{L^2(\Omega)} / \|u\|_{L^2(\Omega)}$ versus $N_\lambda \sim (kh)$. For the h-FOSLS Corollary 4.3 predicts only $O(h^{\min(1+0.5,p+1)})$. The numerical results show, however, for both methods convergence $O(h^{\min(2.5,p+1)})$. The results are visualized in Fig. 4.3.

Remark 4.8 The numerical results of Example 4.3 visualized in Fig. 4.3 indicate that Corollary 4.3 is in fact suboptimal as it predicts only a convergence $O(h^{1.5})$ while we observe $O(h^{\min(2.5,p+1)})$. A starting point for understanding this better convergence behavior could be two observations: first, the duality argument in Theorem 4.4 is based on the regularity $(\boldsymbol{\psi}, v) \in \boldsymbol{H}^2(\Omega) \times H^2(\Omega)$ of the dual solution

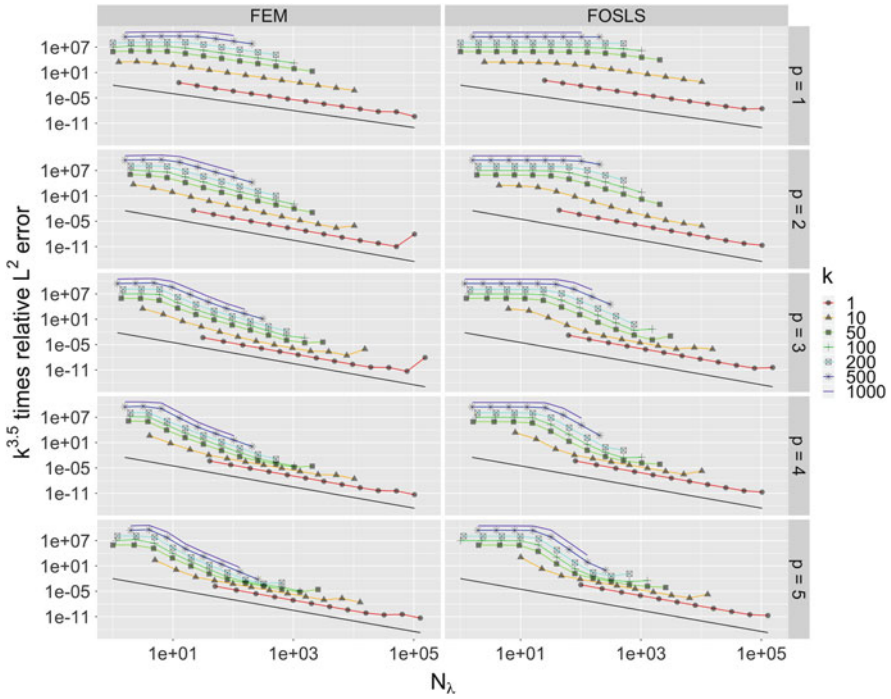


Fig. 4.3 Comparison between the h-FEM (left) and h-FOSLS (right) for $p = 1, \ldots, 5$ as described in Example 4.3. The reference line in black corresponds to $h^{\min(2.5, p+1)}$

$(\boldsymbol{\psi}, v)$ whereas in fact (see the proof of Lemma 4.1) $(\boldsymbol{\psi}, v) \in \boldsymbol{H}^2(\mathrm{div}, \Omega) \times H^2(\Omega)$. Second, a more careful application of the Cauchy-Schwarz inequality (4.36) at the beginning of the proof of Theorem 4.4 is advisable. In this connection, we point to the fact that the terms in the square brackets in (4.36) are not of the same order. To illustrate this, we plot the components

$$e_1 := ik\boldsymbol{e}^{\boldsymbol{\varphi}} + \nabla e^u \quad \text{and} \quad e_2 := ike^u + \nabla \cdot \boldsymbol{e}^{\boldsymbol{\varphi}} \tag{4.39}$$

in Fig. 4.4 for the problem studied in Example 4.3. □

Acknowledgements MB is grateful for the financial support by the Austrian Science Fund (FWF) through the doctoral school *Dissipation and dispersion in nonlinear PDEs* (grant W1245). MB thanks the workgroup of Joachim Schöberl (TU Wien) for help regarding the numerical experiments.

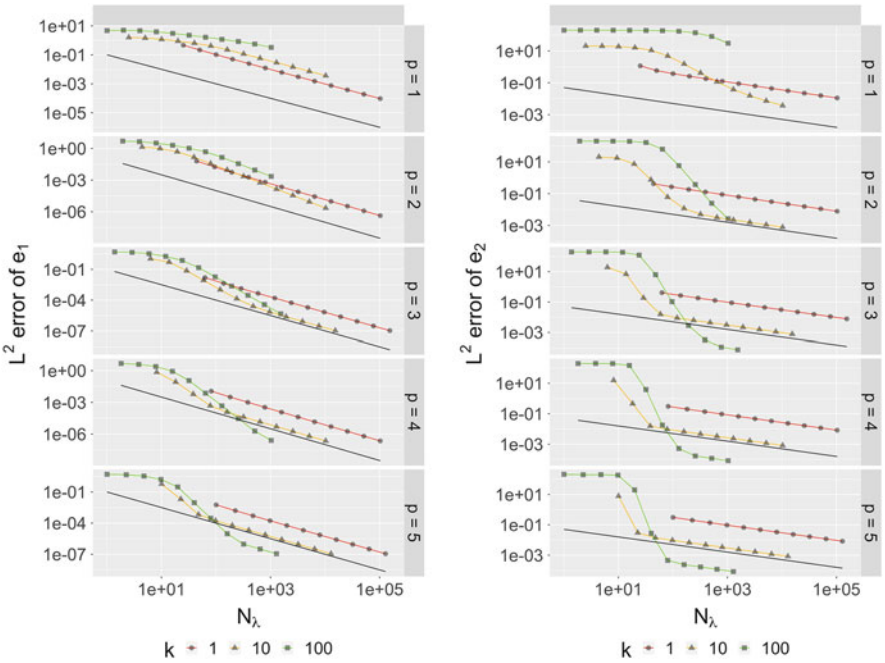


Fig. 4.4 Comparison between the error terms $e_1 := \|ike^{\varphi} + \nabla e^{u}\|_{L^2(\Omega)}$ (left) and $e_2 := \|ike^{u} + \nabla \cdot e^{\varphi}\|_{L^2(\Omega)}$ (right) for $p = 1, \ldots, 5$ as described in Remark 4.8 and Example 4.3. The reference line on the left corresponds to h^1 for $p = 1$ and $h^{1.5}$ for $p > 1$. The reference line on the right corresponds to $h^{1/2}$

References

1. Ainsworth, M.: Discrete dispersion relation for hp-version finite element approximation at high wave number. SIAM J. Numer. Anal. **42**(2), 553–575 (2004)
2. Baskin, D., Spence, E.A., Wunsch, J.: Sharp high-frequency estimates for the Helmholtz equation and applications to boundary integral equations. SIAM J. Math. Anal. **48**(1):229–267 (2016)
3. Chen, H., Qiu, W.: A first order system least squares method for the Helmholtz equation. J. Comput. Appl. Math. **309**, 145–162 (2017)
4. Chen, H., Lu, P., Xu, X.: A hybridizable discontinuous Galerkin method for the Helmholtz equation with high wave number. SIAM J. Numer. Anal. **51**(4), 2166–2188 (2013)
5. Demkowicz, L., Gopalakrishnan, J., Muga, I., Zitelli, J.: Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation. Comput. Methods Appl. Mech. Eng. **213/216**, 126–138 (2012)
6. Esterhazy, S., Melenk, J.M.: On stability of discretizations of the Helmholtz equation. In: Numerical Analysis of Multiscale Problems. Lecture Notes in Computational Science and Engineering, vol. 83, pp. 285–324. Springer, Heidelberg (2012)
7. Esterhazy, S., Melenk, J.M.: An analysis of discretizations of the Helmholtz equation in L^2 and in negative norms. Comput. Math. Appl. **67**(4), 830–853 (2014)
8. Feng, X., Wu, H.: Discontinuous Galerkin methods for the Helmholtz equation with large wave number. SIAM J. Numer. Anal. **47**(4), 2872–2896 (2009)

9. Feng, X., Wu, H.: hp-discontinuous Galerkin methods for the Helmholtz equation with large wave number. Math. Comput. **80**, 1997–2024 (2011)
10. Feng, X., Xing, Y.: Absolutely stable local discontinuous Galerkin methods for the Helmholtz equation with large wave number. Math. Comput. **82**(283), 1269–1296 (2013)
11. Ihlenburg, F.: Finite Element Analysis of Acoustic Scattering. Applied Mathematical Sciences, vol. 132. Springer Verlag, New York (1998)
12. Ihlenburg, F., Babuška, I.: Finite element solution to the Helmholtz equation with high wave number. Part I: The h-version of the FEM. Comput. Math. Appl. **30**, 9–37 (1995)
13. Ihlenburg, F., Babuška, I.: Finite element solution to the Helmholtz equation with high wave number. Part II: The hp-version of the FEM. SIAM J. Numer. Anal. **34**, 315–358 (1997)
14. Jensen, S., Suri, M.: On the l2 error for the p-version of the finite element method over polygonal domains. Comput. Methods Appl. Mech. Eng. **97**(2), 233–243 (1992)
15. Lee, B., Manteuffel, T.A., McCormick, S.F., Ruge, J.: First-order system least-squares for the Helmholtz equation. SIAM J. Sci. Comput. **21**(5), 1927–1949 (2000). Iterative methods for solving systems of algebraic equations (Copper Mountain, CO, 1998)
16. Melenk, J.M.: hp-finite Element Methods for Singular Perturbations. Lecture Notes in Mathematics, vol. 1796. Springer-Verlag, Berlin (2002)
17. Melenk, J.M.: On approximation in meshless methods. In: Frontiers of Numerical Analysis. Universitext, pp. 65–141. Springer, Berlin (2005)
18. Melenk, J.M., Rojik, C: On commuting p-version projection-based interpolation on tetrahedra (2018). arXiv:1802.00197
19. Melenk, J.M., Sauter, S.: Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. Math. Comput. **79**(272), 1871–1914 (2010)
20. Melenk, J.M., Sauter, S.: Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. SIAM J. Numer. Anal. **49**(3), 1210–1243 (2011)
21. Melenk, J.M., Sauter, S.: Wavenumber-explicit hp-FEM analysis for Maxwell's equations with transparent boundary conditions (2018). arXiv:1803.01619
22. Melenk, J.M., Parsania, A., Sauter, S.: General dg-methods for highly indefinite helmholtz problems. J. Sci. Comput. **57**(3), 536–581 (2013)
23. Monk, P.: Finite Element Methods for Maxwell's Equations. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2003)
24. Petrides, S., Demkowicz, L.F.: An adaptive DPG method for high frequency time-harmonic wave propagation problems. Comput. Math. Appl. **74**(8), 1999–2017 (2017)
25. Schöberl, J.: Finite element software NETGEN/NGSolve version 6.2. https://ngsolve.org/
26. Schöberl, J.: NETGEN – an advancing front 2D/3D-mesh generator based on abstract rules. Comput. Vis. Sci. **1**(1), 41–52 (1997)
27. Schwab, C.: P- and Hp-Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics. G.H.Golub and others. Clarendon Press, New York (1998)
28. Zhu, L., Wu, H.: Preasymptotic error analysis of CIP-FEM and FEM for Helmholtz equation with high wave number. Part II: hp version. SIAM J. Numer. Anal. **51**(3), 1828–1852 (2013)

Chapter 5
Numerical Study of Goal-Oriented Error Control for Stabilized Finite Element Methods

Marius Paul Bruchhäuser, Kristina Schwegler, and Markus Bause

Abstract The efficient and reliable approximation of convection-dominated problems continues to remain a challenging task. To overcome the difficulties associated with the discretization of convection-dominated equations, stabilization techniques and a posteriori error control mechanisms with mesh adaptivity were developed and studied in the past. Here we combine the Dual Weighted Residual (DWR) method for goal-oriented error control with stabilized finite element methods. By a duality argument an error representation is derived on that an adaptive strategy is built. The key ingredient of this work is the application of a higher-order discretization of the dual problem in order to make a robust error control for user-chosen quantities of interest feasible. By numerical experiments in 2D and 3D we illustrate that this interpretation of the DWR methodology is capable to resolve layers and sharp fronts with high accuracy and to further reduce spurious oscillations.

5.1 Introduction

From the second half of the last century to nowadays, especially in the pioneering works of the 1980's (cf., e.g., [16, 24]), strong efforts and great progress were made in the development of accurate and efficient approximation schemes for convection-dominated problems; cf., e.g., [38].

The solutions of convection-dominated transport problems are typically characterized by the occurrence of sharp moving fronts and interior or boundary layers. The key challenge for the accurate numerical approximation of these solutions is thus the development of discretization schemes with the ability to capture strong gradients of solutions without producing spurious oscillations or smearing effects. As shown in a comparative study for time-dependent convection-diffusion-reaction equations in [28], conventional stabilization techniques based on standard meshes

M. P. Bruchhäuser (✉) · K. Schwegler · M. Bause
Helmut Schmidt University, Hamburg, Germany
e-mail: bruchhaeuser@hsu-hh.de; bause@hsu-hh.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_5

fail to avoid these oscillations even after a careful fine-tuning of the respective stabilization parameters. As further shown in [28], an alternative in reducing oscillations is obtained by using flux-corrected transport (FCT) schemes [33, 34], that work on the algebraic level. Recently, numerical analysis of these methods were introduced and applied to steady convection-diffusion-reaction equations; cf. [6, 7]. In [14] an hp-adaptive FCT algorithm for unsteady convection equations is presented. Although these FCT schemes show a significant reduction of the unphysical oscillations, smearing effects still arise and can be observed; cf. [14, 28]. Similar results have been perceived in three dimensions; cf. [29].

A further and widespread technique to capture singular phenomena and sharp profiles of solutions is the application of adaptive mesh refinement based on an a posteriori error control mechanism. For a review of a posteriori error estimation techniques for finite element methods and automatic mesh generation we refer, for instance, to the monograph [41]. The design of an adaptive method requires the availability of an appropriate a posteriori error estimator. One possible technique is the commonly used Dual Weighted Residual (DWR) method [4, 11, 12], where the error is estimated in an arbitrary user-chosen target quantity of physical interest. Early studies for stationary convection-reaction equations combined with adaptive mesh refinement have been considered in [22, 36] and [12, Section 3.3]. The DWR approach together with local projection stabilization (LPS) was applied to the steady Navier-Stokes equations in [10] as well as together with LPS and the streamline upwind Petrov-Galerkin (SUPG) method to the nonstationary Navier-Stokes equations in [13, 39]. The one-dimensional case for steady convection-diffusion equations was investigated in [32].

In this work we combine the DWR approach with SUPG stabilized approximations of convection-diffusion-reaction problems. For simplicity, we restrict ourselves to stationary convection-dominated problems here. This is done in order to focus on the interaction of stabilization and error control. Even though several investigations have been done for similar problems, we still expect potential for improvements with respect to accuracy and efficiency, especially considering Péclet numbers that are largely higher than 10^3. Therefore, in contrast to most of the works above, we solve the dual problem by using higher-order finite element techniques. Our motivation also comes through the work of Lube et al. [35], in that the positive impacts of using higher-order finite elements together with stabilized Galerkin methods were investigated. Due to the specific character of convection-dominated problems our computational experience is that the error control needs a particular care in regions with layers and sharp fronts in order to get an accurate quantification of the numerical errors. In numerical experiments we will illustrate the impact of the proper choice of the weights and give a comparison to the common used approximation by higher-order interpolation. The key motivation in this work is to reduce sources of inaccuracies and non-sharp estimates within the error representation as far as possible in order to avoid numerical artefacts. In [19, 37], in particular higher-order finite elements are used to approximate the dual solution for elliptic problems. But in contrast to this work, in [19, 37] only a weak form of the error estimator is used, which is based on a partition-of-unity technique

to avoid the evaluation of strong residuals and jump terms over element edges. Other weak forms of the DWR approach applied to nonlinear partial differential equations or steady convection-diffusion equations have been investigated in [15] or [32], respectively. Here, we follow the classical way of the DWR philosophy, receiving the error representation on every mesh element by a cell-wise integration by parts.

This work is organized as follows. In Sect. 5.2 we introduce our model problem together with some global assumptions and general concepts. In Sect. 5.3 we derive a localized error representation in terms of a target quantity. An adaptive solution algorithm together with some implementational issues is addressed in Sect. 5.4. Finally, in Sect. 5.5 the results of numerical experiments in two and three space dimensions are presented in order to illustrate the feasibility and potential of the proposed approach.

5.2 Problem Formulation and Stabilized Discretization

In this section we first present our model problem. For completion, we briefly sketch the primal and dual stabilized approximation schemes within the DWR framework.

5.2.1 Model Problem and Variational Formulation

In this work we consider the steady linear convection-diffusion-reaction problem

$$-\nabla \cdot (\varepsilon \nabla u) + \mathbf{b} \cdot \nabla u + \alpha u = f \quad \text{in } \Omega\,, \quad u = 0 \quad \text{on } \partial\Omega\,. \tag{5.1}$$

We assume that $\Omega \subset \mathbf{R}^d$, with $d = 2$ or $d = 3$, is a polygonal or polyhedral bounded domain with Lipschitz boundary $\partial\Omega$. For brevity, problem (5.1) is equipped with homogeneous Dirichlet boundary conditions. In our numerical examples in Sect. 5.5, we also consider other types of boundary conditions. In Remark 5.3, the incorporation of nonhomogeneous Dirichlet and Neumann boundary conditions is briefly addressed.

Here, $0 < \varepsilon \ll 1$ is a small positive diffusion coefficient, $\mathbf{b} \in (H^1(\Omega))^d \cap (L^\infty(\Omega))^d$ is the flow field or convection tensor, $\alpha \in L^\infty(\Omega)$ is the reaction coefficient, and $f \in L^2(\Omega)$ is a given outer source of the unknown scalar quantity u. Furthermore, we assume that the following condition is fulfilled:

$$\nabla \cdot \mathbf{b}(\mathbf{x}) = 0 \;\text{ and }\; \alpha(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \Omega\,. \tag{5.2}$$

It is well known that problem (5.1) along with condition (5.2) admits a unique weak solution $u \in V = H_0^1 := \left\{ v \in H^1(\Omega) \middle| \; v|_{\partial\Omega} = 0 \right\}$ that satisfies the following variational formulation; cf., e.g. [2, 27, 38].

Find $u \in V$ such that

$$A(u)(\varphi) = F(\varphi) \quad \forall \varphi \in V \,, \tag{5.3}$$

where the bilinear form $A : V \times V \mapsto \mathbf{R}$ and the linear form $F : V \mapsto \mathbf{R}$ are

$$A(u)(\varphi) := (\varepsilon \nabla u, \nabla \varphi) + (\mathbf{b} \cdot \nabla u, \varphi) + (\alpha u, \varphi) \,, \quad F(\varphi) := (f, \varphi) \,.$$

Here, $(\cdot, \cdot)$ is the inner product of $L^2(\Omega)$ and $\| \cdot \|$ the associated L^2-norm with $\|v\| = (v, v)^{\frac{1}{2}} = (\int_\Omega |v|^2 d\mathbf{x})^{\frac{1}{2}}$.

5.2.2 The Dual Weighted Residual Approach

The DWR method aims at the control of an error in an arbitrary user-chosen target functional J of physical relevance. To get an error representation with respect to this target functional, an additional dual problem has to be solved. Before we focus on this error representation, we introduce the derivation of the dual problem of (5.3) needed below in the DWR approach. For this, we consider the Euler-Lagrangian method of constrained optimization. For some given functional $J : V \mapsto \mathbf{R}$ we consider solving

$$J(u) = \min\{J(v) \,, \; v \in V \,, \; \text{where } A(v)(\varphi) = F(\varphi) \; \forall \varphi \in V\} \,.$$

For this we define the corresponding Lagrangian functional $\mathscr{L} : V \times V \mapsto \mathbf{R}$ by

$$\mathscr{L}(u, z) := J(u) + F(z) - A(u)(z) \,, \tag{5.4}$$

where we refer to $z \in V$ as the dual variable (or Lagrangian multiplier), cf. [4]. We determine a stationary point $\{u, z\} \in V \times V$ of $\mathscr{L}(\cdot, \cdot)$ by the condition that

$$\mathscr{L}'(u, z)(\psi, \varphi) = 0 \quad \forall \{\psi, \varphi\} \in V \times V \,, \tag{5.5}$$

or, equivalently, by the system of equations that

$$A'(u)(\psi, z) = J'(u)(\psi) \quad \forall \psi \in V \,,$$

$$A(u)(\varphi) = F(\varphi) \qquad \forall \varphi \in V \,,$$

where A' is given by $A'(u)(\psi, z) = (\varepsilon \nabla \psi, \nabla z) + (\mathbf{b} \cdot \nabla \psi, z) + (\alpha \psi, z) = A(\psi)(z)$. Applying integration by parts to the convective term along with the condition (5.2)

yields for $A'(u)(\psi, z)$ the representation that

$$A^*(z)(\psi) := A'(u)(\psi, z) = (\varepsilon \nabla z, \nabla \psi) - (\mathbf{b} \cdot \nabla z, \psi) + (\alpha z, \psi)\,. \tag{5.6}$$

Thus we have the following Euler-Lagrange system.

Find $\{u, z\} \in V \times V$ *such that*

$$A(u)(\varphi) = F(\varphi) \quad \forall \varphi \in V\,, \tag{5.7}$$

$$A^*(z)(\psi) = J(\psi) \quad \forall \psi \in V\,. \tag{5.8}$$

5.2.3 Discretization in Space

Here we present the spatial discretization of (5.1). We use Lagrange type finite element spaces of continuous functions that are piecewise polynomials. For the discretization in space, we consider a decomposition $\mathscr{T}_h$ of the domain Ω into disjoint elements K, such that $\bar{\Omega} = \cup_{K \in \mathscr{T}_h} \bar{K}$. Here, we choose the elements $K \in \mathscr{T}_h$ to be quadrilaterals for $d = 2$ and hexahedrals for $d = 3$. We denote by h_K the diameter of the element K. The global space discretization parameter h is given by $h := \max_{K \in \mathscr{T}_h} h_K$. Our mesh adaptation process yields locally refined and coarsened cells, which is enabled by using hanging nodes [17]. We point out that the global conformity of the finite element approach is preserved since the unknowns at such hanging nodes are eliminated by interpolation between the neighboring 'regular' nodes, cf. [4]. The discrete finite element space is defined as usual,

$$V_h^p := \left\{ v \in V \cap C(\bar{\Omega}) \,\middle|\, v|_K \in \mathscr{Q}_h^p(K)\,, \forall K \in \mathscr{T}_h \right\}, \tag{5.9}$$

where $\mathscr{Q}_h^p(K)$ is the space of polynomials that are of degree less than or equal to p with respect to each variable $x_1, \ldots, x_d$.

5.2.4 Streamline Upwind Petrov-Galerkin Stabilization

In order to reduce spurious and non-physical oscillations of the discrete solutions, we apply the SUPG method [16, 23], a well-known residual based stabilization technique for finite element approximations; cf. [1, 8, 28, 38]. Existing a priori error analysis ensure its convergence in the natural norm of the scheme including the control of the approximation error in streamline direction; cf. [38, Thm. 3.27]. Applying the SUPG approach to the discrete counterpart of (5.7) and (5.8) yields the following stabilized discrete system of equations, which can be found in a similar way in [39, Chapter 3.3.1]. We note that here a so called *first dualize and then stabilize* (FDTS) approach is underlying, where the stabilization is applied to the

discrete dual problem only after its derivation via the Euler-Lagrangian method of constrained optimization; cf. Remark 5.2.

Find $\{u_h, z_h\} \in V_h^p \times V_h^{p+s}$, $s \geq 1$, *such that*

$$A_S(u_h)(\varphi_h) = F(\varphi_h) \quad \forall \varphi_h \in V_h^p\,, \tag{5.10}$$

$$A_S^*(z_h)(\psi_h) = J(\psi_h) \quad \forall \psi_h \in V_h^{p+s}\,, s \geq 1\,, \tag{5.11}$$

where the stabilized bilinear forms are given by

$$A_S(u_h)(\varphi_h) := A(u_h)(\varphi_h) + S(u_h)(\varphi_h)\,,$$
$$A_S^*(z_h)(\psi_h) := A^*(z_h)(\psi_h) + S^*(z_h)(\psi_h)\,,$$

and the stabilized terms are defined by

$$S(u_h)(\varphi_h) := \sum_{K \in \mathscr{T}_h} \delta_K(-\nabla \cdot (\varepsilon \nabla u_h) + \mathbf{b} \cdot \nabla u_h + \alpha u_h - f, \mathbf{b} \cdot \nabla \varphi_h)_K\,,$$

$$S^*(z_h)(\psi_h) := \sum_{K \in \mathscr{T}_h} \delta_K^*(-\nabla \cdot (\varepsilon \nabla z_h) - \mathbf{b} \cdot \nabla z_h + \alpha z_h - j(u_h), -\mathbf{b} \cdot \nabla \psi_h)_K\,.$$

Remark 5.1 The proper choice of the stabilization parameters δ_K and δ_K^* is an important issue in the application of the SUPG approach; cf. [25, 26] and the discussion therein. As proposed by the analysis of stabilized finite element methods in [8, 35], we choose the parameter δ_K and δ_K^* as

$$\delta_K, \delta_K^* \sim \min\left\{ \frac{h_K}{p\|\mathbf{b}\|_{L^\infty(K)}}; \frac{h_K^2}{p^4 \varepsilon}; \frac{1}{\alpha} \right\}.$$

Here, the symbol $\sim$ denotes the equivalence up to a multiplicative constant independent of K. This constant has to be understood as a numerical tuning parameter.

Remark 5.2 We note that the stabilization within the dual problem acts in the negative direction of the flow field $\mathbf{b}$; cf. Eq. (5.6). The discrete dual problem (5.11) is based on a *first dualize and then stabilize (FDTS)* principle in that the dual problem of the weak Eq. (5.8) is derived first. The SUPG stabilization is then applied to the discrete counterpart of the dual problem. The alternative strategy, *first stabilize and then dualize (FSTD)*, of transposing the stabilized fully discrete Eq. (5.10) requires differentiation of the stabilization terms. In general, the strategies *FDTS* and *FSTD* do not commute with each other, due to the presence of the stabilization terms in the discrete Lagrangian functional. In our performed numerical experiments the *FSTD* strategy did not show any lack of stability but led to slightly weaker results; cf. [40]. For these reasons we focus on the *FDTS* strategy only in this work.

5.3 Error Estimation

In this section we present a localized (i.e. elementwise) a posteriori error representation for the stabilized finite element approximation in terms of the goal quantity $J(\cdot)$ by using the concepts of the DWR approach introduced in Sect. 5.2.2. In order to keep this work self-contained we briefly summarize the key arguments of the DWR approach applied to the stabilized model problem. We follow the lines of [4, Chapter 6] and [10], where all of the proofs can be found. To start with, we put

$$x := \{u, z\}\,,\ y := \{\psi, \varphi\} \in V \times V\,, \quad x_h := \{u_h, z_h\}\,,\ y_h := \{\psi_h, \varphi_h\} \in V_h^p \times V_h^p\,,$$

and

$$\tilde{S}(x_h)(y_h) := S(u_h)(\varphi_h) + S^*(z_h)(\psi_h)\,.$$

The discrete solution $x_h \in V_h^p \times V_h^p$ then satisfies the variational equation

$$\mathscr{L}'(x_h)(y_h) = \tilde{S}(x_h)(y_h) \quad \forall y_h \in V_h^p \times V_h^p\,. \tag{5.12}$$

Now we develop the error in terms of the Lagrangian functional.

Theorem 5.1 *Let X be a function space and $\mathscr{L} : X \to \mathbf{R}$ be a three times differentiable functional on X. Suppose that $x_c \in X_c$ with some ("continuous") function space $X_c \subset X$ is a stationary point of $\mathscr{L}$. Suppose that $x_d \in X_d$ with some ("discrete") function space $X_d \subset X$, with not necessarily $X_d \subset X_c$, is a Galerkin approximation to x_c being defined by the equation*

$$\mathscr{L}'(x_d)(y_d) = \tilde{S}(x_d)(y_d) \quad \forall y_d \in X_d\,.$$

In addition, suppose that the auxiliary condition $\mathscr{L}'(x_c)(x_d) = 0$ is satisfied. Then there holds the error representation

$$\mathscr{L}(x_c) - \mathscr{L}(x_d) = \frac{1}{2}\mathscr{L}'(x_d)(x_c - y_d) + \frac{1}{2}\tilde{S}(x_d)(y_d - x_d) + \mathscr{R} \quad \forall y_d \in X_d\,,$$

where the remainder $\mathscr{R}$ is defined by $\mathscr{R} = \frac{1}{2}\int_0^1 \mathscr{L}'''(x_d + se)(e, e, e) \cdot s \cdot (s-1)\,\mathrm{d}s$, with the notation $e := x_c - x_d$.

For the subsequent theorem we introduce the primal and dual residuals by

$$\rho(u_h)(\varphi) := F(\varphi) - A(u_h)(\varphi) \qquad \forall \varphi \in V\,, \tag{5.13}$$

$$\rho^*(z_h)(\psi) := J'(u_h)(\psi) - A^*(z_h)(\psi) \quad \forall \psi \in V\,. \tag{5.14}$$

Theorem 5.2 *Suppose that $\{u, z\} \in V \times V$ is a stationary point of the Lagrangian functional $\mathscr{L}$ defined in (5.4) such that (5.5) is satisfied. Let $\{u_h, z_h\} \in V_h^p \times$*

V_h^p denote its Galerkin approximation being defined by (5.10) and (5.11) such that (5.12) is satisfied. Then there holds the error representation that

$$J(u) - J(u_h) = \frac{1}{2}\rho(u_h)(z - \varphi_h) + \frac{1}{2}\rho^*(z_h)(u - \psi_h) + \mathscr{R}_{\tilde{S}} + \mathscr{R}_J \,, \tag{5.15}$$

for arbitrary functions $\{\varphi_h, \psi_h\} \in V_h^p \times V_h^p$, where the remainder terms are $\mathscr{R}_{\tilde{S}} := \frac{1}{2}S(u_h)(\varphi_h + z_h) + \frac{1}{2}S^*(z_h)(\psi_h - u_h)$ and $\mathscr{R}_J := \frac{1}{2}\int_0^1 J'''(u_h + s \cdot e)(e, e, e) \cdot s \cdot (s - 1)\,\mathrm{d}s$, with $e = u - u_h$.

In the error representation (5.15) the continuous solution u is required for the evaluation of the dual residual. The following theorem shows the equivalence of the primal and dual residual up to a quadratic remainder. This observation will be used below to find our final error representation in terms of the goal quantity J and a suitable linearization for its computational evaluation or approximation, respectively.

Theorem 5.3 *Under the assumptions of Theorem 5.2, and with the definitions (5.13) and (5.14) of the primal and dual residual, respectively, there holds that*

$$\rho^*(z_h)(u - \psi_h) = \rho(u_h)(z - \varphi_h) + S(u_h)(\varphi_h - z_h) + S^*(z_h)(u_h - \psi_h) + \Delta\rho_J \,,$$

for all $\{\psi_h, \varphi_h\} \in V_h^p \times V_h^p$, where the remainder term is given by $\Delta\rho_J := -\int_0^1 J''(u_h + s \cdot e)(e, e)\,\mathrm{d}s$ with $e := u - u_h$.

We summarize the results of the previous two theorems in the following corollary.

Corollary 5.1 *Under the assumptions of Theorem 5.2 with the definitions (5.13) and (5.14) of the primal and dual residual, respectively, there holds the error representation that*

$$J(u) - J(u_h) = \rho(u_h)(z - \varphi_h) + S(u_h)(\varphi_h) + \mathscr{R}_J + \frac{1}{2}\Delta\rho_J \,, \tag{5.16}$$

for arbitrary functions $\varphi_h \in V_h^p$, where the remainder term $\mathscr{R}_J$ is given by Theorem 5.2 and the linearization error $\Delta\rho_J$ is defined in Theorem 5.3.

In a final step we present a localized approximation of the error that is then used for the design of the adaptive algorithm. We note that the final result (5.17) is a slight modification of Theorem 4.2 for the case $\varepsilon = 0$ in [36] or Prop. 3.3 for the case $\varepsilon = \alpha = 0$ in [12, Section 3.3]. The difference to these references comes through using a FDTS approach here, cf. Remark 5.2. Similar results can also be found in [22].

Theorem 5.4 (Localized Error Representation) *Let the assumptions of Theorem 5.2 be satisfied. Neglecting the higher-order error terms in (5.16), then there*

holds as a linear approximation the cell-wise error representation

$$J(u) - J(u_h) \doteq \sum_{K \in \mathscr{T}_h} \Big\{ \big(R(u_h), z - \varphi_h\big)_K - \delta_K \big(R(u_h), \mathbf{b} \cdot \nabla \varphi_h\big)_K - \big(E(u_h), z - \varphi_h\big)_{\partial K} \Big\} . \tag{5.17}$$

The cell- and edge-wise residuals are defined by

$$R(u_h)_{|K} := f + \nabla \cdot (\varepsilon \nabla u_h) - \mathbf{b} \cdot \nabla u_h - \alpha u_h \,, \tag{5.18}$$

$$E(u_h)_{|\Gamma} := \begin{cases} \frac{1}{2} \mathbf{n} \cdot [\varepsilon \nabla u_h] & \text{if } \Gamma \subset \partial K \backslash \partial \Omega \,, \\ 0 & \text{if } \Gamma \subset \partial \Omega \,, \end{cases} \tag{5.19}$$

where $[\nabla u_h] := \nabla u_{h|\Gamma \cap K} - \nabla u_{h|\Gamma \cap K'}$ *defines the jump of* ∇u_h *over the inner edges* Γ *with normal unit vector* $\mathbf{n}$ *pointing from* K *to* K'.

Proof The assertion directly follows from (5.16) by neglecting the higher-order remainder terms $\mathscr{R}_J$ and $\Delta \rho_J$ as well as applying integration by parts on each cell $K \in \mathscr{T}_h$ to the diffusion term in the primal residual (5.13). □

Remark 5.3 (Nonhomogeneous Dirichlet and Neumann Boundary Conditions) We briefly address the incorporation of further types of boundary conditions. First, we consider problem (5.1) equipped with the nonhomogeneous Dirichlet condition $u = g_D$ on $\partial \Omega$, for a given function $g \in H^{\frac{1}{2}}(\partial \Omega)$. For this, let $\tilde{g}_D \in H^1(\Omega)$ be an extension of g_D in the sense that the trace of $\tilde{g}_D$ equals g_D on $\partial \Omega$. Further, let the discrete function $\tilde{g}_{D,h}$ be an appropriate finite element approximation of the extension $\tilde{g}_D$. Then, the trace on $\partial \Omega$ of $\tilde{g}_{D,h}$ represents a discretization of g_D. For instance, a nodal interpolation of g_D and an extension in the finite element space can be used. This allows us to recast the weak form of problem (5.1) and its discrete counterpart in terms of $w = u - \tilde{g}_D \in H_0^1(\Omega)$ and $w_h = u_h - \tilde{g}_{D,h} \in V_h^p \subset H_0^1(\Omega)$. The previous calculations and the derivation of the a posteriori error estimator are then done for the weak problem and its discrete counterpart rewritten in terms of w and w_h. This yields the result that

$$J(u) - J(u_h) \doteq \sum_{K \in \mathscr{T}_h} \Big\{ \big(R(u_h), z - \varphi_h\big)_K - \delta_K \big(R(u_h), \mathbf{b} \cdot \nabla \varphi_h\big)_K - \big(E(u_h), z - \varphi_h\big)_{\partial K} \Big\} - \big((g_D - \tilde{g}_{D,h}), \varepsilon \nabla z \cdot \mathbf{n}\big)_{\partial \Omega} \,,$$

where $R(u_h)$ and $E(u_h)$ are given by (5.18) and (5.19), respectively. If a homogeneous Neumann condition is prescribed on a part $\partial \Omega_N$ of the boundary $\partial \Omega = \partial \Omega_D \cup \partial \Omega_N$, with Dirichlet part $\partial \Omega_D$, then the derivation has to be done analogously for the solution space $V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\}$ and its discrete counterpart and the resulting variational problems.

5.4 Practical Aspects

In this section we present some practical aspects regarding the application of the result given in Theorem 5.4 in computational studies of convection-dominated problems. We note that the concepts described here can be generalized to nonstationary problems; cf. [31]. For the nonstationary Navier-Stokes equations along with local projection stabilization (LPS), numerical investigations can be found in [13, 39] for Péclet numbers up to a magnitude of 10^3. The error representation (5.17), rewritten as

$$J(u) - J(u_h) \doteq \sum_{K \in \mathscr{T}_h} \Big\{ \big(R(u_h), z - \varphi_h\big)_K - \delta_K \big(R(u_h), \mathbf{b} \cdot \nabla \varphi_h\big)_K - \big(E(u_h), z - \varphi_h\big)_{\partial K} \Big\} = \eta := \sum_{K \in \mathscr{T}_h} \eta_K \,, \tag{5.20}$$

depends on the discrete primal solution u_h as well as on the exact dual solution z. For the application of (5.20) in computations, the unknown dual solution z has to be approximated, which results in an approximate error indicator $\tilde{\eta}$. This approximation cannot be done in the same finite element space as used for the primal problem, since this would result in an useless vanishing error representation $\tilde{\eta} = 0$, due to Galerkin orthogonality. As a key ingredient of this work, we use an approximation in a finite element space that is of higher-order compared to that one of solving the primal problem, which however leads to higher computational costs; cf. [30] for algorithmic formulations and analyses. In the literature, the application of higher-order interpolation instead of usage of higher-order finite element spaces is often suggested for the DWR approach; cf. [4, 10, 13]. For convection-dominated problems such an interpolation might be defective and lead to a loss of accuracy of the underlying error estimator. Higher-order techniques show more stability and gain with regard to the accuracy of the error estimator, due to the more accurate approximation of the weights. In Example 1 of Sect. 5.5 a comparative study between higher-order interpolation and higher-order finite elements is given.

In order to define the localized error contributions $\tilde{\eta}_K$ we consider a hierarchy of sequentially refined meshes $\mathscr{M}_i$, with $i \geq 1$ indexing the hierarchy. The corresponding finite element spaces are denoted by $V_h^{p+s,i}$, $s \geq 1$, cf. (5.9). We calculate the cell-wise contributions to the linearized error representation (5.20) by means of

$$\tilde{\eta}_K = \big(R(u_h^i), z_h^i - I_h z_h^i\big)_K - \delta_K \big(R(u_h^i), \mathbf{b} \cdot \nabla I_h z_h^i\big)_K - \big(E(u_h^i), z_h^i - I_h z_h^i\big)_{\partial K} \,, \tag{5.21}$$

where the cell and edge residuals are given in (5.18) and (5.19), respectively. By $I_h z_h^i \in V_h^{p,i}$ we denote the nodal based Lagrange interpolation of the higher-order

approximation $z_h^i \in V_h^{p+s,i}$, $s \geq 1$ into the lower order finite element space V_h^p. Our adaptive mesh refinement algorithm based on (5.21) is summarized in the following.

Adaptive Solution Algorithm (Refining and Coarsening)

Initialization Set $i = 0$ and generate the initial finite element spaces for the primal and dual problem.

1. Solve the **primal** problem: Find $u_h^i \in V_h^{p,i}$ such that

$$A_S(u_h^i)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in V_h^{p,i} .$$

2. Solve the **dual** problem: Find $z_h^i \in V_h^{p+s,i} \supset V_h^{p,i}$, $s \geq 1$, such that

$$A_S^*(z_h^i)(\psi_h) = J(\psi_h) \qquad \forall \psi_h \in V_h^{p+s,i}, s \geq 1 .$$

   Here, $V_h^{p+s,i}$ denotes the finite element space of piecewise polynomials of higher-order on the mesh $\mathscr{M}_i$.
3. Evaluate the **a posteriori** error indicator

$$\tilde{\eta} := \sum_{K \in \mathscr{T}_h} \tilde{\eta}_K , \text{ with } \quad \tilde{\eta}_K = \big(R(u_H^i), z_h^i - I_h z_h^i\big)_K - \delta_K \big(R(u_H^i), \mathbf{b} \cdot \nabla I_h z_h^i\big)_K$$
$$- \big(E(u_H^i), z_h^i - I_h z_h^i\big)_{\partial K} ,$$

   where the cell and edge residuals are given in (5.18) and (5.19). By u_H^i we denote the nodal based Lagrange interpolation of u_h^i in $V_h^{p+s,i}$. Further, z_h^i is the computed dual solution and $I_h z_h^i$ is the interpolation of z_h^i in the finite element space $V_h^{p,i}$ of the primal problem.
4. Histogram based refinement strategy:
   Choose $\theta \in (0.25, 5)$. **Put** $\tilde{\eta}_{\max} = \max\limits_{K \in \mathscr{T}_h} |\tilde{\eta}_K|$ and $\mu = \theta \sum_{K \in \mathscr{T}_h} |\tilde{\eta}_K| / \#K$.
   While $\mu > \tilde{\eta}_{\max}$: Set $\mu := \frac{\mu}{2}$. Mark the elements $\tilde{K}$ with $|\tilde{\eta}_{\tilde{K}}| > \mu$ to be refined and those 2% of the elements $\hat{K}$ that provide the smallest contribution to $\tilde{\eta}$ to be coarsened. Generate a new mesh $\mathscr{M}_{i+1}$ by regular coarsening and refinement.
5. Check the stopping condition:
   If $\tilde{\eta}_{\max} < \text{tol}$ or $\tilde{\eta} < \text{tol}$ is satisfied, then the adaptive solution algorithm is terminated; Else, i is increased to $i + 1$ and it is jumped back to Step 1.

Remark 5.4 Regarding the choice of the numerical tuning parameter θ in Step 4 of the previous algorithm we made the computational experience that a value of θ between 0.25 and 5 typically leads to good results. Further, we note that the performance properties of adaptive algorithms are strongly affected by the marking strategy. The so called *Dörfler* marking (cf. [18]) or the marking of the largest local

error indicators represent further popular marking strategies. For a further discussion of this issue we refer to, e.g., [4].

Remark 5.5 According to the adaptive solution algorithm presented above, we use the same mesh for solving the primal and dual problem, more precisely we use the same triangulation for both problems, but different polynomial degrees for the underlying shape functions of the respective finite element space.

5.5 Numerical Studies

In this section we illustrate and investigate the performance properties of the proposed approach of combining the Dual Weighted Residual method with stabilized finite element approximations of convection-dominated problems. We demonstrate the potential of the DWR method with regard to resolving solution profiles admitting sharp layers as they arise in convection-dominated problems. Further, we investigate the mesh adaptation processes by prescribing various target functionals or goal quantities, respectively. For this, standard benchmark problems of the literature for studying the approximation of convection-dominated transport are applied. For the implementation and our numerical computations we use our `DTM++` frontend software [30, Chapter 4] that is based on the open source finite element library `deal.II`; cf. [3, 5]. For measuring the accuracy of the error estimator, we will study in our numerical experiments the effectivity index

$$\mathscr{I}_{\text{eff}} = \left| \frac{\tilde{\eta}}{J(u) - J(u_h)} \right| \tag{5.22}$$

as the ratio of the estimated error $\tilde{\eta}$ of (5.20) over the exact error. Desirably, the index $\mathscr{I}_{\text{eff}}$ should be close to one.

Example 1 (Hump with Circularly Layer, 2d) In the first numerical experiment we focus on studying the accuracy of our error estimator and the impact of approximating the weights of the dual solution within the error indicators (5.21). For this, we consider two different approaches for approximating the dual solution within the error representation and investigate several combinations of polynomial orders for the finite element spaces of the primal and dual solution. We study problem (5.1) with the prescribed solution (cf. [1, 8, 28])

$$u(\mathbf{x}) = 16x_1(1-x_1)x_2(1-x_2) \cdot \left\{ \frac{1}{2} + \frac{\arctan\left(2\varepsilon^{-1/2}\left[r_0^2 - (x_1 - x_1^0)^2 - (x_2 - x_2^0)^2\right]\right)}{\pi} \right\}. \tag{5.23}$$

where $\Omega := (0,1)^2$ and $r_0 = 0.25$, $x_1^0 = x_2^0 = 0.5$. We choose the flow field $\mathbf{b} = (2,3)^\top$ and the reaction coefficient $\alpha = 1.0$. For the solution (5.23) the right-

hand side function f is calculated from the partial differential equation. Boundary conditions are given by the exact solution. Our target quantity is chosen as

$$J_{L^2}(u) = \frac{1}{\|e\|_{L^2(\Omega)}}(e, u)\,. \tag{5.24}$$

In a first computational study we investigate the approximation of the exact dual solution z within the error representation (5.17) on the one hand by higher-order interpolation, and by higher-order finite elements; cf. [4] for more details. In Fig. 5.1 we visualize the respective effectivity indices for varying diffusion coefficients. While the values are quite similar for a comparatively large diffusion coefficient, the difference increases if ε becomes smaller. This confirms our assumption at the beginning, obtaining better results with regard to the accuracy for the underlying error estimator using higher-order finite elements for the approximation of the dual solution. In the sequel, all following examples are performed using this higher-order finite element method and the value $\varepsilon = 10^{-6}$, unless otherwise specified.

In Fig. 5.2a we compare the convergence behavior of the proposed DWR approach with a global mesh refinement strategy. The corresponding solution

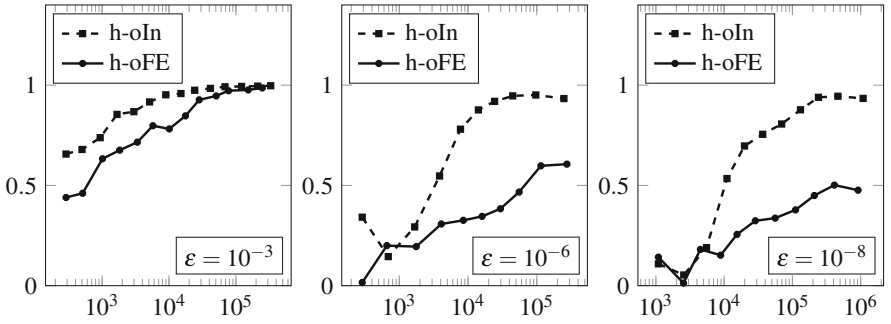


Fig. 5.1 Comparison of effectivity indices (over degrees of freedom) with regard to higher-order interpolation (h-oIn) versus higher-order finite elements (h-oFE) for varying diffusion coefficients ε with target quantity (5.24) for Example 1

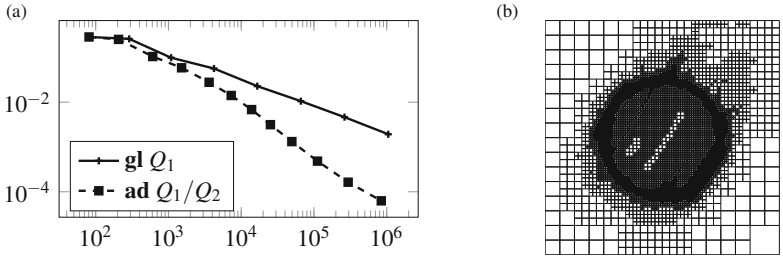


Fig. 5.2 Error comparison and visualization of an adaptive mesh for Example 1 (**a**) L^2-error over degrees of freedom for global and DWR adaptive mesh refinement (**b**) Adaptive mesh for target quantity (24) with 56222 degrees of freedom

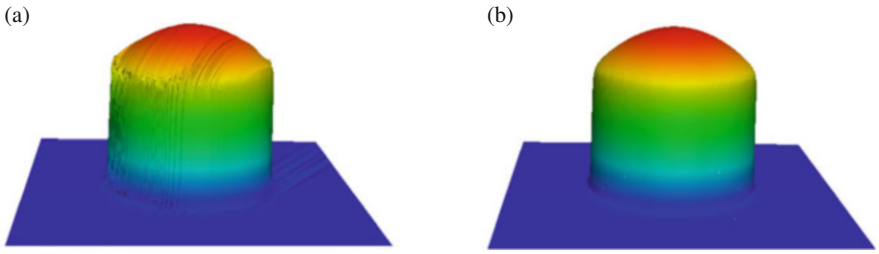


Fig. 5.3 Stabilized solution profile on a globally refined mesh with 66,049 degrees of freedom (**a**) and on an adaptively refined mesh with error control by the target quantity (5.24) with 56,222 degrees of freedom (**b**) for Example 1

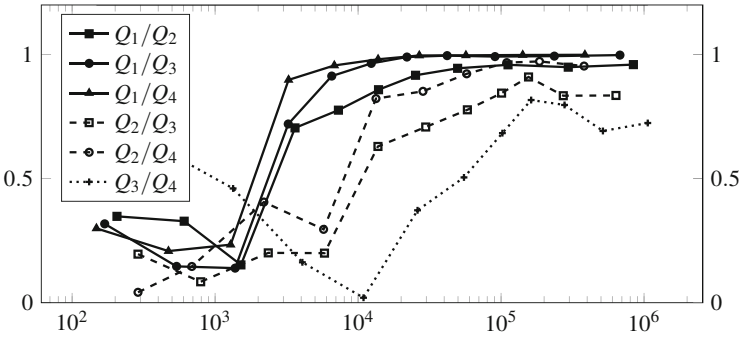


Fig. 5.4 Effectivity indices over degrees of freedom for the target quantity (5.24) and different polynomial degrees for Example 1

profiles are visualized in Fig. 5.3. The adaptively generated mesh is presented in Fig. 5.2b. The DWR based adaptive mesh adaptation is clearly superior to the global refinement in terms of accuracy over degrees of freedom. While the globally refined solution is still perturbed by undesired oscillations within the circular layer and behind the hump in the direction of the flow field **b**, the adaptively computed solution exhibits an almost perfect solution profile for even less degrees of freedom.

In Fig. 5.4 we present the calculated effectivity indices (5.22) for solving the primal and dual problem in different pairs of finite element spaces based on the family of Q_k elements. Considering the Q_1 based approximation of the primal problem, we note that by increasing the polynomial degree of the dual solution from Q_2 to Q_4 the mesh adaptation process reaches the stopping criterion faster and requires less degrees of freedom. This observation is reasonable, since a higher-order approximation of the dual problem is closer to its exact solution of the dual problem, which is part of the error representation (5.17). Thus we conclude that a better approximation of the weights provides a higher accuracy of the error estimator. This observation is also confirmed by the comparison of the pairs of Q_2/Q_3 with Q_2/Q_4 based finite element spaces. Nevertheless, the difference for using higher-order finite elements for solving the dual problem is not that significant,

even less if we take into account the higher computational costs for solving the algebraic form of the dual problem for an increasing order of the piecewise polynomials. Using pairs of Q_k/Q_{k+1} based elements for the approximation of the primal and dual problem, the error estimator gets worse for increasing values of the parameter k. This observation is in good agreement with the results in [19, Example 3]. A reason for this behavior is given by the observation that for increasing values of k the mesh is less refined for the same number of degrees of freedom. Therefore less cells are available to capture the strong gradients of the exact solution. This argues for choosing smaller values of k in the application of our DWR based approach.

Example 2 (Point-Value Error Control, 2d) In this experiment we study the application of our approach for different target functionals within a sequence of decreasing diffusion coefficients. Thereby we aim to analyze the robustness of the approach with respect to the small perturbation parameter ε in (5.1). If convection-dominated problems are considered, or also often for applications of practical interest, local quantities are of greater interest than global ones. The DWR approach offers the appreciable advantage over standard a posteriori error estimators that an error control in an arbitrary user-chosen quantity and not only in the global L^2 norm, as in Example 1, or a norm of energy type can be obtained. Since the error representation is exact up to higher-order terms (cf. Theorem 5.4), robustness with respect to the perturbation parameter ε can be expected to become feasible. Of course, the approximation of the dual solution z in (5.17) adds a source of uncertainty in the error representation. In the sequel, we evaluate the potential of our approach with respect to these topics for different target functionals. As a benchmark problem we consider problem (5.1) for a solution given by (cf. [35, Example 4.2])

$$u(\mathbf{x}) = \frac{1}{2}\left(1 - \tanh\frac{2x_1 - x_2 - 0.25}{\sqrt{5\varepsilon}}\right)$$

with corresponding right-hand side function f. Further, $\Omega = (0,1)^2$, $\alpha = 1.0$, $\mathbf{b} = \frac{1}{\sqrt{5}}(1,2)^\top$. The Dirichlet boundary condition is given by the exact solution. The solution is characterized by an interior layer of thickness $\mathcal{O}(\sqrt{\varepsilon}|\ln\varepsilon|)$. We study the target functionals

$$J_{L^2}(u) = \frac{1}{\|e\|_{L^2(\Omega)}}(e,u)\,, \qquad J_M(u) = \int_\Omega u\,\mathrm{d}\mathbf{x} \qquad \text{and} \qquad J_P(u) = u(\mathbf{x}_e)\,,$$

where $e := u - u_h$ and with a user-prescribed control point $\mathbf{x}_e = \left(\frac{5}{16}, \frac{3}{8}\right)$ that is located in the interior of the layer. In our computations we regularize $J_P(\cdot)$ by

$$J_{rP}(u) = \frac{1}{|B_r|}\int_{B_r} u(\mathbf{x})\,\mathrm{d}\mathbf{x}\,,$$

where the ball B_r is defined by $B_r = \{\mathbf{x} \in \Omega \mid \|\mathbf{x} - \mathbf{x}_e\| < r\}$ with small radius $r > 0$. Here, all test cases are solved by using the Q_1/Q_2 pair of finite elements for the primal and dual problem which is due to the observations depicted in Example 1. In Table 5.1 and Fig. 5.5 we present the effectivity indices of the proposed DWR approach applied to the stabilized approximation scheme (5.10) for a sequence of vanishing diffusion coefficients. For the target functionals $J_{L^2}(\cdot)$ and $J_M(\cdot)$ the effectivity indices nicely converge to one for an increasing number of degrees of freedom. Moreover, the expected convergence behavior is robust with respect to the small diffusion parameter ε. For the more challenging error control of a point-value, which however can be expected to be of higher interest in practice, the effectivity indices also convergences nicely to one. This is in good agreement with effectivity indices for point-value error control that are given in other works of the literature for the Poisson problem only; cf. [4, p. 45]. We note that in the case of a point-value error control the target functional lacks the regularity of the right-hand side term in the dual problem that is typically needed to ensure the existence and regularity of weak solutions; cf. [21, Chapter 6.2]. However, no impact of this lack of regularity is observed in the computational studies. Thus, for all target functionals a robust convergence behavior is ensured for this test case.

For completeness, in Fig. 5.6 we visualize the computed solution profiles and adaptive meshes for an error control based on the local target functional $J_{rP}(\cdot)$ and the global target functional $J_{L^2}(\cdot)$, respectively. This test case nicely illustrates the potential of the DWR approach. For the point-value error control the refined mesh cells are located close to the specified point of interest and along those cells that affect the point-value error by means of transport in the direction of the flow field $\mathbf{b}$. Furthermore, mesh cells without strong impact on the solution close to the control point are coarsened further. Even though a rough approximation of the sharp interface is obtained in downstream direction of the control point, in its neighborhood an excellent approximation of the sharp layer is ensured by the approach. A highly economical mesh along with a high quality in the computation of the user-specified goal quantity is thus obtained. In contrast to this, the global error control of $J_{L^2}(\cdot)$ provides a good approximation of the solution in the whole domain by adjusting the mesh along the complete layer.

Example 3 (Variable Convection Field, 3D) In our last example we apply the approach to a three-dimensional problem case which represents a more challenging task. Moreover, we consider a velocity field $\mathbf{b}$ depending on the space variable $\mathbf{x}$. Precisely, we consider problem (5.1) with the unit cube $\Omega = (0, 1)^3$, $\varepsilon = 10^{-6}$, $\alpha = 1$, $\mathbf{b} = (-x_2, x_1, 0)^\top$ and $f \equiv 0$. The boundary conditions are given by $\frac{\partial u}{\partial \mathbf{n}} = 0$ on $\Gamma_N = \{\mathbf{x} \in \Omega \mid x_1 = 0\}$, $u = 1$ on $\Gamma_{D_1} = \{\mathbf{x} \in \Omega \mid 0.4 \leq x_1 \leq 0.6, x_2 = 0, 0.4 \leq x_3 \leq 0.6\}$, and $u = 0$ on $\Gamma_{D_2} = \partial\Omega \backslash \{\Gamma_N \cup \Gamma_{D_1}\}$. Thus, by the boundary part Γ_{D_1} we model an inflow region (area) where the transport quantity modelled by the unknown u is injected; cf. Fig. 5.7. Γ_N models an outflow boundary. Prescribing a homogeneous Dirichlet condition on Γ_{D_2} is only done for the sake of simplicity. The target functional aims at the control of the solution's mean value in a smaller, inner domain $\Omega_{In} = [0, 0.1] \times [0.4, 0.6] \times [0.4, 0.6]$ close to the outflow boundary,

Table 5.1 Effectivity indices for the target quantities J_{L^2} J_M and J_{rP} and different decreasing perturbation parameters for Example 2

$\varepsilon = 10^{-6}$						$\varepsilon = 10^{-7}$						$\varepsilon = 10^{-8}$					
J_{L^2}		J_M		J_{rP}		J_{L^2}		J_M		J_{rP}		J_{L^2}		J_M		J_{rP}	
dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$	dofs	$\mathscr{I}_{eff}$
4531	0.72	5355	0.14	4334	1.09	1643	0.32	4034	0.12	753	0.73	1477	0.46	2810	0.56	1841	1.00
7603	0.84	9458	1.92	6468	1.16	2772	0.37	8383	6.38	1841	0.99	5668	0.54	10,925	0.90	3301	1.22
13,319	0.92	16,067	1.73	13,851	1.58	5094	0.43	16,314	0.01	6358	1.34	10,005	0.56	22,028	0.05	6411	1.46
24,418	0.98	28,584	0.85	19,670	1.19	10,086	0.53	28,310	0.13	12,674	1.23	17,974	0.59	41,088	1.72	12,904	1.72
42,174	1.00	52,024	3.06	38,002	0.99	21,071	0.67	44,111	0.61	25,558	1.09	41,402	0.66	71,646	0.26	27,039	1.74
76,341	1.01	95,006	1.07	51,603	1.20	46,172	0.84	72,705	1.07	54,549	0.96	90,486	0.78	141,031	1.96	58,254	1.46
126,757	0.99	179,893	1.01	119,171	0.94	103,077	0.97	178,757	0.85	139,531	1.31	253,203	0.90	305,855	1.01	123,436	1.24
224,160	1.01	307,864	1.00	357,046	0.90	240,672	1.02	433,232	1.00	387,749	1.27	502,287	1.00	608,497	1.04	274,188	1.13
409,008	0.99	560,046	1.00	571,577	1.14	580,812	1.00	1003,495	1.00	1181,627	1.01	801,381	1.01	856,320	1.01	691,860	1.08

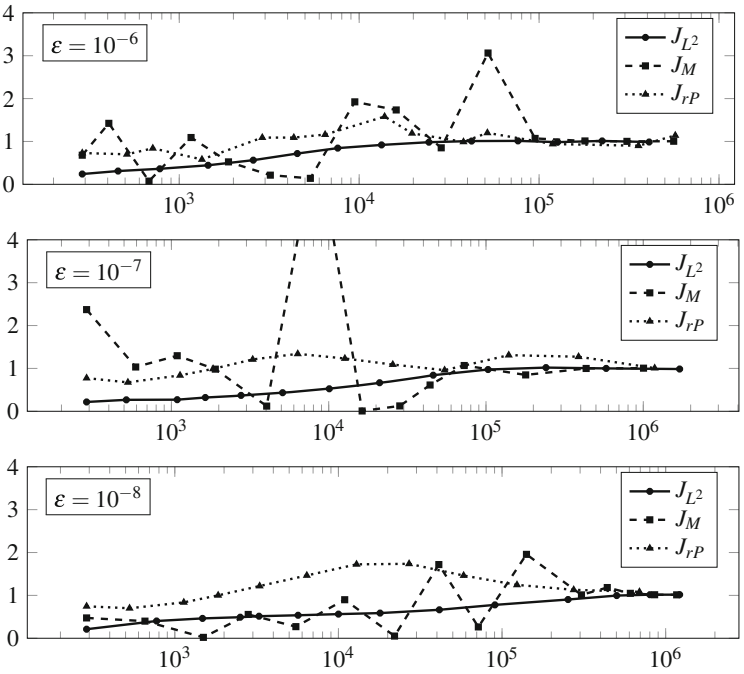


Fig. 5.5 Effectivity indices over degrees of freedom for different target functionals and decreasing diffusion coefficients for Example 2

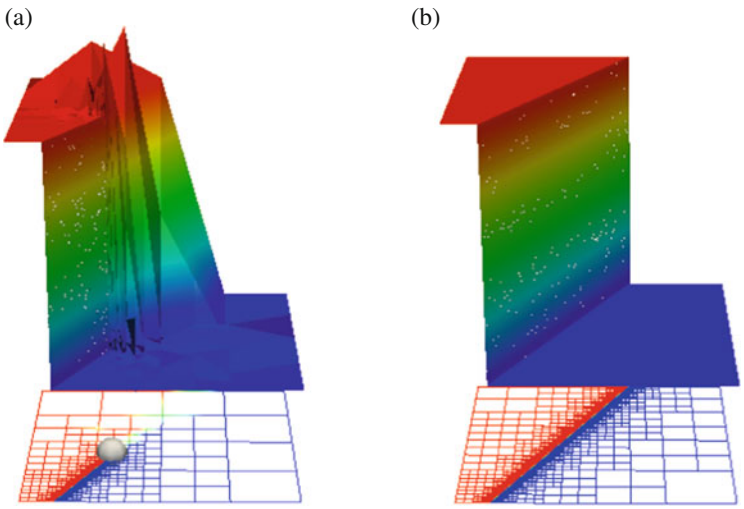


Fig. 5.6 Point-value error control by J_{rP} (**a**) and global error control by J_{L^2} (**b**) by the DWR approach for Example 2

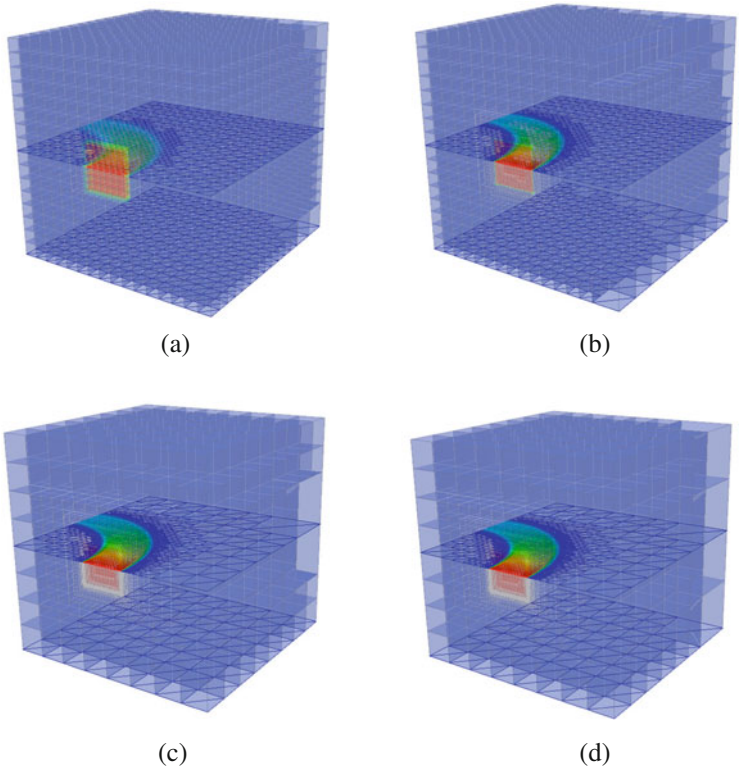


Fig. 5.7 Adaptive grids after first (**a**), third (**b**), fifth (**c**) and seventh (**d**) iteration step of the DWR approach with target functional J_{iM} for Example 3

and is given by

$$J_{iM}(u) = \int_{\Omega_{In}} u \, \mathrm{d}\mathbf{x} \, .$$

In the context of applications, the transport quantity u is thus measured and controlled in the small region of interest Ω_{In}.

Figure 5.7 illustrates the computed adaptively generated meshes for some of the DWR iteration steps. For visualization purposes, two surfaces with corresponding mesh distribution are shown for each of the grids, the bottom surface and the surface in the domain's center with respect to the x_3 direction. We note that the postprocessed solutions are visualized on a grid for the respective surfaces. The cells on the surfaces are triangular-shaped since the underlying visualization software `ParaView` is based on triangular-shaped elements. Similar to the previous test case of a point-value error control, the refinement is located on those cells that contribute to the mean value error control. Here, the cells close to the two inner layers aligned

in the flow direction **b** are strongly refined. This refinement process is obvious since the inner and control domain Ω_{In} is chosen to have exactly the same dimensions as the channel-like extension of the boundary segment Γ_{D_1} along the flow direction into the domain Ω. Outside the inner domain Ω_{In} and the channel-like domain of transport the mesh cells are coarsened for an increasing number of DWR iteration steps.

5.6 Summary and Future Work

In this work we presented an adaptive approach for stabilized finite element approximations of stationary convection-dominated problems. It is based on the Dual Weighted Residual method for goal-oriented a posteriori error control. A *first dualize and then stabilize* philosophy was applied for combining the mesh adaptation process in the course of the DWR approach with the stabilization of the finite element techniques. We used a higher-order approximation of the dual problem instead of a higher-order interpolation of a lower order approximation of the dual solution. A numerical comparison of both approaches was done for different values of the diffusion coefficient. In numerical experiments we could prove that spurious oscillations that typically arise in numerical approximations of convection-dominated problems could be reduced significantly. Robust effectivity indices that are almost one were obtained for the specified test target quantities. We demonstrated the efficiency of the approach also for three space dimensions. The extension to nonstationary problems is our ongoing work. In the nonstationary case higher-order time discretizations can computed efficiently and computationally cheap by new postprocessing techniques, cf. [9, 20].

Acknowledgement The authors wish to thank the anonymous reviewers for their help to improve the presentation of this paper.

References

1. Ahmed, N., John, V.: Adaptive time step control for higher order variational time discretizations applied to convection-diffusion equations. Comput. Methods Appl. Mech. Eng. **285**, 83–101 (2015)
2. Angermann, L.: Balanced a posteriori error estimates for finite-volume type discretizations of convection-dominated elliptic problems. Computing **55**(4), 305–323 (1995)
3. Arndt, D., Bangerth, W., Davydov, D., Heister, T., Heltai, L., Kronbichler, M., Maier, M., Pelteret, J.-P., Turcksin, B., Wells, D.: The `deal.II` library, Version 8.5. J. Numer. Math. **25**(3), 137–146 (2017). https://doi.org/10.1515/jnma-2016-1045
4. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Birkhäuser, Basel (2003)

5. Bangerth, W., Hartmann, R., Kanschat, G.: Deal.II-A general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24/1–24/27 (2007). https://doi.org/10.1145/1268776.1268779
6. Barrenechea, G.R., John, V., Knobloch, P.: Some analytical results for an algebraic flux correction scheme for a steady convection-diffusion equation in one dimension. IMA J. Numer. Anal. **35**(4), 1729–1756 (2015)
7. Barrenechea, G.R., John, V., Knobloch, P.: Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal. **54**(4), 2427–2451 (2016)
8. Bause, M., Schwegler, K.: Analysis of stabilized higher order finite element approximation of nonstationary and nonlinear convection-diffusion-reaction equations. Comput. Methods Appl. Mech. Eng. **209–212**, 184–196 (2012)
9. Bause, M., Köcher, U., Radu, F.A., Schieweck, F.: Post-processed Galerkin approximation of improved order for wave equations. Math. Comput., submitted (2018). arXiv:1803.03005
10. Becker, R.: An optimal-control approach to a posteriori error estimation for finite element discretizations of the Navier–Stokes equations. East-West J. Numer. Math. **8**, 257–274 (2000)
11. Becker, R., Rannacher, R.: Weighted a posteriori error control in FE methods. In: Bock, H.G., et al. (eds.) ENUMATH 97. Proceedings of the 2nd European Conference on Numerical Mathematics and Advanced Applications, pp. 621–637. World Scientific, Singapore (1998)
12. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. In: Iserles, A. (ed.) Acta Numerica, vol. 10, pp. 1–102. Cambridge University, Cambridge (2001)
13. Besier, M., Rannacher, R.: Goal-oriented space-time adaptivity in the finite element galerkin method for the computation of nonstationary incompressible flow. Int. J. Num. Methods Fluids **70**(9), 1139–1166 (2012)
14. Bittl, M., Kuzmin, D.: An hp-adaptive flux-corrected transport algorithm for continuous finite elements. Computing **95**, 27–48 (2013)
15. Braack, M., Ern, A.: A posteriori control of modeling errrors and discretization errors. Multiscale Model. Simul. **1**(2), 221–238 (2003)
16. Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Eng. **32**(1–3), 199–259 (1982)
17. Carey, G.F., Oden, J.T.: Finite Elements, Computational Aspects, Vol. III (The Texas Finite Element Series). Prentice-Hall, Englewood Cliffs (1984)
18. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
19. Endtmayer, B., Wick, T.: A partition-of-unity dual-weighted residual approach for multi-objective goal functional error estimation applied to elliptic problems. Comput. Methods Appl. Math. **17**(4), (2017). https://doi.org/10.1515/cmam-2017-0001
20. Ern, A., Schieweck, F.: Discontinuous Galerkin method in time combined with a stabilized finite element method in space for linear first-order pdes. Math. Comput. **85**, 2099–2129 (2016)
21. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (2010)
22. Houston, P., Rannacher, R., Süli, E.: A posteriori error analysis for stabilised finite element approximations of transport problems. Comput. Methods Appl. Mech. Eng. **190**, 1483–1508 (2000)
23. Hughes, T.J.R., Brooks, A.N.: A multidimensional upwind scheme with no crosswind diffusion. In: Hughes, T.J.R. (ed.) Finite Element Methods for Convection Dominated Flows, AMD, vol. 34, pp. 19–35. American Society of Mechanical Engineers (ASME), New York (1979)
24. Hughes, T.J.R., Mallet, M., Mizukami, A.: A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. Comput. Methods Appl. Mech. Eng. **54**, 341–355 (1986)
25. John, V., Knobloch, P.: Adaptive computation of parameters in stabilized methods for convection-diffusion problems. In: Cangiani, A., et al. (eds.) Numerical Mathematics and Advanced Applications 2011. Springer, Heidelberg (2013)

26. John, V., Knobloch, P., Savescu, S.B.: A posteriori optimization of parameters in stabilized methods for convection-diffusion problems-Part I. Comput. Methods Appl. Mech. Eng. **200**, 2916–2929 (2011)
27. John, V., Novo, J.: A robust SUPG norm a posteriori error estimator for stationary convection-diffusion equations. Comput. Methods Appl. Mech. Eng. **255**, 289–305 (2013)
28. John, V., Schmeyer, E.: Finite element methods for time-dependent convection-diffusion-reaction equations with small diffusion. Comput. Methods Appl. Mech. Eng. **198**, 173–181 (2009)
29. John, V., Schmeyer, E.: On finite element methods for 3D time-dependent convection-diffusion-reaction equations with small diffusion In: Hegarty A., et al. (eds.) BAIL 2008 - Boundary and Interior Layers. Lecture Notes in Computational Science and Engineering, vol. 69. Springer, Heidelberg (2009)
30. Köcher, U.: Variational Space-Time Methods for the Elastic Wave Equation and the Diffusion Equation, Dissertation. Helmut Schmidt University, Hamburg, urn:nbn:de:gbv:705-opus-31129 (2015)
31. Köcher, U., Bruchhäuser, M.P., Bause, M.: Efficient and scalable data structures and algorithms for goal-oriented adaptivity of space-time FEM codes. SoftwareX 100239 (2019). https://doi.org/10.1016/j.softx.2019.100239
32. Kuzmin, D., Korotov, S.: Goal-oriented a posteriori error estimates for transport problems. Math. Comput. Simul. **80**(8), 1674–1683 (2010)
33. Kuzmin, D., Turek, S.: Flux correction tools for finite elements. J. Comput. Phys. **175**, 525–558 (2002)
34. Löhner, R., Morgan, K., Peraire, P., Vahdati, M.: Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. Int. J. Numer. Methods Fluids **7**(10), 1093–1109 (1987)
35. Lube, G., Rapin, G.: Residual-based stabilized higher-order FEM for advection-dominated problems. Comput. Methods Appl. Mech. Eng. **195**, 4124–4138 (2006)
36. Rannacher, R.: A posteriori error estimation in least-squares stabilized finite element schemes. Comput. Methods Appl. Mech. Eng. **166**, 99–114 (1998)
37. Richter, T., Wick, T.: Variational localizations of the dual weighted residual estimator. J. Comput. Appl. Math. **279**, 192–208 (2015)
38. Roos, H.-G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Springer, Berlin (2008)
39. Schmich, M.: Adaptive Finite Element methods for computing nonstationary incompressible Flows, Dissertation. University of Heidelberg, Heidelberg urn:nbn:de:bsz:16-opus-102001 (2009)
40. Schwegler, K.: Adaptive goal-oriented error control for stabilized approximations of convection-dominated problems, Dissertation. Helmut Schmidt University Hamburg (2014). http://edoc.sub.uni-hamburg.de/hsu/volltexte/2014/3086/
41. Verfürth, R.: A Posteriori Error Estimation Techniques for Finite Element Methods. Oxford University, Oxford (2013)

Chapter 6
Uniform Exponential Stability of Galerkin Approximations for a Damped Wave System

Herbert Egger and Thomas Kugler

Abstract We consider the numerical approximation of a linear damped wave system modeling the propagation of pressure waves in a pipeline by Galerkin approximations in space and appropriate time-stepping schemes. By careful energy estimates, we prove exponential decay of the physical energy on the continuous level and uniform exponential stability of semi-discrete and fully discrete approximations obtained by mixed finite element discretization in space and certain one-step methods in time. The validity and limitations of the theoretical results are demonstrated by numerical tests.

6.1 Introduction

The propagation of pressure waves in water or gas pipelines, also known as the water hammer, can be described by hyperbolic systems of the form [5, 20]

$$\partial_t u + \partial_x p + au = 0 \tag{6.1}$$

$$\partial_t p + \partial_x u = 0. \tag{6.2}$$

Here p and u denote the pressure and velocity of the fluid, respectively. The damping parameter $a = a(x)$ accounts for friction at the pipe walls and we assume that a is uniformly positive throughout the chapter. We only consider a one-dimensional model problem in detail, but our results can be generalized to multi-dimensional problems of similar structure in acoustics, elasticity, or electromagnetics, and also to problems on one-dimensional networks.

H. Egger · T. Kugler (✉)
TU Darmstadt, AG Numerics and Scientific Computing, Darmstadt, Germany
e-mail: egger@mathematik.tu-darmstadt.de; kugler@mathematik.tu-darmstadt.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_6

Due to many fields of application, the analysis of damped wave phenomena have attracted significant interest in the literature. When modeling the vibration of a string, the physical energy of the system is given by

$$E^1(t) = \frac{1}{2}\big(\|\partial_t u(t)\|^2 + \|\partial_t p(t)\|^2\big).$$

Replacing $\partial_t p$ by $\partial_x u$ yields the more common form $\frac{1}{2}\big(\|\partial_t u(t)\|^2 + \|\partial_x u(t)\|^2\big)$ of the energy for the one-dimensional wave equation. If the boundary conditions are chosen such that no energy can enter or leave the domain via the boundary, then

$$\frac{d}{dt}E^1(t) = -\int a\ |\partial_t u(t)|^2 dx \leq 0.$$

This shows that kinetic energy is dissipated efficiently by the damping mechanism. If the damping is effective at least on a sub-domain of positive measure [7, 26, 29], then the total energy $E^1(t)$ can be shown to decrease exponentially, i.e.,

$$E^1(t) \leq Ce^{-\alpha t}E^1(0)$$

for some constants C and $\alpha > 0$, A similar result holds if the damping only takes place at the boundary [6, 23]. These decay estimates imply the exponential stability of the system under consideration which is of relevance, e.g. for the control and observation of damped wave systems [6, 15, 30].

A first contribution of this paper is to show that the same decay estimates hold for the problem under investigation and for all energies of the form

$$E^k(t) = \frac{1}{2}\big(\|\partial_t^k u(t)\|^2 + \|\partial_t^k p(t)\|^2\big), \qquad k \geq 0$$

under the assumption that the damping parameter $a(x)$ is bounded and uniformly positive. In particular, our results cover the decay of $E^0(t) = \frac{1}{2}(\|u(t)\|^2 + \|p(t)\|^2)$, which is the physical energy for the acoustic wave propagation problem considered here. This result is derived by carefully adopting classical arguments of [1, 2, 29] to the problem under consideration, in particular addressing the case $k = 0$ and the different choice of boundary conditions. The latter is important for the generalization from a single interval to networks [10]. The uniform positivity assumption on the damping parameter allows us to prove energy decay under minimal regularity conditions and to obtain explicit estimates for the damping rate α depending only on the bounds for the damping parameter. This is useful for the asymptotic analysis of damped wave systems [24] and for the investigation of the parabolic limit behavior.

Also the numerical approximation of damped wave phenomena has attracted significant interest in the literature. Mixed finite element methods have been proven to be particularly well-suited [3, 12, 18, 21] and error estimates for some fully discrete

schemes for the damped wave equation have been obtained in [16, 19, 22, 27]. Our results contribute to this active field of research by proposing and analyzing fully discrete approximation schemes that preserve the exponential stability uniformly with respect to the discretization parameters.

For the discretization in space, we consider a variational formulation of the first order system (6.1)–(6.2) in the spirit of [21] and its Galerkin approximation by a mixed finite element method. A compatibility condition for the approximation spaces for velocity and pressure allows us to establish the uniform exponential decay of the energies $E^k(t)$, $k \geq 0$ also on the semi-discrete level. Extending the results of [12], our analysis also covers higher order approximations and different boundary conditions. For the time discretization we consider certain one-step methods and establish the unconditional and uniform exponential stability for the fully discrete schemes. As a by-product of our stability analysis, one can also obtain error estimates that hold uniformly in time and with respect to the spatial and temporal mesh size.

The rest of the paper is organized as follows: In Sect. 6.2, we formally introduce the problems under investigation and recall some basic results about their well-posedness. In Sect. 6.3, we derive the energy estimates and prove the exponential decay to equilibrium under minimal regularity assumptions. In Sect. 6.4, we provide a characterization of classical solutions via variational principles, which are the starting point for the numerical approximation. Section 6.5 is concerned with a general class of Galerkin discretizations in space, for which we provide uniform stability and error estimates. The discretization by mixed finite elements is discussed as a particular example in Sect. 6.6. In Sect. 6.7, we then investigate the time discretization by a family of one-step methods and we prove unconditional and uniform exponential stability and error estimates for the fully discrete schemes. For illustration of the theoretical results, some numerical tests are presented in Sect. 6.8.

6.2 Preliminaries

We consider the unit interval and denote by $L^p(0,1)$ and $H^1(0,1)$ the usual Lebesgue and Sobolev spaces. The functions in $H_0^1(0,1)$ vanish at the boundary. We denote by $\|f\| = \|f\|_{L^2(0,1)}$ and $\|f\|_1 = \|f\|_{H^1(0,1)}$ the natural norms of these spaces and by $(f,g) = \int_0^1 fg\,dx$ the scalar product of $L^2(0,1)$. By $C^k([0,T];X)$ and $H^k(0,T;X)$ we denote the spaces of functions $f : [0,T] \to X$ with values in a Banach space X having the appropriate smoothness and integrability properties; see [14] for details.

6.2.1 Problem Statement

As a model problem for our consideration, we consider the linear hyperbolic system

$$\partial_t u + \partial_x p + au = 0 \qquad \text{in } (0,1)\times(0,T) \tag{6.3}$$

$$\partial_t p + \partial_x u = 0 \qquad \text{in } (0,1)\times(0,T) \tag{6.4}$$

with homogeneous Dirichlet conditions for the pressure

$$p = 0 \qquad \text{on } \{0,1\}\times(0,T). \tag{6.5}$$

More general boundary conditions or inhomogeneous right-hand side could be taken into account without difficulty. The initial values shall be given by

$$u(0) = u_0 \quad \text{and} \quad p(0) = p_0 \qquad \text{on } (0,1). \tag{6.6}$$

The well-posedness of this initial boundary value problem can be deduced with standard arguments. For later reference, let us summarize the most basic results.

Lemma 6.1 *Let $a \in L^\infty(0,1)$ and $T > 0$. Then for $u_0, p_0 \in L^2(0,1)$, problem (6.3)–(6.6) has a unique mild solution $(u,p) \in C^0([0,T]; L^2(0,1)\times L^2(0,1))$ and*

$$\max_{0\le t\le T} \|u(t)\|^2 + \|p(t)\|^2 \le C\big(\|u_0\|^2 + \|p_0\|^2\big).$$

If $u_0 \in H^1(0,1)$ and $p_0 \in H^1_0(0,1)$, then the solution (u,p) is classical, this means that $(u,p) \in C^1([0,T]; L^2(0,1)\times L^2(0,1)) \cap C^0([0,T]; H^1(0,1)\times H^1_0(0,1))$, and

$$\max_{0\le t\le T} \|\partial_t u(t)\|^2 + \|\partial_t p(t)\|^2 \le C\big(\|\partial_t u(0)\|^2 + \|\partial_t p(0)\|^2\big).$$

The constant C only depends on the bounds for a and the time horizon T.

Proof The results follow from standard results of semigroup theory [14, 25]. □

6.2.2 Stationary Problem

Problems with time inhomogeneous right-hand side and boundary values can be reduced to the homogeneous case by subtracting a solution of the stationary problem

$$\partial_x \bar p + a\bar u = \bar f \qquad \text{in } (0,1), \tag{6.7}$$

$$\partial_x \bar u = \bar g \qquad \text{in } (0,1). \tag{6.8}$$

with boundary conditions

$$\bar{p} = \bar{h} \qquad \text{on } \{0, 1\}. \tag{6.9}$$

We use a bar symbol to denote functions that are independent of time. A similar problem will arise later in the stability analysis for the time-dependent problem.

Lemma 6.2 *Let $0 < a_0 \le a(x) \le a_1$. Then for any $\bar{f}$, $\bar{g} \in L^2(0, 1)$, and $\bar{h} \in \mathbb{R}^2$, problem (6.7)–(6.9) has a unique strong solution $(\bar{u}, \bar{p}) \in H^1(0, 1) \times H_0^1(0, 1)$ and*

$$\|\bar{u}\|_1 + \|\bar{p}\|_1 \le C\big(\|\bar{f}\| + \|\bar{g}\| + |\bar{h}|\big)$$

with a constant C only depending on the bounds for the parameter a.

Proof The solution and the bounds can be computed explicitly here. □

6.3 Energy Estimates

For a detailed stability analysis of the damped wave system, which is one focus of our investigations, we will make use of the following family of generalized energies

$$E^k(t) = \frac{1}{2}\big(\|\partial_t^k u(t)\|^2 + \|\partial_t^k p(t)\|^2\big), \qquad k \ge 0. \tag{6.10}$$

As outlined in the introduction, some of these energies may have a physical interpretation, depending on the application context.

Lemma 6.3 *Let $a \in L^\infty(0, 1)$ and (u, p) be a smooth solution of (6.3)–(6.5). Then*

$$E^k(t) = E^k(s) - \int_s^t \int_0^1 a(x)\, |\partial_t^k u(x, r)|^2 dx\, dr, \qquad 0 \le s \le t \le T.$$

If in addition $a \ge 0$, then the respective energies decay monotonically.

Proof Let us first prove the estimate for $k = 0$ under the assumption that (u, p) is a classical solution. Then $E^0(t)$ is continuously differentiable, and we have

$$\begin{aligned}\frac{d}{dt}E^0(t) &= (\partial_t u(t), u(t)) + (\partial_t p(t), p(t)) \\ &= -(\partial_x u(t), p(t)) - (\partial_x p(t), u(t)) - (au(t), u(t)) = -(au(t), u(t)).\end{aligned}$$

The identity for $k = 0$ follows by integration over time. By density, the estimate also holds for mild solutions and the result for $k \ge 1$ is obtained by differentiation. □

As a next step in the stability analysis, we now show that the energies decay exponentially, provided that the damping is effective everywhere in the domain.

Theorem 6.1 *Let (u, p) be a solution of (6.3)–(6.5) with finite energy $E^k(0)$ for some $k \geq 0$. Moreover, assume that $a_0 \leq a(x) \leq a_1$ for some constants $a_0, a_1 > 0$. Then*

$$E^k(t) \leq 3e^{-\alpha(t-s)} E^k(s), \qquad 0 \leq s \leq t \leq T,$$

with decay rate $\alpha = \frac{4}{3}a_0^3/(8a_0^2 + 4a_0^2 a_1 + 2a_0 a_1 + a_1^4)$ independent of T.

Proof A detailed proof is given in the appendix but we sketch the main arguments already here. Similar as in [29], we define for $\epsilon > 0$ the modified energies

$$E_\epsilon^1(t) = E^1(t) + \epsilon(\partial_t u(t), u(t)).$$

As proven in Lemma 6.14, the two energies $E^1(t)$ and $E_\epsilon^1(t)$ are equivalent for sufficiently small ϵ, more precisely, $\frac{1}{2}E^1(t) \leq E_\epsilon^1(t) \leq \frac{3}{2}E^1(t)$. Under a further restriction on ϵ, one can then show that $\frac{d}{dt}E_\epsilon^1(t) \leq -\frac{2}{3}\epsilon E_\epsilon^1(t)$, from which the result for $k = 1$ follows; see Lemma 6.15, where the precise definition of ϵ and α is given. The case $k = 0$ can be deduced from the one for $k = 1$ by an explicit construction, see Sect. 6.9, and the estimate for $k \geq 2$ finally follows by formal differentiation. □

6.4 Variational Characterization

For the design and the analysis of appropriate discretization schemes, it will be useful to characterize the solutions of the damped wave system via variational principles which are better suited for a systematic approximation [21].

6.4.1 Weak Formulation of the Stationary Problem

Testing (6.7) and (6.8) with appropriate test functions, using integration-by-parts for the first equation and the boundary conditions (6.9), we arrive at the following weak form of the stationary problem. For ease of presentation, we set $\bar{h} = 0$ here.

Problem 6.1 Find $\bar{u} \in H^1(0, 1)$ and $\bar{p} \in L^2(0, 1)$ such that

$$\begin{aligned} -(\bar{p}, \partial_x \bar{v}) + (a\bar{u}, \bar{v}) &= (\bar{f}, \bar{v}) \qquad \text{for all } \bar{v} \in H^1(0, 1) \\ (\partial_x \bar{u}, \bar{q}) &= (\bar{g}, \bar{q}) \qquad \text{for all } \bar{q} \in L^2(0, 1). \end{aligned}$$

As before, we use the bar symbol to denote functions that are independent of time. The existence of a unique weak solution follows almost directly from Lemma 6.2.

Lemma 6.4 *Let* $0 < a_0 \le a(x) \le a_1$. *Then for any* $\bar{f}$, $\bar{g} \in L^2(0,1)$, *Problem 6.1 has a unique solution which coincides with the strong solution of* (6.7)–(6.9) *with* $\bar{h} = 0$. *Moreover, one has* $\|\bar{u}\|_1 + \|\bar{p}\|_1 \le C(\|\bar{f}\| + \|\bar{g}\|)$ *with* C *depending on* a_0 *and* a_1.

Proof Any strong solution of the stationary problem also is a weak solution. Existence therefore follows from Lemma 6.2. Uniqueness follows by testing the variational principle with $\bar{v} = \bar{u}$ and arbitrary test function $\bar{q}$. □

6.4.2 Weak Form of the Instationary Problem

With a similar derivation as for the stationary problem above, one arrives at the following weak formulation for the time-dependent damped wave system.

Problem 6.2 Find $(u, p) \in L^2(0,T; H^1(0,1) \times L^2(0,1)) \cap H^1(0,T; H^1(0,1)' \times L^2(0,1))$ with initial values $u(0) = u_0$ and $p(0) = p_0$, such that for a.e. $t \in (0,T)$ there holds

$$(\partial_t u(t), \bar{v}) - (p(t), \partial_x \bar{v}) + (au(t), \bar{v}) = 0 \qquad \text{for all } \bar{v} \in H^1(0,1) \tag{6.11}$$

$$(\partial_t p(t), \bar{q}) + (\partial_x u(t), \bar{q}) = 0 \qquad \text{for all } \bar{q} \in L^2(0,1). \tag{6.12}$$

Similar as before, the well-posedness of the weak formulation can be deduced from the previous results about existence of classical solutions.

Lemma 6.5 *For any* $u_0 \in H^1(0,1)$ *and* $p_0 \in H_0^1(0,1)$ *the above weak formulation has a unique solution* (u, p) *which coincides with the classical solution of problem* (6.3)–(6.6). *In particular, the a-priori estimates of Lemma 6.1 are valid.*

Proof Since any classical solution is a weak solution, the existence follows from Lemma 6.1. Uniqueness is a consequence of the energy estimates of Lemma 6.3. □

6.5 Galerkin Semi-Discretization

For the discretization in space, we consider Galerkin approximations for the weak formulations stated in the previous section based on approximation spaces

$$V_h \subset H^1(0,1) \qquad \text{and} \qquad Q_h \subset L^2(0,1),$$

which are assumed to be finite dimensional without further mentioning. We start with considering the stationary case and then turn to the time dependent problem.

6.5.1 Discretization of the Stationary Problem

The Galerkin approximation of the stationary problem reads as follows.

Problem 6.3 Find $\bar{u}_h \in V_h \subset H^1(0,1)$ and $\bar{p}_h \in Q_h \subset L^2(0,1)$ such that

$$\begin{aligned} -(\bar{p}_h, \partial_x \bar{v}_h) + (a\bar{u}_h, \bar{v}_h) &= (\bar{f}, \bar{v}_h) \qquad \text{for all } \bar{v}_h \in V_h \\ (\partial_x \bar{u}_h, \bar{q}_h) &= (\bar{g}, \bar{q}_h) \qquad \text{for all } \bar{q}_h \in Q_h. \end{aligned} \tag{6.13}$$

In order to establish the well-posedness of this discretized problem, some compatibility conditions for the approximation spaces are required. We have

Lemma 6.6 *Let $0 < a_0 \le a(x) \le a_1$ and assume that $Q_h = \partial_x V_h$ and $1 \in V_h$. Then for any $\bar{f}, \bar{g} \in L^2(0,1)$, Problem 6.3 has a unique solution which can be bounded by $(\bar{u}_h, \bar{p}_h)$ and $\|\bar{u}_h\|_1 + \|\bar{p}_h\| \le C\big(\|\bar{f}\| + \|\bar{g}\|\big)$ with C depending only on a_0 and a_1.*

Proof Define $b(\bar{v}_h, \bar{q}_h) = (\partial_x \bar{v}_h, \bar{q}_h)$. Then, by choosing $\bar{v}_h(x) = \int_0^x q_h(s)ds$, we get

$$\sup_{\bar{v}_h \in V_h} \frac{b(\bar{v}_h, \bar{q}_h)}{\|\bar{v}_h\|_1} \ge \frac{(\partial_x \int_0^x \bar{q}_h, \bar{q}_h)}{\|\int_0^{\cdot} \bar{q}_h\|_1} \ge \frac{1}{\sqrt{2}} \|\bar{q}_h\| \qquad \text{for all } \bar{q}_h \in Q_h.$$

Next, observe that $N_h = \{\bar{v}_h \in V_h : b(\bar{v}_h, \bar{q}_h) = 0\} = \{\bar{v}_h \in V_h : \partial_x \bar{v}_h = 0\}$. Hence

$$a(\bar{v}_h, \bar{v}_h) = (a\bar{v}_h, \bar{v}_h) \ge a_0 \|\bar{v}_h\|^2 = a_0 \|\bar{v}_h\|_1^2 \qquad \text{for all } \bar{v}_h \in N_h.$$

The assertions then follow from Brezzi's splitting lemma [4]. □

6.5.2 Galerkin Approximation of the Instationary Problem

Let $V_h \subset H^1(0,1)$ and $Q_h \subset L^2(0,1)$ be finite dimensional subspaces and denote by $\pi_h : L^2(0,1) \to V_h$ and $\rho_h : L^2(0,1) \to Q_h$ the respective L^2-orthogonal projections. The semi-discretization for the time dependent problem then reads as follows.

Problem 6.4 Find $(u_h, p_h) \in H^1(0,T; V_h \times Q_h)$ with $u_h(0) = \pi_h u_0$ and $p_h(0) = \rho_h p_0$, such that for a.e. $t \in (0,T)$ there holds

$$\begin{aligned} (\partial_t u_h(t), \bar{v}_h) - (p_h(t), \partial_x \bar{v}_h) + (a u_h(t), \bar{v}_h) &= 0 \qquad \text{for all } \bar{v}_h \in V_h \\ (\partial_t p_h(t), \bar{q}_h) + (\partial_x u_h(t), \bar{q}_h) &= 0 \qquad \text{for all } \bar{q}_h \in Q_h. \end{aligned}$$

By choosing some bases for the spaces V_h and Q_h, this system can be transformed into a linear ordinary differential equation, which yields the following result.

Lemma 6.7 *Let $a \in L^\infty(0,1)$. Then for any $u_0, p_0 \in L^2(0,1)$, Problem 6.4 has a unique solution (u_h, p_h) and $\|u_h(t)\|^2 + \|p_h(t)\|^2 \le C\big(\|u_0\|^2 + \|p_0\|^2\big)$ for all $0 \le t \le T$ with constant C only depending on the bounds for a and the time horizon T.*

Proof Existence and uniqueness follow from the Picard-Lindelöf theorem, and the a-priori estimate follows from the energy identities given in the next section. □

6.5.3 Discrete Energy Estimates

We now present the stability analysis of the Galerkin approximations. Let (u^h, p^h) denote a solution of Problem 6.4. Proceeding in a similar manner as on the continuous level, we define the semi-discrete generalized energies

$$E_h^k(t) = \frac{1}{2}\big(\|\partial_t^k u_h(t)\|^2 + \|\partial_t^k p_h(t)\|^2\big), \qquad k \ge 0.$$

The following energy identities then follow almost directly from the special form of the variational principle underlying and the Galerkin approximation.

Lemma 6.8 *Let $a \in L^\infty(0,1)$ and (u_h, p_h) be a solution of Problem 6.4. Then*

$$E_h^k(t) = E_h^k(s) - \int_s^t \int_0^1 a(x)|\partial_t^k u_h(x,r)|^2 dx\, dr$$

If $a \ge 0$, then the semi-discrete energies are monotonically decreasing.

Proof Since the right-hand side is zero, the discrete solution (u_h, p_h) is always infinitely differentiable with respect to time. For $k = 0$ we then obtain

$$\frac{d}{dt}E_h^0(t) = (\partial_t u_h(t), u_h(t)) + (\partial_t p_h(t), p_h(t)) = -(a u_h(t), u_h(t)),$$

which follows by testing the variational principle with $\bar{v}_h = u_h(t)$ and $\bar{q}_h = p_h(t)$. The result for $k = 0$ then follows by integration over time. The case $k \ge 1$ can be deduced by applying the result for $k = 0$ to the derivatives $(\partial_t^k u_h, \partial_t^k p_h)$. □

Under a mild compatibility condition for the approximation spaces V_h and Q_h, we can also prove exponential decay estimates for the discrete energies.

Theorem 6.2 *Let $0 < a_0 \le a(x) \le a_1$ and assume that $Q_h = \partial_x V_h$ and $1 \in V_h$. Then any solution (u_h, p_h) of Problem 6.4 satisfies*

$$E_h^k(t) \le 3e^{-\alpha(t-s)} E_h^k(s)$$

with decay rate $\alpha > 0$ that can be chosen as on the continuous level.

Proof The proof follows along the same lines as that of Theorem 6.1 and is given in the appendix. The conditions $1 \in V_h$ and $Q_h = \partial_x V_h$ are required for the discrete analogue of Lemma 6.13, which is then used in Lemmas 6.14 and 6.15. The discrete exponential stability thus strongly relies on these compatibility conditions. □

6.6 A Mixed Finite Element Method

Let T_h be a partition of $(0, 1)$ into subintervals of length h. We denote by $P_k(T_h)$ the space of piecewise polynomials of order k. One can now easily define pairs of compatible spaces of arbitrary order with good approximation and stability properties.

Lemma 6.9 *Let $V_h = P_{k+1}(T_h) \cap H^1(0, 1)$ and $Q_h = P_k(T_h)$ with $k \geq 0$. Then*

(i) $Q_h = \partial_x V_h$ *and* $1 \in V_h$,
(ii) $\|\partial_x \pi_h u\| \leq c_1 \|\partial_x u\|$ *for all* $u \in H^1(0, 1)$.

The constants c_1 in the above estimate only depends on the polynomial degree k.

Proof The first assertion follows directly from the construction, and the stability estimate (ii) is well known; see e.g. [28]. □

The mixed finite element method thus yields a uniformly exponentially stable approximation. Standard arguments [8, 17] allow to deduce the usual convergence rate estimates. For the lowest order approximation using P_1 and P_0 elements, one can show by a refined analysis that

$$\|u(t) - u_h(t)\| + \|\rho_h p(t) - p_h(t)\| = O(h^2),$$

i.e., the method actually converges with second order; see [11] for details. Due to the exponential stability, the error estimates are uniform in time.

6.7 Time Discretization

Let $\tau > 0$ be a given time-step and set $t^n = n\tau$ for $n \geq 0$. For ease of notation, we define for any $\theta \in \mathbb{R}$ and any given sequence $\{u_h^n\}_{n \geq 0}$ the symbols

$$u_h^{n,\theta} := \theta u_h^n + (1 - \theta) u_h^{n-1}, \qquad \text{as well as}$$

$$d_\tau^0 u_h^n := u_h^n \qquad \text{and} \qquad d_\tau^{k+1} u_h^n := \frac{d_\tau^k u_h^n - d_\tau^k u_h^{n-1}}{\tau} \quad \text{for } k \geq 0.$$

Note that $d_\tau^k u_h^n$ corresponds to the kth backward difference quotient. To mimic the notation on the continuous level, we will also write $d_{\tau\tau} u_h^n$ instead of $d_\tau^2 u_h^n$.

6.7.1 Fully Discrete Scheme

As before, we assume that $V_h \subset H^1(0,1)$ and $Q_h \subset L^2(0,1)$ are some finite dimensional subspaces. For the time discretization of Problem 6.4, we now consider the following family of fully discrete approximations.

Problem 6.5 Set $u_h^0 = \pi_h u_0$, $p_h^0 = \rho_h p_0$, and for $n \geq 1$, find $(u_h^n, p_h^n) \in V_h \times Q_h$ with

$$(d_\tau u_h^n, \bar{v}_h) - (p_h^{n,\theta}, \partial_x \bar{v}_h) + (a u_h^{n,\theta}, \bar{v}_h) = 0 \qquad \text{for all } \bar{v}_h \in V_h \tag{6.14}$$

$$(d_\tau p_h^n, \bar{q}_h) + (\partial_x u_h^{n,\theta}, \bar{q}_h) = 0 \qquad \text{for all } \bar{q}_h \in Q_h. \tag{6.15}$$

Observe that the system (6.14)–(6.15) can be written equivalently as

$$\tfrac{1}{\tau}(u_h^n, \bar{v}_h) - \theta[(p_h^n, \partial_x \bar{v}_h) + (a u_h^n, \bar{v}_h)] = \tfrac{1}{\tau}(u_h^{n-1}, \bar{v}_h) + (1-\theta)[(p_h^{n-1}, \partial_x \bar{v}_h) - (a u_h^{n-1}, \bar{v}_h)]$$

$$\tfrac{1}{\tau}(p_h^n, \bar{q}_h) + \theta(\partial_x u_h^n, \bar{q}_h) = \tfrac{1}{\tau}(p_h^{n-1}, \bar{q}_h) - (1-\theta)(\partial_x u_h^{n-1}, \bar{q}_h).$$

The well-posedness of the problem of determining (u_h^n, p_h^n) from (u_h^{n-1}, p_h^{n-1}) can then be shown with the same arguments as used in Lemma 6.6, and we obtain

Lemma 6.10 *Let $0 \leq a(x) \leq a_1$ and assume that $Q_h = \partial_x V_h$ and $1 \in V_h$. Then for any u_0, $p_0 \in L^2(0,1)$ and $0 \leq \theta \leq 1$, Problem 6.5 admits a unique solution $\{(u_h^n, p_h^n)\}_{n \geq 0}$. Moreover, $\|u_h^n\| + \|p_h^n\| \leq C(\|u_0\| + \|p_0\|)$ with C independent of n.*

Proof Existence of a unique solution (u_h^n, p_h^n) for (6.14)–(6.15) for any $n \geq 0$ follows from testing with $\bar{v}_h = \bar{u}_h$ and $\bar{q}_h = \bar{p}_h$. The uniform bounds for the solution can again be obtained via energy arguments; see below. □

6.7.2 Discrete Energy Estimates

For the stability analysis of the fully discrete problem, we utilize energy estimates similar as on the continuous and the semi-discrete level. Given a solution $\{(u_h^n, p_h^n)\}$ of Problem 6.5, we define the discrete energies at time t^n by

$$E_h^{k,n} = \frac{1}{2}\big(\|d_\tau^k u_h^n\|^2 + \|d_\tau^k p_h^n\|^2\big), \qquad k \geq 0.$$

By appropriate testing of the fully discrete scheme (6.14)–(6.15) and with similar arguments as on the continuous level, we now obtain the following energy identities.

Lemma 6.11 *Let* $\{(u_h^n, p_h^n)\}$ *be a solution of Problem 6.5. Then*

$$d_\tau E_h^{k,n} = -(\theta - \tfrac{1}{2})\tau\big(\|d_\tau^{k+1} u_h^n\|^2 + \|d_\tau^{k+1} p_h^n\|^2\big) - (a d_\tau^k u_h^{n,\theta}, d_\tau^k u_h^{n,\theta}). \tag{6.16}$$

Proof The result for $k = 0$ follows by setting $\bar{v}_h = u_h^{n,\theta}$ and $\bar{q}_h = p_h^{n,\theta}$ in (6.14)–(6.15). The estimate for $k \geq 1$ reduces to the one for $k = 0$ by observing that, due to linearity of the problem, the differences $(d_\tau^k u_h^n, d_\tau^k p_h^n)$ solve the system (6.14)–(6.15) as well. □

Observe that, without further arguments, a decay of the discrete energy can only be guaranteed, if we require $\theta \geq 1/2$. With similar arguments as already used for the analysis of the time-continuous problem, we can also establish the exponential decay of the energies for the fully discrete setting.

Theorem 6.3 *Let* $0 < a_0 \leq a(x) \leq a_1$ *and assume that* $Q_h = \partial_x V_h$ *and* $1 \in V_h$. *Moreover, let* $\frac{1}{2} < \theta \leq 1$. *Then for any* $0 < \tau \leq \tau_0$ *sufficiently small, there holds*

$$E_h^{k,n} \leq 3e^{-\alpha(n-m)\tau} E_h^{k,m} \qquad \text{for all } k \leq m \leq n$$

with decay rate $\alpha = \frac{2}{3}a_0^3/(8a_0^2 + 4a_0^2 a_1 + 3a_0 a_1 + 4a_1^4)$ *independent of* h *and* τ.

Proof The proof follows with similar arguments as used for the time-continuous case; details are given again in the appendix. An upper bound for the maximal stepsize τ_0 depending only on a_0, a_1, and θ, is given in Lemma 6.18. □

Remark 6.1 A careful inspection of the proof of Theorem 6.3 reveals that the assertion of Theorem 6.3 remains valid if one chooses $\theta = \frac{1}{2} + \lambda\tau$ with λ sufficiently large. As we will illustrate by numerical tests, the uniform and unconditional exponential stability however gets lost for $\theta = 1/2$, i.e., for the Crank-Nicolson scheme. For the lowest order approximation and the choice $\theta = \frac{1}{2} + \lambda\tau$, one can then show that

$$\|u(t^n) - u_h^n\| + \|\rho_h p(t^n) - p_h^n\| = O(h^2 + \tau^2),$$

i.e., the proposed method yields an unconditionally and uniformly exponentially stable second order approximation for the problem under consideration [9].

6.8 Numerical Validation

We now illustrate our theoretical results with some numerical test. To allow for analytic solutions and to guarantee sufficient smoothness, we choose $a \equiv const > 0$.

Table 6.1 Decay of the exact energy $E^0(t^n)$ and the corresponding energies of the semi-discrete and fully discrete approximations obtained with the mixed finite element approximation combined with the implicit Euler method ($\theta = 1$) and the second order scheme ($\theta = \frac{1}{2} + \tau$), respectively

t^n	0	2	4	6	8	10	α
Exact	2.25	2.65e-02	3.13e-04	3.69e-06	4.35e-08	5.12e-10	2.139
$\theta = 1$	2.25	2.66e-02	3.14e-04	3.71e-06	4.39e-08	5.18e-10	2.138
$\theta = \frac{1}{2} + \tau$	2.25	2.65e-02	3.13e-04	3.69e-06	4.34e-08	5.12e-10	2.139

6.8.1 Exponential Convergence

Let us start by comparing the decay behavior of the continuous and the discrete solutions. Using separation of variables, one can see that

$$u(x,t) = e^{-at/2+\sqrt{a^2/4-\pi^2}t}\cos(\pi x) \tag{6.17}$$

and

$$p(x,t) = 1/\pi(-a/2 - \sqrt{a^2/4 - \pi^2})e^{-at/2+\sqrt{a^2/4-\pi^2}t}\sin(\pi x) \tag{6.18}$$

solve the damped wave system (6.3)–(6.6) with $a \equiv const$. For our numerical tests, we choose $a = 10$ and compute the discrete solutions with the mixed finite element approximation with P_1–P_0 elements for the velocity and pressure, and using the θ-scheme with $\theta = 1$ (implicit Euler) and $\theta = \frac{1}{2} + \tau$ (second order). In Table 6.1, we report about the energy decay for the exact, the semi-discrete, and the fully discrete solutions obtained with discretization parameters $h = \tau = 10^{-3}$. As predicted by our theoretical results, the energies decrease exponentially and approximately at the same rates independent of the precise choice of the parameter θ.

6.8.2 Non-Uniform Stability of the Crank-Nicolson Method

As mentioned in Remark 6.1, the unconditional and uniform exponential stability for the fully discrete scheme can be guaranteed for the choice $\theta = \frac{1}{2} + \lambda\tau$ with λ sufficiently large, which yields a second order approximation in time. For $\lambda = 0$, i.e., for $\theta = 1/2$, one obtains the Crank-Nicolson method which is also second order accurate in time but not uniformly exponentially stable, as we demonstrate now.

For our tests, we again set $a = 10$ and as initial values, we choose $u_0 = 0$ and p_0 as the hat function on $[0, 1]$, which ensures that components of all spatial frequencies are present in the solution. We then compute the numerical solutions with the mixed finite element approximation and the θ-scheme with $\theta = \frac{1}{2}$ as well as $\theta = \frac{1}{2} + \tau$. We use a fixed time step $\tau = 10^{-2}$ and different mesh sizes $h = 2^{-k}$ for some values of $k \geq 1$. The evolution of the discrete energies E_h^n is depicted

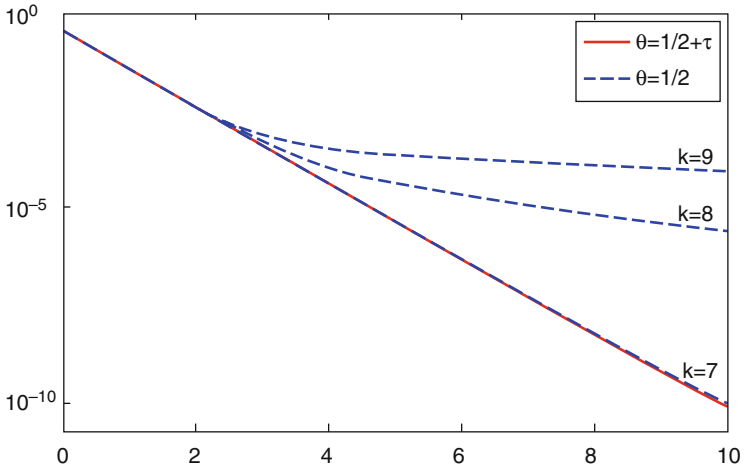


Fig. 6.1 Discrete energies E_h^n for the fully discrete solutions with $\theta = \frac{1}{2}$ (Crank-Nicolson, blue) and $\theta = \frac{1}{2} + \tau$ (second order, red) for fixed time step $\tau = 10^{-2}$ and mesh size $h = 2^{-k}$ with $k = 7, 8, 9$

in Fig. 6.1. As can be seen from the plots, the exponential stability of the Crank-Nicolson scheme with $\theta = \frac{1}{2}$ is lost when h becomes much smaller than τ while the decay of the second order scheme with $\theta = \frac{1}{2} + \tau$ remains uniform. Let us note that the uniform exponential stability could be maintained also for the Crank-Nicolson scheme under a condition $\tau \leq ch$ on the time step, see [13] for results in this direction, and even for explicit Runge-Kutta methods with sufficiently small time steps.

6.8.3 Asymptotic Behavior of the Decay Rate

Our theoretical results allow us to make some predictions about the dependence of the decay rate α on the upper and lower bounds a_0, a_1 for the parameter a. For $a \equiv const$, an analytic expression

$$\alpha = g(a) = a/2 - \mathrm{Re}\sqrt{a^2/4 - \pi^2}$$

for the decay rate of the continuous system can be derived; see for instance [7]. We now illustrate that the correct behavior of the decay rate is reproduced by the semi-discretization and the fully discrete schemes proposed in the previous sections. To do so, we compute the norms of $S_h(t^n) : (u_h(0), p_h(0)) \mapsto (u_h(t^n), p_h(t^n))$ and $S_h^\tau(t^n) : (u_h(0), p_h(0)) \mapsto (u_h^n, p_h^n)$ governing the semi-discrete and discrete evolutions, respectively. In our tests, we set $t^n = 10$ and compute $\|S_h(t^n)\|$ for

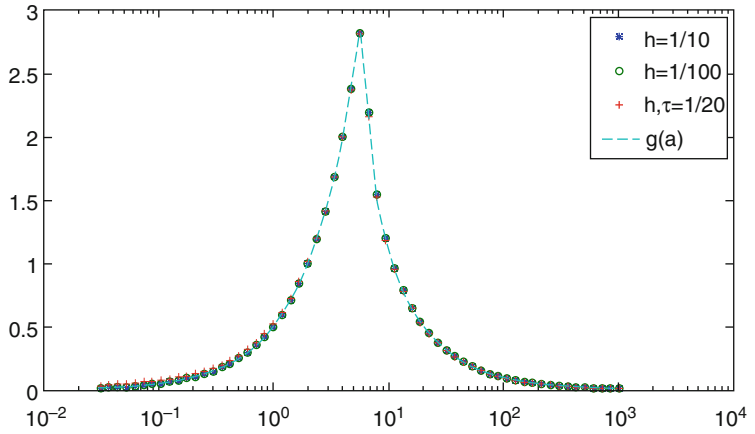


Fig. 6.2 Decay rates for semi-discrete and fully discrete evolution operators $\|S_h(10)\|$ with $h = 1/10$, $h = 1/100$ and $S_h^\tau(10)$ with $h = \tau = 1/20$ for damping parameters $a = 2^{-k}$ with $k = -5, \ldots, 10$

$h = 10^{-1}, 10^{-2}$ and $\|S_h^\tau(t^n)\|$ with $h = \tau = 1/20$ using the second order scheme with $\theta = \frac{1}{2} + \tau$. The damping parameter is chosen from the set $a = 2^{-5}, \ldots, 2^{10}$. The results of our numerical tests are depicted in Fig. 6.2. Already for the coarsest discretization, the numerically observed decay rates are in perfect agreement with the analytical formula over a very large range of parameters a. This illustrates the robustness of our results with respect to the discretization.

6.9 Discussion

In this paper, we considered the systematic numerical approximation of a damped wave system by Galerkin semi-discretization in space and time discretization by certain one-step methods. We derived energy decay estimates on the continuous level and showed that these remain valid uniformly for the semi-discretizations and fully discrete approximations under general assumptions on the approximation spaces and the parameter θ used for the time discretization. Moreover, the estimates are unconditional, i.e., the time step τ can be chosen independently of the discretization spaces. While we only considered here a one-dimensional model problem, our results and methods of proof can in principle also be generalized to multi-dimensional problems and other applications having similar structure. Also non-linearities can be tackled to some point; we refer to [13, 21] for some general analysis in this direction.

Acknowledgements The authors would like to gratefully acknowledge the support by the German Research Foundation (DFG) via grants IRTG 1529, GSC 233, and TRR 154.

Appendix

Auxiliary Results

We start with proving a generalized Poincaré inequality.

Lemma 6.12 *Let $a \in L^2(0,1)$ and $\bar a = \int a dx \neq 0$. Then for any $u \in H^1(0,1)$ we have*

$$\|u\|_{L^2(0,1)} \leq \tfrac{1}{\pi}\big(1 + \tfrac{1}{\bar a}\|a - \bar a\|_{L^2(0,1)}\big)\, \|\partial_x u\|_{L^2(0,1)} + \tfrac{1}{\bar a}\big|\int_0^1 au\, dx\big|,$$

Proof Let $\bar u = \int_0^1 u$ be the average of u and $\|\cdot\|$ the norm of $L^2(0,1)$. Then

$$\|u\| \leq \|u - \bar u\| + \|\bar u\| \leq \tfrac{1}{\pi}\|\partial_x u\| + \|\bar u\|,$$

where we used the standard Poincaré inequality. To bound the last term, observe that

$$\bar u = \tfrac{1}{\bar a}\int_0^1 \bar a u\, dx = \tfrac{1}{\bar a}\int_0^1 (\bar a - a)\, u + au\, dx = \tfrac{1}{\bar a}\int_0^1 (\bar a - a)\, (u - \bar u) + au\, dx.$$

Application of the triangle, Cauchy Schwarz, and Poincaré inequalities yields

$$\|\bar u\| \leq \tfrac{1}{\bar a}\|\bar a - a\|\|u - \bar u\| + \tfrac{1}{\bar a}\big|\int_0^1 au\, dx\big| \leq \tfrac{\|\bar a - a\|}{\bar a \pi}\|\partial_x u\| + \tfrac{1}{\bar a}\big|\int_0^1 au\, dx\big|.$$

The assertion of the lemma now follows by combination of the two estimates. □

An application of this lemma to solutions of the damped wave system yields the following estimate which will be used several times below.

Lemma 6.13 *Let (u, p) be smooth solution of (6.3)–(6.5) and $0 < a_0 \leq a(x) \leq a_1$. Then*

$$\|u(t)\|_{L^2(0,1)} \leq \tfrac{a_1}{a_0}\|\partial_t p(t)\|_{L^2(0,1)} + \tfrac{1}{a_0}\big|\|\partial_t u(t)\|.$$

Proof Using the bounds for the parameter, we obtain from the previous lemma that

$$\|u\|_{L^2(0,1)} \leq \tfrac{a_1}{a_0}\|\partial_x u\|_{L^2(0,1)} + \tfrac{1}{a_0}\big|\int_0^1 au\, dx\big|.$$

Note that this estimate holds for any function $u \in H^1(0,1)$. Using the mixed variational characterization of the solution, we further obtain

$$\big|\int_0^1 au\, dx\big| = |(au, 1)| = |-(\partial_t u, 1) + (p, \partial_x 1)| = |(\partial_t u, 1)| \leq \|\partial_t u\|.$$

Note that the boundary condition on the pressure was used implicitly here. □

Proof of the Theorem 6.1 for k = 1

To establish the decay estimate for the energy $E^1(t) = \frac{1}{2}\big(\|\partial_t u(t)\|^2 + \|\partial_t p(t)\|^2\big)$, let us define the modified energy

$$E^1_\epsilon(t) = E^1(t) + \epsilon(\partial_t u(t), u(t)).$$

We assume that (u, p) is a classical solution of (6.3)–(6.5), such that the energies are finite. As a first step, we will show now that for appropriate choice of ϵ, the two energies E^1 and E^1_ϵ are equivalent.

Lemma 6.14 *Let* $|\epsilon| \le \frac{a_0}{4+2a_1}$. *Then*

$$\tfrac{1}{2}E^1(t) \le E^1_\epsilon(t) \le \tfrac{3}{2}E^1(t).$$

Proof We only have to estimate the additional term in the modified energy. By the Cauchy-Schwarz inequality and the estimate of Lemma 6.13, we get

$$(\partial_t u(t), u) \le \|\partial_t u(t)\|\|u(t)\| \le \tfrac{1}{a_0}\|\partial_t u(t)\|^2 + \tfrac{a_1}{a_0}\|\partial_t u(t)\|\|\partial_t p(t)\|.$$

Using Young's inequality to bound the last term yields

$$|(\partial_t u(t), u(t))| \le \tfrac{2+a_1}{2a_0}\|\partial_t u(t)\|^2 + \tfrac{a_1}{2a_0}\|\partial_t p(t)\|^2 \le \tfrac{2+a_1}{2a_0}\big(\|\partial_t u(t)\|^2 + \|\partial_t p(t)\|^2\big).$$

The bound on ϵ and the definition of $E^1(t)$ further yields $|\epsilon(\partial_t u(t), u(t))| \le \frac{1}{2}E^1(t)$, from which the assertion of the lemma follows via the triangle inequality. □

We can now establish the exponential decay for the modified energy.

Lemma 6.15 *Let* $0 \le \epsilon \le \frac{2a_0^3}{8a_0^2+4a_0^2a_1+2a_0a_1+a_1^4}$. *Then*

$$E^1_\epsilon(t) \le e^{-2\epsilon(t-s)/3}E^1_\epsilon(s).$$

Proof To avoid technicalities, let us assume that the solution is sufficiently smooth first, such that all manipulations are well-defined. By the definition of the modified energy and the energy identity given in Lemma 6.3, we have

$$\begin{aligned}\tfrac{d}{dt}E^1_\epsilon(t) &= \tfrac{d}{dt}E^1(t) + \epsilon\tfrac{d}{dt}(\partial_t u(t), u(t)) \\ &\le -a_0\|\partial_t u(t)\|^2 + \epsilon\tfrac{d}{dt}(\partial_t u(t), u(t)).\end{aligned}$$

The last term can be expanded as

$$\epsilon\tfrac{d}{dt}(\partial_t u(t), u(t)) = \epsilon\|\partial_t u(t)\|^2 + \epsilon(\partial_{tt} u(t), u(t)). \tag{6.19}$$

Using the fact that (u, p) as well as $(\partial_t u, \partial_t p)$ solve the variational principle, we can estimate the last term by

$$\begin{aligned}(\partial_{tt} u(t), u(t)) &= (\partial_t p(t), \partial_x u(t)) - (a\partial_t u(t), u(t)) \\ &= -(\partial_t p(t), \partial_t p(t)) - (a\partial_t u(t), u(t)) \\ &\le -\|\partial_t p(t)\|^2 + a_1 \|\partial_t u(t)\| \|u(t)\|.\end{aligned}$$

Using Lemma 6.13 to bound $\|u(t)\|$ and Young's inequality, we further get

$$\begin{aligned}(\partial_{tt} u(t), u(t)) &\le -\|\partial_t p(t)\|^2 + \tfrac{a_1}{a_0}\|\partial_t u(t)\|^2 + \tfrac{a_1^2}{a_0}\|\partial_t u(t)\| \|\partial_t p(t)\| \\ &\le -\tfrac{1}{2}\|\partial_t p(t)\|^2 + \big(\tfrac{a_1}{a_0} + \tfrac{a_1^4}{2a_0^2}\big)\|\partial_t u(t)\|^2.\end{aligned}$$

Inserting this estimate in (6.19) then yields

$$\frac{d}{dt} E_\epsilon^1(t) \le -\big(a_0 - \epsilon(1 + \tfrac{a_1}{a_0} + \tfrac{a_1^4}{2a_0^2})\big)\|\partial_t u(t)\|^2 - \tfrac{\epsilon}{2}\|\partial_t p(t)\|^2.$$

The two factors are balanced by the choice $\epsilon = \frac{2a_0^3}{3a_0^2 + 2a_0 a_1 + a_1^4}$. In order to satisfy also the condition of the Lemma 6.14, we enlarge the denominator by $5a_0^2 + 4a_0^2 a_1$, which yields the expression for ϵ stated in the lemma. In summary, we thus obtain

$$\tfrac{d}{dt} E_\epsilon^1(t) \le -\epsilon E^1(t) \le -\tfrac{2\epsilon}{3} E_\epsilon^1(t).$$

The result for smooth solutions now follows by integration. The general case is obtained by smooth approximation and continuity; cf. the proof of Lemma 6.3. □

Combination of the previous estimates yields the assertion of Theorem 6.1 for $k = 1$.

Proof of Theorem 6.1 for $k = 0$ and $k \ge 2$

We will first show how the estimate for $k = 0$ can be deduced from that for $k = 1$. Let $u_0, p_0 \in L^2(0, 1)$ be given and consider the following stationary problem

$$\begin{aligned}\partial_x \bar{p} + a\bar{u} &= u_0, \qquad \text{in } (0, 1), \\ \partial_x \bar{u} &= p_0, \qquad \text{in } (0, 1),\end{aligned}$$

with boundary condition $\bar{p} = 0$ on $\{0, 1\}$. Using Lemma 6.2, we readily obtain

Lemma 6.16 *Let $0 < a_0 \le a(x) \le a_1$. Then there exists a unique strong solution $(\bar{u}, \bar{p}) \in H^1(0,1) \times H_0^1(0,1)$ and $\|\bar{u}\|_{H^1(0,1)} + \|\bar{p}\|_{H^1(0,1)} \le C\big(\|u_0\| + \|p_0\|\big)$.*

Let us now define

$$U(t) = \int_0^t u(s)ds - \bar{u} \qquad \text{and} \qquad P(t) = \int_0^t p(s)ds - \bar{p}.$$

Then (U, P) is classical solution of the damped wave system (6.3)–(6.5) with initial values $U(0) = -\bar{u}$ and $P(0) = -\bar{p}$. Applying Theorem 6.1 for $k = 1$ to (U, P) yields

$$\begin{aligned}&\|u(t)\|^2 + \|p(t)\|^2 = \|\partial_t U(t)\|^2 + \|\partial_t P(t)\|^2\\ &\le Ce^{-\alpha(t-s)}\big(\|\partial_t U(s)\|^2 + \|\partial_t P(s)\|^2\big) = Ce^{-\alpha(t-s)}\big(\|u(s)\|^2 + \|p(s)\|^2\big).\end{aligned}$$

This yields the assertion of Theorem 6.1 for $k = 0$. The estimates for $k \ge 2$ follow by simply applying the estimate for $k = 0$ to the derivatives $(\partial_t^k u, \partial_t^k p)$.

Proof of the Theorem 6.3

Let us start with considering the case $k = 1$. To establish the decay estimate for the energy $E_h^{1,n} = \frac{1}{2}(\|d_\tau u_h^n\|^2 + \|d_\tau p_h^n\|^2)$, let us define the modified energy

$$E_{h,\epsilon}^{1,n} = E_h^{1,n} + \epsilon(d_\tau u_h^n, u_h^{n,\theta}).$$

As before, the two energies $E_h^{1,n}$ and $E_{h,\epsilon}^{1,n}$ are equivalent for appropriate choice of ϵ.

Lemma 6.17 *Let $|\epsilon| < \frac{a_0}{4+2a_1}$. Then*

$$\frac{1}{2}E_h^{1,n} \le E_{h,\epsilon}^{1,n} \le \frac{3}{2}E_h^{1,n}.$$

The proof of this assertion follows almost verbatim as that of Lemma 6.14. With similar arguments as on the continuous level, we can then also establish the exponential decay estimate for the modified energy $E_{h,\epsilon}^{1,n}$.

Lemma 6.18 *Let $\frac{1}{2} < \theta \le 1$, $0 < \epsilon \le \epsilon_0 = \frac{2a_0^3}{8a_0^2+4a_0^2a_1+3a_0a_1+4a_1^4}$, and $0 < \tau \le \tau_0$ with*

$$\tau_0 = \frac{1}{\epsilon_0}\frac{\theta - \frac{1}{2}}{\frac{5}{4}\theta^2 + \frac{a_1}{2a_0}\theta^2 + \frac{(1-\theta)^2}{4} + \frac{\theta(1-\theta)}{2}}.$$

Then there holds $E_{h,\epsilon}^{1,n} \le e^{-\epsilon(n-m)\tau/3}E_{h,\epsilon}^{1,m}$ for all $m \le n$.

Note that the maximal step size τ_0 only depends on a_0, a_1, and the choice of θ. Moreover, observe that the condition $\theta > 1/2$ is required here to make τ_0 positive.

Proof Following the arguments of the proof of Lemma 6.15, we start with

$$\begin{aligned} d_\tau E_{h,\epsilon}^{1,n} &= d_\tau E_h^{1,n} + \epsilon d_\tau(d_\tau u_h^n, u_h^{n,\theta}) \\ &\le -a_0 \|d_\tau u_h^{n,\theta}\|^2 - (\theta - \tfrac{1}{2})\tau\big(\|d_{\tau\tau} u_h^n\|^2 + \|d_{\tau\tau} p_h^n\|^2\big) + \epsilon d_\tau(d_\tau u_h^n, u_h^{n,\theta}). \end{aligned}$$

The last term can be expanded as

$$d_\tau(d_\tau u_h^n, u_h^{n,\theta}) = (d_\tau u_h^n, d_\tau u_h^{n,\theta}) + (d_{\tau\tau} u_h^n, u_h^{n-1,\theta}).$$

Using $d_\tau u_h^n = d_\tau u_h^{n,\theta} + (1-\theta)\tau d_{\tau\tau} u_h^n$, the first term of the above expression yields

$$(d_\tau u_h^n, d_\tau u_h^{n,\theta}) \le 2\|d_\tau u_h^{n,\theta}\|^2 + \tfrac{(1-\theta)^2\tau^2}{4}\|d_{\tau\tau} u_h^n\|^2.$$

To estimate the second term, we use the fact that besides (u_h^n, p_h^n) also $(d_\tau u_h^n, d_\tau p_h^n)$ satisfies Eqs. (6.14) and (6.15). This implies

$$\begin{aligned} (d_{\tau\tau} u_h^n, u_h^{n-1,\theta}) &= (d_\tau p_h^{n,\theta}, \partial_x u_h^{n-1,\theta}) - (a d_\tau u_h^{n,\theta}, u_h^{n-1,\theta}) \\ &= -(d_\tau p_h^{n,\theta}, d_\tau p_h^{n-1}) - (a d_\tau u_h^{n,\theta}, u_h^{n-1,\theta}). \end{aligned}$$

Using that $d_\tau p_h^{n-1} = d_\tau p_h^{n,\theta} - \theta\tau d_{\tau\tau} p_h^n$, we see that

$$\begin{aligned} -(d_\tau p_h^{n,\theta}, d_\tau p_h^{n-1}) &= -\|d_\tau p_h^{n,\theta}\|^2 + \theta\tau(d_\tau p_h^{n,\theta}, d_{\tau\tau} p_h^n) \\ &\le -\frac{3}{4}\|d_\tau p_h^{n,\theta}\|^2 + \theta^2\tau^2\|d_{\tau\tau} p_h^n\|^2. \end{aligned}$$

A discrete version of Lemma 6.13 allows us to bound

$$\|u_h^{n-1,\theta}\| \le \tfrac{1}{a_0}\|d_\tau u_h^{n-1}\| + \tfrac{a_1}{a_0}\|d_\tau p_h^{n-1}\|.$$

The remaining term in the above estimate can then be treated by

$$\begin{aligned} -(a d_\tau u_h^{n,\theta}, u_h^{n-1,\theta}) &\le \|d_\tau u_h^{n,\theta}\|\big(\tfrac{a_1}{a_0}\|d_\tau u_h^{n-1}\| + \tfrac{a_1^2}{a_0}\|d_\tau p_h^{n-1}\|\big) \\ &\le \|d_\tau u_h^{n,\theta}\|\big(\tfrac{a_1}{a_0}\|d_\tau u_h^{n,\theta}\| + \theta\tau\tfrac{a_1}{a_0}\|d_{\tau\tau} u_h^n\| + \tfrac{a_1^2}{a_0}\|d_\tau p_h^{n,\theta}\| + \theta\tau\tfrac{a_1^2}{a_0}\|d_{\tau\tau} p_h^n\|\big), \end{aligned}$$

where for the last step, we used the same expansion of p_h^{n-1} as above and a similar formula for u_h^{n-1}. Via Youngs inequalities and basic manipulations, we then obtain

$$\begin{aligned}-(ad_\tau u_h^{n,\theta}, u_h^{n-1,\theta}) \;&\le (\tfrac{3a_0a_1+4a_1^4}{2a_0^2})\|d_\tau u_h^{n,\theta}\|^2 + \tfrac{1}{4}\|d_\tau p_h^{n,\theta}\|^2 \\ &\quad + \tfrac{1}{4}\theta^2\tau^2\|d_{\tau\tau} p_h^n\|^2 + \tfrac{a_1}{2a_0}\theta^2\tau^2\|d_{\tau\tau} u_h^n\|^2.\end{aligned}$$

In summary, we thus arrive at

$$\begin{aligned}(d_{\tau\tau} u_h^n, u_h^{n-1,\theta}) \;&\le \tfrac{3a_0a_1+4a_1^4}{2a_0^2}\|d_\tau u_h^{n,\theta}\|^2 - \tfrac{1}{2}\|d_\tau p_h^{n,\theta}\|^2 \\ &\quad + \tfrac{5}{4}\theta^2\tau^2\|d_{\tau\tau} p_h^n\|^2 + \tfrac{a_1}{2a_0}\theta^2\tau^2\|d_{\tau\tau} u_h^n\|^2.\end{aligned}$$

Putting all estimates together, we finally obtain

$$\begin{aligned}d_\tau E_{h,\epsilon}^{1,n} \;&\le -\big(a_0 - \epsilon\tfrac{4a_0^2+3a_0a_1+4a_1^4}{2a_0^2}\big)\|d_\tau u_h^{n,\theta}\|^2 - \tfrac{\epsilon}{2}\|d_\tau p_h^{n,\theta}\|^2 \\ &\quad -\big(\theta - \tfrac{1}{2} - \epsilon\tau\tfrac{2a_1\theta^2+a_0(1-\theta)^2}{4a_0}\big)\tau\|d_{\tau\tau} u_h^n\|^2 - \big(\theta - \tfrac{1}{2} - \epsilon\tau\tfrac{5\theta^2}{4}\big)\tau\|d_{\tau\tau} p_h^n\|^2.\end{aligned}$$

By the particular choice of τ_0, we may estimate the terms in the second line from above by $-\frac{\epsilon}{2}\theta(1-\theta)\tau^2\big(\|d_{\tau\tau} u_h^n\|^2 + \|d_{\tau\tau} p_h^n\|^2\big)$. The two factors in the first line are balanced by the choice $\epsilon = \frac{2a_0^3}{5a_0^2+3a_0a_1+4a_1^4}$. In order to satisfy also the condition of Lemma 6.17, we enlarge the denominator by $3a_0^2 + 4a_0^2a_1$, and obtain

$$\begin{aligned}d_\tau E_{h,\epsilon}^{1,n} &\le -\epsilon_0\big\{\tfrac{1}{2}\big(\|d_\tau u_h^{n,\theta}\|^2 + \|d_\tau p_h^{n,\theta}\|^2\big) + \theta(1-\theta)\tfrac{\tau^2}{2}\big(\|d_{\tau\tau} u_h^n\|^2 + \|d_{\tau\tau} p_h^n\|^2\big)\big\} \\ &= -\epsilon_0\big(\theta E_h^{1,n} + (1-\theta)E_h^{1,n-1}\big) \le -\epsilon\big(\theta E_h^{1,n} + (1-\theta)E_h^{1,n-1}\big)\end{aligned}$$

for $\epsilon \le \epsilon_0$. By equivalence of the energies stated in Lemma 6.17, this leads to

$$E_{h,\epsilon}^{1,n} \le \tfrac{1-\frac{2}{3}\epsilon(1-\theta)\tau}{1+\frac{2}{3}\epsilon\theta\tau} E_{h,\epsilon}^{1,n-1} \le (1 - \tfrac{\epsilon\tau}{3})E_{h,\epsilon}^{1,n-1} \le e^{-\epsilon\tau/3} E_{h,\epsilon}^{1,n-1},$$

where we used that $\frac{2}{3}\epsilon\theta\tau \le \frac{2}{3}\epsilon_0\theta\tau_0 \le 1$ in the second step, which follows from the definition of τ_0. The assertion of the Lemma now follows by induction. □

Remark 6.2 Let us emphasize that the assertion of Lemma 6.18 holds true also for the choice $\theta = \frac{1}{2} + \lambda\tau$ with λ sufficiently large, but independent of τ.

Using the equivalence of the discrete energies stated in Lemma 6.17, we readily obtain the proof of Theorem 6.3 for the case $k = 1$. The result for $k = 0$ and $k \ge 2$ follows from the one for $k = 1$ with the same arguments as on the continuous level.

References

1. Babin, A.V., Vishik, M.I.: Regular attractors of semigroups and evolution equations. J. Math. Pures Appl. **62**, 441–491 (1983)
2. Babin, A.V., Vishik, M.I.: Attractors of Evolution Equations. Studies in Mathematics and Its Applications, vol. 25. North Holland, Amsterdam (1991)
3. Banks, H.T., Ito, K., Wang, C.: Exponentially stable approximations of weakly damped wave equations. In: Estimation and Control of Distributed Parameter Systems. International Series of Numerical Mathematics, vol. 100, pp. 6–33. Birkhäuser, Basel (1991)
4. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. RAIRO Anal. Numer. **2**, 129–151 (1974)
5. Brouwer, J., Gasser, I., Herty, M.: Gas pipeline models revisited: model hierarchies, non-isothermal models and simulations of networks. Multiscale Model. Simul. **9**, 601–623 (2011)
6. Chen, G.: Energy decay estimates and exact boundary controlability for wave equation in a bounded domain. J. Math. Pures Appl. **5**, 249–274 (1979)
7. Cox, S., Zuazua, E.: The rate at which energy decays in a damped string. Commun. Partial Differ. Equ. **19**, 213–243 (1994)
8. Dupont, T.: l^2 estimates for Galerkin methods for second-order hyperbolic equations. SIAM J. Numer. Anal. **10**, 880–889 (1973)
9. Egger, H., Kugler, T.: Uniform Exponential Stability of Galerkin Approximations for Damped Wave Systems (2015). arXive:1511.08341
10. Egger, H., Kugler, T.: Damped wave systems on networks: exponential stability and uniform approximations. Numer. Math. (2017). https://doi.org/10.1007/s00211-017-0924-4
11. Egger, H., Radu, B.: Super-convergence and post-processing for mixed finite element approximations of the wave equation (2016). arXiv:1608.03818
12. Ervedoza, S.: Observability properties of a semi-discrete 1d wave equation derived from a mixed finite element method on nonuniform meshes. ESAIM Control Optim. Calc. Var. **16**, 298–326 (2010)
13. Ervedoza, S., Zuazua, E.: Uniformly exponentially stable approximations for a class of damped systems. J. Math. Pures Appl. **91**, 20–48 (2009)
14. Evans, L.C.: Partial differential equations. In: Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (1998)
15. Fridman, E.: Observers and initial state recovering for a class of hyperbolic systems via lyapunov method. Automatica **49**, 2250–2260 (2013)
16. Gao, F., Chi, C.: Unconditionally stable difference schemes for a one-space-dimensional linear hyperbolic equation. Appl. Math. Comput. **187**, 1272–1276 (2007)
17. Geveci, T.: On the application of mixed finite element methods to the wave equations. RAIRO Model. Math. Anal. Numer. **22**, 243–250 (1988)
18. Glowinski, R., Kinton, W., Wheeler, M.F.: A mixed finite element formulation for the boundary controllability of the wave equation. Int. J. Numer. Methods Eng. **27**, 623–635 (1989)
19. Grote, M.J., Mitkova, T.: High-order explicit local time-stepping methods for damped wave equations. J. Comput. Appl. Math. **239**, 270–289 (2013)
20. Guinot, V.: Wave propagation in fluids: models and numerical techniques. ISTE and Wiley, London (2008)
21. Joly, P.: Variational methods for time-dependent wave propagation problems. In: Topics in Computational Wave Propagation, LNCSE, vol. 31, pp. 201–264. Springer, Heidelberg (2003)
22. Karaa, S.: Error estimates for finite element approximations of a viscous wave equation. Numer. Func. Anal. Optim. **32**, 750–767 (2011)
23. Lagnese, J.: Decay of solutions of wave equations in a bounded region with boundary dissipation. J. Differential Eq. **50**, 163–182 (1983)
24. Lopez-Gomez, J.: On the linear damped wave equation. J. Differential Eq. **134**, 26–45 (1997)
25. Pazy, A.: Semigroups of linear operators and applications to partial differential equations. In: Applied Mathematical Sciences, vol. 44, Springer, New York (1983)

26. Rauch, J., Taylor, M.: Exponential decay of solutions to hyperbolic equations in bounded domains. Indiana Univ. Math. J. **24**, 79–86 (1974)
27. Rincon, M.A., Copetti, M.I.M.: Numerical analysis for a locally damped wave equation. J. Appl. Anal. Comput. **3**, 169–182 (2013)
28. Schwab, C.: *p*- and *hp*-finite element methods. Numerical Mathematics and Scientific Computation. The Clarendon, Oxford University, New York (1998). Theory and applications in solid and fluid mechanics
29. Zuazua, E.: Stability and decay for a class of nonlinear hyperbolic problems. Asymptotic Anal. **1**, 161–185 (1988)
30. Zuazua, E.: Propagation, observation, and control of waves approximated by finite difference methods. SIAM Rev. **47**, 197–243 (2005)

Chapter 7
Adaptive Algorithm Based on Functional-Type A Posteriori Error Estimate for Reissner-Mindlin Plates

Maxim Frolov and Olga Chistiakova

Abstract This research is devoted to numerical justification of an adaptive mesh refinement algorithm based on functional-type a posteriori local error indicator for Reissner-Mindlin plates. Four stages of this algorithm (solver, estimator, marker and refiner) and its implementation are discussed. A number of numerical experiments for L-shape and skew Reissner-Mindlin plates is provided for verification and efficiency demonstration. It is also shown that efficiency index for functional-type a posteriori error estimate is stable and has acceptable value. As such a technique can be used with in-house implementation of finite element solver as well as with commercial software packages with closed sources, proposed algorithm may be applicable for engineering practice.

7.1 Introduction

In computational mechanics, there is a range of models of different complexity and application area for plate bending. One of the widely used is Reissner-Mindlin plate model, a generalization of the classical Kirchhoff-Love model, which is applicable to plates of small to moderate thickness. Since finite element modeling is important for the industry, there is constant activity in developing numerical methods for solving problems related to Reissner-Mindlin plates and in methods of a posteriori error estimation (see, for example, Beirão da Veiga et al. [4], Pechstein and Schöberl [14], Song and Niu [17], Frolov and Chistiakova [9] for reviews).

Mathematically, this problem is governed by elliptic PDE's. In case of linear statement the respective mathematical model describes the bending of linearly elastic plates of small to moderate thickness in terms of pair of variables in

M. Frolov (✉) · O. Chistiakova
Peter the Great St Petersburg Polytechnic University, St. Petersburg, Russia
e-mail: frolov_me@spbstu.ru

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_7

$\Omega \subset \mathbf{R}^2$: $u \in \mathbf{W}_2^1(\Omega)$-scalar-valued function (displacement) and $\theta \in \mathbf{W}_2^1(\Omega, \mathbf{R}^2)$-vector-valued function (rotations), where $\mathbf{W}_2^1(\ldots)$ are standard denotations of respective Sobolev spaces.

Let t be the thickness of a plate; $\lambda = \frac{Ek}{2(1+\nu)}$; $\varepsilon(\theta) = \frac{1}{2}(\nabla\theta + (\nabla\theta)^T)$; gt^3 represents the transverse loading; C-tensor of bending moduli; E and ν-material constants; k-correction factor. The equilibrium equations in this case are as follows:

$$\begin{cases} -\mathbf{Div}\,(\mathrm{C}\varepsilon(\theta)) = \gamma & \text{in } \Omega\,, \\ -\mathbf{div}\gamma = g & \text{in } \Omega\,, \\ \gamma = \lambda t^{-2}(\nabla u - \theta) & \text{in } \Omega\,. \end{cases} \tag{7.1}$$

For simplicity, boundary conditions of two types are considered. Let Γ_D and Γ_S be two non-intersecting parts of the boundary Γ. Γ_D is a clamped part ($u = 0, \theta = 0$ on Γ_D), Γ_S is a free part ($\partial u/\partial n = \theta \cdot n$, $\mathrm{C}\varepsilon(\theta)n = 0$ on Γ_S, where n is the outward unit normal to the boundary). If Γ_S is a supported part, one has $u = 0$ on Γ_S instead of $\partial u/\partial n = \theta \cdot n$ on Γ_S.

Let $\tilde{u}, \tilde{\theta}$ be some conforming approximations of solution, $\tilde{\gamma} = \lambda t^{-2}(\nabla\tilde{u} - \tilde{\theta})$. Then following errors (deviations) are introduced: $e_{\tilde{u}} = u - \tilde{u}$, $e_{\tilde{\theta}} = \theta - \tilde{\theta}$, $e_{\tilde{\gamma}} = \gamma - \tilde{\gamma}$. Functional approach to a posteriori error estimation is based on the introduction of additional variables with appropriate physical meaning (see Repin [15]). In case of Reissner-Mindlin plates they could be a vector $\tilde{\tilde{y}}$ and a tensor $\tilde{\tilde{\kappa}} = [\tilde{\tilde{\kappa}}^1,\ \tilde{\tilde{\kappa}}^2]$ with the same functional space for implementations:

$$\tilde{\tilde{y}},\ \ \tilde{\tilde{\kappa}}^1,\ \ \tilde{\tilde{\kappa}}^2 \ \in\ \mathbf{H}(\Omega, \mathbf{div}) := \left\{ y \in \mathbf{L}_2(\Omega, \mathbf{R}^2) \mid \mathbf{div} y \in \mathbf{L}_2(\Omega) \right\}\,. \tag{7.2}$$

The very first functional-type estimate for Reissner-Mindlin plates was derived in Repin and Frolov [16] and Frolov et al. [10]. A new variant has appeared in Frolov [8] in the following form:

$$|||e_{\tilde{\theta}}|||^2 + \lambda^{-1}t^2\|e_{\tilde{\gamma}}\|_\Omega^2 \le \hat{a}^2 + \lambda^{-1}t^2\hat{b}^2\,, \tag{7.3}$$

$$\begin{aligned} \hat{a} = {} & |||\mathrm{C}^{-1}\mathbf{sym}(\tilde{\tilde{\kappa}}) - \varepsilon(\tilde{\theta})||| + c_I\|\mathbf{skew}(\tilde{\tilde{\kappa}})\|_\Omega + \\ & + c_{II}c_{III}\sqrt{|\Omega|\ \|g + \mathbf{div}\tilde{\tilde{y}}\|_\Omega^2 \ +\ |\Gamma_\mathrm{S}|\ \|\tilde{\tilde{y}} \cdot n\|_{\Gamma_\mathrm{S}}^2} + \\ & + c_{IV}\sqrt{|\Omega|\ \|\tilde{\tilde{y}} + [\mathbf{div}\tilde{\tilde{\kappa}}^1, \mathbf{div}\tilde{\tilde{\kappa}}^2]\|_\Omega^2 \ +\ |\Gamma_\mathrm{S}|\ \|[\tilde{\tilde{\kappa}}^1 \cdot n, \tilde{\tilde{\kappa}}^2 \cdot n]\|_{\Gamma_\mathrm{S}}^2}\,, \end{aligned} \tag{7.4}$$

$$\hat{b} = \|\tilde{\tilde{y}} - \tilde{\gamma}\|_\Omega + c_{III}\sqrt{|\Omega|\ \|g + \mathbf{div}\tilde{\tilde{y}}\|_\Omega^2 \ +\ |\Gamma_\mathrm{S}|\ \|\tilde{\tilde{y}} \cdot n\|_{\Gamma_\mathrm{S}}^2}\,, \tag{7.5}$$

where

$$\|e_{\tilde{\gamma}}\|_\Omega := \sqrt{\int_\Omega |e_{\tilde{\gamma}}|^2\,\mathrm{d}x}, \qquad |||e_{\tilde{\theta}}||| := \sqrt{\int_\Omega \mathrm{C}\varepsilon(e_{\tilde{\theta}}) : \varepsilon(e_{\tilde{\theta}})\,\mathrm{d}x} \tag{7.6}$$

and the auxiliary constants c_I–c_{IV} are mesh-independent and come from the following standard inequalities:

$$\|\nabla\varphi\|_\Omega^2 \le c_I^2 |||\varphi|||^2 , \quad \|\varphi\|_\Omega^2 \le c_{II}^2 |||\varphi|||^2 ,$$

$$\frac{1}{|\Omega|}\|w\|_\Omega{}^2 + \frac{1}{|\Gamma_S|}\|w\|_{\Gamma_S}{}^2 \le c_{III}^2 \|\nabla w\|_\Omega{}^2 , \quad \frac{1}{|\Omega|}\|\varphi\|_\Omega{}^2 + \frac{1}{|\Gamma_S|}\|\varphi\|_{\Gamma_S}{}^2 \le c_{IV}^2 |||\varphi|||^2$$

$$\forall\varphi \in \left\{\varphi \in \mathbf{W}_2^1(\Omega, \mathbf{R}^2) \mid \varphi = 0 \text{ on } \Gamma_D\right\}, \forall w \in \left\{w \in \mathbf{W}_2^1(\Omega) \mid w = 0 \text{ on } \Gamma_D\right\}. \tag{7.7}$$

These constants are numerically estimated once for given boundary conditions, material and shape parameters of the plate.

Using Cauchy inequality with positive parameter, terms $\hat{a}$ and $\hat{b}$ can be represented without square roots. Note that this estimate is theoretically proven to be guaranteed and reliable.

A measure of a posteriori error majorant efficiency is overestimation of true error. It is provided by well-known quantity—efficiency index I_{eff} which is defined as majorant value divided by relevant norm of difference of given approximate solution and exact solution of the problem. When exact solution is unknown a common practice is to use an approximate solution on a much finer (reference) mesh. That is because as a guaranteed upper estimate is provided by the majorant, using reference mesh could yield minor overestimation of the efficiency index for linear problems, but no underestimation occurs: reference solution always gives a lower estimate of the true error.

Majorant (7.3) can be presented as a sum of local contributions on the elements of a mesh. These local errors can be treated as a posteriori local error indicator and used as a basis of adaptive mesh refinement algorithm. We use the combination of squares of the first terms of $\hat{a}$ and $\hat{b}$ as a counterpart of the square of the error.

7.2 Adaptive Mesh Refinement Algorithm

Main idea of adaptive mesh refinement process is to save computational resources and speed up computations by refining only parts with high relative errors instead of refining all mesh. Thus, it is hoped that the number of finite elements needed to achieve a desired accuracy would be significantly reduced. Adaptive algorithms consist of four main stages (e.g. Dörfler [7], Mekchay and Nochetto [13]): Solve (compute approximate solution on a current finite element mesh), Estimate (compute global error estimate and local indicators for each element), Mark (choose elements with large local errors) and Refine (split marked elements and locally refine the mesh).

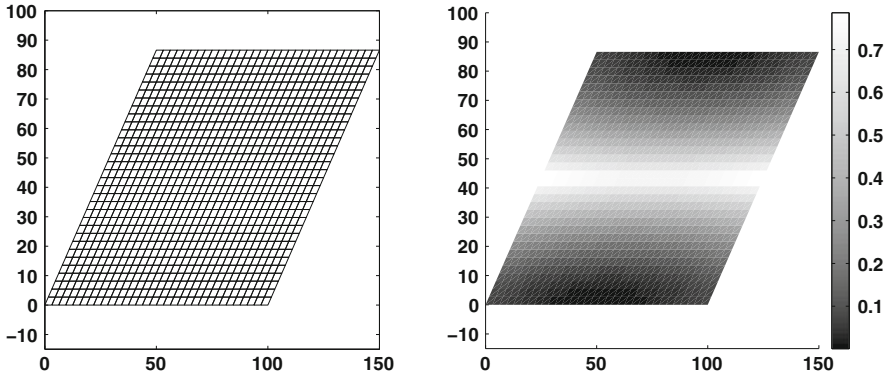


Fig. 7.1 Razzaque plate example: 32×32 mesh

We step by step demonstrate all these stages below on an example of Razzaque plate (Fig. 7.1). This is a skew plate with sides of equal length ($L = 100\,\text{m}$) and base angle 60°. In accordance with Carstensen et al. [5] we consider such a plate with thickness $t = 0.1\,\text{m}$, material constants $E = 1092\,\text{N/m}^2$ and $\nu = 0.3$. The loading is assumed to be uniform and horizontal edges of the plate are hard clamped.

7.2.1 Solver

One choice to solve a plate bending problem is to use commercial software packages like ANSYS as is done in Frolov and Chistiakova [9]. These packages are widely used both by industry and academia for quite a long time. However, to have an opportunity to control all steps of the adaptive algorithm without any "black-boxes" here we use our own implementation of a finite element solver. Applying standard Galerkin approximation was not an option due to well-known locking effect. That is why for this research we have implemented a solver based on well-known quadrilateral element with mixed interpolation of tensorial components (MITC4). These elements were first presented in Bathe and Dvorkin [2] (which is, among others, cited in ANSYS SHELL181-element documentation), are still used for Reissner-Mindlin plates and referred to as classical ones (see Wu and Wang [19], Cen and Shang [6]). Moreover, they are being further developed: for example, MITC4+ element was proposed in Ko et al. [12].

To verify our implementation we compare displacement in the central point of the plate with implementation in [5]. As can be seen in Table 7.1, the difference is relatively small and disappears during mesh refinements. Visualization of obtained displacements on a uniform mesh can be found on the right in Fig. 7.1.

Table 7.1 MITC4 solver (central deflection), error and efficiency index of majorant (7.3), and comparison with results presented in Carstensen et al. [5] for Razzaque skew plate on uniform meshes

	4×4		8×8		16×16		32×32	
Mesh	[5]	Mjr (7.3)	[5]	Mjr (7.3)	[5]	Mjr (7.3)	[5]	Mjr (7.3)
Disp.	0.67	0.64	0.76	0.75	0.78	0.78	0.79	0.79
Err.(%)	16.1	18.3	4.7	4.9	1.7	1.4	0.9	0.9
I_{eff}	0.35*	1.82	0.87*	1.76	0.97*	1.76	0.99*	1.77

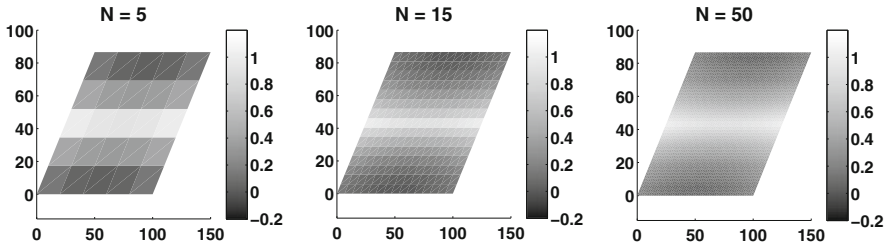


Fig. 7.2 Functional local error indicator for Razzaque plate

7.2.2 Estimator

Recent results show that using standard finite elements can be ineffective in case of functional a posteriori error estimates for problems like (7.1). It is more convenient to use natural approximations for $\mathbf{H}(\Omega, \mathbf{div})$ that are specially designed for mixed finite element methods. So we use the simplest Arnold-Boffi-Falk approximation [1] to implement the estimate. The auxiliary constants were computed via minimization on coarse meshes with overestimation by the factor 1.5 for reliability and set as follows: $c_I = 0.25$, $c_{II} = 0.20$, $c_{III} = 0.35$, $c_{IV} = 0.90$. Distribution of the local error indicator on uniform meshes for 5, 15 and 50 layers can be seen in Fig. 7.2. We compare results with residual-based approach from [5] in the last row of Table 7.1. Unfortunately, absolute values of efficiency index of a posteriori error estimate are not presented in [5] (instead a logarithm of efficiency index ratio on two consequent meshes to the base of two is shown so we mark it with $*$), but we need to mention here that efficiency index of functional a posteriori error estimate is stable and overestimation is in acceptable range, while the same ratio as in [5] is close to 1.00 for all meshes (even for coarse).

7.2.3 Marker

There are a number of strategies to choose elements that should be refined at the next iteration (see, for example, Verfürth [18] and Dörfler [7]). The perfect approach

should have low computational cost ($O(N)$, where N is number of elements in a mesh), should not break mesh structure (e.g. symmetry) and should reliably mark a satisfying number of elements (not all of them, not only a few—both choices will not stop the numerical procedure, but may lead to inefficiency of adaptive algorithm). However, satisfying all three points at once is a non-trivial task, so usually one of the two following popular strategies is used: mark by value (choose all elements with error larger than threshold) or mark by number (choose a certain number of elements with largest error). Marking by value is computationally efficient and keeps the structure of a mesh but has no control on number of elements that will be chosen for refinement. Marking by number strictly defines the number of chosen elements but at the cost of additional sorting operation and risk to break the symmetry of a mesh in case of symmetrical problems. We use another strategy which is as follows: at first, we set a number of thresholds, then on each step for each threshold count the number of elements that would be chosen if we use marking by value approach, then choose one of the thresholds and mark elements by value with it. In this case we keep all the advantages of marking by value strategy combined with an opportunity to control the degree of mesh refinement, which is important in practical applications. An example is presented in Fig. 7.3. At the top row the mesh is coloured by thresholds, while at the bottom row the final binary marking is shown. For the purpose of technique demonstration the thresholds were set as 1/4, 1/2, 3/4, and fraction of elements to refine was chosen to be most close to one third of the current mesh.

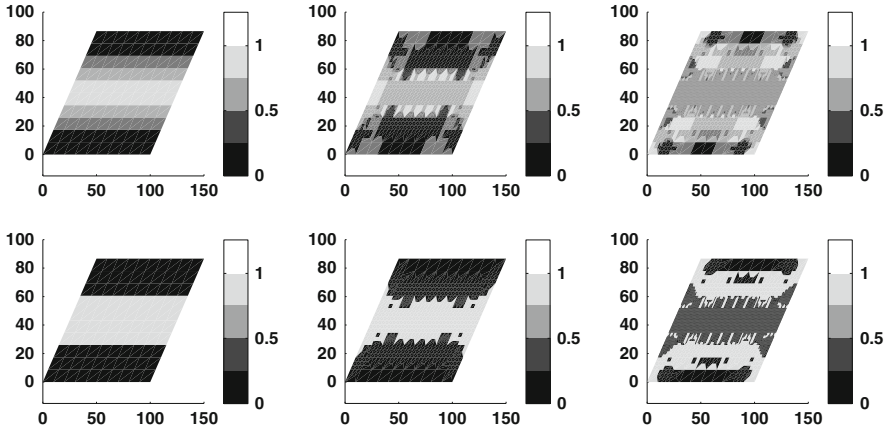


Fig. 7.3 Marking example: upper block—marking with thresholds 1/4, 1/2, 3/4, lower block—selection for refinements

7.2.4 *Refiner*

In general case refinement of quadrilateral elements is not a trivial task as certain properties of a mesh have to be kept. We use an algorithm proposed by Karavaev and Kopysov [11], which is a modification of Schneiders algorithm and can be used for mixed meshes and meshes with constraints. Elements that are marked for refinement are split into nine parts, and those incident to them—into three, seven or eight parts so that no hanging nodes occur during the refinement process. All in all there are five specially designed refinement templates. After that a global mesh smoothing takes place as mesh quality may have a critical impact on the accuracy of the obtained approximate solution. Example of the initial uniform mesh and its transformations during adaptation are presented in Fig. 7.4. Results of error estimation are provided in Table 7.2, where degrees of freedom (DOFs), the relative error (Err.) and the efficiency index are shown.

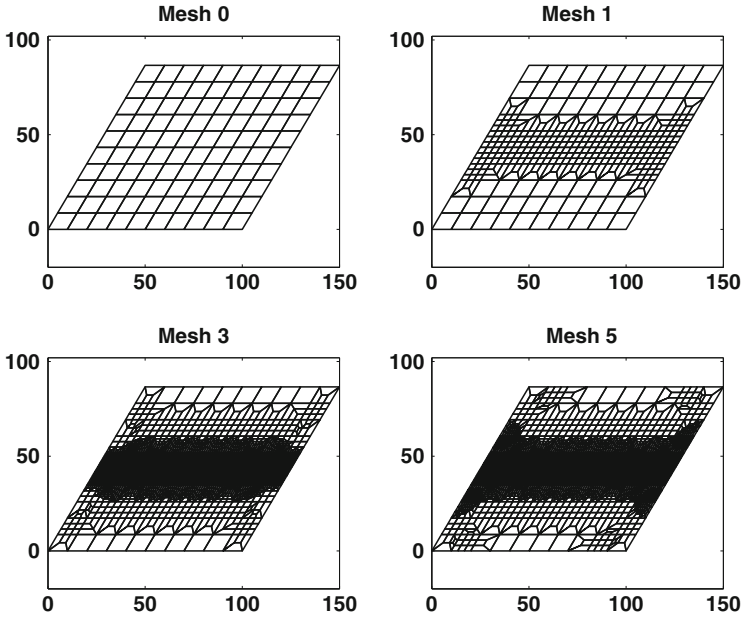


Fig. 7.4 Several mesh adaptation steps for Razzaque plate

Table 7.2 Error and efficiency index of majorant (7.3) for Razzaque skew plate

	Mesh	0	1	2	3	4	5
$t = 0.1$ m	DOFs	100	356	708	1361	1792	2676
	Err.(%)	3.6	1.3	0.9	0.8	0.6	0.5
	I_{eff}	1.8	1.8	1.8	1.8	1.8	1.8

7.3 More Numerical Experiments

Another popular example for numerical testing is L-shaped plate. We assume length of the full side $L = 2\,\text{m}$, material constants $E = 10.92\,\text{N/m}^2$ and $\nu = 0.3$. The loading is again assumed to be uniform. Two thicknesses of the plate ($t = 0.01\,\text{m}$, $t = 0.0001\,\text{m}$) and two types of boundary conditions are considered. In all cases analytical solution is unknown. We set the parameters according to Beirão da Veiga et al. [3] in which a similar problem was considered for triangular meshes. We show that in both cases mesh is mostly refined near the re-entrant corner which agrees with results presented in [3] and local singularity of the problem solution.

7.3.1 L-shape: Clamped Corner Case

At first we assume that only two edges of the plate that form re-entrant corner are hard clamped. All other edges are free. The auxiliary constants for this type of boundary conditions are: $c_I = 1.15$, $c_{II} = 2.15$, $c_{III} = 0.50$, $c_{IV} = 1.30$. Several adaptive mesh refinement steps can be seen in Fig. 7.5. Efficiency index is stable—it is presented in Table 7.3 for a number of adapted meshes.

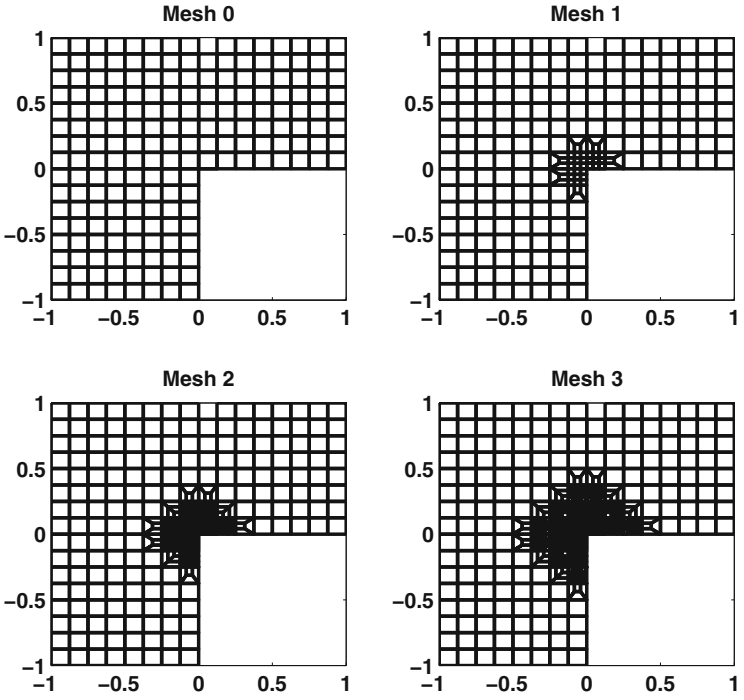


Fig. 7.5 Several mesh adaptation steps for L-shape plate (clamped corner case, $t = 0.0001\,\text{m}$)

Table 7.3 Error and efficiency index of majorant (7.3) for L-shape plate (clamped corner case)

	Mesh	0	1	2	3	4	5	6
$t = 0.01$ m	DOFs	192	264	782	902	1890	4472	5105
	Err.(%)	10.2	9.3	4.9	4.2	2.4	1.6	1.3
	I_{eff}	1.8	1.8	1.8	1.8	1.8	1.8	1.8
$t = 0.0001$ m	DOFs	192	242	752	986	2116	4028	5522
	Err.(%)	12.3	9.6	5.8	5.1	3.7	2.1	1.7
	I_{eff}	1.9	1.9	1.9	1.9	2.0	2.0	2.0

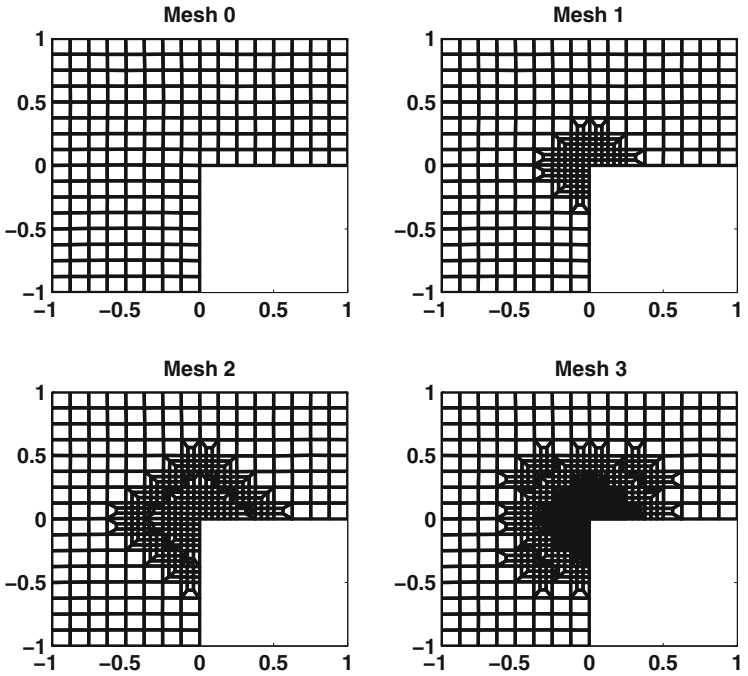


Fig. 7.6 Several mesh adaptation steps for L-shape plate (free corner case, $t = 0.01$ m)

7.3.2 L-shape: Free Corner Case

Vice versa, here we assume hard clamped all the edges except two forming the re-entrant corner. These two edges are free.The auxiliary constants in this case are: $c_I = 2.40$, $c_{II} = 0.60$, $c_{III} = 0.25$, $c_{IV} = 0.35$. Again we present mesh adaptation steps (Fig. 7.6) and the efficiency index (Table 7.4).

Table 7.4 Error and efficiency index of majorant (7.3) for L-shape plate (free corner case)

	Mesh	0	1	2	3	4	5	6
$t = 0.01$ m	DOFs	192	302	506	1496	2954	3872	8498
	Err.(%)	18.2	15.4	13.9	8.2	7.4	6.9	4.1
	I_{eff}	2.1	2.1	2.1	2.1	2.2	2.2	2.2
$t = 0.0001$ m	DOFs	192	302	576	1604	3014	6734	11,615
	Err.(%)	24.8	21.2	18.8	13.3	10.3	7.5	4.9
	I_{eff}	2.4	2.4	2.5	2.5	2.5	2.5	2.6

7.4 Conclusions

In this paper we give an overview of the adaptive mesh refinement approach based on functional-type local error indicator for Reissner-Mindlin plates. Several numerical examples are considered. It is demonstrated that analytically proven reliability of the functional approach, its specific implementation with a certain type of finite elements with combination to some known marking and refinement techniques provide a basis for adaptive algorithms that may be useful in engineering practice.

Acknowledgements This research is supported by the Grant of the President of the Russian Federation MD-1071.2017.1.

References

1. Arnold, D., Boffi, D., Falk, R.: Quadrilateral $H(\mathrm{div})$ finite elements. SIAM J. Numer. Anal. **42**, 2429–2451 (2005)
2. Bathe, K., Dvorkin, E.: A formulation of general shell elements—the use of mixed interpolation of tensoral components. Int. J. Numer. Methods Eng. **22**, 697–722 (1986)
3. Beirão da Veiga, L., Chinosi, C., Lovadina, C., Stenberg, R.: A-priori and a-posteriori error analysis for a family of Reissner-Mindlin plate elements. BIT Numer. Math. **48**, 189–213 (2008)
4. Beirão da Veiga, L., Mora, D., Rivera, G.: Virtual elements for a shear-deflection formulation of Reissner-Mindlin plates. Math. Comput. (2018)
5. Carstensen, C., Xie, X., Yu, G., Zhou, T.: A priori and a posteriori analysis for a locking-free low order quadrilateral hybrid finite element for Reissner-Mindlin plates. Comput. Methods Appl. Mech. Eng. **200**, 1161–1175 (2011)
6. Cen, S., Shang, Y.: Developments of Mindlin-Reissner plate elements. Math. Probl. Eng. **2015**, 12 (2015). Article ID 456740
7. Dörfler, W.: A convergent adaptive algorithm for poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
8. Frolov, M.: Reliable a posteriori error control for solutions to problems of Reissner-Mindlin plates bending. In: Procceding of 10th International Conference. Mesh methods for boundary-value problems and applications, Kazan, Russia, pp. 610–615 (2014)
9. Frolov, M., Chistiakova, O.: A functional-type a posteriori error estimate of approximate solutions for Reissner-Mindlin plates and its implementation. IOP Conf. Ser. Mater. Sci. Eng. **208**, 012043 (2017)

10. Frolov, M., Neittaanmäki, P., Repin, S.: Guaranteed functional error estimates for the Reissner-Mindlin plate problem. J. Math. Sci. **132**, 553–561 (2006). Translated from Problemy Matematicheskogo Analiza **31**, 159–167 (2005)
11. Karavaev, A., Kopysov, S.: A refinement of unstructured quadrilateral and mixed meshes. Vestn. Udmurtsk. Univ. Mat. Mekh. Komp. Nauk. **4**, 62–78 (2013)
12. Ko, Y., Lee, P.S., Bathe, K.: A new 4-node MITC element for analysis of two-dimensional solids and its formulation in a shell element. Comput. Struct. **192**, 34–49 (2017)
13. Mekchay, K., Nochetto, R.H.: Convergence of adaptive finite element methods for general second order linear elliptic PDEs. SIAM J. Numer. Anal. **43**, 1803–1827 (2005)
14. Pechstein, A.S., Schöberl, J.: An analysis of the TDNNS method using natural norms. Numer. Math. **139**, 93–120 (2018)
15. Repin, S.: A posteriori estimates for partial differential equations. Radon Series on Computational and Applied Mathematics, vol. 4. de Gruyter, Berlin (2008)
16. Repin, S., Frolov, M.: Estimation of deviations from the exact solution for the Reissner-Mindlin plate problem. J. Math. Sci. **132**, 331–338 (2006). Translated from Zapiski Nauchnykh Seminarov POMI **310**, 145–157 (2004)
17. Song, S., Niu, C.: A mixed finite element method for the Reissner-Mindlin plate. Bound. Value Probl. **1**, 194 (2016)
18. Verfürth, R.: A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques. Wiley/B.G. Teubner, Chichester/Stuttgart (1996)
19. Wu, C.T., Wang, H.P.: An enhanced cell-based smoothed finite element method for the analysis of Reissner–Mindlin plate bending problems involving distorted mesh. Int. J. Numer. Methods Eng. **95**, 288–312 (2013)

Chapter 8
Wavelet Boundary Element Methods: Adaptivity and Goal-Oriented Error Estimation

Helmut Harbrecht and Manuela Moor

Abstract This article is dedicated to the adaptive wavelet boundary element method. It computes an approximation to the unknown solution of the boundary integral equation under consideration with a rate N_{dof}^{-s}, whenever the solution can be approximated with this rate in the setting determined by the underlying wavelet basis. The computational cost scale linearly in the number N_{dof} of degrees of freedom. Goal-oriented error estimation for evaluating linear output functionals of the solution is also considered. An algorithm is proposed that approximately evaluates a linear output functional with a rate $N_{\mathrm{dof}}^{-(s+t)}$, whenever the primal solution can be approximated with a rate N_{dof}^{-s} and the dual solution can be approximated with a rate N_{dof}^{-t}, while the cost still scale linearly in N_{dof}. Numerical results for an acoustic scattering problem and for the point evaluation of the potential in case of the Laplace equation are reported to validate and quantify the approach.

8.1 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. The discretization of such problems will then amount to a large system of linear equations with a *dense* system matrix. Thus, the numerical solution of such problems requires large amounts of time and computation capacities.

H. Harbrecht (✉)
Departement Mathematik und Informatik, Universität Basel, Basel, Switzerland
e-mail: helmut.harbrecht@unibas.ch

M. Moor
Universitätsspital Basel, Basel, Switzerland
e-mail: manuela.moor@usb.ch

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_8

To overcome this obstruction, several ideas for the efficient approximation of the discrete system have been developed in the last decades. Most prominent examples are the *fast multipole method* [27, 37], the *panel clustering* [29], the *adaptive cross approximation* [3, 4], or *hierarchical matrices* [28, 38], all of which are known to reduce the computational cost to be nearly linear or even linear. A further approach is the *wavelet boundary element method* [7, 14, 31] which employs that the wavelets' vanishing moments lead, in combination with the fact that the kernels of integral operators become smoother when getting farther away from the diagonal, to a quasi-sparse system matrix. Since the number of relevant entries in the system matrix for maintaining the convergence rate of the underlying Galerkin method scales only linearly, wavelet matrix compression leads to a numerical algorithm that has linear cost.

Another issue to be addressed for the efficient discretization of boundary integral equations is the one of adaptivity. For non-smooth geometries or right-hand sides, it is necessary to be able to resolve specific parts of the geometry, while other parts could stay coarse. The *adaptive* wavelet boundary element method has been developed in [16, 25, 33], based on the ideas of related adaptive wavelet methods for local operators from [9, 10, 26]. Assume that the solution of the boundary integral equation to be solved is known and can be approximated with a rate N_{dof}^{-s} in the setting determined by the underlying wavelet basis. Then, the adaptive wavelet boundary element method computes an approximation that converges with a rate N_{dof}^{-s} at a cost expense that scales linearly with N_{dof}. The method is hence *computationally optimal*. Although reliable error estimators for boundary integral operators exist, see e.g., [19], and optimal convergence of traditional boundary element discretizations have been proven, see, e.g., [20, 24], we are not aware of any other boundary element method which is optimal in this sense.

For many applications one is not interested in the unknown solution, but only in a continuous, linear output functional of it. Approximating this new quantity of interest instead is referred to as goal-oriented method. By considering only an output functional, one is able to perform the computation with much less degrees of freedom. This *goal-oriented adaptivity* has intensively been studied in the field of adaptive finite element methods, see e.g. [2, 5, 6, 15, 18, 22, 36] and the references therein. Optimal convergence rates have been analyzed in [22, 36]. For goal-oriented adaptive boundary element methods, only few results can be found [1, 21, 22]. The combination of goal-oriented adaptivity and fast boundary element methods has, however, not been considered yet.

We will present a goal-oriented strategy for the adaptive wavelet boundary element method. The strategy is in accordance with [36] and separately minimizes the error of the primal problem and the error of the dual problem, respectively. One computes two index sets which indicate the possible refinement, one for the primal problem and one for the dual problem. By choosing the smaller index for refinement, the functional evaluation converges at a rate $N_{\mathrm{dof}}^{-(s+t)}$, whenever the primal solution can be approximated at a rate N_{dof}^{-s} and the dual solution can be approximated at a rate N_{dof}^{-t}. The advantage of using the adaptive wavelet boundary

element method instead of a traditional boundary element method as in [20–22] is that the computational cost of the algorithm scales linearly with respect to the number N_{dof} of degrees of freedom.

We would like to mention at this point that the goal-oriented approach is not restricted to linear output functionals, but can also be extended to non-linear output functionals, see e.g. [2] and the references therein. However, for the sake of a thorough convergence analysis, we are considering only a linear output functional here.

The outline is as follows. In Sect. 8.2, we recall the adaptive wavelet boundary element method. Then, in Sect. 8.3, we propose the goal-oriented refinement strategy. Numerical results for an acoustic scattering problem and for point evaluations of the single layer potential in case of the Laplace equation are presented in Sect. 8.4. Finally, concluding remarks are stated in Sect. 8.5.

Throughout this article, in order to avoid the repeated use of generic but unspecified constants, we mean by $C \lesssim D$ 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 $C \gtrsim D$.

8.2 Adaptive Wavelet Methods for Boundary Integral Equations

8.2.1 Problem Formulation

Let $\Omega \subset \mathbb{R}^{n+1}$ be a bounded domain with Lipschitz-smooth boundary $\Gamma = \partial\Omega$. Adaptive wavelet methods rely on an iterative solution method for the *continuous boundary integral equation*

$$(\mathscr{A}u)(\boldsymbol{x}) = \int_\Gamma k(\boldsymbol{x}, \boldsymbol{y})u(\boldsymbol{y})\mathrm{d}\sigma_{\boldsymbol{y}} = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma, \tag{8.1}$$

under consideration, expanded with respect to a wavelet basis. Here, $\mathscr{A} : H^q(\Gamma) \to H^{-q}(\Gamma)$ denotes an elliptic, symmetric, and continuous boundary integral operator[1] of order $2q$ with standard kernel k, satisfying

$$\left|\partial_{\widehat{\boldsymbol{x}}}^{\boldsymbol{\alpha}}\partial_{\widehat{\boldsymbol{y}}}^{\boldsymbol{\beta}}k(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})\right| \le c_{\boldsymbol{\alpha},\boldsymbol{\beta}}\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|^{-(|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|+n+2q)}$$

for all $\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}} \in \Gamma$ with $\widehat{\boldsymbol{x}} \neq \widehat{\boldsymbol{y}}$ provided that $|\boldsymbol{\alpha}| + |\boldsymbol{\beta}| + n + 2q \ge 0$, where the derivation has to be understood with respect to the surface coordinates. We should

[1] In accordance with [23], one might also consider a compact perturbation of an elliptic, symmetric, and continuous boundary integral operator.

remark that the kernel of a boundary integral operator $\mathscr{A}$ of order $2q$ is in general a standard kernel of order $2q$, see, e.g., [34]. This holds especially true for the kernel function associated with the Laplace and the Helmholtz equation, the system of Navier-Lamé equations and the Stokes system.

Having at hand a wavelet basis Ψ for the underlying energy space $H^q(\Gamma)$, the Riesz property

$$\|\Psi \boldsymbol{u}\|_{H^q(\Gamma)} \sim \|\boldsymbol{u}\| \quad \text{for all} \quad \boldsymbol{u} \in \ell^2$$

constitutes an isomorphism between $u \in H^q(\Gamma)$ and $\boldsymbol{u} \in \ell^2$. Especially, (8.1) is equivalent to the well-posed problem of finding $u = \Psi \boldsymbol{u}$ such that the *bi-infinite dimensional system of linear equations*

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}, \quad \text{where} \quad \boldsymbol{A} := \langle \mathscr{A}\Psi, \Psi \rangle \quad \text{and} \quad \boldsymbol{f} := \langle f, \Psi \rangle, \tag{8.2}$$

holds. Suitable wavelet bases on surfaces have, for example, been constructed in [12, 13, 30, 32].

8.2.2 Building Blocks

For the approximate solution of the infinite dimensional system (8.2) of linear equations, one has to perform matrix-vector multiplications by means of adaptive applications of the operator $\boldsymbol{A}$ under consideration. The building blocks COARSE, APPLY, RHS, and SOLVE, which are needed to design an adaptive algorithm of optimal complexity, have been identified in [9, 10].

In order to specify the properties of the building blocks, we shall introduce the approximation spaces

$$\ell^w_\tau = \left\{ \boldsymbol{u} \in \ell^2 : |\boldsymbol{u}|_{\ell^w_\tau} := \sup_{N \in \mathbb{N}} N^{-1/\tau} |\gamma_N(\boldsymbol{u})| < \infty \right\},$$

where $\gamma_N(\boldsymbol{u})$ denotes the N-th largest coefficient in modulus of the vector $\boldsymbol{u}$. It holds $\boldsymbol{u} \in \ell^w_\tau$ whenever $u = \Psi \boldsymbol{u}$ is contained in the Besov space $B^{q+ns}_\tau(\Gamma)$ with $\tau = (s + 1/2)^{-1}$, see, e.g., [17]. We require that the following statements hold true for $\overline{s} := {}^{(d-q)}/_n$, where d denotes the order of polynomials which can be represented exactly by the wavelet discretization:

- *Matrix-vector multiplication:* $\boldsymbol{w}_{\Lambda'} = \text{APPLY}[\varepsilon, \boldsymbol{v}_\Lambda]$. Let $\varepsilon > 0$ and let $\boldsymbol{v}_\Lambda$ consist of $|\Lambda| < \infty$ non-zero coefficients. Then, the output $\boldsymbol{w}_{\Lambda'}$ satisfies

$$\|\boldsymbol{A}\boldsymbol{v}_\Lambda - \boldsymbol{w}_{\Lambda'}\| \leq \varepsilon$$

where, for any $s \leq \overline{s}$, only

$$|\Lambda'| \lesssim \varepsilon^{-1/s} |\boldsymbol{v}_\Lambda|_{\ell_\tau^w}^{1/s}$$

coefficients are non-zero. The number of arithmetic operations and storage locations used by this call is bounded by some absolute multiple of

$$\varepsilon^{-1/s} |\boldsymbol{v}_\Lambda|_{\ell_\tau^w}^{1/s} + |\Lambda| + 1.$$

- *Approximation of the right-hand side:* $\boldsymbol{f}_\Lambda = \textsc{Rhs}[\varepsilon]$. Given $\varepsilon > 0$, the output $\boldsymbol{f}_\Lambda$ satisfies

$$\|\boldsymbol{f} - \boldsymbol{f}_\Lambda\| \leq \varepsilon,$$

and, for any $s \leq \overline{s}$, if $\boldsymbol{u} \in \ell_\tau^w$, then

$$|\Lambda| \lesssim \varepsilon^{-1/s} |\boldsymbol{u}|_{\ell_\tau^w}^{1/s}.$$

The number of arithmetic operations and storage locations used by the call is bounded by some absolute multiple of

$$\varepsilon^{-1/s} |\boldsymbol{u}|_{\ell_\tau^w}^{1/s} + 1.$$

- *Galerkin solver:* $\boldsymbol{w}_\Lambda = \textsc{Solve}[\varepsilon, \Lambda]$. This routine computes the solution $\boldsymbol{w}_\Lambda$ of the system of linear equations

$$\boldsymbol{A}_\Lambda \boldsymbol{u}_\Lambda = \boldsymbol{f}_\Lambda, \quad \text{where} \quad \boldsymbol{A}_\Lambda := \langle \mathscr{A} \Psi_\Lambda, \Psi_\Lambda \rangle \quad \text{and} \quad \boldsymbol{f}_\Lambda := \langle f_\Lambda, \Psi_\Lambda \rangle, \tag{8.3}$$

with accuracy

$$\|\boldsymbol{u}_\Lambda - \boldsymbol{w}_\Lambda\| \leq \varepsilon.$$

The number of arithmetic operations and storage locations used by the call is bounded by some absolute multiple of

$$\varepsilon^{-1/s} |\boldsymbol{u}_\Lambda|_{\ell_\tau^w}^{1/s} + |\Lambda| + 1$$

provided that $\boldsymbol{u}_\Lambda \in \ell_\tau^w$ for some $s \leq \overline{s}$.

- *Coarsening routine:* $\boldsymbol{w}_{\Lambda'} = \textsc{Coarse}[\theta, \boldsymbol{w}_\Lambda]$. This routine produces for given $0 \leq \theta \leq 1$ an index set $\Lambda' \subset \Lambda$ such that the restriction $\boldsymbol{w}_{\Lambda'}$ of the input vector $\boldsymbol{w}_\Lambda$ satisfies

$$\|\boldsymbol{w}_{\Lambda'}\| \leq \theta \|\boldsymbol{w}_\Lambda\|,$$

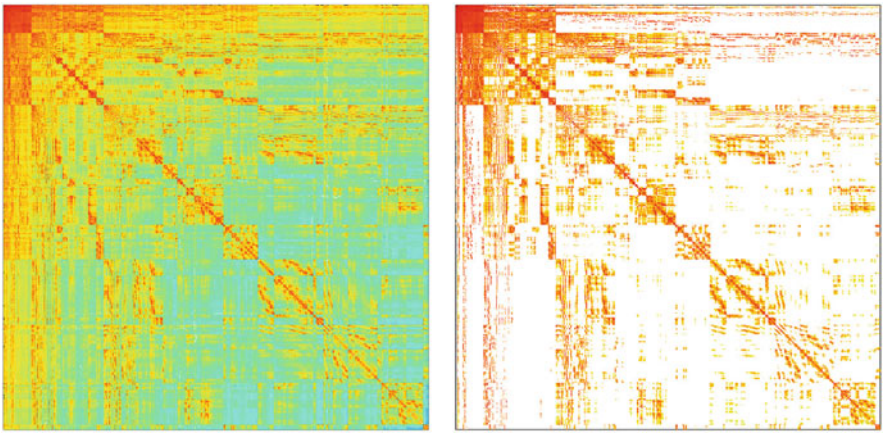


Fig. 8.1 Original matrix (left) and compressed matrix (right)

where $|\Lambda'|$, up to some absolute constant factor, is minimal. The computational complexity is bounded by some absolute multiple of $|\Lambda|$.

Our particular implementation of these building blocks, satisfying all desired properties, is based on piecewise constant wavelets (i.e., $d = 1$). In particular, we restrict the set of active wavelet functions to tree constraints which ensures the method's efficient implementation. Note that the coarsening routine for trees originates from [8], while the realization of RHS requires some a-priori knowledge on the right-hand side f. The matrix-vector multiplication APPLY has been constructed in [16, 33], see also [25] for related results. We mention that the main ingredient is wavelet matrix compression to sparsify the system matrix of the boundary integral operator under consideration, see Fig. 8.1 for an illustration. Straightforward modifications of RHS and APPLY yield finally the routine SOLVE, cf. [26]. We skip all the details here and refer the reader to the cited references.

8.2.3 An Adaptive Boundary Element Method

The specific adaptive algorithm we use has been proposed in [26] and is similar to classical methods which consist of the following steps:

$$\boxed{\text{SOLVE}} \longrightarrow \boxed{\text{ESTIMATE}} \longrightarrow \boxed{\text{MARK}} \longrightarrow \boxed{\text{REFINE}}$$

For a given (finite) index set $\Lambda \subset \ell^2$, we *solve* the Galerkin system (8.3) via $\boldsymbol{u}_\Lambda = \text{SOLVE}[\varepsilon, \Lambda]$ with appropriate accuracy $\varepsilon > 0$. To *estimate* the (infinite) residual $\boldsymbol{r} := \boldsymbol{f} - \boldsymbol{A}\boldsymbol{u}_\Lambda$, we compute an approximation $\boldsymbol{r}_{\Lambda'}$ relative to a finite index set

Algorithm 1: Approximation $\boldsymbol{r}_{\Lambda'} = \textsc{Estimate}[\delta, \boldsymbol{u}_\Lambda]$ of the residual

Data: initial precision δ and approximate solution $\boldsymbol{u}_\Lambda$
do
  ▷ update $\delta \leftarrow \delta/2$;
  ▷ calculate $\boldsymbol{r}_{\Lambda'} = \textsc{Rhs}[\delta] - \textsc{Apply}[\delta, \boldsymbol{u}_\Lambda]$;
while $2\delta > \omega \|\boldsymbol{r}_{\Lambda'}\|$;

$\Lambda \subset \Lambda' \subset \ell^2$ such that

$$(1-\omega)\,\|\boldsymbol{r}_{\Lambda'}\| \le \|\boldsymbol{r}\| \le (1+\omega)\,\|\boldsymbol{r}_{\Lambda'}\| \tag{8.4}$$

for some fixed constant $0 < \omega < 1$. This can be realized by calling

$$\boldsymbol{r}_{\Lambda'} = \textsc{Estimate}[\delta, \boldsymbol{u}_\Lambda]$$

for an appropriately chosen initial precision $\delta > 0$. The routine is defined in Algorithm 1, where the until-clause $2\delta \le \omega\,\|\boldsymbol{r}_{\Lambda'}\|$ ensures that the iteration terminates when (8.4) holds since

$$\|\boldsymbol{r}\| \ge \|\boldsymbol{r}_{\Lambda'}\| - \|\boldsymbol{r} - \boldsymbol{r}_{\Lambda'}\| \ge \|\boldsymbol{r}_{\Lambda'}\| - 2\delta \ge (1-\omega)\,\|\boldsymbol{r}_{\Lambda'}\|$$

on the one hand and

$$\|\boldsymbol{r}\| \le \|\boldsymbol{r}_{\Lambda'}\| + \|\boldsymbol{r} - \boldsymbol{r}_{\Lambda'}\| \le \|\boldsymbol{r}_{\Lambda'}\| + 2\delta \le (1+\omega)\,\|\boldsymbol{r}_{\Lambda'}\|$$

on the other hand.

The supporting index set Λ' of the approximate residual $\boldsymbol{r}_{\Lambda'}$ enlarges the original index set Λ such that the Galerkin solution with respect to Λ' would reduce the current error by a constant factor. Nonetheless, we need to coarsen the index set Λ' for controlling the complexity. This is done by calling the Coarse-routine

$$\boldsymbol{r}_{\Lambda''} = \textsc{Coarse}[\theta, \boldsymbol{r}_{\Lambda'}].$$

for fixed $0 < \theta < 1$ sufficiently small. It combines the steps *mark* and *refine* since the new index set $\Lambda'' \subset \Lambda'$ still enlarges the original index set Λ, which corresponds to mesh refinement. Especially, it holds $\|\boldsymbol{r}_{\Lambda''}\| \sim \|\boldsymbol{r}\|$. Hence, the algorithm's convergence is guaranteed when repeating the procedure again with $\Lambda := \Lambda''$. For all the details of the particular implementation, we refer the reader to [39].

Algorithm 2: The adaptive wavelet boundary element method

Data: initial index set Λ_0, initial precision δ, and parameters $0 < \gamma, \theta < 1$
$\triangleright$ set $\Lambda := \Lambda_0$;
do
  $\triangleright$ compute the Galerkin solution $\boldsymbol{u}_\Lambda = \textsc{Solve}[\gamma\delta, \Lambda]$;
  $\triangleright$ compute the residual $\boldsymbol{r}_{\Lambda'} = \textsc{Estimate}[\delta, \boldsymbol{u}_\Lambda]$ and set $\delta = \|\boldsymbol{r}_{\Lambda'}\|$;
  $\triangleright$ coarse $\boldsymbol{r}_{\Lambda''} = \textsc{Coarse}[\theta, \boldsymbol{r}_{\Lambda'}]$ and update $\Lambda \leftarrow \Lambda''$;

In accordance with [16, 26], having at hand the building blocks COARSE, APPLY, RHS, and SOLVE with the properties specified in Sect. 8.2.2, the following statement provides the optimality of Algorithm 2. Note that the hidden constant depends on the boundary integral operator under consideration, the wavelet basis and the choice of the parameters.

Theorem 8.1 *Let $0 < \gamma, \theta < 1$ be sufficiently small parameters and let $\boldsymbol{u} \in \ell^w_\tau$ with $\tau = (s + 1/2)^{-1}$ for some $s \leq \overline{s}$. Then, Algorithm 2 computes iterates $\boldsymbol{u}_\Lambda$, which satisfy the error estimate*

$$\|\boldsymbol{u} - \boldsymbol{u}_\Lambda\| \lesssim |\Lambda|^{-s},$$

at a computational expense that stays proportional to the number $|\Lambda|$ of degrees of freedom.

Proof The assertion follows from the abstract theory presented in [26] for elliptic, symmetric, and continuous operators (cf. [26, Theorem 2.7]). The extension to compact perturbations of such operators is found in [24].

8.3 Goal-Oriented Adaptivity

We shall motivate the key idea of goal-oriented error estimation. To that end, let $a : V \times V \to \mathbb{R}$ be an elliptic and continuous bilinear form and $f \in V'$. Consider the variational formulation

$$\text{seek } u \in V \text{ such that } a(u, v) = \langle f, v\rangle \text{ for all } v \in V$$

and the associated Galerkin scheme

$$\text{seek } u_\Lambda \in V_\Lambda \text{ such that } a(u_\Lambda, v_\Lambda) = \langle f, v_\Lambda\rangle \text{ for all } v_\Lambda \in V_\Lambda,$$

where $V_\Lambda \subset V$ denotes the trial space. At first glance, we obtain the error estimate

$$|\langle g, u\rangle - \langle g, u_\Lambda\rangle| = |\langle g, u - u_\Lambda\rangle| \leq \|g\|_{V'}\|u - u_\Lambda\|_V$$

for the evaluation of an output functional $g \in V'$. Nonetheless, by introducing the *dual* or *adjoint* solution

$$\text{seek } z \in V \text{ such that } a(v, z) = \langle g, v \rangle \text{ for all } v \in V$$

and observing Galerkin orthogonality, we conclude that

$$|\langle g, u \rangle - \langle g, u_\Lambda \rangle| = \min_{z_\Lambda \in V_\Lambda} |a(u - u_\Lambda, z - z_\Lambda)| \lesssim \min_{z_\Lambda \in V_\Lambda} \|u - u_\Lambda\|_V \|z - z_\Lambda\|_V .$$

This fact greatly improves the error estimate and is exploited in the sequel.

Based on the adaptive wavelet boundary element method, proposed in the previous section, we can formulate a goal-oriented adaptive strategy to efficiently evaluate linear output functionals of the solution to the boundary integral equation (8.1) under consideration. As motivated above, we have now to synchronously approximate the solutions $\boldsymbol{u}, \boldsymbol{z} \in \ell^2$ of the two systems of linear equations,

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f} \quad \text{and} \quad \boldsymbol{A}^\intercal \boldsymbol{z} = \boldsymbol{g}, \tag{8.5}$$

where $\boldsymbol{A}$ and $\boldsymbol{f}$ are defined as in (8.2) and $\boldsymbol{g} = \langle g, \Psi \rangle$ denotes the discretized output functional, where $\langle g, u \rangle = \boldsymbol{g}^\intercal \boldsymbol{u}$. We therefore modify the adaptive wavelet boundary element method from Algorithm 2 as follows.

For a given finite index set Λ, the primal and dual systems of linear equations are solved with sufficient accuracy. This yields the approximations $\boldsymbol{u}_\Lambda$ and $\boldsymbol{z}_\Lambda$ to the primal and dual solution of (8.5), respectively. We then call

$$\boldsymbol{r}_{p,\Lambda_p} = \text{ESTIMATE}_{\text{primal}}[\delta_p, \boldsymbol{u}_\Lambda] \quad \text{and} \quad \boldsymbol{r}_{d,\Lambda_d} = \text{ESTIMATE}_{\text{dual}}[\delta_d, \boldsymbol{z}_\Lambda],$$

which refer to the primal and dual versions of the routine ESTIMATE as outlined in Algorithm 1. The input parameters δ_p and δ_d are initialized at the beginning by a δ_{init} of our choice, and they are modified during the course of the algorithm as outlined in Algorithm 2.

Next, we call

$$\boldsymbol{r}_{p,\Lambda'_p} = \text{COARSE}[\theta, \boldsymbol{r}_{p,\Lambda_p}] \quad \text{and} \quad \boldsymbol{r}_{d,\Lambda'_d} = \text{COARSE}[\theta, \boldsymbol{r}_{d,\Lambda_d}]$$

to compute appropriate refinements $\Lambda'_p, \Lambda'_d \supset \Lambda$ of the original index set Λ. Finally, we choose the smaller of the two index sets Λ_p and Λ_d to update the index set Λ and restart the loop.

The aforementioned goal-oriented adaptive refinement strategy is summarized in Algorithm 3. In accordance with [21, 22, 36], we derive the following result on the goal-oriented wavelet boundary element method, provided that the building blocks COARSE, APPLY, RHS, and SOLVE satisfy the properties specified in Sect. 8.2.2.

Algorithm 3: Goal-oriented refinement strategy

Data: initial index set Λ_0, initial precision δ, and parameters $0 < \gamma, \theta < 1$
$\triangleright$ set $\Lambda := \Lambda_0$ and $\delta_p = \delta_d = \delta$;
do
  $\triangleright$ compute the Galerkin solutions

$$\boldsymbol{u}_\Lambda = \textsc{Solve}_{\text{primal}}[\gamma\delta_p, \Lambda]$$
$$\boldsymbol{z}_\Lambda = \textsc{Solve}_{\text{dual}}[\gamma\delta_d, \Lambda]$$

  $\triangleright$ compute the residuals

$$\boldsymbol{r}_{p,\Lambda_p} = \textsc{Estimate}_{\text{primal}}[\delta_p, \boldsymbol{u}_\Lambda]$$
$$\boldsymbol{r}_{d,\Lambda_d} = \textsc{Estimate}_{\text{dual}}[\delta_d, \boldsymbol{z}_\Lambda]$$

  $\triangleright$ set $\delta_p = \|\boldsymbol{r}_{p,\Lambda_p}\|$ and $\delta_d = \|\boldsymbol{r}_{d,\Lambda_d}\|$;
  $\triangleright$ coarse $\boldsymbol{r}_{p,\Lambda'_p} = \textsc{Coarse}[\theta, \boldsymbol{r}_{p,\Lambda_p}]$ and $\boldsymbol{r}_{d,\Lambda'_d} = \textsc{Coarse}[\theta, \boldsymbol{r}_{d,\Lambda_d}]$;
  $\triangleright$ if $|\Lambda'_d| \le |\Lambda'_p|$, then set $\Lambda = \Lambda'_d$, otherwise set $\Lambda = \Lambda'_p$;

Proposition 8.1 *Let $0 < \gamma, \theta, \omega < 1$ be sufficiently small parameters and let $\boldsymbol{u} \in \ell^w_{\tau_p}$ with $\tau_p = (s + 1/2)^{-1}$ and $\boldsymbol{z} \in \ell^w_{\tau_d}$ with $\tau_d = (t + 1/2)^{-1}$ for some $s, t \le \overline{s}$. Then, the approximation $\boldsymbol{u}_\Lambda$ of the primal solution $\boldsymbol{u}$ and the approximation $\boldsymbol{z}_\Lambda$ of the dual solution $\boldsymbol{z}$, computed by Algorithm 3, satisfy the error estimate*

$$\|\boldsymbol{u} - \boldsymbol{u}_\Lambda\|\|\boldsymbol{z} - \boldsymbol{z}_\Lambda\| \lesssim |\Lambda|^{-(s+t)}. \tag{8.6}$$

The computational expense to compute these approximations scales proportional to the number $|\Lambda|$ of degrees of freedom.

Proof By construction, the norms of the approximate primal residual $\boldsymbol{r}_{p,\Lambda_p}$ and the approximate dual residual $\boldsymbol{r}_{d,\Lambda_d}$ are always equivalent to the respective exact residual:

$$(1-\omega)\,\|\boldsymbol{r}_{p,\Lambda_p}\| \le \|\boldsymbol{r}_p\| \le (1+\omega)\,\|\boldsymbol{r}_{p,\Lambda_p}\|,$$
$$(1-\omega)\,\|\boldsymbol{r}_{d,\Lambda_d}\| \le \|\boldsymbol{r}_d\| \le (1+\omega)\,\|\boldsymbol{r}_{d,\Lambda_d}\|.$$

Therefore, we have

$$\|\boldsymbol{u} - \boldsymbol{u}_\Lambda\|\|\boldsymbol{z} - \boldsymbol{z}_\Lambda\| \sim \|\boldsymbol{r}_{p,\Lambda_p}\|\|\boldsymbol{r}_{d,\Lambda_d}\| \sim \delta_p\delta_d.$$

Thus, $|\Lambda| \lesssim (\delta_p\delta_d)^{-1/(s+t)}$ in accordance with [36, Theorem 5.5], which implies (8.6). The complexity bound is finally an immediate consequence of the optimality of the adaptive wavelet method.

8.4 Numerical Results

8.4.1 Scattering Problems

We shall present numerical results for an acoustic scattering problem with different wavenumbers $\kappa \geq 1$. The choice of larger wavenumbers has a direct effect on the sparsity of the system matrix, as the compression parameters have to be proportionally increased with the wavenumber, see [35]. Therefore, the method cannot be expected to be robust with respect to the wavenumber.

Given a sound-soft scatterer $\Omega \subset \mathbb{R}^3$, we consider the solution u of the exterior Helmholtz equation

$$\Delta u + \kappa u = 0 \text{ in } \mathbb{R}^3 \setminus \overline{\Omega}, \quad u = 0 \text{ on } \Gamma := \partial\Omega. \tag{8.7}$$

The function u consists of an incident and a scattered wave, i.e. $u = u^s + u^i$, where in addition to (8.7) the scattered wave satisfy the Sommerfeld radiation condition

$$\lim_{r\to\infty} r \left(\frac{\partial u^s}{\partial r} - i\kappa u^s \right) = 0 \text{ as } r = \|\boldsymbol{x}\| \to \infty.$$

The incident wave u^i is known and is of the form $\exp(i\kappa \boldsymbol{d}\boldsymbol{x})$, where $\boldsymbol{d}$ denotes the direction (it holds $\|\boldsymbol{d}\| = 1$), and the goal is to compute the scattered wave u^s. When u^s is given, the solution u to the Helmholtz equation (8.7) can be computed as well.

In order to find u, we use the direct ansatz

$$u(\boldsymbol{x}) = u^i(\boldsymbol{x}) - \frac{1}{4\pi} \int_\Gamma k(\boldsymbol{x}, \boldsymbol{y}) \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{y}) \mathrm{d}\sigma_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}. \tag{8.8}$$

Here, $k(\cdot, \cdot)$ denotes the fundamental solution to the Helmholtz equation, given by

$$k(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{4\pi} \frac{e^{i\kappa\|\boldsymbol{x}-\boldsymbol{y}\|}}{\|\boldsymbol{x} - \boldsymbol{y}\|}.$$

In accordance with [11], the unknown Neumann data $\frac{\partial u}{\partial \boldsymbol{n}}$ are obtained by solving the Fredholm boundary integral equation of the second kind (i.e., $q = 0$)

$$\left(\frac{1}{2} + \mathscr{D}^\intercal - i\eta \mathscr{S} \right) \frac{\partial u}{\partial \boldsymbol{n}} = \frac{\partial u^i}{\partial \boldsymbol{n}} - i\eta u^i \quad \text{on } \Gamma, \tag{8.9}$$

where $\mathscr{S}$ and $\mathscr{D}$ are the acoustic single and double layer operators, respectively,

$$(\mathscr{S}v)(\mathbf{x}) = \int_\Gamma k(\boldsymbol{x}, \boldsymbol{y}) v(\boldsymbol{y}) \mathrm{d}\sigma_{\boldsymbol{y}}, \qquad (\mathscr{D}v)(\mathbf{x}) = \int_\Gamma \frac{\partial}{\partial \boldsymbol{n}_{\boldsymbol{y}}} k(\boldsymbol{x}, \boldsymbol{y}) v(\boldsymbol{y}) \mathrm{d}\sigma_{\boldsymbol{y}},$$

and $\eta > 0$ is a parameter which is usually chosen proportional to κ, see [11] for example.

For our numerical computations, we will solve the scattering problem by using the boundary integral equation (8.9) for the wavenumbers $\kappa = 1, 2, 4$, and 8. For the adaptive algorithm, $\omega = 0.5$, $\theta = 0.9$, and $\gamma = 10^{-2}$ were chosen. The arising system of linear equations is solved by means of the GMRES method with diagonal scaling, where the approximate solution from the previous step is used as initial guess and the system is solved up to a relative precision 10^{-5} (we refer the reader to [39] for parameter studies and details of the current implementation).

Having the approximate solution at hand, we have given the Neumann data which can be used to evaluate $u(\boldsymbol{x})$ according to the ansatz (8.8). As scatterer Ω, we consider a drilled cube as seen in Figs. 4, 5, 6, 7. The incident wave is chosen to travel into the direction of $(1, 1, 0)$.

Figure 2 shows the convergence history of the estimated norm of the residual for each different value of κ. We observe a rate of convergence of approximately $N_{\mathrm{dof}}^{-0.5}$, independently of the chosen κ, while the norm of the residual increases as the wavenumber increases. A comparison of the rate of convergence in case of adaptive and uniform refinement is found in Fig. 3 for the wavenumber $\kappa = 8$. We observe approximately the rate of convergence $N_{\mathrm{dof}}^{-0.25}$ in case of uniform refinement, which is only half the rate as in case of adaptive refinement.

For the visualization of the solution, we compute the total field $u(\boldsymbol{x})$ and the scattered field $u^s(\boldsymbol{x})$ in the area $E = \{(x_1, x_2, x_3) : x_3 = 0 \text{ and } x_1, x_2 \in [-2.5, 2.5]\}$. This plane intersects the drilled cube, such that we can illustrate the

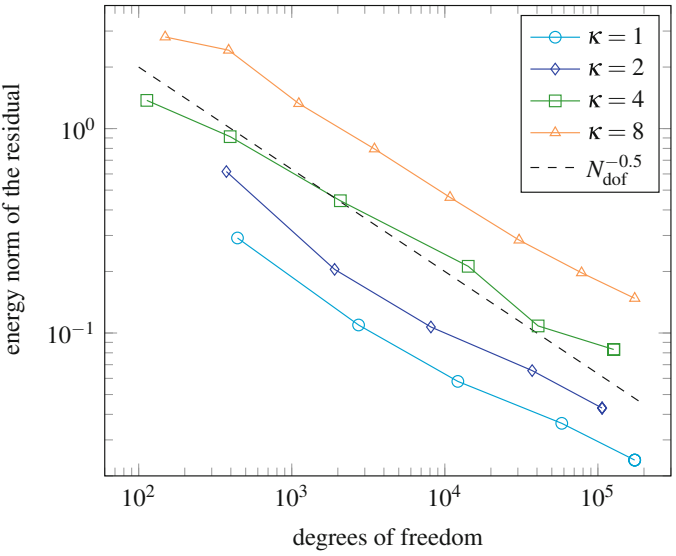


Fig. 2 Energy norm of the residual for the adaptive wavelet method in dependance of the wavenumbers

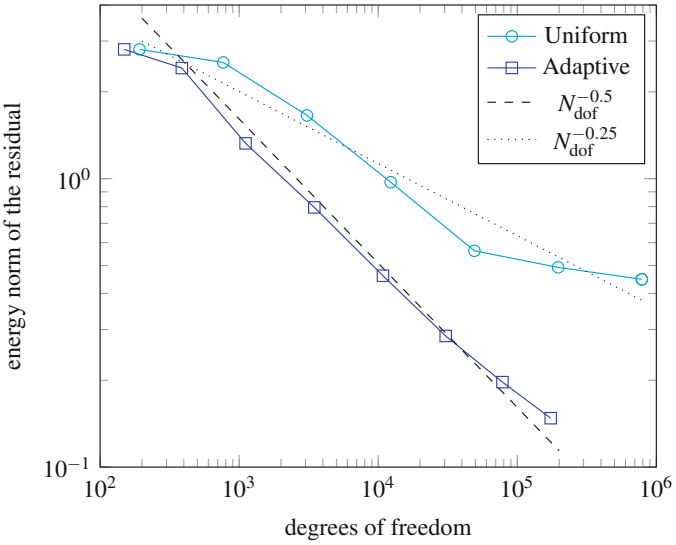


Fig. 3 Norm of the residual for adaptive refinement and for uniform refinement

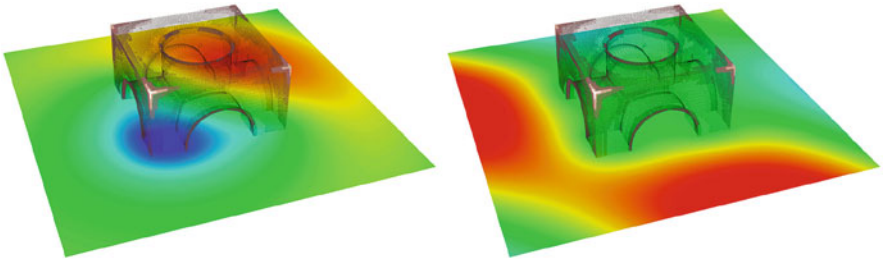


Fig. 4 Scattered field (left) and total field (right) for $\kappa = 1$

pattern which is produced by the scattered wave. In addition to the plane, where $u^s(\boldsymbol{x})$ and $u(\boldsymbol{x})$ are evaluated, we also draw the scatterer in the pictures. In particular, we draw the refinement of the scatterer's surface, where a cluster of wavelets appears in a darker colour first. By looking more closely at the corners and edges of the geometry, we see a lighter colouring, which again indicates even a stronger refinement.

In Fig. 4, we see the scene for $\kappa = 1$. The top corner of the square is the point $(-2.5, -2.5)$, which means that the harmonic wave is travelling upwards. In the left plot of Fig. 4, the scattered field $u^s(\boldsymbol{x})$ is seen and, in the right plot of Fig. 4, the total field $u(\boldsymbol{x})$ is seen. For $\kappa = 1$, we do not observe yet an interesting scattering pattern, as the wavenumber is too small. On the other hand, we already observe that the adaptive wavelet boundary element method refines towards the edges and the vertices of the geometry. This behaviour is expected, since we solve

the scattering problem for the direct formulation involving the Neumann data, which admit a singularity at the non-smooth parts of the geometry. In particular, we observe a refinement on the edges which are illuminated, i.e. the edges which face the incoming wave. On the edges which are at the back of the geometry refinement, the refinement is not that strong.

In Fig. 5, we visualize the scattered field and the total field for $\kappa = 2$. We observe that the wavenumber $\kappa = 2$ is still too small to have a noticeable scattering pattern. Also, we observe again the refinement along the illuminated edges in reference to the incoming wave.

Figure 6 contains the scattered field (left) and total field (right) for the wavenumber $\kappa = 4$. Here, we observe that the wavenumber is chosen just large enough, such that for the first time the wave can enter the inner part of the drilled cube. We observe again the refinement towards the edges and vertices facing the incoming wave, with less refinement in those parts of the geometry which lie on the back of the cube.

In Fig. 7, we draw the scattered field (left) and the total field (right) for the wavenumber $\kappa = 8$. This wavenumber is large enough in order to produce a beautiful scattering pattern. Especially, we see that the wave can travel through the inner part of the drilled cube. We observe again the refinement towards the edges and vertices with more refinement of those parts of the drilled cube which are illuminated by the incoming wave.

We conclude that the adaptive wavelet algorithm produces excellent results for all chosen wavenumbers κ. Adaptivity pays off especially for the direct ansatz, where

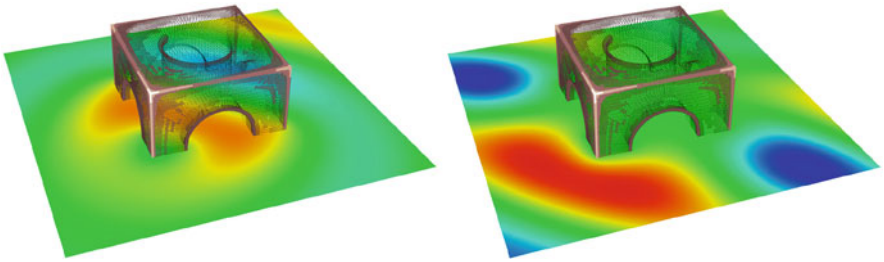


Fig. 5 Scattered field (left) and total field (right) for $\kappa = 2$

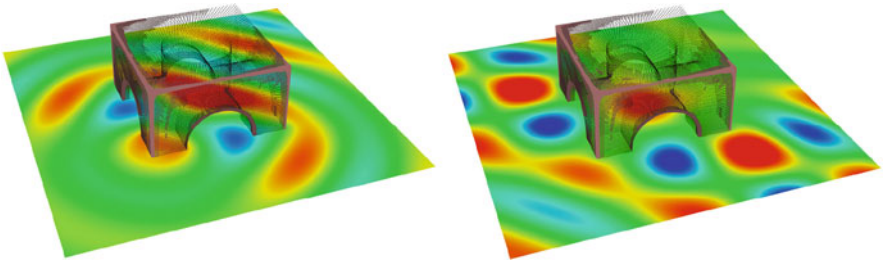


Fig. 6 Scattered field (left) and total field (right) for $\kappa = 4$

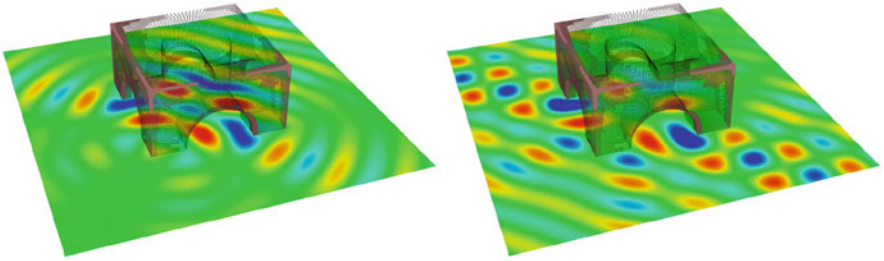


Fig. 7 Scattered field (left) and total field (right) for $\kappa = 8$

the refinement towards the edges and vertices can clearly be observed. To achieve a similar accuracy with uniform mesh refinement, one would need much more degrees of freedom. This would not only take more time to compute, but may not be feasible any more as far as memory consumption is concerned.

8.4.2 Laplace Equation Solved by the Single Layer Operator

Let us present numerical results in order to verify and quantify the goal-oriented adaptive wavelet boundary element method. To this end, consider the Laplace equation

$$\Delta U = 0 \text{ in } \Omega, \quad U = f \text{ on } \Gamma, \tag{8.10}$$

solved inside a bounded domain Ω with boundary $\Gamma = \partial\Omega$. We convert this problem to a boundary integral equation by making the ansatz

$$U = \int_\Gamma \frac{u(\boldsymbol{y})}{\|\cdot - \boldsymbol{y}\|} \mathrm{d}\sigma_{\boldsymbol{y}} \text{ in } \Omega \tag{8.11}$$

for the unknown density $u \in H^{-1/2}(\Gamma)$. Observing that this single layer potential is continuous across the boundary Γ, we arrive at the Fredholm integral equation of the first kind

$$\mathscr{A}u = \int_\Gamma \frac{u(\boldsymbol{y})}{\|\cdot - \boldsymbol{y}\|} \mathrm{d}\sigma_{\boldsymbol{y}} = f \text{ on } \Gamma$$

for the single layer operator $\mathscr{A} : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$. Given an evaluation point $\boldsymbol{x} \in \Omega$, the potential $U(\boldsymbol{x})$ is computed in accordance with (8.11) by

$$g(u) = \int_\Gamma \frac{u(\boldsymbol{y})}{\|\boldsymbol{x} - \boldsymbol{y}\|} \mathrm{d}\sigma_{\boldsymbol{y}}. \tag{8.12}$$

This *potential evaluation* corresponds to the application of a continuous linear functional to the density u. Notice that the integrand becomes weakly singular as $\boldsymbol{x} \in \Omega$ approaches the boundary Γ.

For the following computations, we consider the Fichera vertex $(0,1)^3 \setminus (0,0.5]^3$ as domain of interest. The restriction $f = p|_\Gamma$ of the harmonic polynomial $p(\boldsymbol{x}) = 4x_1^2 - 3x_2^2 - x_3^2$ is chosen as Dirichlet data for the primal problem, which implies that the solution U of the Laplace equation (8.10) coincides with the polynomial p. For given, fixed $\boldsymbol{x} \in \Omega$, we shall apply the goal-oriented adaptive wavelet boundary element method to evaluate the output functional $g(u)$, given by Eq. (8.12). After each iteration of the adaptive algorithm, we compute an approximation $g(u_\Lambda)$ via the scalar product $\boldsymbol{g}_\Lambda^\mathsf{T} \boldsymbol{u}_\Lambda$. We then evaluate the potential error $\| p(\boldsymbol{x}) - \boldsymbol{g}_\Lambda^\mathsf{T} \boldsymbol{u}_\Lambda \|$ with $p(\boldsymbol{x})$ being the analytic solution of the problem under consideration.

8.4.2.1 First Example

We choose the evaluation point for the functional (8.12) as $\boldsymbol{x} = (0.25, 0.25, 0.9)$. This point is located inside Fichera's vertex but close to the top boundary. Moreover, we have chosen the coarsening constant $\theta = 0.5$.

Figure 8 shows the convergence histories of the primal residual, the dual residual and the potential error of the goal-oriented adaptive wavelet boundary element method. We observe that the primal residual has a rate of convergence of $N_{\mathrm{dof}}^{-0.63}$. Whereas, we notice that the dual residual seems to have a rate of convergence of

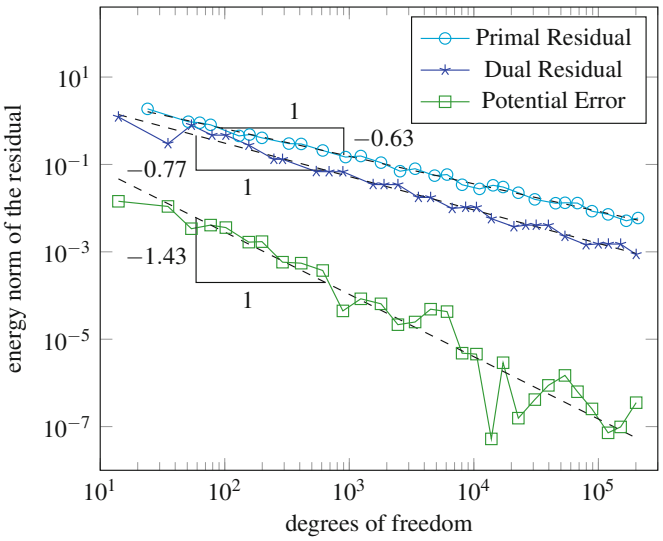


Fig. 8 Norm of the primal residual, norm of the dual residual and potential error versus the number of degrees of freedom in case of the evaluation point $\boldsymbol{x} = (0.25, 0.25, 0.9)$

approximately $N_{\mathrm{dof}}^{-0.75}$, which is significantly better than the rate of convergence for the primal residual. The potential error has a rate of convergence of approximately $N_{\mathrm{dof}}^{-1.43}$, which indeed coincides with the sum of the rates of convergence for the primal and the dual residual.

In Fig. 9, we plot the ratios of the primal residual, the dual residual, and the potential error versus the computation time. It turns out that the computational complexity of the current implementation does not scale linearly but much better than quadratically. Therefore, it does clearly pay off to employ a fast boundary element method. Especially, to compute the solution to a boundary integral equation with more than 200,000 degrees of freedom would have not been possible without matrix compression.

We should also visualize the refinement which is produced by the adaptive algorithm. In Fig. 10, we have visualized the refinement. Since we would not be able to see the refinement by drawing the grid, this picture was produced in the following way: After the code terminated, we assigned to each active wavelet a point in the center of its support which is weighted with 2 to the power of the wavelet's level, achieving that a small wavelet gets assigned a large value. The picture below is thus to be interpreted as: The lighter the colour, the finer are the elements in this area. We observe in Fig. 10 that the mesh refinement takes place on the top of the Fichera vertex, near from where the point $\boldsymbol{x} = (0.25, 0.25, 0.9)$ is located. Also, the algorithm refines towards the edges and vertices of Fichera's vertex.

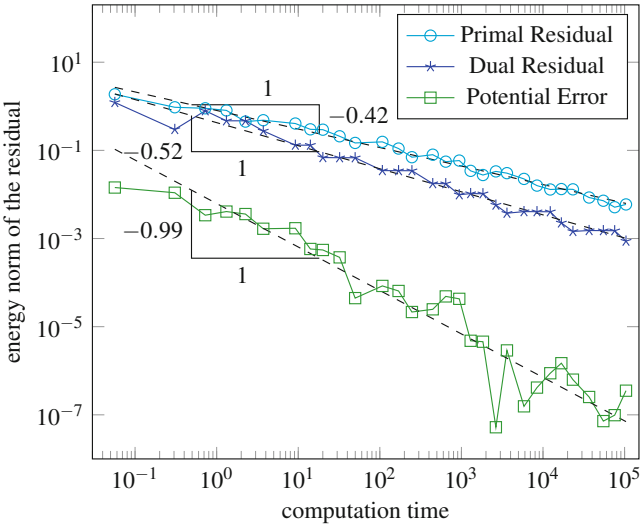


Fig. 9 Norm of the primal residual, norm of the dual residual, and potential error versus computation time in case of the evaluation point $\boldsymbol{x} = (0.25, 0.25, 0.9)$

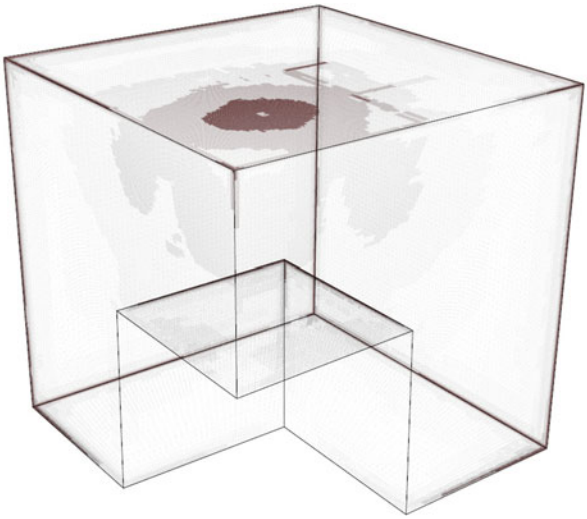


Fig. 10 Adaptive mesh refinement in case of the evaluation point (0.25, 0.25, 0.9)

8.4.2.2 Second Example

For the second example, we move the evaluation point more closely to the boundary, namely we set $\boldsymbol{x} = (0.25, 0.25, 0.95)$, and perform our computations again.

In Fig. 11, we visualize the convergence histories for the primal residual, the dual residual, and the potential error versus the number of degrees of freedom in a log-log scale. We observe that the primal residual and the dual residual show a rate of convergence of $N_{\mathrm{dof}}^{-0.63}$ and $N_{\mathrm{dof}}^{-0.76}$. For the potential error, we observe a rate of convergence of $N_{\mathrm{dof}}^{-1.34}$. This is slightly less than the rate of convergence of the potential error for the evaluation point $\boldsymbol{x} = (0.25, 0.25, 0.9)$. The computing time of the adaptive algorithm does not scale linearly in the number N of degrees of freedom, compare Fig. 12, where the accuracy versus computing times is plotted and the rates are slightly worse than those found in Fig. 11. Nonetheless, we like to repeat that the scaling is much better than a quadratic scaling, which is required by the traditional boundary element method.

To conclude our tests, we finally visualize the mesh refinement produced by the adaptive algorithm in Fig. 13. It is again refined towards the edges and vertices of the geometry and towards the point $\boldsymbol{x} = (0.25, 0.25, 0.95)$. If we compare the refinement on the top of the domain with the refinement from the previous example, we notice that the refinement is slightly more localized here, cf. Fig. 10.

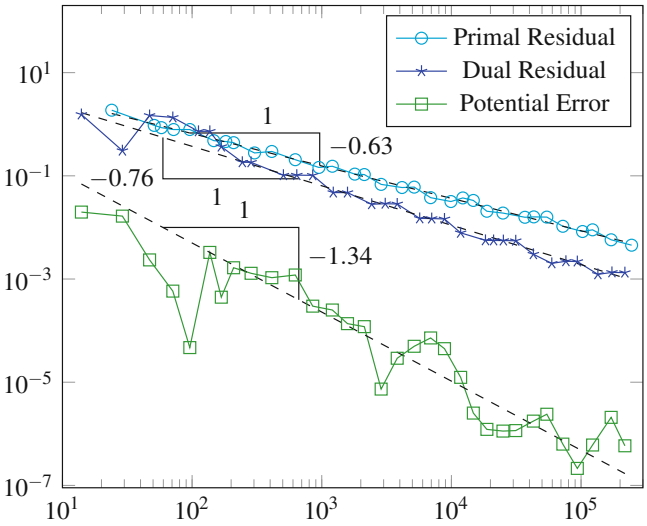


Fig. 11 Norm of the primal residual, norm of the dual residual, and potential error versus the number of degrees of freedom in case of the evaluation point $\boldsymbol{x} = (0.25, 0.25, 0.95)$

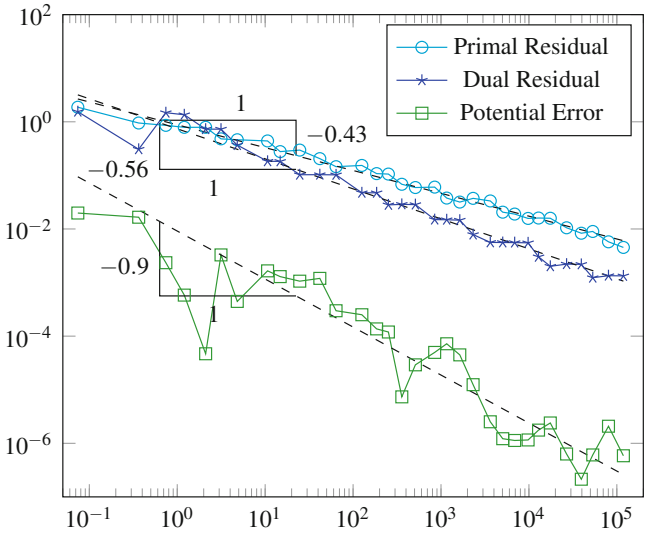


Fig. 12 Norm of the primal residual, norm of the dual residual, and potential error versus computation time in case of the evaluation point $\boldsymbol{x} = (0.25, 0.25, 0.95)$

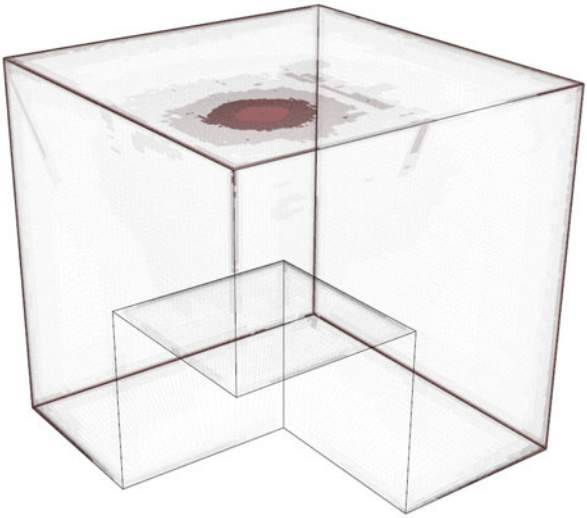


Fig. 13 Adaptive mesh refinement in case of the evaluation point (0.25, 0.25, 0.95)

8.5 Conclusion

In the present article, we presented the adaptive wavelet boundary element method for the rapid solution of boundary integral equations. A goal-oriented strategy for the evaluation of linear output functionals has been proposed as well. The algorithms have been validated and quantified by numerical examples for an acoustic scattering problem and for the point evaluation of the potential in case of the single layer operator for the Laplace equation on Fichera's vertex.

Acknowledgements This research has been supported by the Swiss National Science Foundation (SNSF) through the DACH-project "BIOTOP: Adaptive Wavelet and Frame Techniques for Acoustic BEM".

References

1. Bakry, H.: A goal-oriented a posteriori error estimate for the oscillating single layer integral equation. Appl. Math. Lett. **69**, 133–137 (2017)
2. Bangerth, W., Rannacher, R.: Adaptive finite element method for differential equations. Lectures Math. ETH Zürich, Birkhäuser (2003)
3. Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. **86**, 565–589 (2000)
4. Bebendorf, M., Rjasanow, S.: Adaptive low-rank approximation of collocation matrices. Computing **70**, 1–24 (2003)

5. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
6. Becker, R., Estecahandy, E., Trujillo, D.: Weighted marking for goal-oriented adaptive finite element methods. SIAM J. Numer. Anal. **49**(6), 2451–2469 (2011)
7. Beylkin, G., Coifman, R., Rokhlin, V.: The fast wavelet transform and numerical algorithms. Commun. Pure Appl. Math. **44**, 141–183 (1991)
8. Binev, P., DeVore, R.: Fast computation in adaptive tree approximation. Numer. Math. **97**, 193–217 (2004)
9. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations. Convergence rates. Math. Comput. **70**, 27–75 (2001)
10. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods II. Beyond the elliptic case. Found. Comput. Math. **2**, 203–245 (2002)
11. Colton, D., Kress, R.: Integral equation methods in scattering theory. Wiley, New York (1983)
12. Dahmen, W., Schneider, R.: Composite wavelet bases for operator equations. Math. Comput. **68**, 1533–1567 (1999)
13. Dahmen, W., Schneider, R.: Wavelets on manifolds I. Construction and domain decomposition. Math. Anal. **31**, 184–230 (1999)
14. Dahmen, W., Harbrecht, H., Schneider, R.: Compression techniques for boundary integral equations. Asymptotically optimal complexity estimates. SIAM J. Numer. Anal. **43**(6), 2251–2271 (2006)
15. Dahmen, W., Kunoth, A., Vorloeper, J.: Convergence of adaptive wavelet methods for goal-oriented error estimation. In: Bermudez de Castro, A., Gomez, D., Quintely, P., Salgado, P. (eds.) Numerical Mathematics and Advanced Applications, pp. 39–61. Springer, Berlin (2006)
16. Dahmen, W., Harbrecht, H., Schneider, R.: Adaptive methods for boundary integral equations. Complexity and convergence estimates. Math. Comput. **76**, 1243–1274 (2007)
17. DeVore, R.A.: Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
18. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numer. **4**, 105–158 (1995)
19. Faermann, B.: Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary element methods. Part II: the three-dimensional case. Numer. Math. **92**(3), 467–499 (2002)
20. Feischl, M., Karkulik, M., Melenk, J.M., Praetorius, D.: Quasi-optimal convergence rate for an adaptive boundary element method. SIAM J. Numer. Anal. **51**(2), 1327–1348 (2013)
21. Feischl, M., Gantner, G., Haberl, A., Praetorius, D., Führer, T.: Adaptive boundary element methods for optimal convergence of point errors. Numer. Math. **132**(3), 541–567 (2016)
22. Feischl, M., Praetorius, D., van der Zee, K.G.: An abstract analysis of optimal goal-oriented adaptivity. SIAM J. Numer. Anal. **54**(3), 1423–1448 (2016)
23. Gantumur, T.: An optimal adaptive wavelet method for nonsymmetric and indefinite elliptic problems. J. Comput. Appl. Math. **211**(1), 90–102 (2008)
24. Gantumur, T.: Adaptive boundary element methods with convergence rates. Numer. Math. **124**, 471–516 (2013)
25. Gantumur, T., Stevenson, R.: Computation of singular integral operators in wavelet coordinates. Computing **76**, 77–107 (2006)
26. Gantumur, T., Harbrecht, H., Stevenson, R.: An optimal adaptive wavelet method for elliptic equations without coarsening. Math. Comput. **76**, 615–629 (2007)
27. Greengard, L., Rokhlin, V.: A fast algorithm for particle simulation. J. Comput. Phys. **73**, 325–348 (1987)
28. Hackbusch, W.: A sparse matrix arithmetic based on $\mathscr{H}$-matrices. Part I: Introduction to $\mathscr{H}$-matrices. Computing **64**, 89–108 (1999)
29. Hackbusch, W., Nowak, Z.P.: On the fast matrix multiplication in the boundary element method by panel clustering. Numer. Math. **54**, 463–491 (1989)
30. Harbrecht, H., Schneider, R.: Biorthogonal wavelet bases for the boundary element method. Math. Nachr. **269–270**, 167–188 (2004)

31. Harbrecht, H., Schneider, R.: Wavelet Galerkin schemes for boundary integral equations. Implementation and quadrature. SIAM J. Sci. Comput. **27**(4), 1347–1370 (2006)
32. Harbrecht, H., Stevenson, R.: Wavelets with patchwise cancellation properties. Math. Comput. **75**(256), 1871–1889 (2006)
33. Harbrecht, H., Utzinger, M.: On adaptive wavelet boundary element methods. J. Comput. Math. **36**(1), 90–109 (2018)
34. Hsiao, G.C., Wendland, W.L.: Boundary Integral Equations. Applied Mathematical Sciences, vol. 164. Springer, Berlin (2008)
35. Huybrechs, D., Simoens, J., Vandevalle, S.: A note on wave number dependence of wavelet matrix compression for integral equations with oscillatory kernel. J. Comput. Appl. Math. **172**, 233–246 (2004)
36. Mommer, M.S., Stevenson, R.P.: A goal-oriented adaptive finite element method with convergence rates. SIAM J. Numer. Anal. **47**(2), 861–886 (2009)
37. Rokhlin, V.: Rapid solution of integral equations of classical potential theory. J. Comput. Phys. **60**, 187–207 (1985)
38. Tyrtyshnikov, E.E.: Mosaic skeleton approximation. Calcolo **33**, 47–57 (1996)
39. Utzinger, M.: An Adaptive Wavelet Method for the Solution of Boundary Integral Equations in Three Dimensions. PhD thesis, Universität Basel, Switzerland (2016)

Chapter 9
Comparison Analysis of Two Numerical Methods for Fractional Diffusion Problems Based on the Best Rational Approximations of t^{γ} on [0, 1]

Stanislav Harizanov, Raytcho Lazarov, Svetozar Margenov, Pencho Marinov, and Joseph Pasciak

Abstract The paper is devoted to the numerical solution of algebraic systems of the type $\mathbb{A}^{\alpha}\mathbf{u} = \mathbf{f}$, $0 < \alpha < 1$, where $\mathbb{A}$ is a symmetric and positive definite matrix. We assume that $\mathbb{A}$ is obtained from finite difference or finite element approximations of second order elliptic problems in $\mathbb{R}^d$, $d = 1, 2$ and we have an optimal method for solving linear systems with matrices $\mathbb{A} + c\mathbb{I}$. We study and compare experimentally two methods based on best uniform rational approximation (BURA) of t^{γ} on $[0, 1]$ with the method of Bonito and Pasciak, (Math Comput 84(295):2083–2110, 2015), that uses exponentially convergent quadratures for the Dunford-Taylor integral representation of the fractional powers of elliptic operators. The first method, introduced in Harizanov et al. (Numer Linear Algebra Appl 25(4):115–128, 2018) and based on the BURA $r_{\alpha}(t)$ of $t^{1-\alpha}$ on $[0, 1]$, is used to get the BURA of $t^{-\alpha}$ on $[1, \infty)$ through $t^{-1}r_{\alpha}(t)$. The second method, developed in this paper and denoted by R-BURA, is based on the BURA $r_{1-\alpha}(t)$ of t^{α} on $[0, 1]$ that approximates $t^{-\alpha}$ on $[1, \infty)$ via $r_{1-\alpha}^{-1}(t)$. Comprehensive numerical experiments on some model problems are used to compare the efficiency of these three algorithms depending on α. The numerical results show that R-BURA method performs well for α close to 1 in contrast to BURA, which performs well for α close to 0. Thus, the two BURA methods have mutually complementary advantages.

S. Harizanov · S. Margenov (✉) · P. Marinov
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
e-mail: margenov@parallel.bas.bg

R. Lazarov · J. Pasciak
Department of Mathematics, Texas A&M University, College Station, TX, USA

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_9

9.1 Introduction

9.1.1 Algebraic Problems Under Consideration

Let $\mathbb{R}^N$, N positive integer, be a real N-dimensional vector space with the standard ℓ_2-inner product, $\mathbf{u}^T\mathbf{v}$, for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, and let $\mathbb{A}$ be an $N \times N$ symmetric and positive definite matrix with eigenvalues and eigenvectors $\{(\lambda_i, \boldsymbol{\Psi}_i)\}_{i=1}^N$. We assume that the eigenvectors are orthonormal, that is $\boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_j = \delta_{ij}$, and $0 < \lambda_1 \leq \lambda_2 \leq \ldots \lambda_N$.

For $0 < \alpha < 1$ and given $\mathbf{f} \in \mathbb{R}^N$ we consider the following algebraic problem:

$$\text{find} \quad \mathbf{u} \in \mathbb{R}^N \quad \text{such that} \quad \mathbb{A}^\alpha \mathbf{u} = \mathbf{f} \tag{9.1}$$

where the fractional power $\mathbb{A}^\alpha$ is defined through the eigenvalues and eigenvectors of $\mathbb{A}$

$$\mathbb{A}^\alpha = \mathbb{W}\mathbb{D}^\alpha \mathbb{W}^T, \quad \text{where} \quad \mathbb{A} = \mathbb{W}\mathbb{D}\mathbb{W}^T.$$

Here $\mathbb{W}, \mathbb{D} \in \mathbb{R}^{N\times N}$ are defined as $\mathbb{W} = [\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T, \ldots, \boldsymbol{\Psi}_N^T]$ and $\mathbb{D} = diag(\lambda_1, \ldots, \lambda_N)$. Then $\mathbb{A}^{-\alpha} = \mathbb{W}\mathbb{D}^{-\alpha}\mathbb{W}^T$ and the solution of $\mathbb{A}^\alpha \mathbf{u} = \mathbf{f}$ can be expressed as

$$\mathbf{u} = \mathbb{A}^{-\alpha}\mathbf{f} = \mathbb{W}\mathbb{D}^{-\alpha}\mathbb{W}^T\mathbf{f}. \tag{9.2}$$

For any $\beta \in \mathbb{R}$ we have $\|\mathbf{u}\|_{\mathbb{A}^{\beta+\alpha}} = \|\mathbf{f}\|_{\mathbb{A}^{\beta-\alpha}}$, where $\|\mathbf{u}\|^2_{\mathbb{A}^\gamma} = \mathbf{u}^T\mathbb{A}^\gamma\mathbf{u}$, $\gamma \in \mathbb{R}$.

9.1.2 Motivation and Connection to the Spectral Laplacian

The algebraic problem is motivated by the following boundary value problem: for a given $\mathbf{f} \in L^2(\Omega)$ find $u(x)$ such that

$$(-\Delta)^\alpha u = f, \quad x \in \Omega = (0,1)\times(0,1), \quad u = 0, \quad x \in \partial\Omega. \tag{9.3}$$

The fractional power of Laplacian, $(-\Delta)^\alpha$, is defined through its eigenvalues λ_k and eigenfunctions ψ_k of $-\Delta$ with homogeneous Dirichlet boundary conditions, namely, $\psi_k \in H_0^1(\Omega)$, $(\nabla\psi_k, \nabla v) = \lambda_k(\psi_k, v)$ for all $v \in H_0^1(\Omega)$. Then

$$(-\Delta)^\alpha v := \sum_{k=1}^{\infty} \lambda_k^\alpha (v, \psi_k)\psi_k, \quad \forall v \in D((-\Delta)^\alpha) := \sum_{k=1}^{\infty} \lambda_j^{2\alpha} |(v, \psi_k)|^2 < \infty\,.$$

Remark 9.1 Clearly, for small α the Dirichlet boundary conditions could not make sense. In general, for $f(x) \in H^s(\Omega)$, $0 < s < 1/2$ we have $u \in H^{s+2\alpha}(\Omega)$ (for

more details we refer to [2, Subsection 4.1]). Thus, the trace on the boundary exists for $s + 2\alpha > 1/2$. For example, $f(x) \equiv 1$ has regularity $H^{1/2-\epsilon}(\Omega)$ for any $0 < \epsilon$. Thus, the homogeneous Dirichlet boundary condition makes sense for any positive α. However, if f is just in $L^2(\Omega)$, then the solution u has trace for $\alpha > 1/4$ and the Dirichlet boundary condition is well defined under this condition.

9.1.3 Examples of SPD Matrices Under Consideration

9.1.3.1 Example 1

The first example of such a matrix is $\mathbb{A} \in \mathbb{R}^{N\times N}$, $N = n^2$, that has the following block structure (here $\mathbb{A}_{i,i} \in \mathbb{R}^{n\times n}$, $i = 1, \cdots, n$ and $\mathbb{I}_n$ is the identity matrix in $\mathbb{R}^n$)

$$\mathbb{A} = (n+1)^2 \begin{bmatrix} \mathbb{A}_{1,1} & -\mathbb{I}_n & & & & \\ -\mathbb{I}_n & \mathbb{A}_{2,2} & -\mathbb{I}_n & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ & -\mathbb{I}_n & \mathbb{A}_{i,i} & -\mathbb{I}_n & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ & & & -\mathbb{I}_n & \mathbb{A}_{n,n} & \end{bmatrix}, \quad \mathbb{A}_{i,i} = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ & -1 & 4 & -1 & \\ & & -1 & 4 & \end{bmatrix}. \tag{9.4}$$

This matrix is generated by the finite difference approximation of the following boundary value problem

$$-\Delta u = f, \quad x \in \Omega = (0,1) \times (0,1), \quad u = 0, \quad x \in \partial\Omega \tag{9.5}$$

on a uniform mesh with mesh-size $h = 1/(n+1)$. We are not aware of a rigorous analysis whether the system $\mathbb{A}^\alpha \mathbf{u} = \mathbf{f}$ approximates the problem (9.3). However, we expect this would be the case since, as discussed below in details, the matrix $\mathbb{A}$ can be obtained by the Galerkin approximation of the boundary value problem (9.3).

9.1.3.2 Example 2

We partition $\Omega = (0,1) \times (0,1)$ into squares of size $h = 1/(n+1)$. Let $\mathcal{T}_h$ be obtained by subdividing each square into two triangles by connecting the upper left corner with the lower right corner. On this triangulation we introduce the space $V_h \subset H_0^1(\Omega)$ of continuous piece-wise linear functions. The finite element approximation of (9.5) is: find $u_h \in V_h$ such that

$$a(u_h, v) := \int_\Omega \nabla u_h(x) \cdot \nabla v(x) dx = (f, v) := (\pi_h f, v) \ \ \forall v \in V_h. \tag{9.6}$$

Here $(\cdot,\cdot)$ is the standard L_2-inner product on V_h. We define the operator $\mathrm{A}: V_h \to V_h$ by $(\mathrm{A}u, v) = (z, v)$ for all $v \in V_h$. Then $\mathrm{A}y = z \in V_h$ should have a representation through the nodal basis ψ_k: $z = \sum c_k \psi_k$ and then the operator A is expressed through the global "stiffness" matrix $\{\mathbb{A}\}_{i,k} = a(\psi_i, \psi_k)$ and the global "mass" matrix $\mathbb{M} = \{(\psi_i, \psi_k)\}_{i,k}$ via the relation $\mathrm{A} = \mathbb{M}^{-1}\mathbb{A}$. The matrix $\mathbb{M}$ is not diagonal and has similar sparsity pattern as the "stiffness" matrix $\mathbb{A}$. The algebraic problem

$$\mathrm{A}^{\alpha} u_h = \pi_h f \tag{9.7}$$

is stable and the solution should have a representation $u_h(x) = \sum_{i=1}^{N} u_i \psi_i(x)$. Then $\mathbf{u} = \{u_1, \cdots, u_N\}^T$ approximates $u(x)$ of (9.3) in the nodes of the mesh, see, [2].

In order to get the matrix $\mathbb{A}$ defined by (9.4) (instead of A) we can apply the following approach. First, we introduce the "lumped" mass inner product in V_h, [15, pp. 239–242]. Namely, for $z, v \in V_h$ we define

$$(z, v)_h = \frac{1}{3} \sum_{\tau \in \mathscr{T}_h} \sum_{i=1}^{3} |\tau|\, z(P_i) v(P_i),$$

where P_1, P_2, P_3 are the vertices of the triangle τ and $|\tau|$ is its area. Then the lumped "mass" matrix, defined as

$$\mathbb{M}_h = \{(\psi_i, \psi_k)_h\}_{i,k},$$

is diagonal. Moreover, since the mesh is square, all diagonal elements of $\mathbb{M}_h^{-1}$ are equal to $h^{-2} = (n+1)^2$ and $\mathrm{A}: V_h \to V_h$ is defined by

$$(\mathrm{A}u_h, v)_h = a(u_h, v) \quad \text{gives} \quad \mathrm{A} = \mathbb{M}_h^{-1}\mathbb{A}.$$

Since on a uniform mesh $(\cdot,\cdot)_h$ approximates $(\cdot,\cdot)$ with second order, one may expect that (9.7) approximates the problem (9.3).

Remark 9.2 One can generate matrices with similar structure when solving elliptic problems with Neumann or Robin boundary conditions.

9.1.4 The Concept of the Best Uniform Rational Approximation (BURA)

We shall use the following notation for a class of rational functions:

$$\mathscr{R}(k, m) = \{r(t) : r(t) = P_k(t)/Q_m(t), \text{ where } P_k \in \mathscr{P}_k, \text{ and } Q_m \in \mathscr{P}_m, \text{ monic}\}$$

with $\mathscr{P}_j$ being the set of algebraic polynomials of degree j. The best uniform approximation $r_\alpha(t) \in \mathscr{R}(k,m)$ of $t^{1-\alpha}$ on $[0,1]$ (called further (k,m)-BURA), and its approximation error $E_{\alpha,k,m}$ is defined as follows:

$$r_\alpha(t) := \operatorname*{argmin}_{r\in\mathscr{R}(k,m)} \left\| t^{1-\alpha} - r(t) \right\|_{L^\infty(0,1)}, \quad \varepsilon(t) = r_{1-\alpha}(t) - t^\alpha, \quad E_{\alpha,k,m} := \|\varepsilon(t)\|_{L^\infty(0,1)}.$$

For $m = k$ the existence and uniqueness of $r_\alpha(t)$ is well known, e.g., [11, Chapters 9.1 and 9.2]. Moreover, it is known that both the numerator and the denominator of the minimizer are of exact degree k and the error function $\varepsilon(t)$ possesses $2k+2$ extreme points in $[0,1]$, including the endpoints of the interval.

9.1.5 A Review of Methods and Equations Involving Functions of Matrices

The formula (9.2) could be used in practical computations if the eigenvectors and eigenvalues are explicitly known and Fast Fourier Transform is applicable to perform the matrix vector multiplication with $\mathbb{W}$, thus leading to almost optimal computational complexity, $O(N \log N)$. However, this approach is limited to separable problems with constant coefficients in simple domains and boundary conditions.

This work is related also to the more difficult problem of stable computations of the matrix square root and other functions of matrices, see, e.g. the earlier papers [3, 7, 10], as well as, [4] for some more recent related results. However, in this paper we do not deal with an evaluation of $\mathbb{A}^\alpha$, instead we discuss efficient methods for solving the algebraic system $\mathbb{A}^\alpha \mathbf{u} = \mathbf{f}$, where $\mathbb{A}$ is an SPD matrix generated by an approximation of a second order elliptic operators.

Our research is also connected with the work done in [8, 9], where numerical approximations of a fractional-in-space diffusion equation are considered. In [9], the proposed solver relies on Lanczos method. First, the adaptively preconditioned thick restart Lanczos procedure is applied to a system with $\mathbb{A}$. The gathered spectral information is then used to solve the system with $\mathbb{A}^\alpha$. In [3] an extended Krylov subspace method is proposed. The subspace $K^{k,m}(A,\phi) = span\{A^{-k+1}\phi, \ldots, A^{-1}\phi, \phi, \ldots, A^{m-1}\phi\}$, $m \geq 1, k \geq 1$, is originating by actions of the SPD matrix and its inverse. It is shown that for the same approximation quality, the variant of using the extended subspaces requires about the square root of the dimension of the standard Krylov subspaces using only positive or negative matrix powers. A drawback of this method is the memory required to store the full dense matrix $\mathbb{W}$, and the substantial deterioration of the convergence rate for ill-conditioned matrices. The advantage of the approach discussed in this paper is the robustness and almost optimal computational complexity.

9.1.6 The Main Contributions and Paper Content

We investigate two approaches for an approximate solving of $\mathbb{A}^{-\alpha}\mathbf{f}$ that are based on the best uniform rational approximation (BURA) $r(t)$ of t^{γ}, $\gamma > 0$, on $[0, 1]$. One subclass of such approximations is expressed through the diagonal Walsh table $P_k(t)/Q_k(t)$, i.e. $r \in \mathscr{R}(k, k)$, see, e.g. [14, 16]. Another subclass is the upper diagonal $P_{k+1}(t)/Q_k(t)$, i.e., $r \in \mathscr{R}(k+1, k)$. The first method is introduced in [6], where the BURA $r_\alpha(t)$ of $t^{1-\alpha}$ on $[0, 1]$ introduces a rational approximation of $t^{-\alpha}$ in the form $t^{-1}r_\alpha(t)$ on $[1, \infty)$. Here we develop a new method, denoted by R-BURA, where the best uniform rational approximation $r_{1-\alpha}(t)$ of t^{α} on $[0, 1]$ is used to approximate $t^{-\alpha}$ by $1/r_{1-\alpha}(t)$ on $[1, \infty)$. Both methods reduce solving (9.1) to a number of equations $(\mathbb{A} + c\mathbb{I})\mathbf{u} = \mathbf{F}$.

Our comparative analysis includes also the method proposed in [2] that is based on approximation of the integral representation of the solution of (9.3). Then exponentially convergent quadrature formulae are applied to evaluate numerically the related integrals. In fact, this Q-method leads to a rational approximation as well. An approximation of the boundary value problem with checkerboard right hand side, introduced in [2], is used in the numerical tests of our comparative analysis.

The rest of the paper is organized as follows. In Sect. 9.2 we introduce the basic properties of the solution methods and algorithms used in the paper. The analysis includes error estimates of the BURA, [6], the new R-BURA, and the Q-method of Bonito and Pasciak, [2]. Section 9.3 contains numerical tests for fractional Laplace problems. In the case of the BURA and R-BURA solvers, the impact of scaling is analyzed and experimentally confirmed. Among others, the numerical results complete the proof of concept of the new R-BURA approach, illustrating its advantages in the case of $\alpha \in (1/2, 1)$.

9.2 Description of the Numerical Methods Based on the BURA

9.2.1 The BURA and R-BURA Methods

The BURA Method

In this paper we consider two BURA subclasses (k, k) and $(k+1, k)$, introduced in Sect. 9.1.6. Let $\Lambda := \|\mathbb{A}\|_\infty = \max_{1\le i\le N}\sum_{j=1}^{N}|a_{ij}|$. Following [6], we obtain the rescaled SPD matrix $\mathscr{A} := \Lambda^{-1}\mathbb{A}$ with spectrum in $(0, 1]$. Then the original problem $\mathbb{A}^{\alpha}\mathbf{u} = \mathbf{f}$ can be rewritten as $\mathscr{A}^{\alpha}\mathbf{u} = \Lambda^{-\alpha}\mathbf{f}$. Note that the eigenvalues of $\mathscr{A}$ are $\mu_i := \Lambda^{-1}\lambda_i$, $0 < \mu_i \le 1$, $i = 1, \cdots, N$.

Let $r_\alpha(t)$ be the BURA of $t^{1-\alpha}$ on $[0,1]$ in $\mathscr{R}(k,k)$ or $\mathscr{R}(k+1,k)$. Then

$$\mathbf{u}_r := \Lambda^{-\alpha}\mathscr{A}^{-1}r_\alpha(\mathscr{A})\mathbf{f}, \tag{9.8}$$

are called (k,k)-BURA and $(k+1,k)$-BURA approximation of $\mathbf{u}$, respectively.

Using the spectral decomposition of $\mathbb{A}$, we can derive the following estimation of the BURA error:

$$\frac{\|\mathbf{u}_r-\mathbf{u}\|_2}{\|\mathbf{f}\|_2} = \Lambda^{-\alpha}\max_{\mu_i}\frac{\left|r_\alpha(\mu_i)-\mu_i^{1-\alpha}\right|}{\mu_i} \le \frac{\Lambda^{1-\alpha}E_{\alpha,k,m}}{\lambda_1}. \tag{9.9}$$

Then using [13, Theorem 1] (about the behavior of $E_{\alpha,k.k}$ as $k\to\infty$) we get the following property of the (k,k)-BURA:

$$\lim_{k\to\infty} e^{2\pi\sqrt{(1-\alpha)k}}\|\mathbf{u}_r-\mathbf{u}\|_2 = \frac{4^{2-\alpha}\Lambda^{1-\alpha}}{\lambda_1}\sin(\pi\alpha)\|\mathbf{f}\|_2. \tag{9.10}$$

Since $E_{\alpha,k,k} \ge E_{\alpha,k+1,k} \ge E_{\alpha,k+1,k+1}$ by definition, (9.10) is valid for $(k+1,k)$-BURA, as well.

Below we restricted our experiments to the (k,k)-BURA method. The implementation uses the decomposition of the rational function $t^{-1}r_\alpha(t)$ into a sum of partial fractions

$$\mathbf{u}_r = \sum_{j=0}^{k} c_j(\mathscr{A}-d_j\mathscr{I})^{-1}\mathbf{f} = \Lambda\sum_{j=0}^{k} c_j(\mathbb{A}-\Lambda d_j\mathscr{I})^{-1}\mathbf{f}.$$

Here $0=d_0>d_1\cdots>d_k$ are the poles of $r_\alpha(t)$ plus the additional pole at zero, and $c_j>0$ for every j (see [12] for more details). Obviously, the approximation $\mathbf{u}_r$ is obtained by solving $k+1$ linear systems with nonnegative diagonal shifts of $\mathbb{A}$.

The R-BURA Method

In this approach, we approximate $t^{-\alpha}$ on $[1,\infty)$ by $r_{1-\alpha}^{-1}(t)$, where $r_{1-\alpha}(t)$ is the best rational approximation of t^α on $[0,1]$ in $\mathscr{R}(k,k)$ or $\mathscr{R}(k+1,k)$. where the approximation was by $t^{-1}r_\alpha(t)$. Then

$$\mathbf{u}_r := \Lambda^{-\alpha}r_{1-\alpha}^{-1}(\mathscr{A})\mathbf{f} \tag{9.11}$$

are called the (k,k)-R-BURA and $(k+1,k)$-R-BURA approximation of $\mathbf{u}$, respectively.

9.2.2 Analysis of BURA and R-BURA Methods

For the analysis of the BURA-method we shall need the following properties of $r_\alpha(t)$:

Lemma 9.1 *Let $\alpha \in (0,1)$ and k be a positive integer. Then the best rational approximation (BURA) $r_\alpha(t) \in \mathscr{R}(k,k)$ of $t^{1-\alpha}$ in $[0,1]$ has the following properties:*

(a) $r_\alpha(t)$ is strictly monotonically increasing concave function when $t \in [0,1]$;
(b) $r_\alpha(0) = E_{\alpha,k,k}$.

Proof The second part follows directly from [12, Lemma 2.1], where it is shown that $\eta_1 = 0$ is an extreme point for $t^{1-\alpha} - r_\alpha(t)$ with negative value. The same lemma states that all the k zeros and k poles (denoted by d_j) of r_α are real, pairwise different, non-positive, and interlacing. Thus, for the decomposition of $r_\alpha(t)$ into partial fractions

$$r_\alpha(t) = b_0^* + \sum_{j=1}^{k} \frac{c_j^*}{t - d_j}$$

we have $c_j^*, d_j < 0$, $j = 1, \cdots, m$, for more details, see, e.g. [5, Theorem 1]. Hence,

$$r'_\alpha(t) = \sum_{j=1}^{k} \frac{-c_j^*}{(t-d_j)^2} > 0, \qquad r''_\alpha(t) = \sum_{j=1}^{k} \frac{2c_j^*}{(t-d_j)^3} < 0, \qquad \forall t \in [0,1].$$

The proof is completed.

Applying Lemma 9.1, the (k,k)-R-BURA approximation error is estimated analogously to the BURA error:

$$\frac{\|\mathbf{u}_r - \mathbf{u}\|_2}{\|\mathbf{f}\|_2} \le \Lambda^{-\alpha} \max_{\mu_i} \frac{\left|r_{1-\alpha}(\mu_i) - \mu_i^\alpha\right|}{\mu_i^\alpha r_{1-\alpha}(\mu_i)} \le \frac{\Lambda^{-\alpha} E_{1-\alpha,k,k}}{\mu_1^\alpha r_{1-\alpha}(\mu_1)} \le \frac{E_{1-\alpha,k,k}}{\lambda_1^\alpha r_{1-\alpha}(\mu_1)}.$$

Note that, $r_{1-\alpha}(t)$ has no zeros inside the interval $[0,1]$, therefore the denominator is strictly positive and the error bound is well-defined. On the other hand, $r_{1-\alpha}(0) = E_{1-\alpha,k,k}$ when $h \to 0$, then $\mu_1 \to 0$ and the error deteriorates. The error function $\varepsilon(t) = r_{1-\alpha}(t) - t^\alpha$ has $2k+1$ roots $\{\xi_i\}_1^{2k+1}$ in $(0,1)$, due to the $2k+2$ extreme points, including $\{0,1\}$, see, e.g., [12]. Since $\varepsilon(0) = E_{1-\alpha,k,k} > 0$, we have

$$r_{1-\alpha}(t) \ge t^\alpha, \qquad \forall t \in [0,\xi_1] \cup \bigcup_{i=1}^{k} [\xi_{2i}, \xi_{2i+1}]. \tag{9.12}$$

Therefore, whenever μ_1 is a priori estimated (enough to have a good lower bound for λ_1), we can choose a proper k, such that

$$\frac{\|\mathbf{u}_r - \mathbf{u}\|_2}{\|\mathbf{f}\|_2} \le \frac{E_{1-\alpha,k,k}}{\lambda_1^\alpha \mu_1^\alpha} = \frac{\Lambda^\alpha E_{1-\alpha,k,k}}{\lambda_1^{2\alpha}}, \qquad \mu_1 \in \bigcup_{i=1}^{k} [\xi_{2i}, \xi_{2i+1}]. \tag{9.13}$$

The case $\mu_1 \in [0, \xi_1]$ is more subtle and needs special care. Using $r_{1-\alpha}(t) = t^\alpha + \varepsilon(t)$, with $0 < \varepsilon(t) \le E_{1-\alpha,k,k}$, together with the fact that the function $g(\varepsilon) := \varepsilon/(\mu_1^\alpha + \varepsilon)$ is monotonically increasing for $\varepsilon \ge 0$ we obtain

$$\frac{\left|r_{1-\alpha}(\mu_1) - \mu_1^\alpha\right|}{\mu_1^\alpha r_{1-\alpha}(\mu_1)} \le \frac{E_{1-\alpha,k,k}}{\mu_1^\alpha(\mu_1^\alpha + E_{1-\alpha,k,k})} \le \frac{E_{1-\alpha,k,k}}{\mu_1^\alpha E_{1-\alpha,k,k}}.$$

For every $\mu_i > \xi_1$ we have

$$\frac{\left|r_{1-\alpha}(\mu_i) - \mu_i^\alpha\right|}{\mu_i^\alpha r_{1-\alpha}(\mu_i)} \le \frac{E_{1-\alpha,k,k}}{\mu_i^\alpha r_{1-\alpha}(\xi_1)} \le \frac{E_{1-\alpha,k,k}}{\mu_1^\alpha \xi_1^\alpha}.$$

Therefore, since $\xi_1^\alpha = r_{1-\alpha}(\xi_1) > r_{1-\alpha}(0) = E_{1-\alpha,k,k}$,

$$\frac{\|\mathbf{u}_r - \mathbf{u}\|_2}{\|\mathbf{f}\|_2} \le \frac{E_{1-\alpha,k,k}}{\lambda_1^\alpha \max(\xi_1^\alpha, E_{1-\alpha,k,k})} = \lambda_1^{-\alpha}, \qquad \forall \mu_1 \in [0, \xi_1]. \tag{9.14}$$

Typically $\lambda_1 = \mathrm{O}(1)$ for all h and $\mu_1 \to 0$ as $h \to 0$, thus unlike the BURA case (9.9) the (k,k)-R-BURA relative error is uniformly bounded when k is fixed and $h \to 0$.

The asymptotic behavior of the relative error (9.13) is derived analogously to (9.10):

$$\lim_{k\to\infty} e^{2\pi\sqrt{\alpha k}} \|\mathbf{u}_r - \mathbf{u}\|_2 = \frac{4^{2-\alpha}\Lambda^\alpha}{\lambda_1^{2\alpha}} \sin(\pi\alpha) \|\mathbf{f}\|_2. \tag{9.15}$$

In our experiments, we work with $r_{1-\alpha}$ in $\mathscr{R}(k+1,k)$ and $\mathscr{R}(k+1,k+1)$. Similar to BURA-method, the numerical computation of $\mathbf{u}_r$ involves solving of $k+1$ independent linear systems with nonnegative diagonal shifts of $\mathbb{A}$.

9.2.3 The Q-Method

The solver, proposed by Bonito and Pasciak in [2], incorporates an exponentially convergent quadrature scheme for the approximate computation of an integral

solution representation, i.e., uses the rational function

$$Q_\alpha(t) := \frac{2k' \sin(\pi\alpha)}{\pi} \sum_{\ell=-m}^{M} \frac{e^{2(\alpha-1)\ell k'}}{t + e^{-2\ell k'}}, \qquad t \in (0, \infty),$$

where $m = \lceil (1-\alpha)k \rceil$, $M = \lceil \alpha k \rceil$, $k' = \pi/(2\sqrt{\alpha(1-\alpha)k})$. Since

$$\lceil (1-\alpha)k \rceil + \lceil \alpha k \rceil = \begin{cases} k+1, & \alpha k \notin \mathbb{Z} \\ k, & \alpha k \in \mathbb{Z} \end{cases}$$

Q_α is either a $(k+1, k+1)$ or a $(k+2, k+2)$ rational function. The approximant of $\mathbf{u}_h$ has the form

$$\mathbf{u}_Q := \frac{2k' \sin(\pi\alpha)}{\pi} \sum_{\ell=-m}^{M} e^{2(\alpha-1)\ell k'} \left(\mathbb{A} + e^{-2\ell k'} \mathbb{I} \right)^{-1} \mathbf{f}. \tag{9.16}$$

The parameter $k' > 0$ controls the accuracy of $\mathbf{u}_Q$ and the number of linear systems to be solved. For example, $k' = 1/3$ gives rise to 120 systems for $\alpha = \{0.25, 0.75\}$ and 91 systems for $\alpha = 0.5$ guaranteeing $\|\mathbf{u}_Q - \mathbf{u}\|_2 \approx 10^{-7} \|\mathbf{f}\|_2$. We have

$$\frac{\|\mathbf{u}_Q - \mathbf{u}\|_2}{\|\mathbf{f}\|_2} \le \max_{\lambda_i} \left| Q_\alpha(\lambda_i) - \lambda_i^{-\alpha} \right| \approx \left| Q_\alpha(\lambda_1) - \lambda_1^{-\alpha} \right|. \tag{9.17}$$

Finally, the error analysis, developed in [2] states

$$\lim_{k \to \infty} e^{\pi\sqrt{\alpha(1-\alpha)k}} \|\mathbf{u}_Q - \mathbf{u}\|_2 = \frac{2\sin(\pi\alpha)}{\pi} \left(\frac{1}{\alpha} + \frac{1}{(1-\alpha)\lambda_1} \right) \|\mathbf{f}\|_2. \tag{9.18}$$

Remark 9.3 Varying the quadrature formulae, a family of related methods can be obtained. For example, Gauss-Jacobi quadrature rule is used to approximate the integral representation of the solution in [1].

9.2.4 Comparison of the Three Solvers

Comparing (9.10) with (9.18) we observe exponential decay of both errors with respect to the number of linear systems to be solved. The exponential order of the BURA estimate is at least twice higher than the one for the quadrature rule, but there is a multiplicative factor $\Lambda^{1-\alpha}$ in (9.10), which depends on the mesh size h and $\Lambda \to \infty$ as $h \to 0$. This implies trade-off between numerical accuracy and computational efficiency for the BURA method. The choice of k for $r_\alpha \in \mathscr{R}(k, k)$ should be synchronized with h, while the size of h does not affect the choice of

k for Q_α. Another difference between the two approaches is that the error bound in (9.10) is unbalanced and can be reached only for $\mathbf{f} = \boldsymbol{\Psi}_1$ and only if $\Lambda^{-1}\lambda_1$ is an extreme point for r_α (see (9.9)), while the error bound in (9.18) is balanced. Hence, the BURA error heavily depends on the decomposition of the right-hand-side $\mathbf{f}$ along $\{\boldsymbol{\Psi}_i\}$ and possesses a wide range of values, while the quadrature error is independent on $\mathbf{f}$.

The errors in (9.9) and (9.13) are bounded by the expressions $\frac{\Lambda^{1-\alpha} E_{\alpha,k,k}}{\lambda_1}$ and $\frac{\Lambda^{\alpha} E_{1-\alpha,k,k}}{\lambda_1^{2\alpha}}$. Since $E_{\alpha,k,k}$ is monotonically increasing function with respect to α and $\Lambda, \lambda_1 > 1$, for $\alpha > 0.5$ the R-BURA method provides better theoretical error bounds, while for $\alpha < 0.5$ so does the BURA one. In the case $\alpha = 0.5$ the two approaches behave similarly. The drawback for the R-BURA method is the additional condition on k and h, namely $\mu_1 \in \bigcup_{i=1}^{k} [\xi_{2i}, \xi_{2i+1}]$. On the other hand, if we can guarantee this, then the R-BURA method has some advantage, as we solve one linear system less $k+1$ for BURA vs k for R-BURA, when using the same (k,k) function $r_{0.5}(t)$). Below we provide an experimental comparison of these approaches for various h and $k = \{7, 8, 9\}$.

9.3 Numerical Tests: Comparative Analysis and Proof of Concept

We consider the fractional Laplace problem with homogeneous boundary conditions (9.3) in both 1-D and 2-D. In 1-D we use the well-known eigenvectors and eigenvalues of the corresponding SPD matrix for experimental validation of the theoretical error analysis. In 2-D we investigate the relation between numerical accuracy and computational efficiency of the considered three solvers.

9.3.1 Algorithm for Computing BURA

Following [6] we consider $\alpha = \{0.25, 0.5, 0.75\}$ and investigate methods with similar computational efficiency. The rational functions $r_\alpha(t)$ are computed using the modified Remez algorithm, described in [6, Section 3.2]. In the case $\alpha = 0.25$ we compare (k,k)-BURA with the k-Q-method, $k = 9$. The corresponding numerical solvers incorporate 10, respectively 11 linear systems with positive diagonal shifts of $\mathbb{A}$. In the cases $\alpha = \{0.5, 0.75\}$ we compare (k,k)-BURA, $(k+1,k)$-R-BURA, $(k+1,k+1)$-R-BURA, and k-Q-method for $k = 7$. This gives rise to $k+1$ linear systems with positive diagonal shifts of $\mathscr{A}$ for the first two methods and $k+2$ linear systems with positive diagonal shifts for the last method.

The maximal approximation errors of the involved BURA functions are summarized in Table 9.1. We use $*$ to indicate errors that cannot be computed when the Quadruple-precision floating-point format is applied for the arithmetics. The first four zeros of the associated functions $\varepsilon(t) = r_{1-\alpha}(t) - t^{\alpha}$ are presented in Table 9.2. Note that they are needed only in the analysis of R-BURA setting, thus we exclude $\alpha = 0.25$ where R-BURA behaves worse than BURA.

9.3.2 Numerical Results for 1-D Fractional Laplace Problem

For this problem we have $\lambda_1 = (4/h^2)\sin^2(\pi h/2)$, $\Lambda = 4/h^2$, and $\mu_1 = \sin^2(\pi h/2)$. The numerical results are given in Figs. 9.1 and 9.2.

First, we solve the system $\mathbb{A}^{\alpha}\mathbf{u} = \boldsymbol{\Psi}_1$. The theoretical errors of the three methods, given by $\Lambda^{-\alpha}|\mu_1^{-1}r_{\alpha}(\mu_1) - \mu_1^{-\alpha}|$ of BURA, $\Lambda^{-\alpha}|r_{1-\alpha}(\mu_1) - \mu_1^{\alpha}|/(\mu_1^{\alpha}r_{1-\alpha}(\mu_1))$ of R-BURA, and $|Q_{\alpha}(\lambda_1) - \lambda_1^{-\alpha}|$ of the Q-method, are presented as function of $h \in [10^{-7}, 10^{-2}]$ in Fig. 9.1. The numerical results in each graph are obtained by comparable computational complexity (expressed through the number of systems solved).

The oscillating behavior of the BURA-related error is due to the placement of μ_1 with respect to the extreme points of $\varepsilon(t)$. When $\alpha = 0.75$, $k = 8$ (right plot) and $h < 10^{-4}$ we observe the constant asymptotic behavior of the R-BURA errors towards $\pi^{-2\alpha}$ as $h \to 0$. Similar observation is made for $\alpha = 0.5$ and $h < 2 \cdot 10^{-5}$. Since $\mu_1 \approx \pi^2 h^2/4$, we have that $\mu_1 \approx 2.5 \cdot 10^{-8}$ for $\alpha = 0.75$ and $h = 10^{-4}$, which, as seen from Table 9.2, is close to the first zero ξ_1 of $\varepsilon(t)$ for the corresponding $(8, 7)-$ and $(8, 8)-$ approximations $r_{0.25}(t)$. This asymptotic behavior perfectly agrees with the error estimate (9.14). The same analysis can be made for $\alpha = 0.5$ and $h = 2 \cdot 10^{-5}$, where $\mu_1 \approx 5 \cdot 10^{-10}$. The Q-method errors are independent of h. We observe that for $\alpha = 0.5$ the BURA and R-BURA solvers have comparable accuracy over the whole interval $(0, 1]$. For $\alpha = 0.75$ and $h \in [10^{-4}, 10^{-3}]$, we observe that both $(8, 7)-$ and $(8, 8)$-R-BURA functions give worse relative errors than the $(7, 7)$-BURA function, since $\mu_1 \in [\xi_1, \xi_2]$ (see Table 9.2).

The second set of experiments deals with the error over the whole spectrum of $\mathbb{A}$ and is presented in Fig. 9.2. For $h = 10^{-3}$ and $h = 10^{-6}$ we compute $\Lambda^{-\alpha}|\mu_i^{-1}r_{\alpha}(\mu_i) - \mu_i^{-\alpha}|$, $\Lambda^{-\alpha}|r_{1-\alpha}(\mu_i) - \mu_i^{\alpha}|/(\mu_i^{\alpha}r_{1-\alpha}(\mu_i))$, and $|Q_{\alpha}(\lambda_i) - \lambda_i^{-\alpha}|$ for all i, which is equivalent to letting the right-hand-side in $\mathbb{A}^{\alpha}\mathbf{u} = \mathbf{f}$ run over the eigenvectors of $\mathbb{A}$ ($\mathbf{f} = \boldsymbol{\Psi}_i$). The plots in the first row illustrate the complete spectral decomposition of the error for $h = 10^{-3}$. For $h = 10^{-6}$ we show the spectral error over the first 1% of the eigen-modes in the second row. The unbalanced behavior of the BURA-related errors in contrast to the balanced behavior of the errors of the Q-method is clearly observed. High-frequency modes are practically perfectly reconstructed by the R-BURA methods, while the low-frequency ones lead to larger errors. When $h = 10^{-3}$ and k is chosen accordingly, all BURA-method errors are smaller than the corresponding Q-method errors. When $h = 10^{-6}$ and k is chosen

Table 9.1 Errors $E_{\alpha,k,m}$ of $r_\alpha(t)$ for $t \in [0, 1]$, used in the analysis of BURA and R-BURA numerical methods

α	$E_{\alpha,5,5}$	$E_{\alpha,6,6}$	$E_{\alpha,7,7}$	$E_{\alpha,8,7}$	$E_{\alpha,8,8}$	$E_{\alpha,9,8}$	$E_{\alpha,9,9}$	$E_{\alpha,10,9}$
0.25	2.8676e−5	9.2522e−6	3.2566e−6	1.9500e−6	1.2288e−6	7.5972e−7	4.9096e−7	3.1128e−7
0.50	2.6896e−4	1.0747e−4	4.6037e−5	3.0789e−5	2.0852e−5	*	*	*
0.75	2.7348e−3	1.4312e−3	7.8269e−4	*	*	*	*	*

Table 9.2 First four roots of $\varepsilon(t)$ for $r_{1-\alpha} \in \mathscr{R}(k,m)$

	First four zeros $\xi_1, \xi_2, \xi_3, \xi_4$ of $\varepsilon(t) = r_{1-\alpha}(t) - t^\alpha$							
	$\alpha = 0.50$				$\alpha = 0.75$			
(k,m)	ξ_1	ξ_2	ξ_3	ξ_4	ξ_1	ξ_2	ξ_3	ξ_4
(5, 5)	1.030e−7	6.732e−6	6.592e−5	4.352e−4	2.185e−6	7.269e−5	5.004e−4	2.353e−3
(6, 6)	1.650e−8	1.076e−6	1.053e−5	6.950e−5	4.836e−7	1.609e−5	1.108e−4	5.216e−4
(7, 7)	3.100e−9	1.981e−7	1.932e−6	1.275e−5	1.202e−7	3.999e−6	2.754e−5	1.297e−4
(8, 7)	1.400e−9	8.840e−8	8.644e−7	5.705e−6	6.070e−8	2.019e−6	1.390e−5	6.544e−5
(8, 8)	7.00e−10	4.070e−8	3.967e−7	2.617e−6	3.280e−8	1.091e−6	7.509e−6	3.536e−5

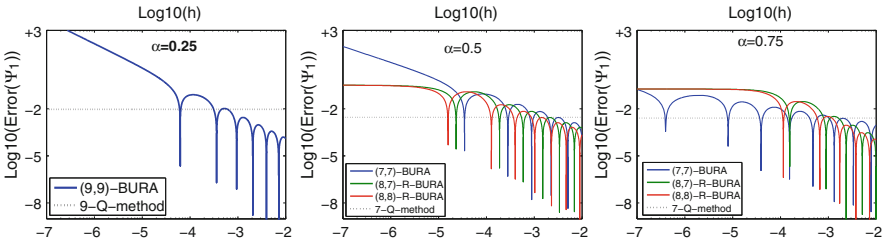


Fig. 9.1 Comparison of the theoretical error bounds $\|\mathbf{u}_r - \mathbf{u}\|_2/\|\mathbf{f}\|_2$, $\mathbf{f} = \boldsymbol{\Psi}_1$, of the three solvers with respect to h for the 1-D fractional Laplacian

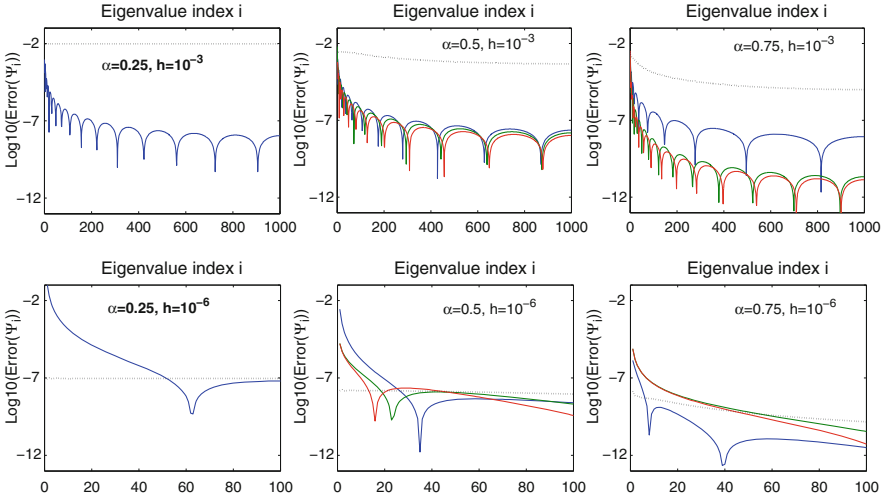


Fig. 9.2 Spectral decomposition of the error for the 1-D fractional Laplace problem with $h = 10^{-3}$ (top) and $h = 10^{-6}$ (bottom). Colors are with respect to the legend of Fig. 9.1. Top: for each i we plot the corresponding errors $\|\mathbf{u}_r - \mathbf{u}_Q\|_2/\|\mathbf{f}\|_2$ for $\mathbf{f} = \boldsymbol{\Psi}_i$. The bottom plots present the corresponding errors for the first 100 eigenvectors, i.e., for $\{\boldsymbol{\Psi}_i\}_1^{100}$

poorly, then the BURA and R-BURA errors on the first several eigenvectors can be significantly larger than the corresponding Q-method errors. However, among a million of eigenvectors, the Q-method outperforms the BURA methods on not more than 50 of them. Comparing BURA to R-BURA approaches, we experimentally confirm that the two methods behave similarly when $\alpha = 0.5$, while R-BURA is better for $\alpha = 0.75$.

9.3.3 2-D Numerical Experiments

We consider the finite difference approximation of (9.3) with two different r.h.s., namely, f_1 and f_2:

$$f_1(x,y) = \begin{cases} 1, \text{ if } (x-0.5)(y-0.5) > 0, \\ -1, \text{ otherwise.} \end{cases} \qquad f_2(x,y) = \cos(\pi h x)\cos(\pi h y) \tag{9.19}$$

The function f_1 has a jump discontinuity along $x = 0.5$ and $y = 0.5$ and has already been used as a test function in this framework [2, 6]. In this case $\lambda_1 \approx 2\pi^2$, $\Lambda = \|\mathbb{A}\|_\infty = 8h^{-2}$, and $\mu_1 = \sin^2(\pi h/2)$. The reference solution $\mathbf{u}_Q$ is generated by the Q-method with $k' = 1/3$ on a fine mesh with mesh-size $h = 2^{-12}$. Note that $\mathbf{u}_Q$ is an approximation to the exact solution $\mathbf{u}$ with six correct digits, $\|\mathbf{u}_Q - \mathbf{u}\|_2/\|\mathbf{f}\|_2 \approx 10^{-7}$ (see [2, Table 3]) . The numerical results are summarized in Tables 9.3, 9.4, 9.5. The presented relative ℓ_2-errors illustrate the theoretical analysis, while ℓ_∞-errors are given as additional information.

The balanced error distribution along the full spectrum for the Q-method gives rise to stable relative errors, independent of h for all α on both examples. The error distribution for BURA and R-BURA methods depends on h and k. From Table 9.3 we see that for $\alpha = 0.25$ and $h \geq 2^{-12}$ the choice $k = 9$ for the (k,k)-BURA method has lower ℓ_2-error than the error of the Q-solver for comparable computational work.

Next set of numerical experiments is presented in Tables 9.4 and 9.5. For $\alpha = 0.5$ and $k = 7$, we have $\mu_1 > \xi_2$ for $h \geq 2^{-12}$ and both $(8,7)$- and $(8,8)$-R-BURA methods are reliable. Their ℓ_2 relative errors are smaller than the corresponding errors for the k-Q-solver on all considered mesh sizes. The $(8,8)$-R-BURA solver is more accurate than the $(8,7)$-R-BURA one. The choice $k = 7$ for the BURA solver is reliable for $h \geq 2^{-11}$, but for $h = 2^{-12}$ a larger k is needed. Like the 1-D case, BURA and R-BURA solvers behave similarly when k is properly chosen.

For $\alpha = 0.75$ and $h < 2^{-10}$, we have $\mu_1 \in [\xi_1, \xi_2]$ for both $(8,7)$- and $(8,8)$-R-BURA methods. As a result the $(8,8)$-R-BURA solution is less accurate than the one obtained by $(7,7)$-BURA for $h = \{2^{-11}, 2^{-12}\}$, while the $(8,7)$-R-BURA solver is outperformed by all the other three for $h = 2^{-12}$. Once we guarantee that $\mu_1 > \xi_2$, the R-BURA approach gives rise to the highest accuracy. Again, the $(8,8)$-R-BURA solution is more accurate than the one obtained by $(8,7)$-R-BURA.

Finally, we present a comparison on numerical accuracy versus computational efficiency for properly chosen k and h. We fix $h = 2^{-10}$ and for each BURA-related method we compute the smallest k, such that the corresponding k-Q-method gives smaller relative ℓ_2-error for $\mathbf{f}_1$ and $\mathbf{f}_2$. When $\alpha = 0.25$ the $(9,9)$-BURA solver has lower accuracy than the 37-Q-method for $\mathbf{f}_1$ and the 36-Q-method for $\mathbf{f}_2$, respectively. This means that, instead of the 10 linear systems incorporated in the BURA-method, we need to solve 39, respectively 37, linear systems for the Q-method. For $\alpha = 0.5$ and both $\{\mathbf{f}_1, \mathbf{f}_2\}$ we need to use $k = 16$ for the Q-method

Table 9.3 Relative errors for various discretization levels and $\alpha = 0.25$

		Checkerboard right-hand-side				Tensor product cosine right-hand-side			
		(k,k)-BURA		k-Q-method		(k,k)-BURA		k-Q-method	
k	h	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞
9	2^{-8}	5.863e−3	5.236e−2	1.080e−2	4.285e−2	2.781e−4	2.600e−3	6.823e−3	9.381e−3
	2^{-9}	2.823e−3	3.234e−2	9.707e−3	2.425e−2	1.441e−4	1.813e−3	6.752e−3	8.586e−3
	2^{-10}	1.253e−3	1.785e−2	9.436e−3	1.870e−2	2.268e−4	1.210e−3	6.726e−3	7.984e−3
	2^{-11}	5.027e−4	7.443e−3	9.381e−3	1.383e−2	4.888e−4	7.707e−4	6.717e−3	7.412e−3
	2^{-12}	4.883e−3	1.019e−2	9.374e−3	9.568e−3	4.425e−3	5.031e−3	6.713e−3	6.790e−3

Table 9.4 Relative errors for $\mathbf{f}_1$ on various discretization levels and $\alpha = \{0.5, 0.75\}$

		Checkerboard right-hand-side							
		(k,k)-BURA		$(k+1,k)$-R-BURA		$(k+1,k+1)$-R-BURA		k-Q-method	
(α, k)	h	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞
(0.50, 7)	2^{-8}	1.383e−3	6.814e−3	1.351e−3	6.820e−3	1.347e−3	6.806e−3	3.113e−3	6.800e−3
	2^{-9}	8.692e−4	3.503e−3	6.777e−4	3.497e−3	6.687e−4	3.497e−3	2.895e−3	4.573e−3
	2^{-10}	7.657e−4	1.808e−3	7.845e−4	1.766e−3	4.660e−4	1.619e−3	2.841e−3	3.552e−3
	2^{-11}	8.243e−4	1.447e−3	1.879e−3	4.204e−3	2.583e−4	6.293e−4	2.830e−3	3.078e−3
	2^{-12}	5.423e−3	1.135e−2	1.976e−3	4.831e−3	1.447e−3	2.861e−3	2.828e−3	2.902e−3
(0.75, 7)	2^{-8}	4.194e−4	1.277e−3	4.226e−4	1.278e−3	4.206e−4	1.272e−3	1.558e−3	2.276e−3
	2^{-9}	2.509e−4	6.038e−4	2.281e−4	6.053e−4	1.967e−4	6.029e−4	1.514e−3	1.984e−3
	2^{-10}	4.264e−4	9.644e−4	2.410e−4	5.451e−4	1.234e−4	2.604e−4	1.503e−3	1.887e−3
	2^{-11}	5.222e−4	1.206e−3	2.128e−4	3.923e−4	5.601e−4	1.185e−3	1.500e−3	1.843e−3
	2^{-12}	6.560e−5	1.420e−4	3.077e−3	6.268e−3	1.316e−3	2.819e−3	1.499e−3	1.823e−3

Table 9.5 Relative errors for $\mathbf{f}_2$ on various discretization levels and $\alpha = \{0.5, 0.75\}$

		Tensor product cosine right-hand-side							
		(k,k)-BURA		$(k+1,k)$-R-BURA		$(k+1,k+1)$-R-BURA		k-Q-method	
(α,k)	h	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞
(0.50, 7)	2^{-8}	1.509e−4	3.299e−4	5.790e−5	1.810e−4	5.031e−5	1.809e−4	1.423e−3	2.901e−3
	2^{9}	2.717e−4	6.065e−4	1.179e−4	2.901e−4	1.075e−4	2.438e−4	1.418e−3	2.867e−3
	2^{-10}	3.432e−4	8.118e−4	3.292e−4	7.450e−4	1.801e−4	4.192e−4	1.415e−3	2.858e−3
	2^{11}	4.545e−4	1.282e−3	8.335e−4	1.817e−3	1.648e−4	4.218e−4	1.415e−3	2.856e−3
	2^{-12}	2.420e−3	5.190e−3	9.188e−4	2.139e−3	6.791e−4	1.635e−3	1.414e−3	2.856e−3
(0.75, 7)	2^{-8}	2.222e−5	5.484e−5	1.893e−5	3.896e−5	3.586e−6	8.906e−6	7.386e−4	1.422e−3
	2^{-9}	7.383e−5	1.836e−4	5.171e−5	1.086e−4	6.334e−6	1.750e−5	7.367e−4	1.420e−3
	2^{-10}	1.861e−4	4.059e−4	1.024e−4	2.215e−4	4.212e−5	9.762e−5	7.358e−4	1.420e−3
	2^{-11}	2.333e−4	5.147e−4	1.125e−4	2.941e−4	2.492e−4	5.215e−4	7.354e−4	1.420e−3
	2^{-12}	9.046e−5	2.771e−4	1.369e−3	2.802e−3	5.864e−4	1.238e−3	7.353e−4	1.420e−3

to get better accuracy than $(7, 7)$-BURA, $k = 16$ for the Q-method to get better accuracy than $(8, 7)$-R-BURA, and $k = 20$ for the Q-method to get better accuracy than $(8, 8)$-R-BURA. For $\mathbf{f}_1$ and $\alpha = 0.75$ we need to use $k = 13$ for the Q-method to get better accuracy than $(7, 7)$-BURA, $k = 17$ for the Q-method to get better accuracy than $(8, 7)$-R-BURA, and $k = 25$ for the Q-method to get better accuracy than $(8, 8)$-R-BURA. Finally, for $\mathbf{f}_2$ and $\alpha = 0.75$ we need to use $k = 13$ for the Q-method to get better accuracy than $(7, 7)$-BURA, $k = 21$ for the Q-method to get better accuracy than $(8, 7)$-R-BURA, and $k = 29$ for the Q-method to get better accuracy than $(8, 8)$-R-BURA. Therefore, with respect to numerical accuracy versus computational efficiency the R-BURA solver for $\alpha = 0.75$ behaves similarly to the BURA solver for $\alpha = 0.25$ and can be up to four times more efficient than the corresponding Q-solver.

9.4 Concluding Remarks

We present a comparative analysis of three methods for solving equations involving fractional powers of elliptic operators, namely, the method of Bonito and Pasciak, [2], BURA method based on the best rational approximation of $t^{1-\alpha}$, [6], and the new method, R-BURA, based on the best rational approximation of t^{α} on $[0, 1]$.

The method of Bonito and Pasciak, [2], uses Sinc quadratures and has exponential convergence with respect to the number of quadrature nodes. The BURA method, [6], has exponential convergence as well, is accurate for α close to 0, and performs well for fixed step-size h. The new method, R-BURA, has also exponential convergence with respect to the degree of the rational approximation for fixed step-size h. In contrast to BURA, R-BURA method performs better for α close to 1. However, the accuracy of both methods deteriorates when $h \to 0$.

We expect that one could be able to construct a method that combines the advantages of these approaches, computational efficiency and exponential convergence rate. Moreover, the proposed algorithms can be used in 3-dimensional computations without any changes. For such computations with the first BURA method, see [6].

The development of solution methods for fractional-in-space diffusion problems in the case of local mesh refinement is a topic of current interest. For instance, for the considered test problems, the goal will be to resolve some clearly expressed boundary and internal layers. For such problems, the BURA methods should be robust with respect to the increased condition number of the matrix $\mathbb{A}$.

Acknowledgements This research has been partially supported by the Bulgarian National Science Fund under grant No. BNSF-DN12/1. The work of R. Lazarov has been partially supported by the grant NSF-DMS #1620318. The work of S. Harizanov has been partially supported by the Bulgarian National Science Fund under grant No. BNSF-DM02/2.

References

1. Aceto, L., Novati, P.: Rational approximation to the fractional Laplacian operator in reaction-diffusion problems. SIAM J. Sci. Comput. **39**(1), A214–A228 (2017)
2. Bonito, A., Pasciak, J.: Numerical approximation of fractional powers of elliptic operators. Math. Comput. **84**(295), 2083–2110 (2015)
3. Druskin, V., Knizhnerman, L.: Extended Krylov subspaces: approximation of the matrix square root and related functions. SIAM J. Matrix Anal. Appl. **19**(3), 755–771 (1998)
4. Filip, S.I., Nakatsukasa, Y., Trefethen, L.N., Beckermann, B.: Rational minimax approximation via adaptive barycentric representations (2018). arXiv:1705.10132v2
5. Harizanov, S., Margenov, S.: Positive approximations of the inverse of fractional powers of SPD M-matrices. In: Control Systems and Mathematical Methods in Economics. Lecture Notes in Economics and Mathematical Systems, vol. 687, pp. 147–163. Springer International Publishing AG (2018)
6. Harizanov, S., Lazarov, R., Margenov, S., Marinov, P., Vutov, Y.: Optimal solvers for linear systems with fractional powers of sparse SPD matrices. Numer. Linear Algebra Appl. **25**(4), 115–128 (2018). https://doi.org/10.1002/nla.2167
7. Higham, N.J.: Stable iterations for the matrix square root. Numer. Algorithms **15**(2), 227–242 (1997)
8. Ilić, M., Liu, F., Turner, I.W., Anh, V.: Numerical approximation of a fractional-in-space diffusion equation, I. Fract. Calc. Appl. Anal. **8**(3), 323–341 (2005)
9. Ilić, M., Turner, I.W., Anh, V.: A numerical solution using an adaptively preconditioned Lanczos method for a class of linear systems related with the fractional Poisson equation. Int. J. Stoch. Anal. **2008**, 104525 (2009)
10. Kenney, C., Laub, A.J.: Rational iterative methods for the matrix sign function. SIAM J. Matrix Anal. Appl. **12**(2), 273–291 (1991)
11. Meinardus, G.: Approximation of Functions: Theory and Numerical Methods. Springer, New York (1967)
12. Saff, E.B., Stahl, H.: Asymptotic distribution of poles and zeros of best rational approximants to x^α on [0, 1]. In: Topics in Complex Analysis. Banach Center Publications, vol. 31. Institute of Mathematics, Polish Academy of Sciences, Warsaw (1995)
13. Stahl, H.: Best uniform rational approximation of x^α on [0, 1]. Bull. Am. Math. Soc. **28**(1), 116–122 (1993)
14. Stahl, H.R.: Best uniform rational approximation of x^α on [0, 1]. Acta Math. **190**(2), 241–306 (2003)
15. Thomée, V.: Galerkin finite element methods for parabolic problems. Springer Series in Computational Mathematics, vol. 25. Springer, Berlin, 2nd edn. (2006)
16. Varga, R.S., Carpenter, A.J.: Some numerical results on best uniform rational approximation of x^α on [0, 1]. Numer. Algorithms **2**(2), 171–185 (1992)

Chapter 10
A Three-Level Extension of the GDSW Overlapping Schwarz Preconditioner in Two Dimensions

Alexander Heinlein, Axel Klawonn, Oliver Rheinbach, and Friederike Röver

Abstract A three-level extension of the GDSW overlapping Schwarz preconditioner in two dimensions is presented, constructed by recursively applying the GDSW preconditioner to the coarse problem. Numerical results, obtained for a parallel implementation using the Trilinos software library, are presented for up to 90,000 cores of the JUQUEEN supercomputer. The superior weak parallel scalability of the three-level method is verified. For large problems and a large number of cores, the three-level method is faster by more than a factor of two, compared to the standard two-level method. The three-level method can also be expected to scale when the classical method will already be out-of-memory.

10.1 The Standard GDSW Preconditioner

Consider the symmetric positive definite and sparse linear system of equations

$$Kx = b \tag{10.1}$$

arising from the weak formulation of a second order scalar elliptic boundary value problem in two dimensions and discretized by low order finite elements. If K is large, it is state of the art to solve (10.1) iteratively by a Krylov method, such as the preconditioned conjugate gradient (PCG) method, in combination with a preconditioner.

A. Heinlein · A. Klawonn
Mathematisches Institut, Universität zu Köln, Köln, Germany
e-mail: alexander.heinlein@uni-koeln.de; axel.klawonn@uni-koeln.de

O. Rheinbach (✉) · F. Röver
Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Freiberg, Germany
e-mail: oliver.rheinbach@math.tu-freiberg.de; oliver.rheinbach@math.tu-bergakademie.de; friederike.roever@math.tu-freiberg.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_10

The GDSW (Generalized Dryja–Smith–Widlund) preconditioner is a two-level overlapping Schwarz domain decomposition preconditioner [20] with an energy-minimizing coarse space introduced in [4, 5]. It can be written in the form

$$M_{\mathrm{GDSW}}^{-1} = \underbrace{\Phi K_0^{-1} \Phi^T}_{\text{Coarse Level}} + \underbrace{\sum_{i=1}^{N} R_i^T \widetilde{K_i}^{-1} R_i}_{\text{First Level}}, \tag{10.2}$$

where

$$K_0 = \Phi^T K \Phi \tag{10.3}$$

is the coarse matrix, and where the matrices

$$\widetilde{K_i} = R_i K R_i^T \tag{10.4}$$

represent the local overlapping subdomain problems. The columns of Φ contain the coarse basis functions, which are chosen as discrete harmonic extensions of functions Φ_Γ defined on the interface Γ of the non-overlapping domain decomposition. To define Φ_Γ, the interface is decomposed into vertices and edges in 2D and vertices, edges, and faces in 3D [4]. Then, the columns of Φ_Γ correspond to the restrictions of the null space of the global Neumann matrix to these interface components. The coarse matrix $K_0 = \Phi^T K \Phi$ can be interpreted as the stiffness matrix assembled using the coarse basis functions. The coarse correction operator $\Phi K_0^{-1} \Phi^T$ is usually not formed explicitly but evaluated from right to left when used in matrix-vector multiplications.

For linear elliptic problems, the condition number of the Schwarz operator is bounded by

$$\kappa(M_{\mathrm{GDSW}}^{-1} K) \leq C \left(1 + \frac{H}{\delta}\right) \left(1 + \log\left(\frac{H}{h}\right)\right)^2, \tag{10.5}$$

where h is the size of a finite element, H the size of a nonoverlapping subdomain, and δ the width of the overlap; see [4].

The crucial advantage of GDSW preconditioners compared to other domain decomposition preconditioners is that they can be constructed in an algebraic fashion from the fully assembled matrix K and without the need of an additional coarse triangulation.

This will also facilitate the construction of the three-level GDSW preconditioner presented in the following section.

10.2 The Three-Level GDSW Preconditioner

For a large number of subdomains, the exact solution of the GDSW coarse problem, i.e., the factorization of K_0 (see (10.3)) by a sparse direct solver, may not be possible, or reasonable, anymore [8, 9]. This is particularly due to the superlinear memory complexity of sparse direct solvers. As a remedy, the GDSW preconditioner can recursively be applied to the coarse problem resulting in the three-level preconditioner discussed in this paper. Three-level BDDC methods in two dimensions [22] are nonoverlapping preconditioners which are closely related to the method presented here. Note that this recursive application of the GDSW preconditioner could also be extended to more than three levels; cf. related multi-level BDDC [2, 15] and Schwarz [14, 18] domain decomposition methods and multigrid [7] methods. An alternative approach for improving the scalability limit is the reduction of the dimension of the GDSW coarse space; cf. [6, 11].

To define the three-level GDSW preconditioner, let us decompose the domain Ω into nonoverlapping subregions Ω_{i0} of diameter H_c; see [22] and Fig. 10.1. Each subregion is decomposed into nonoverlapping subdomains of diameter H.

Extending each subregion Ω_{i0} to Ω'_{i0} by recursively adding layers of subdomains, an overlapping decomposition into subregions is obtained. Similarly, each subdomain Ω_i is extended to Ω'_i, which then overlaps other extended subdomains. The overlap on subregion level is denoted by Δ; the overlap on the subdomain level is denoted by δ, consistent with the notation of the two-level method; see Fig. 10.1.

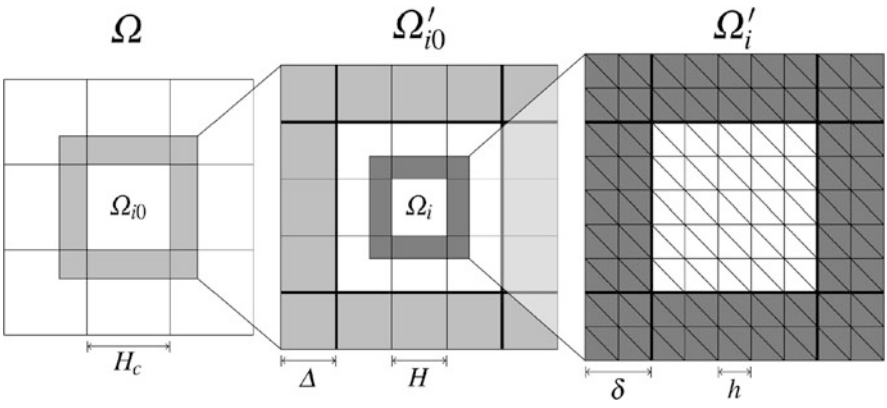


Fig. 10.1 Structured decomposition of the computational domain Ω into nonoverlapping subregions Ω_{i0} (left), a zoom into one overlapping subregion Ω'_{i0} consisting of subdomains Ω_i (middle), and a zoom into one overlapping subdomain Ω'_i (right). Each level of zoom corresponds to one level of the preconditioner

The three-level GDSW preconditioner then is defined as

$$M_{3GDSW}^{-1} = \underbrace{\Phi\Big(\overbrace{\Phi_0 K_{00}^{-1}\Phi_0^T}^{\text{Third Level}} + \overbrace{\sum_{i=1}^{N_0} R_{i0}^T \widetilde{K}_{i0}^{-1} R_{i0}}^{\text{Second Level}}\Big)\Phi^T}_{\text{Coarse Levels}} + \underbrace{\sum_{j=1}^{N} R_j^T \widetilde{K}_j^{-1} R_j}_{\text{First Level}}, \tag{10.6}$$

where

$$K_{00} = \Phi_0^T K_0 \Phi_0, \text{ and } \widetilde{K}_{i0} = R_{i0} K_0 R_{i0}^T.$$

By $V_1, \ldots, V_N$, we denote the local subspaces corresponding to the overlapping subdomains and V^0 denotes the corresponding coarse space. The restriction operators on the subdomain level are defined as $R_i : V^h(\Omega) \to V_i := V^h(\Omega_i')$ for $i = 1, \ldots, N$. On the subregion level, we define the restriction operators to the overlapping subregions Ω_{i0}' as $R_{i0} : V^0 \to V_i^0 := V^0(\Omega_{i0}')$ for $i = 1, \ldots, N_0$. The respective coarse space is denoted as V_{00} and spanned by the coarse basis functions Φ_0.

10.3 Implementation and Software Libraries

Our parallel three-level GDSW implementation is based on the implementation of the GDSW preconditioner described in [8–10]; it is therefore also based on the Trilinos linear algebra package Epetra. A newer version of the implementation described in [9], now based on Xpetra, has recently been added to the Trilinos [13] package ShyLU [16] as part of the FROSch (Fast and Robust Overlapping Schwarz) framework [12]. It has been pushed to the public Trilinos git repository [21] in October 2017.

We here apply our three-level GDSW implementation to a two dimensional Laplace problem on the unit square with homogeneous Dirichlet boundary conditions on $\partial\Omega$. We use piecewise linear finite elements and a structured decomposition of the computational domain; see Fig. 10.1. As a Krylov method, we apply a parallel implementation of the PCG algorithm provided by the Trilinos package Belos [3]. Namely, we use the Belos implementation of a block preconditioned conjugate gradient method (BelosPseudoBlockCG) developed for multiple right hand sides. However, in our numerical experiments in Sect. 10.4, a block size of 1 is used, which reduces the pseudo-block PCG to standard PCG. Opposed to the standard PCG in Belos (BelosCG), this implementation offers the convenient standard condition number estimate using the tridiagonal matrix constructed in the Lanczos process.

When using the Belos pseudo-block PCG, we observed that the number of forward-backward substitutions is always larger by 2 than the number of PCG

iterations. This is because in the Belos pseudo-block PCG implementation after convergence an additional preconditioned residual is computed to prepare processing of the subsequent right hand sides; see the method *solve* of the BelosPseudoBlockCGSolMgr class. This is step is unnecessary if there are no additional right hand sides but is still performed for ease of implementation.

We use the relative stopping criterion $\|r^k\|_2/\|r^0\|_2 \leq 10^{-6}$, where r^k is the k-th residual and r^0 the initial residual. Trilinos version 12.11 (Dev) is used; cf. [13]. We use the IBM XL C/C++ compiler for Blue Gene V.12.1, and Trilinos is linked to the ESSL. Note that in [9] clang 4.7.2 was used on the JUQUEEN.

In this paper, we assume that we have a fast and scalable method to identify vertex and edges degrees of freedom. This cost is therefore neglected in this paper. In [9, Section 4.3], a completely algebraic method was discussed, which only uses the Trilinos Epetra map of the nonoverlapping domain decomposition. However, its superlinear complexity becomes significant beyond 10^4 cores, and it is then preferable to use the infrastructure of a parallel mesh handler for this task, if available.

To solve the sparse linear problems arising in the preconditioner, we always use the sparse direct solver package MUMPS 4.10.0 [1] in symmetric, sequential mode, and interfaced through the Trilinos package Amesos [17]. For the two- as well as the tree-level method, the problem on the coarsest level is thus factorized by MUMPS on the master process. The option to solve the coarsest level problem using the MPI-parallel mode of MUMPS, on a subset of processes (as in [9, Section 4.6]), is not used in this paper.

10.4 Numerical Results on the JUQUEEN Supercomputer

To evaluate the numerical and weak parallel scalability of the three-level preconditioner, we perform numerical experiments on the JUQUEEN BG/Q supercomputer [19] at Jülich Supercomputing Centre, using our implementation based on Trilinos 12.11 (Dev).

We can expect that the advantages of the three-level GDSW preconditioner will become visible only for a large number of subdomains and cores. Although JUQUEEN will be decommissioned and replaced later this year, it is still the supercomputer in Europe with the largest number of cores and thus the most suitable machine for our tests. A node of JUQUEEN has 16 cores (PowerPC A2, 1.6 GHz) and 16 GB of memory. The largest computations presented here will use 5 625 nodes corresponding to 90,000 cores.

In all of our numerical experiments, we maintain a one-to-one correspondence of subdomains to processor cores. Therefore, for our model problem in two dimensions, $(1/H)^2$ always corresponds to the number of subdomains and the number of processor cores. We use one MPI rank for each core. Thus, 16 Gb of memory are available for the 16 MPI processes running on a node. No threading is applied.

In Sect. 10.4.1, we first will investigate the numerical properties of the three-level GDSW method. We will study the dependence of the iteration count and the condition number on the subdomain overlap δ, on the number of degrees of freedom in a subdomain, on the subregion overlap Δ, and on the number of subdomains per subregion.

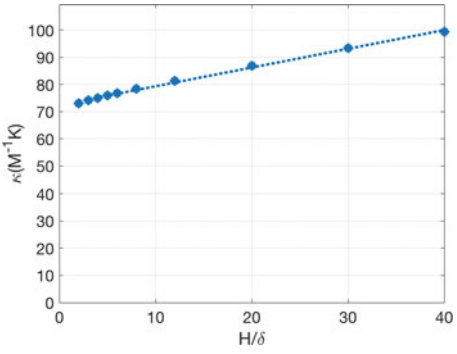
For the weak numerical and parallel scalability, in Sect. 10.4.2, we will consider problems of increasing size while proportionally increasing the number of subdomains and thus also processor cores. For a numerical scalable method, the number of iterations should approach an asymptotic limit, and for a parallel scalable method, the solver time should stay close to constant. We will compare the performance of our new implementation of the three-level GDSW method with the performance of our older standard GDSW implementation.

10.4.1 Numerical Properties of the Three-Level GDSW Preconditioner

First, we present numerical results concerning the condition number of the preconditioned operator using the three-level preconditioner.

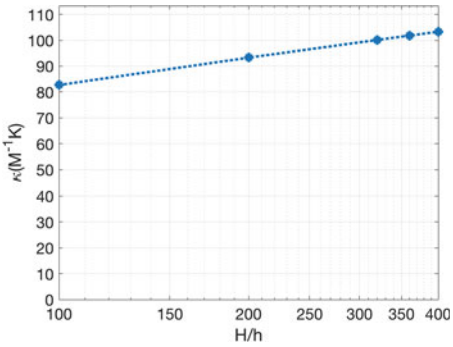
In view of the condition number bound (10.5) for the two-level preconditioner, we consider the three-level GDSW preconditioner while varying the subdomain overlap H/δ. The numerical results in Fig. 10.2 indicate a linear dependence on H/δ, consistent with the bound for the two-level method.

Next, we vary H/h, i.e., the number of degrees of freedom in each subdomain. Here, the results in Fig. 10.3 indicate a polylogarithmic dependence on H/h, again consistent with the bound for the two-level method.



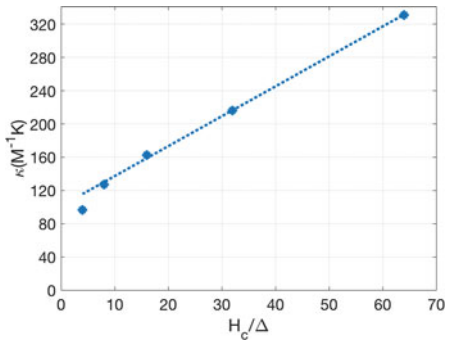
$1/H_c$	$1/H$	H/δ	Iter	$\kappa(M_{3GDSW}^{-1}K)$
32	128	40	84	99.37
32	128	30	82	93.32
32	128	20	78	86.81
32	128	12	77	81.27
32	128	8	76	78.34
32	128	6	75	76.78
32	128	5	76	75.96
32	128	4	76	75.08
32	128	3	76	74.13
32	128	2	75	72.97

Fig. 10.2 Number of PCG iterations *Iter* and estimated condition number $\kappa(M_{3GDSW}^{-1}K)$ using the three-level GDSW preconditioner for our Laplace model problem in two dimensions. Varying size of the overlap H/δ and fixed $1/H_c = 32$, $1/H = 128$, $H/h = 120$, and $H_c/\Delta = 4$; the dotted line is a least square fit of a linear function to the data; the problem has $(120 \times 128 + 1)^2 = 235\,960\,321$ degrees of freedom; we use $128^2 = 16\,384$ processor cores



$1/H_c$	$1/H$	H/h	Iter	$\kappa(M_{3GDSW}^{-1}K)$
16	64	100	75	82.68
16	64	200	81	93.29
16	64	320	86	100.13
16	64	360	87	101.81
16	64	400	89	103.31

Fig. 10.3 Number of PCG iterations *Iter* and estimated condition number $\kappa(M_{3GDSW}^{-1}K)$ using the three-level GDSW preconditioner for our Laplace model problem in two dimensions. Varying H/h and fixed $1/H_c = 16$, $1/H = 64$, $H_c/\Delta = 4$, and $H/\delta = 20$; semi-log plot; the largest problem has $(400 \times 64 + 1)^2 = 655\,411\,201$ degrees of freedom; the dotted line interpolates the data points; we use $64^2 = 4096$ processor cores



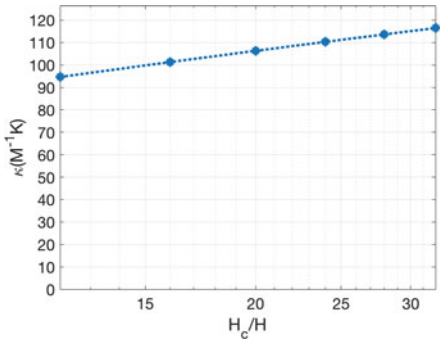
$1/H_c$	$1/H$	H_c/Δ	Iter	$\kappa(M_{3GDSW}^{-1}K)$
4	256	64	137	331.24
4	256	32	113	216.15
4	256	16	98	162.42
4	256	8	84	127.18
4	256	4	73	96.70

Fig. 10.4 Number of PCG iterations *Iter* and estimated condition number $\kappa(M_{3GDSW}^{-1}K)$ using the three-level GDSW preconditioner for our Laplace model problem in two dimensions. Varying subregion overlap H_c/Δ and fixed $1/H_c = 4$, $1/H = 256$, $H/h = 120$, and $H/\delta = 4$; the dotted line is a least square fit of a linear function; the problem has $(256 \times 120 + 1)^2 = 943\,779\,841$ degrees of freedom; we use $256^2 = 65\,536$ processor cores

We also vary the subregion overlap H_c/Δ. The numerical results are visualized in Fig. 10.4 and clearly indicate a linear dependence on H_c/Δ.

Last, we change the number of subdomains in a subregion, i.e., H_c/H. Here, the results in Fig. 10.5 point towards a polylogarithmic dependence on H_c/H.

Our numerical findings thus indicate that the standard bound generalizes to the subregion level. However, the theory corresponding to the numerical results presented here will be presented elsewhere.



$1/H_c$	H_c/H	$1/H$	Iter	$\kappa(M_{3GDSW}^{-1}K)$
8	12	96	77	94.69
8	16	128	80	101.31
8	20	160	83	106.31
8	24	192	84	110.30
8	28	224	86	113.62
8	32	256	87	116.45

Fig. 10.5 Number of PCG iterations *Iter* and estimated condition number $\kappa(M_{3GDSW}^{-1}K)$ using the three-level GDSW preconditioner for our Laplace model problem in two dimensions. Varying H_c/H and fixed $1/H_c = 8$, $H/h = 16$, $H_c/\Delta = 4$, and $H/\delta = 4$; semi-log plot; the dotted line interpolates the data points; the largest problem has $(256 \times 16 + 1)^2 = 16\,785\,409$ degrees of freedom and uses $256^2 = 65\,536$ processor cores

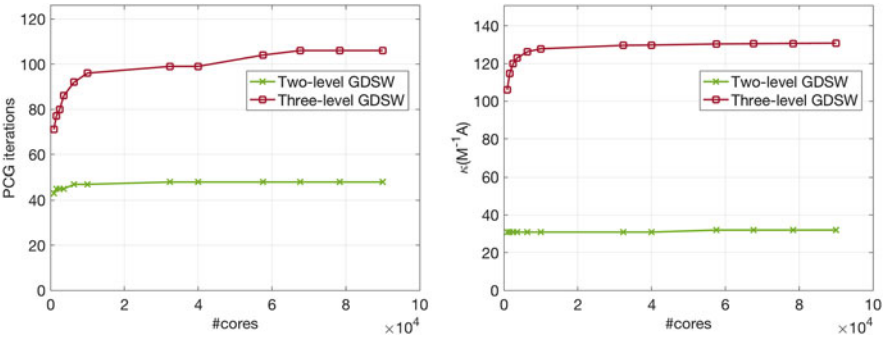


Fig. 10.6 Weak numerical scalability of the two- and three-level methods. Both methods are numerically scalable. See Table 10.1 for the data. (1) Number of PCG iterations for an increasing number of subdomains and processor cores (left). (2) Estimated condition number for an increasing number of subdomains and processor cores (right)

10.4.2 Numerical and Parallel Scalability of the Three-Level GDSW Preconditioner

It can be expected that replacing the direct coarse solver used in the two-level GDSW method [4, 5, 8, 9, 11] by a preconditioner for K_0 will increase the number of Krylov iterations in the three-level method. The results in Fig. 10.6 will show indeed that the condition number can be larger by a factor of more than four and the number of iterations is twice as high for the three-level method.

However, the three-level method can be expected to show a superior range of weak parallel scalability since the coarse matrix K_{00} in three-level GDSW is

Table 10.1 Data corresponding to Fig. 10.7

#Subdomains = #cores	Two-level GDSW					Three-level GDSW				
	Iter	$\kappa(M_{GDSW}^{-1}K)$	Solver time	Setup time	Krylov time	Iter	$\kappa(M_{3GDSW}^{-1}K)$	Solver time	Setup time	Krylov time
900	43	30.82	**44.55 s**	22.18 s	22.37 s	71	105.99	**59.15 s**	21.76 s	37.39 s
1600	45	30.83	**47.51 s**	23.00 s	24.49 s	77	114.73	**63.65 s**	22.79 s	40.85 s
2500	45	30.80	**49.44 s**	24.07 s	25.37 s	80	119.84	**66.13 s**	23.72 s	42.38 s
3600	45	30.84	**50.88 s**	24.37 s	26.51 s	86	122.94	**69.82 s**	23.68 s	46.13 s
6400	47	30.85	**56.59 s**	25.75 s	30.84 s	92	126.22	**73.54 s**	24.08 s	49.46 s
10,000	47	30.85	**69.46 s**	34.62 s	34.84 s	96	127.78	**76.23 s**	24.12 s	52.11 s
32,400	48	30.86	**102.61 s**	41.00 s	61.61 s	99	129.61	**85.27 s**	30.61 s	54.66 s
40,000	48	30.86	**117.26 s**	46.58 s	70.68 s	99	129.75	**87.56 s**	32.84 s	54.72 s
57,600	48	31.93	**146.02 s**	57.42 s	88.60 s	104	130.34	**90.24 s**	32.63 s	57.61 s
67,600	48	31.94	**172.10 s**	69.70 s	102.40 s	106	130.49	**92.91 s**	33.48 s	59.43 s
78,400	48	31.95	**199.08 s**	83.58 s	115.50 s	106	130.63	**94.61 s**	35.09 s	59.52 s
90,000	48	31.95	**220.71 s**	91.41 s	129.30 s	106	130.78	**96.88 s**	36.31 s	60.57 s

By *Iter*, we denote number of PCG iterations, and κ is the condition number of the preconditioned operator
The *Solver Time* is the sum of the *Setup Time* and *Krylov Time*. We have $H/h = 200$, $H/\delta = 20$, $H_c/H = 10$, and $H_c/\Delta = 10$. Also see Figs. 10.6 and 10.7

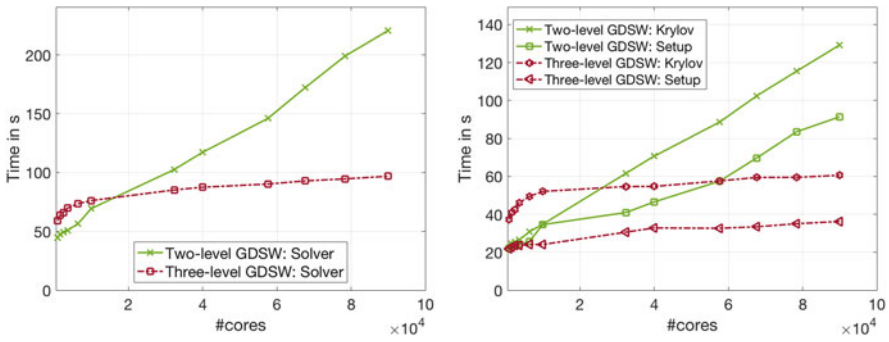


Fig. 10.7 Weak parallel scalability of the two- and three-level methods. See Table 10.1 for the data. (1) Solver Time for the two-level and three-level GDSW method (left). (2) Time for the setup of the preconditioner and for the Krylov iteration for two- and three-level GDSW (right)

significantly smaller than K_0 in the two-level method; see the columns *Size of* K_0 and *Size of* K_{00} in Table 10.2.

Weak Numerical and Parallel Scalability

In Figs. 10.6, 10.7, and Table 10.1, we now present the weak scalability of our implementation of the two-level and three-level GDSW preconditioners.

First, we note that both methods are numerically scalable in the number of iterations; see the column *Iter* in Table 10.1 and Fig. 10.6 (left). This is consistent with the estimated condition number; see $\kappa(M_{GDSW}^{-1}K)$ and $\kappa(M_{3GDSW}^{-1}K)$ in Table 10.1 and Fig. 10.6 (right).

Let us now consider the computing times. By *Solver Time*, we denote the time to solution, which is the sum of the time for the setup of the preconditioner, denoted *Setup Time*, and the time for the PCG iteration, which we denote *Krylov Time*. The *Setup Time* includes the factorizations of the matrices on the different levels by the MUMPS sparse direct solver.

It is clear from Fig. 10.6 and Table 10.1 that the numerical scalability in terms of the PCG iterations and of the condition number is better for the standard two-level method since it seems to approach earlier its asymptotic bound. Moreover, as already mentioned, the number of iterations is twice as high for the three-level method. Therefore, for a small number of cores, with respect to *Solver Time* the three-level method is about 30% slower than the traditional two-level method.

However, as the *Solver Time* in Fig. 10.7 and Table 10.1 shows, the three-level GDSW method exhibits a clearly superior weak parallel scalability compared to the two-level method. As a result, from 32,400 BG/Q cores on, the three-level method is faster than the two-level method. For the largest problem considered in Table 10.1 with 3.6 billion degrees of freedom, the three-level method is more than twice as fast (96.88s *Solver Time*) than the two-level method (220.71s *Solver Time*). The performance advantage of the three-level method

Table 10.2 Cost for solving the problem on the coarsest level, i.e., using K_0 in the standard two-level GDSW preconditioner and using K_{00} in the three-level GDSW preconditioner

	Two-level GDSW				Three-level GDSW			
#Subdomains = #Cores	Size of K_0	Factorization time	Forward-backward	Memory usage	Size of K_{00}	Factorization time	Forward-backward	Memory usage
900	2581	0.10 s	0.68 s	2 Mb	16	<0.01 s	0.03 s	1 Mb
1600	4641	0.18 s	1.29 s	3 Mb	33	<0.01 s	0.04 s	1 Mb
2500	7301	0.28 s	2.03 s	4 Mb	56	<0.01 s	0.05 s	1 Mb
3600	10,561	0.64 s	2.79 s	6 Mb	85	<0.01 s	0.07 s	1 Mb
6400	18,881	1.24 s	5.18 s	11 Mb	161	0.01 s	0.12 s	1 Mb
10,000	29,601	2.03 s	8.21 s	18 Mb	261	0.01 s	0.17 s	1 Mb
32,400	96,481	7.15 s	28.21 s	64 Mb	901	0.03 s	0.56 s	1 Mb
40,000	119,201	9.17 s	34.82 s	78 Mb	1121	0.04 s	0.69 s	1 Mb
57,600	171,841	13.30 s	51.07 s	113 Mb	1633	0.06 s	1.06 s	1 Mb
67,600	201,761	15.87 s	60.97 s	139 Mb	1925	0.07 s	1.28 s	2 Mb
78,400	234,081	18.93 s	70.10 s	158 Mb	2258	0.08 s	1.49 s	2 Mb
90,000	268,801	22.29 s	80.30 s	186 Mb	2581	0.09 s	1.70 s	2 Mb

Here, *Factorization Time* is the time the MUMPS sparse direct solver reports for the sum of symbolic and numerical factorization of K_0 and K_{00}, respectively; *Forward-Backward* is the sum of all times spent in forward-backward substitutions during the Krylov iteration; *Memory Usage* is the estimated amount of memory allocated by MUMPS during the factorization. See Table 10.1 for the corresponding *Solver Time*, *Setup Time* and *Krylov Time*. Also see Figs. 10.8, 10.9, and 10.10

can be expected to increase for a larger number of cores; also see Fig. 10.7 (left).

The Role of the Coarse Problem

The performance difference between both methods comes from the coarse problem. In Table 10.2, we therefore present the computational cost of the factorization and the forward-backward substitutions on the coarsest level of the two methods, i.e., for K_0^{-1} in the two-level GDSW and for K_{00}^{-1} in the three-level GDSW.

To understand the impact of the coarse problem cost on the scalability, let us consider the case of 90,000 cores and subdomains in Table 10.2. Here, the 102.59s (i.e., 80.30s for forward-backward substitutions and 22.29s for the factorization of K_0) for the two-level method represent 46.5% of the total *Solver Time* of 220.71s; see Table 10.1. These 102.59s compare to only 1.79s (i.e., 0.09s for forward-backward substitutions and 1.70s for the factorization of K_{00}) for the three-level method. In comparison, this is faster by a factor of more than 50; also see Fig. 10.8.

Note, that we here only discuss the computational work performed by the sparse direct solver on the coarsest level, i.e., the computational cost for the second level of the three-level method is not included; cf. (10.6). Moreover, MPI communication is not included. Both these aspects will however be considered below.

It is noteworthy that using MUMPS on the BG/Q architecture, in our numerical experiments, the forward-backward substitutions are always significantly more expensive than the factorizations, i.e., for 90,000 cores, in the two-level method, 22.29s are spent for the factorization of K_0, but 80.30s are spent in the 50 forward-backward substitutions with the factors of K_0 while performing the 48 PCG

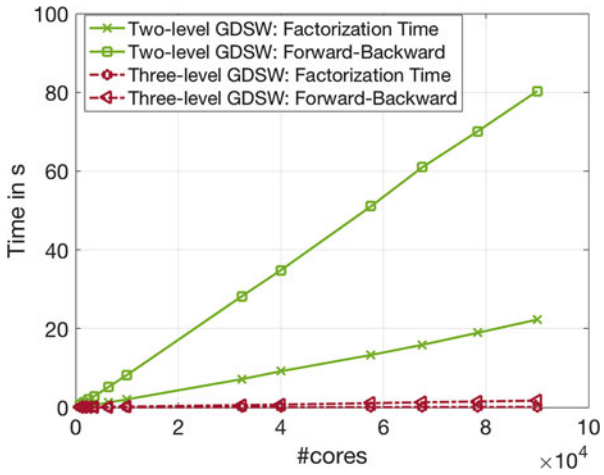


Fig. 10.8 Cost for solving the problem on the coarsest level, i.e., using K_0 in the standard two-level GDSW preconditioner and using K_{00} in the three-level GDSW preconditioner. See Table 10.2 for the data

iterations in Belos; see Table 10.2 and the remark on the Belos pseudo-block PCG implementation in Sect. 10.4.2.

A relatively slow forward-backward substitution, compared to the factorization phase, seems to be characteristic for MUMPS, which uses functions from scalapack and blacs as computational kernels. Other sparse direct solvers such as UMFPACK seem to profit from vendor-provided blas libraries especially for the forward-backward substitution. However, UMFPACK typically uses significantly more memory and is usually not faster than MUMPS in total; see also [9]. The slow forward-backward substitution of MUMPS explains the relatively large portion of the computing time spent in the Krylov iteration observed in Table 10.1.

We also see in Table 10.2 in the column *Memory Usage* that the amount of memory required by MUMPS for the factorization of K_0 grows only slightly faster than linear; also see Fig. 10.9. This is a result of the relatively high sparsity of both coarse matrices, K_0 as well as K_{00}, especially for structured decompositions. Both are sparser than standard low order finite element matrices; see also Fig. 10.10 here, the vertex basis functions do not couple to each other but only to the edge basis functions. Furthermore, the edge basis functions only have support on two subdomains.

Nevertheless, the memory usage will eventually limit the two-level method, since for 90,000 subdomains, MUMPS already uses 186 Mb of memory for the factorization of K_0; see Table 10.2 and Fig. 10.9. Since in our implementation K_0 is factorized on the master, the factors of K_0 as well as the factors of $\widetilde{K}_1$ have to fit into the memory available to MPI rank zero. To avoid out-of-memory errors, assigning a dedicated core or node to the coarse problem would be a possible improvement.

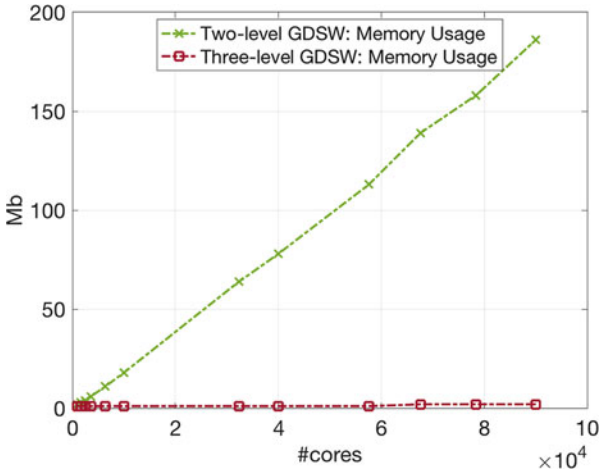


Fig. 10.9 Memory usage of the MUMPS direct solver for the factorization of the coarse matrix K_0 and K_{00} for the two-level and three-level GDSW method, respectively. See Table 10.2 for the data

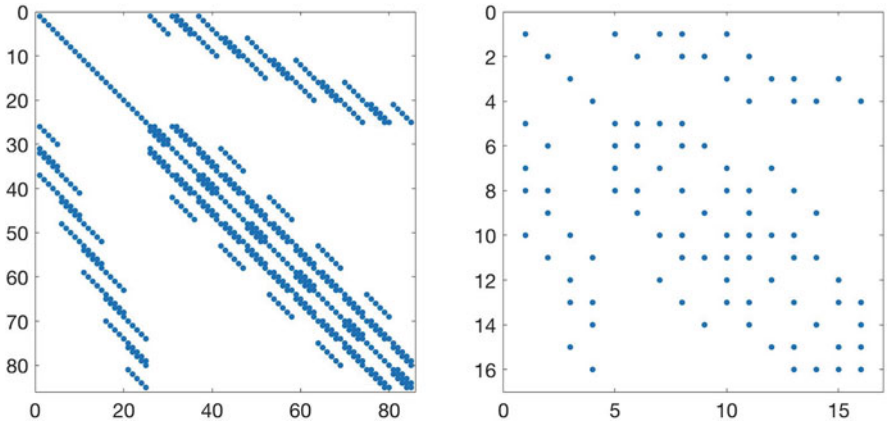


Fig. 10.10 Sparsity of the second level coarse matrix K_0 (left) and the third level coarse matrix K_{00} (right) for our Laplace model problem in two dimension with 36 subdomains and $H/h = 30$. For K_{00}, we have 9 subregions. The vertex basis functions are always numbered first, followed by the edge basis functions

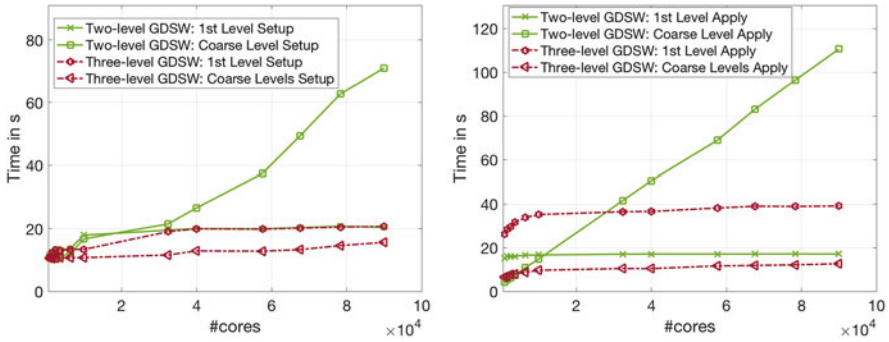


Fig. 10.11 Weak parallel scalability of the setup and application of the preconditioner. See Table 10.3 for the data. (1) Timers corresponding to the setup of the two- and three-level GDSW preconditioner, split into first and coarse level (left). (2) Timers corresponding to the application of the coarse operator and the included direct solves for two- and three-level GDSW (right)

Alternatively, MUMPS can also be used in parallel mode, on a subset of processors, as in [8, Section 4.6].

Weak Parallel Scalability of the First and Coarse Level(s)

Finally, in Fig. 10.11 and Table 10.3, we compare the cost for the setup of the first and coarse levels of the preconditioners as well as the cost for their application during the Krylov iteration.

MPI communication cost is now included, i.e., *Coarse Level Setup* and *Coarse Levels Setup* include the MPI communication to construct the coarse problem, notably for forming the coarse level operator using a triple matrix product; see [8,

Table 10.3 Timers for setup and application of the levels; see Fig. 10.11 for the visualization

#Subdomains = #Cores	Two-level GDSW				Three-level GDSW			
	Setup		Apply		Setup		Apply	
	First level	Coarse level	First level	Coarse level	First level	Coarse levels	First level	Coarse levels
900	11.62 s	10.56 s	15.14 s	4.29 s	11.16 s	10.60 s	26.12 s	6.48 s
1600	12.55 s	10.47 s	15.92 s	5.42 s	12.37 s	10.43 s	28.31 s	6.99 s
2500	13.49 s	10.58 s	15.89 s	6.25 s	13.16 s	10.56 s	29.47 s	7.69 s
3600	13.25 s	11.12 s	16.03 s	7.37 s	13.01 s	10.67 s	31.59 s	8.01 s
6400	13.66 s	12.09 s	16.54 s	10.82 s	13.37 s	10.71 s	33.76 s	8.72 s
10,000	17.91 s	16.71 s	16.67 s	14.81 s	13.37 s	10.75 s	35.18 s	9.64 s
32,400	19.50 s	21.50 s	17.03 s	41.52 s	19.01 s	11.60 s	36.45 s	10.46 s
40,000	19.98 s	26.60 s	17.06 s	50.56 s	19.94 s	12.90 s	36.57 s	10.41 s
57,600	19.92 s	37.50 s	17.06 s	68.88 s	19.83 s	12.80 s	38.16 s	11.65 s
67,600	20.30 s	49.40 s	17.10 s	83.01 s	20.18 s	13.30 s	38.96 s	11.87 s
78,400	20.78 s	62.80 s	17.10 s	96.45 s	20.49 s	14.60 s	38.87 s	12.14 s
90,000	20.41 s	71.00 s	17.09 s	110.60 s	20.71 s	15.60 s	39.07 s	12.66 s

Section 4.5]. Additionally, MPI communication is necessary to send the coarse matrices K_0 or K_{00} to a subset of processes. In this current paper, this subset is always the master. Then, hierarchical Epetra import and export data structures are built to allow the application of K_0^{-1} and K_{00}^{-1} to a distributed Epetra vector; here, intermediate data migration steps are used, in a hierarchical fashion, in order to avoid all-to-one communication; see [8, Section 4.7] for details. Moreover, gather and scatter steps (using Epetra import and export operations) are included in *1st Level Apply* and *Coarse Level Apply* for the two-level method, and in *1st Level Apply*, and *Coarse Levels Apply* for the three-level method.

Note that for the three-level method the *Coarse Levels Setup* and *Coarse Levels Apply* correspond to the sum of the cost on the second and third level. This implies that the computational work for $\widetilde{K}_{i0}^{-1}$ is now included. This is opposed to Table 10.2, where only K_{00}^{-1} was considered, which is part of the computational work performed on the third level.

Also note that the *Factorization Time* presented in Table 10.2 is included in the *Setup Time* of Table 10.3 for the two-level and three-level method. Likewise, the *Forward-Backward* Time in Table 10.2 is included in the *Apply Coarse Level(s)* Time of Table 10.3.

In Fig. 10.11 (left), we see that the setup cost of the first level is almost identical for the two- and three-level methods. This is no surprise as the implementations of the first level are essentially identical.

However, in Fig. 10.11 (right), we see that for the three-level method the application of the first level can be more than twice as expensive as for the two-level method. This is a result of the higher number of PCG iterations in the three-level method and impacts the total solver time because of the relatively slow forward-backward substitutions in MUMPS.

Consistent with our earlier findings, we see in Fig. 10.11 that the scalability of the two-level method suffers mainly from the setup and application of the coarse level, i.e., *Coarse Level Setup* and *Coarse Level Apply* are not scalable for the two-level GDSW preconditioner.

However, for the three-level method the setup of the second and third levels, i.e., *Coarse Levels Setup*, as well as the application of the second and third levels during the Krylov iteration, i.e., *Coarse Levels Apply*, are both parallel scalable in the range of processor cores considered here.

10.4.3 Conclusion

We have shown that the three-level GDSW preconditioner shows good weak scalability with a parallel efficiency of 61.0% considering the *Solver Time* when scaling from 900 to 90,000 processor cores. In comparison, the traditional two-level GDSW method shows a parallel efficiency of only 20.2%.

As a result, for 90,000 cores, the three-level GDSW method outperforms the standard two-level GDSW method by a factor of 2.3. Moreover, the three-level method can be expected to still scale well when the standard method will already be out of memory during the factorization of the coarse matrix K_0.

For even larger coarse problems than considered here, additional levels can be added, and a hybrid version, multiplicative between levels, can be used to reduce the number of iterations as in [9, Section 6.2.3]. Results for a three-level GDSW preconditioner in three dimensions are work in progress.

Acknowledgements This work was supported in part by the German Research Foundation (DFG) through the Priority Programme 1648 "Software for Exascale Computing" (SPPEXA) under grants RH 122/3-2 and KL 2094/4-2.

Also, this work is in part supported under grant RH 122/5-2.

The authors gratefully acknowledge the computing the Gauss Centre for Supercomputing e. V. (www.gauss-centre.eu) for providing computing time on the GCS Supercomputer JUQUEEN BG/Q supercomputer [19] at JSC Jülich. GCS is the alliance of the three national supercomputing centres HLRS (Universität Stuttgart), JSC (Forschungszentrum Jülich), and LRZ (Bayrische Akademie der Wissenschaften).

References

1. Amestoy, P.R., Duff, I.S., L'Excellent, J.-Y., Koster, J.: A fully asynchronous multifrontal solver using distributed dynamic scheduling. SIAM J. Matrix Anal. Appl. **23**(1), 15–41 (2001)
2. Badia, S., Martí n, A.F., Principe, J.: Multilevel balancing domain decomposition at extreme scales. SIAM J. Sci. Comput. **38**(1), C22–C52 (2016)
3. Bavier, E., Hoemmen, M., Rajamanickam, S., Thornquist, H., Amesos2 and Belos: direct and iterative solvers for large sparse linear systems. Sci. Program. **20**(3), 241–255 (2012)
4. Dohrmann, C.R., Klawonn, A., Widlund, O.B.: Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. SIAM J. Numer. Anal. **46**(4), 2153–2168 (2008)
5. Dohrmann, C.R., Klawonn, A., Widlund, O.B.: A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In: Domain Decomposition Methods in Science and Engineering XVII. Lecture Notes in Computational Science and Engineering, vol. 60, pp. 247–254. Springer, Berlin (2008)
6. Dohrmann, C.R., Widlund, O.B.: On the design of small coarse spaces for domain decomposition algorithms. SIAM J. Sci. Comput. **39**(4), A1466–A1488 (2017)
7. Hackbusch, W.: Multigrid Methods and Applications. Springer Series in Computational Mathematics, vol. 4. Springer, Berlin (1985)
8. Heinlein, A.: Parallel Overlapping Schwarz Preconditioners and Multiscale Discretizations with Applications to Fluid-Structure Interaction and Highly Heterogeneous Problems. PhD thesis, Universität zu Köln (2016)
9. Heinlein, A., Klawonn, A., Rheinbach, O.: A parallel implementation of a two-level overlapping Schwarz method with energy-minimizing coarse space based on Trilinos. SIAM J. Sci. Comput. **38**(6), C713–C747 (2016)
10. Heinlein, A., Klawonn, A., Rheinbach, O.: Parallel Two-Level Overlapping Schwarz Methods in Fluid-Structure Interaction, pp. 521–530. Springer International Publishing, Cham (2016)
11. Heinlein, A., Klawonn, A., Rheinbach, O., Widlund, O.: Improving the parallel performance of overlapping Schwarz methods by using a smaller energy minimizing coarse space. 2017. Accepted for publication in the Proceedings of the 24rd International Conference on Domain Decomposition Methods. Springer Lecture Notes in Computational Science and Engineering

12. Heinlein, A., Klawonn, A., Rajamanickam, S., Rheinbach, O.: FROSch: a fast and robust overlapping Schwarz domain decomposition preconditioner based on Xpetra in Trilinos. Technischer Bericht, Universität zu Köln, November 2018. https://kups.ub.uni-koeln.de/9018/
13. Heroux, M.A., Bartlett, R.A., Howle, V.E., Hoekstra, R.J., Hu, J.J., Kolda, T.G., Lehoucq, R.B., Long, K.R., Pawlowski, R.P., Phipps, E.T., Salinger, A.G., Thornquist, H.K., Tuminaro, R.S., Willenbring, J.M., Williams, A., Stanley, K.S.: An overview of the Trilinos Project. ACM Trans. Math. Softw. **31**(3), 397–423 (2005)
14. Kong, F., Cai, X.-C.: A highly scalable multilevel Schwarz method with boundary geometry preserving coarse spaces for 3D elasticity problems on domains with complex geometry. SIAM J. Sci. Comput. **38**(2), C73–C95 (2016)
15. Mandel, J., Sousedík, B., Dohrmann, C.R.: Multispace and multilevel BDDC. Computing **83**(2–3), 55–85 (2008)
16. Rajamanickam, S., Boman, E.G., Heroux, M.A.: Shylu: a hybrid-hybrid solver for multicore platforms. In: 2012 IEEE 26th International Parallel and Distributed Processing Symposium, May 2012, pp. 631–643 (2012)
17. Sala, M., Stanley, K.S, Heroux, M.A: On the design of interfaces to sparse direct solvers. ACM Trans. Math. Softw. **34**(2), 9:1–9:22 (2008)
18. Scacchi, S.: A hybrid multilevel Schwarz method for the bidomain model. Comput. Methods Appl. Mech. Eng. **197**(45–48), 4051–4061 (2008)
19. Stephan, M., Docter, J.: JUQUEEN: IBM Blue Gene/Q® Supercomputer System at the Julich Supercomputing Centre. Journal of Large-Scale Research Facilities **1**, A1 (2015)
20. Toselli, A., Widlund, O.: Domain Decomposition Methods—Algorithms and Theory. Springer Series in Computational Mathematics, vol. 34. Springer, Berlin (2005)
21. Trilinos public git repository. Web, 2018. https://github.com/trilinos/trilinos
22. Tu, X.: Three-level BDDC in two dimensions. Int. J. Numer. Methods Eng. **69**(1), 33–59 (2007)

Chapter 11
A Parallel Multigrid Solver for Multi-Patch Isogeometric Analysis

Christoph Hofer and Stefan Takacs

Abstract Isogeometric Analysis (IgA) is a framework for setting up spline-based discretizations of partial differential equations, which has been introduced around a decade ago and has gained much attention since then. If large spline degrees are considered, one obtains the approximation power of a high-order method, but the number of degrees of freedom behaves like for a low-order method. One important ingredient to use a discretization with large spline degree, is a robust and preferably parallelizable solver. While numerical evidence shows that multigrid solvers with standard smoothers (like Gauss Seidel) does not perform well if the spline degree is increased, the multigrid solvers proposed by the authors and their co-workers proved to behave optimal both in the grid size and the spline degree. In the present paper, the authors want to show that those solvers are parallelizable and that they scale well in a parallel environment.

11.1 Introduction

Isogeometric Analysis (IgA) was originally introduced in the seminal paper [9], aiming to unite the worlds of computer aided design (CAD) and finite element (FEM) simulation. From a technical point of view, it is a framework for setting up spline-based discretizations of partial differential equations. The key idea is that the spline space is typically first defined on the unit square or the unit cube and then mapped to the computational domain using one global geometry function. More complicated domains cannot be represented by just one such geometry function. Instead, the computational domain is decomposed into patches, where each of them

C. Hofer
Doctoral Program Computational Mathematics, University Linz, Linz, Austria
e-mail: christoph.hofer@dkcm.jku.at

S. Takacs (✉)
RICAM, Austrian Academy of Sciences, Linz, Austria
e-mail: stefan.takacs@ricam.oeaw.ac.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_11

is represented by its own geometry function. This is called the *multi-patch case*, in contrast to the *single-patch case*.

As a next step, the linear system resulting from the discretization of the PDE has to be solved. This might be challenging as the condition number of the linear system grows exponentially with the spline degree, where high spline degrees might be desired because of their superior approximation power.

While in early IgA literature, the dependence of methods on the spline degree has not been considered, in the last few years robustness in the spline degree has gained increasing interest. Several (almost) robust approaches or approaches with a mild dependence on the spline degree have been proposed, on the one side for the single-patch case, cf. [1, 4, 7, 8, 11] and references therein, and on the other side as approaches aiming to combine patch-local solvers to a global solver, cf. [2, 3, 10, 12] and references therein.

In [13], we have considered a slightly different approach: We do not aim to combine patch-local solvers to a global solver, but to combine patch-local smoothers to a global smoother which is used within a global multigrid solver. In the present paper, we give some additional remarks on an efficient implementation of the multigrid method, comment on its parallelization and give numerical results.

This paper is organized as follows. First, the model problem and the discretization are discussed in Sect. 11.2. Then, in Sect. 11.3, we recall the formulation of the multigrid solver. Its parallelization is discussed in the following Sect. 11.4. In Sect. 11.5, we give the results of numerical experiments and draw conclusions.

11.2 Model Problem and Isogeometric Discretization

Let $\Omega \subset \mathbb{R}^d$ with $d \in \{2, 3\}$ be a bounded computational domain with Lipschitz boundary. We consider a standard *Poisson model problem*

$$-\Delta u = f \quad \text{in} \quad \Omega\,, \quad u = 0 \quad \text{on} \quad \Gamma_D \qquad \text{and} \qquad \frac{\partial u}{\partial n} = 0 \quad \text{on} \quad \Gamma_N\,,$$

where Γ_D is a subset of $\partial\Omega$ with positive measure and $\Gamma_N := \partial\Omega \backslash \Gamma_D$. The model problem reads in variational form as follows. Given $f \in L_2(\Omega)$, find $u \in H^1_{0,D}(\Omega)$ such that

$$(\nabla u, \nabla v)_{L_2(\Omega)} = (f, v)_{L_2(\Omega)} \qquad \text{for all} \quad v \in H^1_{0,D}(\Omega)\,. \tag{11.1}$$

Here and in what follows, $L_2(\Omega)$ and $H^1(\Omega)$ are the standard Lebesgue and Sobolev spaces with standard norms and $H^1_{0,D}(\Omega) := \{u \in H^1(\Omega) \,:\, u|_{\Gamma_D} = 0\}$.

We preform a standard isogeometric multi-patch discretization as it has been specified in [13]. In the present paper, we try to keep the explanation short and give only an overview. We assume that the computational domain Ω is composed

of K patches Ω_k such that

$$\overline{\Omega} = \bigcup_{k=1}^{K} \overline{\Omega_k} \quad \text{and} \quad \Omega_k \cap \Omega_l = \emptyset \text{ for any } k \neq l \,, \tag{11.2}$$

where each patch Ω_k is a bounded and open domain. We assume that the patches are fully matching, i.e., the intersections $\overline{\Omega_k} \cap \overline{\Omega_l}$ are either empty, common vertices, common edges or common faces. Any of the patches is parameterized by a bijective geometry function

$$\mathbf{G}_k : \widehat{\Omega} := (0,1)^d \to \Omega_k := \mathbf{G}_k(\widehat{\Omega}) \subset \mathbb{R}^d \,.$$

Before we define set of trial functions $V_\ell \subset H^1_{0,D}(\Omega)$, we introduce discretizations living on the parameter domain $\widehat{\Omega}$. Let

$$S_{p,h}(0,1) := \left\{ u \in C^{p-1}(0,1) \ : \ u|_{[hi,h(i+1)]} \text{ is a polynomial of degree } p \ , \ \forall_{i=1,\dots,n} \right\}$$

be the space of univariate splines of maximum smoothness and the space $S_{p,h}(\widehat{\Omega}) := S_{p,h}(0,1) \otimes \cdots \otimes S_{p,h}(0,1)$ be the corresponding tensor-product spline space. The grid size h and the spline degree p might be different for any patch and for any spacial direction; for simplicity, we do not express that in the notation. Based on the discretization living on the parameter domain $\widehat{\Omega}$, we define the function space V_ℓ of isogeometric functions living on the physical domain Ω as follows:

$$V_\ell := \{ u \in C^0(\Omega) \ : \ u \circ \mathbf{G}_k \in S_{p,h_\ell}(\widehat{\Omega}) \} \,. \tag{11.3}$$

We assume to have a *fully matching discretization*, which means that the discretizations agree on the interfaces. A more formal definition of the basis and its discretization is given in [13, Sec. 2]. In Fig. 11.1, a fully matching discretization is depicted, where each node represents one basis function and therefore one degree of freedom (dof). Note that any of the basis function whose associated node lies on one

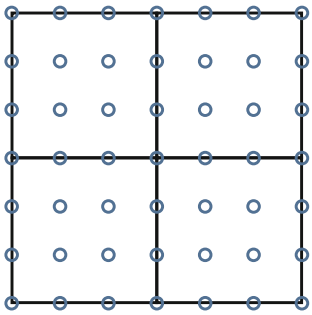


Fig. 11.1 Fully matching discretization

patch, vanishes outside of that patch. Any of the basis functions whose associated node lies within one edge, vanishes outside the union of the edge and the adjacent patches. Finally, any of the basis functions whose associated node coincides with one vertex, vanishes outside the union of that vertex and the adjacent edges and patches. The behavior in three dimensions is completely analogous.

The Galerkin principle yields the following discretized variational problem. Find $u \in V_\ell$ such that

$$a(u, v) = (f, v)_{L_2(\Omega)} \qquad \text{for all } v \in V_\ell \,, \tag{11.4}$$

where

$$a(u, v) := (\nabla u, \nabla v)_{L_2(\Omega)} = \sum_{k=1}^{K} \underbrace{(|\det J_{\mathbf{G}_k}| J_{\mathbf{G}_k}^{-\top} J_{\mathbf{G}_k}^{-1} \nabla \widehat{u}_k, \nabla \widehat{v}_k)_{L^2(\widehat{\Omega})}}_{a_k(u, v) :=} \tag{11.5}$$

for $\widehat{u}_k := u \circ \mathbf{G}_k \in S_{p,h_\ell}(\widehat{\Omega})$ and $\widehat{v}_k := v \circ \mathbf{G}_k \in S_{p,h_\ell}(\widehat{\Omega})$ and where $J_{\mathbf{G}_k}$ is the Jacobian of the geometry map. Using the chosen basis, we obtain a matrix-vector formulation of the discretized problem, which reads as follows. Find $\underline{u} \in \mathbb{R}^N$ such that

$$A_\ell \, \underline{u} = \underline{f} \,. \tag{11.6}$$

Allowing constants that depend on the geometry function, we obtain that the matrix A_ℓ is spectrally equivalent to the matrix $\widehat{A}_\ell$, which discretizes the bilinear form

$$\widehat{a}\,(u, v) := \sum_{k=1}^{K} (\nabla \widehat{u}_k, \nabla \widehat{v}_k)_{L^2(\widehat{\Omega})} \,,$$

where, again, $\widehat{u}_k := u \circ \mathbf{G}_k$ and $\widehat{v}_k := v \circ \mathbf{G}_k$.

11.3 The Multigrid Solver and Its Extension to Three Dimensions

We employ the multigrid solver based on a hierarchy of grids for grid levels $\ell = 0, \ldots, L$, obtained by uniform refinement. Throughout the grid hierarchy, the spline degree p and the corresponding smoothness is kept unchanged. This yields nested spaces: $V_0 \subset V_1 \subset \cdots \subset V_L \subset H^1_{0,D}(\Omega)$, which allows to use the canonical embedding $V_{\ell-1} \to V_\ell$ for the multigrid method; its matrix representation is denoted by P_ℓ. Following the usual pattern, we use its transpose $P_\ell^\top$ as restriction.

One *multigrid cycle* on some grid level ℓ consists of the following steps.

- First, ν *pre-smoothing steps* are applied, where each reads as follows:

$$\underline{u} \leftarrow \underline{u} + \tau L_\ell^{-1}(\underline{f} - A_\ell \underline{u}) \,. \tag{11.7}$$

  The choice of the smoothing operator L_ℓ^{-1} and the damping parameter τ are discussed below.
- Then, the *coarse grid correction* is performed:

$$\underline{u} \leftarrow \underline{u} + \tau P_\ell A_{\ell-1}^{-1} P_\ell^\top (\underline{f} - A_\ell \underline{u}) \,,$$

  where for $\ell > 1$, the application $A_{\ell-1}^{-1}$ is replaced by $\mu = 1$ (V-cycle) or $\mu = 2$ (W-cycle) recursive applications of the multigrid method on the coarser grid level.
- Finally, again ν *post-smoothing steps* (11.7) are applied.

As smoother, an additive Schwarz type combination

$$L_\ell^{-1} := \sum_T P_{\ell,T} L_{\ell,T}^{-1} P_{\ell,T}^\top \tag{11.8}$$

of local smoothing operators $L_{\ell,T}^{-1}$ is proposed, where the dofs are collected based on separating the domain into *pieces*: patches, vertices, edges and, in three dimensions, faces. Here, each dof (basis function) is assigned to exactly one of these pieces as proposed in [13]: Each basis function which is non-vanishing on a vertex, is assigned to that vertex. All other basis functions, which are non-vanishing on an edge, are assigned to that edge. All other basis functions, which are non-vanishing on a face, are assigned to that face. All other basis functions are assigned to the patch in which their support is contained. For an illustration, cf. Fig. 11.2.

Certainly, based on such a one-by-one splitting, the matrix $P_{\ell,T}$ is nothing but a full rank $N \times N_T$ indicator matrix representing the canonical embedding, where N_T is the number of dofs assigned to the piece T and N is the total number of dofs.

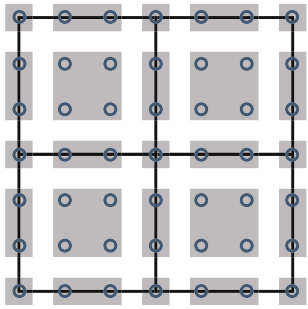


Fig. 11.2 Decomposition into pieces serving as subspaces for the additive Schwarz method

The local smoothing operators are chosen as follows.

- For the patch-interiors, the subspace corrected mass smoother as proposed in [7] is chosen as smoothing operator $L_{\ell,T}^{-1}$.
- For the edges and vertices, in [13] direct solvers have been proposed as smoothers, i.e., $L_{\ell,T} := P_{\ell,T}^{\top} A_\ell P_{\ell,T}$ is the restriction of the matrix A_ℓ to the edge or vertex. To avoid unnecessary communication, we choose an approximation which can be computed directly. Using [13, Lemma 4.1] and [13, eq. (4.16)], we obtain that the restriction of A_ℓ to an edge is spectrally equivalent to

$$L_{\ell,T} := \left(\frac{h_\ell}{p}\right)^{d-1} \mathrm{K}_\ell + \left(\frac{h_\ell}{p}\right)^{d-3} \mathrm{M}_\ell \,,$$

  where K_ℓ and M_ℓ are the corresponding univariate stiffness and mass matrices. Analogously, its restriction to a vertex is a constant in the order of

$$L_{\ell,T} := \left(\frac{h_\ell}{p}\right)^{d-2} .$$

- Three dimensional problems have not been considered in [13], so we have to discuss how to choose the local smoothers for faces. If, as for the edges and vertices, again a direct solver was applied, the overall computational costs would not be optimal anymore. So, again, observe that the the restriction of A_ℓ to a face is spectrally equivalent to

$$L_{\ell,T}^{*} := \left(\frac{h_\ell}{p}\right)^{d-2} \mathscr{K}_\ell + \left(\frac{h_\ell}{p}\right)^{d-4} \mathscr{M}_\ell \,,$$

  where and $\mathscr{K}_\ell = \mathrm{K}_\ell \otimes \mathrm{M}_\ell + \mathrm{M}_\ell \otimes \mathrm{K}_\ell$ and $\mathscr{M}_\ell = \mathrm{M}_\ell \otimes \mathrm{M}_\ell$ are the corresponding stiffness and mass matrices on the face. For $d = 3$, we obtain

$$L_{\ell,T}^{*} = \frac{h_\ell}{p}\left(\mathscr{K}_\ell + \frac{p^2}{h_\ell^2}\mathscr{M}_\ell\right) .$$

  Here, analogously to the case of the patch-interiors, the subspace corrected mass smoother is used. Note that the subspace corrected mass smoother is set up such that it bounds the stiffness matrix $\mathscr{K}_\ell$ from above, cf. [7, eq. (11)]. In the present paper, besides a trivial scaling, the stiffness matrix $\mathscr{K}_\ell$ is augmented by $p^2 h_\ell^{-2}$ times the mass matrix $\mathscr{M}_\ell$. So, we have also to augment the local contributions for the subspace corrected mass smoother, cf. the matrices L_α in [7, Sec. 4.2], in the same way.

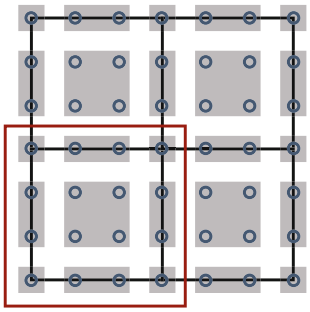


Fig. 11.3 The distribution of the dofs to the processors

11.4 The Parallelization of the Multigrid Solver

The parallelization of the multigrid solver follows the approach presented in [5]. We use MPI,[1] so each processor executes independently the whole algorithm with its local data until communication is explicitly requested.

We assign each of the patches to one of the processors. So, that processor holds the values of all dofs that belong to that patch including its interfaces, cf. Fig. 11.3. This means that the dofs on the interfaces might be assigned to more than one processor.

A vector that occurs in the algorithm, say $\underline{w}$, is stored either in accumulated form (Type I) or distributed form (Type II). We say that a vector is stored in accumulated form if each of the processors holds those parts of the global vector which correspond to the dofs assigned to the processor. We denote such vectors by $\underline{w}_{acc}$. We say that a vector is stored in distributed form if the global vector is the sum of the contributions of all the processors. Such vectors are denoted by $\underline{w}_{dist}$. Again, the processor-local contributions of the distributed vectors are supported only the patches assigned to the processor including their interfaces.

Note that only certain kinds of operations make sense; so we can add accumulated and distributed vectors only to vectors of the same type:

$$\underline{u}_{acc} + \underline{v}_{acc} \to \underline{w}_{acc} \quad \text{and} \quad \underline{u}_{dist} + \underline{v}_{dist} \to \underline{w}_{dist} \,,$$

cf. [5, Sec. 5.3]. As the multi-patch setting is equivalent to a standard approach of non-overlapping domain decomposition, the overall stiffness matrix is assembled on a per-patch basis, i.e., the bilinear forms a_k from (11.5) are evaluated separately yielding matrices $A_{\ell,k}$. Consequently, the global stiffness matrix A_ℓ is the sum of the local contributions $A_{\ell,k}$. This means that the matrix A_ℓ is stored in distributed form, which yields the following mapping type:

$$A_\ell \, \underline{u}_{acc} \to \underline{w}_{dist} \,,$$

[1] Message Passing Interface, see http://mpi-forum.org/.

i.e., A_ℓ can be applied to accumulated vectors and the result of the operation is distributed, cf. [5, Sec. 5.4.1].

Similar to [5, Sec. 7.2.2], the inter-grid transfer operators satisfy

$$P_\ell \, \underline{u}_{acc} \to \underline{w}_{acc} \quad \text{and} \quad P_\ell^\top \, \underline{u}_{dist} \to \underline{w}_{dist}$$

because the prolongation operator has a block-triangular structure as in [5, eq. (5.9)] and the restriction operator has a block-triangular structure as in [5, eq. (5.10)]. The block-triangular structure is obtained because the following statements hold true:

- On each vertex, the prolonged value $\underline{w}_{acc}$ coincides with the coarse-grid value $\underline{u}_{acc}$ of the same vertex.
- On each edge, the prolonged values $\underline{w}_{acc}$ only depend on the coarse-grid values $\underline{u}_{acc}$ on the same edge and on the adjacent vertices.
- On each patch-interior, the prolonged values $\underline{w}_{acc}$ only depend on the coarse-grid values $\underline{u}_{acc}$ on the same patch-interior and on the adjacent edges and vertices.

For three dimensions, completely analogous statements hold true.

The global operator L_ℓ^{-1} is block-diagonal, where each block corresponds to one piece. Note that by construction each piece belongs as a whole to one processor or is shared as a whole by the same processors, so it satisfies both the conditions of [5, eq. (5.9)] and [5, eq. (5.10)]. This shows

$$L_\ell^{-1} \, \underline{u}_{acc} \to \underline{w}_{acc} \quad \text{and} \quad L_\ell^{-1} \, \underline{u}_{dist} \to \underline{w}_{dist} \, ,$$

i.e., this operator can be applied both to distributed and accumulated vectors and it preserves the type of the vector.

As in any iterative solver, we need to accumulate the vectors of interest in each iterate. This we denote using the symbol Σ, which maps as follows:

$$\Sigma \, \underline{u}_{dist} \to \underline{w}_{acc}. \tag{11.9}$$

We note that only a communication between the processors holding neighboring patches is required in order to perform (11.9).

Only the coarsest grid level $\ell = 0$ needs some special treatment. Since the focus of the present paper is set on parallelizing the multigrid solver without changing its mathematical meaning, we perform an exact global solve on the coarsest grid level. This seems to be acceptable as it is done only for the coarsest grid level. So, we are required to communicate the stiffness matrix between all processors such that every processor holds a global stiffness matrix. We set up a direct solver A_0^{-1} for this global stiffness matrix, so its application is perform in the following way

$$\chi_{glob} \, A_0^{-1} \, \Sigma_{glob} \; \underline{u}_{acc} \to \underline{w}_{acc} \, ,$$

where Σ_{glob} denotes the accumulation of vectors where each processor obtains the global vector and χ_{glob} is the restriction of the global vector to the patches assigned to the processor. The latter involves only discarding unnecessary data. We obtain

$$\Sigma_{glob}\, \underline{u}_{dist} \to \underline{w}_{glob} \quad \text{and} \quad \chi_{glob}\, \underline{u}_{glob} \to \underline{w}_{acc}\,.$$

Overall, the parallel multigrid solver looks as follows:

Algorithm 1: Parallel multigrid solver

procedure MULTIGRID$(\ell, \underline{u}_{acc}, \underline{f}_{dist})$
- **for** $i = 1, \ldots, \nu$ **do** ▷ Pre-smoothing
  - $\underline{u}_{acc} \leftarrow \underline{u}_{acc} + \tau\, \Sigma L_\ell^{-1}(\underline{f}_{dist} - A_\ell \underline{u}_{acc})$
- $\underline{r}_{dist} \leftarrow P_\ell^\top(\underline{f}_{dist} - A_\ell \underline{u}_{acc})$
- **if** $\ell = 1$ **then** ▷ Coarse-grid correction
  - $\underline{p}_{acc} \leftarrow \chi_{glob}\, A_0^{-1}\, \Sigma_{glob}\, \underline{r}_{dist}$ ▷ Exact solver for coarsest grid level
- **else**
  - $\underline{p}_{acc} \leftarrow 0$
  - **for** $i = 1, \ldots, \mu$ **do** ▷ $\mu = 1$ is V-cycle; $\mu = 2$ is W-cycle
    - $\underline{p}_{acc} \leftarrow$ MULTIGRID$(\ell - 1, \underline{p}_{acc}, \underline{r}_{dist})$
- $\underline{u}_{acc} \leftarrow \underline{u}_{acc} + P_\ell \underline{p}_{acc}$
- **for** $i = 1, \ldots, \nu$ **do** ▷ Post-smoothing
  - $\underline{u}_{acc} \leftarrow \underline{u}_{acc} + \tau\, \Sigma L_\ell^{-1}(\underline{f}_{dist} - A_\ell \underline{u}_{acc})$
- **return** $\underline{u}_{acc}$

We use our multigrid algorithm as a preconditioner for a standard parallel preconditioned conjugate gradient (PCG) solver. Note that the multigrid preconditioner already takes a distributed residual and returns an accumulated update. So, the preconditioned conjugate gradient solver only needs to accumulate data in order to compute the required scalar products accordingly, cf. [5, Sec. 6.3.1].

11.5 Numerical Experiments

In this section, we present numerical experiments concerning the parallelization of the multigrid solver. The solver was implemented in C++ based on the G+Smo library [6] and, as already mentioned, the parallelization is performed

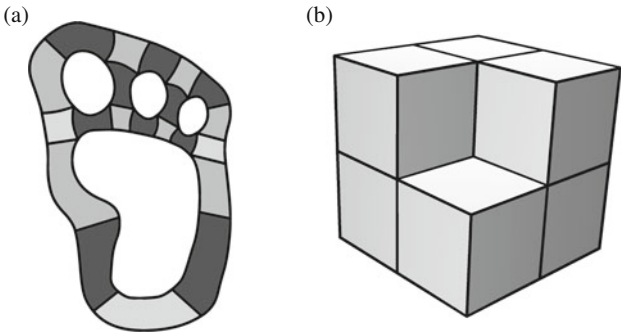


Fig. 11.4 The computational domains. (**a**) The Yeti footprint. (**b**) The Fichera corner

using MPI. All numerical experiments have been done using the HPC Cluster RADON1.[2]

We present timings for setup, assembling and solving. The setup costs include

1. the costs of the setup of the dof-mappers, which describe the relation between the local dof-indices and the global dof-indices,
2. the costs of the grid refinement and the setup of the inter-grid transfer matrices,
3. the costs of the setup of the piece-local smoothers and
4. the costs of the setup of the coarse-grid solver.

Here, our implementation of item 1 requires that each processor knows about the indexing of the global dofs. Also for item 4, the information on all dofs is required, however only on the coarsest grid level. The costs which are typically dominant, i.e., those for assembling and for solving, are presented separately. It is important to note that assembling does not require the any kind of communication between the processors. So its parallelization is trivial. The communication, which is required for the solving phase, is discussed in detail in Sect. 11.4.

We have performed the numerical experiments for two and three dimensions. As two dimensional domain, we use the Yeti footprint (Fig. 11.4a), which has already been considered in [13] and which is also a popular domain for the IETI-DP method, cf. [10]. As three dimensional domain, we consider the Fichera corner (Fig. 11.4b). This domain is often considered as extension of the L-shaped domain to three dimensions; the corresponding numerical experiments show that the proposed method can also be applied to domains without full elliptic regularity.

[2]We use up to 32 out of 68 available nodes, each equipped with 2x Xeon E5-2630v3 "Haswell" CPU (8 cores, 2.4 GHz, 20 MB cache) and 128 GB RAM. More information is available at https://www.ricam.oeaw.ac.at/hpc/.

11.5.1 The Yeti Footprint (2D)

On the Yeti footprint, we solve the model problem

$$\begin{aligned} -\Delta u &= 50\pi^2 \sin(5\pi\ x) \sin(5\pi\ y) && \text{in } \Omega\ , \\ u &= 0 && \text{on } \Gamma_D\ , \\ \frac{\partial}{\partial n} u &= 0 && \text{on } \Gamma_N\ , \end{aligned}$$

where Γ_D is the outer boundary and Γ_N are the four inner boundaries.

The Yeti footprint consists of 21 patches, which can be seen in Fig. 11.4a. Since we need sufficiently many patches for parallelization, we first split each patch uniformly into 16 patches, so we obtain in total $K = 336$ patches. We solve the problem with a conjugate gradient solver, preconditioned with one V-cycle of the multigrid method. We perform $1 + 1$ smoothing steps of the proposed smoother. The damping parameter and the scaling parameter (in the subspace corrected mass smoother) are chosen as in [13], i.e., $\tau = 0.25$ and $\sigma = \frac{1}{0.2} h_\ell^{-2}$.

In Table 11.1, we report on the number of iterations required to reach the desired relative accuracy goal of 10^{-8}. Here, ℓ represents the number of refinement levels and p the spline degree. On the coarsest grid level ($\ell = 0$), the patch-local discretization only consists of global polynomials, i.e., each patch is one element of the discretization. Refinement is done by uniformly refining the patch-local grids, keeping the number of patches unchanged. We observe, as in [13], that the number of iterates is quite robust in the grid size and in the spline degree. The presented numbers have been computed with the serial code. The number of iterates is supposed to be the same if parallelization is applied; however due to some small numerical instabilities, in some cases the parallel code needs one additional iteration (but never more than that). Similar iteration counts are obtained for the W-cycle.

In Table 11.2, we present the strong scaling results. We fix the grid level ℓ and the spline degree p to two typical values. For $\ell = 7$ and $p = 4$, we have 5 768 189 dofs and the corresponding stiffness matrix has $4.6\ 10^8$ non-zero entries. For the case $\ell = 7$ and $p = 8$, the number of dofs increases slightly to 6 125 757, but the stiffness matrix has already $1.8\ 10^9$ non-zero entries. In the first two rows, we compare the costs of the serial code and the parallel code. Here, we obtain that the parallel code is slightly slower during the solving phase which is mainly due to the

Table 11.1 Iteration counts for Yeti footprint, $K = 336$

	p						
ℓ	2	3	4	5	6	7	8
4	45	42	37	33	31	28	25
5	48	44	40	36	33	30	27
6	50	44	41	36	35	33	27
7	51	45	42	37	36	34	28

Table 11.2 Strong scaling behavior for Yeti footprint

	$\ell = 7$, $p = 4$, $K = 336$						$\ell = 7$, $p = 8$, $K = 336$					
	Setup		Assembling		Solving		Setup		Assembling		Solving	
# Proc.	t	s	t	s	t	s	t	s	t	s	t	s
(serial)	217.1	–	522.1	–	4929.5	–	–	–	–	–	–	–
1	220.0	1	520.0	1	5125.9	1	549.0	1	8230.9	1	4729.8	1
2	80.1	2.7	263.8	1.9	1367.6	3.7	225.4	2.4	4158.5	1.9	1250.8	3.7
4	35.3	6.2	131.8	3.9	399.1	12.8	117.0	4.6	2098.5	3.9	409.9	11.5
8	17.1	12.8	66.0	7.8	109.6	46.7	53.9	10.1	1055.9	7.8	140.2	33.7
16	10.7	20.5	33.9	15.3	40.7	125.9	30.0	18.3	543.4	15.1	59.2	79.9
32	8.0	27.5	17.4	29.8	17.1	299.7	17.7	31.0	275.1	29.9	26.8	176.4
64	7.2	30.5	9.4	55.3	10.6	483.5	12.9	42.5	149.7	54.9	13.7	345.2
128	6.0	36.3	5.1	101.3	4.1	1250.2	9.9	55.4	76.2	108.0	7.2	656.9
256	6.3	34.4	3.2	160.0	3.3	1553.3	9.5	57.7	51.4	160.1	6.5	727.6

Table 11.3 Weak scaling behavior for Yeti footprint

	$\ell = 7$, $p = 4$					$\ell = 7$, $p = 8$				
# Proc.	# dofs	It.	Setup	Ass.	Solving/It.	# dofs	It.	Setup	Ass.	Solving/It.
4	360,902	46	1.4	9.2	0.17	383,262	46	6.6	147.3	0.47
16	1,442,569	44	2.4	9.3	0.19	1,531,977	44	8.7	151.2	0.47
64	5,768,189	41	7.2	9.5	0.26	6,125,757	28	13.2	148.7	0.48
256	23,068,573	36	42.4	9.7	0.26	24,498,717	26	54.5	153.8	0.77

fact that the parallel code does not assemble the whole stiffness matrix but works with patch-local stiffness matrices. This allows also to consider the larger problem with $\ell = 7$ and $p = 8$, where the serial code caused memory problems.

In the following rows, we consider the strong scaling behavior. We present in each case the time t in seconds required for setup, assembling and solving and the corresponding speedup s. We observe that the overall method has good strong scaling properties. As the setup phase consists also of parts that are not parallelized, we observe this time does not fall below a few seconds. The assembling phase, which is known to be dominant phase in high-order isogeometric methods, scales almost optimal. Also the solving phase needs rather little communication and is expected to scale well therefore. Indeed, the speedup is much larger than what would be expected. The authors think that this might be explained by some extraordinary caching effects, but here further investigation is required. The extraordinary well behavior of the solver cannot be explained with changed convergence behavior because in all cases, the convergence behavior is identical.

In Table 11.3, we present weak scaling results. We again fix the grid level ℓ and the spline degree p to two typical values. Here, for the case of 4 processors, we consider the initial configuration of $K = 21$ patches; in the following rows we consider 84, 336 and 1344 patches. (So, the third row with 64 processors coincides with the line with 64 processors in Table 11.2.) As the setup phase is not fully

parallelized, the setup times increase if the number of patches is increased. Both, the assembling times and the solving times (per iteration) are rather constant and do not indicate a clear tendency. The solving times also change due to the fact that the required number of iterations decays if the patches are split up. The computational costs for the global-coarse grid solver is negligible in this example; for $K = 1344$ patches the costs are 0.36 s for $p = 4$ and 1.4 s for $p = 7$ and for fewer patches even less.

11.5.2 The Fichera Corner (3D)

On the Fichera corner, we solve the model problem

$$\begin{aligned} -\Delta u &= 75\pi^2 \sin(5\pi\ x) \sin(5\pi\ y) \sin(5\pi\ z) && \text{in } \Omega := (0,2)^3 \backslash [1,2)^3 \,, \\ u &= 0 && \text{on } \Gamma_D := \{(x,y,z) \in \partial\Omega \ :\ xyz = 0\} \,, \\ \frac{\partial}{\partial n} u &= 0 && \text{on } \Gamma_N := \partial\Omega \backslash \Gamma_D \,. \end{aligned}$$

The Fichera corner consists of seven patches, which can be seen in Fig. 11.4b, which are uniformly split into $K = 448$ patches in total. Again, we solve the problem with a conjugate gradient solver preconditioned with one V-cycle of the multigrid method with $1 + 1$ smoothing steps. Again $\tau = 0.25$ and $\sigma = \frac{1}{0.2} h_\ell^{-2}$ are chosen.

In Table 11.4, we report on the number of iterations required to reach the desired relative accuracy goal of 10^{-8}. We observe, as for the Yeti footprint, that the number of iterates is quite robust in the grid size and in the spline degree. The presented numbers have been computed with the serial code. Again, the parallel code yields (almost) the same numbers.

In Table 11.5, we present the strong scaling results. For $\ell = 4$ and $p = 2$, we have $N = 2\,201\,024$ dofs and a stiffness matrix with 2.5 10^8 non-zero entries. The second example with $\ell = 3$ and $p = 4$ yields $N = 596\,288$ dofs and 2.8 10^8 non-zero entries. The timings behave similar as in the two-dimensional case, however the costs of the setup phase are much larger which can be explained by the fact that the interfaces are much larger. (For two dimensional problems, the interfaces consist of $\mathcal{O}(N^{1/2})$ dofs and for three dimensional problems, the interfaces consist

Table 11.4 Iteration counts for Fichera corner, $K = 448$

	p				
ℓ	2	3	4	5	6
1	30	31	31	26	22
2	33	32	33	31	28
3	39	38	37	33	30
4	44	44	42	37	35

Table 11.5 Strong scaling behavior for Fichera corner

	$\ell = 4$, $p = 2$, $K = 448$						$\ell = 3$, $p = 4$, $K = 448$					
	Setup		Assembling		Solving		Setup		Assembling		Solving	
# Proc.	t	s	t	s	t	s	t	s	t	s	t	s
(serial)	179.4	–	260.2	–	4980.7	–	93.5	–	1313.5	–	1252.2	–
1	198.2	1	253.4	1	5091.3	1	109.7	1	1985.9	1	1073.1	1
2	77.7	2.5	127.4	1.9	1355.0	3.7	51.2	2.1	1103.3	1.8	340.0	3.1
4	49.2	4.0	63.5	3.9	395.7	12.8	30.6	3.5	492.2	4.0	110.2	9.7
8	32.9	6.0	32.0	7.9	99.2	51.3	22.8	4.8	214.1	9.2	40.1	26.7
16	28.3	7.0	16.5	15.3	34.4	148.0	27.3	4.0	113.4	17.5	22.7	47.2
32	26.8	7.4	8.4	30.1	12.4	410.5	17.0	6.4	55.5	35.7	8.3	129.2
64	32.2	6.1	5.5	46.0	5.0	1018.2	24.4	4.5	31.2	63.6	6.9	155.5
128	42.3	4.6	2.7	93.8	2.6	1958.1	15.8	6.9	15.4	128.9	3.6	298.0
256	54.1	3.6	1.1	230.3	1.8	2828.5	24.0	4.5	7.9	251.3	3.9	275.1

Table 11.6 Weak scaling behavior for Fichera corner

	$\ell = 4$, $p = 2$					$\ell = 3$, $p = 4$				
# Proc.	# dofs	It.	Setup	Ass.	Solving/It.	# dofs	It.	Setup	Ass.	Solving/It.
1	34,391	28	0.4	4.0	0.07	9317	31	0.6	38.0	0.04
8	275,128	39	1.3	4.5	0.10	74,536	35	1.1	22.8	0.06
64	2,201,024	45	54.4	4.5	0.15	596,288	38	24.1	28.7	0.16
512	17,608,192	46	2071.3	5.1	0.24	4,770,304	35	2343.8	32.4	1.70

of $\mathcal{O}(N^{2/3})$ dofs.) The assembling times seem to be optimal, whereas the solving times again behave extraordinary well.

In Table 11.6, we present weak scaling results. We again fix the grid level ℓ and the spline degree p to two typical values. Here, for the case of 4 processors, we consider the initial configuration of $K = 7$ patches. For the following rows, we consider 56, 448 and 3584 patches. (So, the line with 64 processors coincides with the corresponding line in Table 11.5.) Again, the assembling times do not show any clear tendency, while the solving times (per iteration) are slightly increasing. Only for the last line with 3584 patches, the coarse-grid solver causes problems. For the case $\ell = 3$ and $p = 4$, 1.46 of the 1.70 s required for one iteration of the solver are due to the global solver on the coarsest grid. Again, the setup costs get dominant if the number of patches is increased.

Concluding, we have shown that the robust multi-patch multigrid solver from [13] can be extended to three dimensional domains and that it converges well also in this case. We have observed that the multigrid solver can be parallelized in a natural way yielding very good speedup rates. Certainly, this is not the end of the story and further improvement should be considered in two directions. First, the setup phase becomes a bottleneck if many processors are considered. Here, improvements would be mainly a challenge in terms of implementation and data management. Second, the coarse-grid problem becomes too large if the number

of patches is increased, particularly in the three dimensional case. To resolve that issue, it would be necessary to further coarsen the coarse-grid problem or to consider approximate solvers on the coarsest grid level which certainly would change the mathematical meaning of the algorithm and could, therefore, influence its convergence behavior. Finally, further investigation is required to completely understand the super optimal speedup rates observed in the strong scaling tests.

Acknowledgements The first author would like to thank the Austrian Science Fund (FWF) for the financial support through the DK W1214-04, while the second author was supported by the FWF grant NFN S117-03.

References

1. Collier, N., Pardo, D., Dalcin, L., Paszynski, M., Calo, V.M.: The cost of continuity: a study of the performance of isogeometric finite elements using direct solvers. Comput. Methods Appl. Mech. Eng. **213–216**, 353–361 (2012)
2. da Veiga, L.B., Cho, D., Pavarino, L., Scacchi, S.: Overlapping Schwarz methods for isogeometric analysis. SIAM J. Numer. Anal. **50**(3), 1394–1416 (2012)
3. da Veiga, L.B., Pavarino, L.F., Scacchi, S., Widlund, O.B., Zampini, S.: Isogeometric BDDC preconditioners with deluxe scaling. SIAM J. Sci. Comput. **36**(3), A1118–A1139 (2014)
4. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis. SIAM J. Numer. Anal. **55**(1), 31–62 (2016)
5. Douglas, C., Haase, G., Langer, U.: A Tutorial on Elliptic PDE Solvers and Their Parallelization. SIAM, Philadelphia (2003)
6. Hofer, C., Takacs, S., Mantzaflaris, A., et al.: G+Smo (2018). http://gs.jku.at/gismo
7. Hofreither, C., Takacs, S.: Robust multigrid for isogeometric analysis based on stable splittings of spline spaces. SIAM J. Numer. Anal. **4**(55), 2004–2024 (2017)
8. Hofreither, C., Takacs, S., Zulehner, W.: A robust multigrid method for isogeometric analysis in two dimensions using boundary correction. Comput. Methods Appl. Mech. Eng. **316**, 22–42 (2017)
9. Hughes, T.J.R., Cottrell, J.A., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**(39–41), 4135–4195 (2005)
10. Kleiss, S.K., Pechstein, C., Jüttler, B., Tomar, S.: IETI – Isogeometric tearing and interconnecting. Comput. Methods Appl. Mech. Eng. **247–248**, 201–215 (2012)
11. Sangalli, G., Tani, M.: Isogeometric preconditioners based on fast solvers for the Sylvester equation. SIAM J. Sci. Comput. **38**(6), A3644–A3671 (2016)
12. Takacs, S., Hofer, C., Langer, U.: Inexact dual-primal isogeometric tearing and interconnecting methods. In: Bjørstad, P.E., Brenner, S.C., Halpern, L., Kornhuber, R., Kim, H.H., Rahman, T., Widlund, O.B. (eds.) Domain Decomposition Methods in Science and Engineering XXIV, pp. 397–406 (2019)
13. Takacs, S.: Robust approximation error estimates and multigrid solvers for isogeometric multipatch discretizations. Math. Models Methods Appl. Sci. **28**(10), 1899–1928 (2018)

Chapter 12
On a Renewed Approach to A Posteriori Error Bounds for Approximate Solutions of Reaction-Diffusion Equations

Vadim G. Korneev

Abstract We discuss a new approach to obtaining the guaranteed, robust and consistent a posteriori error bounds for approximate solutions of the reaction-diffusion problems, modelled by the equation $-\Delta u + \sigma u = f$ in Ω, $u|_{\partial\Omega} = 0$, with an arbitrary constant or piece wise constant $\sigma \geq 0$. The consistency of a posteriori error bounds for solutions by the finite element methods assumes in this paper that their orders of accuracy in respect to the mesh size h coincide with those in the corresponding sharp a priori bounds. Additionally, it assumes that for such a coincidence it is sufficient that the testing fluxes possess only the standard approximation properties without resorting to the equilibration. Under mild assumptions, with the use of a new technique, it is proved that the coefficient before the L^2-norm of the residual type term in the a posteriori error bound is $\mathcal{O}(h)$ uniformly for all testing fluxes from admissible set, which is the space $\mathbf{H}(\Omega, \mathrm{div})$. As a consequence of these facts, there is a wide range of computationally cheap and efficient procedures for evaluating the test fluxes, making the obtained a posteriori error bounds sharp. The technique of obtaining the consistent a posteriori bounds was exposed in [arXiv:1711.02054v1 [math.NA] 6 Nov 2017] and very briefly in [*Doklady Mathematics*, **96** (1), 2017, 380–383].

12.1 Introduction

The paper is dedicated to the derivation of the guaranteed robust and computable a posteriori error bounds possessing three properties. Firstly, the bound of the error energy norm must provide the guaranteed accuracy equal in the order to the accuracy of the unimprovable a priori error bound. Secondly, the first property must be achieved with the use of the testing fluxes evaluated by pure approximation instruments, in particular, without resorting to the equilibration of the testing fluxes.

V. G. Korneev (✉)
St. Petersburg State University, St. Petersburg, Russia

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_12

Thirdly, the evaluation of the a posteriori bound must be cheap. For brevity, we term the bounds with the first two properties *consistent*.

A posteriori bounds, suggested in the paper, belong to the family of the error energy norm bounds with two main terms in the right part called diffusion and residual terms. These terms depend on the approximate solution of the boundary value problem, the right part of the problem, and on the testing flux from some admissible set. Effectiveness of such bounds is strongly influenced as by their specific form, so the possibility to pick up a testing flux in the admissible set, providing cheap computations and sufficient accuracy. In spite of the pointed out common features, there were quite a number of suggestions in the literature with differently defined specific forms, i. e., coefficients in front of the norms in the diffusion and residual terms, and differently defined test fluxes.

The main model problem of the paper is

$$\mathfrak{L}u \equiv -\Delta u + \sigma u = f(x), \quad \text{in } \Omega \subset \mathbb{R}^m, \quad m = 2, 3, \quad u\big|_{\partial\Omega} = 0,$$

with the constant or piece wise (element wise) constant reaction coefficient $\sigma \geq 0$. If for simplicity we turn to the case of the constant reaction coefficient, the typical residual term of the majorant[1] looks as $\theta^{1/2}\|\hat{f} - \sigma u_h - \operatorname{div}\mathbf{z}\|_{L_2(\Omega)}$, where u_h is the approximate solution, and $\mathbf{z}$ is the testing flux vector-function. Different authors come to different expressions for θ and, in particular, to (a) $\theta = 1/\sigma$ for all $\sigma > 0$, (b) $\theta =$const for $\sigma \equiv 0$, (c) $\theta = (ch)^2$ for $\sigma \leq (ch)^{-2}$, and (d) on some subdomains (finite elements) the residual equals to zero and, therefore, eliminated. Examples of such majorants are given in the next section.

Majorants related to the cases (a)–(c) behave differently in dependence on the chosen test flux $\mathbf{z}$. Most often, elliptic equations of second order are approximately solved by the FEM's (finite element methods) of the class C, whose solutions belong to the finite dimensional subspaces of $C(\Omega) \cap H^1(\Omega)$. The second derivatives of these solutions, which are needed to calculate the residuals, are not defined. As a consequence, the a posteriori majorants of the error are calculated with the use of smoother testing fluxes, which are found from the FEM fluxes by means of special *flux recovery procedures*. The need to improve the smoothness of numerical fluxes without losses and even, under some conditions, with gaining in the accuracy motivated the development of these procedures, which currently are numerous. Two main classes of them can be distinguished, one aimed completely at achieving highest accuracy without resorting to the equilibration and another which assumes equilibration. For making the equilibration simpler and less dependent on the right part f of the elliptic equation, it is commonly done not for f itself, but for some its approximation $\hat{f}$. The latter is defined element wise, e.g., as the orthogonal L^2-projection of f on the space of each finite element. Among great number of works on the touched topics, we are able to refer to a few, in particular, to Zienkiewicz and Zhu [39], Ainsworth and Oden [1], Babuska et al. [6, 7], Xu and Zhang [36],

[1] In this text, the right part of an a posteriori bound is termed majorant.

Karakashian and Pascal [20], Ern et al. [19], Cheddadi et al. [14], Cai and Zhang [11], and Ainsworth and Vejchodsky [3], where additional extensive bibliography can be found.

For the sequel it is worth to fix the following properties, additional to the smoothing and paid attention at the creation of flux recovery procedures:

- (α) preserving the orders of accuracy, e.g., in the energy and other norms, at least the same as for the numerical fluxes determined by the approximate solution of the problem;
- (α_*) providing (α) and additionally superconvergence of the test fluxes $\mathbf{z}$;
- (β) obtaining the balanced or weakly balanced recovered fluxes;
- (γ) linear or almost linear computational cost.

If only α) or α_*) with or without (γ) are satisfied, then it is easy to see that the majorants in (a) and (b) can be larger in h^{-1} times, and even more in the case (a), than the energy norm of the true error. Therefore, these bounds are inconsistent. The orders of smallness of a posteriori majorants, related to (c) and (d), can be equal to the orders in the corresponding a priori error bounds, but until recently their examples assumed the equilibrated test fluxes.

Obviously, procuring the property (β) complicates flux recovery procedures, makes them directly dependable upon the specific boundary value problem to be solved and, therefore, much less universal.

In this paper, we develop the a posteriori error bounds, which for approximate solutions by the finite element method are consistent with the sharp a priori error bounds on the wide range of testing fluxes satisfying only the property (α). They have additional features, making them quite different from other bounds of the family under consideration. In particular, in their right parts, supposed for the use as the error indicators, they have $ch\|f - \sigma u_h - \operatorname{div}\mathbf{z}\|_{L_2(\Omega)}$ in the residual term, independently of $\mathbf{z} \in \mathbf{H}(\Omega, \operatorname{div}) := \{\mathbf{y} = (y_1, y_2, \ldots, y_m)^\top : y_k \in L^2(\Omega),\ \operatorname{div}\mathbf{y} \in L^2(\Omega)\}$ and $\sigma \in [0, (ch)^{-2}]$. As a consequence, the evaluation of the test fluxes, providing sharp bounds, becomes *a purely approximation problem* and the need to equilibrate fluxes disappears. These two features are meaningful conceptually as establishing a proper correlation between a posteriori and a priori error bounds, but as well they are important for the practice, because the range of the numerical applicability of the sharp and cheap a posteriori error bounds greatly widens. Indeed, a number of simple and cheap flux recovery techniques like local weighted averaging, which have been developed and extensively studied in relation with the residual based a posteriori error indicators, are applicable to the bounds of this paper. They provide not only the standard approximation properties, needed for a consistent a posteriori bound to became sharp, but, under some conditions, also the property α_*) which implies superconvergence of the recovered fluxes, see, e.g., [8, 36]. As is noted by some authors, these flux recovery procedures performed in the practice "astonishingly well", see, e.g., [1, 36].

The approval of the new weights before the norms in the consistent a posteriori error majorants required revision of the proofs. A key change is due to the fact that, at the derivation of the majorants, we take into account the difference in the error

orders of the finite element solutions in L^2 and H^1 norms. This allowed us to come to θ of the optimal order under general conditions in two steps made in Sects. 12.4 and 12.5, respectively. Since the constants in the majorants is an important issue for applications, we study two approaches to their derivation, which influence the values of the constants and simplicity of their evaluation. One of them imposes some restriction on the smoothness of the boundary, because it employs the Aubin-Nitsche trick for a subsidiary problem, governed by the Poisson equation. The constants resulting from another depend only upon properties of the local quasi-interpolation operator $H^1(\delta^{(r)}) \to \mathcal{V}_h(\tau_r)$, where $\mathcal{V}_h(\tau_r)$ is the space induced by the finite element τ_r, $\delta^{(r)}$ is the smallest patch of the finite elements neighboring τ_r. A representative of such operators is the one of Scott and Zhang [34]. However, we do not elaborate on the comparison of the effectiveness of different quasi-interpolation and projection operators, see, e.g., in [9, 10, 30].

The paper is organized as follows. Section 12.2 contains the formulation of the boundary value problem of reaction-diffusion, examples of known error majorants, similar in structure to ones suggested in the paper, and a brief discussion of their consistency. In Sect. 12.3, we derive new general a posteriori majorants for the errors of approximations of the exact solution to the problem by arbitrary sufficiently smooth functions v that satisfy the essential boundary conditions. The cases of an arbitrary piece wise constant and a constant reaction coefficients are considered separately. The majorants are well defined in both cases, if $\sigma \geq 0$, and coincide with the Aubin's [5] type majorant on those subdomains (finite elements), where σ exceeds a certain critical value σ_*. Several versions of the majorant are discussed, related to different ways of defining σ_* and, respectively, coefficients of the majorant. The easiest one corresponds to $\sigma \equiv$ const and Galerkin method with coordinate functions having additional smoothness, e.g., belonging to the space $H^2(\Omega)$. Such coordinate functions find implementation in the isogeometric , see Cottrel etc. [17]. Majorants of Sect. 12.3 are more accurate at least for $\sigma \leq \sigma_*$ than other known general majorants valid for all $\mathbf{z} \in \mathbf{H}(\Omega, \mathrm{div})$ and allow sharpening based upon estimating σ_* for a specific numerical method. In this sense, they are the base for the subsequent sections.

Consistent a posteriori error estimates for solutions by the finite element method, which are the main results of the paper, are presented in Sects. 12.4 and 12.5. Theorems 12.4 and 12.5, proved there on the basis of the results of Sect. 12.3, suggest different approaches to the evaluation of constants in these estimates. For instance, in Theorem 12.5 they are expressed through the constants in the estimates of approximation produced by the quasi-interpolation projection operator of Scott and Zhang [34]. In Sect. 12.5, some other important properties of the derived a posteriori estimates are discussed, in particular, the approximation conditions for test fluxes, sufficient for the a posteriori error estimators to be sharp, cheap flux recovery procedures for computing such test fluxes. The conclusion about efficiency of the majorants is supported also by the example of the inverse like bound.

Below, $\|\phi\|_{H^k(\mathscr{D})}$ is the norm in the Sobolev space $H^k(\mathscr{D})$ on a domain $\mathscr{D}$,

$$\|\phi\|^2_{H^k(\mathscr{D})} = \|\phi\|^2_{L_2(\mathscr{D})} + \sum_{l=1}^{k} |\phi|^2_{H^l(\mathscr{D})}, \quad |\phi|^2_{H^l(\mathscr{D})} = \sum_{|q|=l} \int_{\mathscr{D}} \Big(\frac{\partial^l \phi}{\partial x_1^{q_1} \partial x_2^{q_2} \dots \partial x_m^{q_m}}\Big)^2 dx\,,$$

where $\|\phi\|^2_{L_2(\mathscr{D})} = \int_{\mathscr{D}} \phi^2 dx$ and $q = (q_1, q_2, \dots, q_m)$, $q_k \geq 0$, $|q| = q_1 + q_2 + \dots + q_m$. If $\mathscr{D} = \Omega$, for $\|\cdot\|_{L_2(\Omega)}$, $\|\cdot\|_{H^k(\Omega)}$ and $|\cdot|_{H^k(\Omega)}$ will be also used simpler symbols $\|\cdot\|_0$, $\|\cdot\|_k$ and $|\cdot|_k$, respectively. If on the entire boundary $\partial\Omega$ or on its part Γ_D the homogeneous Dirichlet boundary condition is specified, then for the corresponding subspaces of $H^1(\Omega)$ we use the notations $\mathring{H}^1(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ and $\mathring{H}^1_{\Gamma_D}(\Omega) := \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}$. Besides, we will need the spaces $\mathring{H}^1(\Omega, \Delta) = \{v \in \mathring{H}^1(\Omega) : \Delta v \in L_2(\Omega)\}$ and $\mathring{H}^1_{\Gamma_D}(\Omega, \Delta) = \{v \in \mathring{H}^1_{\Gamma_D}(\Omega) : \Delta v \in L_2(\Omega)\}$.

The finite element space will be denoted $\mathscr{V}_h(\Omega)$ and by definition $\mathring{\mathscr{V}}_h(\Omega) = \{v \in \mathscr{V}_h(\Omega) : v|_{\partial\Omega} = 0\}$.

Everywhere below it is assumed that the assemblage of compatible and, generally speaking, curvilinear finite elements is given on $\Omega \subset \mathbb{R}^m$, $m = 2, 3$, with each finite element occupying domain τ_r, $r = 1, 2, \dots, \mathscr{R}$. Sometimes symbol $\mathscr{R}$ is also used for the *set* of numbers of finite elements. The finite elements are defined by sufficiently smooth mappings $x = \mathscr{X}^{(r)}(\xi) : \tau_\circ \to \tau_r$ of the reference element, defined on the standard triangle or tetrahedron $\tau_\circ$. The span of the coordinate functions of the reference element is the space $\mathscr{P}_p$ of polynomials of degree p. If $p > 1$, we use additionally the notation $\mathscr{V}_h(\Omega) = \mathscr{V}_{h,p}(\Omega)$. If other is not mentioned, it is always assumed that the finite element assemblage satisfies the *generalized conditions of quasiuniformity* with the mesh parameter $h > 0$, which can be understood as the maximum of diameters of finite elements. These conditions can be found, for instance, in Korneev and Langer [26, Section 3.2]. Occasionally for more simplicity the following condition is assumed:

$\mathscr{A}$) The domain Ω is a polygon in $\mathbb{R}^m$, $m = 2, 3$, τ_r are compatible m-dimensional simplices (with flat faces and, respectively, straight edges) forming the triangulation $\mathscr{T}_h$ of Ω, satisfying the conditions of shape regularity.

In the applications, as a rule $\mathscr{V}_h(\Omega) \subset C(\Omega) \cap H^1(\Omega)$. At the same time, in the isogeometric numerical analysis, the smoother spaces $\mathscr{V}_h(\Omega) = \mathscr{V}_h{}^1(\Omega) \subset C^1(\Omega) \cap H^2(\Omega)$, see Cottrell et al. [17] and Langer et al. [28], are used for solving elliptic equations of second order. Superscript l in the notations $\mathscr{V}_h{}^l(\Omega)$, $\mathscr{V}^l_{h,p}(\Omega)$ assumes inclusions of these spaces in $C^l(\Omega) \cap H^{l+1}(\Omega)$.

Matrices and vectors are designated by bold capital and bold lowercase letters, respectively.

12.2 Model Problem, Examples of A Posteriori Error Majorants

We start from a very brief survey of a few well known a posteriori bounds which illustrate some tendencies in the development of such error control instruments and were successfully implemented in the numerical applications. We will concentrate on the consideration of the model problem

$$\begin{gathered}\mathfrak{L}u \equiv -\Delta u + \sigma u = f(x), \quad x = (x_1, x_2, \dots, x_m) \in \Omega \subset \mathbb{R}^m, \\ u\big|_{\Gamma_D} = \psi_D, \quad -\nabla u \cdot \mathbf{n}\big|_{\Gamma_N} = \psi_N,\end{gathered} \tag{12.1}$$

where Γ_D, Γ_N are disjoint simply connected parts of the boundary $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\operatorname{mes}\Gamma_D > 0$, $\mathbf{n}$ is the internal normal to $\partial\Omega$. The reaction coefficient $\sigma \geq 0$ is assumed to be element wise constant, i.e.,

$$\sigma = \sigma_r = \text{const}, \quad x \in \tau_r, \quad r = 1, 2, \dots \mathscr{R}. \tag{12.2}$$

The boundary of Ω and the right part f are always considered as sufficiently smooth, in particular, $f \in L_2(\Omega)$, if the requirements on the smoothness are not formulated differently.

Our primal interest will be the error estimates in the energy norm

$$|\!|\!| v |\!|\!|^2 = \left(|v|_1^2 + \|\sqrt{\sigma} v\|^2_{L_2(\Omega)}\right)^{1/2}, \qquad |v|_1^2 = \int_\Omega \nabla v \cdot \nabla v, \tag{12.3}$$

with $|\!|\!|\cdot|\!|\!|_{\tau_r}$ denoting the restriction to τ_r. For vector-functions $\mathbf{y} = (y_1, y_2, \dots, y_m)^\top$, we introduce also the spaces $\mathbf{L}^2(\Omega) = (L_2(\Omega))^m$, $\mathbf{H}(\Omega, \operatorname{div}) = \{\mathbf{y} \in \mathbf{L}^2(\Omega) : \operatorname{div}\mathbf{y} \in L_2(\Omega)\}$, $\mathbf{W}_{h,p}(\Omega, \operatorname{div}) = \{\mathbf{y} \in \mathbf{H}(\Omega, \operatorname{div}) : y_k\big|_{\tau_r} \in \mathscr{P}_p, \ \forall r \in \mathscr{R}, \ k = 1, \dots, m\}$ and the respective norms. For the norm in $\mathbf{L}^2(\Omega)$ we use additionally the notation $]\!]\mathbf{y}[\![= \|\mathbf{y}\|_{\mathbf{L}^2(\Omega)} = (\int_\Omega \mathbf{y} \cdot \mathbf{y})^{1/2}$.

The a posteriori error majorant of Aubin [5] is one of the earliest:

Theorem 12.1 *Let $f \in L_2(\Omega)$, $0 < \sigma = \text{const}$, $\psi_D \in H^1(\Omega)$, $\psi_N \in L_2(\Gamma_N)$, v be any function of $H^1(\Omega)$ that satisfies the boundary condition on Γ_D. Then, for any $\mathbf{z} \in \mathbf{H}(\Omega, \operatorname{div})$, satisfying on Γ_N the boundary condition $\mathbf{z} \cdot \mathbf{n} = \psi_N$, we have*

$$|\!|\!| v - u |\!|\!|^2 \leq]\!]\nabla v + \mathbf{z}[\![^2 + \frac{1}{\sigma}\|f - \sigma v - \operatorname{div}\mathbf{z}\|^2_{L_2(\Omega)}. \tag{12.4}$$

Proof Estimate (12.4) is a special case of the results of Aubin [5], see Theorem 22 in Introduction and Theorems 1.2, 1.4, 1.6 in Chapter 10. □

Obviously, if $\sigma \to 0$ the majorant of Aubin loses precision and for $\sigma = 0$ makes no sense.

If $\sigma \equiv 0$, one can use the majorant of Repin and Frolov [32]. Let for simplicity, $\Gamma_D = \partial\Omega$, $\psi_D \equiv 0$, and $\sigma \equiv 0$. Then for any $\epsilon > 0$

$$\|\nabla(v-u)\|^2_{\mathbf{L}^2(\Omega)} \le (1+\epsilon)\|\nabla v+\mathbf{z}\|^2_{\mathbf{L}^2(\Omega)}+c_\Omega(1+\frac{1}{\epsilon})\|\operatorname{div}\mathbf{z}-f\|^2_{L_2(\Omega)}\,, \tag{12.5}$$

where v and $\mathbf{z}$ are a function and an arbitrary vector-function from $\mathring{H}^1(\Omega)$ and $\mathbf{H}(\Omega, \operatorname{div})$, respectively, and c_Ω is the constant from the Friedrichs inequality.

Correction of arbitrary vector-function $\mathbf{z} \in \mathbf{H}(\Omega, \operatorname{div})$ into the vector-function $\boldsymbol{\tau}$, satisfying the balance/equilibrium equations, can be done by quite a few rather simple and general techniques. Some of them can be found in Anufriev et al. [4] and Korneev [22]. In particular, it is true for the correction of the finite element flux vector-function ∇u_{fem} into $\boldsymbol{\tau}(u_{\text{fem}})$. This leads to a specific family of a posteriori error bounds. For simplicity, we turn to the same homogeneous Dirichlet problem for the Poisson equation in a two-dimensional convex domain. Let T_k be the projection of the domain Ω on the axis x_{3-k} and the equations of the left and lower parts of the boundary be $x_k = a_k(x_{3-k})$, $x_{3-k} \in T_k$. If β_k are arbitrary bounded functions and $\beta_1 + \beta_2 \equiv 1$, then according to [4, 22]

$$\begin{aligned}&\|\nabla(v-u)\|_{\mathbf{L}_2(\Omega)} \le \|\nabla v+\mathbf{z}\|_{\mathbf{L}_2(\Omega)}+\\ &\textstyle\sum_{k=1,2}\|\int_{a_k(x_{3-k})}^{x_k}\beta_k(f-\operatorname{div}\mathbf{z})(\eta_k, x_{3-k})\,d\eta_k\|_{L_2(\Omega)}\,.\end{aligned} \tag{12.6}$$

On the right of (12.6), we have integrals from the residual and this hopefully will make the majorant more accurate. Besides, there is an additional free function β_1 or β_2 and the right choice of it (for instance, with the use of the found approximate solution $v = u_{\text{fem}}$) can accelerate the process of the minimization of the right part, if such a process is implemented. If to estimate one-dimensional integrals under the sign of the L_2-norm, then we come to the bound similar to (12.5).

There were attempts to modify the majorant of (12.4) with the aim of achieving acceptable accuracy for all $\sigma \ge 0$, see e.g., Repin and Sauter [33] and Churilova [15]. The majorant of the latter, defined for $\forall\,\sigma = \text{const} \ge 0$, has the form

$$|||v-u|||^2 \le (1+\epsilon)|||\nabla v+\mathbf{z}|||^2+\frac{1}{\sigma+\frac{\epsilon}{c_\Omega(1+\epsilon)}}\|f-\sigma v-\operatorname{div}\mathbf{z}\|^2_{L_2(\Omega)}\,. \tag{12.7}$$

One of the efficient majorants for the finite element solutions was developed by Ainsworth and Vejchodsky [2, 3]. For its record, we need additional notations: h_r is the diameter τ_r, $\Pi_r^p : L_2(\tau_r) \to \mathscr{P}_p(\tau_r)$ is the operator of orthogonal projection in $L_2(\tau_r)$.

In Theorems 12.2 and 12.3 below, for simplicity we assume $\Gamma_D = \partial\Omega$, $\psi_D \equiv 0$ and that the condition $\mathscr{A}$) is fulfilled.

Theorem 12.2 *Let $u \in \mathring{H}^1(\Omega)$ be the weak solution of the problem and $u_{\rm fem} \in \mathring{\mathscr{V}}_h(\Omega) = \mathring{\mathscr{V}}^0_{h,1}(\Omega)$ be the solution by the finite element method. Then there exists $\mathbf{z} \in \mathbf{W}_{h,2}(\Omega, \mathrm{div})$ with the following properties:*

(i) $\mathbf{z}$ *is evaluated by the patch wise numerical procedure of linear numerical complexity,*

(ii) for all $x \in \tau_r$ and $r \in \mathscr{R}_* = \{r : \sqrt{\sigma}_r h_r < 1\}$ satisfies the equalities

$$\Pi^1_r f - \sigma_r u_{\rm fem} + \mathrm{div}\,\mathbf{z} = 0\,, \tag{12.8}$$

(iii) for the error $e_{\rm fem} = u - u_{\rm fem}$ and the error indicator $\eta_{\tau_r}(\mathbf{z})$, defined as

$$\eta^2_{\tau_r}(\mathbf{z}) = \|\mathbf{z} - \nabla u_{\rm fem}\|^2_{L_2(\tau_r)}, \quad \forall r \in \mathscr{R}_*\,,$$

$$\eta^2_{\tau_r}(\mathbf{z}) = \|\mathbf{z} - \nabla u_{\rm fem}\|^2_{L_2(\tau_r)} + \tfrac{1}{\sigma_r}\|\Pi^1_r f - \sigma_r u_{\rm fem} + \mathrm{div}\,\mathbf{z}\|^2_{L_2(\tau_r)}\,, \quad \forall r \in \mathscr{R} \setminus \mathscr{R}_*\,, \tag{12.9}$$

there hold the bounds

$$|\!|\!| e_{\rm fem} |\!|\!|^2 \le \sum_{\tau_r \in \mathscr{T}_h} \Big[\eta_{\tau_r}(\mathbf{z}) + \mathrm{osc}_{\tau_r}(f)\Big]^2, \tag{12.10}$$

$$\eta^2_\Omega(\mathbf{z}) = \sum_{r \in \mathscr{R}} \eta^2_{\tau_r}(\mathbf{z}) \le \mathbb{C}\Big[|\!|\!| e_{\rm fem} |\!|\!|^2 + \sum_{r \in \mathscr{R}} \mathrm{osc}^2_{\tau_r}(f)\Big], \tag{12.11}$$

where $\mathrm{osc}_{\tau_r}(f) = \min\left\{\frac{h_r}{\pi}, \frac{1}{\sqrt{\sigma_r}}\right\} \|f - \Pi^1_r f\|_{L_2(\tau_r)}$.

Proof See Ainsworth and Vejchodsky [3] for the proof. In this work the bounds (12.10), (12.11) in a more general form are derived under more general conditions. In particular, $\Gamma_N \neq \varnothing$, the bound (12.11) is proved in the local version, i.e., with $\eta^2_{\tau_r}(\mathbf{z})$ on the left and with the restriction of the right part to the patch $\delta^{(r)}$. □

We present also the error majorant of Cheddadi et al. [14] for approximate solutions of the reaction-diffusion problem by the method of vertex-centered finite volumes. In this work the point wise equilibration is avoided and replaced by the much more flexible equilibration in a weak sense. Let us introduce the notations: $\mathscr{D}_h$ is the dual in respect to $\mathscr{T}_h$ partition of Ω; $\mathscr{S}_h$ is the finer mesh, induced by the partition $\mathscr{D}_h$; D is the polygon with the center in the vertex of triangulation $\mathscr{T}_h$ and containing all simplices of the finer mesh with this vertex, h_D is its diameter; $\mathscr{D}^{\rm int}_h$ is the set of all polygons D, for which $\partial D \cap \partial\Omega = \varnothing$. For additional information about these entities, we refer to [14].

Theorem 12.3 *Let u_h be the solution by the method of vertex-centered finite volumes, $e_h = u - u_h$, vector-function $\mathbf{z} \in \mathbf{H}(\Omega, \mathrm{div})$ satisfy the equalities*

$$(f - \mathrm{div}\,\mathbf{z} - \sigma u_h, 1)_D = 0\,, \quad \forall D \in \mathscr{D}_h^{\mathrm{int}}\,, \tag{12.12}$$

and $\gamma_D = \min(C_D h_D^2, \sigma_D^{-1})$, where C_D is the constant from the Poincaré inequality for the polygon $\mathscr{D}$. Then

$$|||e_h|||^2 \le \eta_\Omega^2(\mathbf{z}) = \sum_{D\in\mathscr{D}_h} \left[\|\nabla u_h + \mathbf{z}\|_{L_2(D)} + \sqrt{\gamma_D}\|f - \sigma u_h - \mathrm{div}\,\mathbf{z}\|_{L_2(D)}\right]^2. \tag{12.13}$$

Proof Theorem is one of the results of [14], see Theorem 4.5. □

Majorants in (12.5)–(12.7) have obvious merits, but are not consistent at the application, e.g., to solutions by the finite element and other mesh methods. If $v = u_{\mathrm{fem}}$ is the finite element solution to the problem (12.1) at $\Gamma_D = \partial\Omega$, $\psi_D \equiv 0$, and $\sigma = 0$, then we can use 12.5). For our purpose, it is sufficient to consider the approximate solutions from the space $\mathscr{V}_h(\Omega) = \mathscr{V}_{h,p}^1(\Omega) \subset C^1(\Omega) \cap H^2(\Omega)$. Under the assumption $u \in H^l(\Omega)$, we have a priori error bounds [16, 21]

$$\|u - v\|_{H^k(\Omega)} \le c h^{l-k} \|u\|_{H^l(\Omega)}\,, \quad k = 1, 2, \quad k \le l \le p + 1\,. \tag{12.14}$$

In particular, if $f \in L_2(\Omega)$ and consequently $l = 2$, see [25], then according to (12.14) the left part of (12.5) is estimated as $\mathscr{O}(h^2)$. At the same time at the choice $\mathbf{z} = -\nabla v$ the first term in the right part of (12.5) vanishes, but $c_\Omega(1 + \frac{1}{\epsilon})\|\mathrm{div}\,\mathbf{z} - f\|^2_{L_2(\Omega)}$ at any $\epsilon > 0$ can be bounded from above only by a constant. The bounds (12.14) are *not improvable* in the order. For $u \in H^2(\Omega)$, $\Omega \in \mathbb{R}^2$, Oganesian and Ruhovets [31] gave the proof by estimating the corresponding Kolmogorov's width. From their results and results on the regularity of solutions of (12.1), it follows existence of such $f \in L_2(\Omega)$ that $u \in H^2(\Omega)$ and the second summand on the right of (12.5) is estimated by the constant *from below*. Therefore, the orders of smallness of the left and the right parts of the a posteriori bound are different, and the value $\mathscr{O}(h^2)$ on the left is estimated by the right part only with the order of unity. If $l > 2$, the left and the right parts are estimated with not equal orders $\mathscr{O}(h^{2(l-k)})$ and $\mathscr{O}(h^{2(l-k-1)})$.

Inconsistency of the majorant (12.7) at $\sigma \le c h^{-\alpha}$, $0 \le \alpha < 2$, $c = \mathrm{const}$, is established in a similar way. Majorant (12.6) is consistent in 1-dimensional case, in general it is also inconsistent. The inconsistency of the above mentioned majorants, obviously, is retained, if finite elements of the class C are used and the test flux is found by some recovery procedure, satisfying only the requirements α), γ).

The equalities of the orders of smallness of the left and right parts of the majorants (12.10) and (12.13) are well provided, as follows, e.g., from (12.11) and similar bound, proved in [14]. However, it is achieved only for the test fluxes, satisfying the additional conditions reflecting the requirement (β), see (12.8) and (12.12).

12.3 A Posteriori Error Majorant Robust for Piece Wise Constant and Constant Reaction Coefficients

In this section, we present a general a posteriori error bound for the approximations of the exact solution to the reaction-diffusion equation with the piecewise constant reaction coefficient. These bounds are not influenced by the origins of the approximations and are robust in the respect to the reaction coefficient. In particular, they are well defined in the vicinity of zero.

12.3.1 Reaction-Diffusion Problem with Piece Wise Constant Reaction Coefficient

Let for some $\sigma_* \equiv \text{const} > 0$, the piece wise constant function $\sigma_\dagger$ be

$$\sigma_\dagger = \begin{cases} \sigma_* , & \forall \sigma \in \Omega_* , \\ \\ \sigma , & \forall \sigma \in \Omega_\sigma , \end{cases}$$

where $\Omega_* := \{x : \sigma(x) < \sigma_*\}$ and $\Omega_\sigma = \Omega \setminus \Omega_*$. The basic assumption for what follows is that for some $\sigma_e \geq 0$ the error $e = u - v$ satisfies the inequality

$$\|(\sigma_\dagger - \sigma)^{1/2} e\|^2_{L_2(\Omega)} \leq \frac{1}{\sigma_e} \|(\sigma_\dagger - \sigma)^{1/2} \nabla e\|^2_{L_2(\Omega)} , \tag{12.15}$$

and, obviously, Ω in it can be replaced by Ω_*. We use the notation $\mathbf{H}_{\psi_N}(\Omega, \mathrm{div}) = \{\mathbf{y} \in \mathbf{H}(\Omega, \mathrm{div}),\ \nabla \mathbf{y} \cdot \mathbf{n}\Big|_{\partial\Omega} = \psi_N\}$.

Lemma 12.1 *Let σ be an arbitrary nonnegative element wise constant function as in 12.2), $f \in L_2(\Omega)$, $\psi_D \in H^1(\Omega)$, $\psi_N \in L_2(\Gamma_N)$. Let also σ_e satisfy (12.15) and $\sigma_* > 0$ be arbitrary. Then for any function $v \in \overset{\circ}{H}{}^1_{\psi_D}(\Omega)$ and any vector-function $\mathbf{z} \in \mathbf{H}_{\psi_N}(\Omega, \mathrm{div})$ we have*

$$|||v - u|||^2 \leq \|\widetilde{\gamma}(\nabla v + \mathbf{z})\|^2_{\mathbf{L}^2(\Omega)} + \|\tfrac{\widetilde{\gamma}}{\sqrt{\sigma_\dagger}}\big(f - \sigma v - \mathrm{div}\, \mathbf{z}\big)\|^2_{L_2(\Omega)} \tag{12.16}$$

with

$$\widetilde{\gamma}^2 = \begin{cases} (\sigma_e + \sigma_*)/(\sigma_e + \sigma) , \ \sigma \in [0, \sigma_*] , \\ \\ 1 , \qquad\qquad\qquad \sigma \geq \sigma_* . \end{cases} \tag{12.17}$$

Proof Two additional features to the proof of the Aubin's bound are a more tricky use of the Cauchy inequality $|ab| \leq \epsilon^{-1} a^2 + \epsilon b^2$, $\forall \epsilon > 0$, and the use of the

inequality (12.15). For an arbitrary v and $\mathbf{z}$ from Lemma 12.1, after integration by parts of the first summand in the square of the energy norm of the error $e = v - u$, we have

$$\begin{aligned} |\!|\!| e |\!|\!|^2 &= \int_\Omega \left[\nabla e \cdot \nabla e + \sigma e^2\right] = \int_\Omega \left[(\nabla v + \mathbf{z}) \cdot \nabla e - (\mathbf{z} + \nabla u) \cdot \nabla e + \sigma e^2\right] = \\ &= \int_\Omega \left\{(\nabla v + \mathbf{z}) \cdot \nabla e + [\operatorname{div}(\mathbf{z} + \nabla u) + \sigma e]\, e\right\} = \\ &= \sum\nolimits_r \int_{\tau_r} \left\{(\nabla v + \mathbf{z}) \cdot \nabla e - [f - \sigma v - \operatorname{div} \mathbf{z}]\, e\right\}. \end{aligned} \tag{12.18}$$

For convenience, we will use also the notations $\kappa = \sqrt{\sigma}$, $\kappa_* = \sqrt{\sigma_*}$, and $\kappa_\dagger = \sqrt{\sigma_\dagger}$. By means of the Cauchy inequality, for $\gamma(x) > 0$ belonging to $L_2(\Omega)$ we find that

$$\begin{aligned} |\!|\!| e |\!|\!|^2 &= \|\nabla e\|^2_{\mathbf{L}^2(\Omega)} + \|\kappa\, e\|^2_{L_2(\Omega)} \le \left[\|\gamma(\nabla v + \mathbf{z})\|^2_{\mathbf{L}^2(\Omega)} + \right. \\ &\left. + \|\tfrac{\gamma}{\kappa_\dagger}(f - \sigma v - \operatorname{div} \mathbf{z})\|^2_{L_2(\Omega)}\right]^{1/2} \times \left[\|\tfrac{1}{\gamma}\nabla e\|^2_{\mathbf{L}^2(\Omega)} + \|\tfrac{\kappa_\dagger}{\gamma} e\|^2_{L_2(\Omega)}\right]^{1/2}. \end{aligned} \tag{12.19}$$

In the case under consideration, it is natural to take γ element wise constant, i.e., $\gamma\big|_{\tau_r} = \gamma_r = \text{const}$, $\forall r \in \mathscr{R}$. The summand in the second multiplier of the right part of (12.19), corresponding to one finite element τ_r for $r \in \mathscr{R}_* := \{r : \tau_r \in \Omega_*\}$, we multiply by γ_r and with the use of some $\beta_r \in (0, 1]$ rearrange as follows:

$$\begin{aligned} &\|\nabla e\|^2_{\mathbf{L}^2(\tau_r)} + \sigma_* \|e\|^2_{L_2(\tau_r)} = |\!|\!| e |\!|\!|^2_{\tau_r} + (\sigma_* - \sigma)\|e\|^2_{L_2(\tau_r)} = \\ &= |\!|\!| e |\!|\!|^2_{\tau_r} + \beta_r(\sigma_* - \sigma)\|e\|^2_{L_2(\tau_r)} + (1 - \beta_r)(\sigma_* - \sigma)\|e\|^2_{L_2(\tau_r)}. \end{aligned} \tag{12.20}$$

Summation of (12.20) over $r \in \mathscr{R}$ yields

$$\begin{aligned} &\|\tfrac{1}{\gamma}\nabla e\|^2_{\mathbf{L}^2(\Omega)} + \|\tfrac{\sigma_\dagger^{1/2}}{\gamma} e\|^2_{L_2(\Omega)} = \sum\nolimits_{r=1}^{\mathscr{R}} \tfrac{1}{\gamma_r^2}\left[|\!|\!| e |\!|\!|^2_{\tau_r} + \right. \\ &\left. + \beta_r \|(\sigma_\dagger - \sigma)^{1/2} e\|^2_{L_2(\tau_r)} + (1 - \beta_r)\|(\sigma_\dagger - \sigma)^{1/2} e\|^2_{L_2(\tau_r)}\right]. \end{aligned} \tag{12.21}$$

Let β be the element wise constant function with the restrictions $\beta\big|_{\tau_r} = \beta_r$. On Ω_* it can be uniquely defined from the equality

$$\sigma\left[1 + (\sigma_\dagger - \sigma)\frac{\beta}{\sigma_e}\right] = (1 - \beta)(\sigma_\dagger - \sigma) + \sigma\,, \quad x \in \Omega_*\,, \tag{12.22}$$

as $\beta = \sigma_e/(\sigma_e + \sigma)$ and continued by the unity on Ω_σ. Estimating the sum of the second terms in the brackets by means of (12.15) and combining (12.20) – (12.22) at $\gamma^2 = \beta$, we conclude that

$$\| \tfrac{1}{\gamma} \nabla e \|^2_{\mathbf{L}^2(\Omega)} + \| \tfrac{\kappa_\dagger}{\gamma} e \|^2_{L_2(\Omega)} \le$$

$$\le \| \tfrac{1}{\gamma} \Big[1 + (\sigma_\dagger - \sigma) \tfrac{\beta}{\sigma_e} \Big]^{1/2} \nabla e \|^2_{\mathbf{L}^2(\Omega)} + \| \tfrac{1}{\gamma} \Big[(1-\beta)(\sigma_\dagger - \sigma) + \sigma \Big]^{1/2} e \|^2_{L_2(\Omega)} \le$$

$$\le \tfrac{\sigma_* + \sigma_e}{\sigma_e} \Big[\| \nabla e \|^2_{\mathbf{L}^2(\Omega)} + \| \kappa e \|^2_{L_2(\Omega)} \Big] . \tag{12.23}$$

Combining (12.19) with (12.23), we come to the bound (12.16). □

Let us make a few remarks in relation to the bound (12.16). Since σ_* and σ_e are constants, one can set $\sigma_* = c_* \sigma_e$, $c_* = \text{const}$, and, therefore, $(\sigma_* + \sigma_e)/\sigma_e = 1 + c_*$. Obviously, $\widetilde{\gamma} \in [1/2, 1]$ and, if $\sigma_* \le \sigma_e$, then $(1 + c_*) \le 2$. Therefore, for $\sigma_* \le \sigma_e$ the bound (12.16) can be written in a ruder, but simpler form

$$|\!|\!| e |\!|\!|^2 \le 2 \Big[\| (\nabla v + \mathbf{z}) \|^2_{\mathbf{L}^2(\Omega)} + \| \frac{1}{\sqrt{\sigma_\dagger}} \big(f - \sigma v - \operatorname{div} \mathbf{z} \big) \|^2_{L_2(\Omega)} \Big] . \tag{12.24}$$

At $\sigma \equiv 0$ the bound, (12.16) differs from (12.5) with $\epsilon = 1$ only by the coefficient before the second term on the right. In Sect. 12.4, we show that for the finite element solutions of the problem in the domains with the sufficiently smooth boundaries this coefficient will have, indeed, the order of $\sigma_e^{-1} = \mathcal{O}(h^2)$ instead of the constant c_Ω in (12.5). In general, e.g., at piece wise constant or constant σ, it is possible to use also rough values $\sigma_e = \sigma_* = 1/c_F$ in (12.16) and (12.24), where $c_F = c_F(\Omega, \Gamma_D)$ is the constant from the Friedrichs type inequality

$$\| (\sigma_\dagger - \sigma)^{1/2} \phi \|^2_{L_2(\Omega)} \le c_F \| (\sigma_\dagger - \sigma)^{1/2} \nabla \phi \|^2_{L_2(\Omega)} , \quad \forall \phi \in \mathring{H}^1_{\Gamma_D}(\Omega) . \tag{12.25}$$

However, in this case minimization of the right parts of the a posteriori bounds with respect to $\mathbf{z}$ becomes often unavoidable.

12.3.2 Reaction-Diffusion Problem with any Constant $\sigma \ge 0$

The case $\sigma = \text{const}$, considered in this subsection, has an independent significance and, besides, will allow us simpler transfer to more specific a posteriori error bounds for the approximate solutions by the finite element method. Below we reformulate the basic assumption (12.15) in a two simpler forms and give the respective version of the preceding Lemma.

In general, when $v \in \mathring{H}^1_{\Gamma_D}(\Omega)$ is any approximation for u, σ_* can be defined from the inequality

$$|u - v|_1^2 / \|u - v\|_0^2 \geq \sigma_* > 0\,. \tag{12.26}$$

If $v = u_G$ is the approximate solution by the Galerkin method in the subspace $\mathscr{V}(\Omega) \subset \mathring{H}^1_{\Gamma_D}(\Omega, \Delta)$, then this condition can be relaxed and it is sufficient that σ_* satisfies

$$|u - v|_1^2 / \|u - Qu\|_0^2 \geq \sigma_* > 0\,. \tag{12.27}$$

where Q is the operator of orthogonal projection $L_2(\Omega) \to \mathscr{V}(\Omega)$, i.e., such that for $\forall \phi \in L_2(\Omega)$ we have $(Q\phi, \psi)_\Omega = (\phi, \psi)_\Omega\,, \quad \forall\, \psi \in \mathscr{V}(\Omega)\,.$

Let function $\hat{f}(x)$, $x \in \Omega$, be such that $\hat{f}(x) = \Pi_r^l f$ for $x \in \tau_r$, $r = 1, 2, \ldots, \mathscr{R}$ and l is some nonnegative integer, e.g., related to the accuracy of the approximation v. In Theorem below domains τ_r can be understood as arbitrary convex subdomains of some partition of the domain

$$\Omega = \text{interior}\{\bigcup_1^{\mathscr{R}} \overline{\tau}_r\}\,, \quad \tau_r \cup \tau_{r'} = \varnothing,\ r \neq r', \quad \text{diam}[\tau_r] = h_r\,,$$

for which the Poincaré inequalities hold, see, e.g., Nazarov and Poborchi [29],

$$\inf_{c \in \mathbb{R}} \|\phi - c\|_{L_2(\tau_r)} \leq \frac{h_r}{\pi} |\phi|_{H^1(\tau_r)}\,, \quad \phi \in H^1(\tau_r)\,.$$

Lemma 12.2 *Let $f \in L_2(\Omega)$, $0 < \sigma = \text{const}$, σ_* satisfy the inequality (12.26), $\Gamma_D = \partial\Omega$, $\psi_D \in H^{1/2}(\partial\Omega)$. Then, for any function $v \in H^1(\Omega)$ that satisfies the boundary condition on Γ_D and any $\mathbf{z} \in \mathbf{H}(\Omega, \text{div})$ we have*

$$|\!|\!| e |\!|\!|^2 \leq \Theta \mathscr{M}\,, \qquad \mathscr{M} = \mathscr{M}(\sigma, \sigma_*, f, v, \mathbf{z})\,, \tag{12.28}$$

where

$$\mathscr{M} = \|\nabla v + \mathbf{z}\|^2_{\mathbf{L}(\Omega)} + \theta \|f - \sigma v - \text{div}\,\mathbf{z}\|^2_{L_2(\Omega)}\,, \tag{12.29}$$

and

$$\Theta = \left\{ \begin{array}{ll} 2/(1 + \mathbb{k})\,, & \forall\, \sigma \in [0, \sigma_*] \\ 1\,, & \forall\, \sigma > \sigma_* \end{array} \right\}, \ \theta = \left\{ \begin{array}{ll} 1/\sigma_*\,, & \forall\, \sigma \in [0, \sigma_*] \\ 1/\sigma\,, & \forall\, \sigma > \sigma_* \end{array} \right\}. \tag{12.30}$$

with $\Bbbk = \sigma/\sigma_*$. *Besides, for* $\sigma \in [0, \sigma_*]$ *and* $\sigma \geq \sigma_*$, *respectively, we have the bounds*

$$\begin{aligned} &|\!|\!| e |\!|\!|^2 \leq \Theta_1 \mathscr{M}(\sigma, \sigma_*, \hat{f}, v, \mathbf{z}) + \textstyle\sum_r \frac{h_r^2}{\varepsilon \pi^2} \int_{\tau_r} (f - \Pi_r^l f)^2 dx\,, \quad \forall\, \varepsilon > 0\,, \\ &|\!|\!| e |\!|\!|^2 \leq \Theta_2 \mathscr{M}(\sigma, \sigma_*, \hat{f}, v, \mathbf{z}) + \textstyle\sum_r \frac{1}{\sigma} \int_{\tau_r} (f - \Pi_r^l f)^2 dx\,, \end{aligned} \tag{12.31}$$

where

$$\Theta_1 = \begin{cases} (2+\varepsilon)/(1+\Bbbk)\,, & 0 \leq \sigma \leq \sigma_*/(1+\varepsilon)\,, \\ 1+\varepsilon\,, & \sigma_*/(1+\varepsilon) \leq \sigma \leq \sigma_*\,, \end{cases}$$

$$\Theta_2 = 1 + 1/(1+\Bbbk^{-1})\,.$$

If $v = u_G$ *is the approximate solution by the method of Galerkin in the space* $\mathscr{V}(\Omega) \subset \overset{\circ}{H}{}^1_{\Gamma_D}(\Omega, \Delta)$, *then the bound (12.28)–(12.30) takes place with* σ_*, *satisfying to the inequality (12.27) and* $\mathbf{z} = \mathbf{z}_G := -\nabla u_G$, *i.e.,*

$$\begin{aligned} &|\!|\!| u_G - u |\!|\!|^2 \leq \Theta \mathscr{M}_G\,, \\ &\mathscr{M}_G = \mathscr{M}_G(\sigma, \sigma_*, f, u_G, \mathbf{z}_G) = \theta \| f - \sigma u_G - \operatorname{div} \mathbf{z}_G \|^2_{L_2(\Omega)}\,. \end{aligned} \tag{12.32}$$

Proof The direct proof of Lemma 12.2 is given in [25]. The proof of the bound (12.28)–(12.30) follows also from (12.16) at $\sigma_* = \sigma_{\mathrm{e}}$. □

Remark 12.1 Other ways of obtaining majorants similar to (12.28)–(12.30) and 12.31) are possible. One can consider the subsidiary problem

$$-\Delta u + \lambda u = f_{\lambda,\sigma}\,, \quad x \in \Omega\,, \quad u\big|_{\partial\Omega} = 0\,, \tag{12.33}$$

with an arbitrary $\lambda \geq \sigma_*$ and $f_{\lambda,\sigma} = f + (\lambda - \sigma)u$, whose solution is the same as for the problem (12.1) at the $\Gamma_N = \varnothing$ and $\psi_D \equiv 0$. In the respective to (12.33) Aubin's majorant for the approximation v of the solution to the problem (12.33), the approximation of the problem (12.1) can be used. The substitution of $f_{\lambda,\sigma} = f + (\lambda - \sigma)u$ in this Aubin's majorant, the application of the Cauchy inequality with ε, the use of the inequality 12.26), and some manipulations produce the subsidiary majorant, depending on λ, ε and β. At that, β plays similar role to the one at the obtaining (12.16). In this way, we come to the set of majorants similar to those in Lemma 12.2, and, obviously, by changing the choice of λ, β, and ε, we can change the weights before the first and second norms in the majorants. If the ratio of the weights is fixed, a suitable majorant can be obtained by the minimization of the subsidiary majorant with respect to λ, β and ε.

The bound (12.28)–(12.30) generalizes the Repin-Frolov's bound (12.5) for $\sigma \equiv 0$ upon interval $\sigma \in [0, \sigma_*]$ with $\sigma_* \leq 1/c_\Omega$ and the coefficients not greater than in (12.5), i.e., $\Theta \leq 2$ and $\theta \leq c_\Omega$. At the same time (12.28)–(12.30) can be considered as an updated Aubin's bound (12.4), coinciding with it at $\sigma \geq \sigma_*$ and well defined for all $\sigma \geq 0$. An additional unique feature of the bound (12.28)–(12.30), if to compare with other attempts [15, 33, 35] to update the Aubin's bound, is that its accuracy can be significantly improved by the use of improved estimates of σ_* for concrete numerical methods.

12.4 Consistent A Posteriori Majorants for Finite Element Method Errors

For specific classes of approximate solutions and, in particular, for solutions by the finite element method, the critical values σ_* of the reaction coefficient in the derived error majorants can be successfully estimated. In this section, we assume that $\sigma \geq 0$ is a piece wise constant function (12.2) satisfying $\mu_1 \leq \sigma \leq \mu_2, \ \forall x \in \Omega$.

Lemma 12.3 *Let $\Gamma_D = \partial\Omega$, $\sigma \geq 0$ be the piece wise constant function satisfying (12.2), $\psi_D \equiv 0$, $f \in H^{-1}(\Omega)$, the finite element assemblage generates the space $\mathcal{V}_{h,p}(\Omega)$, $p \geq 1$, and $e_{\rm fem} = u_{\rm fem} - u$. Then*

$$\|e_{\rm fem}\|_0 \leq c_\dagger h |e_{\rm fem}|_1 \,, \quad c_\dagger = \sqrt{\frac{\mu_2}{\mu_1}} c_\triangle c_{\rm ap} \,, \tag{12.34}$$

with the constants $c_\triangle$, $c_{\rm ap}$, defined below, see (12.38), (12.39).

Proof Let us set $\kappa = +\sqrt{\sigma}$ and introduce notations $u_\circ$, $u_{\rm fe}$ and u_s for the functions minimizing $\|\kappa(u-\phi)\|_0^2$, $|u-\phi|_1^2$, and $h^{-2}\|\kappa(u-\phi)\|_0^2 + |u-\phi|_1^2$, respectively, among all $\phi \in \mathring{\mathcal{V}}_h(\Omega)$ and notations for the respective errors $e_\circ = u_\circ - u$, $e_{\rm fe} = u_{\rm fe} - u$ and $e_s = u_s - u$. Since $u_{\rm fem}$ minimizes $|||u-\phi|||^2$, $\phi \in \mathring{\mathcal{V}}_h(\Omega)$, we conclude that

$$|e_{\rm fem}|_1^2 + \|\kappa\, e_{\rm fem}\|_0^2 \leq |u - \widetilde{u}|_1^2 + \|\kappa\,(u - \widetilde{u})\|_0^2 \,, \tag{12.35}$$

where $\widetilde{u}$ can be any from functions $\widetilde{u} = u_\circ$, $u_{\rm fe}$, u_s. If to take into attention the inequalities $\|e_{\rm fem}\|_0 \geq \|e_\circ\|_0$ and $|e_{\rm fem}|_1 \geq |e_{\rm fe}|_1$, following from the definitions of functions $u_\circ$ and $u_{\rm fe}$, then (12.35) implies

$$\begin{gathered} \|\kappa\, e_{\rm fem}\|_0 \leq \|\kappa\, e_{\rm fe}\|_0 \,, \\ |e_{\rm fem}|_1 \leq |e_\circ|_1 \,. \end{gathered} \tag{12.36}$$

Let $\varphi \in \mathring{H}^1(\Omega)$ be the solution of the problem

$$a_\Omega(v,\varphi) = (v, e_{\rm fe})_\Omega\,, \quad \forall\, v \in \mathring{H}^1(\Omega). \tag{12.37}$$

Obviously, $e_{\rm fe} \in L_2(\Omega)$, and, as a consequence, at sufficiently smooth domain $\varphi \in H^2(\Omega)$ and, see Ladyzhenskaya [27],

$$\|\varphi\|_2 \le c_\triangle \|e_{\rm fe}\|_0\,, \quad c_\triangle = c_\triangle(\Omega)\,. \tag{12.38}$$

The function φ can be approximated by the function $\widetilde{\varphi} \in \mathring{\mathscr{V}}_{h,p}(\Omega)$, for which

$$|\varphi - \widetilde{\varphi}|_1^2 \le c_{\rm ap}^2 h^2 \|\varphi\|_2^2 \le c_\triangle^2 c_{\rm ap}^2 h^2 \|e_{\rm fe}\|_0^2\,. \tag{12.39}$$

Since in Lemma 12.3 only the constant $c_{\rm ap}$ but not the function $\widetilde{\varphi}$ itself is used, one can imply by $\widetilde{\varphi}$ any approximation providing a good constant. Hence, in 2D $\widetilde{\varphi}$ can be the interpolation and, in general, quasi-interpolation of Scott and Zhang [34], or Bartels et al. [9], or one of the functions $\widetilde{\varphi}_\circ$, $\widetilde{\varphi}_{\rm fe}$, $\widetilde{\varphi}_s$ etc.

Estimating $\|e_{\rm fe}\|_0$ by means of the Aubin-Nitsche trick [5] for the problem (12.37) and using the bound (12.39), we get:

$$\begin{gathered}\|e_{\rm fe}\|_0^2 = a_\Omega(e_{\rm fe},\varphi) \le \inf_{w\in\mathring{\mathscr{V}}_h(\Omega)} |a_\Omega(e_{\rm fe},\varphi - w)| \le \\ |e_{\rm fe}|_1 \inf_{w\in\mathring{\mathscr{V}}_h(\Omega)} |\varphi - w|_1 \le |e_{\rm fe}|_1 |\varphi - \widetilde{\varphi}|_1 \le c_\triangle c_{\rm ap} h |e_{\rm fe}|_1 \|e_{\rm fe}\|_0\,.\end{gathered} \tag{12.40}$$

This bound together with the inequality (12.36) and the definitions of functions $e_{\rm fe}$, $e_{\rm fem}$ results in the bound (12.34). □

Theorem 12.4 *Let $\Gamma_D = \partial\Omega$, $\psi_D \equiv 0$, and $u \in \mathring{H}^1(\Omega,\Delta)$. Let also the finite element assemblage generate the space $\mathring{\mathscr{V}}^0_{h,p}(\Omega) \subset \mathring{H}^1(\Omega)$, $p \ge 1$, and $u_{\rm fem}$ be the solution by the finite element method. Then for σ satisfying $0 \le \sigma \le \sigma_* = 1/(c_\dagger h)^2$ with $c_\dagger$, as in (12.34), and any $\mathbf{z} \in \mathbf{H}(\Omega,{\rm div})$ we have*

$$\begin{gathered}|\!|\!| e_{\rm fem} |\!|\!|^2 \le \tfrac{2}{1+{\rm c}_\dagger^2 h^2\sigma}\mathscr{M}^{(1)}_{\rm fem}(\sigma, f, \mathbf{z})\,,\\ \mathscr{M}^{(1)}_{\rm fem}(\sigma, f, \mathbf{z}) = \Big[|\nabla u_{\rm fem} + \mathbf{z}|^2_{\mathbf{L}_2(\Omega)} + {\rm c}_\dagger^2 h^2 \|f - \sigma u_{\rm fem} - {\rm div}\,\mathbf{z}\|^2_{L_2(\Omega)}\Big]\,.\end{gathered} \tag{12.41}$$

Under the condition $\mathscr{A}$), for $\sigma \le \sigma_*/(1+\varepsilon)$ it holds also the bound

$$|\!|\!| e_{\rm fem} |\!|\!|^2 \le \frac{2+\varepsilon}{1+{\rm c}_\dagger^2 h^2\sigma}\mathscr{M}^{(1)}_{\rm fem}(\sigma, \hat{f}, \mathbf{z}) + \sum_r \frac{h_r^2}{\varepsilon\pi^2}\int_{\tau_r}(f - \Pi_r^p f)^2 dx\,, \quad \forall\,\varepsilon > 0\,. \tag{12.42}$$

Proof Since $\mathscr{M}^{(1)}_{\rm fem}(\sigma, f, \mathbf{z}) = \mathscr{M}(\sigma, \sigma_*, f, v, \mathbf{z})$ with σ_* defined according to (12.34), Theorem is a direct consequence of Lemmas 12.2 and 12.3. □

Similar a posteriori error bounds can be derived in a quite similar way for the finite element method solutions to the $2n^{\rm th}$-order elliptic equations, $n \geq 1$, see Korneev [24]. The most complicated for their practical applications is the evaluation of the constant $c_\triangle$. However, in many cases such estimates are well known. For instance, if the domain is convex, then $c_\triangle = 1$, see Ladyzhenskaya [27, (6.5) in ch. II].

Existence of the constant $c_\triangle$ puts some restrictions on the smoothness of the boundary and coefficients of the equation (if they are not constant). At the same time, there is a possibility to avoid the mentioned additional restrictions, except for those related to the suitable approximation operator. If there exists some interpolation type or other approximation operator with locally defined approximations for functions from $H^1(\Omega)$, then constants in the a posteriori bounds may depend only on the local approximation properties of the finite element space. A good example of such an operator is the quasi-interpolation operator of Scott and Zhang [34], which will be used below to illustrate the statement. We start from the description of its properties.

Let $\Omega \subset \mathbb{R}^m$, $m \geq 2$, be a bounded Lipschitz domain, which is the domain of the quasiuniform triangulation $\mathscr{T}_h$ with vertices $x^{(i)}$, $i = 1, 2, \ldots, I$, and simplices τ_r of diameters not greater h. For simplicity it is assumed that faces of simplices are plain and that the following quasiuniformity conditions are fulfilled:

$$0 < \underline{c} \leq \underline{\rho}_r / h_r\,, \quad \hat{\alpha}^{(1)} h \leq h_r \leq h\,, \tag{12.43}$$

where $\underline{\rho}_r$ and h_r are the radius of the largest inscribed sphere and the diameter of simplex τ_r, respectively. To each vertex $x^{(i)}$, we relate $(m-1)$-dimensional simplex $\tau_i^{(m-1)}$, which is the face of one of the simplices τ_r and has $x^{(i)}$ for the vertex. For m vertices of the simplex $\tau_i^{(m-1)}$ we will use also notations $z_l^{(i)}$, $l = 1, 2, \ldots, m$, assuming for definiteness that $z_1^{(i)} = x^{(i)}$. Clearly the choice of the face $\tau_i^{(m-1)}$ is not unique, but for $x^{(i)} \in \partial\Omega$ we always take one of the faces $\tau_i^{(m-1)} \subset \partial\Omega$. We will formulate the result of Scott and Zhang using the simpler notations $\mathscr{V}_\triangle(\Omega)$, $\mathscr{V}_{\rm tr}(\partial\Omega)$, and $\mathring{\mathscr{V}}_\triangle(\Omega)$ for the space of continuous piece wise linear functions $\mathscr{V}^0_{h,1}(\Omega)$, its trace on the boundary, and its subspace of functions, vanishing on the boundary, respectively.

We define functions $\theta_i \in \mathscr{P}_1(\tau_i^{(m-1)})$, satisfying equations

$$\int_{\tau_i^{(m-1)}} \theta_i \lambda_l^{(i)}\, dx = \delta_{1,l}\,, \quad l = 1, 2, .., m\,,$$

where $\lambda_l^{(i)}$ are the barycentric coordinates in $\tau_i^{(m-1)}$, corresponding to the vertices $z_l^{(i)}$, and $\delta_{i,l}$ is the Kronecker's symbol. If $\phi_i \in \mathscr{V}_\triangle(\Omega)$ are the basis functions in

$\mathscr{V}_{\triangle}(\Omega)$, defined by the equalities $\phi_i(x_j) = \delta_{i,j}$, $i, j = 1, 2, .., I$, then for any $v \in H^1(\Omega)$ the quasi-interpolation $\mathscr{I}_h v$ is the function

$$\mathscr{I}_h v = \sum_{i=1}^{I} \left(\int_{\tau_i^{(m-1)}} \theta_i v \, dx \right) \phi_i(x) .$$

Lemma 12.4 *The quasi-interpolation operator $\mathscr{I}_h : H^1(\Omega) \to \mathscr{V}_{\triangle}(\Omega)$ is a projection and has the following properties:*

(a) $\mathscr{I}_h v : H^1(\Omega) \mapsto \mathscr{V}_{\triangle}(\Omega)$ *and, if* $v \in \mathscr{V}_{\triangle}(\Omega)$, *then* $\mathscr{I}_h v = v$,
(b) $(v - \mathscr{I}_h v) \in \mathring{H}^1(\Omega)$, *if* $v|_{\partial\Omega} \in \mathscr{V}_{\mathrm{tr}}(\partial\Omega)$,
(c) $\|v - \mathscr{I}_h v\|_{t,\Omega} \le c_{\mathrm{sz}}(t,s) h^{s-t} \|v\|_{s,\Omega}$ *for* $t = 0, 1$, $s = 1, 2$, *and* $\forall v \in H^s(\Omega)$,
(d) $|\mathscr{I}_h v|_{1,\Omega} \le \check{c}_{\mathrm{sz}} | v |_{1,\Omega}$ *and* $\|\mathscr{I}_h v\|_{1,\Omega} \le \hat{c}_{\mathrm{sz}} \| v \|_{1,\Omega}$, $\forall v \in H^1(\Omega)$,

where $c_{\mathrm{sz}}(s,t)$, $\check{c}_{\mathrm{sz}}$, *and* $\hat{c}$ *are positive constants, depending on* $\underline{c}$.

Proof Scott and Zhang [34] proved a much more general result. In the given form the Lemma was formulated and proved by Xu and Zou [37]. □

Theorem 12.5 *Let* $\Gamma_D = \partial\Omega$, $\psi_D \equiv 0$, $u \in \mathring{H}^1(\Omega, \Delta)$. *Let also the finite element assemblage satisfy the condition* $\mathscr{A}$*) and generate the space* $\mathring{\mathscr{V}}_{\triangle}(\Omega) \subset \mathring{H}^1(\Omega)$ *and* $\mathbf{z} \in \mathbf{H}(\Omega, \mathrm{div})$. *Then at any* $\sigma \equiv \mathrm{const} \in [0, 1/(c_{\mathrm{sz}}(0,1)h)^2]$ *there holds the bound*

$$|||e_{\mathrm{fem}}|||^2 \le \Theta_{\mathrm{sz}} \mathscr{M}^{(2)}_{\mathrm{fem}}(\sigma, f, \mathbf{z}) , \quad \mathscr{M}^{(2)}_{\mathrm{fem}}(\sigma, f, \mathbf{z}) = \mathscr{M}(\sigma, \theta_{\mathrm{sz}}^{-1}, f, u_{\mathrm{fem}}, \mathbf{z}) , \tag{12.44}$$

where

$$\Theta_{\mathrm{sz}} = \frac{1 + \widetilde{c}^2_{\mathrm{sz}}(1,1)}{1 + c^2_{\mathrm{sz}}(0,1) h^2 \sigma} , \quad \theta_{\mathrm{sz}} = c_{\mathrm{sz}}(0,1)^2 h^2 , \tag{12.45}$$

and $\widetilde{c}_{\mathrm{sz}}(1,1)$ *is the constant, depending only upon* $\underline{c}$ *and* $\hat{\alpha}^{(1)}$ *found in (12.43).*

Proof For any $w \in \mathring{\mathscr{V}}^0_{h,1}(\Omega)$ we have the equality

$$|||e_{\mathrm{fem}}|||^2 = \int_{\Omega} \left[\nabla(e_{\mathrm{fem}}) \cdot \nabla(e_{\mathrm{fem}}) + \sigma e_{\mathrm{fem}} e_{\mathrm{fem}} \right] =$$

$$= \int_{\Omega} \left[(\nabla u_{\mathrm{fem}} + \mathbf{z}) \cdot \nabla(e_{\mathrm{fem}} + w) - (\mathbf{z} + \nabla u) \cdot \nabla(e_{\mathrm{fem}} + w) + \right. \tag{12.46}$$

$$\left. + \sigma (u_{\mathrm{fem}} - u)(e_{\mathrm{fem}} + w) \right] .$$

Integration by parts of the second summand in square brackets of the right part and application of the Cauchy inequality with $\epsilon > 0$ result in the inequality

$$\begin{gathered}
|\!|\!| e_{\mathrm{fem}} |\!|\!|^2 = \int_{\Omega} \left[(\nabla u_{\mathrm{fem}} + \mathbf{z}) \cdot \nabla (e_{\mathrm{fem}} + w) + \right. \\
\left. \big(\operatorname{div} \mathbf{z} + \Delta u + \sigma (u_{\mathrm{fem}} - u) \big) (e_{\mathrm{fem}} + w) \right] \leq \\
\left\{ \| \nabla u_{\mathrm{fem}} + \mathbf{z} \|_0^2 + \tfrac{1}{\epsilon} \| f - \sigma u_{\mathrm{fem}} - \operatorname{div} \mathbf{z} \|_0^2 \right\}^{1/2} \times \\
\left\{ \| \nabla (e_{\mathrm{fem}} + w) \|_0^2 + \epsilon \| e_{\mathrm{fem}} + w \|_0^2 \right\}^{1/2}
\end{gathered} \tag{12.47}$$

According to Lemma 12.4 and the definition of the operator Q of L_2-projection upon $\mathring{\mathscr{V}}^0_{h,1}(\Omega)$, for $\phi - Q\phi$ with any $\phi \in H^1(\Omega)$, there are valid the bounds

$$\begin{aligned}
&\| \phi - Q\phi \|_0 \leq \| \phi \|_0 \,, \\
&\| \phi - Q\phi \|_0 \leq c_{\mathrm{sz}}(0,1) h \| \nabla \phi \|_0 \,, \\
&\| \nabla (\phi - Q\phi) \|_0 \leq \widetilde{c}_{\mathrm{sz}}(1,1) \| \nabla \phi \|_0 \,,
\end{aligned} \tag{12.48}$$

in which the constant $\widetilde{c}_{\mathrm{sz}}(1,1)$ depends only on $\underline{c}$ and $\hat{\alpha}^{(1)}$. The proof is needed only for the last bound, and it follows by the relations

$$\| \nabla (\phi - Q\phi) \|_0 \leq \| \nabla (\phi - \mathscr{I}_h \phi) \|_0 + \| \nabla (\mathscr{I}_h \phi - Q\phi) \|_0 \leq \check{c}_{\mathrm{sz}} \| \nabla \phi \|_0 +$$

$$c_{1,0} h^{-1} \| \mathscr{I}_h \phi - Q\phi \|_0 \leq \check{c}_{\mathrm{sz}} \| \nabla \phi \|_0 + c_{1,0} h^{-1} \Big[\| \mathscr{I}_h \phi - \phi \|_0 + \| \phi - Q\phi \|_0 \Big] \leq$$

$$\Big(\check{c}_{\mathrm{sz}} + 2 c_{1,0} c_{\mathrm{sz}}(0,1) \Big) \| \nabla \phi \|_0 = \widetilde{c}_{\mathrm{sz}}(1,1) \| \nabla \phi \|_0 \,,$$

where $c_{1,0}$ is the constant in the inverse inequality $\| \nabla (\mathscr{I}_h \phi - Q\phi) \|_0 \leq c_{1,0} h^{-1} \| \mathscr{I}_h \phi - Q\phi \|_0$. Therefore, $\widetilde{c}_{\mathrm{sz}}(1,1) = \check{c}_{\mathrm{sz}} + 2 c_{1,0} c_{\mathrm{sz}}(0,1)$.

It is worth noting, that the third inequality (12.48), indicating stability in $H^1(\Omega)$ of L_2-projection, was proved by Bramble and Xu [10] differently with the differently defined constant $\widetilde{c}_{\mathrm{sz}}(1,1)$.

For the reason that $w = Qe_{\rm fem} \in \mathring{\mathscr{V}}^0_{h,1}(\Omega)$, it can be adopted $w = Qe_{\rm fem}$. Combining with (12.48) and setting $\epsilon = \sigma_{\rm sz} := (c_{\rm sz}(0,1)h)^{-2}$ lead to the bound

$$\|\nabla(e_{\rm fem}+w)\|_0^2 + \sigma_{\rm sz}\|e_{\rm fem}+w\|_0^2 =$$

$$\|\nabla(e_{\rm fem}+w)\|_0^2 + \beta\sigma\|e_{\rm fem}+w\|_0^2 + +(\sigma_{\rm sz}-\beta\sigma)\|e_{\rm fem}+w\|_0^2 \le \tag{12.49}$$

$$\widetilde{c}^2_{\rm sz}(1,1)\|\nabla e_{\rm fem}\|_0^2 + \beta\sigma\|e_{\rm fem}\|_0^2 + \tfrac{\sigma_{\rm sz}-\beta\sigma}{\sigma_{\rm sz}}\|\nabla e_{\rm fem}\|_0^2 .$$

On the basis of (12.49) we conclude that

$$\|\nabla(e_{\rm fem)}+w\|_0^2 + \sigma_{\rm sz}\|e_{\rm fem}+w\|_0^2 \le \frac{1+\widetilde{c}^2_{\rm sz}(1,1)}{1+\Bbbk}\big[\,\|\nabla e_{\rm fem}\|_0^2 + \sigma\|e_{\rm fem}\|_0^2\,\big] \tag{12.50}$$

with $\Bbbk = \sigma/\sigma_{\rm sz}$. Now from (12.47) and (12.50) the Theorem follows . □

Remark 12.2 Quasi-interpolation operator $\mathscr{I}_h$ is defined in [37] on triangulations, satisfying the conditions of the shape regularity $0 < \underline{c} \le \underline{\rho}_r/h_r$, $h_r \le h$, with preserving the properties (a), (b) and the properties (c), (d) taking the local form

$$c)\ \|v-\mathscr{I}_h v\|_{t,\tau_r} \le c_{\rm sz}(t,s)h_r^{s-t}\|v\|_{s,\delta_r}\,, \quad t=0,1,\ \ s=1,2,\quad \forall v\in H^1(\delta_r)\,,$$

$$d)\ |\mathscr{I}_h v|_{1,\tau_r} \le \check{c}_{\rm sz}|\,v\,|_{1,\delta_r}, \quad \text{and}\ \ \|\mathscr{I}_h v\|_{1,\tau_r} \le \hat{c}_{\rm sz}\|\,v\,\|_{1,\delta_r}, \quad \forall v\in H^1(\delta_r)\,, \tag{12.51}$$

where $c_{\rm sz}(t,s) = {\rm const}$, $\delta_r = {\rm interior}\{\cup_\varkappa \overline{\tau}_\varkappa : \overline{\tau}_\varkappa \cap \overline{\tau}_r \ne \varnothing\}$ and $r = 1,2,\dots,\mathscr{R}$. More over, these authors designed also the quasi-interpolation operators $\mathscr{I}_{h,p} : H^1(\Omega) \to \mathscr{V}_{h,p}(\Omega)$, for which again the properties (a), (b) are preserved, but in (12.51) $s = 1,2,\dots,p+1$. This quasi-interpolation operator allows to expand the a posteriori error bounds (12.44)–(12.45) on the solutions by the finite element methods from the spaces $\mathscr{V}_{h,p}(\Omega)$, $p > 1$.

The bounds (12.41), (12.42) of Theorem 4 essentially use some global properties of the finite element solutions, see Lemma 12.3. In the bound (12.44)–(12.45) of Theorem 12.5, we see only the constants defined by local properties of the quasi-interpolation operator.

12.5 Consistency and Local Effectiveness

For the right parts of the a posteriori error bounds (12.41), (12.44)–(12.45), and (12.42) we introduce the notations $\eta_k,\ k=1,2,$

$$\eta_1^2=\frac{2}{1+\mathrm{c}_\dagger^2h^2\sigma}\mathscr{M}_{\mathrm{fem}}^{(1)}(\sigma,f,\mathbf{z})\,,\quad \eta_2^2=\Theta_{\mathrm{sz}}\mathscr{M}_{\mathrm{fem}}^{(2)}(\sigma,f,\mathbf{z})\,,$$

and $\eta_{1,\varepsilon}$, respectively. The error estimators $\eta_k,\ \eta_{1,\varepsilon}$ are consistent, they are computable with the use of the testing fluxes which make the estimators sharp and which can be calculated by a number of simple flux recovery procedures of linear complexity. These facts lie practically on the surface and are established similarly for all $\eta_k,\ \eta_{1,\varepsilon}$. Hence, they are discussed below briefly for η_1. In support of the effectiveness of the error indicators, we present also the inverse like inequality at some choice of the testing fluxes.

If $u_{\mathrm{fem}}\in\mathscr{Y}_{h,p}(\Omega)\subset H^1(\Omega),\ \mathscr{Y}_{h,p}(\Omega)\not\subseteq H^2(\Omega)$ and $u\in H^l(\Omega),\ 2\le l\le p+1,\ p\ge 1$, then it is natural to require that for the recovered flux $\mathbf{z}$ the estimates of convergence

$$h^{-1}\|\nabla u+\mathbf{z})\|_{\mathbf{L}_2(\Omega)},\,\|\mathrm{div}(\nabla u+\mathbf{z})\|_0\le\tilde{c}_lh^{l-2}\|u\|_{l,\Omega}\,,\qquad \tilde{c}_l=\mathrm{const}\,,\tag{12.52}$$

hold with the orders corresponding to the convergence estimates of $\|e_{\mathrm{fem}}\|_k\le c_{k,l}h^{l-k}\|u\|_l,\ k=0,1$. From these bounds, it easily follows that for $\sigma\le c_\dagger^2h^{-2}$

$$|\!|\!|e_{\mathrm{fem}}|\!|\!|\le c_lh^{l-1}\|u\|_l\,,\quad \eta_1(e_{\mathrm{fem}})\le c_lh^{l-1}\|u\|_l,\qquad c_l,\hat{c}_1=\mathrm{const}\,,\tag{12.53}$$

see [23–25] for additional details and discussion.

Thus, according to (12.53) the a posteriori bound (12.41) is consistent for $\sigma\le c_\dagger^2h^{-2}$, and as the a priori bound is unimprovable in the order, so the same is true for the bound (12.41). In other words, in the class of solutions, for which the a priori bound is unimprovable in the order, there exist such that

$$\eta_1(e_{\mathrm{fem}})\le\mathbb{C}_\star|\!|\!|\,e_{\mathrm{fem}}\,|\!|\!|\,.\tag{12.54}$$

Indeed, let $f\in L_2(\Omega)$, $d=2$, and the domain is such that the inequality $\|v\|_2\le c_\circ\|F\|_0$ holds for solutions of the problems $-\Delta v=F,\ u\Big|_{\partial\Omega}=0$ uniformly in $F\in L_2(\Omega)$, see Ladyzhenskaya [27]. Then the inequalities (12.53) with $l=2$ are fulfilled as well. At the same time, such $f\in L_2(\Omega)$ exists that

$$\begin{gathered}h\|u\|_2\le c_2\,\inf_{\phi\in\mathscr{Y}_h(\Omega)}\|\nabla(u-\phi)\|_0=c_2\,\|\nabla e_{\mathrm{fe}}\|_0\le\\ c_2\,\|\nabla e_{\mathrm{fem}}\|_0\le c_2|\!|\!|\,\nabla e_{\mathrm{fem}}\,|\!|\!|\,,\quad c_2=\mathrm{const}\,.\end{gathered}\tag{12.55}$$

The first of these inequalities follows from the estimates of the N-width of the compact of functions $v \in \mathring{H}^1(\Omega)$, $\|v\|_1 = 1$, for $N = h^{-2}$, see in [31] Ch. 4, Section 4.1. In turn, (12.54) follows from (12.53), (12.55), which together with (12.41) yield the two sided bound

$$|\!|\!| e_{\rm fem} |\!|\!|^2 \le \eta_1(e_{\rm fem}) \le \mathbb{C}_\star |\!|\!| e_{\rm fem} |\!|\!|^2 . \tag{12.56}$$

The error indicators η_k, $\eta_{1,\varepsilon}$ are computable for a wide range of problems and FEM's, and there are numerous works on the flux recovery procedures, which can be used for obtaining the testing fluxes $\mathbf{z}$ satisfying (12.52) and, therefore, making this indicators sharp. Majority of them, following the renown SPR of [39], were created in the relation with the development of the residual type error estimators $\eta = ||\nabla u_{\rm fem} + G u_{\rm fem}||_{\mathbf{L}^2(\Omega)}$, where G is the flux recovery operator. For this reason, these procedures were aimed at achieving as high as possible accuracy, in particular, superconvergence or even ultraconvergence [36, 38]. Under some assumptions on the finite element meshes, smoothness of the exact solutions, finite element discretization and the operator G, these properties were approved by the analysis and by the numerical practice. As is noted by some authors, flux recovery procedures perform "astonishingly well", see, e.g., [1, 36]. Clearly, the superconvergence means that even better than (12.52) in the order bounds hold. For instance, the superconvergence of fluxes produced by the simple and low cost procedures of linear complexity such as

(i) weighted averaging,
(ii) local L^2-projection, and
(iii) the local discrete least squares fitting

for the first order finite element methods was studied in [36]. Under the assumptions $u \in \mathring{H}^1(\Omega) \cap H^2(\Omega) \cap W^3_\infty(\Omega)$ and mild regularity of the mesh, Theorem 4.2 of this work establishes that $||\nabla u_{\rm fem} + G u_{\rm fem}||_{\mathbf{L}^2(\Omega)} = \mathscr{O}(h^{1+\rho})$ with $\rho > 0$ depending on the mesh. Superconverging averaging techniques for p Lagrange elements were studied in [8]. There are many other papers devoted to the expansion of efficient recovery techniques on the more irregular meshes, problems with the discontinuous coefficients etc. It is worth to stress again that the inequalities (12.52) do not assume any superconvergence, but *only* retention by the testing flux $\mathbf{z}$ of the orders of accuracy of the finite element flux $\mathbf{z}_{\rm fem} = -\nabla u_{\rm fem}$. In general, the proof of (12.52) does not differ much from the proofs of similar bounds for approximations by functions from the finite element space. In particular, according to Lemma 4.3 of [1], see also Corollary 4.2 there, under conditions of Theorems 12.4 and 12.5, for the flux recovery procedures (i)–(iii) the inequalities (12.52) are fulfilled.

Obviously also, that for the global $L^2(\Omega)$ orthogonal projection of $\mathbf{z}_{\rm fem}$ upon $\mathscr{V}_{h,\varkappa}(\Omega)$ or $\mathbf{W}_{h,\varkappa}(\Omega, \mathrm{div})$ with $\varkappa = p$ or $\varkappa = p-1$, the proof of (12.52) is very easy while the complexity of such projection is linear.

The bounds of local effectiveness of the a posteriori error majorants, leading as a rule to two-sided error bounds, are paid much attention, see [2, 3, 11–13, 18, 19]. In part, it is for the reason that they are used in the proofs of the convergence of

adaptive algorithms. Here we give an example of the local effectiveness bound for the a posteriori majorant (12.41) applied to the first order finite element method.

Theorem 12.6 *Under conditions of Theorem 12.4, for* $u_{\rm fem} \in \mathcal{V}_{h,1}(\Omega)$ *and the vector-function* $\mathbf{z} \in \mathbf{W}_{h,2}(\Omega, \mathrm{div})$*, defined as* $\mathbf{L}^2$*-projection of the vector-function* $\mathbf{z}_{\rm fem} = -\nabla u_{\rm fem}$ *on the space* $\mathbf{W}_{h,2}(\Omega, \mathrm{div})$*, we have*

$$\mathscr{M}^{(k)}_{\rm fem}(\sigma, f, \mathbf{z}) \le \mathbb{C}\Big[|\!|\!| e_{\rm fem} |\!|\!|^2 + \sum_{r=1}^{\mathscr{R}} \frac{h_r^2}{\pi^2} \int_{\tau_r} (f - \Pi_r^1 f)^2 dx \Big] \tag{12.57}$$

with the constant $\mathbb{C} = \mathbb{C}(\Omega, c_\triangle)$ *and* $k = 1, 2$.

Proof We refer for the proof to Korneev [25]. □

Note that in comparison with (12.54), the inequality (12.57) is weaker in a sense that they assume $\mathbf{z} \in \mathbf{W}_{h,k}(\Omega, \mathrm{div})$ for $k = 1, 2$, respectively. But this is a consequence of the way of the proof, based on indirect use of the equilibrated fluxes.

References

1. Ainsworth, M., Oden, J.T.: A Posteriori Estimation in Finite Element Analysis. Wiley, New York (2000)
2. Ainsworth, M., Vejchodský, T.: Fully computable robust a posteriori error bounds for singularly perturbed reaction-diffusion problems. Numer. Math. **119**(2), 219–243 (2011)
3. Ainsworth, M., Vejchodský, T.: Robust error bounds for finite element approximation of reaction-diffusion problems with non-constant reaction coefficient in arbitrary space dimension. Comput. Methods Appl. Mech. Eng. **281**, 184–199 (2014)
4. Anufriev, I.E., Korneev, V.G., Kostylev, V.S.: Exactly equilibrated fields, can they be efficiently used for a posteriori error estimation? Uchenye zapiski Kazanskogo universiteta, seria Fiziko-matematicheskie nauki (Scientific notes of Kazan State University, series Physical-Mathematical Sciences) **148**(4), 94–143 (2006)
5. Aubin, J.-P.: Approximation of Elliptic Boundary-Value Problems. Wiley-Interscience, New York (1972)
6. Babuska, I., Strouboulis, T.: Finite Element Method and Its Reliability. Oxford University Press, New York (2001)
7. Babuska, I., Witeman, J.R., Strouboulis, T.: Finite elements. An introduction to the method and error estimation. University Press, Oxford (2011)
8. Bank, R.E., Xu, J., Zheng, B.: Superconvergent Derivative Recovery for Lagrange Triangular Elements of Degree p on unstructured grids. SIAM J. Numer. Anal. **45**(5), 2032–2046 (2007)
9. Bartels, S., Nochetto, R.H., Salgado, A.J.: A total variation diminishing interpolation operator and applications. Math. Comput. **84**, 2569–2587 (2015). https://doi.org/10.1090/mcom/2942
10. Bramble, J.H., Xu, J.: Some estimates for a weighted L_2 projection. Math. Comput. **56**(194), 463–476 (1991)
11. Cai, Z., Zhang, S.: Flux recovery and a posteriori error estimators: conforming elements for scalar elliptic equations. SIAM J. Numer. Anal. **48**(2), 578–602 (2010)
12. Carey, V., Carey, G.F.: Flexible patch post-processing recovery strategies for solution enhancement and adaptive mesh refinement. Int. J. Numer. Methods Eng. **87**(1–5), 424–436 (2011)

13. Carstensen, C.E., Merdon, C.: Effective postprocessing for equilibration a posteriori error estimators. Numer. Math. **123**(3), 425–459 (2013)
14. Cheddadi, I., Fučík, R., Prieto, M.I., Vohralik, M.: Guaranteed and robust a posteriori error estimates for singularly perturbed reaction-diffusion problems. ESAIM: Math. Model. Numer. Anal. **43**, 867–888 (2009)
15. Churilova, M.A.: Vychislitel'nye svoistva funktsional'nyh aposteriornyh otsenok dlia stat-sionarnoi zadachi reaktsii-diffuzii (Numerical properties of functional a posteriori bounds for stationary reaction-diffusion problem). Vestnik SPbSU, Seria 1: Matematika, Mehanika, Astronomija **1**(1), 68–78 (2014, in Russian)
16. Ciarlet, P.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
17. Cottrell, J.A., Hughes, J.R., Bazilevs, Y.: Isogeometric Analysis. Toward Integration of CAD and FEA. Wiley, Chichester (2009)
18. Creusé, E., Nicaise, S.: A posteriori error estimator based on gradient recovery by averaging for discontinuous Galerkin methods. J. Comput. Appl. Math. **234**, 2903–2915 (2010)
19. Ern, A., Stephansen, A., Vohralik, M.: Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems. J. Comput. Appl. Math. **234**(1), 114–130 (2009)
20. Karakashian, O.A., Pascal, F.: A posteriori error estimates for a discontinuous Galerkin approximation of second-order problems. SIAM J. Numer. Anal. **41**, 2374–2399 (2003)
21. Korneev, V.: Shemy metoda konechnyh elementov vysokih poriadkov tochnosti (The Finite Element Methods of High Order of Accuracy). Leningrad State University, Leningrad (1977, in Russian).
22. Korneev, V.G.: Simple algorithms for calculation of a posteriori error estimates for approximate solutions of elliptic equations. Uchenye zapiski Kazanskogo universiteta. Seria Fiziko-matematicheskie nauki **154**(4), 11–27 (2011, in Russian)
23. Korneev, V.G.: Robust consistent a posteriori error majorants for approximate solutions of diffusion-reaction equations. Materialy 11-oi mezhdunarodnoi konferentsii "Setochnye metody dlia kraievyh zadach i prilozhenia", Kazan State University, pp. 182–187 (2016, in Russian)
24. Korneev, V.G.: On the accuracy of a posteriori functional error majorants for approximate solutions of elliptic equations. Dokl. Math. **96**(1), 380–383 (2017)
25. Korneev, V.G.: On the error control at numerical solution of reaction-diffusion equations, 6 Nov 2017. arXiv:1711.02054v1
26. Korneev, V.G., Langer U.: Dirichlet-Dirichlet Domain Decomposition Methods for Elliptic Problems, h and hp Finite Element Discretizations. World Scientific, New Jersey (2015)
27. Ladyzhenskaya, O.A.: The Boundary Value Problems of Mathematical Physics. Springer, New York (1985)
28. Langer, U., Moore, S.E., Neumuller, M.: Space-time isogeometric analysis of parabolic evolution problems. Comput. Methods Appl. Mech. Eng. **306**, 342–363 (2016)
29. Nazarov, F.B., Poborchi, S.V.: Neravenstvo Puankare i ego prilozhenia (Poincare inequality and its applications). Publishing House of St. Petersburg State University, St. Petersburg (2012, in Russian)
30. Nochetto, R.H., Otarola, E., Salgado, A.J.: Piecewise polynomial interpolation in Mucken-houpt weighted Sobolev spaces and applications. Numer. Math. **132**, 85–130 (2016). https://doi.org/10.1007/s00211-015-0709-6
31. Oganesian, L.A., Ruhovets, L.A.: Variatsionno-raznostnyie metody reshenia ellipticheskih uravnenii (Variational-difference methods for solution of elliptic equations). Publishing House of Armenian Academy of Sciences of Armenian SSR, Yerevan (1979, in Russian)
32. Repin, S., Frolov, M.: Ob aposteriornyh otsenkah tochnosti priblizhennyh reshenii kraievyh zadach (On a posteriori error bounds for approximate solutions of elliptic boundary value problems). Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki. **42**(12), 1774–1787 (2002, in Russian)

33. Repin, S., Sauter, S.: Functional a posteriori estimates for the reaction-diffusion problem. C. R. Math. Acad. Sci. Paris **343**(5), 349–354 (2006)
34. Scott, L., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**, 483–493 (1990)
35. Vejchodsk'y, T.: Guaranteed and locally computable a posteriori error estimate. IMA J. Numer. Anal. **26**, 525–540 (2006). https://doi.org/10.1093/imanum/dri043
36. Xu, J., Zhang, Z.: Analysis of recovery type a posteriori error estimators for mildly structured grids. Math. Comput. **73**(247), 1139–1152 (2003)
37. Xu, J., Zou, J.: Some nonoverlapping domain decomposition methods. SIAM Rev. **40**(4), 857–914 (1998)
38. Zhang, Z.: Ultracovergence of the patch recovery technique. Math. Comput. **65**(216), 1431–1437 (1996)
39. Zienkiewicz, O.C., Zhu, J.Z.: The superconvergence patch recovery (SPR) and adaptive finite element refinement. Comput. Methods Appl. Mech. Eng. **101**, 207–224 (1992)

Chapter 13
Space-Time Finite Element Methods for Parabolic Evolution Problems with Variable Coefficients

Ulrich Langer, Martin Neumüller, and Andreas Schafelner

Abstract We introduce a completely unstructured, conforming space-time finite element method for the numerical solution of parabolic initial-boundary value problems with variable in space and time, possibly discontinuous diffusion coefficients. Discontinuous diffusion coefficients allow the treatment of moving interfaces. We show stability of the method and an a priori error estimate, including the case of local stabilizations which are important for adaptivity. To study the method in practice, we consider several typical model problems in one, two, and three spatial dimensions. The implementation of our space-time finite element method is fully parallelized with MPI. Extensive numerical tests were performed to study the convergence behavior of the stabilized space-time finite element discretization method and the scaling properties of the parallel AMG-preconditioned GMRES solver that we use to solve the huge system of space-time finite element equations.

13.1 Introduction

When we deal with the simulation of physical problems like transient diffusion problems, heat-conduction problems, or electromagnetic eddy current problems, the governing partial differential equations (PDEs) are often of parabolic type. Thus, the development of efficient numerical schemes for solving parabolic equations is of great importance. The standard approach to the numerical solution of parabolic PDEs uses some time-stepping method applied to the large-scale system of ordinary differential equations arising from a semi-discretization in the spatial variables, e.g., by means of the Finite Element Method (FEM); see, e.g., [42]. Another approach

U. Langer (✉) · M. Neumüller
Institute of Computational Mathematics, Johannes Kepler University Linz, Linz, Austria
e-mail: ulanger@numa.uni-linz.ac.at; neumueller@numa.uni-linz.ac.at

A. Schafelner
Doctoral Program "Computational Mathematics", Johannes Kepler University Linz, Linz, Austria
e-mail: andreas.schafelner@dk-compmath.jku.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_13

first discretizes the parabolic problem with respect to time by some time-stepping method, and then perform a discretization of the resulting elliptic problems in the spatial variables. This approach is sometimes also called Rothe's method, see, e.g., [25]. There are many papers on the more recent continuous or discontinuous Galerkin or Galerkin-Petrov (cG, cGP, dG, dGP) methods based on time-slices; see, e.g., [1, 4, 6, 14, 27, 31–34, 37] and the references therein. These methods are closely related to classical time-integration methods that can be solved in a time-stepping procedure. This is a sequential procedure that is not suited for parallel computing. In order overcome this drawback on massively parallel computers, time-parallel solvers, see, e.g. [12, 15, 18, 45], or time-parallel time-integration methods like PARAREAL [28] have been developed. An excellent historical overview of 50 years of time-parallel integration methods is given in [16]. An alternative approach consists in a full space-time discretization at once by treating time just as another space variable, i.e., we solve a problem with one dimension more. In fact, this approach to the numerical solution of transient problems in space and time simultaneously is not new, but becomes now a really hot topic in connection with the availability of massively parallel computers with many thousands of cores. Besides the overview paper [16] that focuses on time-parallel integration methods, we refer to [41] that provides an overview of the latest developments in this field with focus on completely unstructured space-time methods and simultaneous space-time adaptivity that treats time just as another variable.

In this paper, we will focus on space-time finite element methods that use really unstructured simplicial space-time meshes. The motivation behind this is that, for elliptic problems, there exist plenty of efficient and, most important, parallel solving methods. If we would be able to derive a stable discrete bilinear form, for which we can prove coercivity (ellipticity) with respect to some mesh-dependent norm in the space-time FE-space, then we can expect that we can efficiently solve the space-time problem fully in parallel as in the elliptic case. In this way, we can overcome the curse of sequentiality of the time-stepping methods. Another reason for the space-time approach is the fact that we are not restricted to a special structure of the mesh. This means that we can apply adaptive mesh refinement in space and time simultaneously. Last but not least, we can easily deal with moving interfaces and computational domains, where the coefficients of the PDE and/or the spatial domain $\Omega(t)$ depend on the time as well. Moreover, optimization problems constrained by a parabolic initial-boundary value problem lead to optimality conditions that can very efficiently be solved by space-time methods.

As already mentioned above, these advantages of space-time methods together with the common availability of massively parallel computers have led to a revival of space-time methods. This especially concerns space-time methods that are based on completely unstructured space-time meshes produced, e.g., by simultaneous space-time adaptivity; see [41] for a review of recent publications on this topic. For instance, Steinbach introduced a inf-sup-stable Petrov-Galerkin method [39], whereas Toulopoulos used bubble functions to stabilize a space-time finite element method [43]. In the context of using Isogeometric Analysis (IgA) as space-time discretization method, Langer et al. proposed an upwind-stabilized space-time

method for parabolic evolution equations [26]; see also PhD thesis [29] by Moore. Similar stabilized space-time finite element schemes have recently been developed in [5, 10, 30]. In [5], beside the upwind-stabilization scheme, Bank, Vassilevski and Zikatanov proposed and analyzed new EAFE (edge average finite element) schemes for parabolic convection-diffusion-reaction problems, whereas Devaud and Schwab [10] introduced upwind-stabilized schemes with mesh grading in time dealing with time singularities and hp schemes leading to exponential convergence.

The main aim of this paper is to generalize the results for the stabilized space-time scheme proposed in [26], where the authors use IgA for the discretization, and the corresponding stabilized space-time FE scheme considered by Moore [30] to the case of moving interfaces, i.e., t-dependent, discontinuous diffusion coefficients, and the possibility to choose local (element-wise) upwind test functions of the form $v_h + \theta_K h_K \partial_t v_h$ depending on the mesh-size h_K of an element K from the finite element mesh. This localization of the upwind-stabilization is very important for adaptivity that produces a family of shape regular meshes. We will use an unstructured conforming FEM to discretize the parabolic initial-boundary value problem, which we specify in the following. Let $Q := \Omega \times (0, T)$ be the space-time cylinder, with $\Omega \subset \mathbf{R}^d$, $d \in \{1, 2, 3\}$, being a sufficiently smooth and bounded spatial domain, and $T > 0$ being the final time. Furthermore, let $\Sigma := \partial\Omega \times (0, T)$, $\overline{\Sigma}_0 := \overline{\Omega} \times \{0\}$ and $\overline{\Sigma}_T := \overline{\Omega} \times \{T\}$ such that $\partial Q = \Sigma \cup \overline{\Sigma}_0 \cup \overline{\Sigma}_T$. Then we consider the following model problem that can formally be written as follows: Given f, g, ν and u_0, find u such that (s.t.)

$$\frac{\partial u}{\partial t}(x,t) - \operatorname{div}_x(\nu(x,t)\nabla_x u(x,t)) = f(x,t), \quad (x,t) \in Q, \tag{13.1}$$

$$u(x,t) = g(x,t) = 0, \quad (x,t) \in \Sigma, \tag{13.2}$$

$$u(x,0) = u_0(x), \quad x \in \overline{\Omega}, \tag{13.3}$$

where the diffusion coefficient (reluctivity in electromagnetics) ν is a given uniformly positive and bounded coefficient function. The dependence of ν not only on space but also on time enables us to model moving interfaces. Note that we do not require ν to be smooth. In fact, we will admit discontinuities for ν. For simplicity, we assume homogeneous Dirichlet boundary conditions.

The paper is structured in the following way: In Sect. 13.2, we provide a space-time variational formulation of the parabolic initial boundary value problem (13.1)–(13.3), and we recall some existence, uniqueness and regularity results for weak solutions in appropriate space-time Sobolev spaces. Section 13.3 is devoted to the derivation and analysis of a new locally stabilized space-time finite element scheme. Moreover, we derive a priori discretization error estimates. In Sect. 13.4, we present four typical test cases for which we have performed extensive numerical studies, and we discuss the numerical results. Section 13.5 draws some conclusions, and provides an outlook on the future work.

13.2 The Space-Time Variational Formulation

In order to derive well-posed space-time variational formulations in space-time Sobolev spaces, we follow the classical approach developed in the monograph [24] by Ladyžhenskaya, Solonnikov and Uraltseva, and in the lecture notes [23] by Ladyžhenskaya. Let us first define the proper function spaces.

Definition 13.1 Let $L_2(Q)$ be the space of square integrable functions in the space-time cylinder Q. Then we define the following Sobolev (Hilbert) spaces

$$H_0^1(Q) = W_{2,0}^1(Q) := \{u \in L_2(Q) : \nabla u \in [L_2(Q)]^{d+1} \text{ and } u = 0 \text{ on } \Sigma\},$$
$$H^{1,0}(Q) = W_2^{1,0}(Q) := \{u \in L_2(Q) : \nabla_x u \in [L_2(Q)]^d\},$$
$$H_0^{1,0}(Q) = W_{2,0}^{1,0}(Q) := \{u \in H^{1,0}(Q) : u = 0 \text{ on } \Sigma\},$$

equipped with the usual scalar products and norms, as well as the Banach space

$$V_2(Q) := \{u \in H^{1,0}(Q) : |u|_Q < \infty\},$$

with the subspaces

$$V_{2,0}(Q) := \{u \in H_0^{1,0}(Q) : |u|_Q < \infty\},$$
$$V_2^{1,0}(Q) := \{u \in V_2(Q) : \lim_{\Delta t \to 0} \|u(\cdot, t + \Delta t) - u(\cdot, t)\|_{L_2(\Omega)} = 0, \text{ uniformly on } [0, T]\},$$
$$V_{2,0}^{1,0}(Q) := V_2^{1,0}(Q) \cap H_0^{1,0}(Q),$$

where the norm $|\cdot|_Q$ is defined by

$$|u|_{Q_t} := \max_{0 \le \tau \le t} \|u(\cdot, \tau)\|_{L_2(\Omega)} + \|\nabla_x u\|_{Q_t},$$

and $Q_t = \Omega \times (0, t)$ denotes a truncated space-time cylinder. Here, the appearing differential operators are defined as follows:

$$\nabla = (\nabla_x, \nabla_t)^T, \quad \nabla_x = (\partial_{x_1}, \dots, \partial_{x_d})^T \quad \text{and} \quad \nabla_t = (\partial_t).$$

Multiplying the PDE (13.1) by a test function $v \in \hat{H}_0^1(Q) := \{v \in H_0^1(Q) : v = 0 \text{ on } \Sigma_T\}$, integrating over the complete space-time domain (cylinder) $Q = \Omega \times (0, T)$, integrating by parts with respect to time and space once, and incorporating the initial and boundary conditions, we immediately arrive at the following space-time variational formulation of the initial-boundary value problem (13.1)–(13.3): find a function $u \in H_0^{1,0}(Q)$ such that

$$a(u, v) = l(v), \quad \forall v \in \hat{H}_0^1(Q), \tag{13.4}$$

where the bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$ are defined by the identities

$$a(u,v) = \int_Q (-u\partial_t v + \nu(x,t)\nabla_x u \nabla_x v)\, \mathrm{d}x\mathrm{d}t$$

and

$$l(v) = \int_\Omega u_0(x)\, v(x,0)\, \mathrm{d}x + \int_Q f v\, \mathrm{d}x\mathrm{d}t,$$

respectively. A solution u of the space-time variational (13.4) is called generalized (weak) solution of the parabolic initial-boundary value problem (13.1)–(13.3) in the space $u \in H_0^{1,0}(Q)$.

Under the assumptions that

$$u_0 \in L_2(\Omega) \quad \text{and} \quad f \in L_{2,1}(Q) := \{v : Q \to \mathbf{R} : \int_0^T \|v(\cdot,t)\|_{L_2(\Omega)}\, \mathrm{d}t < \infty\}, \tag{13.5}$$

and that

$$0 < \underline{\nu} \le \nu(x,t) \le \overline{\nu}, \quad \text{for almost all } (x,t) \in Q, \tag{13.6}$$

with positive constants $\underline{\nu}$ and $\overline{\nu}$, the following theorem was proven by means of Galerkin's method and appropriate a priori estimates in [23]:

Theorem 13.1 ([23, Chapter III, Thm. 3.1]) *Under the conditions* (13.5) *and* (13.6)*, the space-time variational problem* (13.4) *has at least one generalized (weak) solution in* $H_0^{1,0}(Q)$.

Definition 13.2 ([23, Chapter III]) A generalized solution $u \in H_0^{1,0}(Q)$ of the space-time variational problem (13.4) is a called a generalized solution in $V_{2,0}^{1,0}(Q)$, if $u \in V_{2,0}^{1,0}(Q)$ and if it fulfills the energy-balance equation

$$\frac{1}{2}\|u(\cdot,t)\|^2_{L_2(\Omega)} + \int_{Q_t} \nu(x,\tau)|\nabla_x u|^2\, \mathrm{d}x\mathrm{d}\tau = \frac{1}{2}\|u(\cdot,0)\|^2_{L_2(\Omega)} + \int_{Q_t} f\, u\, \mathrm{d}x\mathrm{d}\tau.$$

and the identity

$$\int_\Omega u(x,t)\ \ v(x,t)\, \mathrm{d}x - \int_\Omega u_0 v(x,0)\, \mathrm{d}x + \int_{Q_t} -u\partial_t v + \nu\nabla_x u \nabla_x v\, \mathrm{d}x\mathrm{d}\tau = \int_{Q_t} f\, v\, \mathrm{d}x\mathrm{d}\tau,$$

for all $v \in H_0^1(Q)$ and any $t \in (0,T)$.

Theorem 13.2 ([23, Chapter III, Thm. 3.2]) *If the assumptions* (13.5) *and* (13.6) *are fulfilled, then any generalized solution of the space-time variational problem* (13.4) *in* $H_0^{1,0}(Q)$ *is the generalized solution in* $V_{2,0}^{1,0}(Q)$ *and it is unique in* $H_0^{1,0}(Q)$.

Corollary 13.1 *If the assumptions* (13.5) *and* (13.6) *hold, then there exists a unique generalized solution* $u \in V_{2,0}^{1,0}(Q)$ *to the the space-time variational problem* (13.4).

Remark 13.1 For the case $\nu = 1$, $f \in L_2(Q)$ and $u_0 \in H_0^1(\Omega)$, Ladyžhenskaya proved in [23, Chapter III, Thm. 2.1] that the generalized solution u of (13.4) belongs to space $H_0^{\Delta,1}(Q) = W_{2,0}^{\Delta,1}(Q) = \{v \in H_0^1(Q) : \Delta_x v \in L_2(Q)\}$, and u continuously depends on t in the norm of the space $H_0^1(\Omega)$. If $\partial\Omega \in C^2$, then $u \in W_{2,0}^{2,1}(Q)$.

More regularity results can already be found in the classical monograph [24] and in the more recent references [46] and [22]. The last reference provides an overview on maximal parabolic regularity results; see also [27]. The space-time finite element scheme that we are going to derive is consistent for solutions u of (13.4) that have at least piecewise partial time derivative $\partial_t u$ in L_2 and fluxes $\nu\nabla_x u$ in $H(\mathrm{div}_x) = \{v = (v_1, \ldots, v_d) \in [L_2(Q)]^d : \mathrm{div}_x v \in L_2(Q)\}$. This is ensured in the case of maximal parabolic regularity where $\partial_t u \in L_2(Q)$ and $\mathrm{div}_x(\nu\nabla_x u) \in L_2(Q)$, i.e. $\partial_t u - \mathrm{div}_x(\nu\nabla_x u) = f$ in $L_2(Q)$. We emphasize that we need this property only element-wise for deriving a consistent scheme.

13.3 The Space-Time Finite Element Scheme

From the previous section, we know that there exists a unique generalized solution of the initial-boundary value problem (13.1) in $H_0^{1,0}(Q) \cap V_{2,0}^{1,0}(Q)$ that may have more regularity due to more regularity of the data, see Remark 13.1 and the references mentioned above. The goal of this section is to derive a consistent and stable space-time finite element scheme with a discrete (mesh-dependent) bilinear form $a_h(\cdot,\cdot)$ that is coercive (elliptic) on the space-time finite element spaces and bounded on extended spaces with respect to appropriately chosen, mesh-dependent norms. These properties ensure existence and uniqueness of a finite element solution, and, together with appropriate interpolation respectively approximation error estimates, a priori discretization error estimates for sufficiently smooth solutions.

Similar to Langer et al. in [26], we use special time-upwind test functions, but in contrast to [26] the time-upwind test functions are now locally scaled by the element mesh-size in order to handle adaptivity. First, we need a regular or, at least, a shape regular triangulation $\mathscr{T}_h$ of the space-time cylinder Q; see, e.g., [7, 9] for details. We now formally define this triangulation as $\mathscr{T}_h := \{K : K \subset Q, K \text{ open}\}$ such that $\overline{Q} = \bigcup_{K \in \mathscr{T}_h} \overline{K}$, with $K \cap K' = \emptyset$ for $K \neq K' \in \mathscr{T}_h$, and the usual conditions

imposed on a regular or a shape regular triangulation are fulfilled [7, 9]. On each of these elements K, we now define individual *time-upwind test functions*

$$v_{h,K}(x,t) := v_h(x,t) + \theta_K h_K \partial_t v_h(x,t), \text{ for all } (x,t) \in K,$$

where θ_K is a local positive parameter that will be defined later, and $h_K := diam(K)$. Here, v_h is some test function from a standard conforming space-time finite element space $V_{0h} = \{v \in C(\overline{Q}) : v(x_K(\cdot)) \in \mathbf{P}_p(\hat{K}), \forall K \in \mathscr{T}_h, v = 0 \text{ on } \overline{\Sigma} \cup \overline{\Sigma_0}\}$, where $x_K(\cdot)$ is the map from the reference element $\hat{K}$ to the finite element $K \in \mathscr{T}_h$, and $\mathbf{P}_p(\hat{K})$ is the space of polynomials of the degree p on the reference element $\hat{K}$. For simplicity, throughout this paper and, in particular, in our numerical experiments in Sect. 13.4, we use affine-linear mappings $x_K(\cdot)$ and simplicial elements. From now on, unless specified otherwise, all functions depend on both space and time variables. So, we can omit the arguments.

In this section, we will also use the following spaces:

$$H^{1,1}_{0,\underline{0}}(Q) := \{u \in L_2(Q) : \nabla_x u \in [L_2(Q)]^d,\ \partial_t u \in L_2(Q) \text{ and } u|_{\Sigma \cup \Sigma_0} = 0\},$$

$$H^{2,1}_{0,\underline{0}}(\mathscr{T}_h) := \{v \in H^{1,1}_{0,\underline{0}}(Q) : v|_K \in H^{2,1}(K),\ \forall K \in \mathscr{T}_h\},$$

$$W^1_\infty(\mathscr{T}_h) := \{v \in L_\infty(Q) : v|_K \in W^1_\infty(K),\ \forall K \in \mathscr{T}_h\},$$

where $H^{2,1}(K) := \{v \in L_2(K) : \partial_t v, \partial_{x_i} v, \partial_{x_i}\partial_{x_j} v \in L_2(K) \text{ and } \partial_t v \in L_2(K)\}$. For the sake of convenience, we now consider homogeneous initial conditions, i.e., $u_0 = 0$ on Ω. Furthermore, we assume that $\nu \in W^1_\infty(\mathscr{T}_h)$, and that the PDE has a sufficiently smooth solution u, e.g., $u \in H^{2,1}_{0,\underline{0}}(\mathscr{T}_h)$; cf. also our discussion in Sect. 13.2. Now we first multiply the PDE (13.1) by the space-time test function $v_{h,K}$, and then integrate over a single element K. Summing up over all elements and applying integration by parts in the principle term, we obtain

$$\sum_{K \in \mathscr{T}_h} \int_K (\partial_t u v_h + \theta_K h_K \partial_t u \partial_t v_h + \nu \nabla_x u \cdot \nabla_x v_h + \theta_K h_K \nu \nabla_x u \cdot \nabla_x(\partial_t v_h))\ \mathrm{d}(x,t)$$

$$- \sum_{K \in \mathscr{T}_h} \int_{\partial K} (\nu \nabla_x u \cdot n_x v_h + \theta_K h_K \nu \nabla_x u \cdot n_x \partial_t v_h)\ \mathrm{d}s_{(x,t)} = l_h(v_h) \qquad (13.7)$$

with the linear form

$$l_h(v_h) := \sum_{K \in \mathscr{T}_h} \int_K f\ (v_h + \theta_K h_K \partial_t v_h)\ \mathrm{d}(x,t). \qquad (13.8)$$

For the exact solution u of (13.1), we know that the fluxes have to be continuous across the boundaries of the elements $K \in \mathscr{T}_h$. This observation means that

$$\sum_{K \in \mathscr{T}_h} \int_{\partial K} \nu \nabla_x u \cdot n_x v_h\ \mathrm{d}s_{(x,t)} = 0$$

for all test functions $v_h \in V_{0h}$. We mention that v_h is zero on Σ, and that n_x vanishes on $\overline{\Sigma} \cup \overline{\Sigma}_0$. Therefore, the first boundary term completely disappears from (13.7), but, in general, not the second term, since $\theta_K h_K$ varies from element to element. We now arrived at the consistency identity

$$a_h(u, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h}, \tag{13.9}$$

that holds for a sufficiently smooth solution u, e.g., $u \in H^{2,1}_{0,\underline{0}}(\mathscr{T}_h)$, where the discrete (mesh-dependent) bilinear form $a_h(\cdot,\cdot)$ is defined by the identity

$$\begin{aligned} a_h(u, v_h) := & \sum_{K \in \mathscr{T}_h} \int_K (\partial_t u \, v_h + \theta_K h_K \, \partial_t u \, \partial_t v_h) \, \mathrm{d}(x,t) \\ & + \sum_{K \in \mathscr{T}_h} \int_K (\nu \nabla_x u \cdot \nabla_x v_h + \theta_K h_K \, \nu \nabla_x u \cdot \nabla_x (\partial_t v_h)) \, \mathrm{d}(x,t) \\ & - \sum_{K \in \mathscr{T}_h} \int_{\partial K} \theta_K h_K \, \nu \nabla_x u \cdot n_x \, \partial_t v_h \, \mathrm{d}s_{(x,t)}, \end{aligned} \tag{13.10}$$

and the linear form $l_h(\cdot)$ is defined by (13.8), with given $\nu \in W^1_\infty(\mathscr{T}_h)$ and $f \in L_2(Q)$.

Remark 13.2 We can derive a scheme that is equivalent to (13.10). In particular, instead of applying integration by parts on both principal terms, we only apply it to the first principal term and keep the second. Hence, we obtain another consistency identity

$$\tilde{a}_h(u, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h},$$

that holds for a solution u of (13.4) that only belongs to $H^{L,1}_{0,\underline{0}}(\mathscr{T}_h) := \{u \in L_2(Q) : (\partial_t u)|_K \in L_2(K), (\nu \nabla_x u)|_K \in H(\mathrm{div}_x, K), \; u = 0 \text{ on } K \cap (\Sigma \cup \Sigma_0) \; \forall K \in \mathscr{T}_h\}$, where

$$\begin{aligned} \tilde{a}_h(u, v_h) := & \sum_{K \in \mathscr{T}_h} \int_K (\partial_t u \, v_h + \theta_K h_K \, \partial_t u \, \partial_t v_h) \, \mathrm{d}(x,t) \\ & + \sum_{K \in \mathscr{T}_h} \int_K (\nu \nabla_x u \cdot \nabla_x v_h - \theta_K h_K \, \mathrm{div}_x (\nu \nabla_x u) \partial_t v_h) \, \mathrm{d}(x,t) \end{aligned} \tag{13.11}$$

with given $\nu \in W^1_\infty(\mathscr{T}_h)$ and $f \in L_2(Q)$, and l_h as in (13.8). We mention that $u \in H^{L,1}_{0,\underline{0}}(\mathscr{T}_h)$ is ensured in the case of maximal parabolic regularity where u belongs $H^{L,1}(Q) := \{v \in H^1(Q) : Lu := \mathrm{div}_x(\nu \nabla_x u) \in L_2(Q)\}$.

Remark 13.3 If the test functions $v_h \in V_{0h}$ are continuous and piecewise linear ($p = 1$), then the term in (13.10) containing $\nabla_x(\partial_t v_h)$ vanishes in all elements $K \in \mathscr{T}_h$, since it only contains mixed second order derivatives of the test functions.

Now we look for a Galerkin approximation $u_h \in V_{0h}$ to the generalized solution u of the initial boundary value problem (13.1)–(13.3) using the variational identity (13.9), i.e., find $u_h \in V_{0h}$ such that

$$a_h(u_h, v_h) = l_h(v_h), \quad \forall v_h \in V_{0h}, \tag{13.12}$$

with $a_h(\cdot,\cdot)$ and $l_h(\cdot)$ as defined above by (13.10) and (13.8), respectively. In Sect. 13.2, we already showed existence and uniqueness of a weak solution to the initial-boundary value problem (13.1)–(13.3). However, our finite element scheme (13.12) is based on a mesh-dependent bilinear form $a_h(\cdot,\cdot)$. Thus, we have to investigate the stability of the space-time finite element scheme. More precisely, we will show ellipticity of the bilinear form $a_h(\cdot,\cdot) : V_{0h} \times V_{0h} \to \mathbf{R}$ w.r.t. the mesh-dependent norm

$$\|v_h\|_h^2 := \sum_{K\in\mathscr{T}_h} \left[\|\nu^{1/2}\nabla_x v_h\|_{L_2(K)}^2 + \theta_K h_K \|\partial_t v_h\|_{L_2(K)}^2\right] + \frac{1}{2}\|v_h\|_{L_2(\Sigma_T)}^2. \tag{13.13}$$

This implies existence and uniqueness of the finite element solution $u_h \in V_{0h}$ of (13.12). For the following derivations, we assume that our triangulation $\mathscr{T}_h$ of Q is *shape regular* such that local approximation error estimates are available [7, 9]. A shape regular triangulation $\mathscr{T}_h$ of Q is called quasi-uniform, if there exists a constant c_u such that

$$h_K \le h \le c_u h_K, \quad \text{for all } K \in \mathscr{T}_h,$$

where $h = \max_{K\in\mathscr{T}_h} h_K$. Moreover, we introduce localized bounds for our coefficient function ν, i.e.,

$$\underline{\nu}_K \le \nu(x,t) \le \overline{\nu}_K, \quad \text{for almost all } (x,t) \in K \text{ and for all } K \in \mathscr{T}_h, \tag{13.14}$$

where $\underline{\nu}_K \ge \underline{\nu}$ and $\overline{\nu}_K \le \overline{\nu}$ are positive constants on every $K \in \mathscr{T}_h$. In the following, we need some inverse inequalities for functions from finite element spaces.

Lemma 13.1 *There exist generic positive constants $c_{I,1}$ and $c_{I,2}$ such that*

$$\|v_h\|_{L_2(\partial K)} \le c_{I,1} h_K^{-1/2} \|v_h\|_{L_2(K)}, \tag{13.15}$$

$$\|\nabla v_h\|_{L_2(K)} \le c_{I,2} h_K^{-1} \|v_h\|_{L_2(K)} \tag{13.16}$$

for all $v_h \in V_{0h}$ and for all $K \in \mathscr{T}_h$.

Proof For (13.15); see e.g. [11, 35], and for (13.16) see e.g. [7, 9, 11]. □

From $\nabla = (\nabla_x, \partial_t)^T$ and (13.16), we can immediately deduce

$$\|\partial_t v_h\|_{L_2(K)} \leq c_{I,2} h_K^{-1} \|v_h\|_{L_2(K)}. \tag{13.17}$$

The above inequalities hold for the standard L_2-norms. However, we will also need such a result in some scaled norm.

Lemma 13.2 *Let $\nu \in W_\infty^1(\mathscr{T}_h)$ be a given, uniformly positive function. Then*

$$\|v\|_{L_2^\nu(K)}^2 = \int_K \nu(x,t)\, |v(x,t)|^2 \, \mathrm{d}(x,t)$$

is a norm, and the inverse estimate

$$\|\partial_t v_h\|_{L_2^\nu(K)} \leq \|\nabla v_h\|_{L_2^\nu(K)} \leq c_{I,\nu} h_K^{-1} \|v_h\|_{L_2^\nu(K)}$$

holds for all $v_h \in V_{0h}$ and for all $K \in \mathscr{T}_h$, with $c_{I,\nu} := (\overline{\nu}_K / \underline{\nu}_K)^{1/2} c_{I,2}$.

Proof See [36]. □

We note that, in practical applications, it is clear that $1 \leq \overline{\nu}_K / \underline{\nu}_K$ is close to 1. Below, we will also need the estimate

$$\|\partial_t \partial_{x_i} v_h\|_{L_2^\nu(K)} \leq c_{I,\nu} h_K^{-1} \|\partial_{x_i} v_h\|_{L_2^\nu(K)}, \tag{13.18}$$

which obviously holds for all $v_h \in V_{0h}$ and for all $K \in \mathscr{T}_h$. Moreover, we need the following inverse inequality.

Lemma 13.3 *Let $\nu \in W_\infty^1(\mathscr{T}_h)$ be a given uniformly positive function. Let $W_h|_K := \{w_h : w_h = \nabla_x v_h, v_h \in V_{0h}|_K\}$. Then the inverse estimate*

$$\|\mathrm{div}_x(\nu w_h)\|_{L_2(K)} \leq c_{I,3} h_K^{-1} \|\nu w_h\|_{L_2(K)}, \forall w_h \in W_h|_K \tag{13.19}$$

holds, where $c_{I,3}$ is a positive constant that is independent of h_K.

Proof First, we know that $V_{0h}|_K$ is a finite-dimensional space spanned by the local shape functions $\{p^{(i)}\}_{i \in \overline{\omega}_K}$, where $\overline{\omega}_K$ is the index set of local degrees of freedom. Hence, the space $W_h|_K$ is also finite-dimensional and spanned by the generating system $\{\nabla_x p^{(i)}\}_{i \in \overline{\omega}_K}$. Moreover, for a fixed ν, each product $z_h := \nu\, w_h$ can be represented by means of a non-necessary unique linear combination $\{\nu\, \nabla_x p^{(i)}\}_{i \in \overline{\omega}_K}$ on K. We denote this space by $Z_h(K) := \mathrm{span}_{i \in \overline{\omega}_K}\{\nu\, \nabla_x p^{(i)}\}$. Using Cauchy's inequality, we obtain

$$\begin{aligned}
\|\mathrm{div}_x z_h\|_{L_2(K)}^2 &= \int_K |\mathrm{div}_x z_h|^2 \, \mathrm{d}(x,t) = \int_K |\sum_{i=1}^d \partial_{x_i} z_{h,i}|^2 \, \mathrm{d}(x,t) \\
&\leq d \int_K \sum_{i=1}^d |\partial_{x_i} z_{h,i}|^2 \, \mathrm{d}(x,t) = d \sum_{i=1}^d \|\partial_{x_i} z_{h,i}\|_{L_2(K)}^2,
\end{aligned}$$

for all $z_h \in Z_h(K)$. Now, by a simple scaling argument, we can estimate each element in the sum, and obtain

$$d\sum_{i=1}^{d}\|\partial_{x_i} z_{h,i}\|^2_{L_2(K)} \le d\sum_{i=1}^{d} C^2 h_K^{-2}\|z_{h,i}\|^2_{L_2(K)} = dC^2 h_K^{-2}\|z_h\|^2_{L_2(K)},$$

where C is a positive constant that is independent of h_K. Taking the square root and setting $c_{I,3} := C\sqrt{d}$ closes the proof. □

Lemma 13.3 gives information how the two norms in (13.19) scale w.r.t. the mesh-size h_K. However, the estimate (13.19) is not sharp w.r.t. the constant.

Lemma 13.4 *Let the assumptions of Lemma 13.3 hold. Then*

$$\|\mathrm{div}_x(vw_h)\|_{L_2(K)} \le c_{opt}\|vw_h\|_{L_2(K)}, \forall w_h \in W_h|_K,$$

with $c^2_{opt} = \sup_{0\neq z_h\in Z_h(K)} \frac{\|\mathrm{div}_x(z_h)\|^2_{L_2(K)}}{\|z_h\|^2_{L_2(K)}} \le C^2 d h_K^{-1}$.

Proof See [36]. □

Remark 13.4 We note that the constant c_{opt} in Lemma 13.4 is not only optimal, but also computable. If $z_h \in Z_h(K)$, then, by definition, we have the representation

$$z_h(x,t) = \sum_{j\in\tilde{\omega}_K} \tilde{z}_j \tilde{q}^{(j)}, \tag{13.20}$$

where $\{\tilde{q}^{(j)}\}_{j\in\tilde{\omega}_K}$ forms some basis of $Z_h(K)$. Once some basis is chosen, we can rewrite

$$\|z_h\|^2_{L_2(K)} = (z_h, z_h)_{L_2(K)} \quad \text{and} \quad \|\mathrm{div}_x z_h\|^2_{L_2(K)} = \underbrace{(\mathrm{div}_x z_h, \mathrm{div}_x z_h)_{L_2(K)}}_{=:b(z_h,z_h)}$$

in the form

$$(y_h, z_h)_{L_2(K)} = (M_h \underline{y}, \underline{z}) \quad \text{and} \quad b(y_h, z_h) = (B_h \underline{y}, \underline{z}),$$

with the element mass matrix $M_h = (M_{ij} = (\tilde{q}^{(j)}, \tilde{q}^{(i)})_{L_2(K)})_{i,j\in\tilde{\omega}_K}$ and the element div_x-stiffness matrix $B_h = (B_{ij} = b(\tilde{q}^{(j)}, \tilde{q}^{(i)})_{L_2(K)})_{i,j\in\tilde{\omega}_K}$, respectively. Here, the vectors $\underline{y}$ and $\underline{z}$ are the vector of coefficients in the representation (13.20) w.r.t. the chosen basis $\{\tilde{q}^{(j)}\}_{j\in\tilde{\omega}_K}$. Using this matrix representation, we immediately get

$$c^2_{opt} = \sup_{0\neq z_h\in Z_h(K)} \frac{\|\mathrm{div}_x(z_h)\|^2_{L_2(K)}}{\|z_h\|^2_{L_2QT(K)}} = \sup_{\underline{z}\in\mathbf{R}^{N_K=|\tilde{\omega}_K|}} \frac{(B_h\underline{z},\underline{z})_{\ell_2}}{(M_h\underline{z},\underline{z})_{\ell_2}}.$$

Hence, c_{opt}^2 is the *largest eigenvalue* of the generalized eigenvalue problem

$$B_h \underline{z} = \lambda M_h \underline{z},$$

that can easily be computed.

Now, we are in the position to proof the following coercivity lemma that is crucial for our approach.

Lemma 13.5 *There exits a positive constant μ_c such that*

$$a_h(v_h, v_h) \geq \mu_c \|v_h\|_h^2, \quad \forall v_h \in V_{0h},$$

with $\mu_c = \min_{K \in \mathscr{T}_h} \left\{1 - c_{I,3}\sqrt{\frac{\overline{\nu}_K \theta_K}{4h_K}}\right\} \geq \frac{1}{2}$ *provided that* $\theta_K \leq \frac{h_K}{c_{I,3}^2 \overline{\nu}_K}$. *For instance,* $\theta_K = \frac{h_K}{c_{I,3}^2 \overline{\nu}_K}$ *yields* $\mu_c = \frac{1}{2}$.

Proof Integration by parts in the last term of (13.10) yields

$$a_h(v_h, v_h) = \sum_{K \in \mathscr{T}_h} \left[\int_K \frac{1}{2} \partial_t (v_h^2) \, \mathrm{d}(x,t) + \theta_K h_K \|\partial_t v_h\|_{L_2(K)}^2 + \int_K \nu |\nabla_x v_h|^2 \, \mathrm{d}(x,t) \right.$$
$$\left. - \int_K \theta_K h_K \, \mathrm{div}_x(\nu \nabla_x v_h) \partial_t v_h \, \mathrm{d}(x,t) \right].$$

Now using Gauss' theorem and the facts that v_h is continuous across the element boundary and that $n_t = 0$ on Σ, we obtain

$$a_h(v_h, v_h) = \frac{1}{2}\left(\|v_h\|_{L_2(\Sigma_T)}^2 - \|v_h\|_{L_2(\Sigma_0)}^2\right) + \sum_{K \in \mathscr{T}_h} \left[\theta_K h_K \|\partial_t v_h\|_{L_2(K)}^2 \right.$$
$$\left. + \int_K \nu |\nabla_x v_h|^2 \, \mathrm{d}(x,t) - \int_K \theta_K h_K \, \mathrm{div}_x(\nu \nabla_x v_h) \partial_t v_h \, \mathrm{d}(x,t) \right].$$

The first, second and third term already appear in the definition of our mesh-dependent norm (13.13). It remains to estimate the last term. Using the Cauchy-Schwarz inequality, Lemma 13.3, and a scaled Young's inequality, we arrive at the estimate

$$|\theta_K h_K \int_K \mathrm{div}_x(\nu \nabla_x v_h) \, \partial_t v_h \, \mathrm{d}(x,t)|$$
$$\leq c_{I,3}\left(\frac{\varepsilon \overline{\nu}_K \theta_K}{2h_K} \|\nabla_x v_h\|_{L_2^\nu(K)}^2 + \frac{1}{2\varepsilon} \theta_K h_K \|\partial_t v_h\|_{L_2(K)}^2\right).$$

This estimate and the fact that $v_h = 0$ on Σ_0 immediately yield the estimate

$$a_h(v_h, v_h) \geq \frac{1}{2}\|v_h\|^2_{L_2(\Sigma_T)} + \sum_{K \in \mathscr{T}_h} \Big[\big(1 - \frac{c_{I,3}}{2\varepsilon}\big)\theta_K h_K \|\partial_t v_h\|^2_{L_2(K)} + \big(1 - \varepsilon \frac{c_{I,3}\overline{\nu}_K \theta_K}{2h_K}\big)\|\nabla_x v_h\|^2_{L_2^\nu(K)}\Big].$$

We now choose $\varepsilon = \sqrt{h_K/(\theta_K \overline{\nu}_K)}$ and obtain

$$\begin{aligned} a_h(v_h, v_h) &\geq \min_{K \in \mathscr{T}_h} \left(1 - c_{I,3}\sqrt{\frac{\theta_K\, \overline{\nu}_K}{4h_K}}\right) \\ &\quad \times \left(\sum_{K \in \mathscr{T}_h} \big[\|\nabla_x v_h\|^2_{L_2^\nu(K)} + \theta_K h_K \|\partial_t v_h\|^2_{L_2(K)}\big] + \frac{1}{2}\|v_h\|^2_{L_2(\Sigma_T)}\right) \\ &\geq \mu_c \|v_h\|_h^2, \end{aligned}$$

which concludes the proof. □

Remark 13.5 The above proof holds for any polynomial degree $p \geq 1$ and any fixed, uniformly positive $\nu \in L_\infty(Q)$. However, for the special case $p = 1$ and $\nu|_K = const$, the proof is trivial since $\partial_t(\nabla_x v_h) \equiv 0$ and $\nu|_K \Delta_x v_h \equiv 0$. Hence, the identity

$$a_h(v_h, v_h) = \sum_{K \in \mathscr{T}_h} \frac{1}{2}\int_{\partial K} v_h^2\, n_t \, \mathrm{d}s_{(x,t)} + \theta_K h_K \|\partial_t v_h\|^2_{L_2(K)} + \|\nu^{1/2}\nabla_x v_h\|^2_{L_2(K)} = \|v_h\|_h^2$$

holds, i.e., $\mu_c = 1$. Thus, for this special case, the choice of θ_K has no influence on the coercivity (ellipticity) of the space-time finite element method.

Lemma 13.5 already ensures uniqueness of the finite element solution $u_h \in V_{0h}$. Furthermore, the space V_{0h} is finite-dimensional. Hence, uniqueness implies existence of finite element solution $u_h \in V_{0h}$ of (13.9). For the special case of uniform meshes and uniform θ, i.e., $h_K = h$ and $\theta_K = \theta$ for all $K \in \mathscr{T}_h$, and $\nu \equiv 1$, a proof for ellipticity with a *mesh-independent* constant was done by Langer, Moore and Neumüller in [26] and by Moore in [29]. For a second special case, where θ_K vanishes, i.e., $\theta_K = \theta = 0$ for all $K \in \mathscr{T}_h$, Steinbach has shown existence and uniqueness of solutions to both the continuous and discrete version of (13.9) on the basis of Banach-Nečas-Babuška's theorem in [39]. In addition, both papers also include a priori discretization error estimates, where Steinbach's estimate is based on a discrete inf-sup condition. To derive an a priori error estimate w.r.t. the mesh dependent norm (13.13), we need to show that our bilinear form $a_h(\cdot, \cdot)$ is uniformly

bounded on $V_{0h,*} \times V_{0h}$, where $V_{0h,*} = H_0^{1,0}(Q) \cap H^2(\mathscr{T}_h) + V_{0h}$ is equipped with the norm

$$\|v\|_{h,*}^2 = \frac{1}{2}\|v\|_{L_2(\Sigma_T)}^2 + \sum_{K \in \mathscr{T}_h} \Big[\theta_K h_K \|\partial_t v\|_{L_2(K)}^2 + \|\nabla_x v\|_{L_2^\nu(K)}^2 + (\theta_K h_K)^{-1}\|v\|_{L_2(K)}^2 + \theta_K h_K |v|_{H^2(K)}^2\Big] \quad (13.21)$$

Moreover, we will make use of the following scaled trace inequality.

Lemma 13.6 *There exists a positive constants $c_{Tr} > 0$ such that*

$$\|v\|_{L_2(\partial K)}^2 \le 2c_{Tr}^2 h_K^{-1}\big(\|v\|_{L_2(K)}^2 + h_K^2\|\nabla v\|_{L_2(K)}^2\big) \quad (13.22)$$

for all $v \in H^1(K)$ and for all $K \in \mathscr{T}_h$.

Proof See, e.g., [35]. □

Lemma 13.7 *The bilinear form $a_h(\cdot,\cdot)$ is uniformly bounded on $V_{0h,*} \times V_{0h}$, i.e.,*

$$|a_h(u, v_h)| \le \mu_b \|u\|_{h,*}\, \|v_h\|_h, \quad \forall\, u \in V_{0h,*},\, v_h \in V_{0h},$$

where $\mu_b = \max_{K\in\mathscr{T}_h}\big\{2(1+\theta_K h_K^{-1} c_{Tr}^2 \overline{\nu}_K^2 \underline{\nu}_K^{-1}), 2c_{Tr}^2\overline{\nu}_K^2, 2+c_{I,1}^2, 1+(c_{I,\nu}\theta_K)^2\big\}^{1/2}$ that is uniformly bounded provided that $\theta_K = \mathscr{O}(h_K)$.

Proof We will estimate the bilinear form (13.10) term by term. Since $V_{0h} \subset H_{0,0}^{1,1}(Q)$, we can apply integration by parts and the Cauchy-Schwarz inequality to the first term, and obtain

$$\left|\sum_{K\in\mathscr{T}_h}\int_K \partial_t u v_h \,\mathrm{d}(x,t)\right| \le \sum_{K\in\mathscr{T}_h}\Big[\big((\theta_K h_K)^{-1}\|u\|_{L_2(K)}^2\big)^{1/2}\big(\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2\big)^{1/2}\Big] + \big(\|u\|_{L_2(\Sigma_T)}^2\big)^{1/2}\big(\|v_h\|_{L_2(\Sigma_T)}^2\big)^{1/2}.$$

For the second and third term, applying the Cauchy-Schwarz inequality to each term of the sum, we immediately get the estimates

$$\left|\theta_K h_K \int_K \partial_t u \partial_t v_h \,\mathrm{d}(x,t)\right| \le \big(\theta_K h_K\|\partial_t u\|_{L_2(K)}^2\big)^{1/2}\big(\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2\big)^{1/2},$$

$$\left|\int_K \nu\nabla_x u \nabla_x v_h \,\mathrm{d}(x,t)\right| \le \big(\|\nabla_x u\|_{L_2^\nu(K)}^2\big)^{1/2}\big(\|\nabla_x v_h\|_{L_2^\nu(K)}^2\big)^{1/2},$$

respectively. For the fourth term, we use again Cauchy-Schwarz' inequality, the inverse estimate (13.18), and obtain

$$\begin{aligned}
&\left|\theta_K h_K \int_K \nu \nabla_x u \nabla_x(\partial_t v_h)\, \mathrm{d}(x,t)\right| \\
&\leq \left(\|\nabla_x u\|^2_{L_2^\nu(K)}\right)^{1/2} \left((\theta_K h_K)^2 \sum_{i=1}^d c_{I,\nu}^2 h_K^{-2} \|\partial_{x_i} v_h\|^2_{L_2^\nu(K)}\right)^{1/2} \\
&= \left(\|\nabla_x u\|^2_{L_2^\nu(K)}\right)^{1/2} \left((c_{I,\nu}\theta_K)^2 \|\nabla_x v_h\|^2_{L_2^\nu(K)}\right)^{1/2}.
\end{aligned}$$

For the last term, we apply Cauchy-Schwarz' inequality, the trace inequalities (13.15) and (13.22), and get

$$\begin{aligned}
&\left|\theta_K h_K \int_{\partial K} \nu \nabla_x u \cdot n_x \partial_t v_h \, \mathrm{d}s_{(x,t)}\right| \\
&\leq \left(2\theta_K \overline{\nu}_K^2 c_{Tr}^2 h_K^{-1}\left[\|\nabla_x u\|^2_{L_2(K)} + h_K^2 \sum_{i=1}^d \|\nabla \partial_{x_i} u\|^2_{L_2(K)}\right]\right)^{1/2} \left(\theta_K h_K c_{I,1}^2 \|\partial_t v_h\|^2_{L_2(K)}\right)^{1/2} \\
&\leq \left(2\theta_K c_{Tr}^2 \frac{\overline{\nu}_K^2}{\underline{\nu}_K} h_K^{-1} \|\nabla_x u\|^2_{L_2^\nu(K)} + 2c_{Tr}^2 \overline{\nu}_K^2 \theta_K h_K |u|^2_{H^2(K)}\right)^{1/2} \left(c_{I,1}^2 \theta_K h_K \|\partial_t v_h\|^2_{L_2(K)}\right)^{1/2}.
\end{aligned}$$

Now combining the above estimates, applying Cauchy's inequality and gathering all similar items, we finally arrive at the estimate

$$\begin{aligned}
&|a_h(u, v_h)| \\
&\leq \left(\|u\|^2_{L_2(\Sigma_T)} + \sum_{K\in\mathscr{T}_h} \left[\theta_K h_K \|\partial_t u\|^2_{L_2(K)} + 2(1+\theta_K c_{Tr}^2 \frac{\overline{\nu}_K^2}{\underline{\nu}_K} h_K^{-1}) \|\nabla_x u\|^2_{L_2^\nu(K)}\right.\right. \\
&\qquad \left.\left. + (\theta_K h_K)^{-1} \|u\|^2_{L_2(K)} + 2c_{Tr}^2 \overline{\nu}_K^2 \theta_K h_K |u|^2_{H^2(K)}\right]\right)^{1/2} \\
&\qquad \times \left(\|v_h\|^2_{L_2(\Sigma_T)} + \sum_{K\in\mathscr{T}_h} \left[(2+c_{I,1}^2)\theta_K h_K \|\partial_t v_h\|^2_{L_2(K)}\right.\right. \\
&\qquad\qquad \left.\left. + \big(1+(c_{I,1}\theta_K)^2\big) \|\nabla_x v_h\|^2_{L_2^\nu(K)}\right]\right)^{1/2} \\
&\leq \mu_b \|u\|_{h,*} \|v_h\|_h,
\end{aligned}$$

with $\mu_b := \max_{K\in\mathscr{T}_h} \left\{2(1+\theta_K h_K^{-1} c_{Tr}^2 \frac{\overline{\nu}_K^2}{\underline{\nu}_K}), 2c_{Tr}^2 \overline{\nu}_K^2, 2 + c_{I,1}^2, 1 + (c_{I,\nu}\theta_K)^2\right\}^{1/2}$. Choosing now $\theta_K = \mathscr{O}(h_K)$ ensures the uniform boundedness of the constant μ_b.

□

Remark 13.6 Choosing θ_K as in Lemma 13.5, i.e., $\theta_K = h_K/(c_{I,3}^2 \overline{\nu}_K)$, we obtain $\mu_c = 1/2$ and $\mu_b = \max_{K\in\mathscr{T}_h} \left\{2(1 + \frac{\overline{\nu}_K c_{Tr}^2}{\underline{\nu}_K c_{I,3}^2}), 2c_{Tr}^2 \overline{\nu}_K^2, 2 + c_{I,1}^2, 1 + (\frac{c_{I,\nu} h_K}{c_{I,3}^2 \overline{\nu}_K})^2\right\}^{1/2}$.

Remark 13.7 If we consider the bilinear form from Remark 13.2, we can derive an equivalent statement, but in a different norm $\|\cdot\|_{h,*}$ defined as

$$\|v\|_{h,*}^2 = \frac{1}{2}\|v\|_{L_2(\Sigma_T)}^2 + \sum_{K\in\mathscr{T}_h} \big[\theta_K h_K \|\partial_t v\|_{L_2(K)}^2 + \|\nabla_x v\|_{L_2^\nu(K)}^2 + (\theta_K h_K)^{-1}\|v\|_{L_2(K)}^2 + \theta_K h_K \|\mathrm{div}_x(\nu\nabla_x v)\|_{L_2(K)}^2\big].$$

By the same arguments as in the proof above, we estimate the first three terms in (13.11) by

$$\begin{aligned}
\left|\sum_{K\in\mathscr{T}_h}\int_K \partial_t u v_h \,\mathrm{d}(x,t)\right| &\le \sum_{K\in\mathscr{T}_h}\big[\big((\theta_K h_K)^{-1}\|u\|_{L_2(K)}^2\big)^{1/2}\big(\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2\big)^{1/2}\big] \\
&\quad + \big(\|u\|_{L_2(\Sigma_T)}^2\big)^{1/2}\big(\|v_h\|_{L_2(\Sigma_T)}^2\big)^{1/2}, \\
\left|\theta_K h_K \int_K \partial_t u \partial_t v_h \,\mathrm{d}(x,t)\right| &\le \big(\theta_K h_K\|\partial_t u\|_{L_2(K)}^2\big)^{1/2}\big(\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2\big)^{1/2}, \\
\left|\int_K \nu\nabla_x u\nabla_x v_h \,\mathrm{d}(x,t)\right| &\le \big(\|\nabla_x u\|_{L_2^\nu(K)}^2\big)^{1/2}\big(\|\nabla_x v_h\|_{L_2^\nu(K)}^2\big)^{1/2}.
\end{aligned}$$

For the fourth term, we just apply the Cauchy-Schwarz inequality to each term of the sum to obtain

$$\left|\theta_K h_K \int_K \mathrm{div}_x\,(\nu\nabla_x u)\,\partial_t v_h \,\mathrm{d}(x,t)\right| \le \left(\theta_K h_K \|\mathrm{div}_x\,(\nu\nabla_x u)\,\|_{L_2(K)}^2\right)^{1/2}\left(\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2\right)^{1/2}.$$

Now, combining the above estimates, applying Cauchy's inequality and reordering the terms, we finally obtain the estimate

$$\begin{aligned}
|a_h(u,\,v_h)| &\le \left(\|u\|_{L_2(\Sigma_T)}^2 + \sum_{K\in\mathscr{T}_h}\big[\theta_K h_K\|\partial_t u\|_{L_2(K)}^2 + \|\nabla_x u\|_{L_2^\nu(K)}^2 + (\theta_K h_K)^{-1}\|u\|_{L_2(K)}^2 + \theta_K h_K\|\mathrm{div}_x\,(\nu\nabla_x u)\,\|_{L_2(K)}^2\big]\right)^{1/2} \\
&\quad\times\left(\|v_h\|_{L_2(\Sigma_T)}^2 + \sum_{K\in\mathscr{T}_h}\big[3\theta_K h_K\|\partial_t v_h\|_{L_2(K)}^2 + \|\nabla_x v_h\|_{L_2^\nu(K)}^2\big]\right)^{1/2} \\
&\le 3\|u\|_{h,*}\|v_h\|_h.
\end{aligned}$$

Thus, the bilinear form (13.11) is bounded for all choices of θ_K.

Remark 13.8 As in Remark 13.5, we can provide a simplified estimate for the special case $p = 1$ and $\nu|_K = \nu_K = const$. The first three terms can be estimated as in the above proof. The fourth term completely vanishes, since $\nabla_x(\partial_t v_h) = 0$. For the fifth term, we use the fact that $\partial_t v_h = const$ on $K \in \mathscr{T}_h$, Gauss' theorem and the Cauchy-Schwarz inequality, obtaining

$$|\theta_K h_K \int_{\partial K} \nu_K \nabla_x u \cdot n_x \partial_t v_h \, \mathrm{d}s_{(x,t)}| \leq \left(\theta_K h_K \nu_K^2 \|\Delta_x u\|^2_{L_2(K)}\right)^{1/2} \left(\theta_K h_K \|\partial_t v_h\|^2_{L_2(K)}\right)^{1/2}.$$

Gathering the terms from the proof and the above estimate, we get

$$\begin{aligned}
|a_h(u, v_h)| \leq & \left(\|u\|^2_{L_2(\Sigma_T)} + \sum_{K \in \mathscr{T}_h} \left[\theta_K h_K \|\partial_t u\|^2_{L_2(K)} + \|\nabla_x u\|^2_{L_2^\nu(K)} \right.\right. \\
& \left.\left. + (\theta_K h_K)^{-1} \|u\|^2_{L_2(K)} + \nu_K^2 \theta_K h_K \|\Delta_x u\|^2_{L_2(K)} \right] \right)^{1/2} \\
& \times \left(\|v_h\|^2_{L_2(\Sigma_T)} \right. \\
& \left. + \sum_{K \in \mathscr{T}_h} \left[3\theta_K h_K \|\partial_t v_h\|^2_{L_2(K)} + \|\nabla_x v_h\|^2_{L_2^\nu(K)} \right] \right)^{1/2} \\
\leq & \max_{K \in \mathscr{T}_h} \{3, \nu_K^2\}^{1/2} \|u\|_{h,*} \|v_h\|_h .
\end{aligned}$$

We immediately deduce that this new constant $\tilde{\mu}_b = \max_{K \in \mathscr{T}_h} \{3, \nu_K^2\}^{1/2}$ is also independent of h_K for all choices of positive θ_K, $K \in \mathscr{T}_h$.

Coercivity, boundedness, and consistency of the bilinear form $a_h(\cdot, \cdot)$ immediately yield a Céa-like estimate of the discretization error in the norm $\|\cdot\|_h$ by the best approximation error in the norm $\|\cdot\|_{h,*}$.

Lemma 13.8 *Let the assumptions of the coercivity Lemma 13.5 and the boundedness Lemma 13.7 hold, and let the solution u of the space-time variational problem (13.4) belong to $H^{2,1}_{0,\underline{0}}(\mathscr{T}_h)$. Then the discretization error estimate*

$$\|u - u_h\|_h \leq \left(1 + \frac{\mu_b}{\mu_c}\right) \inf_{v_h \in V_{0h}} \|u - v_h\|_{h,*} \tag{13.23}$$

hold, where $u_h \in V_{0h}$ denotes the solution of the space-time finite element scheme (13.12), and the norms $\|\cdot\|_h$ and $\|\cdot\|_{h,*}$ are defined by (13.13) and (13.21), respectively.

Proof First, from the consistency identity (13.9) and the space-time finite element scheme (13.12), we immediately deduce Galerkin orthogonality, i.e.,

$$a_h(u - u_h, v_h) = 0, \quad \forall v_h \in V_{0h}. \tag{13.24}$$

We start with the triangle inequality for the discretization error, i.e.,

$$\|u - u_h\|_h \le \|u - v_h\|_h + \|v_h - u_h\|_h.$$

Applying ellipticity proved in Lemma 13.5, the Galerkin orthogonality (13.24) and the generalized boundedness from Lemma 13.7 to the second term, we get

$$\begin{aligned}\mu_c \|v_h - u_h\|_h^2 &\le a_h(v_h - u_h, v_h - u_h) = a_h(v_h - u, v_h u - u_h)\\ &\le \mu_b \|v_h - u\|_{h,*} \|v_h - u_h\|_h.\end{aligned}$$

Inserting this estimate in the triangle inequality above, we obtain

$$\|u - u_h\|_h \le \|u - v_h\|_h + \frac{\mu_b}{\mu_c} \|v_h - u\|_{h,*}. \tag{13.25}$$

Since $\|u - v_h\|_h \le \|v_h - u\|_{h,*}$, we immediately get the Céa-like estimate (13.23). □

Remark 13.9 Remark 13.7 immediately implies that the Céa-like estimate (13.23) is also valid for solutions u from $H^{L,1}_{0,\underline{0}}(\mathscr{T}_h)$ provided that the norm $\|\cdot\|_{h,*}$ is now defined as in Remark 13.7.

To obtain a priori error estimates w.r.t. to the mesh dependent norm (13.13), we need approximation respectively interpolation error estimates for the finite element spaces V_{0h} w.r.t. the norm (13.21), which we summarize in the next Lemmas. Moreover, we need the *broken Sobolev space*

$$H^l(\mathscr{T}_h) := \{v \in L_2(Q) : v|_K \in H^l(K) \ \forall K \in \mathscr{T}_h\},$$

equipped with the *broken Sobolev (semi-)norm*

$$|v|^2_{H^l(\mathscr{T}_h)} := \sum_{K \in \mathscr{T}_h} |v|^2_{H^l(K)} \quad \text{and} \quad \|v\|^2_{H^l(\mathscr{T}_h)} := \sum_{K \in \mathscr{T}_h} \|v\|^2_{H^l(K)},$$

where l is some positive integer. For further details on such spaces, we refer to [11, 35].

Lemma 13.9 *Let l and k be positive integers with $l \ge k > (d+1)/2$, and let $v \in V_0 \cap H^k(Q) \cap H^l(\mathscr{T}_h)$, where $\{\mathscr{T}_h\}_{h>0}$ is a shape regular family of subdivisions of Q. Then there exists an interpolation operator Π_h, mapping from $V_0 \cap H^k(Q)$ to V_{0h}, such that*

$$\|v - \Pi_h v\|_{L_2(K)} \le C\, h_K^s\, |v|_{H^s(K)}, \tag{13.26}$$

$$\|\nabla(v - \Pi_h v)\|_{L_2(K)} \le C\, h_K^{s-1}\, |v|_{H^s(K)}, \tag{13.27}$$

$$|v - \Pi_h v|_{H^2(K)} \le C\, h_K^{s-2}\, |v|_{H^s(K)}, \tag{13.28}$$

where C is some generic constant independent of h_K and v, $s = \min\{l, p+1\}$, and p denotes the polynomial degree of the finite element shape functions on the reference element, and $V_0 = H^{1,1}_{0,\underline{0}}(Q)$.

Proof See e.g. [8, Theorem 4.4.4] or [9, Theorem 3.1.6]. □

Lemma 13.10 *Let the assumptions of Lemma 13.9 hold. Then the following interpolation error estimates are valid:*

$$\|v - \Pi_h v\|_{L_2(\Sigma_T)} \leq c_1 \Big(\sum_{\substack{K \in \mathscr{T}_h \\ \partial K \cap \Sigma_T \neq \emptyset}} h_K^{2s-1} |v|^2_{H^s(K)} \Big)^{1/2}, \tag{13.29}$$

$$\|v - \Pi_h v\|_h \leq c_2 \Big(\sum_{K \in \mathscr{T}_h} h_K^{2(s-1)} |v|_{H^s(K)} \Big)^{1/2}, \tag{13.30}$$

$$\|v - \Pi_h v\|_{h,*} \leq c_3 \Big(\sum_{K \in \mathscr{T}_h} h_K^{2(s-1)} |v|_{H^s(K)} \Big)^{1/2}, \tag{13.31}$$

with positive constants c_1, c_2 and c_3 that do not depend on v or h_K provided that $\theta_K = \mathscr{O}(h_K)$ for all $K \in \mathscr{T}_h$.

Proof We start with the first estimate (13.29). We use the scaled trace inequality (13.22), and the interpolation error estimates (13.26) and (13.27), obtaining

$$\begin{aligned}
\|v - \Pi_h v\|^2_{L_2(\Sigma_T)} &= \sum_{\substack{K \in \mathscr{T}_h \\ \partial K \cap \Sigma_T \neq \emptyset}} \|v - \Pi_h v\|^2_{L_2(\partial K \cap \Sigma_T)} \leq \sum_{\substack{K \in \mathscr{T}_h \\ \partial K \cap \Sigma_T \neq \emptyset}} \|v - \Pi_h v\|^2_{L_2(\partial K)} \\
&\leq \sum_{\substack{K \in \mathscr{T}_h \\ \partial K \cap \Sigma_T \neq \emptyset}} \big[2c^2_{Tr} h_K^{-1} (\|v - \Pi_h v\|^2_{L_2(K)} + h_K^2 \|\nabla (v - \Pi_h v)\|^2_{L_2(K)}) \big] \\
&\leq c^2_{Tr} C^2 \sum_{\substack{K \in \mathscr{T}_h \\ \partial K \cap \Sigma_T \neq \emptyset}} [h_K^{2s-1} |v|_{H^s(K)}].
\end{aligned}$$

To prove (13.30), we use definition (13.13), assumption (13.14), the interpolation error estimate (13.27), and the above estimate (13.29), and obtain

$$\begin{aligned}
\|v - \Pi_h v\|^2_h &= \sum_{K \in \mathscr{T}_h} \big[\theta_K h_K \|\partial_t (v - \Pi_h v)\|^2_{L_2(K)} + \|\nabla_x (v - \Pi_h v)\|^2_{L_2^\nu(K)} \big] \\
&\quad + \frac{1}{2} \|v - \Pi_h v\|^2_{L_2(\Sigma_T)} \\
&\leq \sum_{K \in \mathscr{T}_h} \Big[(C^2 \theta_K h_K + \overline{\nu}_K C^2 + c_1^2 h_K) h_K^{2(s-1)} |v|^2_{H^s(K)} \Big].
\end{aligned}$$

For the last estimate (13.31), we use definition (13.21), the above estimate (13.30), and the interpolation error estimate (13.28), obtaining

$$\begin{aligned}\|v - \Pi_h v\|_{h,*}^2 &= \|v - \Pi_h v\|_h^2 \\ &\quad + \sum_{K \in \mathscr{T}_h} \left[(\theta_K h_K)^{-1} \|v - \Pi_h v\|_{L_2(K)}^2 + \theta_K h_K |v - \Pi_h v|_{H^2(K)}^2 \right] \\ &\leq \sum_{K \in \mathscr{T}_h} \left(c_2^2 + h_K \theta_K^{-1} C^2 + \theta_K h_K^{-1} C^2 \right) h_K^{2(s-1)} |v|_{H^s(K)}^2 .\end{aligned}$$

The special choice $\theta_K = \mathscr{O}(h_K)$ ensures that the constant c_3 is independent of h_K. □

Remark 13.10 The strong assumption $v \in H^k(Q)$ with $k > (d+1)/2$ is needed for the Lagrangian interpolation operator. However, in practical applications, where usually different materials occur, this requirement is too restrictive. In this case, the space-time cylinder $\overline{Q} = \bigcup_{i=1}^M \overline{Q_i}$ can be split into subdomains Q_i, which correspond to different materials. On each such subdomain Q_i, we can assume some regularity for the solution u, e.g., $u \in H^{\mathbf{l}}(\mathscr{T}(Q)) := \{v \in L_2(Q) : v|_{Q_i} \in H^{l_i}(Q_i),$ for all $i = 1, \ldots, M\}$ with some $\mathbf{l} = (l_1, \ldots, l_M) > 1$. For a similar case, Duan, Li, Tan and Zheng have shown a localized interpolation error estimate of the form

$$\|\nabla(v - I_h v)\|_{L_2(Q)} \leq C \sum_{i=1}^M h_i^{s_i - 1} \|v\|_{H^{s_i}(Q_i)},$$

in [13], where I_h is a special quasi-interpolation operator, and $s_i = \min\{l_i, p+1\}$.

Now we can formulate the following a priori estimate for the discretization error.

Theorem 13.3 *Let l and k be positive integers with $l \geq k > (d+1)/2$, $u \in V_0 \cap H^k(Q) \cap H^l(\mathscr{T}_h)$ be the exact solution, and $u_h \in V_{0h}$ be the solution of the finite element scheme (13.12). Furthermore, let the assumptions of the Lemmas 13.5 (coercivity), 13.7 (boundedness) and 13.10 (interpolation error estimates) be fulfilled. Then the a priori error estimate*

$$\|u - u_h\|_h \leq c \left(\sum_{K \in \mathscr{T}_h} h_K^{2(s-1)} |u|_{H^s(K)}^2 \right)^{1/2} \tag{13.32}$$

holds with $s = \min\{l, p+1\}$ and some generic positive constant c.

Proof Setting $v_h = \Pi_h u$ in (13.25), and using the interpolation error estimates (13.30) and (13.31), we obtain

$$\begin{aligned}\|u - u_h\|_h &\leq \|u - \Pi_h u\|_h + \frac{\mu_b}{\mu_c}\|\Pi_h u - u\|_{h,*} \\ &\leq \big(c_2 + c_3\frac{\mu_b}{\mu_c}\big)\Big(\sum_{K\in\mathscr{T}_h} h_K^{2(s-1)}|u|^2_{H^s(K)}\Big)^{1/2},\end{aligned}$$

which proves estimate (13.32) with $c = c_2 + c_3(\mu_b/\mu_c)$. □

Now we proceed with solving the discrete variational problem (13.12) that is nothing but one huge system of linear algebraic equations. Indeed, let $\{p^{(i)} : i = 1, \ldots, N_h\}$ be the finite element nodal basis of V_{0h}, i.e., $V_{0h} = \text{span}\{p^{(1)}, \ldots, p^{(N_h)}\}$, where N_h is the number of all space-time unknowns (dofs). Then we can express the approximate solution u_h in terms of this basis, i.e., $u_h(x,t) = \sum_{i=1}^{N_h} u_i\, p^{(i)}(x,t)$. Furthermore, each basis function is a valid test function. Thus, we obtain N_h equations from (13.12). We can rewrite this system in terms of a system of linear algebraic equations

$$\mathbf{K}_h\mathbf{u}_h = \mathbf{f}_h, \tag{13.33}$$

with $\mathbf{K}_h = (a_h(p^{(j)}, p^{(i)}))_{i,j=1,\ldots,N_h}$, $\mathbf{u}_h = (u_j)_{j=1,\ldots,N_h}$, $\mathbf{f}_h = (l_h(p^{(i)}))_{i=1,\ldots,N_h}$. The system matrix is non-symmetric, but positive definite due to Lemma 13.5. Indeed,

$$(\mathbf{K}_h\mathbf{v}_h, \mathbf{v}_h) = a_h(v_h, v_h) \geq \mu_c\|v_h\|_h^2 > 0$$

for all $V_{0h} \ni v_h \leftrightarrow \mathbf{v}_h \in \mathbf{R}^{N_h} : \mathbf{v}_h \neq 0$. In dependence on the dimension N_h, the linear system (13.33) of algebraic equations can efficiently be solved by means of a sparse direct solver (e.g., sparse LU-factorization) or an iterative solver (e.g., preconditioned GMRES). In particular, it turns out that parallel versions of the GMRES preconditioned by algebraic multigrid can solve large-scale systems with several millions of unknowns on distributed memory computers with several hundreds of cores in a few seconds, also see Example 13.1 in Sect. 13.4.

13.4 Implementation and Numerical Results

We implemented our conforming space-time finite element scheme with the help of MFEM [21], a C++ library for finite elements. The resulting linear systems were then solved by means of the GMRES method, preconditioned by one V-algebraic multigrid (AMG) cycle of *BoomerAMG*. As a stopping criterion we used the reduction of the initial residual by a factor of 10^{-8}. These methods were provided

by the solver library hypre.[1] We note that both libraries are already fully parallelized with MPI. All numerical tests were performed on the RADON1[2] high performance computing cluster at Linz. The initial (spatial) meshes were created by NETGEN [38], and the space-time meshes were obtained by means of an algorithm provided by Karabelas and Neumüller [19]. For visualization we either use GLVis [20] or ParaView [2].

Example 13.1 For the first example, we consider the unit (hyper-)cube $Q = (0,1)^{d+1}$, with $d = 2, 3$, as space-time cylinder, and choose the diffusion coefficient $\nu \equiv 1$. The manufactured function

$$u(x,t) = \prod_{i=1}^{d} \sin(x_i \pi) \sin(t\pi)$$

is chosen as the exact solution, where the right-hand side is computed accordingly. This solution is highly smooth, and thus fulfills all regularity assumptions made for deriving the a prior error estimate (13.32) with optimal rates. Hence, we really can expect optimal convergence rates provided that we choose θ_K as in Remark 13.4, i.e., on each element $K \in \mathscr{T}_h$, we numerically solve a small generalized eigenvalue problem with LAPACK [3]. Indeed, Fig. 13.1b shows optimal convergence rates for all tested polynomial degrees and spatial dimensions. Moreover, we can observe from Fig. 13.1c that the preconditioned GMRES method has an optimal strong scaling behavior for systems with $N_h = 4\,601\,025$ ($p = 1, 2$) and $N_h = 5\,764\,804$ ($p = 3$) unknowns in the case $d = 3$, i.e., $Q = (0,1)^4$. The stagnation of the scaling rate at 256 cores is due to an increased communication overhead since the problems become too small on each processor (only $\sim$15 000 dofs).

Example 13.2 Let us now consider an example with a moving interface in the unit hyper-cube $Q := (0,1)^{d+1}$, with $d = 2, 3$. The moving interface is defined by the discontinuous diffusion coefficient

$$\nu(x,t) = \begin{cases} 1 \times 10^2, & \text{for } 2x_1 - t < \frac{1}{2}, \\ 7 \times 10^5, & \text{for } 2x_1 - t > \frac{1}{2}; \end{cases}$$

see Fig. 13.2 (left). We choose the function

$$u(x,t) = \begin{cases} \sin\left(9\pi \left(2x_1 - t - \frac{1}{2}\right)^2 (x_1 - x_1^2)\right) \sin(4\pi t) g(x), & \text{for } 2x_1 - t \le \frac{1}{2}, \\ \sin\left(40\pi \left(2x_1 - t - \frac{1}{2}\right)^2 (x_1 - x_1^2)(t - t^2)\right) g(x), & \text{else}, \end{cases}$$

[1] https://www.llnl.gov/casc/hypre/.

[2] https://www.ricam.oeaw.ac.at/hpc/.

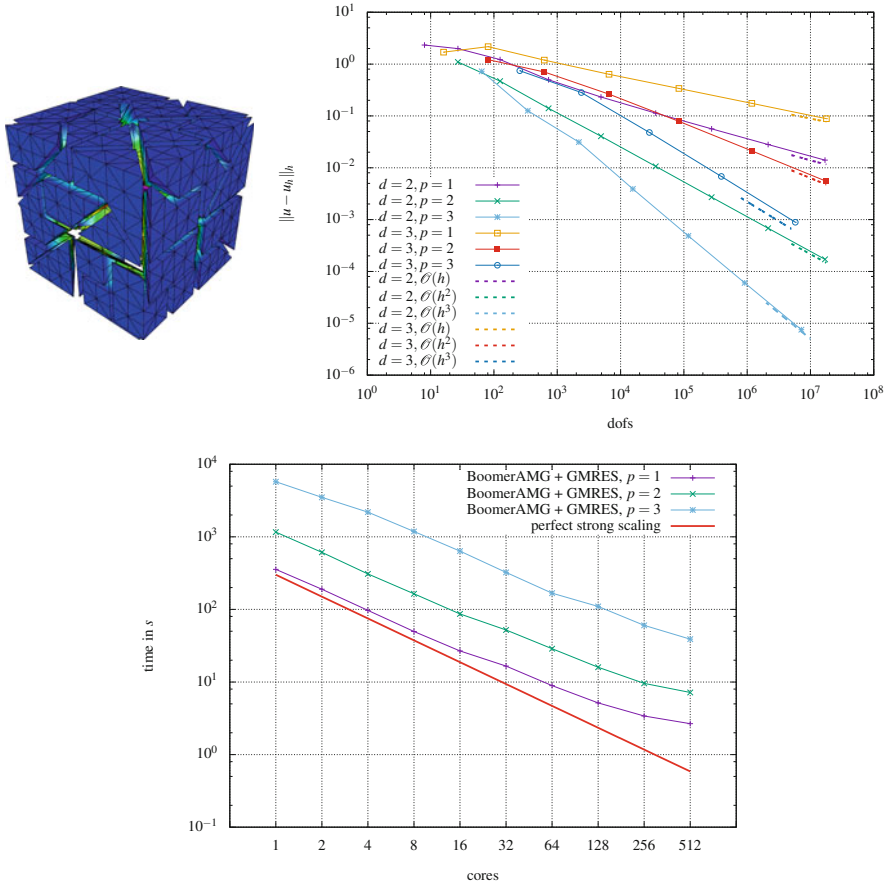


Fig. 13.1 Decomposition of the space-time cylinder into 64 subdomains for parallel computing (left); Error rates in the $\|\,.\,\|_h$-norm (right); Strong scaling of the solver for $d = 3$ and $N_h = 4\,601\,025, 4\,601\,025, 5\,764\,801$ for $p = 1, 2, 3$ (below)

with $g(x) = \prod_{i=2}^{d} \sin(\pi x_i)$, as our exact solution, and compute the corresponding right-hand side and initial data. This manufactured solution fulfills the interface conditions since this function and its first derivatives are 0 at the interface. Since this function is smooth on both sides of the moving interface, we expect optimal convergence rates; cf. Theorem 13.3. Indeed, for linear, quadratic and cubic shape functions, we observe optimal rates provided that we choose θ_K according to Remark 13.4; see Fig. 13.2 (right).

Example 13.3 For the third example, we consider the exact solution

$$u(x,t) = \left(x_1^2 - x_1\right)\left(x_2^2 - x_2\right)\left(t^2 - t\right) \mathrm{e}^{-100\left((x_1-0.25)^2+(x_2-0.25)^2+(t-0.25)^2\right)}$$

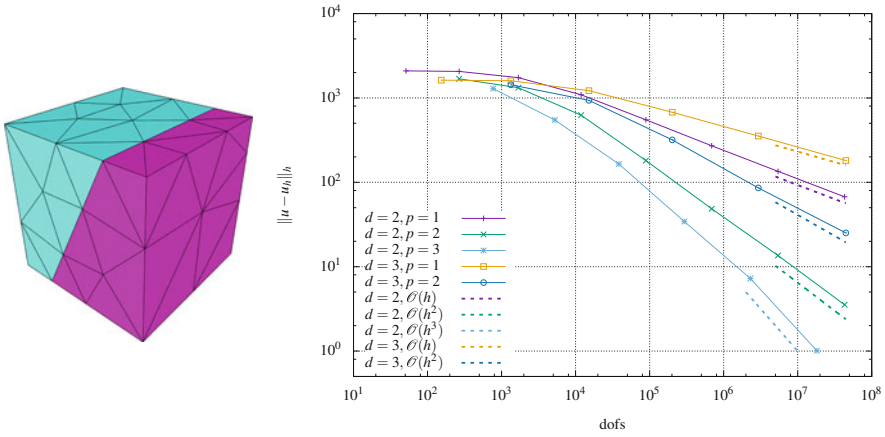


Fig. 13.2 Initial space-time mesh and diffusion coefficient $\nu(x,t)$ in color (left); Error rates in the $\|\,.\,\|_h$-norm (right)

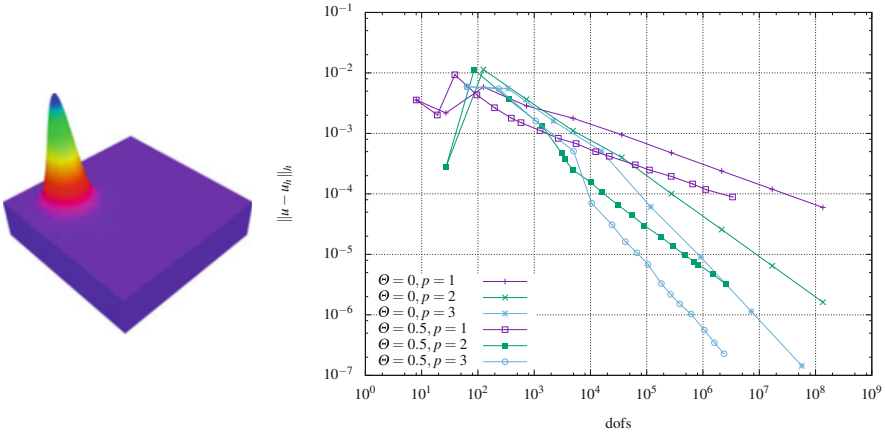


Fig. 13.3 Plot of the exact solution at $t = 0.25$ (left); Convergence rates in the $\|\,.\,\|_h$-norm (right)

in the unit cube $Q = (0,1)^3$, i.e. $d = 2$, and compute the initial and boundary conditions as well as the right-hand side accordingly, where we set $\nu \equiv 1$. This function is almost zero everywhere in the space-time cylinder Q except a small area around $(0.25, 0.25, 0.25)$; see Fig. 13.3 (left). This motivates the use of an a posteriori error estimator. In particular, we use the residual-based error indicator proposed by Steinbach and Yang in [40]. For each element $K \in \mathcal{T}_h$, we compute the error indicator

$$\eta_K := \left(h_K^2 \|R_h(u_h)\|^2_{L_2(K)} + h_K \|J_h(u_h)\|^2_{L_2(\partial K)} \right)^{1/2},$$

where u_h is the solution of the finite element scheme (13.12), and

$$R_h(u_h) := f + \mathrm{div}_x(\nu \nabla_x u_h) - \partial_t u_h \quad \text{in } K,$$
$$J_h(u_h) := [\nu \nabla_x u_h]_e \qquad\qquad \text{on } e \subset \partial K.$$

Here, $[\,.\,]_e$ denotes the jump across one face $e \subset \partial K$. We use a maximum marking strategy, i.e., given a parameter $\Theta \in [0, 1]$, we mark all elements whose error indicator fulfills the condition

$$\eta_K \geq \Theta \max_{K \in \mathcal{T}_h} \eta_K.$$

Unless stated otherwise, we set $\Theta = 0.5$. We note that uniform refinement is achieved by setting $\Theta = 0$. The exact solution in this example is smooth. Hence, we expect optimal convergence rates for uniform refinement after some pre-asymptotic range, which we indeed observe for all tested polynomial degrees; c.f., Fig. 13.3. For adaptive refinement, we get a better error w.r.t. to the absolute value and optimal convergence rates.

Example 13.4 For the fourth and last example, we consider the exact solution

$$u(x, t) = \sin\left(\frac{1}{\frac{1}{10\pi} + \sqrt{x_1^2 + x_2^2 + t^2}}\right),$$

in the unit cube $Q = (0, 1)^3$, i.e. $d = 2$, and compute the initial and boundary conditions as well as the right-hand side accordingly, where we set $\nu \equiv 1$. This function has a highly oscillatory behavior near the origin $(0, 0, 0)$ and is smooth everywhere else in the space-time cylinder Q; see Fig. 13.3 (left). This again motivates the use of an a posteriori error estimator. We use the same setup as in Example 13.3, i.e., the residual-based error indicator by Steinbach and Yang with a maximum marking strategy. For adaptive refinement, we recover the optimal rates for all polynomial degrees tested, whereas only reduced rates are observed for $p = 2, 3$; c.f., Fig. 13.4 (right). Moreover, we only need 47 330 dofs to obtain an energy error of the same magnitude as for 135 005 697 dofs after uniform refinement in the case $p = 1$.

13.5 Conclusions and Future Work

In this paper, following the classical books [23] and [24], we recalled that the parabolic initial boundary value problem (13.1)–(13.3) has a unique generalized (weak) solution in $H_0^{1,0}(Q)$ that even belongs to $V_{2,0}^{1,0}(Q)$. Already Ladyžhenskaya

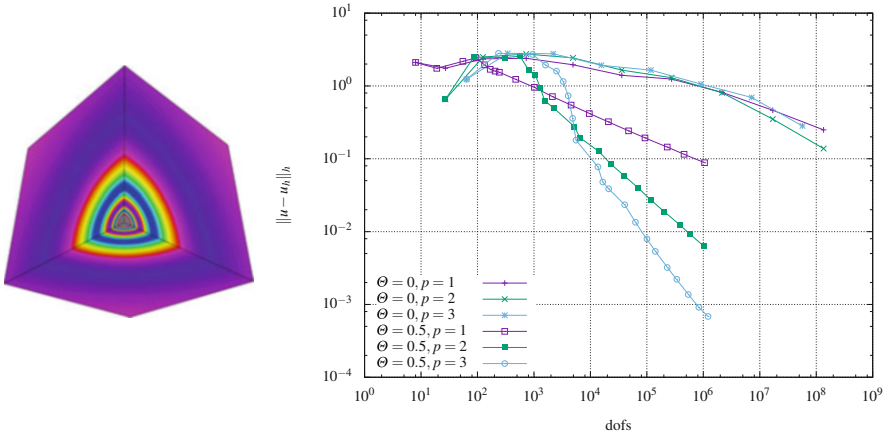


Fig. 13.4 Plot of the exact solution (left); Convergence rates in the $\|\,.\,\|_h$-norm (right)

proved that, in the case $\nu = 1$, the solution u even belongs to $H^{\Delta,1}(Q)$ provided that the right-hand side $f \in L_2(Q)$ and initial conditions $u_0 \in H_0^1(\Omega)$; see [23]. This setting of the so-called maximal parabolic regularity was also considered in this paper. We again mention that we only need this property element-wise to construct a consistent and stable space-time finite element scheme. We proceeded with deriving a stable space-time finite element scheme, for which we showed coercivity (ellipticity) and boundedness on the finite element spaces respectively extended finite element spaces. These properties together with consistency and standard interpolation or quasi-interpolation error estimates led to a priori discretization error estimates in the corresponding mesh-dependent norm with optimal rates. We performed several numerical experiments with four test problems possessing different features. The first example has a smooth solution that led to optimal convergence rate as predicted by the theory. Moreover, due to the ellipticity of the bilinear form $a_h(\cdot,\cdot)$, the AMG precondition GMRES is a very efficient parallel solver. The second example has a moving interface that is given by a discontinuous diffusion coefficient $\nu(x,t)$ depending on both x and t. In the third and fourth example, we studied adaptivity based on the a posteriori residual error indicator proposed in [40]. It is clear that the interplay of adaptivity and fast parallel iterative solvers will lead to the most efficient completely unstructured adaptive space-time solvers for complicated initial-boundary value problems for linear and even non-linear parabolic partial differential equations. Adaptive Space-Time Finite Element Methods and Solvers can be useful for solving eddy current problems with moving and non-moving parts like in electrical machines. In many practical applications, one is interested in optimal control or in optimal design of electrical machines; see, e.g., [17]. Adaptive Space-Time Finite Element Methods are especially suited for solving the optimality system that is nothing but a coupled PDE system living in the space-time cylinder Q; see, e.g., [44].

Acknowledgements This work was supported by the Austrian Science Fund (FWF) under the grant W1214, project DK4. This support is gratefully acknowledged. Furthermore, the authors would like to express their thanks to the anonymous referees for their helpful hints and valuable suggestions.

References

1. Ahmed, N., Matthies, G.: Numerical study of SUPG and LPS methods combined with higher order variational time discretization schemes applied to time-dependent linear convection–diffusion–reaction equations. J. Sci. Comput. **67**(3), 988–1018 (2016)
2. Ahrens, J., Geveci, B., Law, C.: ParaView: An End-User Tool for Large Data Visualization. Elsevier, San Diego (2005)
3. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide, 3rd edn. SIAM, Philadelphia (1999)
4. Aziz, A.K., Monk, P.: Continuous finite elements in space and time for the heat equation. Math. Comput. **52**, 255–274 (1989)
5. Bank, R.E., Vassilevski, P., Zikatanov, L.: Arbitrary dimension convection-diffusion scheme for space-time discretizations. J. Comput. Appl. Math. **310**, 19–31 (2017)
6. Bause, M., Radu, F.A., Köcher, U.: Error analysis for discretizations of parabolic problems using continuous finite elements in time and mixed finite elements in space. Numer. Math. **137**(4), 773–818 (2017)
7. Braess, D.: Finite Elemente; Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie. Springer-Verlag, Berlin (2007)
8. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15, 3rd edn. Springer, New York (2008)
9. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland Publishing, Amsterdam (1978)
10. Devaud, D., Schwab, C.: Space–time hp-approximation of parabolic equations. Calcolo **55**(3), 35 (2018)
11. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 69. Springer, Heidelberg (2012)
12. Dörfler, W., Findeisen, S., Wieners, C.: Space-time discontinuous Galerkin discretizations for linear first-order hyperbolic evolution. Comput. Methods Appl. Math. **16**, 409–428 (2016)
13. Duan, H., Li, S., Tan, R.C.E., Zheng, W.: A delta-regularization finite element method for a double curl problem with divergence-free constraint. SIAM J. Numer. Anal. **50**(6), 3208–3230 (2012)
14. Ern, A., Schieweck, F.: Discontinuous Galerkin method in time combined with a stabilized finite element method in space for linear first-order PDEs. Math. Comput. **85**(301), 2099–2129 (2016)
15. Gander, M., Neumüller, M.: Analysis of a new space-time parallel multigrid algorithm for parabolic problems. SIAM J. Sci. Comput. **38**(4), A2173–A2208 (2016)
16. Gander, M.J.: 50 years of time parallel integration. In: Multiple Shooting and Time Domain Decomposition, pp. 69–114. Springer Verlag, Heidelberg (2015)
17. Gangl, P., Langer, U., Laurain, A., Meftahi, H., Sturm, K.: Shape optimization of an electric motor subject to nonlinear magnetostatics. SIAM J. Sci. Comput. **37**(6), B1002–B1025 (2015)
18. Hackbusch, W.: Parabolic multigrid methods. In: Glowinski, R., Lions, J.L. (eds.) Computing Methods in Applied Sciences and Engineering VI, pp. 189–197. North-Holland, Amsterdam (1984)

19. Karabelas, E., Neumüller, M.: Generating admissible space-time meshes for moving domains in d+1-dimensions. Technical Report 07, Institute of Computational Mathematics, Johannes Kepler University Linz (2015)
20. Kolev, T., Dobrev, V.: GLVis: OpenGL finite element visualization tool. https://github.com/glvis/glvis, 6 (2010)
21. Kolev, T., Dobrev, V.: MFEM: modular finite element methods library. https://github.com/mfem/mfem, 6 (2010)
22. Kunstmann, P.C., Weis, L.: Maximal L_p-regularity for Parabolic Equations, Fourier Multiplier Theorems and H^∞-functional Calculus, pp. 65–311. Springer-Verlag, Berlin (2004)
23. Ladyžhenskaya, O.A.: The Boundary Value Problems of Mathematical Physics. Applied Mathematical Sciences, vol. 49. Springer-Verlag, New York (1985)
24. Ladyzhenskaya, O.A., Solonnikov, V.A., Uraltseva, N.N.: Linear and Quasilinear Equations of Parabolic Type. AMS, Providence (1968)
25. Lang, J.: Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems: Theory, Algorithm, and Applications. LNCSE, vol. 16. Springer-Verlag, Berlin (2001)
26. Langer, U., Moore, S.E., Neumüller, M.: Space-time isogeometric analysis of parabolic evolution problems. Comput. Methods Appl. Mech. Eng. **306**, 342–363 (2016)
27. Leykekhman, D., Vexler, B.: Discrete maximal parabolic regularity for Galerkin finite element methods. Numer. Math. **135**, 923–952 (2017)
28. Lions, J.-L., Maday, Y., Turinici, G.: A "parareal" in time discretization of PDE's. C. R. Acad. Sci. Paris I **332**, 661–668 (2001)
29. Moore, S.: Nonstandard discretization strategies in isogeometric analysis for partial differential equations. PhD thesis, Johannes Kepler University Linz, Linz (2017)
30. Moore, S.: A stable space–time finite element method for parabolic evolution problems. Calcolo **55**, 18 (2018). First Online: 16 April 2018.
31. Neumüller, M.: Space-time Methods; Fast Solvers and Applications. Monographic Series TU Graz: Computation in Engineering and Science, vol. 20. TU Graz, Graz (2013)
32. Neumüller, M., Steinbach, O.: Refinement of flexible space-time finite element meshes and discontinuous Galerkin methods. Comput. Vis. Sci. **14**(5), 189–205 (2011)
33. Otoguro, Y., Takizawa, K., Tezduyar, T.E.: Space-time VMS computational flow analysis with isogeometric discretization and a general-purpose NURBS mesh generation method. Comput. Fluids **158**, 189–200 (2017)
34. Rhebergen, S., Cockburn, B., van der Vegt, J.J.W.: A space-time discontinuous Galerkin method for the incompressible Navier-Stokes equations. J. Comput. Phys. **233**, 339–358 (2013)
35. Rivière, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations. Theory and Implementation. Frontiers in Applied Mathematics, vol. 35. Society for Industrial and Applied Mathematics, Philadelphia (2008)
36. Schafelner, A.: Space-time finite element methods for parabolic inital-boundary problems. Master's thesis, Johannes Kepler University Linz (2017)
37. Schieweck, F.: *A*-stable discontinuous Galerkin-Petrov time discretization of higher order. J. Numer. Math. **18**(1), 25–57 (2010)
38. Schöberl, J.: NETGEN: an advancing front 2D/3D-mesh generator based on abstract rules. Comput. Vis. Sci. **1**(1), 41–52 (1997)
39. Steinbach, O.: Space-time finite element methods for parabolic problems. Comput. Methods Appl. Math. **15**(4), 551–566 (2015)
40. Steinbach, O., Yang, H.: Comparison of algebraic multigrid methods for an adaptive space-time finite-element discretization of the heat equation in 3d and 4d. Numer. Linear Algebra Appl. **25**(3) (2018). First published online as e2143, 7 February 2018
41. Steinbach, O., Yang, H.: Space–time finite element methods for parabolic evolution equations: discretization, a posteriori error estimation, adaptivity and solution. In: U. Langer, O. Steinbach (eds.) Space-Time Methods: Application to Partial Differential Equations. Radon Series on Computational and Applied Mathematics, vol. 25. pp. 226–259, de Gruyter (2019)
42. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer Series in Computational Mathematics, vol. 25, 2nd edn. Springer-Verlag, Berlin (2006)

43. Toulopoulos, I.: Stabilized space-time finite element methods of parabolic evolution problems. Technical Report 2017–19, RICAM, Linz (2017)
44. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications. AMS, Providence (2010)
45. Vandewalle, S.: Parallel Multigrid Waveform Relaxation for Parabolic Problems. Teubner Skripten zur Numerik. Teubner, Stuttgart (1993)
46. Wloka, J.: Partial Differential Equations. Cambridge University Press, Cambridge (1987)

Chapter 14
ACA Improvement by Surface Segmentation

Sergej Rjasanow and Steffen Weißer

Abstract In this paper, we present a modification of the clustering procedure for the fast Boundary Element Method (BEM), based on hierarchical techniques for the matrix decomposition and Adaptive Cross Approximation (ACA). An initial segmentation of the surface elements is shown to be a reasonable tool to prevent problematic blocks which appear on surfaces with edges. It leads to significantly easier control of the Partial ACA algorithm and our numerical results show perfect convergence of all numerical quantities corresponding to the theory of BEM. In particular, third order convergence is reached for the gradient of the solution inside the domain.

14.1 Introduction

The Boundary Element Method (BEM) leads to fully populated matrices and, therefore, to asymptotically non-optimal memory requirements as well as a non-optimal number of numerical operations. The ACA algorithms can be efficiently used for BEM matrices, where they are applied to the admissible blocks of the matrices in the hierarchical $H-$matrix format. The $H-$matrices were introduced by W. Hackbusch in [11], see also a recent monograph [12], and the first variant of the ACA was published by M. Bebendorf in [1]. Meanwhile, there are many generalisations and improvements of the original strategy and we only mention [3–5, 13, 18, 19]. For a more detailed discussion on the fully pivoted ACA algorithm we refer to [17]. Unfortunately, the convergence of the ACA approximation is non-monotone in general. Especially the most natural but heuristic stopping criterion of the partially pivoted ACA algorithm sometimes exhibits jumps of several orders of magnitude. This behaviour is well known in the literature and there are more sophisticated pivoting strategies, which try to improve the situation,

S. Rjasanow (✉) · S. Weißer
Saarland University, Department of Mathematics, Saarbrücken, Germany
e-mail: rjasanow@num.uni-sb.de; weisser@num.uni-sb.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_14

see, e.g., [2, 10]. Other methods combine kernel and matrix-based approximation techniques to overcome this issue, see [6, 7]. This behaviour is amplified if the BEM discretisation contains sharp edges and corners and they are not resolved by the clusters of the $H-$matrices. In contrast to the previously cited literature, we suggest a surface segmentation before clustering in order to prevent surface clusters to contain geometric edges. Thus, our main idea is not to modify the original ACA by Bebendorf in [1], see also [17], and not to develop new pivoting strategies, but to prevent ACA from approximating blocks with extremely non-smooth background, especially due to clusters of surface triangles containing edges of the geometry.

For the surface segmentation, we propose a rather simple and efficient approach which is suitable for our purpose. We detect geometric edges in the surface triangulation and perform a region growing algorithm to form surface segments on each side of the geometric edges. In a second step, we might reduce the number of found segments. In the literature of computer sciences, there are alternative algorithms for this segmentation process applying the watershed algorithm or clustering the surface normals, see, e.g., [14, 22, 23].

The paper is organised as follows. A short description of the problem and of the BEM is presented in Sect. 14.2. In Sect. 14.3, the ACA approximation is presented in a form of the fully pivoted algorithm. Furthermore, possible control problems by application of the partially pivoted ACA algorithm are discussed. A corresponding numerical example is given for illustration. A segmentation procedure for triangulated surfaces and the corresponding algorithms are presented in Sect. 14.4. Finally, in Sect. 14.5, we show numerical results for an example with segmentation of the surface and comment on the improvement. The ACA accelerated BEM with surface segmentation exhibit optimal cubic convergence of the numerical solution as well as of its gradient.

14.2 Boundary Element Method

The BEM can efficiently be applied to partial differential equations with constant coefficients. As model problem, we consider the Laplace equation in a bounded domain $\Omega \subset \mathbb{R}^3$ with Dirichlet boundary conditions on $\Gamma = \partial\Omega$:

$$-\Delta u(\boldsymbol{x}) = 0 \quad \text{for } \boldsymbol{x} \in \Omega\,, \qquad \gamma_0 u(\boldsymbol{x}) = g(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Gamma\,. \tag{14.1}$$

14.2.1 Boundary Integral Formulation

The solution of this problem is given by the representation formula

$$u(\boldsymbol{x}) = \int_\Gamma u^*(\boldsymbol{x}, \boldsymbol{y}) t(\boldsymbol{y}) ds_{\boldsymbol{y}} - \int_\Gamma \gamma_{1,\boldsymbol{y}} u^*(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{y}) ds_{\boldsymbol{y}}\,, \tag{14.2}$$

for $\boldsymbol{x} \in \Omega$, where $t = \gamma_1 u$ is the unknown Neumann datum. The fundamental solution u^* of the Laplace equation is

$$u^*(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{4\pi} \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|} \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3 . \tag{14.3}$$

For sufficiently regular u, the Neumann trace $\gamma_1 u$ is defined as

$$t(\boldsymbol{x}) = \gamma_1 u(\boldsymbol{x}) = (\gamma_0 \operatorname{grad} u(\boldsymbol{x}), \boldsymbol{n}_{\boldsymbol{x}}) \quad \text{for } \boldsymbol{x} \in \Gamma .$$

By applying the interior trace operator γ_0 to the representation formula (14.2) and using the jump relations (see e.g. [15]) and the Dirichlet boundary condition, we obtain the boundary integral equation

$$\int_\Gamma u^*(\boldsymbol{x}, \boldsymbol{y}) t(\boldsymbol{y}) ds_{\boldsymbol{y}} = \frac{1}{2} g(\boldsymbol{x}) + \int_\Gamma \gamma_{1,\boldsymbol{y}} u^*(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{y}) ds_{\boldsymbol{y}} \quad \text{for } \boldsymbol{x} \in \Gamma . \tag{14.4}$$

Thus, we have to solve a first kind boundary integral equation to find the Neumann datum $t \in H^{-1/2}(\Gamma)$ such that

$$(Vt)(\boldsymbol{x}) = \frac{1}{2} g(\boldsymbol{x}) + (Kg)(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Gamma , \tag{14.5}$$

where $V : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ denotes the single layer potential, which is self-adjoint and positive definite, and $K : H^{1/2}(\Gamma) \to H^{1/2}(\Gamma)$ the double layer potential.

14.2.2 Boundary Element Discretisation

To obtain a boundary element discretisation of the problem, we first assume that the domain Ω is a polyhedron and we mesh its boundary Γ by a conforming surface triangulation Γ_h with N_{BEM} plane triangles τ_ℓ and M_{BEM} nodes $\boldsymbol{x}_i$, see Fig. 14.1. We use the piece-wise constant functions

$$\Psi = \left(\psi_1, \ldots, \psi_{N_{\mathrm{BEM}}}\right),$$

where ψ_ℓ is 1 on triangle τ_ℓ and 0 outside τ_ℓ, as basis and test functions for the discretised single layer potential. For the double layer potential, in contrast, we utilise piece-wise linear basis functions, i.e.

$$\Phi = \left(\varphi_1, \ldots, \varphi_{M_{\mathrm{BEM}}}\right),$$

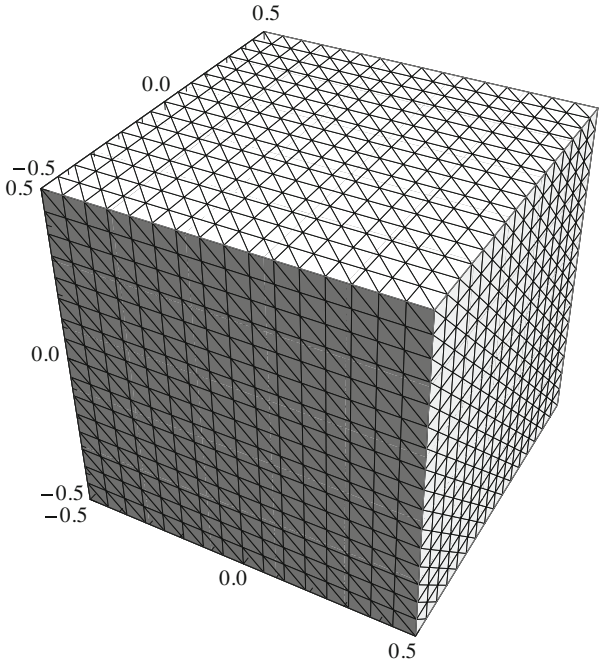


Fig. 14.1 BEM discretisation for $N_{\mathrm{BEM}} = 3072$ and $M_{\mathrm{BEM}} = 1538$

where $\varphi_j(\boldsymbol{x}_i) = \delta_{ij}$ and φ_j is linear on each τ_ℓ, and piece-wise constant test functions. Thus the BEM Galerkin procedure leads to a system of linear equations

$$V_h \underline{t} = \Big(\frac{1}{2} M_h + K_h\Big)\underline{g}\,,$$

where the Galerkin approximation of the unknown Neumann datum is sought in the space span Ψ of piece-wise constant functions as

$$t_h = \Psi \underline{t}\,, \quad \underline{t} \in \mathbb{R}^{N_{\mathrm{BEM}}}\,,$$

and

$$g_h = \Phi \underline{g}\,, \quad \underline{g} \in \mathbb{R}^{M_{\mathrm{BEM}}}$$

denotes the L_2-projection of the given Dirichlet datum onto the space span Φ of piece-wise linear functions. The matrix V_h of the above system is symmetric and positive definite. Thus, the CG method can be applied for large systems. If the dimension N_{BEM} is moderate, the Cholesky decomposition provided by LAPACK

is the best solution strategy. The practical realisation consists of the following steps. First, the BEM matrices are computed, namely the double layer potential matrix

$$K_h \in \mathbb{R}^{N_{\mathrm{BEM}} \times M_{\mathrm{BEM}}}, \quad K_h[k,j] = \frac{1}{4\pi} \int_{\tau_k} \int_{\Gamma_h} \frac{(\boldsymbol{x}-\boldsymbol{y}, \boldsymbol{n}_{\boldsymbol{y}})}{|\boldsymbol{x}-\boldsymbol{y}|^3} \varphi_j(\boldsymbol{y})\, ds_{\boldsymbol{y}}\, ds_{\boldsymbol{x}},$$

the single layer potential matrix

$$V_h \in \mathbb{R}^{N_{\mathrm{BEM}} \times N_{\mathrm{BEM}}}, \quad V_h[k,\ell] = \frac{1}{4\pi} \int_{\tau_k} \int_{\tau_\ell} \frac{1}{|\boldsymbol{x}-\boldsymbol{y}|}\, ds_{\boldsymbol{y}}\, ds_{\boldsymbol{x}},$$

the mixed mass matrix

$$M_h \in \mathbb{R}^{N_{\mathrm{BEM}} \times M_{\mathrm{BEM}}}, \quad M_h[k,j] = \int_{\tau_k} \int_{\Gamma_h} \varphi_j(\boldsymbol{y})\, ds_{\boldsymbol{y}}\, ds_{\boldsymbol{x}},$$

and the linear mass matrix

$$M_h^{(1)} \in \mathbb{R}^{M_{\mathrm{BEM}} \times M_{\mathrm{BEM}}}, \quad M_h^{(1)}[i,j] = \int_{\Gamma_h} \int_{\Gamma_h} \varphi_j(\boldsymbol{y}) \varphi_i(\boldsymbol{x})\, ds_{\boldsymbol{y}}\, ds_{\boldsymbol{x}}.$$

Note that the matrices M_h and $M_h^{(1)}$ are sparse while the matrices K_h and V_h are dense and require probably an additional approximation technique. The final preparation step is the L_2-projection of the Dirichlet boundary condition g onto the space of piece-wise linear functions φ_j. This is equivalent to the numerical solution of the linear system

$$M_h^{(1)} \underline{g} = \underline{b}^{(1)}, \quad \underline{b}^{(1)} \in \mathbb{R}^{M_{\mathrm{BEM}}}, \quad \underline{b}_i^{(1)} = \int_{\Gamma_h} g(\boldsymbol{x}) \varphi_i(\boldsymbol{x})\, ds_{\boldsymbol{x}}.$$

The matrix $M_h^{(1)}$ is symmetric, positive definite and well conditioned. Thus, this system is solved with only a few CG iterations without preconditioning up to the computer accuracy. In the following, we give approximation properties of the BEM and refer to specialised literature for the details. The error of the numerical solution t_h behaves in the $L_2(\Gamma_h)$-norm linear, i.e.

$$\|t - t_h\|_{L_2(\Gamma_h)} = \mathcal{O}(h),$$

where

$$h = \max_{1 \le \ell \le N_{\mathrm{BEM}}} \operatorname{diam} \tau_\ell.$$

For the L_2-projection of the Dirichlet boundary condition g we have

$$\|g - g_h\|_{L_2(\Gamma_h)} = \mathcal{O}(h^2)\,.$$

The numerical solution obtained for $\boldsymbol{x} \in \Omega$ by the approximate representation formula

$$u_h(\boldsymbol{x}) = \int\limits_{\Gamma_h} u^*(\boldsymbol{x}, \boldsymbol{y}) t_h(\boldsymbol{y}) ds_{\boldsymbol{y}} - \int\limits_{\Gamma_h} \gamma_{1,\boldsymbol{y}} u^*(\boldsymbol{x}, \boldsymbol{y}) g_h(\boldsymbol{y}) ds_{\boldsymbol{y}} \tag{14.6}$$

is very accurate and it can be differentiated in order to obtain an approximation of the derivatives of u. More precisely, it holds cubic point-wise convergence for $x \in \Omega$ such that

$$|u(\boldsymbol{x}) - u_h(\boldsymbol{x})| = \mathcal{O}(h^3) \quad \text{and} \quad |\nabla u(\boldsymbol{x}) - \nabla u_h(\boldsymbol{x})| = \mathcal{O}(h^3)\,.$$

For more details on the error analysis, and especially on the smoothness requirements on the Dirichlet boundary data g, we refer to [21] and [20]. If the domain Ω is not a polyhedron, an additional error will appear due to the approximation of the boundary Γ by a system of plane triangles Γ_h. Such an approximation of Γ by Γ_h is, for instance, considered in [16] and the more recent paper [9]. In the numerical experiments of this paper, we apply the BEM for a cube domain and observe the optimal rates of convergence as shown above.

14.3 Adaptive Cross Approximation

The BEM matrices V_h and K_h are dense and, therefore, require an amount of computer memory and a computational time which both are quadratic with respect to N_{BEM}. Thus for dimensions $N_{\mathrm{BEM}} \simeq 20{,}000$ an ACA approximation of these matrices is in many cases more efficient leading to almost linear algorithms.

Let $A \in \mathbb{R}^{N \times M}$ be a given matrix.

Algorithm 14.1 (Fully Pivoted ACA)

1. Initialisation

$$R_0 = A\,, \quad S_0 = 0\,.$$

2. For $i = 0, 1, 2, \ldots$ compute

2.1. position of the pivot element

$$(k_{i+1}, \ell_{i+1}) = \mathrm{ArgMax}\,|(R_i)_{k\ell}|\,,$$

2.2. normalising constant

$$\gamma_{i+1} = \left((R_i)_{k_{i+1}\ell_{i+1}}\right)^{-1},$$

2.3. new vectors

$$u_{i+1} = \gamma_{i+1} R_i e_{\ell_{i+1}}, \quad v_{i+1} = R_i^\top e_{k_{i+1}},$$

2.4. new residual

$$R_{i+1} = R_i - u_{i+1} v_{i+1}^\top,$$

2.5. new approximation

$$S_{i+1} = S_i + u_{i+1} v_{i+1}^\top.$$

In Algorithm 14.1, e_j denotes the jth column of the identity matrix I. A natural stopping criterion for Algorithm 14.1 is

$$\frac{\|A - S_i\|_F}{\|A\|_F} = \frac{\|R_i\|_F}{\|A\|_F} \leq \varepsilon_{\mathrm{ACA}} \tag{14.7}$$

for some given $\varepsilon_{\mathrm{ACA}}$. The Frobenius norm of the matrix $R_0 = A$ is computed in Step 2.1 for $i = 0$ while the norm of R_i is calculated for $i = 1, 2, \ldots$. However, if the matrix A is not given and its computation is either too expensive or it is too big to be stored in the computer memory, the Partial ACA algorithm can be applied, see [17] for more details. For this algorithm, the matrix is approximated with crosses formed by rows and columns, which are computed successively. Starting by generating an arbitrary chosen row, the corresponding column is defined by the index of the largest absolute value in the row. This cross is identified by the index pair. All further pivot positions, i.e. index pairs, are defined on the crosses of the residuum R_{i+1}. Here, the positions are selected where the values are maximal. Thus, only a few rows and columns of the matrix are generated and the Partial ACA algorithm is very fast. The main problem is to define an effective stopping strategy. A natural generalisation of the criterion (14.7) is

$$\frac{\|S_{i+1} - S_i\|_F}{\|S_{i+1}\|_F} \leq \varepsilon_{\mathrm{ACA}}, \tag{14.8}$$

which works fine in most cases. But there are some exceptions. If the generated rows or columns are linearly dependent, then the residuum will be exactly zero and the algorithm should be restarted by an arbitrary not yet generated row. The algorithm cannot be stopped directly when the criterion (14.8) is fulfilled for the first time. We have to check all the already computed information of the matrix to guarantee a reliable stopping criterion.

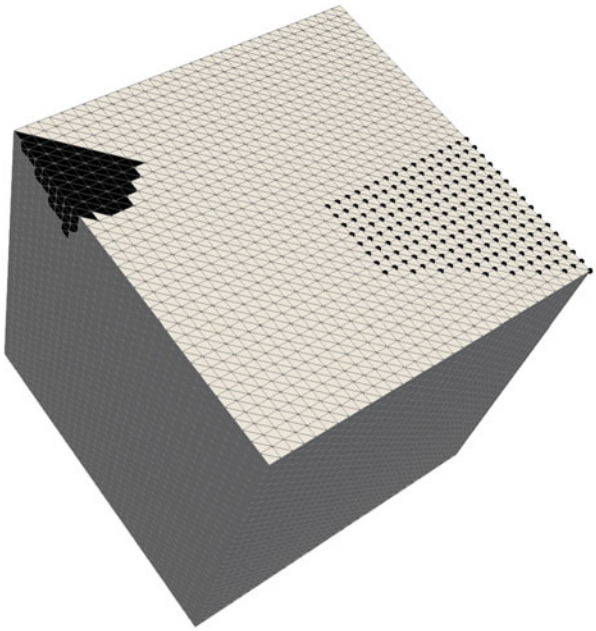


Fig. 14.2 A critical admissible cluster pair with $N = 114$ and $M = 189$

A stable practical implementation of the Partial ACA algorithm is a highly non-trivial task due to many possible exceptional situations. We try to illustrate this by a simple and well known example leading to some of the possible difficulties, cf., e.g., [2, 7]. We consider the unit cube $\Omega = (-1/2, 1/2)^3$, its uniform surface discretisation into triangles with $N_{\mathrm{BEM}} = 12{,}288$ and $M_{\mathrm{BEM}} = 6146$ and generate the double layer potential matrix K_h in the hierarchical form. This matrix contains several zero blocks, since the kernel of the double layer potential vanishes if $x - y$ is orthogonal to n_y. The surface elements and nodes are divided in two systems of clusters with 4107 and 2063 clusters, correspondingly. A system of cluster pairs is constructed leading to 93,149 pairs. Corresponding to the criterion

$$\min(\mathrm{diam}Cl_1, \mathrm{diam}Cl_2) \leq \eta \mathrm{dist}(Cl_1, Cl_2)$$

with $\eta = 0.8$, the pairs are marked as admissible. One of the admissible clusters is shown in Fig. 14.2. The convergence behaviour of the Full ACA algorithm is shown in Fig. 14.3, where a $\log_{10}$-plot of the error (14.7) is presented. It is almost monotone and any accuracy can be reached. However, the convergence is not very fast. The Partial ACA algorithm from [1] (see also [17] for a more detailed description) behaves at first sight similarly but there is a big difference. In Fig. 14.4 the error estimate (14.8) is shown together with the exact relative error in the Frobenius norm. Two times in steps 5–9 and 20–24 the error estimate (14.8) decreases very

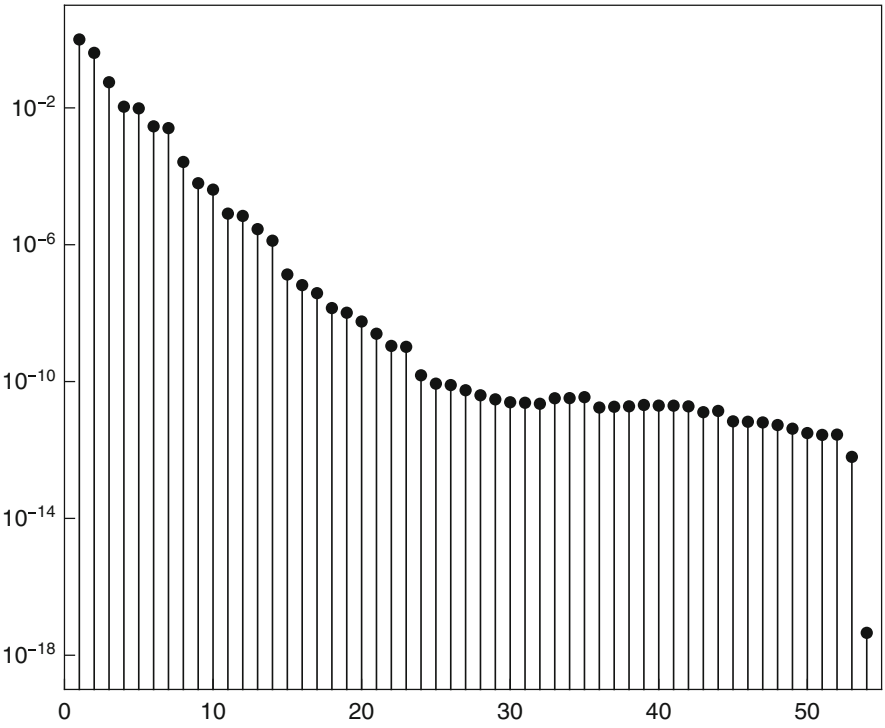


Fig. 14.3 Relative error in the Frobenius norm for the full ACA

fast indicating accuracy better than 10^{-6} after 9 steps and better than 10^{-10} after 24 steps. The real error in contrast stagnates during these steps and it is only of the order 10^{-2} or 10^{-8}. In both cases the estimate (14.8) jumps afterwards to the correct value. A more detailed explanation of this effect is as follows. The partially pivoted ACA algorithm has chosen a row or column with a lot of zeros. Consequently, $\|S_{i+1} - S_i\|$ is very small although the true error is larger. Furthermore, the strategy cannot choose an appropriate pivot element on the cross due to the dominating number of zero entries and since there is no information for the selection. This exceptional case has to be detected and the algorithm should be restarted. However, also the Partial ACA algorithm reaches an arbitrary accuracy if it is not stopped due to error estimate (14.8) too early.

The main reasons for such exceptional cases as, e.g., described above are of course the plane geometry leading to a lot of exact zeros in the matrix, the geometric edges of the domain and, in particular, the clusters containing these edges. We cannot prevent edges and plane parts on realistic surfaces, but we can prevent clusters of elements containing such edges. Consequently, we avoid the appearance of critical matrix blocks for the criterion (14.8). Whereas the clustering of nodes is done as usual, we explain a clustering for the triangular elements that respects

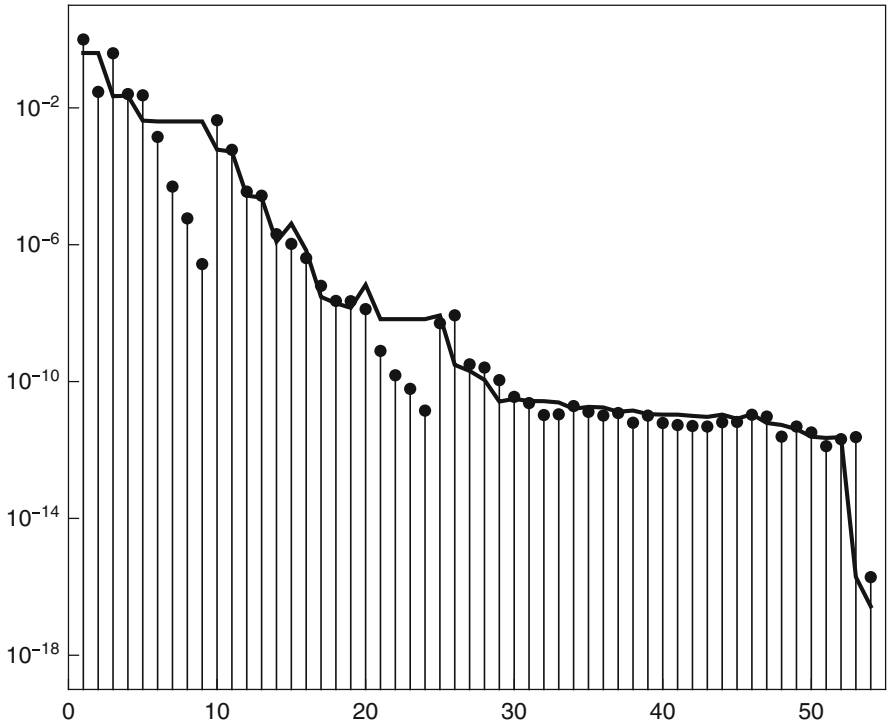


Fig. 14.4 Relative error in Frobenius norm for the partial ACA

geometric edges in the next section. After the use of this surface segmentation, the most simple, original, partially pivoted ACA performs very well, see Sect. 14.5. Thus, this strategy is designed for approximation spaces in the BEM that are spanned by element basis functions such as piecewise polynomials that are usually used for the discretisation of $H^{-1/2}(\Gamma)$, for instance.

14.4 Surface Segmentation

Instead of clustering the triangles directly, the idea is to perform beforehand a preprocessing step in which the surface is decomposed into several segments that do not contain any edges of the geometry. This surface segmentation guarantees that adjacent triangles of the same segment do not have a large dihedral angle[1] and thus one of the critical situations for the Partial ACA algorithm is prevented.

[1] Angle between the normal vectors to the triangles, see [8].

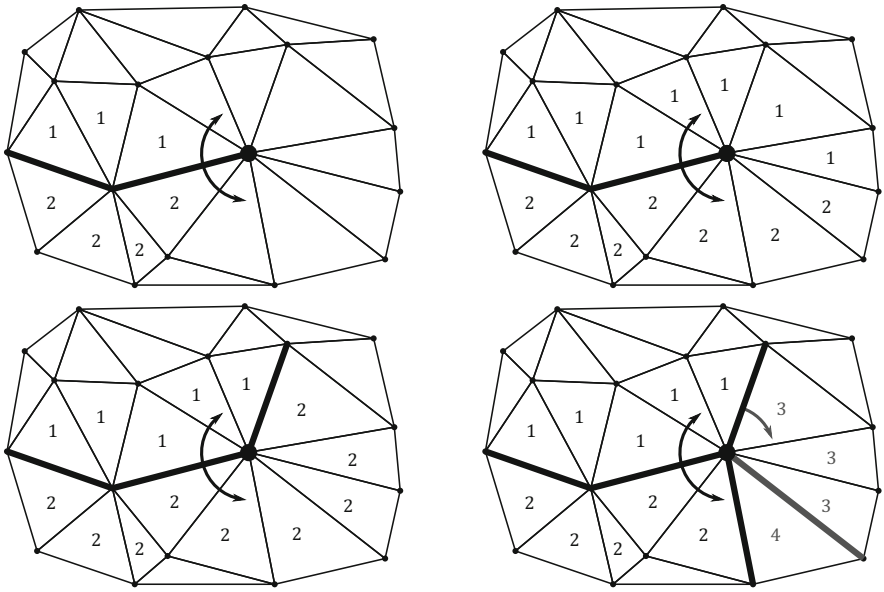


Fig. 14.5 Part of surface triangulation with detected geometric edges in bold and assigned segment numbers to the triangles; the Algorithm 14.2 started at the left bold edge and evolves to the right; initial state of Algorithm 14.3 (top left) and three possible cases with no, one and multiple edges to follow

After this preliminary segmentation, the clustering can be built on top of the initial decomposition of the surface triangles.

In the following, this procedure is described in more details. Here, we have to take care on various types of geometric edges which may occur. We may have edges which are open or closed curves on the surface, but they may also branch as seen for the cube or in Fig. 14.5, for instance. The surface mesh is usually represented as a set of triangles described by vertices. But in the forthcoming algorithms we additionally need the edges of the triangles and the neighbourhood relations of the geometric objects. This information can be constructed in linear complexity from the initial mesh data, and it is stored in dynamic data structures in our implementation. Many mesh libraries, however, already support the needed functionality.

In order to perform the surface decomposition, each triangle is assigned to a set of elements belonging to the same segment. The borders of these segments are formed by edges of triangles. There are two kinds of such borders. They are either located aligned with edges of the geometry or between two segments meeting on a flat part of the surface. The aim of the forthcoming algorithm is to divide the surface triangulation of the geometry into segments such that the edges in the geometry are resolved and each segment consists of a smooth part of the surface only. For this reason, we proceed as follows: We seek the geometric edges in the surface, follow them and assign the adjacent triangles of each side a number which defines the

segment. Afterwards, we complete the segments in a uniform way. Consequently, we define the borders of the segments along the edges of the geometry explicitly and give a fully automatic procedure to specify the borders within the smooth part of the surface.

We present the three essential steps in an algorithmic fashion, which are given, however, in a more verbal form. The full technical details would spoil the presentation, therefore we refer the interested reader for more insights and in order to test the algorithms to our implementation that is available on github.com.[2] Let $T_{\texttt{dihed}}$ be a user specified threshold parameter. All edges for which the dihedral angle α of the adjacent triangles exceeds this threshold, i.e. $\alpha > T_{\texttt{dihed}}$, are assumed to be geometric edges. The implementation of these algorithms makes extensive use of data structures realising dynamic lists.

Algorithm 14.2 (Detect Geometric Edges)

```
1. For all edges E
 1.1. if E not yet processed
  1.1.1. mark E as processed
  1.1.2. if E is geometric edge
      1.1.2.1. assign adjacent triangles new segment
               numbers
      1.1.2.2. follow E to the left (→ Alg. 3)
      1.1.2.3. follow E to the right (→ Alg. 3)
```

The main routine is Algorithm 14.2 followed by the forthcoming Algorithm 14.4. Here, Algorithm 14.2 makes use of Algorithm 14.3 in `1.1.2.2` and `1.1.2.3`, which is a function call of a recursive procedure in the implementation. In step `1.1.2.1`, we assign new segment numbers to the triangles. In most cases these triangles have no number yet. In some exceptional cases they might be already assigned, but then we just overwrite the number.

Algorithm 14.3 detects the whole geometric edge starting from a single edge in the discretisation. Therefore, it has to distinguish between the different possible cases while following an edge. This is illustrated in Fig. 14.5, where the detected geometric edges are marked in bold. The Algorithm 14.3 has started at the left bold edge in the figure and evolves to the right. Now (top left), the edges next to the bold node have to be checked. Here, the geometric edge might end (top right), it might continue in one direction (bottom left) or it might branch (bottom right).

[2]Rjasanow, S. and Weißer, S.: Surface Segmentation (Version 1.0). Saarland University, Saarbrücken, Germany (2018). See https://github.com/s-weisser/surface-segmentation.

Algorithm 14.3 (Follow Edge E to the Left/Right)

1. Let $E_\mathrm{f} = E_\mathrm{b} = E$ and let L be the list of edges adjacent to the left/right node of E in clockwise order
2. Until E_f ($\neq E$) is geom. edge and E_b ($\neq E$) is geom. edge or $E_\mathrm{f} = E_\mathrm{b}$ ($\neq E$)
   - 2.1. if E_f is not geom. edge
     - 2.1.1. set E_f as next edge in L from the front
     - 2.1.2. assign segment number to triangle enclosed by E_f and the previous E_f if not already assigned
   - 2.2. if E_b is not geom. edge
     - 2.2.1. set E_b as next edge in L from the back
     - 2.2.2. assign segment number to triangle enclosed by E_b and the previous E_b if not already assigned
3. If $E_\mathrm{f} \neq E_\mathrm{b}$ (Fig. 14.5 bottom right black)
   - 3.1. iterate over remaining edges E' in L starting from E_f
     - 3.1.1. check if E' is geom. edge
     - 3.1.2. assign segment numbers to triangles enclosed by E' and the previous E' if not already assigned (Fig. 14.5 bottom right grey)
   - 3.2. mark geom. edges E_f, E_b and those E' of 3.1.1. as processed and follow them recursively in the order they were detected if they have not been already processed (→ Alg. 3)
4. Else if $E_\mathrm{f} = E_\mathrm{b}$ is geom. edge (Fig. 14.5 bottom left)
   - 4.1. mark E_f as processed and follow E_f if it has not been already processed (→ Alg. 3)
5. Else ($E_\mathrm{f} = E_\mathrm{b}$ is no geom. edge, Fig. 14.5 top right)
   - 5.1. return

The algorithm assigns the segment numbers such that connected triangles which belong to the same smooth part of the surface, i.e., which lie on the same side of the geometric edge, have the same number. Thus, the segment numbers of the previous recursion step have to be advanced along the detected geometric edge. Also new segment numbers may occur in case that the edge branches, cf. Fig. 14.5 (bottom right). After the execution of Algorithm 14.2, all triangles adjacent to geometric

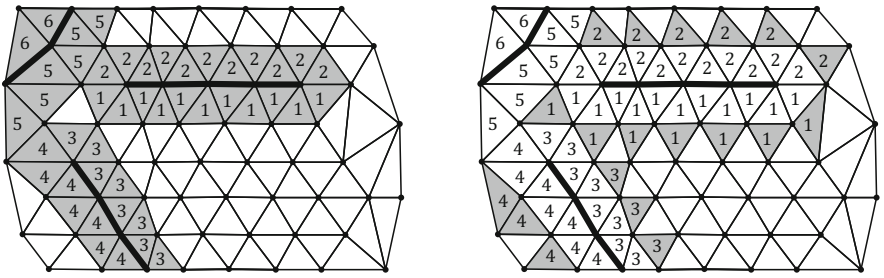


Fig. 14.6 Part of surface triangulation with geometric edges in bold and assigned segment numbers to the triangles; initial state (left) and after one run of the loop (right) of the Algorithm 14.4, only the neighbourhoods of the shaded triangles are checked

edges are assigned to segments, but there are also triangles which have not been assigned yet, see Fig. 14.6 (left). This is the initialisation of Algorithm 14.4, which completes the segments. In order to make the algorithm efficient, we only loop over segments and triangles that might influence the procedure, see shaded triangles in Fig. 14.6. More precisely, we only consider segments which still have the possibility to be enlarged and which are not already surrounded by borders of other segments. This can be seen in Fig. 14.6 for the segment numbers 5 and 6 in the right sketch. Furthermore, we only consider triangles within the segments which have been added in the previous step, cf. Fig. 14.6.

Algorithm 14.4 (Complete Segments)

1. For all segments S which might be enlarged

 1.1. for all triangles τ_k in S added in the previous step (or given in the initialisation, respectively)

  1.1.1. for all neighbouring triangles τ_ℓ of τ_k

   1.1.1.1. assign τ_ℓ to S if τ_ℓ is not already assigned

  1.1.2. if no triangle has been added to S, then S cannot be enlarged further

The Algorithm 14.2 involving Algorithm 14.3 has linear complexity, since each edge is processed only once due to the marking. The recursion depth is uniformly bounded because of the regularity of the closed surface mesh in the sense of Ciarlet and since marked edges are not processed twice. Arguing with the regularity of the mesh once more, we see that Algorithm 14.4 has also linear complexity.

Fig. 14.7 A complicated surface with open geometric edges

In Fig. 14.7, we give an example of the surface segmentation for a non-trivial domain. The Algorithm 14.2 has been performed with $T_{\texttt{dihed}} = \pi/3$ and it has produced an output of 20 segments. It is possible to run an additional post-processing that reduces the number of segments by gluing those, which share boundaries on flat parts of the geometry only. This option has been implemented in the example code available on github.com. For this example, the number of segments is reduce from 20 to 9.

14.5 Numerical Examples

In this section, we illustrate the ACA approximation, the accuracy as well as the convergence of the BEM for a series of uniform discretisations of the cube [4] $\Omega = (-1/2, +1/2)^3$. The surface mesh for $N_{\texttt{BEM}} = 3072$ and $M_{\texttt{BEM}} = 1538$ is shown in Fig. 14.1. The analytic solution of the problem is

$$u(\boldsymbol{x}) = \frac{1}{4\pi\,|\boldsymbol{x} - \boldsymbol{x}^*|}, \quad \text{for } \boldsymbol{x} \in \Omega$$

Table 14.1 ACA approximation of the Galerkin matrices K_h and V_h

N_{BEM}	M_{BEM}	$\varepsilon_{\mathrm{ACA}}$	MByte (K_h)	%	MByte (V_h)	%
192	98	$1.0 \cdot 10^{-4}$	0.14	100.0	0.28	51.5
768	386	$1.0 \cdot 10^{-5}$	2.21	97.8	2.06	45.8
3 072	1 538	$1.0 \cdot 10^{-6}$	27.70	76.8	19.40	26.9
12 288	6 146	$1.0 \cdot 10^{-7}$	239.71	40.1	150.50	13.1
49 152	24 578	$1.0 \cdot 10^{-9}$	1 607.06	17.4	993.80	5.4
196 608	98 306	$1.0 \cdot 10^{-10}$	20 459.62	13.9	7 043.02	2.4

Table 14.2 Accuracy of the BEM

N_{BEM}	M_{BEM}	*Iter*	$ND - Error$	CF_1	$DD - Error$	CF_2
192	98	46	$7.52 \cdot 10^{-2}$	–	$1.58 \cdot 10^{-3}$	–
768	386	71	$3.02 \cdot 10^{-2}$	2.49	$3.55 \cdot 10^{-4}$	4.46
3 072	1 538	95	$1.27 \cdot 10^{-2}$	2.38	$8.34 \cdot 10^{-5}$	4.26
12 288	6 146	129	$5.60 \cdot 10^{-3}$	2.27	$2.01 \cdot 10^{-5}$	4.15
49 152	24 578	175	$2.49 \cdot 10^{-3}$	2.25	$4.94 \cdot 10^{-6}$	4.07
196 608	98 306	237	$1.22 \cdot 10^{-3}$	2.04	$1.22 \cdot 10^{-6}$	4.05

with $\boldsymbol{x}^* = (2, 0, 0)^\top$. Thus, the function u is harmonic in Ω and its trace $g(\boldsymbol{x}) = \gamma_0 u(\boldsymbol{x})$ will be used as the Dirichlet boundary condition in (14.1). In Table 14.1, we show the results of the ACA approximation of both BEM matrices K_h and V_h in MBytes and in percentage of the memory required by full matrices. As usual, the approximation of the single layer potential matrix V_h is significantly better. Note that the ACA accuracy $\varepsilon_{\mathrm{ACA}}$ was increased by a factor 10 for each mesh refinement except in the fourth step, where an increase by a factor 100 was necessary to guarantee the cubic point-wise convergence of the gradient of the numerical solution, see Fig. 14.9. In the next Table 14.2, the relative L_2-error of the computed Neumann datum as well as the accuracy of the L_2-projection of the Dirichlet datum are presented. The expected linear and quadratic convergence can be seen. Furthermore, in the third column, the number of CG iterations for the iterative solution of the linear system with the matrix V_h up to the accuracy $\varepsilon_{\mathrm{CG}} = 10^{-15}$ is shown. As usual, the computing time required by solving the linear system is negligible. Finally, in Figs. 14.8 and 14.9, we show the convergence history for the Neumann (left) and Dirichlet data (right) as well as the point-wise accuracy of the solution (left) and of its gradient (right) in the centre of the cube obtained by the representation formula (14.6). The point-wise convergence, which is as expected cubic, can be clearly seen. Note that the plots in Figs. 14.8 and 14.9

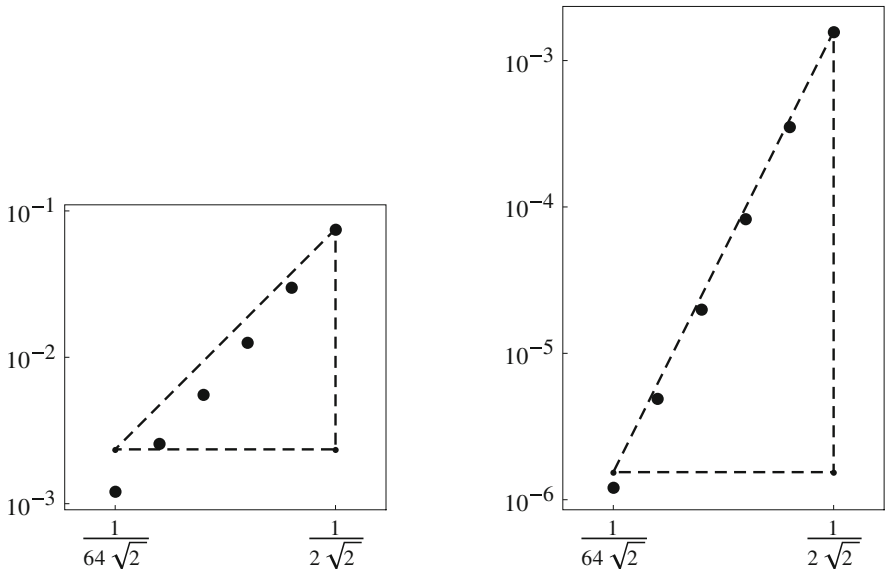


Fig. 14.8 Accuracy of the BEM, Neumann (left) and Dirichlet (right) datum

are double logarithmic and the horizontal axis shows the values of $\log_{10} h$, where h changes from $\sqrt{2}/4$ to $\sqrt{2}/128$ in five refinement steps. The dashed triangles in these plots represent the ideal linear, quadratic and cubic convergence, respectively, while the thick dots indicate the real numerical values of the error. However, without the segmentation of the surface and by the use of the simple stopping criterion (14.8), the partially pivoted ACA algorithm stops too early several times during the approximation of the double layer potential. This happens starting at the dimension $N_{\mathtt{BEM}} = 12\,288$ and $M_{\mathtt{BEM}} = 6\,146$ as reported in Sect. 14.3. The result is the loss of cubic convergence for the last three steps of the refinement since the error is dominated here by the matrix approximation. For example, the point error for $N_{\mathtt{BEM}} = 49\,152$ and $M_{\mathtt{BEM}} = 24\,578$ is only $8.45 \cdot 10^{-10}$ instead of $2.20 \cdot 10^{-10}$, which is obtained with the segmentation of the surface and gives the correct order of convergence, see Fig. 14.9.

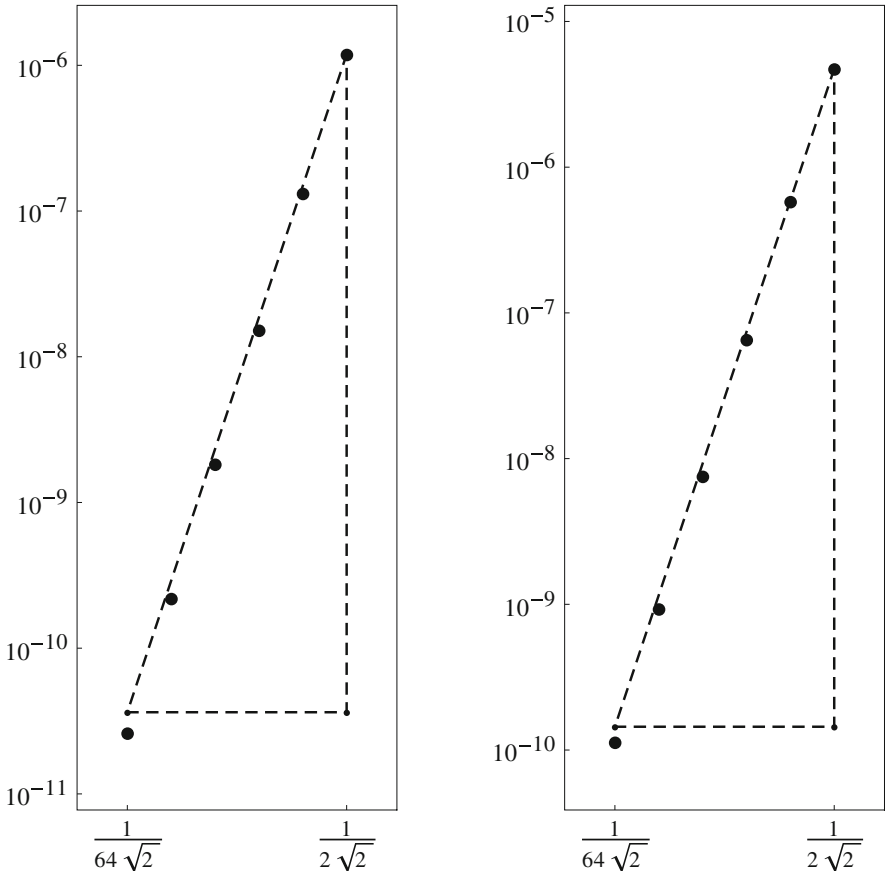


Fig. 14.9 Accuracy of the BEM solution (left) and of its gradient (right)

References

1. Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. **86**(4), 565–589 (2000)
2. Bebendorf, M.: Hierarchical Matrices. Lecture Notes in Computational Science and Engineering, vol. 63. Springer-Verlag, Berlin (2008). A means to efficiently solve elliptic boundary value problems
3. Bebendorf, M., Venn, R.: Constructing nested bases approximations from the entries of non-local operators. Numer. Math. **121**(4), 609–635 (2012)
4. Bebendorf, M., Kühnemund, A., Rjasanow, S.: An equi-directional generalization of adaptive cross approximation for higher-order tensors. Appl. Numer. Math. **74**, 1–16 (2013)
5. Bebendorf, M., Kuske, C., Venn, R.: Wideband nested cross approximation for Helmholtz problems. Numer. Math. **130**(1), 1–34 (2015)
6. Börm, S., Christophersen, S.: Approximation of integral operators by Green quadrature and nested cross approximation. Numer. Math. **133**(3), 409–442 (2016)

7. Börm, S., Grasedyck, L.: Hybrid cross approximation of integral operators. Numer. Math. **101**(2), 221–249 (2005)
8. Gellert, W., Gottwald, S., Hellwich, M., Kästner, H., Küstner, H. (eds.) The VNR Concise Encyclopedia of Mathematics, 2nd edn. Van Nostrand Reinhold, New York (1989)
9. Grande, J., Reusken, A.: A higher order finite element method for partial differential equations on surfaces. SIAM J. Numer. Anal. **54**(1), 388–414 (2016)
10. Grasedyck, L.: Adaptive recompression of $\mathscr{H}$-matrices for BEM. Computing **74**(3), 205–223 (2005)
11. Hackbusch, W.: A sparse matrix arithmetic based on $\mathscr{H}$-matrices. I. Introduction to $\mathscr{H}$-matrices. Computing **62**(2), 89–108 (1999)
12. Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis. Springer Series in Computational Mathematics, vol. 49. Springer, Berlin (2015)
13. Kravcenko, M., Maly, L., Merta, M., Zapletal, J.: Parallel assembly of ACA BEM matrices on Xeon Phi clusters. In: Parallel Processing and Applied Mathematics. Part I, Lecture Notes in Computer Science, vol. 10777, pp. 101–110. Springer, Cham (2018)
14. Mangan, A., Whitaker, R.: Partitioning 3d surface meshes using watershed segmentation. IEEE Trans. Vis. Comput. Graph. **5**(4), 308–321 (1999)
15. McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)
16. Nédélec, J.C.: Curved finite element methods for the solution of singular integral equations on surfaces in R^3. Comput. Methods Appl. Mech. Eng. **8**(1), 61–80 (1976)
17. Rjasanow, S., Steinbach, O.: The Fast Solution of Boundary Integral Equations. Mathematical and Analytical Techniques with Applications to Engineering. Springer, New York (2007)
18. Rjasanow, S., Weggler, L.: ACA accelerated high order BEM for Maxwell problems. Comput. Mech. **51**(4), 431–441 (2013)
19. Rjasanow, S., Weggler, L.: Matrix valued adaptive cross approximation. Math. Methods Appl. Sci. **40**(7), 2522–2531 (2017)
20. Sauter, S., Schwab, C.: Boundary Element Methods. Springer Series in Computational Mathematics. Springer, Berlin (2010)
21. Steinbach, O.: Numerical Approximation Methods for Elliptic Boundary Value Problems. Springer, New York (2008)
22. Vieira, M., Shimada, K.: Surface mesh segmentation and smooth surface extraction through region growing. Comput. Aided Geom. Des. **22**(8), 771–792 (2005)
23. Yamauchi, H., Lee, S., Lee, Y., Ohtake, Y., Belyaev, A., Seidel, H.P.: Feature sensitive mesh segmentation with mean shift. In: Proceedings of the International Conference on Shape Modeling and Applications 2005, pp. 236–243. Cambridge, USA (2005)

Chapter 15
First Order Error Correction for Trimmed Quadrature in Isogeometric Analysis

Felix Scholz, Angelos Mantzaflaris, and Bert Jüttler

Abstract In this work, we develop a specialized quadrature rule for trimmed domains, where the trimming curve is given implicitly by a real-valued function on the whole domain. We follow an error correction approach: In a first step, we obtain an adaptive subdivision of the domain in such a way that each cell falls in a predefined base case. We then extend the classical approach of linear approximation of the trimming curve by adding an error correction term based on a Taylor expansion of the blending between the linearized implicit trimming curve and the original one. This approach leads to an *accurate* method which improves the convergence of the quadrature error by one order compared to piecewise linear approximation of the trimming curve. It is at the same time *efficient*, since essentially the computation of one extra one-dimensional integral on each trimmed cell is required. Finally, the method is *easy to implement*, since it only involves one additional line integral and refrains from any point inversion or optimization operations. The convergence is analyzed theoretically and numerical experiments confirm that the accuracy is improved without compromising the computational complexity.

15.1 Introduction

A common representation of a Computer-Aided Design (CAD) model is a boundary representation (B-rep), which typically consists of trimmed tensor-product NURBS patches. A trimmed surface patch consists of a tensor-product surface and a set of

F. Scholz (✉)
Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Linz, Austria
e-mail: felix.scholz@ricam.oeaw.ac.at

A. Mantzaflaris · B. Jüttler
Institute of Applied Geometry, Johannes Kepler University Linz, Linz, Austria
e-mail: angelos.mantzaflaris@inria.fr; bert.juettler@jku.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_15

trimming curves on the surface that represent the boundary of the actual surface. Therefore, it represents only a part of the full tensor-product surface yet no explicit parametric representation is available. In this paper we are interested in applying numerical integration on a trimmed surface patch.

Computing integrals over trimmed domains both *efficiently* and *accurately* remains a challenging problem, notably for use in the frame of isogeometric analysis (IgA) [11]. The latter computational framework aims at a unification of the representations used in CAD and in numerical simulation, therefore operating directly on trimmed patches. The reader is referred to [18] for a recent review on trimming in CAD and IgA. We remark that different CAD representations are possible, e.g. subdivision surface-based models or T-spline models, see [1, 2, 12].

Multiple challenges arise for IgA on trimmed domains. One issue is the efficient coupling between adjacent trimmed patches. To this end, a finite cell method with weak coupling has been proposed in [24], a tearing and interconnecting approach was recently studied in [32] as well as Discontinuous Galerkin (DG) methods [9, 10, 31]. Another issue is the numerical stability of the trimmed basis functions, since basis functions with a tiny support can appear around the trimmed boundary. Several modified bases have been considered, such as immersed B-splines [25] and extended B-splines [19] to overcome the issue. Finally, the problem of applying numerical quadrature on a trimmed patch is a challenge in its own right.

One first approach to integrating over trimmed surfaces is to place quadrature points on the full surface and set the weights of the points lying outside the trimmed domain to zero. However, this method has no guarantees and the integration error cannot be controlled easily. In engineering practice local adaptive quadrature is used on top of it, that is, subdivision is performed around the trimmed region and quadrature nodes are placed in each sub-cell [8, 13, 28]. This approach can generate an extensive number of quadrature points, thus posing efficiency barriers.

Another approach is to perform a reparameterization (either globally, or locally at the element level), also known as "untrimming"; this puts more effort in the geometric side and results in tensor-product patches, which can be handled in an efficient way [27]. However, it is known that exact reparameterization is not feasible and approximate solutions result in cracks or overlaps in the model which require special treatment, e.g. by means of DG methods [9, 31]. In [14], the authors use base cases for the trimmed elements and perform local untrimming, using the intersection points of the trimmed curve and the boundary; see also [29] for some applications of this machinery in optimization. In [26] a local untrimming on the element level is performed by a projection method, which can be interpolation or least-squares fitting.

When the trimming curves are complicated and have high degree, it is typical to compute a piecewise linear approximation of the boundary to simplify further processing [1, 3, 15, 23]. However, when it comes to numerical integration, the geometry approximation error accumulates in the final result, and deteriorates the overall approximation order. Alternatively, in [20] the linearization is avoided, and a quadrature rule is constructed for each trimmed element by solving a moment-

fitting, non-linear system to obtain quadrature nodes and weights, which are all contained inside the domain.

The problem of integrating over trimmed domains also arises in the context of geometrically unfitted finite element methods [4]. In these methods, the solution of a PDE is approximated on a background finite element mesh that is cut by the boundary of the computational domain. A method for quadrature in this context is presented in [16]. The author uses piecewise linear approximation of the boundary and applies a transformation of the finite element mesh in order to improve the approximation order. For the case of Cut Finite Element Methods (CutFEM, [5]), a method for numerical integration using boundary integrals and Taylor approximation in a Nitsche formulation was developed in [6].

In the present work we develop an *efficient and accurate* quadrature rule for the approximation of integrals over trimmed domains. The trimming curve is given implicitly by a real-valued level function on the whole domain. This does not impose any restrictions, since any trimming curve can be converted into this format by employing implicitization techniques, which can be either exact or approximate ones, see, e.g., [7, 30]. Our method is based on an error correction approach. In a first step, we obtain an adaptive subdivision of the domain in such a way that each cell falls in a predefined base case. We then extend the classical approach of linear approximation of the trimming curve by adding an error correction term based on a Taylor expansion of the blending between the linearized implicit trimming curve and the original one.

In terms of *accuracy*, the method improves the local quadrature error in each cell by two orders of magnitude compared to the piecewise linear approximation of the trimming curve, thus providing an extra order of convergence globally. In particular, cubic order of convergence is achieved with a negligible additional computational cost. The *efficiency* of the method is achieved by the fact that it requires solely the evaluation of the trimming function at the vertices of the cells and the quadrature nodes, and refrains from any kind of point inversion or non-linear solving. Furthermore, our method is *easy to implement*, since the resulting nodes for the correction term are simply one-dimensional Gauss nodes and their corresponding weights are given by a direct computation. Moreover, we do not need to test for quadrature points outside the integration domain or treat them in a different way. Overall, it is straight-forward to upgrade existing codes to incorporate our method.

In the next section, we state the problem of trimmed quadrature with a implicitly defined trimming curve. In Sect. 15.3 we explain the first step of the method, the subdivision of the domain into quadrature cells belonging to certain base cases. Section 15.4 deals with the piecewise linear approximation of the trimming function which is then extended by the first order error correction in Sect. 15.5. We analyze the convergence behavior of our method theoretically and experimentally in Sects. 15.6 and 15.7, respectively.

15.2 Problem Formulation

Throughout this paper, we consider integrals of bivariate functions on trimmed domains. More precisely, we assume that a sufficiently smooth function

$$f : [0, 1]^2 \to \mathbb{R} \tag{15.1}$$

is given, which is defined on the entire unit square. In addition, we restrict the unit square by trimming with an implicitly defined curve $\tau(x, y) = 0$, which is defined by another smooth bivariate function

$$\tau : [0, 1]^2 \to \mathbb{R}. \tag{15.2}$$

This results in the trimmed domain

$$\Omega_\tau = \{(x, y) \in [0, 1]^2 : \tau(x, y) \geq 0\}. \tag{15.3}$$

We seek for quadrature rules that provide approximate values of the integral

$$I_\tau f = \int_{\Omega_\tau} f(x, y) \mathrm{d}y\mathrm{d}x. \tag{15.4}$$

These quadrature rules shall take the form

$$Q_\tau f = \sum_i w_i f(x_i, y_i) \tag{15.5}$$

with a finitely many quadrature nodes (x_i, y_i) and associated weights w_i. Two comments about this problem are in order:

1. As described in the introduction, this problem originates in isogeometric analysis, where one needs to solve it in order to perform isogeometric discretizations of partial differential equations on trimmed patches. Typically, the function f then takes the form

$$f(x, y) = DB_j(x, y)\bar{D}B_k(x, y)K(x, y) \tag{15.6}$$

where the functions B_i are bivariate tensor-product B-splines or polynomial segments thereof and the kernel K reflects the influence of the geometry mapping (i.e., the parameterization of the computational domain by a NURBS surface) and the coefficient functions of the PDE. In many situations, this results in a piecewise rational function f.

2. Usually, the trimming functions in CAD and not given implicitly but by low degree parametric curves. It is then possible to convert these curves into implicit form by invoking suitable implicitization techniques, which can be either exact or approximate ones, see e.g. [7, 30]

Our approach to finding a quadrature rule consists of two steps. First we subdivide the domain to reach a certain discretization size, while simultaneously ensuring that we arrive at a sufficiently simple configuration on each cell. Second we evaluate the contribution of each cell to the total value of the integral. These steps will be discussed in the next three sections.

15.3 Adaptive Subdivision of the Domain

Given a step size h, we subdivide the domain uniformly until the cell size does not exceed h. Subsequently, we perform adaptive subdivision until each resulting cell K is an instance of one of the five base cases depicted in Fig. 15.1. The resulting set of quadrature cells will be denoted by $\mathscr{K}$.

More precisely, we use only evaluations of the trimming function at the cell vertices to identify the base cases, similar to the marching cubes algorithm [21]. Consequently, the method does not detect branches of the trimming curve that leave and re-enter the cell within the same edge. In order to illustrate this fact, Fig. 15.2 shows two instances of each trimmed base case.

Zero values at the vertices are treated as positive numbers. Consequently, there are only two sign distributions that do not represent a base case, see Fig. 15.3. These situations are dealt with by uniformly subdividing the corresponding cell. This process is guaranteed to terminate if no singularities of the trimming curve are present (which is always the case in practice).

We summarize our adaptive subdivision approach to the generation of quadrature cells $\mathscr{K}$ in the following algorithm.

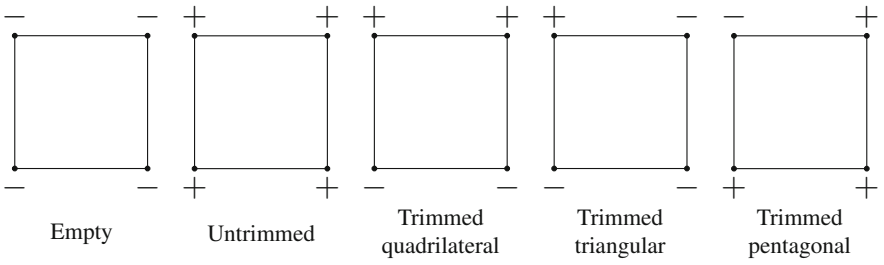


Fig. 15.1 Sign distributions of the trimming function τ for all five base cases (up to rotations)

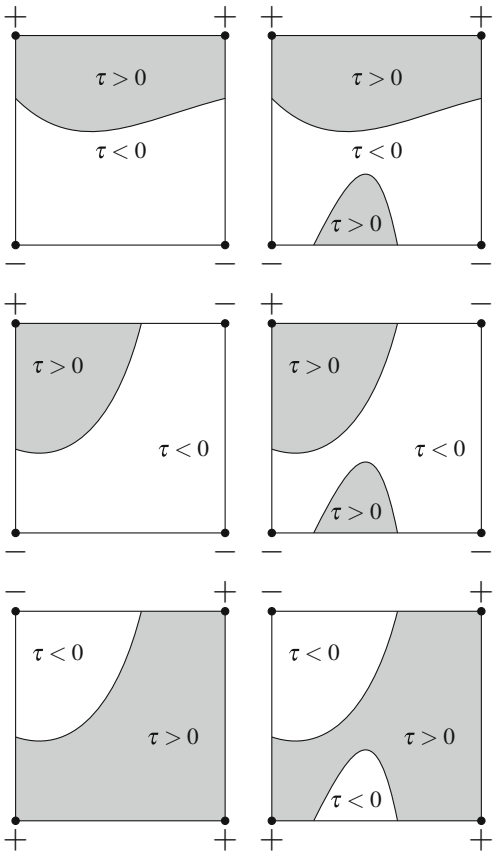


Fig. 15.2 Two instances of each trimmed base case (curved quadrilateral, triangle and pentagon)

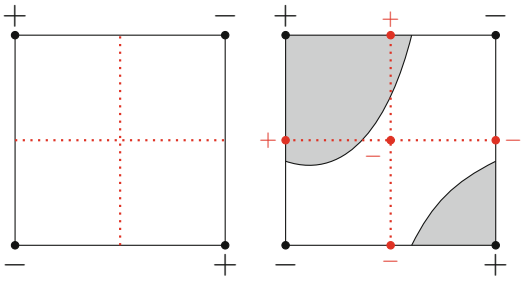


Fig. 15.3 Left: Sign distribution that does not represent a base case. Right: Uniform subdivision (quadsection) of this cell

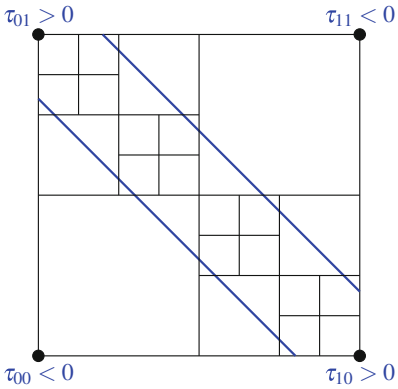


Fig. 15.4 Quadrature cells for $h = \frac{1}{2}$. The trimming curve is shown in blue

- **QuadratureCells**, input: τ, $h > 0$
- Let the initial cell be $K = [0, 1]^2$.
- Repeat for each untreated cell K:

  if $Size(K) < h$ and $BaseCase(\tau, K)$

  Report K

  else

  $K_1, \ldots, K_4 = Quadsect(K)$

  Mark $K_1, \ldots, K_4$ as untreated.

Figure 15.4 shows an instance of quadrature cells generated by the algorithm. The zero set of the trimming function consists of two parallel lines, which are parallel to one of the square's diagonals. The parameter h was chosen as $\frac{1}{2}$. For this specific instance of τ, the algorithm needs one or two additional subdivision steps at the northwest and southeast corners of the domain.

15.4 Linearized Trimmed Quadrature

We perform the quadrature individually on each cell $K \in \mathscr{K}$. Thus, we need to approximate the integral

$$\int_{K_\tau} f(x, y) \mathrm{d}y\mathrm{d}x, \tag{15.7}$$

where

$$K_\tau = \{(x, y) \in K : \tau(x, y) > 0\}. \tag{15.8}$$

This approximation is trivial for the first two base cases: The value of the integral equals zero in the first case, and it is approximated by a tensor-product Gauss rule in the second one.

In order to perform this approximation in the remaining three base cases, we replace τ by another function σ. That function is chosen such that the integral

$$\int_{K_\sigma} f(x, y)\mathrm{d}y\mathrm{d}x \tag{15.9}$$

over the region

$$K_\sigma = \{(x, y) \in K : \sigma(x, y) > 0\} \tag{15.10}$$

enclosed by the zero-level set of σ admits a simple evaluation. This is achieved by using a suitable linear approximation of τ, that is, the level set $\sigma(x, y) = 0$ is simply a straight line segment.

The evaluation of (15.9) using numerical quadrature is based on the intersections of $\sigma(x, y) = 0$ with the boundary of the cell. We determine these intersections by linear interpolation of the function values of τ at the vertices of the cell. Figure 15.5 shows examples for this approximation. For this choice of σ, the linearized integral (15.9) belongs to the same base case as the original integral (15.7).

For the quadrilateral case we proceed as follows: First, we construct a bilinear parameterization of K_σ. Second, we transform the integral to the associated parameter domain and evaluate its value using Gauss quadrature with n evaluations per parametric direction, where n is chosen by the user. The triangular case is dealt with analogously, by using a parameterization with a singularity at the involved cell vertex. In the pentagonal case, the identity

$$K_\sigma = K \setminus K_{-\sigma} \tag{15.11}$$

allows to evaluate (15.9) by combining results for the untrimmed and the triangular case.

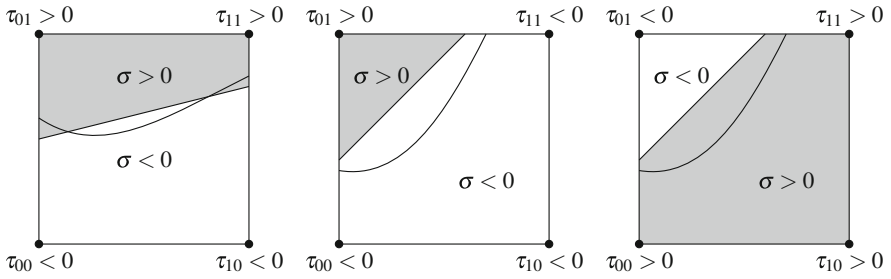


Fig. 15.5 A simple approximation of the three trimmed base cases

The resulting linearized trimmed quadrature rule will be referred to as

$$\mathrm{LT}(h, n), \tag{15.12}$$

where h is the maximum cell size and n denotes the number of Gauss nodes. Clearly, the values generated by the LT rule converge to the true integral as h is decreased. As we shall see later, however, the rather rough approximation of the quadrature domain limits the order of convergence.

15.5 First Order Correction

We improve the order of convergence of the LT rule by adding an error correction term for the last three base cases. This term is found by performing a Taylor expansion.

Throughout this section, we consider a fixed trimmed cell

$$K = [x_0, x_0 + h] \times [y_0, y_0 + h] \in \mathscr{K} \tag{15.13}$$

which is either a quadrangular or triangular base case, see Fig. 15.5. Recall that the pentagonal case is solved by considering the complementary triangular domain.

Linear blending of the trimming function τ and its linear approximation σ leads to the subsets

$$K_{\sigma+u(\tau-\sigma)} = \{(x, y) \in K : \sigma(x, y) + u(\tau(x, y) - \sigma(x, y)) > 0\}, \quad u \in [0, 1], \tag{15.14}$$

that define the function

$$F(u) = \int_{K_{\sigma+u(\tau-\sigma)}} f(x, y)\mathrm{d}y\mathrm{d}x. \tag{15.15}$$

It attains the exact value of (15.4) for $u = 1$, while the LT rule is based on the approximate evaluation of $F(0)$. We improve the accuracy by adding a correction term that is based on the first two terms of the Taylor series

$$F(1) = F(0) + F'(0) + R_1(1) \tag{15.16}$$

of F around $u = 0$, where R_1 denotes the remainder.

In order to compute $F'(0)$, we observe that the level set of the function obtained by linear blending defines a function $y = c_u(x)$ or $x = c_u(y)$ for sufficiently small values of u, where the projection of the trimming curve onto the x or y axis specifies the domain

$$[a_u, b_u], \tag{15.17}$$

respectively. If both choices are possible, we choose the one with the larger domain for $u = 0$.

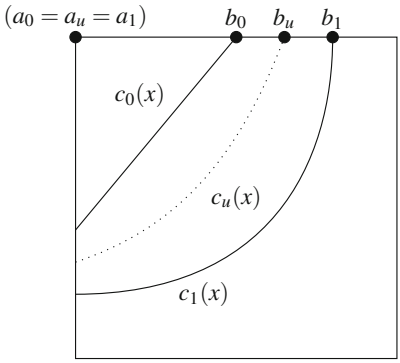


Fig. 15.6 Linear blending of τ and σ in the triangular base case

Without loss of generality, we consider the first case

$$[a_u, b_u] \subset [x_0, x_0 + h] \tag{15.18}$$

where the function satisfies

$$\sigma(x, c_u(x)) + u(\tau(x, c_u(x)) - \sigma(x, c_u(x))) = 0, \tag{15.19}$$

see Fig. 15.6. By differentiating (15.19) we observe that

$$\frac{\partial}{\partial u} c_u(x) = \frac{-\tau(x, c_u(x)) + \sigma(x, c_u(x))}{\frac{\partial \sigma}{\partial y}(x, c_u(x)) + u(\frac{\partial \tau}{\partial y}(x, c_u(x)) - \frac{\partial \sigma}{\partial y}(x, c_u(x)))}. \tag{15.20}$$

The function F in (15.15) can be rewritten as

$$F(u) = \int_{a_u}^{b_u} \int_{c_u(x)}^{y_0+h} f(x, y) \mathrm{d}y \mathrm{d}x. \tag{15.21}$$

Its first derivative thus evaluates to

$$\begin{aligned} F'(u) \; = \; & -\int_{a_u}^{b_u} f(x, c_u(x)) \frac{\partial}{\partial u} c_u(x) \mathrm{d}x \\ & -\frac{\mathrm{d}}{\mathrm{d}u} a_u \int_{c_u(a_u)}^{y_0+h} f(a_u, y) \mathrm{d}y + \frac{\mathrm{d}}{\mathrm{d}u} b_u \int_{c_u(b_u)}^{y_0+h} f(b_u, y) \mathrm{d}y \end{aligned} \tag{15.22}$$

The integration limit satisfies $b_u = x_0 + h$ or

$$c_u(b(u)) = y_0 + h. \tag{15.23}$$

Consequently, the third term in (15.22) vanishes since either the integral or the factor in front of it take value zero. Similarly, the second term vanishes as well.

Finally we use (15.20), (15.22) and the fact that σ vanishes on the graph of c_0,

$$\sigma(x, c_0(x)) = 0, \tag{15.24}$$

to rewrite the first order correction term

$$F'(0) = \frac{1}{\frac{\partial}{\partial y}\sigma} \int_{a_0}^{b_0} f(x, c_0(x))\tau(x, c_0(x))\mathrm{d}x \tag{15.25}$$

as a univariate integral over the linearized trimming curve. An approximate value is computed using a Gauss rule with k quadrature nodes.

Note that we only used the zero level set of σ in Sect. 15.5. The value of the correction term depends on the gradient of the linearized trimming function, which we did not discuss so far. In fact, we would like to choose $\nabla\sigma$, and more specifically $\frac{\partial\sigma}{\partial y}$, so that it approximates sufficiently well $\nabla\tau$. A good choice is to set $\frac{\partial\sigma}{\partial y}$ using finite differences over the values of τ at the cell's vertices. In the quadrilateral case we use the average

$$\frac{\partial\sigma}{\partial y} = \frac{1}{2}\left(\frac{\tau_{01} - \tau_{00}}{h} + \frac{\tau_{11} - \tau_{10}}{h}\right) = \frac{\tau_{01} - \tau_{00} + \tau_{11} - \tau_{10}}{2h} \tag{15.26}$$

of the finite differences over the two edges that intersect the trimming curve. In the triangular case we use the finite difference

$$\frac{\partial\sigma}{\partial y} = \frac{\tau_{01} - \tau_{00}}{h} \tag{15.27}$$

over the edge which intersects the trimming curve.

The resulting corrected linearized trimmed quadrature rule (with first order correction term) will be referred to as

$$\mathrm{CLT}(h, n, k), \tag{15.28}$$

where h is the maximum cell size, and n resp. k are the numbers of bivariate resp. univariate quadrature nodes. In addition, we use

$$\mathrm{CLT}_K(h, n, k), \tag{15.29}$$

to denote the value contributed by an individual cell $K \in \mathscr{K}$.

If the trimming function τ is linear and thus $\sigma = \tau$, then

$$\tau(x, c_0(x)) = 0. \tag{15.30}$$

Consequently, the correction term (15.25) vanishes. In this case, CLT and LT give equivalent results.

15.6 Convergence Result

In this section we will show that the first order error correction in the CLT rule improves the convergence by one order with respect to the non-corrected LT rule. More precisely, we will prove this result for a slightly modified version of CLT, obtained by adapting the quadrature cells, which we denote as CLT*. Intuitively, we construct the modified version of CLT as follows: We look at all cells where the gradient of the trimming function is "almost" horizontal or vertical, for some point. In this case we fuse two adjacent cells in such a way that the trimmed cell becomes of the quadrilateral base case which is easier to treat theoretically than the triangular case.

First we prove two technical results about the local errors in the trimmed cells of the quadrilateral base case (Lemma 15.1) and of the triangular base case (Lemma 15.2). Second we combine them with the known approximation properties of the employed Gauss rules to estimate the global quadrature error in Theorem 15.1.

Both lemmas consider a rectangular cell (not necessarily a square) K of size h and a trimming function τ defined on it. We derive error bounds that applies to all cells of these base cases.

We say that the cell satisfies the assumptions (about the base cases) *in the strong sense* if the trimming curve crosses the boundary in exactly two points.

Lemma 15.1 *Assume that a rectangular cell K fulfills the assumptions of the quadrilateral base case in the strong sense, and the trimming function τ and its linear approximation σ satisfy the inequalities*

$$\left|\frac{\partial \tau}{\partial y}(x, y)\right| \geq C_1 \quad , \quad \forall (x, y) \in K \tag{15.31}$$

and

$$\|\sigma - \tau\|_{L^\infty(K)} \leq C_2 h^2, \tag{15.32}$$

$$\|\nabla\sigma - \nabla\tau\|_{L^\infty(K)} \leq C_3 h \tag{15.33}$$

for certain positive constants C_1, C_2, C_3. Then there exists a constant

$$C_{\text{quad}}(C_1, C_2, C_3, f) \tag{15.34}$$

which depends solely on these three constants and f, such that the corrected trimmed quadrature on this cell fulfills for $n = k = 2$

$$|I_{\tau,K} f - \text{CLT}_K(h, 2, 2) f| \leq C_{\text{quad}} h^4. \tag{15.35}$$

Proof We denote the rectangular cell by $K = [x_0, x_0 + \alpha h] \times [y_0, y_0 + \beta h]$ where (x_0, y_0) is the lower left vertex and $\alpha, \beta > 0$. Since by the monotonicity assump-

tion (15.31) the trimming curve $\tau(x, y) = 0$ can be written as a graph, the same is true for all intermediate curves if h is sufficiently small. By differentiating (15.22) and using that both $a_u = x_0$ and $b_u = x_0 + \alpha h$ are constant, we obtain, for all $u \in [0, 1]$,

$$F''(u) = -\int_{a_u}^{b_u} \frac{\partial}{\partial y} f(x, c_u(x)) \left(\frac{\partial}{\partial u} c_u(x)\right)^2 + f(x, c_u(x)) \frac{\partial^2}{\partial u^2} c_u(x) \mathrm{d}x \tag{15.36}$$

We observe that under the assumptions (15.31)–(15.33)

$$\left|\frac{\partial}{\partial u} c_u(x)\right| = \frac{|\tau(x, c_u(x)) - \sigma(x, c_u(x))|}{|(1-u)\frac{\partial}{\partial y}\sigma + u\frac{\partial}{\partial y}\tau(x, c_u(x))|} \leq C' h^2. \tag{15.37}$$

where C' depends on C_1, C_2, C_3. By differentiating (15.19) twice we obtain

$$\frac{\partial^2}{\partial u^2} c_u(x) = \frac{-2\frac{\partial}{\partial u} c_u \left(\frac{\partial}{\partial y}\tau - \frac{\partial}{\partial y}\sigma\right) - u\left(\frac{\partial}{\partial u} c_u\right)^2 \frac{\partial^2}{\partial y^2}\tau}{(1-u)\frac{\partial}{\partial y}\sigma + u\frac{\partial}{\partial y}\tau(x, c_u(x))} \tag{15.38}$$

and thus using again the assumptions on τ and σ we get

$$\left|\frac{\partial^2}{\partial u^2} c_u(x)\right| \leq C'' h^3 \tag{15.39}$$

where C'' depends again on C_1, C_2, C_3. By estimating the integral by the supremum we conclude that for all $u \in [0, 1]$

$$F''(u) \leq C''' h^4. \tag{15.40}$$

The result is obtained by combining Taylor's theorem with the approximation properties of the employed Gauss rules for the bi- and univariate quadrature. □

Lemma 15.2 *Assume that a rectangular cell K satisfies the assumptions of the triangular base case (in the strong sense) and that in addition to the assumptions (15.31)–(15.33) in Lemma 15.1 there is a constant C_4 independent of h, such that*

$$\left|\frac{\partial \tau}{\partial x}(x, y)\right| \geq C_4 > 0 \quad , \quad \forall (x, y) \in K \tag{15.41}$$

Then, there exists a constant $C_{\text{triangle}}(C_1, C_2, C_3, C_4, f)$, such that the corrected trimmed quadrature on this cell fulfills

$$|I_{\tau,K} f - \mathrm{CLT}_K(h, 2, 2) f| \leq C_{\text{triangle}} h^4. \tag{15.42}$$

Proof Again, we write $K = [x_0, x_0 + \alpha h] \times [y_0, y_0 + \beta h]$.. In the triangular case, b_u in (15.22) is not constant but defined implicitly by

$$c_u(b_u) = y_0 + \beta h. \tag{15.43}$$

For the second derivative of F this means that an additional term appears which depends on the derivative of b_u:

$$\begin{aligned} F''(u) = &- \int_{a_u}^{b_u} \frac{\partial}{\partial y} f(x, c_u(x)) \left(\frac{\partial}{\partial u} c_u(x) \right)^2 + f(x, c_u(x)) \frac{\partial^2}{\partial u^2} c_u(x) \mathrm{d}x \\ &- f(b_u, c_u(b_u)) \frac{\partial}{\partial u} c_u(b_u) \frac{\mathrm{d}}{\mathrm{d}u} b_u. \end{aligned} \tag{15.44}$$

In order to estimate the last term in (15.44), we compute

$$\frac{\mathrm{d}}{\mathrm{d}u} b_u = - \frac{\frac{\partial}{\partial u} c_u(b_u)}{\frac{\partial}{\partial x} c_u(b_u)}. \tag{15.45}$$

Differentiating (15.19) with respect to x leads to

$$\frac{\partial}{\partial x} c_u(x) = - \frac{\frac{\partial}{\partial x}\sigma + u \left(\frac{\partial}{\partial x}\tau(x, c_u(x)) - \frac{\partial}{\partial x}\sigma \right)}{\frac{\partial}{\partial y}\sigma + u \left(\frac{\partial}{\partial y}\tau(x, c_u(x)) - \frac{\partial}{\partial y}\sigma \right)}. \tag{15.46}$$

Using assumption (15.41) we conclude

$$\left| \frac{\partial}{\partial x} c_u(x) \right| \geq C'''' > 0 \tag{15.47}$$

and thus

$$\left| \frac{\mathrm{d}}{\mathrm{d}u} b_u \right| \leq C''''' h^2. \tag{15.48}$$

Therefore, in view of (15.37) and (15.39) the last term in (15.36) satisfies

$$f(b_u, c_u(b_u)) \frac{\partial}{\partial u} c_u(b_u) \frac{\mathrm{d}}{\mathrm{d}u} b_u \leq C''''' h^4 \tag{15.49}$$

and the result follows. □

Unfortunately, even with these two results at hand, we cannot analyze the quadrature rule CLT directly. This is due to two reasons: First, we cannot guarantee that all the triangular cells in the subdivision $\mathscr{K}$ satisfy the assumption of Lemma 15.2. Second, there may be trimmed cells which are not base cases in the strong sense. Both problems are resolved by suitably modifying the quadrature rule.

More precisely, we construct a modified quadrature rule CLT* by replacing the subdivision $\mathscr{K}$ of Ω with a new subdivision $\mathscr{K}^\star$. We begin by using creating a uniform grid of cells of size $\frac{h}{2}$. We assume that h is small enough, such that all cells belong to one of the base cases. The subdivision $\mathscr{K}^0$ consists of all cells that possess a non-empty intersection with the integration domain. Note that this also includes cells where the trimming curve crosses the same edge twice.

We will obtain $\mathscr{K}^\star$ by merging some of the trimmed cells in $\mathscr{K}^0$.

First, we define the constants C_1 and C_4 that are to be used in Lemmas 15.1 and 15.2 as

$$C_1 = C_4 = \frac{1}{4} \min_{\tau(x,y)=0} \|\nabla \tau(x, y)\|_2. \tag{15.50}$$

If h is sufficiently small this means that each *trimmed* cell $K \in \mathscr{K}^0$ belongs to one of three classes:

1. Class H (horizontal gradient): For all $(x, y) \in K$ we have

$$\left|\frac{\partial}{\partial x}\tau(x, y)\right| \geq C_4, \text{ and } \left|\frac{\partial}{\partial y}\tau(x_0, y_0)\right| < C_1 \text{ for at least one } (x_0, y_0) \in K. \tag{15.51}$$

2. Class V (vertical gradient): For all $(x, y) \in K$ we have

$$\left|\frac{\partial}{\partial y}\tau(x, y)\right| \geq C_1, \text{ and } \left|\frac{\partial}{\partial x}\tau(x_0, y_0)\right| < C_4 \text{ for at least one } (x_0, y_0) \in K. \tag{15.52}$$

3. Class D (diagonal gradient): For all $(x, y) \in K$ we have

$$\left|\frac{\partial}{\partial x}\tau(x, y)\right| \geq C_4 \text{ and } \left|\frac{\partial}{\partial y}\tau(x, y)\right| \geq C_1. \tag{15.53}$$

This is illustrated in Fig. 15.7. If a cell K belongs to class H (resp. class V), then also all of its neighbors are either in class H (resp. V) or in class D. To obtain the modified subdivision $K^\star$ we merge all pairs of vertically adjacent cells where one of them is in class H. Similarly, we merge all pairs of horizontally adjacent cells where one of them is in class V. The remaining cells are kept. Note that this results in rectangular cells of maximum size h, since at most two cells will be merged, due to the restricted range of the gradients. The modified rule CLT* is obtained by applying CLT to the modified subdivision $\mathscr{K}^\star$.

Next, we prove the convergence result for the modified rule CLT*. The modified quadrature cells in CLT* ensure that all triangular cells belong to class D, where we have a bound on both partial derivatives, cf. (15.53). Moreover, all cells satisfy the assumptions about the base cases in the strong sense.

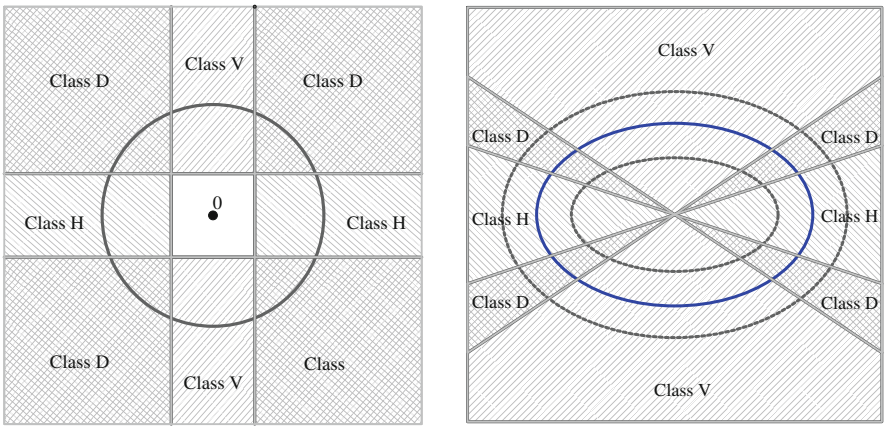


Fig. 15.7 Illustration of the regions of cell classes V, H and D. Left: The regions defined by $\nabla\tau(x, y)$ for (x, y) over the union of the trimmed cells. The black circle has radius $4C_1 = \min_{\tau(x,y)=0} \|\nabla\tau(x, y)\|_2$. It is visible that $\nabla\tau(x, y)$ lies outside the square in the middle. Right: An example trimming curve (an ellipse, shown in blue). The dotted offsets enclose the region that contains trimmed cells

Theorem 15.1 *Assume $\tau \in C^2([0, 1]^2)$ such that the constant $C_1 = C_4$ as defined in (15.50) is positive, and $f \in C^4([0, 1]^2)$. Then, there exists a constant $C_{\tau,f}$, such that the* CLT* *rule with $n = 2$ and $k = 2$ satisfies*

$$|I_\tau f - \mathrm{CLT}^\star(h, 2, 2) f| \leq C_{\tau,f} h^3. \tag{15.54}$$

Proof By the construction of CLT*, the subdivision $\mathscr{K}^\star$ consists of untrimmed quadrature cells, trimmed quadrilateral cells of classes H, V and D, and trimmed triangular and pentagonal cells of class D. In the trimmed quadrilateral cells we can always apply Lemma 15.1, while in the trimmed triangular cells of class D we can apply Lemma 15.2. Moreover, we can treat the trimmed pentagonal cells of class D by applying Lemma 15.2 to the complement of the quadrature domain. In both cases the constants C_2 and C_3 are obtained by linear approximation of τ. They depend on the second derivative of τ whose norm is bounded.

The number of trimmed cells does not exceed $C_5 \frac{1}{h}$ for some constant C_5. Moreover, we can use the same constants C_{quad} and C_{triangle} for all trimmed cells. Indeed, these constants depend on the values $C_1, \ldots, C_4$, which are determined by derivatives of the trimming function. Consequently, a global upper bound for these constants exists and depends solely on the trimming function τ. Moreover, we may use an upper bound on the derivatives of f. We conclude

$$\sum_{K \in \mathscr{K}^\star_{\text{trimmed}}} |I_{\tau,K} f - \mathrm{CLT}_K(h, 2, 2) f| \leq C_5 \max\{C_{\text{quad}}, C_{\text{triangle}}\} h^3. \tag{15.55}$$

Since we use $n = 2$ Gauss nodes in each direction for the untrimmed cells, the local error is bounded by $C_{\text{Gauss}} h^5$ in each of these cells for some constant C_{Gauss}. Since there are at most $\frac{1}{h^2}$ untrimmed cells, we have

$$\sum_{K \in \mathscr{K}^\star_{\text{untrimmed}}} |I_{\tau,K} f - \text{CLT}_K(h, 2, 2) f| \leq C_{\text{Gauss}} h^3. \tag{15.56}$$

The result (15.54) is implied by these two inequalities, since the various constants depend on τ and f only. □

We conjecture that in the triangular base case the influence of the last term in (15.44) is canceled by the corresponding term in the adjacent cell. Consequently, in practice it suffices to use CLT instead of CLT*. This is supported by our numerical experiments, where CLT is tested.

15.7 Numerical Experiments

We implemented the method in C++ using the G+Smo library [17]. In this section, we will test the approximation properties of the linearized trimmed quadrature rule CLT as well as the linearized trimmed quadrature rule LT on a number of trimmed geometries. Implementing the modified rule CLT* that was constructed in Sect. 15.6 for the theoretical analysis can be computationally inefficient in practice since it is necessary to estimate the gradient of τ in all cells for deciding which cells are to be joined. Additionally, one needs to find an upper bound of the gradient's norm along the trimming curve. However, we also implemented CLT* in a particular case in order to make a direct comparison with CLT. In the numerical experiments we observe that the theoretical error estimate for CLT* (Theorem 15.1) still holds for the original CLT quadrature rule.

15.7.1 Ellipse

As a first example, we use our method to compute the volume of an ellipse implicitly defined by

$$\tau(x, y) = -\frac{(x - 0.5)^2}{a^2} - \frac{(y - 0.5)^2}{b^2} + 1 > 0, \tag{15.57}$$

where we set $a = 0.45$ and $b = 0.2$ in our experiment. Figure 15.8 shows the result of the subdivision of this ellipse after some steps of refinement. In each cell, the trimming function was approximated by a linear function as described in Sect. 15.4. Note that in the pentagonal case we integrate over the remaining triangle

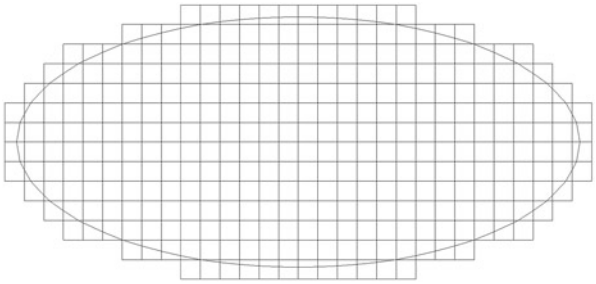


Fig. 15.8 Subdivision of the ellipse after some steps of refinement with approximate linear trimming curve in each cell

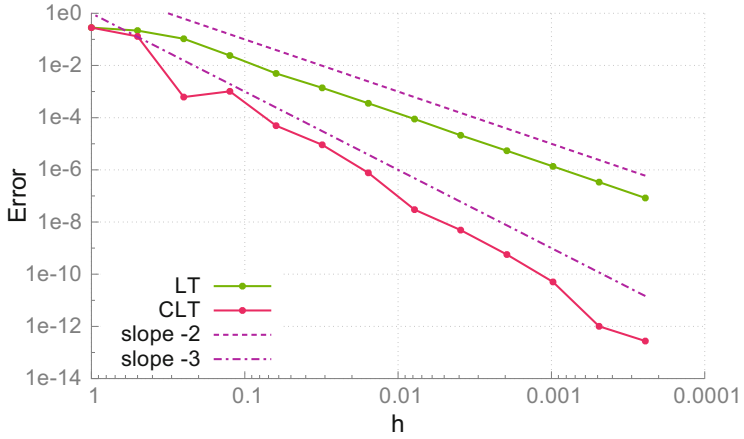


Fig. 15.9 Absolute error in LT and CLT for the area enclosed by the ellipse

and subtract from the full integral. In the left plot in Fig. 15.9 we show the quadrature error for different values of h when computing the area enclosed by the ellipse with the simple trimmed quadrature $LT(h, 1)$ described in Sect. 15.4 and with the trimmed quadrature by first order correction $CLT(h, 1, 2)$ described in Sect. 15.5. We observe that the first order error correction results in an additional order of convergence with respect to h, confirming the theoretical result from Theorem 15.1.

Next, we show the computation times for both quadrature rules in Fig. 15.10. We observe that applying the error correction does not result in a significant increase in complexity compared to the linearized trimmed quadrature. The complexity for both LT and CLT increases linearly with the number of cells.

Finally, we compare the CLT quadrature rule with the modified quadrature rule CLT* which was defined in Sect. 15.6 for the theoretical analysis of the convergence. For the ellipse given by the trimming function (15.57) the exact value of the

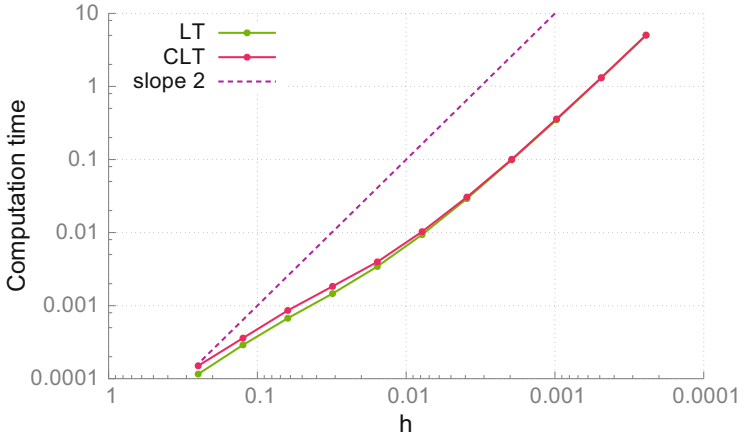


Fig. 15.10 Computation times for the approximation of the area enclosed by the ellipse

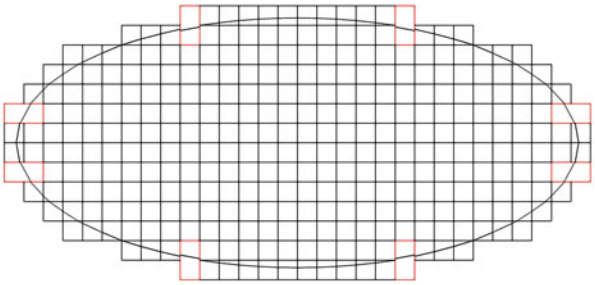


Fig. 15.11 Modified subdivision of the ellipse for CLT* after five steps of refinement. Joined cells are highlighted

constants $C_1 = C_4$ in the conditions (15.51), (15.52) and (15.53) is

$$C_1 = \frac{1}{4} \min_{\tau(x,y)=0} \|\nabla \tau(x, y)\|_2 = \frac{10}{9}.$$

We sample the gradient of τ on a fine grid in each cell in order to check these conditions. In Fig. 15.11 we show the modified subdivision of the ellipse after some steps of refinement. Eight cells were joined in this case, they are highlighted in the picture. Figure 15.12 shows the error of both CLT and CLT* when computing the area of the ellipse. We observe that the quadrature errors for both methods have the same asymptotic behavior.

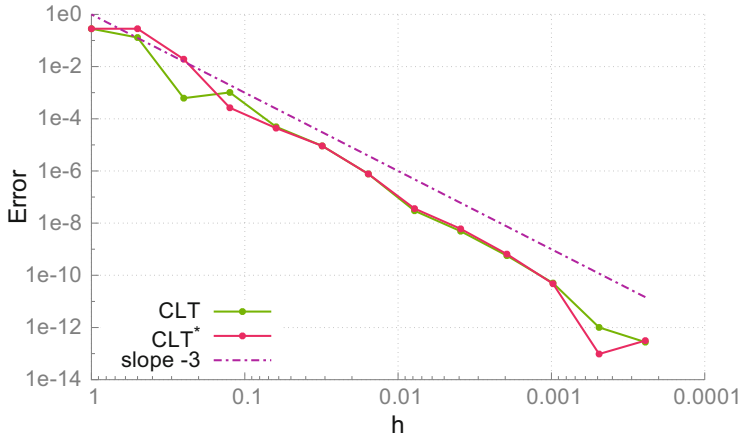


Fig. 15.12 Absolute error in CLT and CLT* for the area enclosed by the ellipse

15.7.2 Perforated Quarter Annulus

In our next example, we will approximate the area of a quarter annulus which is trimmed with three circles. The domain is represented by Non-Uniform Rational B-Splines (NURBS), which is a powerful technique for the representation of complex geometric shapes, see [22] for more details. The quarter annulus is represented exactly as a NURBS domain of polynomial degrees $(p_1, p_2) = (1, 2)$ with no interior knots.[1] Figure 15.13 shows the piecewise linear approximation of the trimming curve in the computational and in the parametric domain after some steps of refinement.

Since we perform the computation on the parametric domain, we approximate the integral

$$\int_{\Omega_{\gamma \circ G}} |\det J_G(x, y)| \mathrm{d}y\mathrm{d}x, \tag{15.58}$$

where $G : [0.1]^2 \to \mathbb{R}^2$ is the NURBS parameterization of the quarter annulus and $J_G(x, y)$ is the Jacobian matrix of G at the point (x, y). The trimming function $\gamma : \mathbb{R}^2 \to \mathbb{R}$ on the physical domain is defined as the product of the implicit representations of the three circles.

In the first plot in Fig. 15.14 we show the convergence rate of the quadrature rules with and without error correction. We chose $n = 2$ Gauss nodes for the linearized quadrature and $k = 2$ Gauss nodes for the correction term. As in the case of the

[1]The weights were chosen as $(\omega_{0,0}, \omega_{1,0}, \omega_{0,1}, \omega_{1,1}, \omega_{0,2}, \omega_{1,2}) = \left(1, 1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 1, 1\right)$ and the corresponding control points as (1, 0), (2, 0), (1, 1), (2, 2), (0, 1), (0, 2).

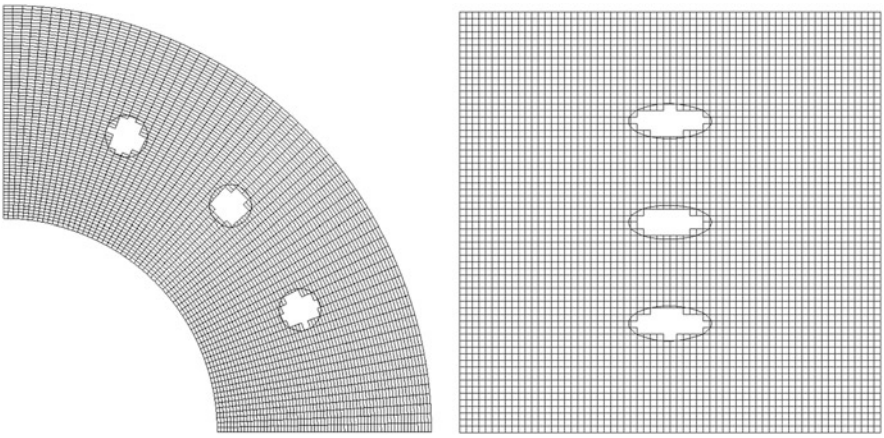


Fig. 15.13 Subdivision of the perforated quarter annulus and its corresponding parametric domain after some steps of refinement with approximate linear trimming curve in each cell

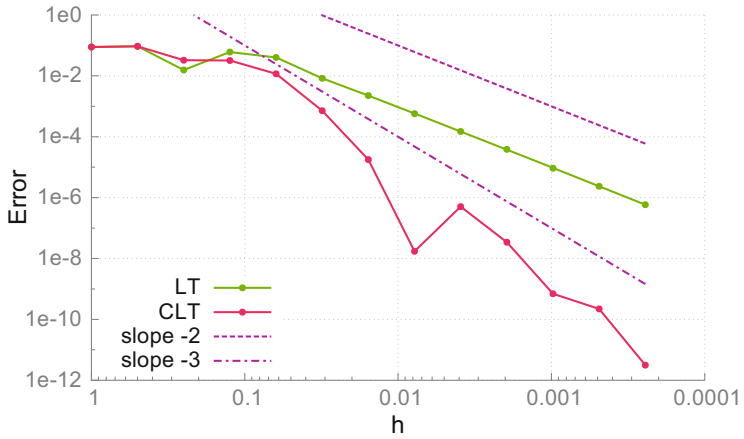


Fig. 15.14 Absolute error in LT and CLT for the area of the perforated quarter annulus for $n = 2, k = 2$

ellipse, we observe that the first order error correction term in CLT results in an additional order of convergence compared to the linearized quadrature in LT.

Since we only use one error correction term, the convergence error cannot be improved by additional Gauss nodes in the bivariate and univariate quadrature. This is confirmed by the second plot in Fig. 15.15 which shows the same experiment as in the first plot but for $n = 3$ and $k = 3$.

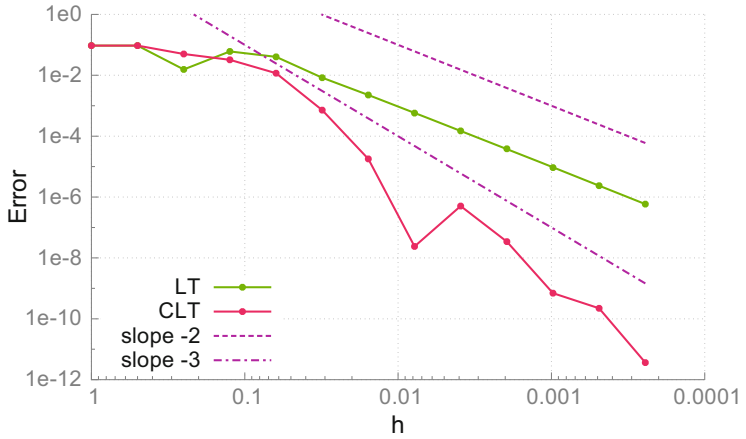


Fig. 15.15 Absolute error in LT and CLT for the area of the perforated quarter annulus for $n = 3, k = 3$

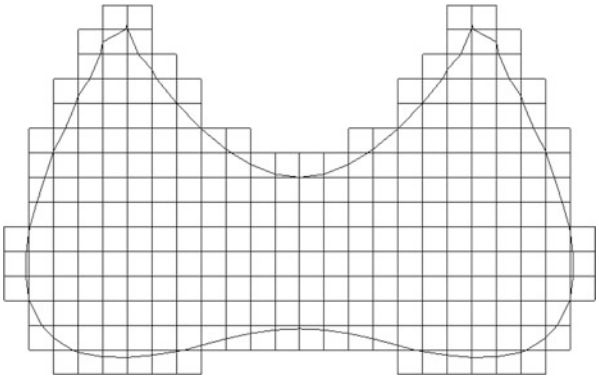


Fig. 15.16 Subdivision of the bicuspid with approximate linear trimming curve in each cell

15.7.3 Singular Case: Bicuspid Curve

Figure 15.16 shows a linear approximation of the bicuspid curve which is an algebraic curve given by

$$(x^2 - a^2)(x - a)^2 + (y^2 - a^2)^2 = 0 \tag{15.59}$$

for some $a > 0$. It has two cusps that hinder the improvement of the approximation order by the error correction term. In Fig. 15.17 we show the error convergence of the approximation of the area enclosed by the bicuspid, where the reference value was computed with a lower value of h. We observe that the error correction in CLT does not improve the convergence rate in this case, however, the absolute error is lower.

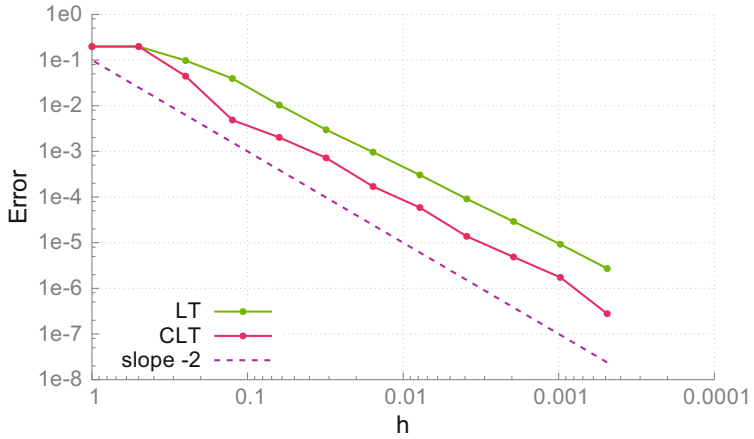


Fig. 15.17 Absolute error for the computation of the area enclosed by the bicuspid curve using LT and CLT

15.8 Conclusion

We presented a novel method for quadrature on trimmed two-dimensional domains which was shown to be accurate, efficient and easy to implement. In particular, cubic convergence was both proved and observed in our experiments, while the computation times were not compromised.

Two generalizations are possible. On the one hand, more terms of the Taylor expansion (15.16) can be added to the result in order to further improve the order of convergence of the quadrature error. This is necessary for applying the method to isogeometric discretizations of higher approximation power. Our conjecture is that each term of the Taylor expansion raises the order of convergence by one.

On the other hand, the method has the potential to be generalized to three-dimensional domains. We expect that the generalization will follow the same main ideas as in the two-dimensional case. However, the number of base cases will be higher.

Acknowledgements The authors gratefully acknowledge the support provided by the Austrian Science Fund (FWF) through project NFN S11708 and by the European Research Council (ERC), project GA 694515.

References

1. Bandara, K., Rüberg, T., Cirak, F.: Shape optimisation with multiresolution subdivision surfaces and immersed finite elements. Comput. Methods Appl. Mech. Eng. **300**(Supplement C), 510–539 (2016)

2. Bazilevs, Y., Calo, V., Cottrell, J., Evans, J., Hughes, T., Lipton, S., Scott, M., Sederberg, T.: Isogeometric analysis using T-splines. Comput. Methods Appl. Mech. Eng. **199**(5), 229–263 (2010)
3. Beer, G., Marussig, B., Zechner, J.: A simple approach to the numerical simulation with trimmed CAD surfaces. Comput. Methods Appl. Mech. Eng. **285**, 776–790 (2015)
4. Bordas, S.P.A., Burman, E.N., Larson, M.G., Olshanskii, M.A. (eds.) Geometrically Unfitted Finite Element Methods and Applications. Lecture Notes in Computational Science and Engineering. Springer International Publishing, Cham (2017)
5. Burman, E., Claus, S., Hansbo, P., Larson, M.G., Massing, A.: CutFEM: discretizing geometry and partial differential equations. Int. J. Numer. Methods Eng. **104**(7), 472–501 (2015)
6. Burman, E., Hansbo, P., Larson, M.G.: A cut finite element method with boundary value correction. Math. Comput. **87**, 633–657 (2018)
7. Cox, D.A., Sederberg, T.W., Chen, F.: The moving line ideal basis of planar rational curves. Comput. Aided Geom. Des. **15**(8), 803–827 (1998)
8. Gander, W., Gautschi, W.: Adaptive quadrature—revisited. BIT Numer. Math. **40**(1), 84–101 (2000)
9. Hofer, C., Langer, U., Toulopoulos, I.: Discontinuous Galerkin isogeometric analysis of elliptic diffusion problems on segmentations with gaps. SIAM J. Sci. Comput. **38**(6), A3430–A3460 (2016)
10. Hofer, C., Toulopoulos, I.: Discontinuous Galerkin isogeometric analysis for parametrizations with overlapping regions. RICAM-Report 2017–17 (2017)
11. Hughes, T., Cottrell, J., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**(39–41), 4135–4195 (2005)
12. Jüttler, B., Mantzaflaris, A., Perl, R., Rumpf, M.: On numerical integration in isogeometric subdivision methods for PDEs on surfaces. Comput. Methods Appl. Mech. Eng. **302**, 131–146 (2016)
13. Kamensky, D., Hsu, M.C., Schillinger, D., Evans, J.A., Aggarwal, A., Bazilevs, Y., Sacks, M.S., Hughes, T.J.: An immersogeometric variational framework for fluid-structure interaction: application to bioprosthetic heart valves. Comput. Methods Appl. Mech. Eng. **284**, 1005–1053 (2015). Isogeometric Analysis Special Issue
14. Kim, H., Seo, Y., Youn, S..: Isogeometric analysis for trimmed CAD surfaces. Comput. Methods Appl. Mech. Eng. **198**(37–40), 2982–2995 (2009)
15. Kumar, S., Manocha, D.: Efficient rendering of trimmed nurbs surfaces. Comput. Aided Des. **27**(7), 509–521 (1995). Display and visualisation
16. Lehrenfeld, C.: High order unfitted finite element methods on level set domains using isoparametric mappings. Comput. Methods Appl. Mech. Eng. **300**, 716–733 (2016)
17. Mantzaflaris, A., Scholz, F., others (see website): G+Smo (Geometry plus Simulation modules) v0.8.1 (2018). http://gs.jku.at/gismo
18. Marussig, B., Hughes, T.: A review of trimming in isogeometric analysis: challenges, data exchange and simulation aspects. Arch. Comput. Methods Eng. **25**, 1–69 (2017)
19. Marussig, B., Hiemstra, R., Hughes, T.: Improved conditioning of isogeometric analysis matrices for trimmed geometries. Comput. Methods Appl. Mech. Eng. **334**, 79–110 (2018)
20. Nagy, A., Benson, D.: On the numerical integration of trimmed isogeometric elements. Comput. Methods Appl. Mech. Eng. **284**, 165–185 (2015)
21. Newman, T.S., Yi, H.: A survey of the marching cubes algorithm. Comput. Graph. **30**(5), 854–879 (2006)
22. Pigl, L.A., Tiller, W.: The NURBS Book. Springer-Verlag, Berlin (1997)
23. Rüberg, T., Cirak, F.: A fixed-grid b-spline finite element technique for fluid-structure interaction. Int. J. Numer. Methods Fluids **74**(9), 623–660 (2014)
24. Ruess, M., Schillinger, D., Özcan, A., Rank, E.: Weak coupling for isogeometric analysis of non-matching and trimmed multi-patch geometries. Comput. Methods Appl. Mech. Eng. **269**, 46–71 (2014)

25. Sanches, R., Bornemann, P., Cirak, F.: Immersed b-spline (i-spline) finite element method for geometrically complex domains. Comput. Methods Appl. Mech. Eng. **200**(13), 1432–1445 (2011)
26. Schmidt, R., Wüchner, R., Bletzinger, K.U.: Isogeometric analysis of trimmed NURBS geometries. Comput. Methods Appl. Mech. Eng. **241–244**, 93–111 (2012)
27. Scholz, F., Mantzaflaris, A., Jüttler, B.: Partial tensor decomposition for decoupling isogeometric Galerkin discretizations. Comput. Methods Appl. Mech. Eng. **336**, 485–506 (2018)
28. Seiler, A., Jüttler, B.: Reparameterization and adaptive quadrature for the isogeometric discontinuous Galerkin method. In: Mathematical Methods for Curves and Surfaces, pp. 251–269. Springer International Publishing, Cham (2017)
29. Seo, Y.D., Kim, H.J., Youn, S.K.: Isogeometric topology optimization using trimmed spline surfaces. Comput. Methods Appl. Mech. Eng. **199**(49–52), 3270–3296 (2010)
30. Taubin, G.: Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **13**(11), 1115–1138 (1991)
31. Zhang, H., Mo, R., Wan, N.: An IGA discontinuous Galerkin method on the union of overlapped patches. Comput. Methods Appl. Mech. Eng. **326**, 446–480 (2017)
32. Zhu, X., Ma, Z., Hu, P.: Nonconforming isogeometric analysis for trimmed CAD geometries using finite-element tearing and interconnecting algorithm. Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci. **231**(8), 1371–1389 (2017)

Chapter 16
A Space–Time Finite Element Method for the Linear Bidomain Equations

Olaf Steinbach and Huidong Yang

Abstract In this work, we study a Galerkin–Petrov space–time finite element method for a linear system of parabolic–elliptic equations with in general anisotropic conductivity matrices, which may be considered as a simplified version of the nonlinear bidomain equations. The discretization is based on a stable space–time variational formulation employing continuous and piecewise linear finite elements in both spatial and temporal directions simultaneously. We show stability of the space–time formulation on both the continuous and discrete level for such a coupled problem under a rather general condition on the conductivity matrices. We further discuss the construction of a monolithic algebraic multigrid (AMG) method for solving the coupled linear system of algebraic equations globally. Numerical experiments are performed to demonstrate the convergence of the space–time finite element approximations, and the performance of the AMG method with respect to the mesh discretization parameter. Finally, we apply the space–time finite element method to the nonlinear bidomain equations in order to show the applicability of the proposed approach.

16.1 Introduction

The modelling of the electrical activity of the human heart relies on the Maxwell equations when neglecting the time derivative in Faraday's law. Hence we may use a scalar potential to describe the electric field, where the potential inside a cell is called intracellular potential, while the potential exterior to a cell is called extracellular potential. When the cells are at rest, there is a potential difference

O. Steinbach (✉)
Institut für Angewandte Mathematik, TU Graz, Graz, Austria
e-mail: o.steinbach@tugraz.at

H. Yang
Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria
e-mail: huidong.yang@oeaw.ac.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_16

across the cell membrane which is called the transmembrane potential. Hence, using the continuity equation and Ohm's law, and considering the ionic current exciting the cell, this results in a coupled system of nonlinear parabolic and elliptic partial differential equations, and a system of ordinary differential equations to describe the ionic current via cellular state variables. For a more detailed discussion of the mathematical model we refer to, e.g., [6, 10, 25].

As model problem we consider the simplified linear bidomain equations to find the transmembrane potential u_T and the extracellular potential u_e satisfying the linear parabolic–elliptic system

$$C_m \partial_t u_T(x,t) - \mathrm{div}_x[M_i(x)\nabla_x u_T(x,t)] - \mathrm{div}_x[M_i(x)\nabla_x u_e(x,t)] = s_i(x,t), \quad (16.1)$$

$$-\mathrm{div}_x[M_i(x)\nabla_x u_T(x,t)] - \mathrm{div}_x[(M_i(x)+M_e(x))\nabla_x u_e(x,t)] = s_e(x,t) \quad (16.2)$$

for $(x,t) \in Q := \Omega \times (0,T)$, where $\Omega \subset \mathbb{R}^n$ is assumed to be Lipschitz, $n = 2, 3$, with homogeneous Dirichlet boundary conditions $u_T = 0$ and $u_e = 0$ on the lateral boundary $\Sigma := \partial\Omega \times (0,T)$, and a given initial condition $u_T = 0$ in $\Omega, t = 0$. Note that inhomogeneous data can be handled via a standard homogenisation approach by using a suitable extension. Moreover, s_i and s_e are some given current sources, M_i and M_e are, in general anisotropic, conductivity matrices, that are assumed to be symmetric and positive definite, and satisfying

$$\mu\,(M_i(x)v, v) \le (M_e(x)v, v) \le \overline{\mu}\,(M_i(x)v, v) \quad \text{for all } v \in \mathbb{R}^n \quad (16.3)$$

uniformly for $x \in \Omega$ for some $0 < \mu \le \overline{\mu}$. Finally, C_m is the capacitance of the cell membrane.

In this simplified model we have neglected the nonlinear coupling term among the two potentials and the third variable, a vector of cellular state variables, via dropping the nonlinear ionic current term and the related nonlinear system of ordinary differential equations. In [6], a similar linear model problem is considered, with Robin boundary conditions instead of Dirichlet conditions. Admittedly, the bidomain reaction–diffusion system in its simplified form (16.1) and (16.2) will be a reasonable starting point towards the construction of robust monolithic algebraic multigrid (AMG) methods for the fully coupled nonlinear bidomain equations. For general concepts of AMG methods, we refer to [1, 2, 20].

As it is well known, most conventional methods for discretizing the nonlinear bidomain equations, i.e. the coupled parabolic–elliptic equations and the system of ordinary differential equations to describe the ionic current, are proper combinations of explicit/implicit time stepping methods and finite element methods with respect to the temporal and spatial directions, respectively; see, e.g., [3, 4, 10, 13–19, 22, 27].

Recently, a space–time discontinuous Galerkin finite element discretization on arbitrary simplex meshes has been employed to approximate the solution of the bidomain equations and the coupled electro-mechanical system [6, 7]. Such a space–time discontinuous Galerkin scheme has been investigated for the heat equation in

[11, 12], by treating the time as another variable, and adding an upwind with respect to the time derivative.

On the other hand, iterative and parallel solution methods for the bidomain equations have been considerably studied in the past years, that are mainly based on time stepping methods and operator splitting schemes. Related references on the splitting solution methods for the bidomain equations can be found, e.g., in [10, 13, 26].

In [3], the system of ordinary differential equations is solved by a combination of an exact solution of related scalar linear ordinary differential equations, and the explicit Euler method for the remaining equations. The coupled reaction–diffusion part is tackled as an elliptic problem via a NURBS-based isogeometric discretization [5] in space and a semi-implicit scheme in time, i.e., an implicit Euler method for the diffusion term, and an explicit treatment for the nonlinear reaction term. The optimal convergence rate of two-level additive Schwarz preconditioners for the resulting linear system is shown. Earlier results for multilevel Schwarz preconditioners for the bidomain parabolic–parabolic and parabolic–elliptic formulations (both become elliptic problems after temporal discretization) can be found in [14, 15]. A similar operator splitting scheme has been used in [27], where the resulting discrete bidomain elliptic equations at each time step are solved by a balancing Neumann–Neumann preconditioned conjugate gradient method.

In [19], an explicit Euler method has been used to solve the parabolic equation and nonlinear system of ordinary differential equations at each time step, and the remaining elliptic problem is solved by an algebraic multigrid method.

Block factorized preconditioners for the coupled bidomain reaction–diffusion 2×2 system in a semi-implicit time stepping method have been investigated in [17, 18], where an AMG method is used to approximate the blocks.

Very recently, a monolithic scheme has been studied in [6] for the fully coupled nonlinear bidomain equations, that are discretized by using a space–time discontinuous Galerkin finite element scheme in the space–time domain. On each Newton iteration, the linearized system is reduced to the Schur complement equation with respect to the two potential variables. Further, discrete stability conditions for both the linear and nonlinear problems are shown therein, with respect to specially chosen DG-norms.

In this work, we follow a continuous Galerkin–Petrov space–time finite element discretization scheme [23] for approximating the solution of the model problem (16.1) and (16.2) in the space–time domain. The resulting linear system of algebraic equations is then solved by a monolithic AMG method.

The remainder of this paper is organized as follows. In Sect. 16.2, we present a stable space–time variational formulation of the model problem (16.1) and (16.2), and in Sect. 16.3 we discuss the related continuous Galerkin–Petrov space–time finite element method. The monolithic algebraic multigrid method for the solution of the coupled linear system of algebraic equations is discussed in Sect. 16.4. Some numerical results are provided in Sects. 16.5 and 16.6 where we also consider the nonlinear system including the ionic current. Finally, some conclusions are drawn in Sect. 16.7.

16.2 Space–Time Variational Formulations

Let us define the function spaces

$$X := \Big\{ v \in L^2(0,T; H_0^1(\Omega)) \cap H^1(0,T; H^{-1}(\Omega)), v(x,0) = 0 \text{ for } x \in \Omega \Big\},$$

$$Y := L^2(0,T; H_0^1(\Omega))$$

to consider the Galerkin–Petrov variational formulation of the Dirichlet boundary value problem (16.1) and (16.2) to find $(u_T, u_e) \in X \times Y$ such that

$$a(u_T, u_e; v_T, v_e) = \int_0^T \int_\Omega \Big[s_i \, v_T + s_e \, v_e \Big] dx \, dt \tag{16.4}$$

is satisfied for all $(v_T, v_e) \in Y \times Y$ where the bilinear form is given as

$$\begin{aligned} a(u_T, u_e; v_T, v_e) := & \int_0^T \int_\Omega \Big[C_m \partial_t u_T \, v_T + (M_i \nabla_x u_T, \nabla_x v_T) + (M_i \nabla_x u_e, \nabla_x v_T) \Big] dx \, dt \\ & + \int_0^T \int_\Omega \Big[(M_i \nabla_x u_T, \nabla_x v_e) + ((M_i + M_e) \nabla_x u_e, \nabla_x v_e) \Big] dx \, dt \, . \end{aligned}$$

To establish unique solvability of the space–time variational formulation (16.4) we need to have a related stability estimate for the involved bilinear form. For this we first consider an ellipticity estimate for the spatial part.

Lemma 16.1 *For $u_T, u_e, v_T, v_e \in Y$ we consider the spatial bilinear form*

$$\begin{aligned} a_S(u_T, u_e; v_T, v_e) := & \int_0^T \int_\Omega \Big[(M_i \nabla_x u_T, \nabla_x v_T) + (M_i \nabla_x u_e, \nabla_x v_T) \Big] dx \, dt \\ & + \int_0^T \int_\Omega \Big[(M_i \nabla_x u_T, \nabla_x v_e) + ((M_i + M_e) \nabla_x u_e, \nabla_x v_e) \Big] dx \, dt \end{aligned} \tag{16.5}$$

and Assumption (16.3). Then, for $(v_T, v_e) \in Y \times Y$ there holds the ellipticity estimate

$$a_S(v_T, v_e; v_T, v_e) \geq c_S \, \| (v_T, v_e) \|^2_{Y \times Y}$$

with the positive constant

$$c_S = 1 + \frac{\mu}{2} - \sqrt{\frac{\mu^2}{4} + 1} > 0,$$

and with respect to the norm

$$\|(v_T, v_e)\|^2_{Y\times Y} := \int_0^T \int_\Omega \Big[(M_i \nabla_x v_T, \nabla_x v_T) + (M_i \nabla_x v_e, \nabla_x v_e)\Big] dx\, dt\,.$$

Proof Using Assumption (16.3) we have, for some $\gamma > 0$,

$$\begin{aligned}
&a_S(v_T, v_e; v_T, v_e) = \\
&= \int_0^T \int_\Omega \Big[(M_i \nabla_x v_T, \nabla_x v_T) + 2\,(M_i \nabla_x v_e, \nabla_x v_T) + ((M_i + M_e)\nabla_x v_e, \nabla_x v_e)\Big] dx\, dt \\
&\geq \int_0^T \int_\Omega \Big[(M_i \nabla_x v_T, \nabla_x v_T) + 2\,(M_i \nabla_x v_e, \nabla_x v_T) + (1+\mu)(M_i \nabla_x v_e, \nabla_x v_e)\Big] dx\, dt \\
&= \int_0^T \int_\Omega \left[\left(1 - \frac{1}{\gamma}\right)(M_i \nabla_x v_T, \nabla_x v_T) + \Big(1 + \mu - \gamma\Big)(M_i \nabla_x v_e, \nabla_x v_e)\right] dx\, dt \\
&\qquad + \int_0^T \int_\Omega \left(M_i \nabla_x \left(\frac{1}{\sqrt{\gamma}} v_T + \sqrt{\gamma} v_e\right), \nabla_x \left(\frac{1}{\sqrt{\gamma}} v_T + \sqrt{\gamma} v_e\right)\right) dx\, dt \\
&\geq \Big(1 + \mu - \gamma^*\Big) \int_0^T \int_\Omega \Big[(M_i \nabla_x v_T, \nabla_x v_T) + (M_i \nabla_x v_e, \nabla_x v_e)\Big] dx\, dt
\end{aligned}$$

if

$$1 - \frac{1}{\gamma^*} = 1 + \mu - \gamma^*$$

is satisfied, i.e.

$$\gamma^* = \frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} + 1}, \quad c_S = 1 + \mu - \gamma^* = 1 + \frac{\mu}{2} - \sqrt{\frac{\mu^2}{4} + 1} > 0.$$

□

In fact, the bilinear form (16.5) induces a norm in $Y \times Y$, i.e. for $(v_T, v_e) \in Y \times Y$ we have

$$\|(v_T, v_e)\|^2_M := a_S(v_T, v_e; v_T, v_e)$$

satisfying

$$c_S\, \|(v_T, v_e)\|^2_{Y\times Y} \leq \|(v_T, v_e)\|^2_M \leq \left(1 + \frac{\overline{\mu}}{2} + \sqrt{\frac{\overline{\mu}^2}{4} + 1}\right) \|(v_T, v_e)\|^2_{Y\times Y}\,.$$

Hence we can define $(w_T, w_e) \in Y \times Y$ as the unique solution of the variational formulation

$$a_S(w_T, w_e; v_T, v_e) = a(u_T, u_e; v_T, v_e) \quad \text{for all } (v_T, v_e) \in Y \times Y, \tag{16.6}$$

where $(u_T, u_e) \in X \times Y$ is given. For the latter we introduce the norm

$$\|(u_T, u_e)\|_{X\times Y} := \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a(u_T, u_e; v_T, v_e)}{\|(v_T, v_e)\|_M}, \tag{16.7}$$

and where we can write the bilinear form, by using integration by parts, as

$$\begin{aligned} a(u_T, u_e; v_T, v_e) = {} & \langle C_m \partial_t u_T - \mathrm{div}_x[M_i \nabla_x u_T] - \mathrm{div}_x[M_i \nabla_x u_e], v_T\rangle_Q \\ & + \langle -\mathrm{div}_x[M_i \nabla_x u_T] - \mathrm{div}_x[(M_i + M_e)\nabla_x u_e], v_e\rangle_Q . \end{aligned}$$

Recall that $\langle \cdot, \cdot \rangle_Q$ denotes the duality pairing as extension of the L^2 inner product in the space time domain. The norm (16.7) is indeed the adjoint norm of the partial differential operator in (16.1) and (16.2) applied to (u_T, u_e). Although the norm definition (16.7) already implies a related stability condition, we will present a proof in order to establish the forthcoming relation (16.9).

Lemma 16.2 *For all $(u_T, u_e) \in X \times Y$ there holds the stability condition*

$$\|(u_T, u_e)\|_{X\times Y} \leq \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a(u_T, u_e; v_T, v_e)}{\|(v_T, v_e)\|_M} . \tag{16.8}$$

Proof For the unique solution $(w_T, w_e) \in Y \times Y$ of the variational problem (16.6) we first have

$$\begin{aligned} \|(w_T, w_e)\|_M^2 &= a_S(w_T, w_e; w_T, w_e) \\ &= a(u_T, u_e; w_T, w_e) \leq \|(u_T, u_e)\|_{X\times Y} \|(w_T, w_e)\|_M , \end{aligned}$$

i.e.

$$\|(w_T, w_e)\|_M \leq \|(u_T, u_e)\|_{X\times Y} .$$

On the other hand,

$$\begin{aligned} \|(u_T, u_e)\|_{X\times Y} &= \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a(u_T, u_e, v_T, v_e)}{\|(v_T, v_e)\|_M} \\ &= \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a_S(w_T, w_e, v_T, v_e)}{\|(v_T, v_e)\|_M} \leq \|(w_T, w_e)\|_M , \end{aligned}$$

implying

$$\|(w_T, w_e)\|_M = \|(u_T, u_e)\|_{X\times Y}\,.$$

Hence we conclude

$$\|(u_T, u_e)\|^2_{X\times Y} = \|(w_T, w_e)\|^2_M = a_S(w_T, w_e; w_T, w_e) = a(u_T, u_e; w_T, w_e), \tag{16.9}$$

and therefore

$$\|(u_T, u_e)\|_{X\times Y} = \frac{a(u_T, u_e; w_T, w_e)}{\|(w_T, w_e)\|_M} \le \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a(u_T, u_e; v_T, v_e)}{\|(v_T, v_e)\|_M}$$

follows. □

Since the norm (16.7) is defined as adjoint norm of the partial differential operator applied on (u_T, u_e) we may ask for equivalent norms which are probably simpler to handle. Hence we introduce the space

$$\mathbb{Y}_0 := \Big\{(v_T, v_e) \in Y \times Y : \\ \langle M_i \nabla_x v_T, \nabla_x \phi_e\rangle_{L^2(Q)} + \langle (M_i + M_e)\nabla_x v_e, \nabla_x \phi_e\rangle_{L^2(Q)} = 0 \;\forall \phi_e \in Y\Big\}$$

and the norm

$$\|C_m \partial_t u_T\|_{Y'} := \sup_{0\neq(v_T,v_e)\in\mathbb{Y}_0} \frac{\langle C_m \partial_t u_T, v_T\rangle_Q}{\|(v_T, v_e)\|_M}. \tag{16.10}$$

Corollary 16.1 *For $(u_T, u_e) \in X \times Y$ there holds the stability condition*

$$\frac{1}{\sqrt{2}}\Big[\|(u_T, u_e)\|^2_M + \|C_m \partial_t u_T\|^2_{Y'}\Big]^{1/2} \le \sup_{0\neq(v_T,v_e)\in Y\times Y} \frac{a(u_T, u_e; v_T, v_e)}{\|(v_T, v_e)\|_M}\,. \tag{16.11}$$

Proof We start to consider, by using (16.9),

$$\begin{aligned}
\|(u_T, u_e)\|^2_{X\times Y} &= a(u_T, u_e; w_T, w_e)\\
&= a(u_T, u_e; u_T, u_e) + a(u_T, u_e; w_T - u_T, w_e - u_e)\\
&= \langle C_m \partial_t u_T, u_T\rangle_{L^2(Q)} + a_S(u_T, u_e; u_T, u_e) + a_S(w_T, w_e; w_T - u_T, w_e - u_e)\\
&\ge a_S(u_T, u_e; u_T, u_e) + a_S(w_T, w_e; w_T - u_T, w_e - u_e)\\
&= a_S(u_T, u_e; u_T, u_e) + a_S(w_T - u_T, w_e - u_e; w_T - u_T, w_e - u_e)\\
&\qquad + a_S(u_T, u_e; w_T - u_T, w_e - u_e)
\end{aligned}$$

$$\geq \|(u_T,u_e)\|_M^2 + \|(w_T-u_T,w_e-u_e)\|_M^2 - \|(u_T,u_e)\|_M \|(w_T-u_T,w_e-u_e)\|_M$$
$$\geq \frac{1}{2}\Big[\|(u_T,u_e)\|_M^2 + \|(w_T-u_T,w_e-u_e)\|_M^2\Big].$$

It remains to compute

$$\begin{aligned}\|(w_T-u_T,w_e-u_e)\|_M^2 &= a_S(w_T-u_T,w_e-u_e;w_T-u_T,w_e-u_e)\\ &= a_S(w_T,w_e;w_T-u_T,w_e-u_e) - a_S(u_T,u_e;w_T-u_T,w_e-u_e)\\ &= a(u_T,u_e;w_T-u_T,w_e-u_e) - a_S(u_T,u_e;w_T-u_T,w_e-u_e)\\ &= \langle C_m\partial_t u_T, w_T-u_T\rangle_Q\\ &= \langle C_m\partial_t u_T, z_T\rangle_Q,\end{aligned}$$

where $(z_T,z_e) := (w_T-u_T,w_e-u_e) \in Y\times Y$ is the unique solution of the variational problem

$$a_S(z_T,z_e;v_T,v_e) = \langle C_m\partial_t u_T, v_T\rangle_Q \quad \text{for all } (v_T,v_e)\in Y\times Y, \tag{16.12}$$

i.e.

$$\int_0^T\int_\Omega \Big[(M_i\nabla_x z_T,\nabla_x v_T) + (M_i\nabla_x z_e,\nabla_x v_T)\Big]dx\,dt = \int_0^T\int_\Omega C_m\partial_t u_T\, v_T\,dx\,dt,$$
$$\int_0^T\int_\Omega \Big[(M_i\nabla_x z_T,\nabla_x v_e) + ((M_i+M_e)\nabla_x z_e,\nabla_x v_e))\Big]dx\,dt = 0.$$

Hence we conclude

$$\|(z_T,z_e)\|_M^2 = a_S(z_T,z_e;z_T,z_e) = \langle C_m\partial_t u_T, z_T\rangle_Q,$$

i.e.

$$\|(z_T,z_e)\|_M = \frac{\langle C_m\partial_t u_T,z_T\rangle_Q}{\|(z_T,z_e)\|_M} \leq \sup_{0\neq(v_T,v_e)\in\mathbb{Y}_0}\frac{\langle C_m\partial_t u_T,v_T\rangle_Q}{\|(v_T,v_e)\|_M} =: \|C_m\partial_t u_T\|_{Y'}.$$

On the other hand,

$$\begin{aligned}\|C_m\partial_t u_T\|_{Y'} &= \sup_{0\neq(v_T,v_e)\in\mathbb{Y}_0}\frac{\langle C_m\partial_t u_T,v_T\rangle_Q}{\|(v_T,v_e)\|_M}\\ &= \sup_{0\neq(v_T,v_e)\in\mathbb{Y}_0}\frac{a_S(z_T,z_e;v_T,v_e)}{\|(v_T,v_e)\|_M} \leq \|(z_T,z_e)\|_M\end{aligned}$$

implies

$$\|C_m \partial_t u_T\|_{Y'} = \|(z_T, z_e)\|_M ,$$

where $(z_T, z_e) \in Y \times Y$ solves the variational problem (16.12), i.e. the norm (16.10) is induced by the Schur complement operator of the system (16.12) when eliminating z_e from the second equation. □

Remark 16.1 Instead of the parabolic–elliptic system (16.1) and (16.2) we may also consider the related Schur complement system when eliminating the extracellular potential u_e. This results in a parabolic evolution equation with the bounded and elliptic Schur complement operator, and applying arguments as for the standard heat equation, see, e.g., [23], we would conclude a similar stability estimate as given in (16.11).

16.3 A Galerkin–Petrov Space–Time Finite Element Method

We decompose the space–time cylinder $Q = \Omega \times (0, T) \subset \mathbb{R}^{n+1}$ into simplicial finite elements q_ℓ, i.e. $Q_h = \cup_{\ell=1}^N \overline{q}_\ell$. For simplicity, we assume that Ω is polygonal or polyhedral, i.e., $\overline{Q} = Q_h$. The finite element spaces are given by $X_h = S_h^1(Q_h) \cap X$ and $Y_h = X_h$ with $S_h^1(Q_h) = \text{span}\{\varphi_i\}_{i=1}^M$ being the span of piecewise linear and continuous basis functions φ_i.

The conforming discrete Galerkin–Petrov variational formulation of (16.4) is to find $(u_{T,h}, u_{e,h}) \in X_h \times Y_h \subset X \times Y$ such that

$$a(u_{T,h}, u_{h,e}; v_{T,h}, v_{e,h}) = \int_0^T \int_\Omega \Big[s_i \, v_{T,h} + s_e \, v_{e,h} \Big] dx \, dt \tag{16.13}$$

is satisfied for all $(v_{T,h}, v_{e,h}) \in Y_h \times Y_h$, where we assume $X_h \subset Y_h$. Analogously as in [23, Theorem 3.1] we can show a discrete inf–sup condition which ensures unique solvability of (16.13). Related to the variational formulation (16.12) we define an approximate solution $(z_{T,h}, z_{e,h}) \in Y_h \times Y_h$ of the variational problem

$$a_S(z_{T,h}, z_{e,h}; v_{T,h}, v_{e,h}) = \langle C_m \partial_t u_{T,h}, v_{T,h} \rangle_Q \quad \text{for all } (v_{T,h}, v_{e,h}) \in Y_h \times Y_h, \tag{16.14}$$

and as in (16.10) we define the discrete norm

$$\|C_m \partial_t u_{T,h}\|_{Y',h} = \|(z_{T,h}, z_{e,h})\|_M \leq \|(z_T, z_e)\|_M = \|C_m \partial_t u_{T,h}\|_{Y'} .$$

Now we are in a position to prove, as in [23, Theorem 3.1], a discrete stability condition for the bilinear form $a(\cdot, \cdot; \cdot, \cdot)$.

Theorem 16.1 *Assume $X_h \subset X$, $Y_h \subset Y$, and $X_h \subset Y_h$. Then there holds the discrete stability condition*

$$\frac{1}{2\sqrt{2}}\left[\|(u_{T,h},u_{e,h})\|_M^2+\|C_m\partial_t u_{T,h}\|_{Y',h}^2\right]^{1/2} \tag{16.15}$$
$$\leq \sup_{0\neq(v_{T,h},v_{e,h})\in Y_h\times Y_h}\frac{a(u_{T,h},u_{e,h};v_{T,h},v_{e,h})}{\|(v_{T,h},v_{e,h})\|_M} \quad \textit{for all } (u_{T,h},u_{e,h})\in X_h\times Y_h.$$

Proof For $(u_{T,h},u_{e,h})\in X_h\times Y_h$ let $(z_{T,h},z_{e,h})\in Y_h\times Y_h$ be the unique solution of the variational problem (16.14). We then consider

$$\begin{aligned}
a(u_{T,h},u_{e,h};u_{T,h}+z_{T,h},u_{e,h}+z_{e,h}) &= \langle C_m\partial_t u_{T,h},u_{T,h}\rangle_Q + a_S(u_{T,h},u_{e,h};u_{T,h},u_{e,h})\\
&\quad +\langle C_m\partial_t u_{T,h},z_{T,h}\rangle_Q + a_S(u_{T,h},u_{e,h};z_{T,h},z_{e,h})\\
&\geq a_S(u_{T,h},u_{e,h};u_{T,h},u_{e,h})+a_S(z_{T,h},z_{e,h};z_{T,h},z_{e,h})+a_S(u_{T,h},u_{e,h};z_{T,h},z_{e,h})\\
&\geq \|(u_{T,h},u_{e,h})\|_M^2+\|(z_{T,h},z_{e,h})\|_M^2-\|(u_{T,h},u_{e,h})\|_M\|(z_{T,h},z_{e,h})\|_M\\
&\geq \frac{1}{2}\left[\|(u_{T,h},u_{e,h})\|_M^2+\|(z_{T,h},z_{e,h})\|_M^2\right]\\
&\geq \frac{1}{2}\left[\|(u_{T,h},u_{e,h})\|_M^2+\|C_m\partial_t u_{T,h}\|_{Y',h}^2\right].
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
\|(u_{T,h}+z_{T,h},u_{e,h}+z_{e,h})\|_M^2 &\leq \Big(\|(u_{T,h},u_{e,h})\|_M+\|(z_{T,h},z_{e,h})\|_M\Big)^2\\
&\leq 2\left(\|(u_{T,h},u_{e,h})\|_M^2+\|(z_{T,h},z_{e,h})\|_M^2\right)\\
&= 2\left(\|(u_{T,h},u_{e,h})\|_M^2+\|C_m\partial_t u_{T,h}\|_{Y',h}^2\right),
\end{aligned}$$

and therefore

$$\begin{aligned}
&a(u_{T,h},u_{e,h};u_{T,h}+z_{T,h},u_{e,h}+z_{e,h})\\
&\quad\geq \frac{1}{2\sqrt{2}}\left[\|(u_{T,h},u_{e,h})\|_M^2+\|C_m\partial_t u_{T,h}\|_{Y',h}^2\right]^{1/2}\|(u_{T,h}+z_{T,h},u_{e,h}+z_{e,h})\|_M\,.
\end{aligned}$$

follows which implies the assertion. □

The discrete stability condition (16.15) implies unique solvability of the Galerkin–Petrov finite element formulation (16.13). As in [23, Theorem 3.2] we then conclude

Cea's lemma,

$$\left[\|(u_T - u_{T,h}, u_e - u_{e,h})\|_M^2 + \|C_m \partial_t (u_T - u_{T,h})\|_{Y',h}^2\right]^{1/2}$$
$$\leq \inf_{(v_{T,h}, v_{e,h}) \in X_h \times Y_h} \left[\|(u_T - v_{T,h}, u_e - v_{e,h})\|_M^2 + \|C_m \partial_t (u_T - v_{T,h})\|_{Y'}^2\right]^{1/2},$$

and as in [23, Theorem 3.3] we can prove the following convergence result.

Theorem 16.2 *Let $(u_T, u_e) \in X \times Y$ and $(u_{T,h}, u_{e,h}) \in X_h \times Y_h$ be the unique solutions of the variational formulations (16.4) and (16.13), respectively. Let $Y_h = X_h = S_h^1(Q_h) \cap X$. Assume $(u_{T,h}, u_{e,h}) \in H^2(Q) \times H^2(Q)$. Then there holds the energy error estimate*

$$\|u_T - u_{T,h}\|_{L^2(0,T;H_0^1(\Omega))} + \|u_e - u_{e,h}\|_{L^2(0,T;H_0^1(\Omega))} \leq c\, h \left[|u_T|_{H^2(Q)} + |u_e|_{H^2(Q)}\right]. \tag{16.16}$$

From the definition of the bilinear form

$$a(u_T, u_e; v_T, v_e) = \langle C_m \partial_t u_T, v_T \rangle_Q + a_S(u_T, u_e; v_T, v_e)$$

we conclude, by using Lemma 16.1,

$$\begin{aligned} a(v_{T,h}, v_{e,h}; v_{T,h}, v_{e,h}) &= \langle C_m \partial_t v_{T,h}, v_{T,h} \rangle_Q + a_S(v_{T,h}, v_{e,h}; v_{T,h}, v_{e,h}) \\ &\geq \frac{1}{2} C_m \|u_{T,h}(T)\|_{L^2(\Omega)}^2 + \|(v_{T,h}, v_{e,h})\|_M^2 > 0, \end{aligned}$$

i.e. the stiffness matrix of the space time finite element variational formulation (16.4) is positive definite which is desirable for algebraic multigrid methods [1, 2, 20].

16.4 A Monolithic Algebraic Multigrid Method

The coupled system of linear equations arises from the variational formulation (16.13),

$$Ax = b. \tag{16.17}$$

Here x denotes the vector of coefficients of the finite element approximations for the transmembrane potential u_T and the extracellular potential u_e. In fact, we use a pointwise ordering of unknowns, which means at each node, we have two potential degrees of freedom. This approach has been utilized in the AMG methods for

solving fluid and elasticity problems in some monolithic fluid–structure interaction solvers [9].

For coarsening, we use a simple matrix graph based AMG coarsening strategy [8] to generate the hierarchical matrices on coarse levels, see the algorithm in [24, Section 3.2.1]. This coarsening strategy usually leads to a very low operator and grid complexity, approximately 1.2 and 1.1, respectively, in our numerical experiments. Here, grid complexity denotes the total number of degrees of freedom on all levels divided by the number of degrees of freedom on the finest level; operator complexity is the total number of nonzero entries in all matrices on all levels, divided by the number of nonzero entries on the finest level matrix. We refer to [2] for more details. More sophisticated AMG coarsening strategies for the space–time finite element discretization of parabolic equations are reported in [24], that may be considered for such a coupled system in the near future, and help to improve the AMG convergence rate. In addition, we need to have a proper smoother for such a coupled system. For the current being, we employ blockwise ILU [21] as a smoother for such a nonsymmetric system.

16.5 Numerical Results

In the following numerical example we set $\Omega = (0,1)^2$ and $T = 1$, i.e., the computational domain is a unit cube, $Q = (0,1)^3$. For studying the estimated order of convergence (eoc), we consider the exact solution

$$u_T(x,t) = x_1(1-x_1)x_2(1-x_2)t(1-t),$$

$$u_e(x,t) = \sin(\pi x_1)\sin(\pi x_2)\sin(\pi t).$$

We run simulations on 6 mesh refinement levels with tetrahedral elements. On the coarsest level, there are 250 degrees of freedom (#Dofs). The mesh on the next level is obtained by subdividing each tetrahedron on the previously coarser level into 8 smaller tetrahedra. On the finest level, there are 4, 293, 378 degrees of freedom.

The conductivity matrices are given by

$$M_i = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}, \quad M_e = \begin{bmatrix} 4.95 & 0.05 \\ 0.05 & 4.95 \end{bmatrix},$$

which are diagonally dominant and therefore positive definite. To check assumption (16.3) we compute $\mu = 12.5$ and $\overline{\mu} = 49$.

The estimated order of convergence (eoc) in $L^2(0,T;H_0^1(\Omega))$- and $L^2(Q)$-norms are shown in Tables 16.1 and 16.2 for u_T and u_e, respectively. In the numerical results we observe an almost linear convergence rate in the $L^2(0,T;H_0^1(\Omega))$-norm as predicted by the theory. Further, we see a second order convergence rate in the $L^2(Q)$-norm for u_T, and a bit less for u_e.

Table 16.1 Estimated order of convergence (eoc) of $\|u_T - u_{T,h}\|_{L^2(0,T;H_0^1(\Omega))}$ and $\|u_T - u_{T,h}\|_{L^2(Q)}$

#Dofs	$\|u_T - u_{T,h}\|_{L^2(0,T;H_0^1(\Omega))}$	eoc	$\|u_T - u_{T,h}\|_{L^2(Q)}$	eoc
250	1.01e−1	–	1.96e−2	–
1,458	4.76e−2	1.09	8.15e−3	1.26
9,826	1.78e−2	1.42	2.29e−3	1.83
71,874	7.93e−3	1.16	5.65e−4	2.02
549,250	4.15e−3	0.93	1.40e−4	2.02
4,293,378	2.22e−3	0.90	3.60e−5	1.95

Table 16.2 Estimated order of convergence (eoc) of $\|u_e - u_{e,h}\|_{L^2(0,T;H_0^1(\Omega))}$ and $\|u_e - u_{e,h}\|_{L^2(Q)}$

#Dofs	$\|u_e - u_{e,h}\|_{L^2(0,T;H_0^1(\Omega))}$	eoc	$\|u_e - u_{e,h}\|_{L^2(Q)}$	eoc
250	7.35e−1	–	1.02e−1	–
1,458	4.10e−1	0.84	3.69e−2	1.47
9,826	2.10e−1	0.96	1.14e−2	1.69
71,874	1.06e−1	1.00	3.60e−3	1.67
549,250	5.27e−2	1.00	1.83e−3	1.60
4,293,378	2.64e−2	1.00	4.02e−4	1.56

Table 16.3 AMG performance

#Dofs	#It	Time (s)	Opt Comp	Grid Comp
250	4	0.003 s	1.33	1.22
1,458	5	0.025 s	1.24	1.21
9,826	6	0.8 s	1.19	1.18
71,874	7	18 s	1.17	1.16
549,250	12	497 s	1.16	1.15
4,293,378	18	14,343 s	1.15	1.15

For the AMG solver, we set the relative residual norm 10^{-11} as a stopping criterion. In the smoothing steps, we apply Richardson iterations to the blockwise ILU preconditioned system. For the current being, we set the relative residual error 0.08 as a stopping criterion for the Richardson iterations in order to achieve multigrid convergence. This requires different smoothing steps on different mesh levels. Future work will concentrate on finding more robust smoothers for such coupled systems. In Table 16.3, we show the number of AMG iterations (#It), the computational time in seconds (s), the operator complexity (Opt Comp), and the grid complexity (Grid Opt). As observed, we obtain a reasonable AMG performance in terms of AMG iterations. Although the operator/grid complexity is low, the computational time is rather high due to the costly ILU smoother and various smoothing steps. This requires further investigations.

16.6 An Extension to the Nonlinear Model

In this section, we extend the space–time finite element method for the linear model to the fully nonlinear bidomain equations: Find the transmembrane potential u_T, the extracellular potential u_e, and the cellular state variable v, satisfying the system of the nonlinear bidomain equations

$$\begin{aligned} C_m \partial_t u_T(x,t) + I(u_T(x,t), v(x,t)) \\ -\mathrm{div}_x[M_i(x)\nabla_x u_T(x,t)] - \mathrm{div}_x[M_i(x)\nabla_x u_e(x,t)] = s_i(x,t), \end{aligned} \tag{16.18}$$

$$-\mathrm{div}_x[M_i(x)\nabla_x(x,t)] - \mathrm{div}_x[(M_i(x) + M_e(x))\nabla_x u_e(x,t)] = s_e(x,t), \tag{16.19}$$

$$\partial_t v(x,t) + H(u_T(x,t), v(x,t)) \ = s_v(x,t) \tag{16.20}$$

for $(x,t) \in Q$, with Dirichlet boundary conditions $u_T = g_T$, $u_e = g_e$ on the lateral boundary $\Sigma := \partial\Omega \times (0,T)$, and given initial conditions $u_T = u_0$, $v = v_0$ in Ω, $t = 0$. Here, we use the FitzHugh–Nagumo (FHN) model

$$I(u_T(x,t), v(x,t)) = c_1 u_T(x,t)(u_T(x,t) - u_{th})(u_T(x,t) - 1) + c_2 v(x,t), \tag{16.21}$$

$$H(u_T(x,t), v(x,t)) = b(dv(x,t), u_T(x,t)) \tag{16.22}$$

with given positive constants c_1, c_2, u_{th}, b, and d.

In this example, the conductivity matrices are given by

$$M_i = \begin{bmatrix} 0.75 & 0.15 \\ 0.15 & 0.75 \end{bmatrix}, \quad M_e = \begin{bmatrix} 1.25 & 0.30 \\ 0.30 & 1.25 \end{bmatrix},$$

and the constants are $c_1 = 0.175$, $c_2 = 0.03$, $u_{th} = 0.12$, $b = d = 10$. We use the exact solutions

$$\begin{aligned} u_T(x,t) &= x_1(1-x_1)x_2(1-x_2)t(1-t), \\ u_e(x,t) &= \sin(\pi x_1)\sin(\pi x_2)\sin(\pi t), \\ v(x,t) &= \cos(\pi x_1)\cos(\pi x_2)\cos(\pi t). \end{aligned}$$

The estimated order of convergence (eoc) in $L^2(0,T; H_0^1(\Omega))$- and $L_2(Q)$-norms are shown in Tables 16.4, 16.5, and 16.6 for u_T, u_e and v, respectively. As in the numerical example for the linear case, we see a linear convergence rate in the

Table 16.4 Estimated order of convergence (eoc) of $\|u_T - u_{T,h}\|_{L^2(0,T;H_0^1(\Omega))}$ and $\|u_T - u_{T,h}\|_{L_2(Q)}$

#Dofs	$\|u_T - u_{T,h}\|_{L_2(0,T;H_0^1(\Omega))}$	eoc	$\|u_T - u_{T,h}\|_{L^2(Q)}$	eoc
375	1.53e−2	–	9.60e−4	–
2,187	8.22e−3	0.90	2.90e−4	1.72
1,4739	3.99e−3	1.04	9.55e−5	1.60
107,811	1.96e−3	1.03	2.60e−5	1.87
823,875	9.74e−4	1.00	6.90e−6	1.91

Table 16.5 Estimated order of convergence (eoc) of $\|u_e - u_{e,h}\|_{L^2(0,T;H_0^1(\Omega))}$ and $\|u_e - u_{e,h}\|_{L_2(Q)}$

#Dofs	$\|u_e - u_{e,h}\|_{L_2(0,T;H_0^1(\Omega))}$	eoc	$\|u_e - u_{e,h}\|_{L^2(Q)}$	eoc
375	7.37e−1	–	1.08e−1	–
2,187	4.11e−1	0.84	3.98e−2	1.44
1,4739	2.11e−1	0.96	1.23e−2	1.69
107,811	1.06e−1	1.00	3.83e−3	1.68
823,875	5.28e−2	1.00	1.25e−3	1.61

Table 16.6 Estimated order of convergence (eoc) of $\|v - v_{e,h}\|_{L^2(0,T;H_0^1(\Omega))}$ and $\|v - v_{e,h}\|_{L^2(Q)}$

#Dofs	$\|v - v_{e,h}\|_{L_2(0,T;H_0^1(\Omega))}$	eoc	$\|v - v_{e,h}\|_{L_2(Q)}$	eoc
375	9.74e−1	–	3.90e−2	–
2,187	4.61e−1	1.08	8.79e−3	2.15
1,4739	2.22e−1	1.05	2.08e−3	2.08
107,811	1.10e−1	1.02	5.08e−4	2.03
823,875	5.48e−2	1.00	1.28e−4	1.99

$L^2(0,T;H_0^1(\Omega))$-norm. Further, we observe a quadratic convergence rate in the $L_2(Q)$-norm for V_{tm}, and a bit less for u_e.

16.7 Conclusions

In this contribution we have applied a continuous Galerkin–Petrov space–time finite element method [23] to a linear system of parabolic–elliptic equations, which may be considered as a simplified model towards the fully coupled nonlinear bidomain equations. It requires further development in order to apply such a space–time finite method to the full model which includes the nonlinearity and the cellular state variables. Then, for an accurate resolution of the wave type potentials the use of adaptive refined finite element meshes in the space–time domain seems to be mandatory, and motivates the proposed approach.

Under a rather general condition on the conductivities we have shown the stability of the space–time finite element method for the model problem. The linear

order of convergence for both potential variables with respect to the spatial energy norm has been confirmed by numerical results.

A monolithic AMG method has been utilized to solve the coupled system of algebraic equations up to about 4.3 million degrees of freedom, which on the one hand, already shows quite nice performance with respect to the AMG iterations, and on the other hand, demands further exploration on finding more robust and efficient smoothers.

Acknowledgements This work has been supported by the Austrian Science Fund (FWF) under the Grant SFB Mathematical Optimization and Applications in Biomedical Sciences.

References

1. Brandt, A., McCormick, S.F., Ruge, J.W.: Algebraic multigrid (AMG) for sparse matrix equations. In: Evans, D.J. (ed.) Sparsity and Its Applications, pp. 257–284. Cambridge University Press, Cambridge (1984)
2. Briggs, W., Henson, V., McCormick, S.: A Multigrid Tutorial, 2nd edn. SIAM, Philadelphia (2000)
3. Charawi, L.A.: Isogeometric overlapping Schwarz preconditioners for the bidomain reaction–diffusion system. Comput. Methods Appl. Mech. Eng. **319**, 472–490 (2017)
4. Chen, H., Li, X., Wang, Y.: A splitting preconditioner for a block two-by-two linear system with applications to the bidomain equations. J. Comput. Appl. Math. **321**, 487–498 (2017)
5. Hughes, T.J.R., Cottrell, J.A., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**(39), 4135–4195 (2005)
6. Karabelas, E.: Space–time discontinuous Galerkin methods for cardiac electromechanics. PhD thesis, Technische Universität Graz (2015)
7. Karabelas, E., Steinbach, O.: Space–time DG methods for the coupled electro–mechanical activation of the human heart. Proc. Appl. Math. Mech. **14**(1), 839–840 (2014)
8. Kickinger, F.: Algebraic multigrid for discrete elliptic second–order problems. In: Hackbush, W. (ed.) Multigrid Methods V. Proceedings of the 5th European Multigrid Conference. Lecture Notes in Computational Sciences and Engineering, vol. 3, pp. 157–172. Springer, Berlin (1998)
9. Langer, U., Yang, H.: Robust and efficient monolithic fluid–structure–interaction solvers. Int. J. Numer. Methods Eng. **108**(4), 303–325 (2016)
10. Linge, S., Sundnes, J., Hanslien, M., Lines, G.T., Tveito, A.: Numerical solution of the bidomain equations. Philos. Trans. R. Soc. A **367**(1895), 1931–1950 (2009)
11. Neumüller, M.: Space–time methods: fast solvers and applications. PhD thesis, Technische Universität Graz (2013)
12. Neumüller, M., Steinbach, O.: Refinement of flexible space–time finite element meshes and discontinuous Galerkin methods. Comput. Vis. Sci. **14**, 189–205 (2011)
13. Pathmanathan, P., Bernabeu, M.O., Bordas, R., Cooper, J., Garny, A., Pitt-Francis, J.M., Whiteley, J.P., Gavaghan, D.J.: A numerical guide to the solution of the bidomain equations of cardiac electrophysiology. Prog. Biophys Mol. Biol. **102**(2), 136–155 (2010)
14. Pavarino, L.F., Scacchi, S.: Multilevel additive Schwarz preconditioners for the bidomain reaction–diffusion system. SIAM J. Sci. Comput. **31**(1), 420–443 (2008)
15. Pavarino, L.F., Scacchi, S.: Parallel multilevel Schwarz and block preconditioners for the bidomain parabolic–parabolic and parabolic–elliptic formulations. SIAM J. Sci. Comput. **33**(4), 1897–1919 (2011)

16. Pennacchio, M., Simoncini, V.: Efficient algebraic solution of reaction–diffusion systems for the cardiac excitation process. J. Comput. Appl. Math. **145**(1), 49–70 (2002)
17. Pennacchio, M., Simoncini, V.: Algebraic multigrid preconditioners for the bidomain reaction–diffusion system. Appl. Numer. Math. **59**(12), 3033–3050 (2009)
18. Pennacchio, M., Simoncini, V.: Non-symmetric algebraic multigrid preconditioners for the bidomain reaction–diffusion system. In: Kreiss, G.G., Lötstedt, P., Målqvist, A., Neytcheva, M. (eds.) Numerical Mathematics and Advanced Applications 2009, pp. 729–736. Springer, Berlin (2010)
19. Plank, G., Liebmann, M., dos Santos, R.W., Vigmond, E.J., Haase, G.: Algebraic multigrid preconditioner for the cardiac bidomain model. IEEE Trans. Biomed. Eng. **54**(4), 585–596 (2007)
20. Ruge, J.W., Stüben, K.: Algebraic multigrid (AMG). In: McCormick, S.F. (ed.) Multigrid Methods, pp. 73–130. SIAM, Philadelphia (1987)
21. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM, Philadelphia (2003)
22. Southern, J.A., Plank, G., Vigmond, E.J., Whiteley, J.P.: Solving the coupled system improves computational efficiency of the bidomain equations. IEEE Trans. Biomed. Eng. **56**(10), 2404–2412 (2009)
23. Steinbach, O.: Space–time finite element methods for parabolic problems. Comput. Methods Appl. Math. **15**, 551–566 (2015)
24. Steinbach, O., Yang, H.: Comparison of algebraic multigrid methods for an adaptive space–time finite-element discretization of the heat equation in 3D and 4D. Numer. Linear Algebra Appl. **25**, e2143 (2018)
25. Sundnes, J., Lines, G.T., Cai, X., Nielsen, B.F., Mardal, K.-A., Tveito, A.: Computing the Electrical Activity of the Heart. Springer, Berlin (2006)
26. Vigmond, E.J., dos Santos, R.W., Prassl, A.J., Deo, M., Plank, G.: Solvers for the cardiac bidomain equations. Prog. Biophys. Mol. Biol. **96**(1–3), 3–18 (2008)
27. Zampini, S.: Balancing Neumann–Neumann methods for the cardiac bidomain model. Numer. Math. **123**(2), 363–393 (2013)

Chapter 17
A Stabilized Space–Time Finite Element Method for the Wave Equation

Olaf Steinbach and Marco Zank

Abstract We consider a space–time variational formulation of the wave equation by including integration by parts also in the time variable. A standard finite element discretization by using lowest order piecewise linear continuous functions then requires a CFL condition to ensure stability. To overcome this restriction, and following the work of Zlotnik (Convergence rate estimates of finite-element methods for second-order hyperbolic equations. In: Numerical methods and applications, pp. 155–220. CRC, Boca Raton, 1994), we consider, in the case of tensor–product space–time discretizations, a stabilized variational problem which is unconditionally stable. We provide a stability and error analysis, and some numerical results which confirm the theoretical findings.

17.1 Introduction

While for the analysis of parabolic and hyperbolic partial differential equations a variety of approaches such as Fourier and Laplace methods, semigroup theory, or Galerkin methods, is available, see, for example, [9–11, 14, 21, 22], standard approaches for the numerical solution are in most cases based on semi–discretizations where the discretization in space and time is split accordingly, see, e.g., [19] for parabolic problems, and [5, 6, 15] for hyperbolic equations. More recently, there exist space–time approaches as for example in [1, 2, 12, 13, 16, 17, 20] for parabolic problems, and [3, 4, 7, 8, 23] for hyperbolic equations, see also [18] where the space–time discretization of the wave equation requires some CFL condition.

In this work we introduce a stabilized finite element method for a second order ordinary differential equation and we transfer this approach to the corresponding hyperbolic partial differential equation. As model problem we consider the Dirichlet

O. Steinbach (✉) · M. Zank
Institut für Angewandte Mathematik, TU Graz, Graz, Austria
e-mail: o.steinbach@tugraz.at; zank@math.tugraz.at

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_17

problem for the wave equation,

$$\left.\begin{aligned} \partial_{tt} u(x,t) - \Delta_x u(x,t) &= f(x,t) && \text{for } (x,t) \in Q := \Omega \times (0,T), \\ u(x,t) &= 0 && \text{for } (x,t) \in \Sigma := \Gamma \times (0,T), \\ u(x,0) = \partial_t u(x,0) &= 0 && \text{for } x \in \Omega, \end{aligned}\right\} \tag{17.1}$$

where $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded domain with Lipschitz boundary $\Gamma = \partial\Omega$, $T > 0$ is a finite time and f is a given right-hand side. The variational formulation of (17.1) is to find $u \in H^{1,1}_{0;0,}(Q) := L^2(0,T; H^1_0(\Omega)) \cap H^1_{0,}(0,T; L^2(\Omega)) \subset H^1(Q)$ such that

$$-\langle \partial_t u, \partial_t w \rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x w \rangle_{L^2(Q)} = \langle f, w \rangle_Q \tag{17.2}$$

is satisfied for all $w \in H^{1,1}_{0;,0}(Q) := L^2(0,T; H^1_0(\Omega)) \cap H^1_{,0}(0,T; L^2(\Omega)) \subset H^1(Q)$, where $f \in L^2(Q)$ is given. Here, we use the standard Sobolev and Bochner spaces with the subspaces

$$H^1_{0,}(0,T; L^2(\Omega)) := \left\{ v \in H^1(0,T; L^2(\Omega)) : \; v(\cdot,0) = 0 \right\}$$

and

$$H^1_{,0}(0,T; L^2(\Omega)) := \left\{ v \in H^1(0,T; L^2(\Omega)) : \; v(\cdot,T) = 0 \right\}.$$

Furthermore, $\langle \cdot, \cdot \rangle_Q$ denotes the duality pairing as extension of the inner product in $L^2(Q)$. Note that the initial condition $u(\cdot,0) = 0$ is considered in the strong sense, whereas the initial condition $\partial_t u(\cdot,0) = 0$ is incorporated in a weak sense. It is well–known that for $f \in L^2(Q)$ there exists a unique solution $u \in H^{1,1}_{0;0,}(Q)$ of the variational formulation (17.2), see [9, Theorem 3.2 in Chapter IV], and [18].

Nevertheless, a conforming tensor–product space–time discretization of (17.2) by piecewise multilinear continuous functions requires the CFL condition [18]

$$h_t \le \frac{1}{\sqrt{d}} h_x, \tag{17.3}$$

where h_t and h_x are the uniform mesh sizes in time and space, see also Remark 17.1 in Sect. 17.3. To gain a deeper understanding of the CFL condition (17.3) a corresponding scalar ordinary differential equation

$$\partial_{tt} u(t) + \mu\, u(t) = f(t) \quad \text{for } t \in (0,T), \quad u(0) = \partial_t u(0) = 0, \tag{17.4}$$

where $\mu > 0$ and f are given, is analyzed and an unconditionally stable finite element method for (17.4) is introduced. Note that $\mu > 0$ is related to the eigenvalues of the Laplace operator for homogeneous Dirichlet conditions.

Instead of the variational formulation (17.2) we consider a stabilized formulation which generalizes the approach of [23]. This stabilization is first discussed for the scalar differential equation (17.4), and then transferred to the wave equation (17.1). The rest of this paper is organized as follows: In Sect. 17.2 we consider the second order ordinary differential equation (17.4), where we show unique solvability. In addition, we prove a discrete inf–sup condition to get an unconditionally stable numerical scheme and error estimates. We present some numerical examples to illustrate the theoretical results. In Sect. 17.3 we extend the ideas of Sect. 17.2 to the scalar wave equation, where we get an unconditionally stable finite element method for the wave equation. We present a numerical analysis of the discretization scheme, an error estimate with respect to $\|\cdot\|_{L^2(Q)}$ and we provide some numerical results for illustration.

17.2 Second Order Ordinary Differential Equations

As a model problem we consider the second order linear equation for $\mu > 0$,

$$\partial_{tt} u(t) + \mu\, u(t) = f(t) \quad \text{for } t \in (0,T), \qquad u(0) = \partial_t u(0) = 0, \tag{17.5}$$

and the variational formulation to find $u \in H^1_{0,}(0,T)$ such that

$$a(u,w) = \langle f, w\rangle_{(0,T)} \tag{17.6}$$

is satisfied for all $w \in H^1_{,0}(0,T)$, where $T > 0$ and $f \in [H^1_{,0}(0,T)]'$ are given, and where the bilinear form is

$$a(u,w) := -\langle \partial_t u, \partial_t w\rangle_{L^2(0,T)} + \mu \langle u, w\rangle_{L^2(0,T)}.$$

Note that $\langle\cdot,\cdot\rangle_{(0,T)}$ denotes the duality pairing as extension of the inner product in $L^2(0,T)$, and the Sobolev spaces

$$H^1_{0,}(0,T) := \Big\{ v \in H^1(0,T) \colon\ v(0) = 0 \Big\},$$
$$H^1_{,0}(0,T) := \Big\{ v \in H^1(0,T) \colon\ v(T) = 0 \Big\}$$

are endowed with the inner products

$$\langle u, v\rangle_{H^1_{0,}(0,T)} := \langle u, v\rangle_{H^1_{,0}(0,T)} := \int_0^T \partial_t u(t) \partial_t v(t)\, dt,$$

and with the induced norm

$$|u|^2_{H^1(0,T)} := \|\partial_t u\|^2_{L^2(0,T)} = \int_0^T [\partial_t u(t)]^2\,dt.$$

The dual space $[H^1_{,0}(0,T)]'$ is characterized as completion of $L^2(0,T)$ with respect to the norm

$$\|f\|_{[H^1_{,0}(0,T)]'} := \sup_{0\neq w\in H^1_{,0}(0,T)} \frac{\langle f, w\rangle_{(0,T)}}{|w|_{H^1(0,T)}}.$$

For $v \in H^1_{0,}(0,T)$ we define $w(t) = (\overline{\mathscr{H}}_T v)(t) := v(T) - v(t)$, i.e. $w \in H^1_{,0}(0,T)$. Then the variational formulation (17.6) is equivalent to the variational formulation to find $u \in H^1_{0,}(0,T)$ such that

$$-\langle \partial_t u, \partial_t \overline{\mathscr{H}}_T v\rangle_{L^2(0,T)} + \mu\langle u, \overline{\mathscr{H}}_T v\rangle_{L^2(0,T)} = \langle f, \overline{\mathscr{H}}_T v\rangle_{(0,T)} \tag{17.7}$$

is satisfied for all $v \in H^1_{0,}(0,T)$. Since the bilinear form

$$-\langle \partial_t u, \partial_t \overline{\mathscr{H}}_T v\rangle_{L^2(0,T)} = \langle \partial_t u, \partial_t v\rangle_{L^2(0,T)} \quad \text{for } u, v \in H^1_{0,}(0,T)$$

implies an elliptic operator $A : H^1_{0,}(0,T) \to [H^1_{0,}(0,T)]'$, unique solvability of the variational formulation (17.7) follows by using some compact perturbation argument, and injectivity, see [18, Theorem 4.7].

Next, we consider a conforming finite element discretization for the variational formulation (17.7). For a time interval $(0,T)$ and a discretization parameter $N \in \mathbf{N}$ we define nodes

$$0 = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = T,$$

finite elements $\tau_\ell = (t_{\ell-1}, t_\ell)$ of local mesh size $h_\ell = t_\ell - t_{\ell-1}$, $\ell = 1,\ldots,N$, the global mesh size $h = \max h_\ell$ and a related finite element space $S^1_h(0,T) = \operatorname{span}\{\varphi_k\}_{k=0}^N$ of piecewise linear continuous functions, where the basis functions φ_k are the usual hat functions. The Galerkin–Bubnov finite element discretization of the variational formulation (17.7) is to find $u_h \in V_h := S^1_h(0,T) \cap H^1_{0,}(0,T) = \operatorname{span}\{\varphi_k\}_{k=1}^N$ such that

$$-\langle \partial_t u_h, \partial_t \overline{\mathscr{H}}_T v_h\rangle_{L^2(0,T)} + \mu\langle u_h, \overline{\mathscr{H}}_T v_h\rangle_{L^2(0,T)} = \langle f, \overline{\mathscr{H}}_T v_h\rangle_{(0,T)} \tag{17.8}$$

is satisfied for all $v_h \in V_h$. It turns out that for a sufficiently small mesh size

$$h \le \frac{2\sqrt{3}}{(2+\sqrt{\mu}T)\mu T} \tag{17.9}$$

there holds the discrete stability condition [18, Theorem 4.13]

$$c(\mu, T)\, |u_h|_{H^1(0,T)} \leq \sup_{0 \neq v_h \in V_h} \frac{a(u_h, \overline{\mathscr{H}}_T v_h)}{|v_h|_{H^1(0,T)}} \quad \text{for all } u_h \in V_h,$$

implying the error estimate [18, Theorem 4.14], when assuming $u \in H^2(0, T)$,

$$|u - u_h|_{H^1(0,T)} \leq c(\mu, T)\, h\, |u|_{H^2(0,T)}\,.$$

When considering the stability of the finite element scheme (17.8) in the case of a uniform mesh, i.e. when analyzing the root condition, instead of (17.9) we conclude the weaker mesh assumption [18]

$$h \leq \sqrt{\frac{12}{\mu}}\,.$$

To overcome the mesh condition (17.9) we will stabilize the numerical scheme in (17.8) for which we need the following technical lemmata, where the trapezoidal rule is used analogously as in [23, Chapter 2]. In addition to $S_h^1(0, T)$ we also use the finite element space $S_h^0(0, T)$ of piecewise constant functions.

Lemma 17.1 *For all $f \in L^2(0, T)$ there holds*

$$\partial_t I_h \int_0^t f(s)ds = Q_h^0 f = \partial_t I_h \int_T^t f(s)ds, \tag{17.10}$$

where $I_h\colon C[0, T] \to S_h^1(0, T)$ is the piecewise linear nodal interpolation operator, and $Q_h^0\colon L^2(0, T) \to S_h^0(0, T)$ denotes the L^2 projection on the piecewise constant finite element space $S_h^0(0, T)$.

Proof For $t \in \tau_\ell = (t_{\ell-1}, t_\ell)$, $\ell = 1, \ldots, N$, we have

$$\partial_t I_h \int_0^t f(s)ds = \frac{1}{h_\ell}\left[\int_0^{t_\ell} f(s)ds - \int_0^{t_{\ell-1}} f(s)ds\right] = \frac{1}{h_\ell}\int_{t_{\ell-1}}^{t_\ell} f(s)ds = Q_h^0 f,$$

and

$$\partial_t I_h \int_T^t f(s)ds = \frac{1}{h_\ell}\left[\int_T^{t_\ell} f(s)ds - \int_T^{t_{\ell-1}} f(s)ds\right] = \frac{1}{h_\ell}\int_{t_{\ell-1}}^{t_\ell} f(s)ds = Q_h^0 f.$$

□

Lemma 17.2 *For all* $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ *and* $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ *there holds the representation*

$$\langle u_h, w_h \rangle_{L^2(0,T)} = \frac{1}{12} \sum_{\ell=1}^{N} h_\ell^2 \langle \partial_t u_h, \partial_t w_h \rangle_{L^2(\tau_\ell)} + \langle u_h, Q_h^0 w_h \rangle_{L^2(0,T)}, \tag{17.11}$$

where $Q_h^0 \colon L^2(0,T) \to S_h^0(0,T)$ *denotes the* L^2 *projection on the piecewise constant finite element space* $S_h^0(0,T)$.

Proof We consider the L^2 projection Q_h^0 on the finite element space $S_h^0(0,T)$, and with the error representation of the trapezoidal rule we obtain for each finite element τ_ℓ, $\ell = 1, \ldots, N$,

$$\begin{aligned} Q_h^0 \int_T^t w_h(s)\,ds &= \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} \int_T^t w_h(s)\,ds\,dt \\ &= \frac{1}{2} \left[\int_T^{t_{\ell-1}} w_h(s)\,ds + \int_T^{t_\ell} w_h(s)\,ds \right] - \frac{h_\ell^2}{12}\,\partial_t w_{h|\tau_\ell} \\ &= Q_h^0 I_h \int_T^t w_h(s)\,ds - \frac{h_\ell^2}{12}\,\partial_t w_{h|\tau_\ell}\,. \end{aligned}$$

Using integration by parts and (17.10) we further have

$$\begin{aligned} \int_0^T \partial_t u_h(t)\, I_h \int_T^t w_h(s)\,ds\,dt &= -\int_0^T u_h(t)\,\partial_t I_h \int_T^t w_h(s)\,ds\,dt \\ &= -\int_0^T u_h(t)\, Q_h^0 w_h(t)\,dt\,. \end{aligned}$$

With this we then conclude, by using integration by parts and the local definition of the L^2 projection Q_h^0,

$$\begin{aligned} \langle u_h, w_h \rangle_{L^2(0,T)} &= \int_0^T u_h(t)\,\partial_t \int_T^t w_h(s)\,ds\,dt = -\int_0^T \partial_t u_h(t) \int_T^t w_h(s)\,ds\,dt \\ &= -\sum_{\ell=1}^{N} \int_{t_{\ell-1}}^{t_\ell} \partial_t u_h(t)\, Q_h^0 \int_T^t w_h(s)ds\,dt \\ &= \sum_{\ell=1}^{N} \frac{h_\ell^2}{12} \int_{t_{\ell-1}}^{t_\ell} \partial_t u_h(t)\,\partial_t w_h(t)\,dt - \sum_{\ell=1}^{N} \int_{t_{\ell-1}}^{t_\ell} \partial_t u_h(t)\, Q_h^0 I_h \int_T^t w_h(s)\,ds\,dt \\ &= \frac{1}{12} \sum_{\ell=1}^{N} h_\ell^2 \langle \partial_t u_h, \partial_t w_h \rangle_{L^2(\tau_\ell)} - \sum_{\ell=1}^{N} \int_{t_{\ell-1}}^{t_\ell} \partial_t u_h(t)\, I_h \int_T^t w_h(s)\,ds\,dt \end{aligned}$$

$$= \frac{1}{12}\sum_{\ell=1}^{N} h_\ell^2 \langle \partial_t u_h, \partial_t w_h\rangle_{L^2(\tau_\ell)} - \int_0^T \partial_t u_h(t)\, I_h \int_T^t w_h(s) ds\, dt$$

$$= \frac{1}{12}\sum_{\ell=1}^{N} h_\ell^2 \langle \partial_t u_h, \partial_t w_h\rangle_{L^2(\tau_\ell)} + \int_0^T u_h(t)\, \partial_t I_h \int_T^t w_h(s) ds\, dt$$

$$= \frac{1}{12}\sum_{\ell=1}^{N} h_\ell^2 \langle \partial_t u_h, \partial_t w_h\rangle_{L^2(\tau_\ell)} + \langle u_h, Q_h^0 w_h\rangle_{L^2(0,T)},$$

i.e. the representation (17.11). □

Now we are in a position to find an alternative representation of the bilinear form $a(\cdot,\cdot)$.

Corollary 17.1 *For $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ and $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ we have*

$$\begin{aligned} a(u_h, w_h) &= -\langle \partial_t u_h, \partial_t w_h\rangle_{L^2(0,T)} + \mu \langle u_h, w_h\rangle_{L^2(0,T)} \\ &= -\langle \partial_t u_h, \partial_t w_h\rangle_{L^2(0,T)} + \sum_{\ell=1}^{N} \frac{\mu h_\ell^2}{12} \langle \partial_t u_h, \partial_t w_h\rangle_{L^2(\tau_\ell)} + \mu \langle u_h, Q_h^0 w_h\rangle_{L^2(0,T)} \\ &= \sum_{\ell=1}^{N} \left(\frac{\mu h_\ell^2}{12} - 1 \right) \langle \partial_t u_h, \partial_t w_h\rangle_{L^2(\tau_\ell)} + \mu \langle u_h, Q_h^0 w_h\rangle_{L^2(0,T)}. \end{aligned} \tag{17.12}$$

Motivated by the representation (17.12) we now define the perturbed bilinear form

$$a_h(u_h, w_h) := -\langle \partial_t u_h, \partial_t w_h\rangle_{L^2(0,T)} + \mu \langle u_h, Q_h^0 w_h\rangle_{L^2(0,T)} \tag{17.13}$$

for $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ and $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$, and we consider the perturbed variational formulation to find $\widetilde{u}_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ such that

$$a_h(\widetilde{u}_h, w_h) = \langle f, w_h\rangle_{(0,T)} \tag{17.14}$$

is satisfied for all $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$.

Lemma 17.3 *The perturbed bilinear form (17.13) is bounded, i.e. we have*

$$\left| a_h(u_h, w_h) \right| \le \left(1 + \frac{1}{2}\mu T^2 \right) |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)}$$

for all $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$, $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$.

Proof With the Cauchy–Schwarz inequality, the L^2 stability of Q_h^0, and with the Poincaré inequality we have

$$\begin{aligned}\left|a_h(u_h, w_h)\right| &\leq |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)} + \mu \, \|u_h\|_{L^2(0,T)} \|Q_h^0 w_h\|_{L^2(0,T)} \\ &\leq \left(1 + \frac{1}{2}\mu T^2\right) |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)}\end{aligned}$$

for $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$, $w_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$, and so the assertion. □

To prove a discrete stability condition for the perturbed bilinear form (17.13) we need the following lemma which is analogous to [23, Theorem 2.1].

Lemma 17.4 *For a given $z_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ represented by*

$$z_h(t) = \sum_{i=0}^{N} z_i \varphi_i(t) \quad \text{with } z_N = 0$$

and a fixed index $j \in \{0, \ldots, N-1\}$ there exists a function $\overline{v}_h^j \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ with the following properties:

i. For $t \in [0, t_j]$ we have $\overline{v}_h^j(t) = 0$.
ii. For $\ell = j+1, \ldots, N$ we have

$$\langle \partial_t \overline{v}_h^j, \partial_t z_h \rangle_{L^2(\tau_\ell)} = \frac{1}{2}\left(z_\ell^2 - z_{\ell-1}^2\right)$$

as well as

$$\langle \overline{v}_h^j, Q_h^0 z_h \rangle_{L^2(\tau_\ell)} = \frac{1}{2}\left(\int_{t_j}^{t_\ell} z_h(s)ds\right)^2 - \frac{1}{2}\left(\int_{t_j}^{t_{\ell-1}} z_h(s)ds\right)^2 .$$

iii. There holds the estimate

$$|\overline{v}_h^j|_{H^1(0,T)} \leq \|z_h\|_{L^2(0,T)} .$$

Proof For $z_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ we consider the piecewise linear interpolation of the antiderivative, i.e. for $t \in [0,T]$ we define

$$\overline{v}_h(t) := \sum_{k=0}^{N} \left(\int_0^{t_k} z_h(s)ds\right) \varphi_k(t) = I_h \int_0^t z_h(s)ds \, , \quad \overline{v}_h \in S_h^1(0,T) \cap H_{0,}^1(0,T).$$

From (17.10) the relation $\partial_t \overline{v}_h = Q_h^0 z_h$ follows. For a fixed index $j \in \{0, \dots, N-1\}$ we now define

$$z_h^j(t) = \sum_{i=0}^{N} z_i^j \varphi_i(t), \quad z_i^j = \begin{cases} (-1)^{j-i} z_j & \text{for } i = 0, \dots, j, \\ z_i & \text{for } i = j+1, \dots, N. \end{cases}$$

Note that $z_h^j \in S_h^1(0,T) \cap H_{,0}^1(0,T)$, and according to z_h^j we introduce $\overline{v}_h^j$ satisfying $\partial_t \overline{v}_h^j = Q_h^0 z_h^j$. In particular for $j > 0$ and $t \in \tau_\ell$ for $\ell = 1, \dots, j$ we then have

$$\partial_t \overline{v}_h^j(t) = Q_h^0 z_h^j(t) = \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} z_h^j(s)\, ds = \frac{1}{2}\left(z_{\ell-1}^j + z_\ell^j\right) = 0,$$

and due to $\overline{v}_h^j(0) = 0$ we conclude $\overline{v}_h^j(t) = 0$ for $t \in [0, t_j]$, i.e. *i.*

To prove *ii.* we compute for $\ell = j+1, \dots, N$

$$\begin{aligned} \langle \partial_t \overline{v}_h^j, \partial_t z_h \rangle_{L^2(\tau_\ell)} &= \langle Q_h^0 z_h^j, \partial_t z_h \rangle_{L^2(\tau_\ell)} \\ &= \frac{1}{2}\left(z_{\ell-1}^j + z_\ell^j\right)\left(z_\ell - z_{\ell-1}\right) \\ &= \frac{1}{2}\left(z_{\ell-1} + z_\ell\right)\left(z_\ell - z_{\ell-1}\right) = \frac{1}{2}\left(z_\ell^2 - z_{\ell-1}^2\right) \end{aligned}$$

as well as

$$\begin{aligned} \langle \overline{v}_h^j, Q_h^0 z_h \rangle_{L^2(\tau_\ell)} &= \int_{t_{\ell-1}}^{t_\ell} I_h \int_0^t z_h^j(s) ds\; Q_h^0 z_h(t)\, dt \\ &= Q_h^0 z_{h|\tau_\ell} \int_{t_{\ell-1}}^{t_\ell} \left[\int_0^{t_{\ell-1}} z_h^j(s) ds\, \varphi_{\ell-1}(t) + \int_0^{t_\ell} z_h^j(s) ds\, \varphi_\ell(t) \right] dt \\ &= \frac{1}{h_\ell} \int_{t_{\ell-1}}^{t_\ell} z_h(s) ds\; \frac{1}{2} h_\ell \left[\int_0^{t_{\ell-1}} z_h^j(s) ds + \int_0^{t_\ell} z_h^j(s) ds \right] \\ &= \frac{1}{2} \int_{t_{\ell-1}}^{t_\ell} z_h(s) ds \left[\int_{t_j}^{t_{\ell-1}} z_h(s) ds + \int_{t_j}^{t_\ell} z_h(s) ds \right] \\ &= \frac{1}{2} \left(\int_{t_{\ell-1}}^{t_\ell} z_h(s)\, ds \right)^2 + \int_{t_{\ell-1}}^{t_\ell} z_h(s)\, ds \int_{t_j}^{t_{\ell-1}} z_h(s)\, ds \\ &= \frac{1}{2} \left(\int_{t_{\ell-1}}^{t_\ell} z_h(s)\, ds + \int_{t_j}^{t_{\ell-1}} z_h(s)\, ds \right)^2 - \frac{1}{2} \left(\int_{t_j}^{t_{\ell-1}} z_h(s)\, ds \right)^2 \\ &= \frac{1}{2} \left(\int_{t_j}^{t_\ell} z_h(s)\, ds \right)^2 - \frac{1}{2} \left(\int_{t_j}^{t_{\ell-1}} z_h(s) ds \right)^2. \end{aligned}$$

From the L^2 stability of Q_h^0 we finally conclude the third assertion, i.e.

$$|\overline{v}_h^j|_{H^1(0,T)} = |\overline{v}_h^j|_{H^1(t_j,T)} = \|Q_h^0 z_h^j\|_{L^2(t_j,T)}$$
$$= \|Q_h^0 z_h\|_{L^2(t_j,T)} \le \|Q_h^0 z_h\|_{L^2(0,T)} \le \|z_h\|_{L^2(0,T)}.$$

□

Lemma 17.5 *The variational formulation to find $z_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ such that*

$$a_h(v_h, z_h) = \langle g, \partial_t v_h \rangle_{L^2(0,T)} \tag{17.15}$$

is satisfied for all $v_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ is uniquely solvable, where $g \in L^2(0,T)$ is given. Moreover, the stability estimate

$$\|z_h\|_{L^2(0,T)} \le 2T \, \|g\|_{L^2(0,T)} \tag{17.16}$$

holds for any mesh with maximal mesh size h.

Proof The finite element stiffness matrix of the variational problem (17.15) is upper triangular with positive diagonal elements and hence, there exists a unique solution $z_h \in S_h^1(0,T) \cap H_{,0}^1(0,T)$ of (17.15).

For the stability estimate we consider for an index $j \in \{0, \dots, N-1\}$ the function $\overline{v}_h^j \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ as given in Lemma 17.4. Plugging $\overline{v}_h^j$ into (17.15) and by using the properties of Lemma 17.4 this gives

$$\langle g, \partial_t \overline{v}_h^j \rangle_{L^2(0,T)} = a_h(\overline{v}_h^j, z_h)$$
$$= -\langle \partial_t \overline{v}_h^j, \partial_t z_h \rangle_{L^2(0,T)} + \mu \langle \overline{v}_h^j, Q_h^0 z_h \rangle_{L^2(0,T)}$$
$$= -\sum_{\ell=j+1}^{N} \langle \partial_t \overline{v}_h^j, \partial_t z_h \rangle_{L^2(\tau_\ell)} + \mu \sum_{\ell=j+1}^{N} \langle \overline{v}_h^j, Q_h^0 z_h \rangle_{L^2(\tau_\ell)}$$
$$= -\frac{1}{2} \sum_{\ell=j+1}^{N} \left(z_\ell^2 - z_{\ell-1}^2 \right) + \frac{\mu}{2} \sum_{\ell=j+1}^{N} \left(\left(\int_{t_j}^{t_\ell} z_h(s) ds \right)^2 - \left(\int_{t_j}^{t_{\ell-1}} z_h(s) ds \right)^2 \right)$$
$$= \frac{1}{2} z_j^2 + \frac{\mu}{2} \left(\int_{t_j}^{T} z_h(s) ds \right)^2 .$$

This result yields, with the Cauchy–Schwarz inequality, and the use of the properties of Lemma 17.4,

$$\begin{aligned}
\|z_h\|^2_{L^2(0,T)} &= \sum_{\ell=1}^{N} \|z_h\|^2_{L^2(\tau_\ell)} = \sum_{\ell=1}^{N} \frac{h_\ell}{3}\Big(z_\ell^2 + z_\ell z_{\ell-1} + z_{\ell-1}^2\Big) \le \frac{1}{2}\sum_{\ell=1}^{N} h_\ell\Big(z_\ell^2 + z_{\ell-1}^2\Big) \\
&\le \frac{1}{2}\sum_{j=1}^{N-1} h_j z_j^2 + \frac{1}{2}\sum_{j=0}^{N-1} h_{j+1} z_j^2 \\
&\le \sum_{j=1}^{N-1} h_j \,\langle g, \partial_t \overline{v}_h^j\rangle_{L^2(0,T)} + \sum_{j=0}^{N-1} h_{j+1}\,\langle g, \partial_t \overline{v}_h^j\rangle_{L^2(0,T)} \\
&\le \sum_{j=1}^{N-1} h_j \|g\|_{L^2(0,T)} |\overline{v}_h^j|_{H^1(0,T)} + \sum_{j=0}^{N-1} h_{j+1}\|g\|_{L^2(0,T)} |\overline{v}_h^j|_{H^1(0,T)} \\
&\le 2T\,\|g\|_{L^2(0,T)}\|z_h\|_{L^2(0,T)},
\end{aligned}$$

i.e. the assertion. □

Lemma 17.6 *For each $u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ there holds the discrete inf–sup condition*

$$\frac{1}{1+\sqrt{2}\mu T^2}\,|u_h|_{H^1(0,T)} \le \sup_{0\neq w_h \in S_h^1(0,T)\cap H_{,0}^1(0,T)} \frac{|a_h(u_h, w_h)|}{|w_h|_{H^1(0,T)}}\,.$$

Proof For a fixed function $u_h \in S_h^1(0,T)\cap H_{0,}^1(0,T)$ let $w_h \in S_h^1(0,T)\cap H_{,0}^1(0,T)$ be the unique solution of (17.15) for $g := \partial_t u_h \in L^2(0,T)$, i.e. we have

$$a_h(v_h, w_h) = \langle \partial_t u_h, \partial_t v_h\rangle_{L^2(0,T)} \tag{17.17}$$

for all $v_h \in S_h^1(0,T)\cap H_{0,}^1(0,T)$. For the particular choice $v_h(t) = w_h(0) - w_h(t)$ with $v_h \in S_h^1(0,T)\cap H_{0,}^1(0,T)$ we obtain

$$\langle \partial_t w_h, \partial_t w_h\rangle_{L^2(0,T)} - \mu\langle w_h - w_h(0), Q_h^0 w_h\rangle_{L^2(0,T)} = -\langle \partial_t u_h, \partial_t w_h\rangle_{L^2(0,T)},$$

and hence we conclude, by using the Cauchy–Schwarz and Poincaré inequalities, and the L^2 stability of the L^2 projection Q_h^0,

$$\begin{aligned}
|w_h|^2_{H^1(0,T)} &= -\langle \partial_t u_h, \partial_t w_h\rangle_{L^2(0,T)} + \mu\,\langle w_h - w_h(0), Q_h^0 w_h\rangle_{L^2(0,T)} \\
&\le |u_h|_{H^1(0,T)}|w_h|_{H^1(0,T)} + \mu\,\|w_h - w_h(0)\|_{L^2(0,T)}\|Q_h^0 w_h\|_{L^2(0,T)}
\end{aligned}$$

$$\leq |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)} + \frac{1}{\sqrt{2}} \mu T |w_h|_{H^1(0,T)} \|w_h\|_{L^2(0,T)}$$
$$\leq \left(1 + \sqrt{2}\mu T^2\right) |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)},$$

where in the last step we used the stability estimate (17.16).

The choice $v_h = u_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ in (17.17) and the estimate above yield

$$a_h(u_h, w_h) = |u_h|^2_{H^1(0,T)} \geq \frac{1}{1 + \sqrt{2}\mu T^2} |u_h|_{H^1(0,T)} |w_h|_{H^1(0,T)}$$

and hence the discrete inf–sup condition follows. □

Theorem 17.1 *Let the unique solution u of (17.6) satisfy $u \in H^s(0,T)$ for some $s \in [1,2]$. There exists a unique solution $\widetilde{u}_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ of the Galerkin–Petrov finite element discretization (17.14) satisfying*

$$|u - \widetilde{u}_h|_{H^1(0,T)} \leq \left[1 + \left(1 + \frac{1}{2}\mu T^2\right)\left(1 + \sqrt{2}\mu T^2\right)\right] C_1(s)\, h^{s-1} |u|_{H^s(0,T)}$$
$$+ \frac{1}{12}\mu \left(1 + \sqrt{2}\mu T^2\right) h^2\, C_2\, |u|_{H^1(0,T)},$$

where the constants $C_1, C_2 > 0$ are coming from standard interpolation error and stability estimates.

Proof For any $v_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)$ we first have

$$|u - \widetilde{u}_h|_{H^1(0,T)} \leq |u - v_h|_{H^1(0,T)} + |\widetilde{u}_h - v_h|_{H^1(0,T)}$$

and it remains to bound the second term. With the discrete inf–sup condition in Lemma 17.6 and using the Galerkin orthogonality for the variational formulations (17.6) and (17.14), we first have

$$\frac{1}{1 + \sqrt{2}\mu T^2} |\widetilde{u}_h - v_h|_{H^1(0,T)} \leq \sup_{0 \neq w_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)} \frac{|a_h(\widetilde{u}_h - v_h, w_h)|}{|w_h|_{H^1(0,T)}}$$
$$= \sup_{0 \neq w_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)} \frac{|a_h(\widetilde{u}_h, w_h) - a_h(v_h, w_h)|}{|w_h|_{H^1(0,T)}}$$
$$= \sup_{0 \neq w_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)} \frac{|a(u, w_h) - a_h(v_h, w_h)|}{|w_h|_{H^1(0,T)}}$$
$$= \sup_{0 \neq w_h \in S_h^1(0,T) \cap H_{0,}^1(0,T)} \frac{|a(u - v_h, w_h) + a(v_h, w_h) - a_h(v_h, w_h)|}{|w_h|_{H^1(0,T)}}.$$

Now, with the boundedness of the bilinear form $a(\cdot,\cdot)$ and the Poincaré inequality we further conclude

$$\begin{aligned} a(u-v_h,w_h) &= -\langle \partial_t(u-v_h),\partial_t w_h\rangle_{L^2(0,T)} + \mu\langle u-v_h,w_h\rangle_{L^2(0,T)} \\ &\le |u-v_h|_{H^1(0,T)}|w_h|_{H^1(0,T)} + \mu\|u-v_h\|_{L^2(0,T)}\|w_h\|_{L^2(0,T)} \\ &\le \Big(1+\frac{1}{2}\mu T^2\Big)|u-v_h|_{H^1(0,T)}|w_h|_{H^1(0,T)}\,. \end{aligned}$$

Moreover, using the representation (17.12) we can estimate the consistency error by

$$\begin{aligned} |a(v_h,w_h)-a_h(v_h,w_h)| &= \frac{1}{12}\mu\left|\sum_{\ell=1}^{N} h_\ell^2\langle\partial_t v_h,\partial_t w_h\rangle_{L^2(\tau_\ell)}\right| \\ &\le \frac{1}{12}\mu\, h^2\,|v_h|_{H^1(0,T)}|w_h|_{H^1(0,T)}. \end{aligned}$$

Hence we have

$$\frac{1}{1+\sqrt{2}\mu T^2}|\widetilde{u}_h-v_h|_{H^1(0,T)} \le \Big(1+\frac{1}{2}\mu T^2\Big)|u-v_h|_{H^1(0,T)} + \frac{1}{12}\mu\, h^2\,|v_h|_{H^1(0,T)}\,,$$

and therefore

$$\begin{aligned} |u-\widetilde{u}_h|_{H^1(0,T)} \le \left[1+\Big(1+\frac{1}{2}\mu T^2\Big)\Big(1+\sqrt{2}\mu T^2\Big)\right]|u-v_h|_{H^1(0,T)} \\ +\frac{1}{12}\mu\,\Big(1+\sqrt{2}\mu T^2\Big)h^2\,|v_h|_{H^1(0,T)} \end{aligned}$$

follows. In particular for the piecewise linear nodal interpolation $v_h = I_h u$ we have

$$\|u-I_h u\|_{H^1(0,T)} \le C_1(s)\,h^{s-1}\,|u|_{H^s(0,T)}, \quad \|I_h u\|_{H^1(0,T)} \le C_2\,|u|_{H^1(0,T)}\,.$$

□

As numerical example for the Galerkin finite element methods (17.8) and (17.14) we consider a uniform discretization of the time interval $(0,T)$ with $T=10$ and a mesh size $h=T/N$. For $\mu=1000$ we consider the solution $u(t)=\sin^2\left(\frac{5}{4}\pi t\right)$ and we compute the appearing integrals for the related right–hand side in (17.8) and (17.14) by the usage of high order integration rules.

In Table 17.1 we present the results for the stabilized variational formulation (17.14) which is unconditionally stable, and where the error estimate in the energy norm of Theorem 17.1 is confirmed. In addition we also present the error in $L^2(0,T)$ where we observe a second order convergence, as expected. But at this

Table 17.1 Numerical results for the stabilized variational formulation (17.14), $\mu = 1000$, $T = 10$

N	h	$\|u - \widetilde{u}_h\|_{L^2(0,10)}$	eoc	$\|u - \widetilde{u}_h\|_{H^1(0,10)}$	eoc
4	2.5000000	1.7722e+00	0.00	9.0867e+00	0.00
8	1.2500000	6.0704e+00	−1.78	2.0130e+01	−1.15
16	0.6250000	1.2687e+00	2.26	9.4204e+00	1.10
32	0.3125000	5.7861e+00	−2.19	6.0121e+01	−2.67
64	0.1562500	3.3966e−01	4.09	6.1941e+00	3.28
128	0.0781250	7.6647e−02	2.15	2.2955e+00	1.43
256	0.0390625	2.0315e−02	1.92	9.4091e−01	1.29
512	0.0195312	5.2649e−03	1.95	4.1539e−01	1.18
1024	0.0097656	1.3365e−03	1.98	1.9803e−01	1.07
2048	0.0048828	3.3682e−04	1.99	9.7671e−02	1.02
4096	0.0024414	8.4229e−05	2.00	4.8663e−02	1.01
8192	0.0012207	2.1057e−05	2.00	2.4310e−02	1.00
16,384	0.0006104	5.2644e−06	2.00	1.2152e−02	1.00
32,768	0.0003052	1.3161e−06	2.00	6.0758e−03	1.00

Table 17.2 Numerical results for the variational formulation (17.8), $\mu = 1000$, $T = 10$

N	h	$\|u - u_h\|_{L^2(0,10)}$	eoc	$\|u - u_h\|_{H^1(0,10)}$	eoc
4	2.5000000	7.0573e+01	0.00	9.8785e+01	0.00
8	1.2500000	1.6871e+03	−4.58	3.7166e+03	−5.23
16	0.6250000	9.1421e+07	−15.73	3.7247e+08	−16.61
32	0.3125000	2.3915e+15	−24.64	1.9496e+16	−25.64
64	0.1562500	1.6337e+22	−22.70	2.9536e+23	−23.85
128	0.0781250	3.1417e−02	78.78	1.7859e+00	77.13
256	0.0390625	9.2885e−03	1.76	8.2361e−01	1.12
512	0.0195312	2.4767e−03	1.91	3.9567e−01	1.06
1024	0.0097656	6.3105e−04	1.97	1.9532e−01	1.02
2048	0.0048828	1.5839e−04	1.99	9.7325e−02	1.00
4096	0.0024414	3.9633e−05	2.00	4.8620e−02	1.00
8192	0.0012207	9.9106e−06	2.00	2.4304e−02	1.00
16,384	0.0006104	2.4778e−06	2.00	1.2152e−02	1.00
32,768	0.0003052	6.1946e−07	2.00	6.0757e−03	1.00

point we do not include any further discussion of error estimates in $L^2(0, T)$ since this is behind the scope of this contribution.

In Table 17.2 we present the related results for the variational formulation (17.8) without stabilization. We observe that we have convergence for a sufficiently small mesh size only. Note that $\sqrt{12/\mu} \approx 0.1095$.

17.3 Wave Equation

Instead of the ordinary differential equation (17.5) we now consider the wave equation (17.1) and the related variational formulation (17.2), and we aim to extend the results of Sect. 17.2. For $u \in H^{1,1}_{0;0,}(Q)$ and $w \in H^{1,1}_{0;,0}(Q)$ we define the bilinear form

$$a(u, w) := -\langle \partial_t u, \partial_t w \rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x w \rangle_{L^2(Q)} .$$

The Hilbert spaces $H^{1,1}_{0;0,}(Q)$ and $H^{1,1}_{0;,0}(Q)$, defined in Sect. 17.1, are endowed with the inner product

$$\langle u, v \rangle_{H^{1,1}_{0;0,}(Q)} = \langle u, v \rangle_{H^{1,1}_{0;,0}(Q)} := \int_0^T \int_\Omega \Big[\partial_t u(x,t) \partial_t v(x,t) + \nabla_x u(x,t) \cdot \nabla_x v(x,t) \Big] dx\, dt,$$

where the induced norm

$$|u|^2_{H^1(Q)} := \int_0^T \int_\Omega \Big[|\partial_t u(x,t)|^2 + |\nabla_x u(x,t)|^2 \Big] dx\, dt$$

is well–defined due to Poincaré inequalities with respect to space and time. As in [9] we have unique solvability of (17.2) when assuming $f \in L^2(Q)$, in particular we have, see [18],

$$|u|_{H^1(Q)} \leq \frac{1}{\sqrt{2}}\, T\, \|f\|_{L^2(Q)}.$$

Next, we examine a conforming finite element discretization for the variational formulation (17.2) in the case where $\Omega = (0, L)$ is an interval for $d = 1$, or Ω is polygonal for $d = 2$, or Ω is polyhedral for $d = 3$. For a tensor–product ansatz we consider a sequence $(\mathscr{T}_N)_{N \in \mathbb{N}}$ of admissible decompositions

$$\overline{Q} = \overline{\mathscr{T}_N} = \overline{\Omega} \times [0, T] = \bigcup_{i=1}^{N_x} \overline{\omega}_i \times \bigcup_{\ell=1}^{N_t} \overline{\tau}_\ell$$

with $N := N_x \cdot N_t$ space–time elements, where the time intervals $\tau_\ell = (t_{\ell-1}, t_\ell)$ with mesh size $h_{t,\ell}$ are defined via the decomposition

$$0 = t_0 < t_1 < t_2 < \cdots < t_{N_t - 1} < t_{N_t} = T$$

of the time interval $(0, T)$. For the spatial domain Ω we consider an admissible and globally quasi–uniform decomposition into finite elements ω_i with mesh size $h_{x,i}$ which can be represented by using standard maps with respect to related reference elements. Here, the spatial elements ω_i are intervals for $d = 1$, triangles or

quadrilaterals for $d = 2$, and tetrahedra or hexahedra for $d = 3$. Next, we consider the finite element space

$$Q_h^1(Q) := V_{h_x}(\Omega) \otimes S_{h_t}^1(0,T)$$

of piecewise multilinear continuous functions, i.e.

$$V_{h_x}(\Omega) = \operatorname{span}\{\psi_j\}_{j=1}^{M_x} \subset H_0^1(\Omega), \quad S_{h_t}^1(0,T) = \operatorname{span}\{\varphi_\ell\}_{\ell=0}^{N_t} \subset H^1(0,T).$$

In fact, $V_{h_x}(\Omega)$ is either the space $S_{h_x}^1(\Omega)$ of piecewise linear continuous functions on intervals ($d = 1$), triangles ($d = 2$), and tetrahedra ($d = 3$), or $V_{h_x}(\Omega)$ is the space $Q_{h_x}^1(\Omega)$ of piecewise linear/bilinear/trilinear continuous functions on intervals ($d = 1$), quadrilaterals ($d = 2$), and hexahedra ($d = 3$). A finite element function $u_h \in Q_h^1(Q)$ admits the representation

$$u_h(x,t) = \sum_{\ell=0}^{N_t} \sum_{j=1}^{M_x} u_j^\ell \psi_j(x) \varphi_\ell(t) \quad \text{for } (x,t) \in Q. \tag{17.18}$$

The Galerkin–Petrov finite element discretization of the variational formulation (17.2) is to find $u_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ such that

$$a(u_h, w_h) = -\langle \partial_t u_h, \partial_t w_h \rangle_{L^2(Q)} + \langle \nabla_x u_h, \nabla_x w_h \rangle_{L^2(Q)} = \langle f, w_h \rangle_Q \tag{17.19}$$

is satisfied for all $w_h \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$. After an appropriate ordering of the degrees of freedom, the discrete variational formulation (17.19) is equivalent to the linear system $K_h \underline{u} = \underline{f}$ with the system matrix

$$K_h := -M_{h_x} \otimes A_{h_t} + A_{h_x} \otimes M_{h_t} \in \mathbb{R}^{M_x \cdot N_t \times M_x \cdot N_t}, \tag{17.20}$$

where M_{h_x}, $A_{h_x} \in \mathbb{R}^{M_x \times M_x}$ are the mass and stiffness matrix with respect to space, and M_{h_t}, $A_{h_t} \in \mathbb{R}^{N_t \times N_t}$ are the mass and stiffness matrix with respect to time, respectively.

Remark 17.1 With the help of a von Neumann type stability analysis [5] for the matrix (17.20) of the Galerkin–Petrov finite element method (17.19) for a uniform discretization in time with mesh size h_t and a uniform discretization in space with mesh size h_x for piecewise linear/bilinear/trilinear continuous functions on intervals ($d = 1$), squares ($d = 2$), or cubes ($d = 3$) we can show stability of the corresponding numerical scheme, if the condition

$$h_t \le \frac{1}{\sqrt{d}} h_x$$

is satisfied, see [18].

From Remark 17.1 we conclude that we have only conditional stability of (17.19). To stabilize the numerical scheme in (17.19) we use as in (17.12) again Zlotnik's idea [23] and we prove the following representation.

Lemma 17.7 *For all $u_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ and $w_h \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$ the bilinear form in (17.19) has the representation*

$$a(u_h, w_h) = -\langle \partial_t u_h, \partial_t w_h \rangle_{L^2(Q)} + \sum_{m=1}^{d} \langle \partial_{x_m} u_h, Q_{h_t}^0 \partial_{x_m} w_h \rangle_{L^2(Q)} \tag{17.21}$$
$$+ \sum_{m=1}^{d} \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}^2}{12} \langle \partial_t \partial_{x_m} u_h, \partial_t \partial_{x_m} w_h \rangle_{L^2(\Omega \times \tau_\ell)},$$

where $Q_{h_t}^0 : L^2(0,T) \to S_{h_t}^0(0,T)$ denotes the L^2 projection with respect to time on the space $S_{h_t}^0(0,T)$ of piecewise constant functions.

Proof Let $u_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ and $w_h \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$ be given. With the representation (17.18) we have for $(x,t) \in Q$

$$u_h(x,t) = \sum_{\ell=1}^{N_t} \sum_{j=1}^{M_x} u_j^\ell \psi_j(x) \varphi_\ell(t) = \sum_{j=1}^{M_x} U_{j,h}(t) \psi_j(x), \quad U_{j,h}(t) = \sum_{\ell=1}^{N_t} u_j^\ell \varphi_\ell(t),$$

as well as

$$w_h(x,t) = \sum_{\ell=0}^{N_t-1} \sum_{j=1}^{M_x} w_j^\ell \psi_j(x) \varphi_\ell(t) = \sum_{j=1}^{M_x} W_{j,h}(t) \psi_j(x), \quad W_{j,h}(t) = \sum_{\ell=0}^{N_t-1} w_j^\ell \varphi_\ell(t).$$

Hence we have, for $m = 1, \ldots, d$, and by using (17.11),

$$\langle \partial_{x_m} u_h, \partial_{x_m} w_h \rangle_{L^2(Q)} = \sum_{i=1}^{M_x} \sum_{j=1}^{M_x} \int_0^T U_{i,h}(t) W_{j,h}(t) dt \int_\Omega \partial_{x_m} \psi_i(x) \partial_{x_m} \psi_j(x)\, dx$$
$$= \sum_{i=1}^{M_x} \sum_{j=1}^{M_x} \left[\frac{1}{12} \sum_{\ell=1}^{N_t} h_{t,\ell}^2 \langle \partial_t U_{i,h}, \partial_t W_{j,h} \rangle_{L^2(\tau_\ell)} + \langle U_{i,h}, Q_{h_t}^0 W_{j,h} \rangle_{L^2(0,T)} \right]$$
$$\cdot \int_\Omega \partial_{x_m} \psi_i(x) \partial_{x_m} \psi_j(x)\, dx$$
$$= \langle \partial_{x_m} u_h, Q_{h_t}^0 \partial_{x_m} w_h \rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}^2}{12} \langle \partial_t \partial_{x_m} u_h, \partial_t \partial_{x_m} w_h \rangle_{L^2(\Omega \times \tau_\ell)}.$$

□

Due to the representation (17.21) we define for $u_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ and $w_h \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$ the perturbed bilinear form

$$\begin{aligned} a_h(u_h, w_h) &:= -\langle \partial_t u_h, \partial_t w_h \rangle_{L^2(Q)} + \sum_{m=1}^{d} \langle \partial_{x_m} u_h, Q_{h_t}^0 \partial_{x_m} w_h \rangle_{L^2(Q)} \\ &= -\langle \partial_t u_h, \partial_t w_h \rangle_{L^2(Q)} + \sum_{m=1}^{d} \langle Q_{h_t}^0 \partial_{x_m} u_h, \partial_{x_m} w_h \rangle_{L^2(Q)}, \end{aligned}$$

and we consider the perturbed variational problem to find $\widetilde{u}_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ such that

$$a_h(\widetilde{u}_h, w_h) = \langle f, w_h \rangle_Q \tag{17.22}$$

is satisfied for all $w_h \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$.

To prove the existence and uniqueness of a solution $\widetilde{u}_h$ of (17.22) we show the following lemma, which is analogous to Lemma 17.4.

Lemma 17.8 *For a given $v_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ represented by*

$$v_h(x,t) = \sum_{\ell=0}^{N_t} V_{\ell,h}(x)\varphi_\ell(t), \quad V_{\ell,h}(x) = \sum_{j=1}^{M_x} v_j^\ell \psi_j(x) \quad \text{for } (x,t) \in Q$$

with $V_{0,h}(x) = 0$ for $x \in \Omega$ and $V_{\ell,h} \in V_{h_x}(\Omega)$, and for a fixed index $j \in \{1, \dots, N_t\}$ there exists a function $\overline{z}_h^j \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$ with the following properties:

i. For $(x,t) \in \overline{\Omega} \times [t_j, T]$ we have $\overline{z}_h^j(x,t) = 0$.

ii. For $\ell = 1, \dots, j$ and for $x \in \overline{\Omega}$ we have

$$\langle \partial_t \overline{z}_h^j(x,\cdot), \partial_t v_h(x,\cdot) \rangle_{L^2(\tau_\ell)} = \frac{1}{2}\Big([V_{\ell-1,h}(x)]^2 - [V_{\ell,h}(x)]^2 \Big),$$

and for $m = 1, \dots, d$

$$\begin{aligned} &\langle \partial_{x_m} \overline{z}_h^j(x,\cdot), Q_{h_t}^0 \partial_{x_m} v_h(x,\cdot) \rangle_{L^2(\tau_\ell)} = \\ &\qquad = \frac{1}{2}\left(\int_{t_{\ell-1}}^{t_j} \partial_{x_m} v_h(x,s) ds \right)^2 - \frac{1}{2}\left(\int_{t_\ell}^{t_j} \partial_{x_m} v_h(x,s) ds \right)^2. \end{aligned}$$

iii. For $x \in \overline{\Omega}$ there holds the estimate

$$\|\partial_t \overline{z}_h^j(x,\cdot)\|_{L^2(0,T)} \le \|v_h(x,\cdot)\|_{L^2(0,T)}\,.$$

Proof For $v_h \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ we define

$$u_h^j(x,t) = \sum_{i=0}^{N_t} U_{i,h}^j(x)\varphi_i(t), \quad U_{i,h}^j(x) := \begin{cases} V_{i,h}(x) & \text{for } i = 0,\dots,j, \\ (-1)^{j-i} V_{j,h}(x) & \text{for } i = j+1,\dots,N_t, \end{cases}$$

and for $(x,t) \in Q$ we set

$$\overline{z}_h^j(x,t) := -\sum_{k=0}^{N_t} \int_T^{t_k} u_h^j(x,s)\,ds\varphi_k(t) = -I_{h_t} \int_T^t u_h^j(x,s)\,ds,$$

where $I_{h_t}\colon C[0,T] \to S_{h_t}^1(0,T)$ is the interpolation operator with respect to time. Note that $\overline{z}_h^j \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$. For $x \in \Omega$ it follows from relation (17.10) that

$$\partial_t \overline{z}_h^j(x,\cdot) = -Q_{h_t}^0 u_h^j(x,\cdot).$$

In particular for $j < N_t$, $x \in \overline{\Omega}$, and for $t \in \tau_\ell$ for $\ell = j+1,\dots,N_t$ we then have

$$-\partial_t \overline{z}_h^j(x,t) = Q_{h_t}^0 u_h^j(x,t) = \frac{1}{h_{t,\ell}} \int_{t_{\ell-1}}^{t_\ell} u_h^j(x,s)\,ds = \frac{1}{2}\left(U_{\ell-1,h}^j(x) + U_{\ell,h}^j(x)\right) = 0,$$

and due to $\overline{z}_h^j(x,T) = 0$ we conclude $\overline{z}_h^j(x,t) = 0$ for $t \in [t_j, T]$, i.e. *i.*

To prove *ii.* we first compute for $x \in \overline{\Omega}$ and for $\ell = 1,\dots,j$

$$\begin{aligned}
\langle \partial_t \overline{z}_h^j(x,\cdot),\, \partial_t v_h(x,\cdot)\rangle_{L^2(\tau_\ell)} &= \int_{t_{\ell-1}}^{t_\ell} \partial_t \overline{z}_h^j(x,t)\,\partial_t v_h(x,t)\,dt \\
&= -\frac{1}{2}\left(U_{\ell-1,h}^j(x) + U_{\ell,h}^j(x)\right) \int_{t_{\ell-1}}^{t_\ell} \partial_t v_h(x,t)\,dt \\
&= -\frac{1}{2}\left(U_{\ell-1,h}^j(x) + U_{\ell,h}^j(x)\right)\left(V_{\ell,h}(x) - V_{\ell-1,h}(x)\right) \\
&= \frac{1}{2}\left(V_{\ell-1,h}(x) + V_{\ell,h}(x)\right)\left(V_{\ell-1,h}(x) - V_{\ell,h}(x)\right) \\
&= \frac{1}{2}\Big([V_{\ell-1,h}(x)]^2 - [V_{\ell,h}(x)]^2\Big).
\end{aligned}$$

Moreover, for $m = 1, \ldots, d$ we have for $x \in \Omega$ and for $\ell = 1, \ldots, j$

$$\begin{aligned}
&\langle \partial_{x_m} \overline{z}_h^j(x,\cdot),\, Q_{h_t}^0\, \partial_{x_m} v_h(x,\cdot)\rangle_{L^2(\tau_\ell)} = Q_{h_t}^0\, \partial_{x_m} v_h(x,\cdot)_{|\tau_\ell} \int_{t_{\ell-1}}^{t_\ell} \partial_{x_m} \overline{z}_h^j(x,t)\, dt \\
&= -Q_{h_t}^0\, \partial_{x_m} v_h(x,\cdot)_{|\tau_\ell} \int_{t_{\ell-1}}^{t_\ell} \partial_{x_m} \left[\int_T^{t_{\ell-1}} u_h^j(x,s) ds\, \varphi_{\ell-1}(t) + \int_T^{t_\ell} u_h^j(x,s) ds\, \varphi_\ell(t) \right] dt \\
&= -\frac{1}{h_{t,\ell}} \int_{t_{\ell-1}}^{t_\ell} \partial_{x_m} v_h(x,t) dt\, \frac{1}{2} h_{t,\ell} \left[\int_T^{t_{\ell-1}} \partial_{x_m} u_h^j(x,s) ds + \int_T^{t_\ell} \partial_{x_m} u_h^j(x,s) ds \right] \\
&= -\frac{1}{2} \left[\int_{t_j}^{t_\ell} \partial_{x_m} v_h(x,t) dt - \int_{t_j}^{t_{\ell-1}} \partial_{x_m} v_h(x,t) dt \right] \\
&\qquad\qquad \cdot \left[\int_{t_j}^{t_{\ell-1}} \partial_{x_m} v_h^j(x,s) ds + \int_{t_j}^{t_\ell} \partial_{x_m} v_h^j(x,s) ds \right] \\
&= \frac{1}{2} \left(\int_{t_j}^{t_{\ell-1}} \partial_{x_m} v_h^j(x,s) ds \right)^2 - \frac{1}{2} \left(\int_{t_j}^{t_\ell} \partial_{x_m} v_h^j(x,s) ds \right)^2 .
\end{aligned}$$

From the L^2 stability of $Q_{h_t}^0$ and by using

$$\begin{aligned}
\| \partial_t \overline{z}_h^j(x,\cdot) \|_{L^2(0,T)} &= \| \partial_t \overline{z}_h^j(x,\cdot) \|_{L^2(0,t_j)} = \| Q_{h_t}^0 u_h^j(x,\cdot) \|_{L^2(0,t_j)} \\
&= \| Q_{h_t}^0 v_h(x,\cdot) \|_{L^2(0,t_j)} \le \| Q_{h_t}^0 v_h(x,\cdot) \|_{L^2(0,T)} \le \| v_h(x,\cdot) \|_{L^2(0,T)}
\end{aligned}$$

for $x \in \overline{\Omega}$ we finally conclude the third property. □

With the last lemma we are now able to prove existence, uniqueness, and stability of a solution $\widetilde{u}_h$ of (17.22).

Lemma 17.9 *For $f_0 \in [H^1_{,0}(0,T; L^2(\Omega))]'$ and $f_1, f_2 \in L^2(Q)$ the variational formulation to find $w_h \in Q_h^1(Q) \cap H^{1,1}_{0;0,}(Q)$ such that*

$$a_h(w_h, v_h) = \langle f_0, v_h \rangle_Q + \langle f_1, \partial_t v_h \rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} h_{t,\ell}^2 \langle f_2, \partial_t v_h \rangle_{L^2(\Omega \times \tau_\ell)} \tag{17.23}$$

is satisfied for all $v_h \in Q_h^1(Q) \cap H^{1,1}_{0;,0}(Q)$ is uniquely solvable, and there holds the stability estimate

$$\| w_h \|_{L^2(Q)} \le 2\, T \left\{ \| f_0 \|_{[H^1_{,0}(0,T;L^2(\Omega))]'} + \| f_1 \|_{L^2(Q)} + h_t^2\, \| f_2 \|_{L^2(Q)} \right\}. \tag{17.24}$$

Proof Let $w_h^0 \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ be any solution of the homogeneous variational formulation (17.23) with $f_i \equiv 0$, and with the representation (17.18), i.e.

$$w_h^0(x,t) = \sum_{\ell=0}^{N_t} W_{\ell,h}^0(x)\varphi_\ell(t) \quad \text{for } (x,t) \in Q, \quad W_{\ell,h}^0 \in V_{h_x}(\Omega), \quad W_{0,h}^0(x) = 0.$$

For an index $j \in \{1,\dots,N_t\}$ we now consider an element $\overline{z}_h^j \in Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$ as given in Lemma 17.8. Plugging $\overline{z}_h^j$ into (17.23) and by using the properties of Lemma 17.8 this gives

$$\begin{aligned}
0 &= a_h(w_h^0, \overline{z}_h^j) \\
&= -\langle \partial_t w_h^0, \partial_t \overline{z}_h^j \rangle_{L^2(Q)} + \sum_{m=1}^{d} \langle Q_{h_t}^0 \partial_{x_m} w_h^0, \partial_{x_m} \overline{z}_h^j \rangle_{L^2(Q)} \\
&= -\sum_{\ell=1}^{j} \langle \partial_t w_h^0, \partial_t \overline{z}_h^j \rangle_{L^2(\Omega\times\tau_\ell)} + \sum_{m=1}^{d}\sum_{\ell=1}^{j} \langle Q_{h_t}^0 \partial_{x_m} w_h^0, \partial_{x_m} \overline{z}_h^j \rangle_{L^2(\Omega\times\tau_\ell)} \\
&= -\int_\Omega \sum_{\ell=1}^{j} \left(\frac{1}{2}[W_{\ell-1,h}^0(x)]^2 - \frac{1}{2}[W_{\ell,h}^0(x)]^2 \right) dx \\
&\quad + \sum_{m=1}^{d} \int_\Omega \sum_{\ell=1}^{j} \left(\frac{1}{2} \left(\int_{t_{\ell-1}}^{t_j} \partial_{x_m} w_h^0(x,s) ds \right)^2 - \frac{1}{2} \left(\int_{t_\ell}^{t_j} \partial_{x_m} w_h^0(x,s) ds \right)^2 \right) dx \\
&= \frac{1}{2} \int_\Omega [W_{j,h}^0(x)]^2 dx + \frac{1}{2} \sum_{m=1}^{d} \int_\Omega \left(\int_0^{t_j} \partial_{x_m} w_h^0(x,s) ds \right)^2 dx.
\end{aligned}$$

This result yields, with the Cauchy–Schwarz inequality and the use of the properties of Lemma 17.8,

$$\begin{aligned}
\|w_h^0\|_{L^2(Q)}^2 &= \sum_{\ell=1}^{N_t} \|w_h^0\|_{L^2(\Omega\times\tau_\ell)}^2 \\
&= \int_\Omega \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}}{3} \left([W_{\ell,h}^0(x)]^2 + W_{\ell,h}^0(x) W_{\ell-1,h}^0(x) + [W_{\ell-1,h}^0(x)]^2 \right) dx \\
&\le \int_\Omega \sum_{j=1}^{N_t} \frac{h_{t,j}}{2} [W_{j,h}^0(x)]^2 dx + \int_\Omega \sum_{j=1}^{N_t-1} \frac{h_{t,j+1}}{2} [W_{j,h}^0(x)]^2 dx \;\le\; 0,
\end{aligned}$$

which implies $w_h^0 \equiv 0$. By using

$$\dim Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q) = \dim Q_h^1(Q) \cap H_{0;,0}^{1,1}(Q)$$

we therefore conclude unique solvability of the variational formulation (17.23) for any right–hand side $f_0 \in [H_{,0}^1(0,T;L^2(\Omega))]'$ and $f_1,\, f_2 \in L^2(Q)$. Following the approach as above we then obtain

$$\langle f_0, \overline{z}_h^j\rangle_Q + \langle f_1, \partial_t \overline{z}_h^j\rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} h_{t,\ell}^2 \langle f_2, \partial_t \overline{z}_h^j\rangle_{L^2(\Omega\times\tau_\ell)}$$
$$= a_h(w_h, \overline{z}_h^j) \geq \frac{1}{2}\int_\Omega [W_{j,h}(x)]^2\, dx$$

and

$$\begin{aligned}
\|w_h\|_{L^2(Q)}^2 &\leq \int_\Omega \sum_{j=1}^{N_t} \frac{h_{t,j}}{2}[W_{j,h}(x)]^2\, dx + \int_\Omega \sum_{j=1}^{N_t-1} \frac{h_{t,j+1}}{2}[W_{j,h}(x)]^2\, dx \\
&\leq \sum_{j=1}^{N_t} h_{t,j} \left\{ \langle f_0, \overline{z}_h^j\rangle_Q + \langle f_1, \partial_t \overline{z}_h^j\rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} h_{t,\ell}^2 \langle f_2, \partial_t \overline{z}_h^j\rangle_{L^2(\Omega\times\tau_\ell)} \right\} \\
&\quad + \sum_{j=1}^{N_t-1} h_{t,j+1} \left\{ \langle f_0, \overline{z}_h^j\rangle_Q + \langle f_1, \partial_t \overline{z}_h^j\rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} h_{t,\ell}^2 \langle f_2, \partial_t \overline{z}_h^j\rangle_{L^2(\Omega\times\tau_\ell)} \right\} \\
&\leq \sum_{j=1}^{N_t} h_{t,j} \left\{ \|f_0\|_{[H_{,0}^1(0,T;L^2(\Omega))]'} + \|f_1\|_{L^2(Q)} + h_t^2 \|f_2\|_{L^2(Q)} \right\} \|\partial_t \overline{z}_h^j\|_{L^2(Q)} \\
&\quad + \sum_{j=1}^{N_t-1} h_{t,j+1} \left\{ \|f_0\|_{[H_{,0}^1(0,T;L^2(\Omega))]'} + \|f_1\|_{L^2(Q)} + h_t^2 \|f_2\|_{L^2(Q)} \right\} \|\partial_t \overline{z}_h^j\|_{L^2(Q)} \\
&\leq 2T \left\{ \|f_0\|_{[H_{,0}^1(0,T;L^2(\Omega))]'} + \|f_1\|_{L^2(Q)} + h_t^2 \|f_2\|_{L^2(Q)} \right\} \|w_h\|_{L^2(Q)},
\end{aligned}$$

and hence the stability estimate is proven. □

As a consequence of Lemma 17.9 we obtain unique solvability of the variational formulation (17.22), and the stability estimate

$$\|\widetilde{u}_h\|_{L^2(Q)} \leq 2\,T\, \|f\|_{[H_{,0}^1(0,T;L^2(\Omega))]'}.$$

To derive an estimate for the L^2 error $\|u - \widetilde{u}_h\|_{L^2(Q)}$ we introduce, as in [2, Section 2], a space–time projection $Q_{h_t}^1 Q_{h_x}^1 v \in Q_h^1(Q) \cap H_{0;0,}^{1,1}(Q)$ when $v \in H_{0;0,}^{1,1}(Q)$ is given. First, the H_0^1 projection $Q_{h_x}^1 : L^2(0,T;H_0^1(\Omega)) \to V_{h_x}(\Omega) \otimes$

$L^2(0,T)$ is defined by

$$\int_0^T \int_\Omega \nabla_x Q^1_{h_x} v(x,t) \cdot \nabla_x v_{h_x}(x,t)\, dx\, dt \;=\; \int_0^T \int_\Omega \nabla_x v(x,t) \cdot \nabla_x v_{h_x}(x,t)\, dx\, dt \tag{17.25}$$

for all $v_{h_x} \in V_{h_x}(\Omega) \otimes L^2(0,T)$. Note that we have the stability estimate

$$\|\nabla_x Q^1_{h_x} v\|_{L^2(Q)} \le \|\nabla_x v\|_{L^2(Q)},$$

and, when assuming $v \in L^2(0,T;H^{1+s}(\Omega))$ for some $s \in [0,1]$, the standard error estimate

$$\|v - Q^1_{h_x} v\|_{L^2(Q)} \le c\, h_x^{1+s}\, \|v\|_{L^2(0,T;H^{1+s}(\Omega))}, \tag{17.26}$$

if Ω is sufficiently regular. Second, we define the $H^1_{0,}$ projection $Q^1_{h_t} : H^1_{0,}(0,T;L^2(\Omega)) \to L^2(\Omega) \otimes S^1_{h_t}(0,T) \cap H^1_{0,}(0,T)$ by

$$\int_0^T \int_\Omega \partial_t Q^1_{h_t} v(x,t)\, \partial_t v_{h_t}(x,t)\, dx\, dt \;=\; \int_0^T \int_\Omega \partial_t v(x,t)\, \partial_t v_{h_t}(x,t)\, dx\, dt \tag{17.27}$$

for all $v_{h_t} \in L^2(\Omega) \otimes S^1_{h_t}(0,T) \cap H^1_{0,}(0,T)$. Again we have the stability estimate

$$\|\partial_t Q^1_{h_t} v\|_{L^2(Q)} \le \|\partial_t v\|_{L^2(Q)},$$

and, when assuming $v \in H^{1+s}(0,T;L^2(\Omega))$ for some $s \in [0,1]$, the standard error estimate

$$\|v - Q^1_{h_t} v\|_{L^2(Q)} \le c\, h_t^{1+s}\, \|v\|_{H^{1+s}(0,T;L^2(\Omega))}\,. \tag{17.28}$$

The next lemma shows that $Q^1_{h_t} Q^1_{h_x} v \in Q^1_h(Q) \cap H^{1,1}_{0;0,}(Q)$ is well–defined under some regularity assumptions on v, and that the projection operators in space and time commute, see also [2, Lemma 2.1].

Lemma 17.10 *For a given function $v \in H^{1,1}_{0;0,}(Q)$ satisfying $\partial_t v \in L^2(0,T;H^1_0(\Omega))$ and $\partial_{x_m} v \in H^1_{0,}(0,T;L^2(\Omega))$, $m = 1,\dots,d$, there hold the following relations:*

i. $\partial_t Q^1_{h_x} v = Q^1_{h_x} \partial_t v \in V_{h_x}(\Omega) \otimes L^2(0,T)$,

ii. $\partial_{x_m} Q^1_{h_t} v = Q^1_{h_t} \partial_{x_m} v \in L^2(\Omega) \otimes S^1_{h_t}(0,T) \cap H^1_{0,}(0,T)$, $m = 1,\dots,d$,

iii. $Q^1_{h_t} Q^1_{h_x} v = Q^1_{h_x} Q^1_{h_t} v \in Q^1_h(Q) \cap H^{1,1}_{0;0,}(Q)$. *In particular, the space–time projections $Q^1_{h_t} Q^1_{h_x} v$ and $Q^1_{h_x} Q^1_{h_t} v$ are well–defined.*

Moreover, there holds the error estimate

$$\|v-Q^1_{h_t}Q^1_{h_x}v\|_{L^2(Q)} \le \|v-Q^1_{h_t}v\|_{L^2(Q)}+\|v-Q^1_{h_x}v\|_{L^2(Q)}+c\,h_x\,h_t\,\|\partial_t\nabla_x v\|_{L^2(Q)}.$$

Proof For $\partial_t v \in L^2(0,T;H^1_0(\Omega))$ we consider $Q^1_{h_x}\partial_t v \in V_{h_x}(\Omega)\otimes L^2(0,T)$ as the unique solution of the variational formulation

$$\int_0^T\int_\Omega \nabla_x Q^1_{h_x}\partial_t v(x,t)\cdot\nabla_x v_{h_x}(x,t)\,dx\,dt = \int_0^T\int_\Omega \nabla_x \partial_t v(x,t)\cdot\nabla_x v_{h_x}(x,t)\,dx\,dt$$

for all $v_{h_x}\in V_{h_x}(\Omega)\otimes C^\infty_0(0,T)$. By using integration by parts in time twice this gives

$$\begin{aligned}\int_0^T\int_\Omega \nabla_x Q^1_{h_x}\partial_t v(x,t)\cdot\nabla_x v_{h_x}(x,t)\,dx\,dt &= -\int_0^T\int_\Omega \nabla_x v(x,t)\cdot\nabla_x\partial_t v_{h_x}(x,t)\,dx\,dt\\ &= -\int_0^T\int_\Omega \nabla_x Q^1_{h_x}v(x,t)\cdot\nabla_x\partial_t v_{h_x}(x,t)\,dx\,dt\\ &= \int_0^T\int_\Omega \nabla_x\partial_t Q^1_{h_x}v(x,t)\cdot\nabla_x v_{h_x}(x,t)\,dx\,dt\end{aligned}$$

for all $v_{h_x}\in V_{h_x}(\Omega)\otimes C^\infty_0(0,T)$. Since $C^\infty_0(0,T)$ is dense in $L^2(0,T)$ this holds true for all $v_{h_x}\in V_{h_x}(\Omega)\otimes L^2(0,T)$, i.e. *i.* The proof of *ii.* follows in the same manner.

To prove *iii.* we first note that $Q^1_{h_x}v\in V_{h_x}(\Omega)\otimes H^1_{0,}(0,T)\subset H^1_{0,}(0,T;L^2(\Omega))$, and so $Q^1_{h_t}Q^1_{h_x}v\in Q^1_h(Q)\cap H^{1,1}_{0;0,}(Q)$ is well–defined. Analogously we have that $Q^1_{h_x}Q^1_{h_t}v\in Q^1_h(Q)\cap H^{1,1}_{0;0,}(Q)$ is well–defined. Now, with *i.*, *ii.*, and the definitions (17.25), (17.27) we have that

$$\begin{aligned}\langle\partial_t\nabla_x Q^1_{h_t}Q^1_{h_x}v,\partial_t\nabla_x v_h\rangle_{L^2(Q)} &= \langle\partial_t Q^1_{h_t}\nabla_x Q^1_{h_x}v,\partial_t\nabla_x v_h\rangle_{L^2(Q)}\\ &= \langle\partial_t\nabla_x Q^1_{h_x}v,\partial_t\nabla_x v_h\rangle_{L^2(Q)}\\ &= \langle\partial_t\nabla_x v,\partial_t\nabla_x v_h\rangle_{L^2(Q)}\end{aligned}$$

as well as

$$\langle\partial_t\nabla_x Q^1_{h_x}Q^1_{h_t}v,\partial_t\nabla_x v_h\rangle_{L^2(Q)} = \langle\partial_t\nabla_x v,\partial_t\nabla_x v_h\rangle_{L^2(Q)}$$

for all $v_h\in Q^1_h(Q)\cap H^{1,1}_{0;0,}(Q)$, i.e. *iii.*

The error estimate follows from the triangle inequality

$$\begin{aligned}\|v-Q^1_{h_t}Q^1_{h_x}v\|_{L^2(Q)} &\le \|v-Q^1_{h_t}v\|_{L^2(Q)}+\|Q^1_{h_t}(v-Q^1_{h_x}v)\|_{L^2(Q)}\\ &\le \|v-Q^1_{h_t}v\|_{L^2(Q)}+\|v-Q_{h_x}v\|_{L^2(Q)}+\|(I-Q^1_{h_t})(I-Q^1_{h_x})v\|_{L^2(Q)}\end{aligned}$$

$$\leq \|v - Q^1_{h_t} v\|_{L^2(Q)} + \|v - Q_{h_x} v\|_{L^2(Q)} + c\, h_t\, \|\partial_t (I - Q^1_{h_x}) v\|_{L^2(Q)}$$
$$\leq \|v - Q^1_{h_t} v\|_{L^2(Q)} + \|v - Q_{h_x} v\|_{L^2(Q)} + c\, h_t\, h_x\, \|\partial_t \nabla_x v\|_{L^2(Q)}.$$

□

Now we are in a position to prove an error estimate for the approximate solution $\widetilde{u}_h$.

Theorem 17.2 *Let $u \in H^{1,1}_{0;0,}(Q)$ be the unique solution of the variational formulation (17.2) satisfying $\partial_t u \in L^2(0,T; H^1_0(\Omega))$ and $\partial_{x_m} u \in H^1_{0,}(0,T; L^2(\Omega))$, $m = 1, \ldots, d$, and $\Delta_x u \in H^1_{0,}(0,T; L^2(\Omega))$. Then, the unique solution $\widetilde{u}_h \in Q^1_h(Q) \cap H^{1,1}_{0;0,}(Q)$ of the Galerkin–Petrov finite element discretization (17.22) satisfies the error estimate*

$$\|u - \widetilde{u}_h\|_{L^2(Q)} \leq$$
$$\leq \|u - Q^1_{h_t} u\|_{L^2(Q)} + \|u - Q^1_{h_x} u\|_{L^2(Q)} + c\, h_x\, h_t \|\partial_t \nabla_x u\|_{L^2(Q)}$$
$$+2T \Big\{ \|\Delta_x (u - Q^1_{h_t} u)\|_{[H^1_{,0}(0,T;L^2(\Omega))]'} + \|\partial_t (Q^1_{h_x} u - u)\|_{L^2(Q)} + \frac{h_t^2}{12} \|\partial_t \Delta_x u\|_{L^2(Q)} \Big\}.$$

Proof Since the solution u fulfills the assumptions of Lemma 17.10, the space–time projection $Q^1_{h_t} Q^1_{h_x} u \in Q^1_h(Q) \cap H^{1,1}_{0;0,}(Q)$ is well–defined. When using the representation (17.21), the properties of $Q^1_{h_t} Q^1_{h_x}$ as given in Lemma 17.10, and applying integration by parts, we conclude for all $w_h \in Q^1_h(Q) \cap H^{1,1}_{0;,0}(Q)$ that

$$\begin{aligned}
a_h(\widetilde{u}_h - Q^1_{h_t} Q^1_{h_x} u, w_h) &= a_h(\widetilde{u}_h, w_h) - a_h(Q^1_{h_t} Q^1_{h_x} u, w_h) \\
&= a(u, w_h) - a_h(Q^1_{h_t} Q^1_{h_x} u, w_h) \\
&= a(u, w_h) - a(Q^1_{h_t} Q^1_{h_x} u, w_h) + \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}^2}{12} \langle \partial_t \nabla_x Q^1_{h_t} Q^1_{h_x} u, \partial_t \nabla_x w_h \rangle_{L^2(\Omega \times \tau_\ell)} \\
&= -\langle \partial_t u, \partial_t w_h \rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x w_h \rangle_{L^2(Q)} + \langle \partial_t Q^1_{h_t} Q^1_{h_x} u, \partial_t w_h \rangle_{L^2(Q)} \\
&\qquad - \langle \nabla_x Q^1_{h_t} Q^1_{h_x} u, \nabla_x w_h \rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}^2}{12} \langle \partial_t \nabla_x Q^1_{h_t} Q^1_{h_x} u, \partial_t \nabla_x w_h \rangle_{L^2(\Omega \times \tau_\ell)} \\
&= -\langle \partial_t u, \partial_t w_h \rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x w_h \rangle_{L^2(Q)} + \langle \partial_t Q^1_{h_x} u, \partial_t w_h \rangle_{L^2(Q)} \\
&\qquad - \langle \nabla_x Q^1_{h_t} u, \nabla_x w_h \rangle_{L^2(Q)} + \sum_{\ell=1}^{N_t} \frac{h_{t,\ell}^2}{12} \langle \partial_t \nabla_x Q^1_{h_t} u, \partial_t \nabla_x w_h \rangle_{L^2(\Omega \times \tau_\ell)}
\end{aligned}$$

$$= \langle \partial_t (Q^1_{h_x} u - u), \partial_t w_h \rangle_{L^2(Q)} + \langle \nabla_x (u - Q^1_{h_t} u), \nabla_x w_h \rangle_{L^2(Q)}$$
$$+ \sum_{\ell=1}^{N_t} \frac{h^2_{t,\ell}}{12} \langle \partial_t \nabla_x Q^1_{h_t} u, \partial_t \nabla_x w_h \rangle_{L^2(\Omega \times \tau_\ell)}$$
$$= \langle \partial_t (Q^1_{h_x} u - u), \partial_t w_h \rangle_{L^2(Q)} + \langle -\Delta_x (u - Q^1_{h_t} u), w_h \rangle_{L^2(Q)}$$
$$- \sum_{\ell=1}^{N_t} \frac{h^2_{t,\ell}}{12} \langle \partial_t \Delta_x Q^1_{h_t} u, \partial_t w_h \rangle_{L^2(\Omega \times \tau_\ell)}.$$

In particular we observe that $\widetilde{u}_h - Q^1_{h_t} Q^1_{h_x} u$ is the unique solution of (17.23) in the case

$$f_0 = -\Delta_x (u - Q^1_{h_t} u), \quad f_1 = \partial_t (Q^1_{h_x} u - u), \quad f_2 = -\frac{1}{12} \partial_t \Delta_x Q^1_{h_t} u.$$

Therefore, the stability estimate (17.24) and the stability of $Q^1_{h_t}$ in $H^1_{0,}(0,T)$ give

$$\|\widetilde{u}_h - Q^1_{h_t} Q^1_{h_x} u\|_{L^2(Q)} \le 2T \left\{ \|\Delta_x (u - Q^1_{h_t}) u\|_{[H^1_{,0}(0,T;L^2(\Omega))]'} \right.$$
$$\left. + \|\partial_t (Q^1_{h_x} u - u)\|_{L^2(Q)} + \frac{h_t^2}{12} \|\partial_t \Delta_x Q^1_{h_t} u\|_{L^2(Q)} \right\}$$
$$\le 2T \left\{ \|\Delta_x (u - Q^1_{h_t}) u\|_{[H^1_{,0}(0,T;L^2(\Omega))]'} \right.$$
$$\left. + \|\partial_t (Q^1_{h_x} u - u)\|_{L^2(Q)} + \frac{h_t^2}{12} \|\partial_t \Delta_x u\|_{L^2(Q)} \right\}.$$

With the last estimate, the triangle inequality, and the error estimate of Lemma 17.10 we finally obtain

$$\|u - \widetilde{u}_h\|_{L^2(Q)} \le \|u - Q^1_{h_t} Q^1_{h_x} u\|_{L^2(Q)} + \|\widetilde{u}_h - Q^1_{h_t} Q^1_{h_x} u\|_{L^2(Q)}$$
$$\le \|u - Q^1_{h_t} u\|_{L^2(Q)} + \|u - Q^1_{h_x} u\|_{L^2(Q)} + c\, h_x\, h_t \|\partial_t \nabla_x u\|_{L^2(Q)}$$
$$+ 2T \left\{ \|\Delta_x (u - Q^1_{h_t} u)\|_{[H^1_{,0}(0,T;L^2(\Omega))]'} + \|\partial_t (Q^1_{h_x} u - u)\|_{L^2(Q)} + \frac{h_t^2}{12} \|\partial_t \Delta_x u\|_{L^2(Q)} \right\}.$$

□

By using the error estimates (17.28) for the $H^1_{0,}$ projection $Q^1_{h_t}$ and (17.26) for the H^1_0 projection $Q^1_{h_x}$ we now conclude from Theorem 17.2 the following error estimate.

Corollary 17.2 *Let the assumptions of Theorem 17.2 be satisfied. If in addition the unique solution u of (17.2) is sufficiently smooth and Ω is sufficiently regular, we*

obtain the error estimate

$$\|u-\widetilde{u}_h\|_{L^2(Q)} \le c\,h_x^2\left(\|u\|_{L^2(0,T;H^2(\Omega))}+\|\partial_t u\|_{L^2(0,T;H^2(\Omega))}\right) \tag{17.29}$$
$$+c\,h_x\,h_t\|\partial_t\nabla_x u\|_{L^2(Q)}+c\,h_t^2\left(\|\partial_{tt}u\|_{L^2(Q)}+\|\partial_{tt}\Delta_x u\|_{L^2(Q)}+\|\partial_t\Delta_x u\|_{L^2(Q)}\right).$$

As a numerical example for the Galerkin–Petrov finite element method (17.22) we consider the one–dimensional spatial domain $\Omega = (0,1)$, i.e. we have the rectangular space–time domain $Q := \Omega \times (0,T) := (0,1)\times(0,10)$. The discretization is done with respect to nonuniform meshes as shown in Fig. 17.1 where we apply a uniform refinement strategy. Note that these meshes do not fulfill the CFL condition (17.3). As exact solutions we choose for $(x,t)\in Q$

$$u_1(x,t) = \sin(\pi x)\sin^2\left(\frac{5}{4}\pi t\right), \quad u_2(x,t) = \sin(\pi x)t^2(10-t)^{3/4}.$$

The appearing integrals to compute the related right–hand side in (17.22) are calculated by using high order quadrature rules. The numerical results for the smooth solution u_1 are given in Table 17.3 where we observe unconditional stability and quadratic convergence in $\|\cdot\|_{L^2(Q)}$, as predicted by the error estimate (17.29). Moreover we have linear convergence when measuring the error in $|\cdot|_{H^1(Q)}$. Note that such an error estimate can be shown by using the $H^1(Q)$ projection, an inverse inequality, and the error estimate (17.29). For the singular solution u_2 the related results are given in Table 17.4 where we observe a reduced order of convergence in $\|\cdot\|_{L^2(Q)}$ and in $|\cdot|_{H^1(Q)}$, respectively. These convergence rates correspond to the reduced Sobolev regularity $u_2 \in H^{5/4-\varepsilon}(Q)$, $\varepsilon > 0$.

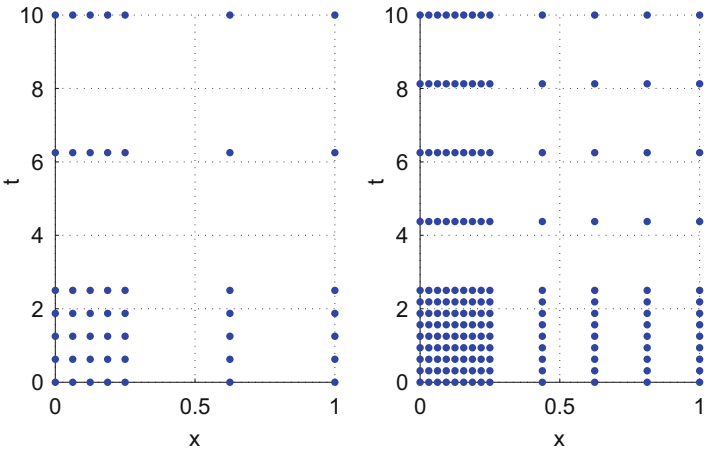


Fig. 17.1 Nonuniform meshes: starting mesh and the mesh after one uniform refinement step

Table 17.3 Numerical results of (17.22) for $Q = (0,1) \times (0,10)$ and for u_1

dof	$h_{x,\max}$	$h_{x,\min}$	$h_{t,\max}$	$h_{t,\min}$	$\|u_1 - \widetilde{u}_{1,h}\|_{L^2(Q)}$	eoc	$\|u_1 - \widetilde{u}_{1,h}\|_{H^1(Q)}$	eoc
30	0.37500	0.06250	3.75000	0.62500	3.579e+00	0.00	1.289e+01	0.00
132	0.18750	0.03125	1.87500	0.31250	1.975e+00	0.86	9.849e+00	0.39
552	0.09375	0.01562	0.93750	0.15625	9.213e−01	1.10	6.534e+00	0.59
2256	0.04688	0.00781	0.46875	0.07812	6.829e−01	0.43	5.210e+00	0.33
9120	0.02344	0.00391	0.23438	0.03906	2.466e−01	1.47	2.848e+00	0.87
36,672	0.01172	0.00195	0.11719	0.01953	7.029e−02	1.81	1.435e+00	0.99
147,072	0.00586	0.00098	0.05859	0.00977	1.819e−02	1.95	7.159e−01	1.00
589,056	0.00293	0.00049	0.02930	0.00488	4.588e−03	1.99	3.576e−01	1.00
2,357,760	0.00146	0.00024	0.01465	0.00244	1.149e−03	2.00	1.788e−01	1.00
9,434,112	0.00073	0.00012	0.00732	0.00122	2.875e−04	2.00	8.938e−02	1.00
37,742,592	0.00037	0.00006	0.00366	0.00061	7.189e−05	2.00	4.469e−02	1.00

Table 17.4 Numerical results of (17.22) for $Q = (0,1) \times (0,10)$ and for u_2

dof	$h_{x,\max}$	$h_{x,\min}$	$h_{t,\max}$	$h_{t,\min}$	$\|u_2 - \widetilde{u}_{2,h}\|_{L^2(Q)}$	eoc	$\|u_2 - \widetilde{u}_{2,h}\|_{H^1(Q)}$	eoc
30	0.37500	0.06250	3.75000	0.62500	7.836e+01	0.00	3.173e+02	0.00
132	0.18750	0.03125	1.87500	0.31250	2.166e+01	1.86	1.191e+02	1.41
552	0.09375	0.01562	0.93750	0.15625	5.487e+00	1.98	5.225e+01	1.19
2256	0.04688	0.00781	0.46875	0.07812	1.777e+00	1.63	2.696e+01	0.95
9120	0.02344	0.00391	0.23438	0.03906	6.476e−01	1.46	1.593e+01	0.76
36,672	0.01172	0.00195	0.11719	0.01953	3.001e−01	1.11	1.076e+01	0.57
147,072	0.00586	0.00098	0.05859	0.00977	1.393e−01	1.11	8.077e+00	0.41
589,056	0.00293	0.00049	0.02930	0.00488	6.156e−02	1.18	6.452e+00	0.32
2,357,760	0.00146	0.00024	0.01465	0.00244	2.650e−02	1.22	5.308e+00	0.28
9,434,112	0.00073	0.00012	0.00732	0.00122	1.126e−02	1.23	4.423e+00	0.26
37,742,592	0.00037	0.00006	0.00366	0.00061	4.758e−03	1.24	3.704e+00	0.26

Remark 17.2 The Galerkin–Petrov finite element method (17.22) seems to fulfill a kind of conservation of the total energy

$$E(t) := \frac{1}{2}\|\partial_t u(\cdot,t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|\nabla_x u(\cdot,t)\|^2_{L^2(\Omega)}, \quad t \in [0,T].$$

As illustration we consider a solution of the homogeneous wave equation, i.e.

$$u_3(x,t) = (\cos(\pi t) + \sin(\pi t))\sin(\pi x) \quad \text{for } (x,t) \in Q := (0,1) \times (0,10)$$

with the total energy

$$E(t) = \frac{\pi^2}{2} \quad \text{for } t \in [0,10].$$

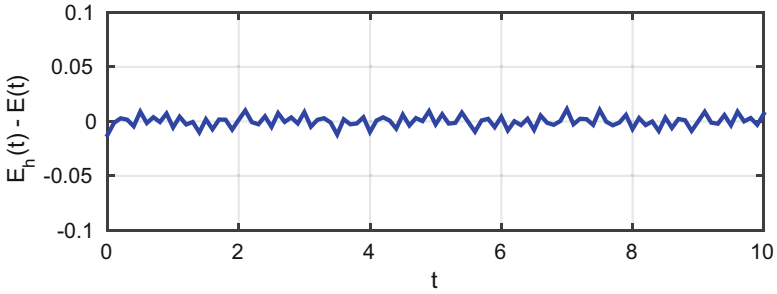


Fig. 17.2 Difference of the total energy $E(t) = \frac{\pi^2}{2}$ and $E_h(t)$ for the solution u_3 for a uniform mesh

Here, the initial condition $u_3(x, 0) = \sin(\pi x)$, $x \in \Omega$, is treated via homogenization, while the initial condition $\partial_t u_3(x, 0) = \pi \sin(\pi x)$, $x \in \Omega$, is incorporated in a weak sense. For the solution u_3 we compute the discrete total energy

$$E_h(t) := \frac{1}{2}\|\partial_t \widetilde{u}_h(\cdot, t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|\nabla_x \widetilde{u}_h(\cdot, t)\|^2_{L^2(\Omega)}, \ t \in [0, T].$$

In Fig. 17.2 the difference

$$E_h(t) - E(t) = E_h(t) - \frac{\pi^2}{2} \quad \text{for } t \in [0, 10]$$

is plotted pointwise for the refinement level with uniform mesh sizes $h_t = \frac{10}{6 \cdot 2^{10}}$ and $h_x = \frac{1}{6 \cdot 2^{10}}$. Note that $\partial_t \widetilde{u}_h$ is piecewise constant in time. Probably due to the used space–time approximation we observe some oscillations within the finite element accuracy, but no energy loss occurs.

References

1. Andreev, R.: Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. IMA J. Numer. Anal. **33**(1), 242–260 (2013)
2. Aziz, A.K., Monk, P.: Continuous finite elements in space and time for the heat equation. Math. Comput. **52**(186), 255–274 (1989)
3. Bales, L., Lasiecka, I.: Continuous finite elements in space and time for the nonhomogeneous wave equation. Comput. Math. Appl. **27**(3), 91–102 (1994)
4. Bangerth, W., Geiger, M., Rannacher, R.: Adaptive Galerkin finite element methods for the wave equation. Comput. Methods Appl. Math. **10**(1), 3–48 (2010)
5. Cohen, G.C.: Higher-Order Numerical Methods for Transient Wave Equations. Scientific Computation. Springer, Berlin (2002)
6. Cohen, G.C., Pernet, S.: Finite Element and Discontinuous Galerkin Methods for Transient Wave Equations. Scientific Computation. Springer, Dordrecht (2017)

7. Dörfler, W., Findeisen, S., Wieners, C.: Space-time discontinuous Galerkin discretizations for linear first-order hyperbolic evolution systems. Comput. Methods Appl. Math. **16**(3), 409–428 (2016)
8. French, D.A., Peterson, T.E.: A continuous space-time finite element method for the wave equation. Math. Comput. **65**(214), 491–506 (1996)
9. Ladyzhenskaya, O.A.: The Boundary Value Problems of Mathematical Physics. Applied Mathematical Sciences, vol. 49. Springer, New York (1985)
10. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications. Travaux et Recherches Mathématiques, vol. 1, no. 17 Dunod, Paris (1968)
11. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications. Travaux et Recherches Mathématiques, vol. 2, no. 18 Dunod, Paris (1968)
12. Mollet, C.: Stability of Petrov-Galerkin discretizations: application to the space-time weak formulation for parabolic evolution problems. Comput. Methods Appl. Math. **14**(2), 231–255 (2014)
13. Neumüller, M.: Space-time methods: fast solvers and applications. Ph.D. Thesis, TU Graz (2013)
14. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Applied Mathematical Sciences, vol. 44. Springer, New York (1983)
15. Peterseim, D., Schedensack, M.: Relaxing the CFL condition for the wave equation on adaptive meshes. J. Sci. Comput. **72**, 1196–1213 (2017)
16. Schwab, C., Stevenson, R.: Space-time adaptive wavelet methods for parabolic evolution problems. Math. Comput. **78**(267), 1293–1318 (2009)
17. Steinbach, O.: Space-time finite element methods for parabolic problems. Comput. Methods Appl. Math. **15**(4), 551–566 (2015)
18. Steinbach, O., Zank, M.: Coercive space–time finite element methods for initial boundary value problems (in review)
19. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer Series in Computational Mathematics, vol. 25, 2nd edn. Springer, Berlin (2006)
20. Urban, K., Patera, A.T.: An improved error bound for reduced basis approximation of linear parabolic problems. Math. Comput. **83**(288), 1599–1615 (2014)
21. Wloka, J.: Partielle Differentialgleichungen. B. G. Teubner, Stuttgart (1982)
22. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. II/A. Springer, New York (1990)
23. Zlotnik, A.A.: Convergence rate estimates of finite-element methods for second-order hyperbolic equations. In: Numerical Methods and Applications, pp. 155–220. CRC, Boca Raton (1994)

Chapter 18
An Optimal Order CG-DG Space-Time Discretization Method for Parabolic Problems

Igor Voulis

Abstract We consider a space-time discretization method for second-order parabolic problems with inhomogeneous (time-dependent) Dirichlet boundary conditions. A combination of a temporal discontinuous Galerkin scheme and a spatial continuous Galerkin scheme is used. In previous work it has been established that the standard semi-discrete temporal scheme has to be modified to obtain an optimal error bound. Here we extend this modification to a fully discrete scheme. For this modified discretization an optimal error bound for the energy norm is derived. Results of experiments confirm the theoretically predicted optimal convergence rates. We are able to pinpoint why the standard CG-DG space-time method (without any modifications) has suboptimal convergence behavior. The method presented here avoids this suboptimality in a way which is computationally very cheap.

18.1 Introduction

Galerkin finite element methods are a popular discretization technique for many classes of ordinary and partial differential equations. See [3, 4, 8, 11, 16] for an overview. For parabolic partial differential equations the discontinuous Galerkin (DG) finite element for time discretization is commonly combined with a spatial continuous Galerkin (CG) method. Such a method was introduced in [12]. Further methods of this type have been studied in [5–7, 14–16].

To the best of our knowledge, the error analyses available in the literature only treat parabolic problems with *homogeneous* constraints. Inhomogeneous linear constraints have a significant impact on the behavior of the convergence order of the DG time discretization of abstract parabolic problems. This topic is treated in

I. Voulis (✉)
RWTH-Aachen University, Institut fuer Geometrie und Praktische Mathematik, Aachen, Germany
e-mail: voulis@igpm.rwth-aachen.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_18

[17]. A modification needs to be introduced to obtain an optimal convergence order. Similar effects (so-called order reduction) are known to occur in other numerical schemes, such as Runge-Kutta methods [13, Sect. 2.12] and Rosenbrock methods [1].

In this paper we discuss how to apply the modified method described in [17] to a second-order parabolic equation [9, 18] with an inhomogeneous (time-dependent) Dirichlet boundary condition. We combine the DG time discretization with a standard CG spatial discretization, see [8]. This results in a fully discrete space-time finite element method.

The main results of the paper are the following. We introduce the fully discrete CG-DG space-time scheme for a parabolic problem with *in*homogeneous constraints. For the resulting space-time method we *prove an optimal discretization error bound* for the global energy norm. We also discuss *why* a modification is necessary to obtain errors of optimal order.

The paper is organized as follows. In Sect. 18.2 we consider a class of abstract parabolic problems which is relevant for the further analysis. We recall the DG time discretization scheme for such problems from [17] and we state the relevant results.

In Sect. 18.3 we introduce the spatial discretization for second-order parabolic equations with inhomogeneous Dirichlet boundary conditions. This gives rise to a modified space-time method, which uses the temporal DG scheme and a spatial CG scheme. We give an error analysis of the CG-DG space-time scheme. We derive an optimal order error bound in the energy norm.

In Sect. 18.4 this scheme is used to perform numerical experiments and the results of a few experiments are presented. Our theoretical results are confirmed and we see optimal superconvergence results at temporal nodes. We also show that without a proper treatment of the Dirichlet boundary conditions the results are no longer optimal. We discuss the source of this suboptimality. We conclude with an outlook in Sect. 18.5.

18.2 Temporal Discretisation of a Parabolic Problem with Linear Constraints

In this section we consider a class of parabolic problems with linear constraints which has been treated in [17]. In the first subsection we recall the temporal discretization and in the second subsection we recall some convergence results. These results will play an important role in the following sections.

Let $\mathscr{U}$, $\mathscr{H}$ be real separable Hilbert spaces with a dense continuous embedding $\mathscr{U} \hookrightarrow \mathscr{H}$. The norms are denoted by $\|\cdot\|_{\mathscr{U}}$, $\|\cdot\|_{\mathscr{H}}$, respectively. Let $\mathscr{U}_0$ be a closed subspace of $\mathscr{U}$ and let $\mathscr{H}_0 = \overline{\mathscr{U}_0}^{\mathscr{H}}$. We assume that these spaces induce a Gelfand triple $\mathscr{U}_0 \hookrightarrow \mathscr{H}_0 \cong \mathscr{H}_0{}' \hookrightarrow \mathscr{U}_0{}'$. Let $a : \mathscr{U} \times \mathscr{U} \to \mathbf{R}$ be a symmetric

continuous bilinear form on $\mathscr{U}$, which is coercive on $\mathscr{U}_0$:

$$|a(u,v)| \leq \Gamma \|u\|_{\mathscr{U}} \|v\|_{\mathscr{U}} \quad \text{for all } u, v \in \mathscr{U}, \tag{18.1}$$

$$a(v,v) \geq \gamma \|v\|^2_{\mathscr{U}} \quad \text{for all } v \in \mathscr{U}_0, \tag{18.2}$$

with $\gamma > 0$. The corresponding operator is denoted by $A : \mathscr{U} \to \mathscr{U}'$, $Au(v) = a(u,v)$. Let $\mathscr{Q}$ be a Hilbert space. Let $B : \mathscr{U} \to \mathscr{Q}$ be a surjective continuous mapping with $\mathscr{U}_0 = \ker B$. Let $I = (0,T)$ be a given time interval. For given $f \in L^2(I;\mathscr{U}_0{}')$, $g \in H^1(I;\mathscr{Q})$ we consider the following abstract parabolic problem: find $u \in L^2(I;\mathscr{U}) \cap H^1(I;\mathscr{U}_0{}')$ such that $u(0) = u_0$ and

$$u' + Au = f \quad \text{in } L^2(I;\mathscr{U}_0{}'), \tag{18.3}$$

$$Bu = g \quad \text{in } L^2(I;\mathscr{Q}). \tag{18.4}$$

If we assume that $u_0 \in \mathscr{U}$ and $Bu_0 = g(0)$, then (by a standard lifting argument) this problem is well-posed [15, 18].

In the next two subsection we will consider the temporal discretization method for (18.3) and (18.4) and the error bounds for this method. In Sect. 18.3 these abstract results will be applied to second-order parabolic problems with *inhomogeneous* Dirichlet boundary conditions, which we introduce now. Let $\Omega \subset \mathbf{R}^d$ be a bounded domain. Let $\mathscr{U} = H^1(\Omega)$, $\mathscr{H} = L^2(\Omega)$ and let $B : u \mapsto u|_{\partial\Omega}$ be the trace operator. Then we have $\mathscr{Q} = H^{1/2}(\partial\Omega)$, $\mathscr{U}_0 = H^1_0(\Omega)$ and $\mathscr{H}_0 = L^2(\Omega)$. For simplicity we will assume that Ω is a convex polyhedron. We will consider the following problem: find u such that $u(0) = u_0$ and

$$u' + Au = f \quad \text{in } L^2(I;H^{-1}(\Omega)), \tag{18.5}$$

$$u|_{\partial\Omega} = g \quad \text{in } L^2(I;H^{1/2}(\partial\Omega)). \tag{18.6}$$

If we take $A = -\Delta$, we obtain the heat equation.

In the literature one can find analyses of DG time discretizations applied to this parabolic problem, combined with a conforming finite element method in space, e.g., [16, Chapter 12]. In this paper we consider the modification which is necessary to deal with problems with *in*homogeneous boundary conditions.

18.2.1 Temporal Discretization

In this subsection we briefly recall the DG time discretization for the parabolic problem (18.3) and (18.4).

We take a fixed $q \in \mathbf{N}$, $q \geq 1$. In the temporal discretization we will use polynomials of degree $q-1$, this space of polynomials will be denoted by $\mathscr{P}_{q-1}$. For $N \in \mathbf{N}$, introduce $0 = t_0 < \cdots < t_N = T$, $I_n = (t_{n-1}, t_n]$, and $k = |I_n| = \frac{T}{N} \leq 1$

is the fixed step-size for $n = 1, \ldots, N$. The assumption that the step-size is fixed is merely made for simplicity. The same arguments apply for a variable step-size. We define the broken spaces

$$\mathscr{P}^b(I) := \bigoplus_{n=1}^{N} \mathscr{P}_{q-1}(I_n) \subset L^2(I), \quad H^{1,b}(I) := \bigoplus_{n=1}^{N} H^1(I_n),$$

$$\mathscr{P}^b(I; \mathscr{H}) := \mathscr{P}^b(I) \otimes \mathscr{H}, \quad H^{1,b}(I; \mathscr{H}) := H^{1,b}(I) \otimes \mathscr{H}.$$

For $U \in H^{1,b}(I; \mathscr{H})$ and $n = 1, \ldots, N$ we will write $U^n = U|_{I_n}(t_n)$, $U_+^{n-1} = \lim_{t \downarrow t_{n-1}} U|_{I_n}(t)$. We define U' by taking the derivative on each interval I_n:

$$U' = \sum_{n=1}^{N} U|'_{I_n} \chi_{I_n},$$

with χ_{I_n} the characteristic function for I_n. We define the following bilinear form on $H^{1,b}(I; \mathscr{H})$ which corresponds to a discrete time derivative:

$$(Y, X) \mapsto D_{\mathscr{H}}(Y, X) := \int_I (Y', X)_{\mathscr{H}} + \sum_{n=1}^{N-1} (Y_+^n - Y^n, X_+^n)_{\mathscr{H}} + (Y_+^0, X_+^0)_{\mathscr{H}}.$$

The following projection plays a crucial role. See also [15] and [16, pp. 207–208].

Definition 18.1 We define a projection $\Pi_q : H^{1,b}(I) \to \mathscr{P}^b(I)$. For any $w \in H^{1,b}(I)$ we define $\Pi_q w$ by the following relationship:

$$(\Pi_q w)^n = w^n,$$

$$\int_{I_n} (\Pi_q w(t) - w(t)) t^j \, dt = 0 \quad \text{for all } j = 0, \ldots, q-2,$$

for all n=1, …, N (for q=1 only the first condition is used). Finally let $\mathscr{H}$ be a Hilbert space with an (orthogonal) basis $\{h_j | j \in \mathbf{N}\}$. Let $\{b_i | i \in \mathbf{N}\}$ be an (orthogonal) basis for $H^{1,b}(I)$. For $u \in H^{1,b}(I; \mathscr{H}) = H^{1,b}(I) \otimes \mathscr{H}$ we define

$$\Pi_q u := \Pi_q \otimes 1_{\mathscr{H}} u = \sum_{i,j \in \mathbf{N}} u_{i,j} (\Pi_q b_i) \otimes h_j$$

where we have used that $\{b_i \otimes h_j | i, j \in \mathbf{N}\}$ is a basis of $H^{1,b}(I) \otimes \mathscr{H}$ and thus u can be expressed as $u = \sum_{i,j \in \mathbf{N}} u_{i,j} b_i \otimes h_j$ for some uniquely defined coefficients $u_{i,j}$ for $i, j \in \mathbf{N}$.

We now describe the DG time discretizations of (18.3) and (18.4). We introduce the notation

$$K(u,v) := D_{\mathscr{H}}(u,v) + \int_I a(u,v), \qquad u,v \in H^{1,b}(I;\mathscr{H}) \cap L^2(I;\mathscr{U}).$$

The discrete (in time) version of (18.3) and (18.4) reads: find $U \in \mathscr{P}^b(I;\mathscr{U})$, such that

$$K(U,X) = f(X) + (u_0, X_+^0)_{\mathscr{H}} \quad \text{for all } X \in \mathscr{P}^b(I;\mathscr{U}_0), \tag{18.7}$$

$$BU = \Pi_q g. \tag{18.8}$$

Remark 18.1 We have the following characterization of the projection Π_q. Let $w \in H^{1,b}(I;\mathscr{H})$, then $W = \Pi_q w$ is the unique solution of the following problem: find $W \in \mathscr{P}^b(I;\mathscr{H})$ such that

$$D_{\mathscr{H}}(W,X) = D_{\mathscr{H}}(w,X) \quad \text{for all } X \in \mathscr{P}^b(I;\mathscr{H}). \tag{18.9}$$

A proof can be found in [16, pp. 207–208]. From this property (and by replacing $\mathscr{H}$ by $\mathscr{Q}$) we can conclude that (18.8) is equivalent to $D_{\mathscr{Q}}BU = D_{\mathscr{Q}}g$, which is the discrete formulation of $Bu' = g'$ with $Bu(0) = g(0)$. This essential property is used to obtain the optimal error bounds which we will state in the next subsection. A similar approach was used in [2] to obtain accurate Runge-Kutta discretizations of DAEs.

18.2.2 Optimal Error Bounds for the Time-Discrete Formulation

In this subsection we recall the optimal discretization error bounds for the semi-discrete formulation (18.7) and (18.8). These results are from [17].

Theorem 18.1 ([17, Theorem 4.1]) *Let u be the solution of (18.3) and (18.4) with $u \in H^1(I;\mathscr{U})$ and $U \in \mathscr{P}^b(I;\mathscr{U})$ the solution of (18.7) and (18.8). The following holds:*

$$\|u(T) - U(T)\|_{\mathscr{H}} + \|u - U\|_{L^2(I;\mathscr{U})} \leq (1 + c_\gamma \Gamma)\|u - \Pi_q u\|_{L^2(I;\mathscr{U})}$$

with $c_\gamma := \max\{\frac{1}{\gamma}, 2\}$.

Theorem 18.2 ([17, Theorem 4.6]) *Assume that the solution u of (18.3) and (18.4) has regularity $u \in H^m(I;\mathscr{U})$ and let $U \in \mathscr{P}^b(I;\mathscr{U})$ be the solution of (18.7) and*

(18.8). *For any* $1 \leq \ell \leq 2q-1$, $1 \leq m \leq q$, *we have*

$$\|u(T) - U(T)\|_{\mathscr{H}} \leq ck^{\frac{\ell-1}{2}+m} \left(\int_I |u^{(m)}|_\ell^2 \right)^{\frac{1}{2}}$$

for some $c > 0$ *which depends only on* q, γ *and* Γ. *The* $|\cdot|_\ell$*—seminorm is defined using the spectral decomposition of* $A|_{\mathscr{U}_0}$*:*

$$|u|_\ell := \sup_{x \in \mathscr{D}(A|_{\mathscr{U}_0}^{\ell/2})} \frac{(u, A^{\frac{\ell}{2}}x)_{\mathscr{H}}}{\|x\|_{\mathscr{H}}}, \quad u \in \mathscr{H}, \quad \ell \in \mathbf{N},$$

where $\mathscr{D}(A|_{\mathscr{U}_0}^{\ell/2})$ *denotes the space of all* x *such that* $A|_{\mathscr{U}_0}^{\ell/2} x \in \mathscr{H}$. *If* $\{(\lambda_n, v_n) | n \in \mathbf{N}\}$ *is the set of eigenvalues and corresponding orthogonal eigenvectors of* $A|_{\mathscr{U}_0}$, *then*

$$\mathscr{D}(A|_{\mathscr{U}_0}^{\ell/2}) = \left\{ x = \sum_{n \in \mathbf{N}} a_n v_n \,\middle|\, (a_n)_n \in \mathbf{R}^{\mathbf{N}}, \sum_{n \in \mathbf{N}} a_n^2 \lambda_n^\ell < \infty \right\}.$$

18.3 Error Analysis for the Fully Discrete Formulation

In this section we consider the second-order parabolic problem (18.5) and (18.6). We first introduce a standard spatial discretization. We then give an error analysis for the spatial discretization for the problem with *in*homogeneous constraints. Analyses for the homogeneous problem can be found in classical literature, e.g. [8, Sect. 6.1.4]. In the second subsection we use this discretization and the DG-discretisation from the previous section to formulate the fully discrete scheme. Using Theorem 18.1, we obtain an optimal error bound for the energy norm for the fully discrete formulation.

18.3.1 Spatial Discretization

We assume that we have a discretization parameter h and a family of shape regular, geometrically conformal meshes $\{\mathbf{T}_h\}_{h>0}$. We take a H^1-conforming piecewise polynomial space

$$\mathscr{U}^h = \{u \in H^1(\Omega) \mid u|_{T_s} \in \mathscr{P}_\ell(T_s) \text{ for all } T_s \in \mathbf{T}_h\}.$$

Let $\mathscr{U}_0^h = H_0^1(\Omega) \cap \mathscr{U}^h$. We denote the standard nodal basis of $\mathscr{U}^h$ by $\{\psi_1, \ldots \psi_{N_s}\} = \Psi$. From this we extract a nodal basis of $\mathscr{U}_0^h$: $\Psi_0 = \{\psi \mid \psi \in \Psi \cap H_0^1(\Omega)\}$ and we take $\Psi_{\partial\Omega} = \Psi \setminus \Psi_0$. We treat the constraint $u|_{\partial\Omega} = g$ by

taking $G_h = \sum_{\psi \in \Psi_{\partial\Omega}} c(\psi)\psi$ such that $G_h(x) = g(x)$ for all nodal points $x \in \partial\Omega$. This defines a mapping $\mathbf{I}_h^\partial : C(\partial\Omega) \to \mathscr{U}_h$, where $\mathbf{I}_h^\partial g = G_h$, see [8, Sect. 3.2.2].

For a given $u \in H^1(\Omega)$ with $g = u|_{\partial\Omega} \in C(\partial\Omega)$, consider the solution $u_h \in \mathscr{U}^h$ of the discrete stationary problem (Ritz projection with constraints)

$$a(u_h, v_h) = a(u, v_h) \quad \text{for all } v_h \in \mathscr{U}_0^h, \tag{18.10}$$

$$u_h|_{\partial\Omega} = (\mathbf{I}_h^\partial g)|_{\partial\Omega}. \tag{18.11}$$

We assume that for this stationary problem, the discrete solution u_h is an approximation of u of optimal order.

Remark 18.2 (Assumption on the Ritz Projection with Constraints) For $u \in H^{\ell+1}(\Omega)$ with $u|_{\partial\Omega} \in C(\partial\Omega)$, let $u_h \in \mathscr{U}^h$ be the solution of (18.10) and (18.11). We have the following error bound for all $h > 0$

$$\|u - u_h\|_{L^2(\Omega)} + h\|u - u_h\|_{H^1(\Omega)} \le ch^{\ell+1}\|u\|_{H^{\ell+1}(\Omega)} \tag{18.12}$$

for some constant $c > 0$ which is independent of h and u.

Remark 18.3 A sufficient condition on $\Omega, \{\mathbf{T}_h\}_{h>0}, \ell$ and A under which this assumption is true, is given in [8, Corollary 3.29]. In particular, this assumption holds if we take $d = 2$ or $d = 3$ and $\ell \ge 1$, $A = -\Delta$. See [8, Example 3.30].

Lemma 18.1 *Let $f \in L^2(I; H^{-1}(\Omega))$, $g \in H^1(I; C(\partial\Omega))$ and $u_0 \in H^1(\Omega)$. Assume that $u_0^h \in \mathscr{U}^h$ is obtained from u_0 by solving the corresponding stationary problem (18.10) and (18.11). There exists a unique solution $u_h \in H^1(I; \mathscr{U}^h)$ of the following problem*

$$(u_h', v_h)_{L^2(I\times\Omega)} + \int_I a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in L^2(I; \mathscr{U}_0^h), \tag{18.13}$$

$$u_h|_{I\times\partial\Omega} = (\mathbf{I}_h^\partial g)|_{I\times\partial\Omega}, \tag{18.14}$$

$$u_h(0) = u_0^h. \tag{18.15}$$

Moreover, let us assume that the assumption in Remark 18.2 is satisfied and that the solution u of (18.5) and (18.6) has the regularity $u \in H^1(I; H^{\ell+1}(\Omega))$ which also requires more regularity on the given data. Under these assumptions we have the bound

$$\|u_h(T) - u(T)\|_{L^2(\Omega)} + \|u_h - u\|_{L^2(I;H^1(\Omega))} \le E_h(u), \tag{18.16}$$

where

$$E_h(u) := C(h^\ell\|u\|_{L^2(I;H^{\ell+1}(\Omega))} + h^{\ell+1}\|u'\|_{L^2(I;H^{\ell+1}(\Omega))}) \tag{18.17}$$

for some $C > 0$ which is independent of h and u. We also have the following bound

$$\|u_h - u\|_{L^\infty(I;H^1(\Omega))} \leq E_h^\infty(u), \tag{18.18}$$

where

$$E_h^\infty(u) := \hat{C}(h^\ell \|u\|_{L^\infty(I;H^{\ell+1}(\Omega))} + h^{\ell+1}\|u'\|_{L^2(I;H^{\ell+1}(\Omega))}) \tag{18.19}$$

for some $\hat{C} > 0$ which is independent of h and u.

Proof The problem (18.13)–(18.15) is a special case of the abstract problem (18.3) and (18.4) and thus has a unique solution. Using that $\mathscr{U}^h$ is finite dimensional, we find that $u_h \in H^1(I;\mathscr{U}^h)$.

Let, for all $t \in I$, $\tilde{u}_h(t) \in \mathscr{U}^h$ be the solution of the stationary problem

$$\begin{aligned} a(\tilde{u}_h(t), v_h) &= a(u(t), v_h) \quad \text{for all } v_h \in \mathscr{U}_0^h, \\ \tilde{u}_h(t)|_{\partial\Omega} &= (\mathbf{I}_h^\partial g(t))|_{\partial\Omega}. \end{aligned}$$

Since u_h satisfies (18.13) and u satisfies (18.5), we have

$$(u_h' - u', w_h)_{L^2(I\times\Omega)} + \int_I a(u_h - u, w_h) = 0 \quad \text{for all } w_h \in L^2(I;\mathscr{U}_0^h). \tag{18.20}$$

Define $v_h = u_h - \tilde{u}_h$ and $e = u - \tilde{u}_h$. If we take $w_h = v_h$ in (18.20), then we find

$$(v_h', v_h)_{L^2(I\times\Omega)} + \int_I a(v_h, v_h) - (e', v_h)_{L^2(I\times\Omega)} - \int_I a(e, v_h) = 0.$$

Using $v_h(0) = 0$, coercivity (18.2) and $a(e(t), v_h(t)) = 0$ for all $t \in I$, we find

$$\begin{aligned} \frac{1}{2}\|v_h(T)\|_{L^2(\Omega)}^2 + \gamma\|v_h\|_{L^2(I;H^1(\Omega))}^2 &\leq (v_h', v_h)_{L^2(I\times\Omega)} + \int_I a(v_h, v_h) = (e', v_h)_{L^2(I\times\Omega)} \\ &\leq \|e'\|_{L^2(I;H^{-1}(\Omega))}\|v_h\|_{L^2(I;H^1(\Omega))}. \end{aligned}$$

Combining this with Young's inequality, we get

$$\|v_h(T)\|_{L^2(\Omega)} + \|v_h\|_{L^2(I;H^1(\Omega))} \leq c\|e'\|_{L^2(I;L^2(\Omega))}$$

for some c which only depends on γ. Using that $\|e(T)\|_{L^2(\Omega)} \leq \tilde{c}\|e\|_{H^1(I;L^2(\Omega))}$ for some constant $\tilde{c} > 0$, we get

$$\begin{aligned} \|u_h(T) - u(T)\|_{L^2(\Omega)} + \|u_h - u\|_{L^2(I;H^1(\Omega))} &\leq \|v_h(T)\|_{L^2(\Omega)} + \|v_h\|_{L^2(I;H^1(\Omega))} \\ &\quad + \|e(T)\|_{L^2(\Omega)} + \|e\|_{L^2(I;H^1(\Omega))} \\ &\leq c'(\|e'\|_{L^2(I;L^2(\Omega))} + \|e\|_{L^2(I;H^1(\Omega))}) \end{aligned}$$

for some $c' > 0$ which depends only on γ and $\tilde{c}$. Applying (18.12) concludes the proof of (18.16).

If we take any $t \in I$, $w_h = \chi_{(0,t)} v_h'$ and $u_h - u = u_h - \tilde{u}_h - (u - \tilde{u}_h) = v_h - e$ in (18.20), then we find

$$(v_h', v_h')_{L^2((0,t)\times\Omega)} + \int_0^t a(v_h, v_h') - (e', v_h')_{L^2((0,t)\times\Omega)} - \int_0^t a(e, v_h') = 0.$$

Using $v_h(0) = 0$, $\int_0^t a(v_h, v_h') = a(v_h(t), v_h(t))$ and $a(e(\tau), v_h'(\tau)) = 0$ for all $\tau \in I$, we find

$$a(v_h(t), v_h(t)) + (v_h', v_h')_{L^2((0,t)\times\Omega)} = (e', v_h')_{L^2((0,t)\times\Omega)}.$$

Using coercivity (18.2) and the Cauchy inequality, we get

$$\gamma \|v_h(t)\|^2_{H^1(\Omega)} \le a(v_h(t), v_h(t)) \le \|e'\|_{L^2((0,t)\times\Omega)} \|v_h'\|_{L^2((0,t)\times\Omega)} - \|v_h'\|^2_{L^2((0,t)\times\Omega)}.$$

By applying Young's inequality, we get

$$\gamma \|v_h(t)\|^2_{H^1(\Omega)} \le \frac{1}{4} \|e'\|^2_{L^2((0,t)\times\Omega)}.$$

We conclude that for any $t \in I$

$$\|v_h(t)\|_{H^1(\Omega)} \le \frac{1}{2\sqrt{\gamma}} \|e'\|_{L^2((0,t)\times\Omega)} \le \frac{1}{2\sqrt{\gamma}} \|e'\|_{L^2(I;L^2(\Omega))}.$$

Applying (18.12) and the triangle inequality concludes the proof of (18.18). □

18.3.2 Fully Discrete Formulation

We now apply the theory from Sect. 18.2.1 to the parabolic problem (18.13)–(18.15). We take $\mathscr{U} = (\mathscr{U}^h, \|\cdot\|_{H^1(\Omega)})$ and $\mathscr{H} = (\mathscr{U}^h, \|\cdot\|_{L^2(\Omega)})$. The fully discrete version of (18.5) and (18.6) now reads: find $U_h \in \mathscr{P}^b(I; \mathscr{U}^h)$ such that

$$K(U_h, X_h) = \int_I f(X_h) + (u_0^h, (X_h)_+^0)_{L^2(\Omega)} \quad \text{for all } X_h \in \mathscr{P}^b(I; \mathscr{U}_0^h), \tag{18.21}$$

$$U_h|_{I\times\partial\Omega} = (\Pi_q \mathbf{I}_h^\partial g)|_{I\times\partial\Omega}. \tag{18.22}$$

This formulation is consistent with (18.13)–(18.15), which is in turn consistent with (18.5) and (18.6). We have the following error bound.

Theorem 18.3 *Assume that the assumption in Remark 18.2 holds. Let* $1 \leq m \leq q$. *Let* u *be the solution of* (18.5) *and* (18.6) *with* $u \in H^1(I; H^{\ell+1}(\Omega)) \cap H^m(I; H^1(\Omega))$ *and let* $U_h \in \mathscr{P}^b(I; \mathscr{U}^h)$ *be the solution of* (18.21) *and* (18.22). *The following holds:*

$$\|u(T) - U_h(T)\|_{L^2(\Omega)} + \|u - U_h\|_{L^2(I;H^1(\Omega))} \leq ck^m \|u^{(m)}\|_{L^2(I;H^1(\Omega))} + cE_h(u) + c\sqrt{T}E_h^\infty(u), \tag{18.23}$$

for some $c > 0$, *independent of* k, h *and* u. *The optimal spacial error bounds* E_h *and* E_h^∞ *are defined as in Lemma 18.1.*

Proof Let u_h be the solution of (18.13)–(18.15). If we apply Theorem 18.1, then we find

$$\|u_h(T) - U_h(T)\|_{L^2(\Omega)} + \|u_h - U_h\|_{L^2(I;H^1(\Omega))} \leq (1 + c_\gamma \Gamma)\|u_h - \Pi_q u_h\|_{L^2(I;H^1(\Omega))}. \tag{18.24}$$

Recall the following properties of Π_q. From [15, Lemma 3.10] we get

$$\begin{aligned}\|\Pi_q(u - u_h)\|^2_{L^2(I;H^1(\Omega))} &\leq c_1 \sum_{n=1}^{N} (\|u - u_h\|^2_{L^2(I_n;H^1(\Omega))} + k\|u^n - u_h^n\|^2_{H^1(\Omega)}) \\ &\leq c_1 \|u - u_h\|^2_{L^2(I;H^1(\Omega))} + c_1 T \|u - u_h\|^2_{L^\infty(I;H^1(\Omega))}\end{aligned}$$

for some $c_1 > 0$. From [15, Theorem 3.10] we get the optimal bound

$$\|\Pi_q u - u\|_{L^2(I;H^1(\Omega))} \leq c_2 k^m \|u^{(m)}\|_{L^2(I;H^1(\Omega))}$$

for some $c_2 > 0$. Combining these two results, we get

$$\begin{aligned}\|u_h - \Pi_q u_h\|_{L^2(I;H^1(\Omega))} &\leq \|u_h - u - \Pi_q u_h + \Pi_q u\|_{L^2(I;H^1(\Omega))} + \|u - \Pi_q u\|_{L^2(I;H^1(\Omega))} \\ &\leq \|u_h - u\|_{L^2(I;H^1(\Omega))} + \|\Pi_q u - \Pi_q u_h\|_{L^2(I;H^1(\Omega))} \\ &\quad + c_2 k^m \|u^{(m)}\|_{L^2(I;H^1(\Omega))} \\ &\leq (1 + \sqrt{c_1})\|u - u_h\|_{L^2(I;H^1(\Omega))} + \sqrt{c_1 T}\|u - u_h\|_{L^\infty(I;H^1(\Omega))} \\ &\quad + c_2 k^m \|u^{(m)}\|_{L^2(I;H^1(\Omega))}.\end{aligned}$$

Combining this bound with Lemma 18.1 and (18.24), we now find

$$\|u_h(T) - U_h(T)\|_{L^2(\Omega)} + \|u_h - U_h\|_{L^2(I;H^1(\Omega))} \leq c_3 E_h(u) + c_3\sqrt{T}E_h^\infty(u) + c_3 k^m \|u^{(m)}\|_{L^2(I;H^1(\Omega))}$$

for some $c_3 > 0$, independent of k, h and u. Combining this with (18.16) and the triangle inequality completes the proof. □

The fully discrete problem (18.21) and (18.22) gives rise to a system of linear equations. Considering the structure of $D_{L^2(\Omega)}$, we see that solving (18.21) and (18.22) is equivalent to solving

$$\int_{I_n} (W'_{h,n}, X_h)_{L^2(\Omega)} + ((W_{h,n})_+^{n-1}, (X_h)_+^{n-1})_{L^2(\Omega)} + \int_{I_n} a(W_{h,n}, X_h) \tag{18.25}$$
$$= \int_{I_n} f(X_h) + (W_{h,n-1}^{n-1}, (X_h)_+^{n-1})_{L^2(\Omega)} - \int_{I_n} (\mathbf{I}_h^\partial g', X_h)_{L^2(\Omega)} - \int_{I_n} a(\Pi_q \mathbf{I}_h^\partial g, X_h)$$

for all $X_h \in \mathscr{P}_{q-1}(I_n; \mathscr{U}_0^h)$ and for all $n = 1, \ldots, N$, where $W_{h,n} = (U_h - \Pi_q \mathbf{I}_h^\partial g)|_{I_n}$ and $W_{h,0}^0 = u_0^h - \mathbf{I}_h^\partial g(0)$, where we have used that $D_{L^2(\Omega)}(\Pi_q \mathbf{I}_h^\partial g, X_h) = D_{L^2(\Omega)}(\mathbf{I}_h^\partial g, X_h)$ for all $X_h \in \mathscr{P}_{q-1}(I_n; \mathscr{U}_0^h)$, see Remark 18.1. Using (18.25), $W_{h,n} \in \mathscr{P}_{q-1}(I_n; \mathscr{U}_0^h)$ can be determined sequentially in $n = 1, \ldots, N$. Note that g' does not need to be computed explicitly to compute the term $\int_{I_n} (\mathbf{I}_h^\partial g', X_h)_{L^2(\Omega)}$, instead we can use g and the equality

$$\int_{I_n} (\mathbf{I}_h^\partial g', X_h)_{L^2(\Omega)} = (\mathbf{I}_h^\partial g^n, X_h^n)_{L^2(\Omega)} - (\mathbf{I}_h^\partial g_+^{n-1}, (X_h)_+^{n-1})_{L^2(\Omega)} - \int_{I_n} (\mathbf{I}_h^\partial g, X_h')_{L^2(\Omega)}.$$

18.4 Numerical Experiments

We consider a heat equation with a known analytic solution to validate the results from the error analysis numerically. We use this example not only to validate our theoretical results but also to explain why one does *not* obtain optimal results without the projection Π_q in (18.22). A second example is given in order to gain further insight in the behavior of the method. This second example is selected in such a way that the temporal error should dominate.

The method was implemented in the software package DROPS [10], using formulation (18.25). Note that g' in (18.25) is not known explicitly, therefore the term $\int_{I_n} (\mathbf{I}_h^\partial g', X_h)_{L^2(\Omega)}$ is implemented using integration by parts.

We consider the heat equation, i.e. (18.5) and (18.6) with $A = -\Delta$. We take $\Omega = [0, 1]^3$ and a time interval $I = [0, 1]$. We take

$$u = \exp(t(x^3 + y^2 + z)) \sin(4tz)$$

and we discretize the problem

$$u' - \Delta u = f, \tag{18.26}$$
$$u|_{\partial\Omega} = g, \tag{18.27}$$

for the appropriate right hand sides. Note that $g \neq 0$ and is time-dependent. We take a temporal step size $k = \frac{1}{N}$ and $q = 2$. For the discretization in space we take a triangulation of Ω. To obtain the triangulation $\mathbf{T}_h$ the domain Ω is divided into cubes with side length $h := \frac{1}{N_S}$ and each of the cubes is divided into six tetrahedra. We use the finite element space $\mathscr{U}^h$ which was introduced in the previous section, with $\ell = 2$. Let U_h be the discrete solution obtained by the method from previous section. The error $\|u - U_h\|_{L^2(I;H^1(\Omega))}$ is given in Table 18.1. In this table we see that the error is of order $\mathscr{O}(k^2+h^2)$, which is optimal, as predicted by Theorem 18.3. Note that the spatial error initially appears to have a higher convergence rate, this is most likely due to the third order term $h^3\|u'\|_{L^2(I;H^3(\Omega))}$ in (18.17). If we omit the projection Π_q in (18.22), then we obtain a suboptimal results of order $\mathscr{O}(k^2h^{-1/2} + h^2)$ in Table 18.2. Note that for a fixed temporal discretization refinement level k, the discretization error in space diverges with a rate $\mathscr{O}(h^{-1/2})$. This can be seen in the first column and it will be seen even more clearly in the next example. Here we see that along the diagonal we only have a space-time convergence order of 1.5.

Remark 18.4 The $h^{-1/2}$ behavior stems from the following phenomenon. Omitting the projection Π_q in (18.22), results in the following problem for the discrete

Table 18.1 Error in $L^2(I; H^1(\Omega))$-norm between u and the solution of (18.21) and (18.22)

$N_S \backslash N$	8	16	32	64	128	256	EOC_S
3	1.39E−01	1.10E−01	1.08E−01	1.07E−01	1.07E−01	1.07E−01	
6	8.91E−02	3.00E−02	2.14E−02	2.07E−02	2.07E−02	2.07E−02	2.4
12	8.66E−02	2.19E−02	6.32E−03	3.53E−03	3.28E−03	3.26E−03	2.7
24	8.65E−02	2.17E−02	5.44E−03	1.43E−03	5.71E−04	4.66E−04	2.8
48	8.65E−02	2.17E−02	5.42E−03	1.36E−03	3.47E−04	1.05E−04	2.2
EOC_T		2.0	2.0	2.0	2.0	1.7	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

Table 18.2 Error in $L^2(I; H^1(\Omega))$-norm between u and the solution of (18.21) and (18.22), if we omit the projection Π_q

$N_S \backslash N$	8	16	32	64	128	256	EOC_S
3	2.73E−01	1.26E−01	1.09E−01	1.07E−01	1.07E−01	1.07E−01	
6	4.65E−01	1.21E−01	3.75E−02	2.23E−02	2.08E−02	2.07E−02	2.4
12	7.38E−01	1.88E−01	4.82E−02	1.29E−02	4.64E−03	3.38E−03	2.6
24	1.10	2.79E−01	7.06E−02	1.80E−02	4.68E−03	1.30E−03	1.4
48	1.60	4.04E−01	1.02E−01	2.57E−02	6.54E−03	1.68E−03	−0.4
EOC_T		2.0	2.0	2.0	2.0	2.0	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

solution $\tilde{U}^h$

$$K(\tilde{U}_h - \mathbf{I}_h^\partial g, X_h) = K(u - \mathbf{I}_h^\partial g, X_h) \quad \text{for all } X_h \in \mathscr{P}^b(I; \mathscr{U}_0^h), \tag{18.28}$$

$$(\tilde{U}_h - \mathbf{I}_h^\partial g)|_{I\times\partial\Omega} = 0. \tag{18.29}$$

The analysis from Theorem 18.3 can still be applied, but now it involves the term $u_h - \mathbf{I}_h^\partial g$. Therefore the term $k^2\|(u - \mathbf{I}_h^\partial g)^{(2)}\|_{L^2(I;H^1(\Omega))}$ appears in the error bound. Take any $m \in \mathbf{N}$. By construction $(\mathbf{I}_h^\partial g)^{(m)}$ vanishes on $I \times [h, 1-h]^3$ and $(\mathbf{I}_h^\partial g)^{(m)}(t,x) = u^{(m)}(t,x)$ for all nodal points $x \in \partial\Omega$ and all $t \in I$. It follows that

$$\|\nabla(\mathbf{I}_h^\partial g)^{(m)}(t,x)\|_{\mathbf{R}^3} = \mathscr{O}(h^{-1}) \text{ for } x \in \Omega \setminus [h, 1-h]^3, t \in I.$$

Using that $|\Omega \setminus [h, 1-h]^3| = \mathscr{O}(h)$, we can conclude that for all $t \in I$

$$\|(\mathbf{I}_h^\partial g)^{(m)}(t)\|^2_{H^1(\Omega)} = \|(\mathbf{I}_h^\partial g)^{(m)}(t)\|^2_{H^1(\Omega\setminus[h,1-h]^3)} = \mathscr{O}(h^{-1}).$$

This implies that $\|(\mathbf{I}_h^\partial g)^{(m)}\|_{L^2(I;H^1(\Omega))} = \mathscr{O}(h^{-1/2})$. Noting that $\|u^{(2)}\|_{L^2(I;H^1(\Omega))}$ is bounded, we can conclude that $k^2\|(u - \mathbf{I}_h^\partial g)^{(2)}\|_{L^2(I;H^1(\Omega))} = \mathscr{O}(k^2h^{-1/2})$.

We have not proven superconvergence for the fully discrete method. However based on Theorem 18.2 and the bound (18.12) for the stationary case, we expect a $\mathscr{O}(k^3 + h^3)$ convergence order for the maximal nodal error $\max_{n=1,\dots,N}\|u^n - U_h^n\|_{L^2(\Omega)}$. This is consistent with the numerical results in Table 18.3, and can be seen along the diagonal. However, the spatial error dominates and therefore we do not see a clear temporal convergence order. We will address this again in the next example. If we omit the projection operator Π_q, we again obtain results that are suboptimal (Table 18.4). In particular we do *not* observe superconvergence.

We have established that if we omit the projection Π_q, then both for the global error and for the nodal error we do not have an optimal method. This results in errors which are sometimes orders of magnitude larger then the (optimal) errors which we can obtain by using the projection. It should also be noted that using the projection Π_q has a negligible contribution to the computational cost.

Table 18.3 Maximal nodal error $\max_{n=1,\dots,N}\|u^n - U_h^n\|_{L^2(\Omega)}$, where U_h is the solution of (18.21) and (18.22)

$N_S \backslash N$	8	16	32	64	128	256	EOC_S
3	8.11E−02	8.10E−02	8.10E−02	8.10E−02	8.10E−02	8.10E−02	
6	1.15E−02	1.10E−02	1.10E−02	1.10E−02	1.10E−02	1.10E−02	2.9
12	3.84E−03	1.58E−03	1.41E−03	1.41E−03	1.41E−03	1.41E−03	3.0
24	3.60E−03	7.52E−04	2.18E−04	1.78E−04	1.77E−04	1.77E−04	3.0
48	3.59E−03	7.33E−04	1.30E−04	3.02E−05	2.23E−05	2.21E−05	3.0
EOC_T		2.3	2.5	2.1	0.4	0.0	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

Table 18.4 Maximal nodal error $\max_{n=1,\dots,N} \|u^n - \tilde{U}_h^n\|_{L^2(\Omega)}$, where $\tilde{U}_h$ is the solution of (18.21) and (18.22), if we omit the projection Π_q

$N_S \backslash N$	8	16	32	64	128	256	EOC_S
3	1.14E−01	8.44E−02	8.14E−02	8.11E−02	8.10E−02	8.10E−02	
6	1.01E−01	2.51E−02	1.22E−02	1.11E−02	1.10E−02	1.10E−02	2.9
12	1.18E−01	2.65E−02	5.87E−03	1.87E−03	1.44E−03	1.41E−03	3.0
24	1.27E−01	2.90E−02	6.39E−03	1.36E−03	3.32E−04	1.87E−04	2.9
48	1.32E−01	3.03E−02	6.76E−03	1.46E−03	3.08E−04	6.66E−05	1.5
EOC_T		2.1	2.2	2.2	2.2	2.2	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

Table 18.5 Error in $L^2(I; H^1(\Omega))$-norm between u and the solution of (18.21) and (18.22)

$N_S \backslash N$	4	8	16	32	64	128	EOC_S
3	5.93E−02	1.49E−02	3.72E−03	9.29E−04	2.32E−04	5.80E−05	
6	5.93E−02	1.49E−02	3.72E−03	9.29E−04	2.32E−04	5.80E−05	0.0
12	5.93E−02	1.49E−02	3.72E−03	9.29E−04	2.32E−04	5.80E−05	0.0
24	5.93E−02	1.49E−02	3.72E−03	9.29E−04	2.32E−04	5.80E−05	0.0
48	5.93E−02	1.49E−02	3.72E−03	9.29E−04	2.32E−04	5.80E−05	0.0
EOC_T		2.0	2.0	2.0	2.0	2.0	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

Table 18.6 Error in $L^2(I; H^1(\Omega))$-norm between u and the solution of (18.21) and (18.22), if we omit the projection Π_q

$N_S \backslash N$	4	8	16	32	64	128	EOC_S
3	9.08E−02	2.45E−02	6.66E−03	1.81E−03	4.86E−04	1.28E−04	
6	1.45E−01	3.78E−02	1.00E−02	2.69E−03	7.27E−04	1.95E−04	−0.6
12	2.15E−01	5.52E−02	1.43E−02	3.73E−03	9.90E−04	2.65E−04	−0.4
24	3.12E−01	7.92E−02	2.02E−02	5.17E−03	1.34E−03	3.50E−04	−0.4
48	4.47E−01	1.13E−01	2.85E−02	7.22E−03	1.84E−03	4.72E−04	−0.4
EOC_T		2.0	2.0	2.0	2.0	2.0	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

In order to see some effects more clearly, we repeat the experiment with

$$u = (x + ty - z + \cos(t))^2 \sin(2t)$$

and the appropriate right hand sides in (18.26) and (18.27). Here we have $u \in H^1(I; \mathscr{U}^h)$, therefore we expect no spatial error. The error $\|u - U_h\|_{L^2(I;H^1(\Omega))}$ is given in Table 18.5. We see that there is no spatial error and we have an optimal temporal convergence order $\mathcal{O}(k^2)$. If we omit the projection Π_q, then we see in Table 18.6 that refining the spatial discretization results in a “convergence"

Table 18.7 Maximal nodal error $\max_{n=1,\ldots,N} \|u^n - U_h^n\|_{L^2(\Omega)}$, where U_h is the solution of (18.21) and (18.22)

$N_S \backslash N$	4	8	16	32	64	128	EOC_S
3	1.59E−03	2.95E−04	4.75E−05	6.98E−06	9.67E−07	1.28E−07	
6	1.61E−03	2.98E−04	4.81E−05	7.10E−06	9.92E−07	1.33E−07	−0.1
12	1.61E−03	2.99E−04	4.82E−05	7.11E−06	9.94E−07	1.33E−07	0.0
24	1.61E−03	2.99E−04	4.82E−05	7.11E−06	9.99E−07	1.33E−07	0.0
48	1.61E−03	2.99E−04	4.82E−05	7.11E−06	9.99E−07	1.34E−07	0.0
EOC_T		2.4	2.6	2.8	2.8	2.9	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

Table 18.8 Maximal nodal error $\max_{n=1,\ldots,N} \|u^n - \tilde{U}_h^n\|_{L^2(\Omega)}$, where $\tilde{U}_h$ is the solution of (18.21) and (18.22), if we omit the projection Π_q

$N_S \backslash N$	4	8	16	32	64	128	EOC_S
3	2.14E−02	4.19E−03	7.53E−04	1.26E−04	1.96E−05	2.86E−06	
6	2.79E−02	5.82E−03	1.14E−03	2.10E−04	3.63E−05	5.91E−06	−1.0
12	3.15E−02	6.84E−03	1.41E−03	2.80E−04	5.33E−05	9.63E−06	−0.7
24	3.34E−02	7.38E−03	1.57E−03	3.24E−04	6.50E−05	1.27E−05	−0.4
48	3.43E−02	7.65E−03	1.64E−03	3.47E−04	7.26E−05	1.59E−05	−0.3
EOC_T		2.2	2.2	2.2	2.3	2.2	

The estimated temporal (spatial) order of convergence EOC_T (EOC_S) is computed using the last row (column)

order $\mathcal{O}(h^{-1/2})$ for a fixed temporal discretization. This has been explained in Remark 18.4.

We also consider the maximal nodal error for this example. We see in Table 18.7 that we have no spatial error and a temporal convergence which approaches the optimal order $\mathcal{O}(k^3)$. Omitting the projection Π_q results in a negative spatial convergence order. The temporal convergence order is also suboptimal, see Table 18.8. In this case we see a convergence of second order along the diagonal.

18.5 Conclusion and Outlook

We have studied the CG-DG space-time discretization methods for second-order parabolic equations with inhomogeneous Dirichlet boundary conditions. It has been noted in [17] that a modification of the standard semi-discrete DG time discretization method is necessary to obtain optimal error bounds. Here we have proven that the fully discrete scheme with this modification has an optimal convergence rate in the energy norm. Numerical experiments confirm the predicted optimal convergence. Furthermore, we see in the numerical experiments that without this modification the (standard) CG-DG method yields suboptimal results. We are able to pinpoint the source of this suboptimality.

We consider the following topics to be of interest for future research. The optimal superconvergence results have been proven in the semi-discrete setting and observed numerically. It is not clear, yet, how to derive a superconvergence result for the fully discrete scheme. In this paper we presented the analysis for second-order parabolic equations with Dirichlet boundary conditions. This analysis can be extended to other problems, for example, problems with other boundary conditions. Furthermore, the analysis can be extended to parabolic problems where the constraints are (partially) treated by means of a Lagrange multiplier. This is the case for the Stokes problem. In this case the error analysis for the pressure Lagrange multiplier is of interest.

References

1. Alonso-Mallo, I., Cano, B.: Spectral/Rosenbrock discretizations without order reduction for linear parabolic problems. Appl. Numer. Math. **41**(2), 247–268 (2002)
2. Altmann, R., Zimmer, C.: Runge-Kutta methods for linear semi-explicit operator differential-algebraic equations. Math. Comput. **87**(309), 149–174 (2017)
3. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer, Berlin (2013)
4. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer, Berlin (2012)
5. Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems I: a linear model problem. SIAM J. Numer. Anal. **28**(1), 43–77 (1991)
6. Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems II: optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$. SIAM J. Numer. Anal. **32**(3), 706–740 (1995)
7. Eriksson, K., Johnson, C., Larsson, S.: Adaptive finite element methods for parabolic problems VI: analytic semigroups. SIAM J. Numer. Anal. **35**(4), 1315–1325 (1998)
8. Ern, A., Guermond, J.L.: Theory and Practice Of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
9. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (2010)
10. Groß, S., Peters, J., Reichelt, V., Reusken, A.: The DROPS package for numerical simulations of incompressible flows using parallel adaptive multigrid techniques. Preprint 227, IGPM, RWTH Aachen University (2002)
11. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods. Springer, New York (2008)
12. Jamet, P.: Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain. SIAM J. Numer. Anal. **15**(5), 912–928 (1978)
13. Kennedy, C.A., Carpenter, M.H.: Diagonally implicit Runge-Kutta methods for ordinary differential equations. A review. Tech. Rep. NASA/TM–2016–219173, NASA (2016)
14. Larsson, S., Thomée, V., Wahlbin, L.B.: Numerical solution of parabolic integro-differential equations by the discontinuous Galerkin method. Math. Comput. **67**(221), 45–71 (1998)
15. Schötzau, D., Schwab, C.: Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method. SIAM J. Numer. Anal. **38**(3), 837–875 (2000)
16. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer Series in Computational Mathematics, vol. 25. Springer, Berlin (2006)
17. Voulis, I., Reusken, A.: Discontinuous Galerkin time discretization methods for parabolic problems with linear constraints. J. Numer. Math. (2018). https://doi.org/10.1515/jnma-2018-0013
18. Wloka, J.: Partial Differential Equations. Cambridge University Press, Cambridge (1987)

Chapter 19
A Framework for Efficient Hierarchic Plate and Shell Elements

Michael Weise

Abstract The Mindlin–Reissner plate model is widely used for the elastic deformation simulation of moderately thick plates. Shear locking occurs in the case of thin plates, which means slow convergence with respect to the mesh size. The Kirchhoff plate model does not show locking effects, but is valid only for thin plates. One would like to have a method suitable for both thick and thin plates. Several approaches are known to deal with the shear locking in the Mindlin–Reissner plate model. In addition to the well known Mixed Interpolation of Tensorial Components (MITC) elements and other approaches based on a mixed formulation, hierarchic methods have been developed in the recent years. These are based on the Kirchhoff model and add terms to account for shear deformations. We present some of these methods and develop a new hierarchic plate formulation. This new model can be discretised by a combination of C^0 and C^1 finite elements. Numerical tests show that the new formulation is locking free and numerically efficient. We also give an extension of the model to a hierarchic Naghdi shell based on a Koiter shell formulation with unknowns in Cartesian coordinates.

19.1 Introduction

The Mindlin–Reissner plate model is widely used for the elastic deformation simulation of moderately thick plates. The weak formulation of the model features functions from H^1 and can thus be discretised with C^0 finite elements. This simple approach works if the considered plate is rather thick, but it leads to problems when used with thin plates. One observes very slow convergence with respect to the mesh size. Meshes sufficiently fine for a thick plate may yield results which are several orders of magnitude too small in case of a thin plate. This effect is called shear locking in the engineering literature due to the fact that the plate seems to be

M. Weise (✉)
TU Chemnitz, Chemnitz, Germany
e-mail: weise.michael@aol.de

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5_19

stiffer than it is with an insufficiently fine mesh. This is only one of several locking phenomena which can be observed in different use cases and for different elements such as volumetric locking, also called Poisson locking, for incompressible material or membrane locking for shells. See for example [14] for an overview.

On the other hand, the Kirchhoff plate model does not show such locking effects, but is valid only for thin isotropic plates. It excludes out-of-plane shear deformations, which are negligible for thin plates but relevant for thick plates. The weak formulation features functions from H^2 and therefore requires C^1 finite elements for a conforming discretisation.

One would like to have a method suitable for both thick and thin plates. Several approaches are known to deal with the shear locking in the Mindlin–Reissner model. In addition to the well known MITC elements, see for example [10], and other approaches based on a mixed formulation, hierarchic methods have been developed in the recent years. We focus on hierarchic methods in this article and present two new formulations. In fact we will not really discuss specific elements, but different formulations of the plate theory which may then be discretised with suitable elements.

We will present and discuss numerical examples achieved with two combinations of C^0 and C^1 elements: linear Lagrange ansatz functions and reduced Hsieh–Clough–Tocher (rHCT) ansatz functions for triangular elements and bilinear Lagrange ansatz functions and Bogner–Fox–Schmit (BFS) ansatz functions for rectangular elements. See for example [3, 9, 26] for details on rHCT elements and [5, 18] for details on BFS elements.

The article is structured as follows. Section 19.2 deals with plate formulations. A simple benchmark problem is given in Sect. 19.2.1. Section 19.2.2 introduces some basic concepts and formulas for the plate problem. Sections 19.2.3–19.2.10 present known and new plate formulations and give a short assessment of their performance with conforming elements based on the benchmark problem from Sect. 19.2.1. A more thorough numerical comparison is presented in Sect. 19.2.11. Section 19.3 presents the basic concepts of the Naghdi and Koiter shell theories in Sects. 19.3.1 and 19.3.2 and extends the new plate formulations to shells in Sects. 19.3.3 and 19.3.4 with numerical examples in Sect. 19.3.5. The article is concluded in Sect. 19.4.

Remark 19.1 What we call Mindlin–Reissner plate throughout this article should more accurately be called Mindlin plate. Reissner's plate formulation is actually slightly different. We still use the term Mindlin–Reissner plate because of its wide spread in the literature.

19.2 Plate Theory

19.2.1 A Simple Benchmark Problem

Before presenting the different plate formulations we define a benchmark problem which will be used for a first assessment of each theory in the following sections. A deeper analysis and comparison of the numerical performance is then given in Sect. 19.2.11.

We consider a square plate of length 1 with isotropic material, i. e. same material behaviour in all space directions. All edges are hard clamped which means no deflection and no bending angle. This reads $w = \theta = 0$ in the variables defined in the following section. A thick plate with thickness $d = 10^{-1}$ and a thin plate with thickness $d = 10^{-3}$ are subjected to a scaled load of d^3. See Fig. 19.1 for the qualitative resulting deflection in the thin plate case. Rough reference solutions for the maximum deflection are 1.60×10^{-2} for the thick plate and 1.38×10^{-2} for the thin plate. They have been computed with an adaptive overkill solution, see Sect. 19.2.11 for more details.

The first assessment is done with uniform refinements of one quadratic finite element comprising the whole plate as initial coarse mesh. The full comparison in

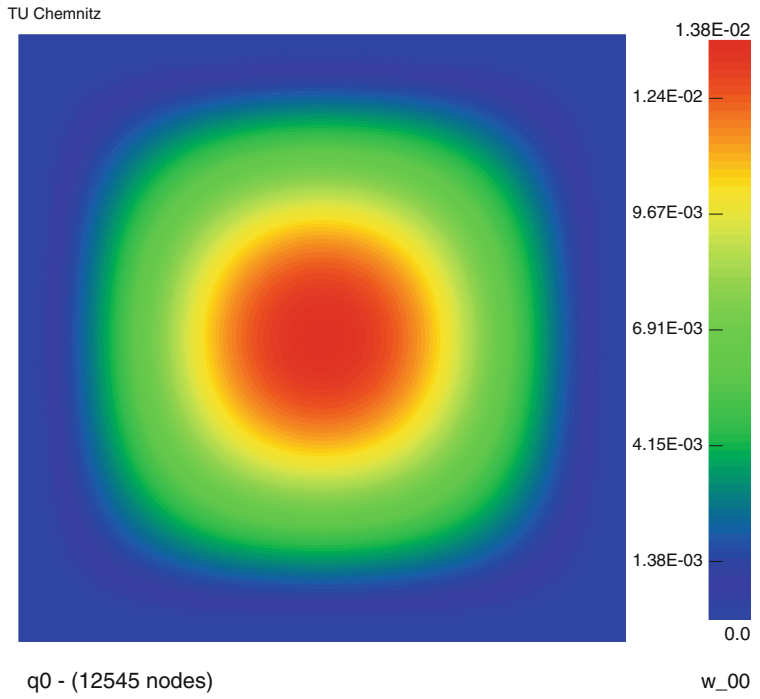


Fig. 19.1 Plate deflection under load

Sect. 19.2.11 also features triangular elements. C^0 linear Lagrange ansatz functions and C^1 reduced Hsieh–Clough–Tocher ansatz functions are used for triangular elements and C^0 bilinear Lagrange ansatz functions and C^1 Bogner–Fox–Schmit ansatz functions are used for quadratic elements.

The FE system is solved with the preconditioned conjugate gradient method. Hierarchic preconditioning is used, see [16]. The relative decrease of the residual in the preconditioned norm

$$(r_k^\mathsf{T} w_k)^{1/2} < tol\,(r_0^\mathsf{T} w_0)^{1/2}$$

with residuals and preconditioned residuals r_k, w_k of iteration k and $tol = 10^{-4}$ serves as stopping criterion. The initial residual r_0 is computed from a zero solution on the initial coarse mesh and from an interpolated solution of the next coarser mesh on all refined meshes.

19.2.2 Basic Assumptions and Formulas

Figure 19.2 depicts a deformed plate with extremely exaggerated thickness viewed from the side. The Mindlin hypothesis states that a straight line vertical to the plate midsurfaces remains a straight line after deformation, possibly with a different angle than before. The angles θ_1 and θ_2 between the original line and the same line after deformation projected onto the x_1x_3-plane and the x_2x_3-plane, respectively, are collected in the bending angle vector $\theta = [\theta_1, \theta_2]^\mathsf{T}$. We abbreviate the spatial derivative $\partial\bullet/\partial x_i$ of any object $\bullet$ with an index $\bullet_{,i}$ throughout this section. The 2D gradient of the vertical deflection w, $\nabla w = [w_{,1}, w_{,2}]^\mathsf{T} = [\partial w/\partial x_1, \partial w/\partial x_2]^\mathsf{T}$, can also be viewed as two angles in the x_1x_3-plane and the x_2x_3-plane. This allows for the definition of the shear angle vector γ which contains the shear angles in the same planes by $\gamma = \nabla w - \theta$ or, equivalently,

$$\nabla w = \theta + \gamma. \tag{19.1}$$

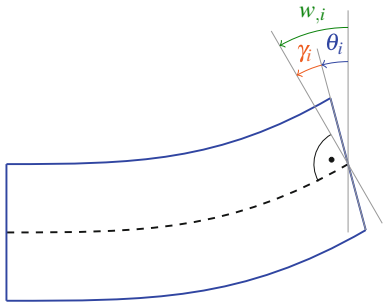


Fig. 19.2 Deformed plate from one side; thickness exaggerated

The 3D deformation of a point in the plate domain $(\eta^1, \eta^2, \tau) \in \omega \times [-d/2, d/2]$, $\omega \subset \mathbf{R}^2$ under the Mindlin hypothesis may then be described by

$$u^{3\mathrm{D}}(x_1, x_2, x_3) = \begin{bmatrix} -x_3\theta_1(x_1, x_2) \\ -x_3\theta_2(x_1, x_2) \\ w(x_1, x_2) \end{bmatrix}. \tag{19.2}$$

We collect the unknowns for any of the following plate formulations in a vector called u. Plugging the deformation ansatz (19.2) into the 3D elasticity bilinear form and integrating over the thickness direction variable x_3 results in the problem

$$\text{find } u \in \mathbf{V} \text{ with } a(u, \widetilde{u}) = l(\widetilde{u}) \quad \forall\, \widetilde{u} \in \mathbf{V}_{\text{test}} \tag{19.3}$$

with the bilinear form

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\theta) : \mathscr{C}^{\mathrm{b}} : \varepsilon(\widetilde{\theta})\, \mathrm{d}\omega + \int_\omega \gamma \cdot \mathsf{C}^{\mathrm{s}} \cdot \widetilde{\gamma}\, \mathrm{d}\omega, \tag{19.4}$$

the linear form

$$l(\widetilde{u}) = \int_\omega p\, \widetilde{w}\, \mathrm{d}\omega, \tag{19.5}$$

and appropriate FE ansatz and test spaces $\mathbf{V}$ and $\mathbf{V}_{\text{test}}$. The linearised 2D strain tensor $\varepsilon(\bullet) = \frac{1}{2}\big(\nabla \bullet + (\nabla \bullet)^{\mathsf{T}}\big)$ is given as the symmetrised 2D gradient. The fourth order material bending tensor $\mathscr{C}^{\mathrm{b}}$ and the second order shear tensor C^{s} are proportional to d^3 and d, respectively, due to the thickness integration.

The given formulation of the problem and the bilinear and linear forms features the three variables w, θ, γ. The relation $\nabla w = \theta + \gamma$ from (19.1) allows the elimination of either θ or γ from the problem. This leads to differing formulations with different numerical behaviour which are explored in the following sections.

Remark 19.2 In this article we concentrate on the "classical" plate formulations by Mindlin and Kirchhoff, which are based on non-proofable hypotheses. In contrast, two structured strategies for the derivation of refined plate theories have been developed: the so-called "direct approach" and the "consistent approach". The direct approach is based on the Cosserat theory which employs deformable directors defined directly on a surface. An overview is given for example in [2]. In the consistent approach, all quantities are developed into a series in the thickness direction with respect to a suitable basis and then truncated at different orders. See for example the works of Kienzler, Schneider, and co-workers [13, 23]. The Kirchhoff plate theory of Sect. 19.2.4 is included as a special case in this theory, but not the Mindlin plate theory.

Table 19.1 Results for the standard Mindlin–Reissner plate formulation with C^0 elements

	Bilin., thick plate		Bilin., thin plate		Biquadr., thick plate		Biquadr., thin plate	
#El	$w_{\max}$	#it	$w_{\max}$	#it	$w_{\max}$	#it	$w_{\max}$	#it
4	2.44E−3	2	2.44E−7	2	8.82E−3	4	1.39E−6	4
16	6.67E−3	9	9.26E−7	9	1.49E−2	24	4.95E−6	42
64	1.17E−2	24	3.43E−6	38	1.59E−2	42	4.88E−3	441
256	1.46E−2	40	1.34E−5	114	1.60E−2	42	1.08E−2	1444
1024	1.56E−2	48	5.31E−5	286	1.60E−2	35	1.23E−2	227
4096	1.59E−2	49	2.10E−4	662	1.60E−2	33	1.29E−2	891
16,384	1.60E−2	47	8.01E−4	1489	1.60E−2	32	1.32E−2	1426

19.2.3 The Standard Mindlin–Reissner Plate Formulation (MRs)

The standard formulation of the Mindlin–Reissner plate problem follows from eliminating $\gamma = \nabla w - \theta$ from the three variable formulation (19.2)–(19.5). This leads to the bilinear form

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\theta) : \mathscr{C}^{\mathrm{b}} : \varepsilon(\widetilde{\theta})\, \mathrm{d}\omega + \int_\omega (\nabla w - \theta) \cdot \mathsf{C}^{\mathrm{s}} \cdot (\nabla \widetilde{w} - \widetilde{\theta})\, \mathrm{d}\omega$$

with the unknowns $u = [w, \theta_1, \theta_2]^{\mathsf{T}}$. All unknowns are featured with derivatives up to first order and are thus assumed to be in $H^1(\omega)$. This allows for a simple discretisation with C^0 finite elements.

Results for the example problem from Sect. 19.2.1 with bilinear and biquadratic rectangular elements are shown in Table 19.1. One observes reasonable h-convergence for the thick plate case but very slow convergence for the thin plate case. This effect is called thickness locking or shear locking. Standard finite elements are not capable of resolving the first (bending) term and the second (shear) term of the bilinear form appropriately for small thickness d. As d tends to zero, the shear term dominates due to the scaling of the bending term with d^3 and the shear term with d. The shear term is imbalanced in the sense that it consists of a function value and derivative value.

19.2.4 The Kirchhoff Plate

The plate model of Kirchhoff can be viewed as a special case of the Mindlin–Reissner plate without the allowance for a shear angle. The condition $\gamma = 0$ gives $\theta = \nabla w$ from (19.1). The geometric interpretation of this ansatz is that orthogonal

Table 19.2 Results for the Kirchhoff plate problem with C^1 BFS elements

	Thick plate		Thin plate	
#El	$w_{\max}$	#it	$w_{\max}$	#it
4	1.45E−2	2	1.45E−2	2
16	1.38E−2	8	1.38E−2	8
64	1.38E−2	23	1.38E−2	22
256	1.38E−2	35	1.38E−2	35
1024	1.38E−2	29	1.38E−2	29
4096	1.38E−2	29	1.38E−2	28
16,384	1.38E−2	31	1.38E−2	83

line elements prior to deformation are still orthogonal to the new plate midsurface after deformation. The problem formulation changes to

$$u^{3\mathrm{D}} = \begin{bmatrix} -x_3 w_{,1} \\ -x_3 w_{,2} \\ w \end{bmatrix}, \quad a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w) : \mathscr{C}^{\mathrm{b}} : \varepsilon(\nabla \widetilde{w}) \, \mathrm{d}\omega, \quad l(\widetilde{u}) = \int_\omega p \, \widetilde{w} \, \mathrm{d}\omega$$

with the single unknown $u = w$. The bilinear form features second derivatives which calls for $w \in H^2(\omega)$. Thus, C^1 finite elements are needed for a conforming discretisation.

Results for the example problem from Sect. 19.2.1 with BFS elements are shown in Table 19.2. Even very coarse meshes yield a quite accurate solution for the thin plate. No shear locking is observed because no shear term is present. The thick plate example converges equally fast, but to the same value of 1.38×10^{-2} while the reference solution for the thick plate is about 1.60×10^{-2}. This is due to the fact that the Kirchhoff model neglects the shear term and is thus only suitable for thin plates.

19.2.5 The Mindlin–Reissner Plate in Hierarchic Formulation (MRh)

The desire for a suitable plate formulation for thick and thin plates has lead to numerous reformulations of the Mindlin–Reissner plate problem with the goal of the exclusion of locking. A hierarchic approach can be found in [11] based on earlier works, see the references therein. The cited work uses isogeometric analysis (finite elements with non-uniform rational B-splines used for geometry definition and as basis functions) but the formulation given there can also be discretised with standard finite elements.

Table 19.3 Results for the hierarchic Mindlin–Reissner formulation with C^1 BFS elements for w and C^0 bilinear elements for γ

	Thick plate		Thin plate	
#El	$w_{\max}$	#it	$w_{\max}$	#it
4	1.45E−2	2	1.45E−2	2
16	1.53E−2	10	1.38E−2	9
64	1.56E−2	32	1.38E−2	24
256	1.58E−2	73	1.38E−2	45
1024	1.59E−2	130	1.38E−2	68
4096	1.60E−2	214	1.38E−2	92
16,384	1.60E−2	423	1.38E−2	108

The idea is to eliminate $\theta = \nabla w - \gamma$ from the three variable formulation (19.2)–(19.5) instead of γ. This leads to

$$u^{\mathrm{3D}} = \begin{bmatrix} -x_3(w_{,1} - \gamma_1) \\ -x_3(w_{,2} - \gamma_2) \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w - \gamma) : \mathscr{C}^{\mathrm{b}} : \varepsilon(\nabla \widetilde{w} - \widetilde{\gamma}) \, \mathrm{d}\omega + \int_\omega \gamma \cdot \mathsf{C}^{\mathrm{s}} \cdot \widetilde{\gamma} \, \mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p \, \widetilde{w} \, \mathrm{d}\omega$$

with the variables $u = [w, \gamma_1, \gamma_2]^{\mathsf{T}}$. First derivatives of γ and second derivatives of w call for $\gamma \in H^1(\omega)$ and $w \in H^2(\omega)$. Thus, a discretisation with C^0 elements for γ and C^1 elements for w is needed. The name "hierarchic" comes from the fact that the bilinear form includes the Kirchhoff bending term and can thus be viewed as a hierarchic extension of the Kirchhoff model.

Results for the example problem from Sect. 19.2.1 are shown in Table 19.3. One observes good h-convergence to the right solutions for the thick and the thin plate. The hierarchic reformulation has rendered the problem locking free and the method performs well for thick and thin plates. There is, however, one major drawback: the iteration numbers needed to reduce the initial residual below the defined threshold are significantly higher for the thick plate than for the thin plate and also much higher than with the standard Mindlin–Reissner formulation. A fair and thorough comparison between the pure C^0 elements used there and the partially C^1 combined elements here also needs to involve some error and computation time measurement and is carried out in Sect. 19.2.11. But the reduced comparison given here still serves to show that it might be useful to search for further formulations even though the hierarchic formulation is locking free.

19.2.6 The Mindlin–Reissner Plate in a Rotation Free Formulation by Oesterle, Ramm and Bischoff (ORB)

The locking phenomenon is caused by the imbalance of function values and derivative values in the standard formulation. The authors of [22] try to overcome this issue with a so-called rotation free formulation. All variables of this formulation represent displacements and not rotation angles. Two additional plate deflections $w_{\mathrm{sb}}, w_{\mathrm{bs}}$ whose derivatives in x_1 respectively x_2 represent the shear angles need to be introduced for an equivalent formulation to the Mindlin–Reissner plate. The ansatz

$$w = w_{\mathrm{b}} + w_{\mathrm{sb}} + w_{\mathrm{bs}},\ \gamma = (w_{\mathrm{sb},1}, w_{\mathrm{bs},2})^{\mathsf{T}},\ \theta = \nabla w - \gamma = \nabla w_{\mathrm{b}} + (w_{\mathrm{bs},1}, w_{\mathrm{sb},2})^{\mathsf{T}}$$

yields the formulation

$$u^{3\mathrm{D}} = \begin{bmatrix} u_1 - x_3(w_{\mathrm{b},1} + w_{\mathrm{bs},1}) \\ u_2 - x_3(w_{\mathrm{b},2} + w_{\mathrm{sb},2}) \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon \begin{pmatrix} w_{\mathrm{b},1} + w_{\mathrm{bs},1} \\ w_{\mathrm{b},2} + w_{\mathrm{sb},2} \end{pmatrix} : \mathscr{C}^{\mathrm{b}} : \varepsilon \begin{pmatrix} \widetilde{w}_{\mathrm{b},1} + \widetilde{w}_{\mathrm{bs},1} \\ \widetilde{w}_{\mathrm{b},2} + \widetilde{w}_{\mathrm{sb},2} \end{pmatrix} \mathrm{d}\omega$$

$$+ \int_\omega \begin{pmatrix} w_{\mathrm{sb},1} \\ w_{\mathrm{bs},2} \end{pmatrix} \cdot \mathsf{C}^{\mathrm{s}} \cdot \begin{pmatrix} \widetilde{w}_{\mathrm{sb},1} \\ \widetilde{w}_{\mathrm{bs},2} \end{pmatrix} \mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p\,(\widetilde{w}_{\mathrm{b}} + \widetilde{w}_{\mathrm{sb}} + \widetilde{w}_{\mathrm{bs}})\,\mathrm{d}\omega.$$

All three variables $w_{\mathrm{b}}, w_{\mathrm{sb}}, w_{\mathrm{bs}}$ are present with second derivatives and need to be discretised with C^1 elements. Unlike the source [22] which uses isogeometric analysis we again employ standard C^1 finite elements.

Results for the example problem from Sect. 19.2.1 are shown in Table 19.4. The conjugate gradient method was stopped if the stopping criterion was not met after 5000 iterations. One observes good h-convergence to the right solutions for the thick

Table 19.4 Results for the rotation free Mindlin–Reissner formulation with C^1 BFS elements for w, w_{sb} and w_{bs}

	Thick plate		Thin plate	
#El	$w_{\max}$	#it	$w_{\max}$	#it
4	1.63E−2	3	1.45E−2	3
16	1.56E−2	34	1.38E−2	14
64	1.58E−2	265	1.38E−2	63
256	1.59E−2	910	1.38E−2	204
1024	1.60E−2	2367	1.38E−2	559
4096	1.60E−2	5000	1.38E−2	1241
16,384	1.60E−2	5000	1.38E−2	2837

and the thin plate. The rotation free reformulation has rendered the problem locking free and the method performs well for thick and thin plates. The iteration numbers, however, are extremely high in comparison with the other methods. With shear locking gone but very low numerical efficiency this method does not look promising for practical use. Maybe there exist better solvers for this problem formulation, but at least our preconditioned solver is inefficient in this case.

19.2.7 The Plate Formulation of Endo and Kimura (EKs)

The authors of [12] argue that the bending and shear deformations of the Mindlin–Reissner plate formulation can not be determined uniquely and therefore propose a different formulation. Like in the standard formulation they eliminate $\gamma = \nabla w - \theta$ from the three variable formulation (19.2)–(19.5) and make the additional assumption $\theta = \nabla w_\mathrm{b}$ for some w_b. In consequence it holds $\gamma = \nabla w_\mathrm{s}$ for $w_\mathrm{s} = w - w_\mathrm{b}$. Just like the formulation of the previous section the absence of rotation angle variables leads to a rotation free and thus also locking free formulation. One gets

$$u^{3\mathrm{D}} = \begin{bmatrix} u_1 - x_3 w_{\mathrm{b},1} \\ u_2 - x_3 w_{\mathrm{b},2} \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w_\mathrm{b}) : \mathscr{C}^\mathrm{b} : \varepsilon(\nabla \widetilde{w}_\mathrm{b})\,\mathrm{d}\omega + \int_\omega (\nabla w - \nabla w_\mathrm{b}) \cdot \mathsf{C}^\mathrm{s} \cdot (\nabla \widetilde{w} - \nabla \widetilde{w}_\mathrm{b})\,\mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p\,\widetilde{w}\,\mathrm{d}\omega.$$

Both variables w and w_b are present with second derivatives and need to be discretised with C^1 elements. One expects a slightly different solution than with the standard Mindlin–Reissner plate due to the additional condition that θ needs to be a gradient.

Numerical results are shown in Table 19.5. One observes good h-convergence for both the thick and the thin plate. The solution of the thick plate is different from the

Table 19.5 Results for the Endo–Kimura plate formulation with C^1 BFS elements for w and w_b

	Thick plate		Thin plate	
#El	w_max	#it	w_max	#it
4	1.63E−2	3	1.45E−2	3
16	1.55E−2	24	1.38E−2	119
64	1.56E−2	49	1.38E−2	1320
256	1.57E−2	96	1.38E−2	3132
1024	1.57E−2	103	1.38E−2	2666
4096	1.57E−2	139	1.38E−2	834
16,384	1.57E−2	142	1.38E−2	483

Mindlin–Reissner plate. The additional gradient condition leads to a slightly stiffer behaviour. It is unclear which solution is the "better" one compared to the behaviour of real plates. Iteration numbers are good for the thin plate but very high for the thick plate.

19.2.8 The Endo–Kimura Plate in Hierarchic Formulation (EKh)

The ansatz $\theta = \nabla w_\mathrm{b}, \gamma = w_\mathrm{s}$ of Endo and Kimura can also be combined with the hierarchic ansatz of eliminating θ instead of γ from the three variable formulation (19.2)–(19.5). This yields

$$u^{3\mathrm{D}} = \begin{bmatrix} u_1 - x_3(w_{,1} - w_{\mathrm{s},1}) \\ u_2 - x_3(w_{,2} - w_{\mathrm{s},2}) \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w - \nabla w_\mathrm{s}) : \mathscr{C}^\mathrm{b} : \varepsilon(\nabla \widetilde{w} - \nabla \widetilde{w}_\mathrm{s})\, \mathrm{d}\omega + \int_\omega (\nabla w_\mathrm{s}) \cdot \mathsf{C}^\mathrm{s} \cdot (\nabla \widetilde{w}_\mathrm{s})\, \mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p\, \widetilde{w}\, \mathrm{d}\omega.$$

The numerical results in Table 19.6 show good h-convergence to the same solution like the standard Endo–Kimura formulation. Iteration numbers are good for the thick plate but very high for the thin plate. This is exactly the other way around than with standard Endo–Kimura in the previous section. It appears that the difficulty of the problem has shifted from the thick to the thin plate, or from the shear to the bending term.

Table 19.6 Results for the hierarchic Endo–Kimura plate formulation with C^1 BFS elements for w and w_s

	Thick plate		Thin plate	
#El	w_max	#it	w_max	#it
4	1.63E−2	3	1.45E−2	3
16	1.55E−2	17	1.38E−2	9
64	1.56E−2	103	1.38E−2	28
256	1.57E−2	374	1.38E−2	48
1024	1.57E−2	706	1.38E−2	59
4096	1.57E−2	1383	1.38E−2	100
16,384	1.57E−2	2829	1.38E−2	139

19.2.9 First New Formulation: Endo–Kimura Plate Decoupled (EKd)

Another kind of elimination which was not present in [12] is possible. With the ansatz of Endo and Kimura one can also eliminate $\nabla w = \nabla w_\mathrm{b} + \nabla w_\mathrm{s}$ and $w = w_\mathrm{b} + w_\mathrm{s}$ from the equations. Only the variables w_b and w_s remain and one arrives at

$$u^{3\mathrm{D}} = \begin{bmatrix} u_1 - x_3 w_{\mathrm{b},1} \\ u_2 - x_3 w_{\mathrm{b},2} \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w_\mathrm{b}) : \mathscr{C}^\mathrm{b} : \varepsilon(\nabla \widetilde{w}_\mathrm{b})\, \mathrm{d}\omega + \int_\omega (\nabla w_\mathrm{s}) \cdot \mathsf{C}^\mathrm{s} \cdot (\nabla \widetilde{w}_\mathrm{s})\, \mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p\, (\widetilde{w}_\mathrm{b} + \widetilde{w}_\mathrm{s})\, \mathrm{d}\omega.$$

This formulation has completely decoupled bending and shear terms. The bending deflection w_b is the same as in the Kirchhoff theory and an additional shear deflection w_s may be calculated separately and added for a total deflection $w = w_\mathrm{b} + w_\mathrm{s}$ afterwards. We solve, however, the complete system. An analogous decoupled formulation is not directly obtainable from the standard Mindlin–Reissner formulation. A conforming discretisation can be achieved with C^1 elements for w_b and C^0 elements for w_s.

The numerical results in Table 19.7 show good h-convergence to the same solution like the standard Endo–Kimura formulation. Iteration numbers are very good for both the thick and the thin plate.

The decoupling has eliminated the convergence problems of the other two Endo–Kimura formulations. We have obtained a formulation which is suitable for both thick and thin plates and is numerically efficient in both cases. The only drawback is that we do not get the solution of the Mindlin–Reissner plate but of a slightly stiffer problem.

Table 19.7 Results for the decoupled Endo–Kimura plate formulation with C^1 BFS elements for w_b and C^0 bilinear elements for w_s

	Thick plate		Thin plate	
#El	w_max	#it	w_max	#it
4	1.69E−2	2	1.45E−2	2
16	1.58E−2	10	1.38E−2	9
64	1.58E−2	22	1.38E−2	22
256	1.57E−2	26	1.38E−2	35
1024	1.57E−2	24	1.38E−2	29
4096	1.57E−2	21	1.38E−2	29
16,384	1.57E−2	20	1.38E−2	33

19.2.10 Second New Formulation: Hierarchic Mindlin–Reissner Based on Endo–Kimura (MREK)

We propose to extend the formulation of the previous section with an additional term to relax the gradient condition again. In this way we arrive at a Mindlin–Reissner formulation which can be viewed as hierarchically based on the decoupled Endo–Kimura formulation. The ansatz

$$w = w_\mathrm{b} + w_\mathrm{s}, \quad \nabla w = \theta + \gamma, \quad \theta = \nabla w_\mathrm{b} - \varrho, \quad \gamma = \nabla w_\mathrm{s} + \varrho$$

and elimination of w yields

$$u^{3\mathrm{D}} = \begin{bmatrix} u_1 - x_3(w_{\mathrm{b},1} - \varrho_1) \\ u_2 - x_3(w_{\mathrm{b},2} - \varrho_2) \\ w \end{bmatrix},$$

$$a(u, \widetilde{u}) = \int_\omega \varepsilon(\nabla w_\mathrm{b} - \varrho) : \mathscr{C}^\mathrm{b} : \varepsilon(\nabla \widetilde{w}_\mathrm{b} - \widetilde{\varrho})\,\mathrm{d}\omega + \int_\omega (\nabla w_\mathrm{s} + \varrho) \cdot \mathsf{C}^\mathrm{s} \cdot (\nabla \widetilde{w}_\mathrm{s} + \widetilde{\varrho})\,\mathrm{d}\omega,$$

$$l(\widetilde{u}) = \int_\omega p\,(\widetilde{w}_\mathrm{b} + \widetilde{w}_\mathrm{s})\,\mathrm{d}\omega.$$

A conforming discretisation can be achieved with C^1 elements for w_b and C^0 elements for ϱ_1, ϱ_2 and w_s.

The numerical results in Table 19.8 show good h-convergence to the same solution like for the Mindlin–Reissner model. Iteration numbers are not as good as for the decoupled Endo–Kimura formulation but still good. They are comparable for the thick and thin plate and grow slowly with the number of elements. In comparison with the hierarchic Mindlin–Reissner formulation from Sect. 19.2.5, the iteration numbers for the thin plate are slightly bigger and those of the thick plate are somewhat higher at lower element numbers but are growing slower and thus they are lower than those of the hierarchic Mindlin–Reissner formulation at higher element numbers.

Table 19.8 Results for the Endo–Kimura based Mindlin–Reissner plate formulation with C^1 BFS elements for w_b and C^0 bilinear elements for w_s and ϱ

	Thick plate		Thin plate	
#El	$w_{\max}$	#it	$w_{\max}$	#it
4	1.69E−2	2	1.45E−2	2
16	1.60E−2	9	1.38E−2	10
64	1.60E−2	26	1.38E−2	25
256	1.60E−2	46	1.38E−2	50
1024	1.60E−2	60	1.38E−2	68
4096	1.60E−2	100	1.38E−2	100
16,384	1.60E−2	130	1.38E−2	132

19.2.11 Comparison of Numerical Results

The numerical example from Sect. 19.2.1 is explored in more detail in this section. We consider the error estimator for the energy norm of [6] for MITC elements and neglect the terms which are zero for our standard elements. The error contribution of an element T then reads

$$\eta_T^2 = h_T^2 \left(\|\mathrm{div}(\mathscr{C}^{\mathrm{b}} : \varepsilon(\theta_h)) + \mathsf{C}^{\mathrm{s}} \cdot \gamma_h\|_{L^2(T)} + (d^2 + h_T^2)\|p + \mathrm{div}(\mathsf{C}^{\mathrm{s}} \cdot \gamma_h)\|_{L^2(T)} \right)$$
$$+ \frac{1}{2} \sum_{E \in \mathbf{E}(T)} h_E \left(\|[\mathscr{C}^{\mathrm{b}} : \varepsilon(\theta_h)]_E \cdot \mathbf{n}_E\|_{L^2(E)} + (d^2 + h_E^2)\|[\mathsf{C}^{\mathrm{s}} \cdot \gamma_h]_E \cdot \mathbf{n}_E\|_{L^2(E)} \right)$$

for a finite element solution θ_h and γ_h (adapted to the actual used theory for each of the Mindlin–Reissner and Endo–Kimura plate formulations). The formula involves the element diameter h_T, the set of edges $\mathbf{E}(T)$ of the element T, the edge length h_E, the jump $[\bullet]_E$ of a quantity over the edge E and a fixed unit normal $\mathbf{n}_E$ of edge E.

Computations were done with our TU Chemnitz adaptive FEM software SPC written in Fortran, see [17]. The software features adaptive FEM but uniform refinement was employed for this tests for good comparability. The plate module of SPC features rectangular and triangular elements. The coarse mesh for our example with rectangular elements is one unit square; the coarse mesh for triangular elements is the unit square divided by one of its diagonals. Square element use Bogner–Fox–Schmit (BFS) quadrangles for C^1 parts and Lagrange elements for C^0 parts of the different formulations as detailed in the sections above. Triangular elements use reduced Hsieh–Clough–Tocher (rHCT) elements for C^1 parts and Lagrange elements for C^0 parts of the different formulations as detailed in the sections above. BFS elements feature 4 degrees of freedom (dofs) per node and unknown in the ansatz, rHCT elements 3 dofs per node and unknown. This leads to unknowns per node as given in Table 19.9.

Figure 19.3 collects the development of the error estimator for the thick and thin plates in case of square elements. The first row shows its reduction over the number

Table 19.9 Continuous and discretised unknowns of the plate formulations

Formulation	C^0 unkn.	C^1 unkn.	dofs/node (rectangle)	dofs/node (triangle)
MRs	3	0	3	3
MRh	2	1	6	5
ORB	0	3	12	9
EKs	0	2	8	6
EKh	0	2	8	6
EKd	1	1	5	4
MREK	3	1	7	6

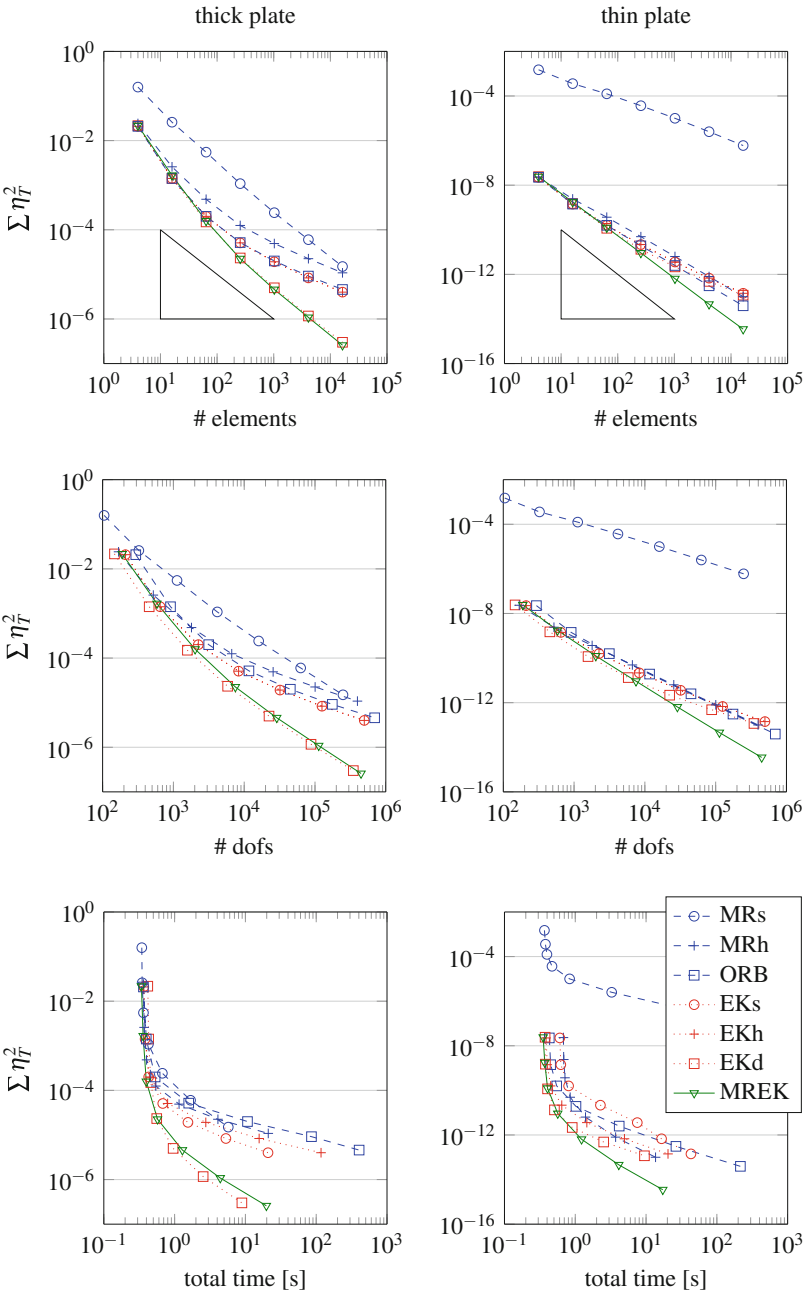


Fig. 19.3 Squared relative error estimator for square elements; triangle shows rate 1/(# elements)

of elements. All formulations show h-convergence with best results for MREK and EKd for the thick plate and MREK for the thin plate.

The second row shows the error reduction over the total number of unknowns. The unknowns are approximately linked to the element numbers by a fixed factor because of the uniform refinement. Thus, the graphs do look quite similar for both the thick and the thin plate compared to those of the first row.

The third row shows the error reduction over the total time summed over all refinements. These two are the most important graphs because only they factor in the iteration numbers via the computation time. EKd is best closely followed by MREK and they are clearly more efficient than the other formulations for the thick plate. For the thin plate this changes to MREK as the most efficient formulation followed by EKd and than the other formulations.

The results for triangular elements are collected in Fig. 19.4. The performance differences are less pronounced in this case than in the square elements case. EKd is the most efficient formulation for both the thick and the thin plate. MREK, EKh and MRh are also performing quite well in both cases. EKs is very good for thick plates but rather inefficient for the thin plate.

In summary, MREK is the most efficient Mindlin–Reissner formulation and EKd the most efficient Endo–Kimura formulation for thick and thin as well as square and triangular elements. We conclude that both our new formulations are the formulations of choice for flexibly usable thick and thin Mindlin–Reissner or Endo–Kimura plate elements. The MRh formulation comes close in performance to MREK in case of triangular elements but is clearly outperformed by MREK in case of square elements.

19.3 Shell Theory

19.3.1 The Naghdi Shell

We consider a shell, a thin-walled, possibly curved structure in 3D. The shell domain is defined by the midsurface $\mathbf{y}$ and the thickness d (significantly smaller than the midsurface dimensions) via

$$\{\mathbf{x}(\eta^1, \eta^2, \tau) = \mathbf{y}(\eta^1, \eta^2) + \tau \mathbf{a}_3(\eta^1, \eta^2) : (\eta^1, \eta^2) \in \omega \subset \mathbf{R}^2, \tau \in [-\tfrac{d}{2}, \tfrac{d}{2}]\}$$

with the tangential vectors $\mathbf{a}_1, \mathbf{a}_2$ and the unit normal vector $\mathbf{a}_3$ of the midsurface given by

$$\mathbf{a}_i = \mathbf{y}_{,i} = \frac{\partial \mathbf{y}}{\partial \eta^i},\ i = 1, 2\,, \quad \mathbf{a}_3 = \tfrac{\mathbf{a}_1 \times \mathbf{a}_2}{\|\mathbf{a}_1 \times \mathbf{a}_2\|}.$$

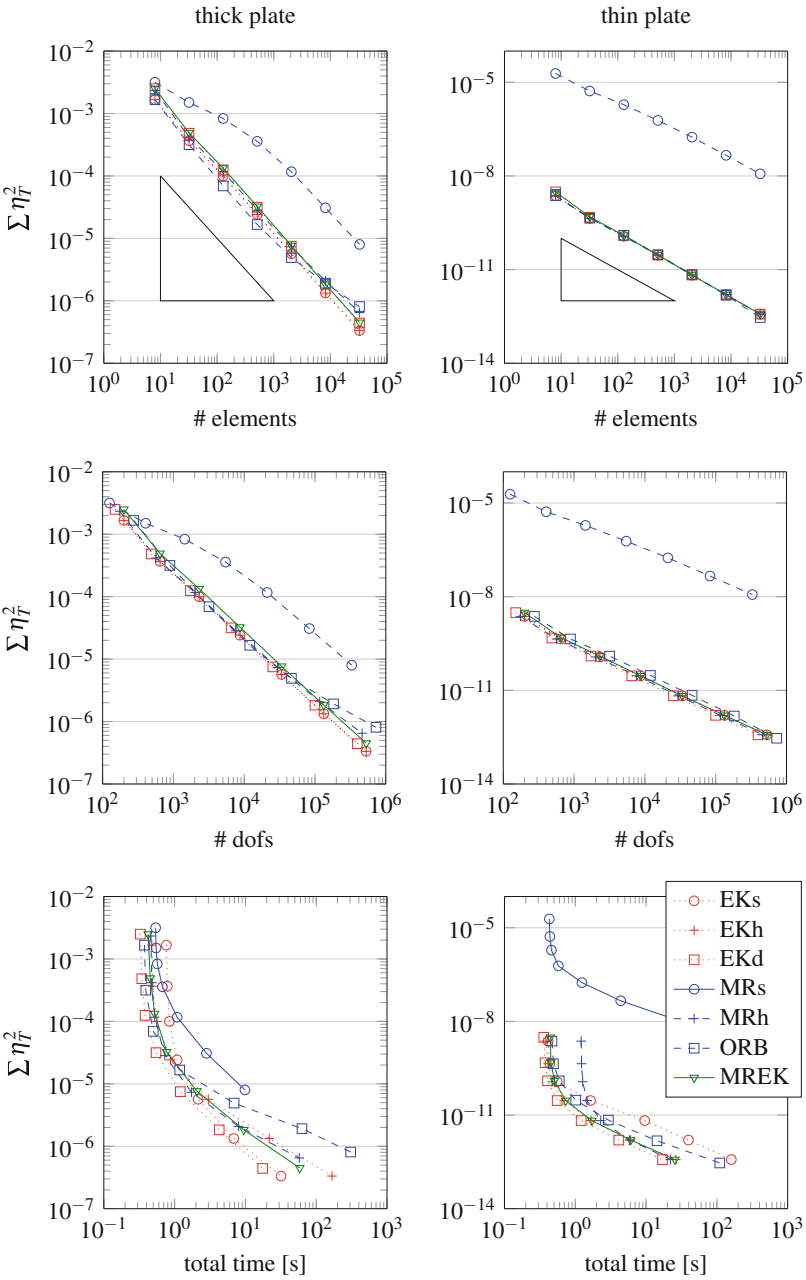


Fig. 19.4 Squared relative error estimator for triangle elements; triangle shows rate 1/(# elements)

These vectors are called covariant basis and form a biorthogonal system with the contravariant basis $\mathbf{a}^1, \mathbf{a}^2$ and $\mathbf{a}^3 = \mathbf{a}_3$. We abbreviate the parameter derivative $\partial\bullet/\partial\eta^i$ of any object $\bullet$ with an index $\bullet_{,i}$ throughout this section.

The solution of a shell deformation problem with 3D-FEM is not suitable due to the small thickness (either elements with bad aspect ratios occur or an extreme amount of elements is required). Therefore, special shell models are needed. One popular example is the Naghdi shell model, see [21] or for example [7, Sect. 4.2.2] and the references therein. It employs the Mindlin–Reissner hypotheses (normal lines remain straight after deformation, no change of thickness, plane state of stress) with the deformation ansatz

$$\mathbf{u}^{3\mathrm{D}} := \mathbf{u}(\eta^1, \eta^2) + \tau\boldsymbol{\theta}(\eta^1, \eta^2) \quad \text{with} \quad \mathbf{u} = u_1\mathbf{a}^1 + u_2\mathbf{a}^2 + u_3\mathbf{a}_3,\ \boldsymbol{\theta} = \theta_1\mathbf{a}^1 + \theta_2\mathbf{a}^2$$

in covariant coordinates with the midsurface translation $\mathbf{u}$ and the rotation $\boldsymbol{\theta}$.

19.3.1.1 Naghdi Shell in Covariant Coordinates

We collect the five unknown covariant coordinates in a vector

$$u = [u_1, u_2, u_3, \theta_1, \theta_2]^{\mathsf{T}}.$$

The weak formulation of the shell problem reads

$$a(u, \widetilde{u}) = f(\widetilde{u}) \quad \forall\, \widetilde{u}$$

with test functions $\widetilde{u}$ from an appropriate space, the bilinear form $a(\cdot,\cdot)$ accounting for the inner virtual work of the elastic deformation and the linear functional $f(\cdot)$ accounting for the load of the test functions. The bilinear form can be written as a sum of membrane, bending and shear terms with according strains in the form

$$a(u, \widetilde{u}) = a^{\mathrm{m}}(u, \widetilde{u}) + a^{\mathrm{b}}(u, \widetilde{u}) + a^{\mathrm{s}}(u, \widetilde{u}), \tag{19.6}$$

$$a^{\mathrm{m}}(u, \widetilde{u}) = \int_\omega \varepsilon^{\mathrm{m}}(u) : \mathscr{C}^{\mathrm{m}} : \varepsilon^{\mathrm{m}}(\widetilde{u})\,\mathrm{dS}, \tag{19.7}$$

$$a^{\mathrm{b}}(u, \widetilde{u}) = \int_\omega \varepsilon^{\mathrm{b}}(u) : \mathscr{C}^{\mathrm{b}} : \varepsilon^{\mathrm{b}}(\widetilde{u})\,\mathrm{dS}, \tag{19.8}$$

$$a^{\mathrm{s}}(u, \widetilde{u}) = \int_\omega \varepsilon^{\mathrm{s}}(u) : \mathscr{C}^{\mathrm{m}} : \varepsilon^{\mathrm{s}}(\widetilde{u})\,\mathrm{dS} \tag{19.9}$$

with the surface element $\mathrm{dS} = \|\mathbf{a}_1 \times \mathbf{a}_2\|\mathrm{d}\eta^1\mathrm{d}\eta^2$. The membrane, bending and shear strain tensors

$$\varepsilon^{\mathrm{m}}(u) = \Sigma_{ij}\, \varepsilon^{\mathrm{m}}_{ij}(u)\mathbf{a}^i\mathbf{a}^j,\ \varepsilon^{\mathrm{b}}(u) = \Sigma_{ij}\, \varepsilon^{\mathrm{b}}_{ij}(u)\mathbf{a}^i\mathbf{a}^j,\ \varepsilon^{\mathrm{s}}(u) = \Sigma_i\, \varepsilon^{\mathrm{s}}_i(u)(\mathbf{a}^i\mathbf{a}_3 + \mathbf{a}_3\mathbf{a}^i)$$

have the covariant coordinates

$$\begin{aligned}
\varepsilon^{\mathrm{m}}_{ij}(u) &= \tfrac{1}{2}(u_{i|j} + u_{j|i}) - b_{ij}u_3 \quad \text{(not dependent on } \theta_i\text{)},\\
\varepsilon^{\mathrm{b}}_{ij}(u) &= \tfrac{1}{2}(\theta_{i|j} + \theta_{j|i} - \Sigma_k\, b^k_i u_{k|j} - \Sigma_k\, b^k_j u_{k|i}) + c_{ij}u_3,\\
\varepsilon^{\mathrm{s}}_{i}(u) &= \tfrac{1}{2}(\theta_i + u_{3,i} + \Sigma_j\, b^j_i u_j).
\end{aligned} \tag{19.10}$$

The notation above uses the first fundamental form $a_{ij} = \mathbf{a}_i \cdot \mathbf{a}_j$ or $a^{ij} = \mathbf{a}^i \cdot \mathbf{a}^j$, the second fundamental form $b_{ij} = \mathbf{a}_{i,j} \cdot \mathbf{a}_3 = -\mathbf{a}_i \cdot \mathbf{a}_{3,j}$ or $b^j_i = \Sigma_k\, b_{ik}a^{kj}$, the third fundamental form $c_{ij} = \Sigma_l\, b^l_i b_{lj} = \Sigma_{kl}\, b_{ik}a^{kl}b_{lj}$ and the covariant derivative $u_{i|j} = u_{i,j} - \Sigma_k\, \Gamma^k_{ij}u_k$ with the Christoffel symbols $\Gamma^k_{ij} = \mathbf{a}_{i,j} \cdot \mathbf{a}^k = \Sigma_l\, a^{kl}(a_{il,j} + a_{jl,i} - a_{ij,l})/2$. Indices take the values 1 and 2 whenever a (multi-)sum Σ without an index range appears. The membrane and bending material tensors $\mathscr{C}^{\mathrm{m}}$ and $\mathscr{C}^{\mathrm{b}}$ are obtained by integrating the material tensor $\mathscr{C}$ and $\tau^2\mathscr{C}$, respectively, over the thickness of the shell. If the material is constant over the thickness this leads to $\mathscr{C}^{\mathrm{m}} = d\mathscr{C}$, $\mathscr{C}^{\mathrm{b}} = \frac{d^3}{12}\mathscr{C}$. For a complete derivation of the above formulas we again refer to [7].

The action of the second order tensor $\varepsilon^{\mathrm{s}}(u) = \Sigma_i\, \varepsilon^{\mathrm{s}}_i(u)(\mathbf{a}^i\mathbf{a}_3 + \mathbf{a}_3\mathbf{a}^i)$ with the fourth order material tensor $\mathscr{C}^{\mathrm{m}}$ may be expressed by a replacement first order tensor $2\mathbf{e}^{\mathrm{s}}(u) = \Sigma_i\, \varepsilon^{\mathrm{s}}_i(u)\mathbf{a}^i$ with an appropriately chosen reduced second order shear tensor C^{s} which reads

$$a^{\mathrm{s}}(u, \widetilde{u}) = \int_\omega \mathbf{e}^{\mathrm{s}}(u) \cdot \mathsf{C}^{\mathrm{s}} \cdot \mathbf{e}^{\mathrm{s}}(\widetilde{u})\, \mathrm{dS}. \tag{19.11}$$

The shell midsurface can be given as a non-uniform rational B-spline (NURBS) surface for practical applications like outlined in [8] and [27].

19.3.1.2 Coordinate Free Naghdi Shell Formulation

The traditional Naghdi shell formulation of the previous section uses the covariant coordinates as unknowns. The shell formulation itself can also be written in a coordinate free formulation which is independent of a coordinate system for the unknowns. With the surface gradient ∇_{S} defined by

$$\nabla_{\mathrm{S}} \bullet = \Sigma_i \mathbf{a}^i \bullet_{,i}$$

the equivalent of the angle decomposition (19.1) from plate theory reads $\nabla_{\mathrm{S}}\mathbf{u} \cdot \mathbf{a}_3 = \boldsymbol{\theta} + \boldsymbol{\gamma}$. With the first and second fundamental tensors $\mathsf{A} = a_{ij}\mathbf{a}^i\mathbf{a}^j$ and $\mathsf{B} = b_{ij}\mathbf{a}^i\mathbf{a}^j$ the shell strains may then be expressed by

$$\begin{aligned}
2\varepsilon^{\mathrm{m}} &= \nabla_{\mathrm{S}}\mathbf{u} \cdot \mathsf{A} + \mathsf{A} \cdot (\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}},\\
2\varepsilon^{\mathrm{s}} &= \boldsymbol{\gamma}\mathbf{a}_3 + \mathbf{a}_3\boldsymbol{\gamma}, \quad 2\mathbf{e}^{\mathrm{s}} = \boldsymbol{\gamma},\\
2\varepsilon^{\mathrm{b}} &= -\nabla_{\mathrm{S}}\boldsymbol{\theta} \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_{\mathrm{S}}\boldsymbol{\theta})^{\mathsf{T}} - \nabla_{\mathrm{S}}\mathbf{u} \cdot \mathsf{B} - \mathsf{B} \cdot (\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}}.
\end{aligned} \tag{19.12}$$

The elimination of $\boldsymbol{\gamma} = \nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \boldsymbol{\theta}$ changes the shear strain tensor to

$$2\varepsilon^s = (\nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \boldsymbol{\theta})\mathbf{a}_3 + \mathbf{a}_3(\nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \boldsymbol{\theta}), \quad 2\mathbf{e}^s = \nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \boldsymbol{\theta}. \tag{19.13}$$

See also [20] for a full derivation (with $\boldsymbol{\theta}$ oriented differently, $\boldsymbol{\theta}_{\text{here}} = -\boldsymbol{\theta}_{\text{there}}$). The bilinear form stays as in (19.6) and (19.11).

19.3.2 The Koiter Shell

While the Naghdi shell is the shell equivalent to the Mindlin–Reissner plate, the Koiter shell is the equivalent to the Kirchhoff plate. The additional hypothesis reads

$$\boldsymbol{\theta} = \nabla_S \mathbf{u} \cdot \mathbf{a}_3, \quad \boldsymbol{\gamma} = \nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \boldsymbol{\theta} = 0$$

for shells. It is a valid assumption only for thin shells.

The Koiter shell model can be derived in different ways from 3D elasticity or directly as a 2D model. We refer to [1] and [24] for further reading on this topic.

19.3.2.1 Coordinate Free Koiter Shell Formulation

Direct insertion of the hypothesis into the coordinate free Naghdi shell formulation yields $\mathbf{e}^s = 0$ and

$$\begin{aligned} 2\varepsilon^m &= \nabla_S \mathbf{u} \cdot \mathsf{A} + \mathsf{A} \cdot (\nabla_S \mathbf{u})^{\mathsf{T}}, \\ 2\varepsilon^b &= -\nabla_S(\nabla_S \mathbf{u} \cdot \mathbf{a}_3) \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_S(\nabla_S \mathbf{u} \cdot \mathbf{a}_3))^{\mathsf{T}} - \nabla_S \mathbf{u} \cdot \mathsf{B} - \mathsf{B} \cdot (\nabla_S \mathbf{u})^{\mathsf{T}} \\ &= -\nabla_S \nabla_S \mathbf{u} \cdot \mathbf{a}_3 \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_S \nabla_S \mathbf{u} \cdot \mathbf{a}_3)^{\mathsf{T}} \end{aligned}$$

with the chain rule $\nabla_S(\nabla_S \mathbf{u} \cdot \mathbf{a}_3) = \nabla_S \nabla_S \mathbf{u} \cdot \mathbf{a}_3 + \nabla_S \mathbf{u} \cdot \nabla_S \mathbf{a}_3 = \nabla_S \nabla_S \mathbf{u} \cdot \mathbf{a}_3 - \nabla_S \mathbf{u} \cdot \mathsf{B}$. See also [19] and [20, Sect. 5.2] for a full derivation.

19.3.2.2 Koiter Shell in Cartesian Coordinates

Writing the Koiter shell in covariant coordinates like the Naghdi shell leads to formulas which are not so easy to implement for general geometries, see [7, Sect. 4.2.3]. A conforming discretisation of this formulation requires C^0 elements for both tangential coordinates and C^1 elements for the normal coordinate.

Alternatively, one can use Cartesian coordinates for the unknowns, see [4]. The coordinate free formulation from above can be reformulated to

$$\begin{aligned} 2\varepsilon^{\mathrm{m}} &= \nabla_{\mathrm{S}}\mathbf{u}\cdot\mathsf{A}+\mathsf{A}\cdot(\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}} \\ &= \mathbf{u}_{,j}\cdot\mathbf{a}_i+\mathbf{u}_{,i}\cdot\mathbf{a}_j, \\ 2\varepsilon^{\mathrm{b}} &= -\nabla_{\mathrm{S}}\nabla_{\mathrm{S}}\mathbf{u}\cdot\mathbf{a}_3\cdot\mathsf{A}-\mathsf{A}\cdot(\nabla_{\mathrm{S}}\nabla_{\mathrm{S}}\mathbf{u}\cdot\mathbf{a}_3)^{\mathsf{T}} \\ &= 2(\mathbf{u}_{,ij}\cdot\mathbf{a}_3-\Gamma_{ij}^{k}\mathbf{u}_{,k}\cdot\mathbf{a}_3)\mathbf{a}^i\mathbf{a}^j. \end{aligned}$$

There is no need for further derivative evaluation if the unknowns are given in Cartesian coordinates $\mathbf{u}=u_x\mathbf{e}_x+u_y\mathbf{e}_y+u_z\mathbf{e}_z$. The Cartesian basis vectors $\mathbf{e}_x,\mathbf{e}_y,\mathbf{e}_z$ are fixed in space and thus the derivatives carry over directly to the unknown coordinates themselves;

$$\begin{aligned} \mathbf{u}_{,i} &= (\tfrac{\partial}{\partial\eta^i}u_x)\mathbf{e}_x+(\tfrac{\partial}{\partial\eta^i}u_y)\mathbf{e}_y+(\tfrac{\partial}{\partial\eta^i}u_z)\mathbf{e}_z, \\ \mathbf{u}_{,ij} &= (\tfrac{\partial^2}{\partial\eta^i\partial\eta^j}u_x)\mathbf{e}_x+(\tfrac{\partial^2}{\partial\eta^i\partial\eta^j}u_y)\mathbf{e}_y+(\tfrac{\partial^2}{\partial\eta^i\partial\eta^j}u_z)\mathbf{e}_z. \end{aligned}$$

This formulation can then easily be implemented. A conforming discretisation requires C^1 finite elements for all three coordinates u_x,u_y,u_z.

19.3.3 The Principle of Endo–Kimura Applied to Shells

The basic idea of Endo and Kimura can also be applied to the Naghdi shell formulation. The total deformation $\mathbf{u}$ is split into a bending and a shear part with $\mathbf{u}=\mathbf{u}_{\mathrm{b}}+\mathbf{u}_{\mathrm{s}}$. We start from the coordinate free Naghdi formulation of (19.12) and (19.13)

$$\begin{aligned} 2\varepsilon^{\mathrm{m}} &= \nabla_{\mathrm{S}}\mathbf{u}\cdot\mathsf{A}+\mathsf{A}\cdot(\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}}, \\ 2\mathbf{e}^{\mathrm{s}} &= \nabla_{\mathrm{S}}\mathbf{u}\cdot\mathbf{a}_3-\boldsymbol{\theta}, \\ 2\varepsilon^{\mathrm{b}} &= -\nabla_{\mathrm{S}}\boldsymbol{\theta}\cdot\mathsf{A}-\mathsf{A}\cdot(\nabla_{\mathrm{S}}\boldsymbol{\theta})^{\mathsf{T}}-\nabla_{\mathrm{S}}\mathbf{u}\cdot\mathsf{B}-\mathsf{B}\cdot(\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}} \end{aligned}$$

and insert the assumption $\boldsymbol{\theta}=\nabla_{\mathrm{S}}\mathbf{u}_{\mathrm{b}}\cdot\mathbf{a}_3$ to obtain

$$\begin{aligned} 2\varepsilon^{\mathrm{m}} &= \nabla_{\mathrm{S}}\mathbf{u}\cdot\mathsf{A}+\mathsf{A}\cdot(\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}}, \\ 2\mathbf{e}^{\mathrm{s}} &= \nabla_{\mathrm{S}}\mathbf{u}\cdot\mathbf{a}_3-\nabla_{\mathrm{S}}\mathbf{u}_{\mathrm{b}}\cdot\mathbf{a}_3, \\ 2\varepsilon^{\mathrm{b}} &= -\nabla_{\mathrm{S}}(\nabla_{\mathrm{S}}\mathbf{u}_{\mathrm{b}}\cdot\mathbf{a}_3)\cdot\mathsf{A}-\mathsf{A}\cdot(\nabla_{\mathrm{S}}(\nabla_{\mathrm{S}}\mathbf{u}_{\mathrm{b}}\cdot\mathbf{a}_3))^{\mathsf{T}}-\nabla_{\mathrm{S}}\mathbf{u}\cdot\mathsf{B}-\mathsf{B}\cdot(\nabla_{\mathrm{S}}\mathbf{u})^{\mathsf{T}} \end{aligned}$$

as the shell equivalent to the standard Endo–Kimura plate formulation. Hierarchic and decoupled versions can again be obtained by using the unknowns $\mathbf{u}$ and $\mathbf{u}_{\mathrm{s}}$ with

elimination of $\mathbf{u}_\mathrm{b}$ for the hierarchic version or by using $\mathbf{u}_\mathrm{b}$ and $\mathbf{u}_\mathrm{s}$ with elimination of $\mathbf{u}$ for the decoupled version. Just like in the special case of a plate the additional gradient condition leads to a slightly stiffer solution than that of the Naghdi shell.

The decoupled version reads

$$\begin{aligned}
2\varepsilon^\mathrm{m} &= \nabla_\mathrm{S}(\mathbf{u}_\mathrm{b} + \mathbf{u}_\mathrm{s}) \cdot \mathsf{A} + \mathsf{A} \cdot (\nabla_\mathrm{S}(\mathbf{u}_\mathrm{b} + \mathbf{u}_\mathrm{s}))^\mathsf{T}, \\
2\mathbf{e}^\mathrm{s} &= \nabla_\mathrm{S}\mathbf{u}_\mathrm{s} \cdot \mathbf{a}_3, \\
2\varepsilon^\mathrm{b} &= -\nabla_\mathrm{S}(\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} \cdot \mathbf{a}_3) \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_\mathrm{S}(\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} \cdot \mathbf{a}_3))^\mathsf{T} \\
&\quad -\nabla_\mathrm{S}(\mathbf{u}_\mathrm{b} + \mathbf{u}_\mathrm{s})) \cdot \mathsf{B} - \mathsf{B} \cdot (\nabla_\mathrm{S}(\mathbf{u}_\mathrm{b} + \mathbf{u}_\mathrm{s})))^\mathsf{T}
\end{aligned}$$

and the product rule $\nabla_\mathrm{S}(\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} \cdot \mathbf{a}_3) = \nabla_\mathrm{S}\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} + \nabla_\mathrm{S}\mathbf{u}_\mathrm{b}\mathsf{B}$ together with $\mathsf{B} \cdot \mathsf{A} = \mathsf{B}$ yields the simplified bending strain

$$2\varepsilon^\mathrm{b} = -(\nabla_\mathrm{S}\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} \cdot \mathbf{a}_3) \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_\mathrm{S}\nabla_\mathrm{S}\mathbf{u}_\mathrm{b} \cdot \mathbf{a}_3)^\mathsf{T} - \nabla_\mathrm{S}\mathbf{u}_\mathrm{s} \cdot \mathsf{B} - \mathsf{B} \cdot (\nabla_\mathrm{S}\mathbf{u}_\mathrm{s})^\mathsf{T}.$$

The bending strain features both $\mathbf{u}_\mathrm{b}$ and $\mathbf{u}_\mathrm{s}$. Therefore, this formulation is actually not totally decoupled like in the special case of a plate. We still keep the name "decoupled" to link it with the corresponding plate formulation.

We have not yet specified if $\mathbf{u}_\mathrm{s}$ is from $\mathbf{R}^3$ or a smaller subspace. From now on we use $\mathbf{u}_\mathrm{s} \in \mathrm{span}\{\mathbf{a}_3\}$ because it is the equivalent of the plate deflection w_s. In contrast, $\mathbf{u}_\mathrm{b}$ is from the whole $\mathbf{R}^3$ because it emulates not only the deflection but also the membrane deformations. It is sensible to use the normal coordinate $\mathbf{u}_\mathrm{s} = w_\mathrm{s}\mathbf{a}_3$ as an unknown with this assumption. The coordinates of $\mathbf{u}_\mathrm{b}$ on the other hand are best expressed in Cartesian coordinates via $\mathbf{u}_\mathrm{b} = u_{\mathrm{b}x}\mathbf{e}_x + u_{\mathrm{b}y}\mathbf{e}_y + u_{\mathrm{b}z}\mathbf{e}_z$. This yields the covariant coordinates

$$\begin{aligned}
2\varepsilon^\mathrm{m}_{ij} &= \mathbf{u}_{\mathrm{b},i} \cdot \mathbf{a}_j + \mathbf{u}_{\mathrm{b},j} \cdot \mathbf{a}_i - 2w_\mathrm{s}b_{ij}, \\
2e^\mathrm{s}_i &= w_{\mathrm{s},i}, \\
2\varepsilon^\mathrm{b}_{ij} &= -2\mathbf{u}_{\mathrm{b},ij} \cdot \mathbf{a}_3 + 2\Sigma_k\, \Gamma^k_{ij}\mathbf{u}_{\mathrm{b},k} \cdot \mathbf{a}_3 + 2w_\mathrm{s}c_{ij}
\end{aligned}$$

which can be used for implementing the decoupled Endo–Kimura shell with C^1 elements for the three Cartesian coordinates of $\mathbf{u}_\mathrm{b}$ and C^0 elements for the normal coordinate w_s of $\mathbf{u}_\mathrm{s}$.

19.3.4 Hierarchic Naghdi Shell Formulation Based on Endo–Kimura

A relaxation of the decoupled Endo–Kimura shell formulation to obtain again the Naghdi shell solution can be done analogously to the relaxation of the according

plate model in Sect. 19.2.10. Adding the relaxation angle $\boldsymbol{\varrho} = \varrho_1 \mathbf{a}^1 + \varrho_2 \mathbf{a}^2$ with $\boldsymbol{\theta} = \nabla_S \mathbf{u}_b \cdot \mathbf{a}_3 - \boldsymbol{\varrho}$ and $\boldsymbol{\gamma} = \nabla_S \mathbf{u}_s \cdot \mathbf{a}_3 + \boldsymbol{\varrho}$ results in the shear and bending strains

$$
\begin{aligned}
2\mathbf{e}^s &= \nabla_S \mathbf{u}_s \cdot \mathbf{a}_3 + \boldsymbol{\varrho}, \\
2\varepsilon^b &= -(\nabla_S \nabla_S \mathbf{u}_b \cdot \mathbf{a}_3) \cdot \mathsf{A} - \mathsf{A} \cdot (\nabla_S \nabla_S \mathbf{u}_b \cdot \mathbf{a}_3)^T - \nabla_S \mathbf{u}_s \cdot \mathsf{B} - \mathsf{B} \cdot (\nabla_S \mathbf{u}_s)^T \\
&\quad + \nabla_S \boldsymbol{\varrho} \cdot \mathsf{A} + \mathsf{A} \cdot (\nabla_S \boldsymbol{\varrho})^T
\end{aligned}
$$

with the covariant coordinates

$$
\begin{aligned}
2e_i^s &= w_{s,i} + \varrho_i, \\
2\varepsilon_{ij}^b &= -2\mathbf{u}_{b,ij} \cdot \mathbf{a}_3 + 2\Sigma_k \, \Gamma_{ij}^k \mathbf{u}_{b,k} \cdot \mathbf{a}_3 + 2w_s c_{ij} + \varrho_{i,j} + \varrho_{j,i} - 2\Sigma_k \, \Gamma_{ij}^k \varrho_k .
\end{aligned}
$$

The membrane strains are the same as above. A conforming discretisation may be achieved with C^1 elements for the three Cartesian coordinates of $\mathbf{u}_b$ and C^0 elements for the normal coordinate w_s of $\mathbf{u}_s$ and the two covariant coordinates of $\boldsymbol{\varrho}$.

19.3.5 Numerical Example: Scordelis–Lo Roof

The Scordelis–Lo roof was used to validate the exactness of the models and the implementation. The problem with the vertical deflection u_z^{3D} for is illustrated in Fig. 19.5. A roof section cut out from a right circular cylinder with an opening angle of 80° rests on a rigid diaphragm with its curved edges. The normal and tangential deformation with respect to the spanning circle are zero at the curved edges (together with their first derivatives in tangential direction); the straight edges are free. The shell thickness is 0.25, the cylinder radius 25 and the length 50. Isotropic material

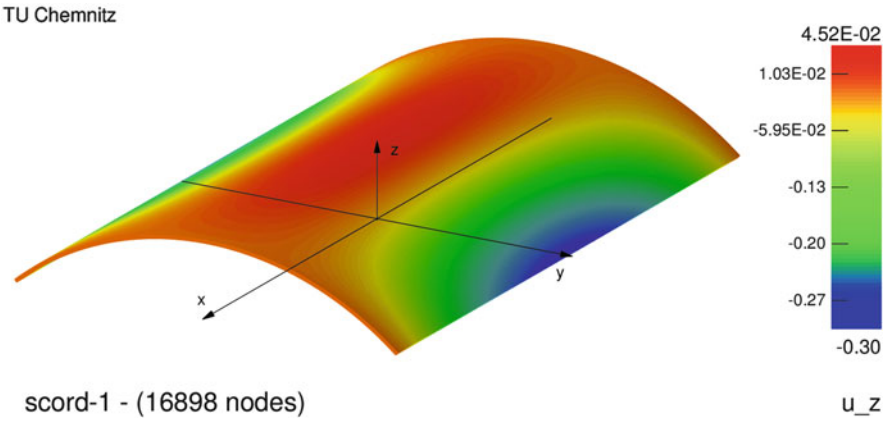


Fig. 19.5 Scordelis-Lo roof with vertical deflection u_z^{3D}

Table 19.10 Results and iteration numbers for Scordelis–Lo roof

	Naghdi bilin.		Naghdi biquadr.		EK shell		Hier. Naghdi/EK	
#El	Max$\lvert u_z^{3D}\rvert$	#it	Max$\lvert u_z^{3D}\rvert$	#it	Max$\lvert u_z^{3D}\rvert$	#it	Max$\lvert u_z^{3D}\rvert$	#it
4	2.02E−3	9	3.74E−2	36	0.2138	77	0.2809	29
16	7.72E−3	34	0.1472	267	0.2959	643	0.2960	653
64	1.99E−2	108	0.2829	661	0.3025	1011	0.3025	1047
256	3.76E−2	255	0.3007	545	0.3030	1065	0.3031	744
1024	7.27E−2	587	0.3026	423	0.3030	605	0.3032	978
4096	0.1485	1184	0.3038	452	0.3030	709	0.3034	1629
16,384	0.2362	1764	0.3043	534	0.3030	670	0.3036	2788

with $E = 4.32 \cdot 10^8$ and $\nu = 0$ is used and a vertical load per unit surface of -90 is applied. The absolute value of the maximum vertical deflection of $\max|u_z^{3D}| = 0.3024$, which occurs at the middle of the free edges, is suggested to be used as a benchmark in [15]. Our own overkill solution with 137 962 adaptively refined biquadratic Lagrangian elements obtained $\max|u_z^{3D}| = 0.3043$. The same value for the first four decimal places was obtained with the standard Naghdi shell model using 16,384 uniform biquadratic Lagrangian elements, see Table 19.10. Further uniform refinements were not possibly on the used machine due to running out of memory.

We did not make use of symmetries and simulated the whole domain. Results obtained with uniform refinement of bilinear as well as biquadratic Lagrangian elements for the standard Naghdi formulation and a combination of bilinear Lagrangian and Bogner–Fox–Schmit elements for the C^0 and C^1 parts of the Endo–Kimura shell formulation as well as the hierarchic Naghdi shell formulation based on Endo–Kimura are collected in Table 19.10. Iteration numbers of all methods are much higher than for their counterparts in the plate example. The iteration numbers of the Koiter shell in cartesian coordinates with BFS elements which are not shown in the table are about the same as those of the Endo–Kimura shell. The bilinear elements show slow h-convergence but the h-convergence of the other methods appears satisfying.

In order to conduct a locking study we also considered a thinner Scordelis–Lo roof with thickness 0.025 and load -0.9. Results are collected in Table 19.11. We observe that both Endo–Kimura shell formulations have the best h-convergence but at extreme iteration numbers. A better preconditioner tailored for these formulations which might reduce these iteration numbers would be desirable. The Koiter shell in cartesian unknowns with BFS elements shows iteration numbers in the same order of magnitude.

Table 19.11 Results and iteration numbers for thinner Scordelis–Lo roof

	Naghdi biquadr.		EK shell		Hier. Naghdi/EK	
#El	Max$\lvert u_z^{3D}\rvert$	#it	Max$\lvert u_z^{3D}\rvert$	#it	Max$\lvert u_z^{3D}\rvert$	#it
4	3.77E−3	36	0.1707	59	0.2983	30
16	2.16E−2	271	0.2181	1317	0.2184	1347
64	0.1107	1310	0.3067	5382	0.3067	6071
256	0.2404	2949	0.8183	9494	0.3183	9655
1024	0.3020	2712	0.3204	5615	0.3205	5636
4096	0.3022	501	0.3206	3019	0.3206	3050
16,384	0.3197	3426	0.3206	4545	0.3206	5272

19.4 Conclusion and Outlook

We have compared several known and two new plate formulations with respect to their locking behaviour and their numerical efficiency combined with a conforming discretisation. The two new formulations for the Endo–Kimura variant and the original Mindlin–Reissner plate turned out to be locking free and among the most efficient methods for both thick and thin plates. Further experiments with other preconditioners should be conducted to round out this picture.

There are currently no existence results for the new formulations, which started out just as numerical experiments. Providing existence results and a priori error estimates is desirable after the shown promising experiments.

Extensions of both new formulations to shells have been presented. The example problem has shown that the solution of shell problems is more challenging and a better preconditioner might be needed. Nevertheless both methods are locking free and thus suitable for thick and thin shells.

An application to anisotropic plates is directly possible by use of an anisotropic material tensor. The presented theory does not rely on any restrictions or special material formulations except for the usual symmetries of the material tensor, which are fulfilled also by anisotropic ones. Plates and shells with varying material properties over the thickness can also be covered with effective 2D material tensors which stem from integration in the thickness direction of the 3D material tensor. Especially piecewise constant material properties with respect to the plate or shell thickness play an important rule and are analysed under the name of laminate theory. Adaptive FEM for anisotropic laminated Kirchhoff plates has been addressed in [25] and [28].

References

1. Altenbach, H., Zhilin, P.A.: The Theory of Simple Elastic Shells, pp. 1–12. Springer, Berlin (2004)
2. Altenbach, J., Altenbach, H., Eremeyev, V.A.: On generalized cosserat-type theories of plates and shells: a short review and bibliography. Arch. Appl. Mech. **80**(1), 73–92 (2010)

3. Bernadou, M., Hassan, K.: Basis functions for general Hsieh-Clough-Tocher triangles, complete or reduced. Int. J. Numer. Methods Eng. **17**(5), 784–789 (1981)
4. Blouza, A., Le Dret, H.: Existence and uniqueness for the linear Koiter model for shells with little regularity. Q. Appl. Math. **57**(2), 317–337 (1999)
5. Bogner, F.K., Fox, R.L., Schmit, L.A.: The generation of interelement compatible stiffness and mass matrices by the use of interpolation formulas. In: Proceedings of the Conference on Matrix Methods in Structural Mechanics, pp. 397–444 (1965)
6. Carstensen, C., Hu, J.: A posteriori error analysis for conforming MITC elements for reissner-mindlin plates. Math. Comput. **77**(262), 611–632 (2008)
7. Chapelle, D., Bathe, K.: The Finite Element Analysis of Shells: Fundamentals. Computational Fluid and Solid Mechanics. Springer, Heidelberg (2003)
8. Cho, M., Choi, J., Roh, H.Y.: Integration of shell FEA with geometric modeling on NURBS surface representation for practical applications. Comput. Model. Eng. Sci. **33**(1), 17–47 (2008)
9. Ciarlet, P.G.: Interpolation error estimates for the reduced Hsieh-Clough-Tocher triangle. Math. Comput. **32**(142), 335–344 (1978)
10. Dvorkin, E.N.: Nonlinear analysis of shells using the MITC formulation. Arch. Comput. Meth. Eng. **2**(2), 1–50 (1995)
11. Echter, R.: Isogeometric analysis of shells. Dissertation, Bericht Nr. 59, Institut für Baustatik und Baudynamik der Universität Stuttgart (2013)
12. Endo, M., Kimura, N.: An alternative formulation of the boundary value problem for the Timoshenko beam and Mindlin plate. J. Sound Vibr. **301**(1), 355–373 (2007)
13. Kienzler, R., Schneider, P.: Consistent theories of isotropic and anisotropic plates. J. Theor. Appl. Mech. **50**(3), 755–768 (2012)
14. Koschnick, F.: Geometrische Locking-Effekte bei Finiten Elementen und ein allgemeines Konzept zu ihrer Vermeidung. Dissertation, Technische Universität München (2004)
15. MacNeal, R.H., Harder, R.L.: A proposed standard set of problems to test finite element accuracy. Finite Elem. Anal. Des. **1**(1), 3–20 (1985)
16. Meyer, A.: Hierarchical preconditioners for higher order elements and applications in computational mechanics. Preprint Series of the SFB 393 SFB393/99-02, Technische Universität Chemnitz (1999)
17. Meyer, A.: Programmer's manual for adaptive finite element code SPC–PM 2Ad. Preprint Series of the SFB 393 SFB393/01-18, Technische Universität Chemnitz (2001)
18. Meyer, A.: Hierarchical preconditioners and adaptivity for Kirchhoff–plates. Chemnitz Scientific Computing Preprints CSC/08-03, Technische Universität Chemnitz (2008)
19. Meyer, A.: The Koiter shell equation in a coordinate free description - extended. Chemnitz Scientific Computing Preprints CSC/13-01, Technische Universität Chemnitz (2013)
20. Meyer, A.: The linear Naghdi shell equation in a coordinate free description. Chemnitz Scientific Computing Preprints CSC/13-03, Technische Universität Chemnitz (2013)
21. Naghdi, P.: Foundations of elastic shell theory. Prog. Solid Mech. **4**, 1–90 (1963)
22. Oesterle, B., Ramm, E., Bischoff, M.: A shear deformable, rotation-free isogeometric shell formulation. Comput. Methods Appl. Mech. Eng. **307**(suppl. C), 235–255 (2016)
23. Schneider, P., Kienzler, R., Böhm, M.: Modeling of consistent second order plate theories for anisotropic materials. Zeitschrift für Angewandte Mathematik und Mechanik **94**(1–2), 21–42 (2014)
24. Steigmann, D.J.: Koiter's shell theory from the perspective of three-dimensional nonlinear elasticity. J. Elast. **111**(1), 91–107 (2013)
25. Weise, M.: Adaptive FEM for fibre-reinforced 3D structures and laminates. Dissertation, Technische Universität Chemnitz (2014). http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-150439
26. Weise, M.: Simplified calculation of rHCT basis functions for an arbitrary splitting. Chemnitz Scientific Computing Preprints CSC/15-01, Technische Universität Chemnitz (2015)

27. Weise, M.: Adaptive FEM for NURBS surface shells. Proc. Appl. Math. Mech. **16**(1), 773–774 (2016)
28. Weise, M.: Residual error estimation for anisotropic Kirchhoff plates. Appl. Numer. Math. **125**, 10–22 (2018)

Index

T. Apel et al. (eds.), *Advanced Finite Element Methods with Applications*,
Lecture Notes in Computational Science and Engineering 128,
https://doi.org/10.1007/978-3-030-14244-5

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Tutorials
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). Tutorials are lecture notes typically arising via summer schools or similar events, which are used to teach graduate students. These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgement on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

– at least 100 pages of text;
– a table of contents;
– an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
– a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact the Editor for CSE at Springer at the planning stage, see *Addresses* below.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Only works in English will be considered. For evaluation purposes, manuscripts may be submitted in print or electronic form, in the latter case, preferably as pdf- or zipped ps-files. Authors are requested to use the LaTeX style files available from Springer at http://www.springer.com/gp/authors-editors/book-authors-editors/manuscript-preparation/5636 (Click on LaTeX Template → monographs or contributed books).

For categories ii) and iii) we strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. Electronic material can be included if appropriate. Please contact the publisher.

Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online.

5. The following terms and conditions hold. Categories i), ii) and iii):

Authors receive 50 free copies of their book. No royalty is paid.
Volume editors receive a total of 50 free copies of their volume to be shared with authors, but no royalties.

Authors and volume editors are entitled to a discount of 40 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Springer secures the copyright for each volume.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
griebel@ins.uni-bonn.de

David E. Keyes
Mathematical and Computer Sciences
and Engineering
King Abdullah University of Science
and Technology
P.O. Box 55455
Jeddah 21534, Saudi Arabia
david.keyes@kaust.edu.sa

and

Department of Applied Physics
and Applied Mathematics
Columbia University
500 W. 120 th Street
New York, NY 10027, USA
kd2112@columbia.edu

Risto M. Nieminen
Department of Applied Physics
Aalto University School of Science
and Technology
00076 Aalto, Finland
risto.nieminen@aalto.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
dirk.roose@cs.kuleuven.be

Tamar Schlick
Department of Chemistry
and Courant Institute
of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA
schlick@nyu.edu

Editor for Computational Science
and Engineering at Springer:

Martin Peters
Springer-Verlag
Mathematics Editorial IV
Tiergartenstrasse 17
69121 Heidelberg, Germany
martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*
2. H.P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.
3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*
4. P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*
5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*
6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.
7. R. von Schwerin, ***M**ulti **B**ody **S**ystem **SIM**ulation.* Numerical Methods, Algorithms, and Software.
8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
9. T.J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*
10. H.P. Langtangen, A.M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing*.
11. B. Cockburn, G.E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.
12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.
13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*
14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*
15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*
16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.
17. B.I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*
18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*
19. I. Babuška, P.G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*
20. T.J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.
21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.
23. L.F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H.H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T.J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M.A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H.P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B.N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III*.

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV*.

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science.*

67. I.H. Tuncer, Ü. Gülcat, D.R. Emerson, K. Matsuno (eds.), *Parallel Computational Fluid Dynamics 2007.*

68. S. Yip, T. Diaz de la Rubia (eds.), *Scientific Modeling and Simulations*.

69. A. Hegarty, N. Kopteva, E. O'Riordan, M. Stynes (eds.), *BAIL* 2008 – *Boundary and Interior Layers.*

70. M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XVIII.*

71. B. Koren, C. Vuik (eds.), *Advanced Computational Methods in Science and Engineering.*

72. M. Peters (ed.), *Computational Fluid Dynamics for Sport Simulation.*

73. H.-J. Bungartz, M. Mehl, M. Schäfer (eds.), *Fluid Structure Interaction II - Modelling, Simulation, Optimization.*

74. D. Tromeur-Dervout, G. Brenner, D.R. Emerson, J. Erhel (eds.), *Parallel Computational Fluid Dynamics 2008.*

75. A.N. Gorban, D. Roose (eds.), *Coping with Complexity: Model Reduction and Data Analysis.*

76. J.S. Hesthaven, E.M. Rønquist (eds.), *Spectral and High Order Methods for Partial Differential Equations.*

77. M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*.

78. Y. Huang, R. Kornhuber, O.Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XIX*.

79. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations V*.

80. P.H. Lauritzen, C. Jablonowski, M.A. Taylor, R.D. Nair (eds.), *Numerical Techniques for Global Atmospheric Models*.

81. C. Clavero, J.L. Gracia, F.J. Lisbona (eds.), *BAIL 2010 – Boundary and Interior Layers, Computational and Asymptotic Methods*.

82. B. Engquist, O. Runborg, Y.R. Tsai (eds.), *Numerical Analysis and Multiscale Computations*.

83. I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (eds.), *Numerical Analysis of Multiscale Problems*.

84. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*.

85. J. Blowey, M. Jensen (eds.), *Frontiers in Numerical Analysis - Durham 2010*.

86. O. Kolditz, U.-J. Gorke, H. Shao, W. Wang (eds.), *Thermo-Hydro-Mechanical-Chemical Processes in Fractured Porous Media - Benchmarks and Examples*.

87. S. Forth, P. Hovland, E. Phipps, J. Utke, A. Walther (eds.), *Recent Advances in Algorithmic Differentiation*.

88. J. Garcke, M. Griebel (eds.), *Sparse Grids and Applications*.

89. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VI*.

90. C. Pechstein, *Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems*.

91. R. Bank, M. Holst, O. Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XX*.

92. H. Bijl, D. Lucor, S. Mishra, C. Schwab (eds.), *Uncertainty Quantification in Computational Fluid Dynamics*.

93. M. Bader, H.-J. Bungartz, T. Weinzierl (eds.), *Advanced Computing*.

94. M. Ehrhardt, T. Koprucki (eds.), *Advanced Mathematical Models and Numerical Techniques for Multi-Band Effective Mass Approximations*.

95. M. Azaïez, H. El Fekih, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2012*.

96. F. Graziani, M.P. Desjarlais, R. Redmer, S.B. Trickey (eds.), *Frontiers and Challenges in Warm Dense Matter*.

97. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Munich 2012*.

98. J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXI*.

99. R. Abgrall, H. Beaugendre, P.M. Congedo, C. Dobrzynski, V. Perrier, M. Ricchiuto (eds.), *High Order Nonlinear Numerical Methods for Evolutionary PDEs - HONOM 2013*.

100. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VII*.

101. R. Hoppe (ed.), *Optimization with PDE Constraints - OPTPDE 2014*.

102. S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab, H. Yserentant (eds.), *Extraction of Quantifiable Information from Complex Systems*.

103. A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso (eds.), *Numerical Mathematics and Advanced Applications - ENUMATH 2013*.

104. T. Dickopf, M.J. Gander, L. Halpern, R. Krause, L.F. Pavarino (eds.), *Domain Decomposition Methods in Science and Engineering XXII*.

105. M. Mehl, M. Bischoff, M. Schäfer (eds.), *Recent Trends in Computational Engineering - CE2014*. Optimization, Uncertainty, Parallel Algorithms, Coupled and Complex Problems.

106. R.M. Kirby, M. Berzins, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM'14*.

107. B. Jüttler, B. Simeon (eds.), *Isogeometric Analysis and Applications 2014*.

108. P. Knobloch (ed.), *Boundary and Interior Layers, Computational and Asymptotic Methods – BAIL 2014*.

109. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Stuttgart 2014*.

110. H. P. Langtangen, *Finite Difference Computing with Exponential Decay Models*.

111. A. Tveito, G.T. Lines, *Computing Characterizations of Drugs for Ion Channels and Receptors Using Markov Models*.

112. B. Karazösen, M. Manguoğlu, M. Tezer-Sezgin, S. Göktepe, Ö. Uğur (eds.), *Numerical Mathematics and Advanced Applications - ENUMATH 2015*.

113. H.-J. Bungartz, P. Neumann, W.E. Nagel (eds.), *Software for Exascale Computing - SPPEXA 2013-2015*.

114. G.R. Barrenechea, F. Brezzi, A. Cangiani, E.H. Georgoulis (eds.), *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*.

115. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VIII*.

116. C.-O. Lee, X.-C. Cai, D.E. Keyes, H.H. Kim, A. Klawonn, E.-J. Park, O.B. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXIII*.

117. T. Sakurai, S.-L. Zhang, T. Imamura, Y. Yamamoto, Y. Kuramashi, T. Hoshi (eds.), *Eigenvalue Problems: Algorithms, Software and Applications in Petascale Computing*. EPASA 2015, Tsukuba, Japan, September 2015.

118. T. Richter (ed.), *Fluid-structure Interactions*. Models, Analysis and Finite Elements.

119. M.L. Bittencourt, N.A. Dumont, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2016*. Selected Papers from the ICOSAHOM Conference, June 27-July 1, 2016, Rio de Janeiro, Brazil.

120. Z. Huang, M. Stynes, Z. Zhang (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016.*

121. S.P.A. Bordas, E.N. Burman, M.G. Larson, M.A. Olshanskii (eds.), *Geometrically Unfitted Finite Element Methods and Applications*. Proceedings of the UCL Workshop 2016.

122. A. Gerisch, R. Penta, J. Lang (eds.), *Multiscale Models in Mechano and Tumor Biology*. Modeling, Homogenization, and Applications.

123. J. Garcke, D. Pflüger, C.G. Webster, G. Zhang (eds.), *Sparse Grids and Applications - Miami 2016*.

124. M. Schäfer, M. Behr, M. Mehl, B. Wohlmuth (eds.), *Recent Advances in Computational Engineering*. Proceedings of the 4th International Conference on Computational Engineering (ICCE 2017) in Darmstadt.

125. P.E. Bjørstad, S.C. Brenner, L. Halpern, R. Kornhuber, H.H. Kim, T. Rahman, O.B. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXIV*. 24th International Conference on Domain Decomposition Methods, Svalbard, Norway, February 6–10, 2017.

126. F.A. Radu, K. Kumar, I. Berre, J.M. Nordbotten, I.S. Pop (eds.), *Numerical Mathematics and Advanced Applications – ENUMATH 2017*.

127. X. Roca, A. Loseille (eds.), 27th International Meshing Roundtable.

128. Th. Apel, U. Langer, A. Meyer, O. Steinbach (eds.), Advanced Finite Element Methods with Applications. Selected Papers from the 30th Chemnitz Finite Element Symposium 2017.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/3527

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: www.springer.com/series/7417

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition
2. A. Quarteroni, F. Saleri, P. Gervasio, *Scientific Computing with MATLAB and Octave.* 4th Edition
3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition
4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.
5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.
6. H. P. Langtangen, *A Primer on Scientific Programming with Python.* 5th Edition
7. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai, *Elements of Scientific Computing.*
8. B. Gustafsson, *Fundamentals of Scientific Computing.*
9. M. Bader, *Space-Filling Curves.*
10. M. Larson, F. Bengzon, *The Finite Element Method: Theory, Implementation and Applications.*
11. W. Gander, M. Gander, F. Kwok, *Scientific Computing: An Introduction using Maple and MATLAB.*
12. P. Deuflhard, S. Röblitz, *A Guide to Numerical Modelling in Systems Biology*.
13. M. H. Holmes, *Introduction to Scientific Computing and Data Analysis*.
14. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with MATLAB/Octave.
15. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with Python.
16. H.P. Langtangen, S. Linge, *Finite Difference Computing with PDEs* - A Modern Software Approach.
17. B. Gustafsson, *Scientific Computing from a Historical Perspective*.
18. J. A. Trangenstein, *Scientific Computing*. Volume I - Linear and Nonlinear Equations.

19. J. A. Trangenstein, *Scientific Computing*. Volume II - Eigenvalues and Optimization.

20. J. A. Trangenstein, *Scientific Computing*. Volume III - Approximation and Integration.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/5151