

hp-finite elements for pde-constrained
optimal control problems
with focus on
distributed control and fast solvers

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Katharina Hofer

aus

Wien

Bonn 2016

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Sven Beuchler
 2. Gutachter: Prof. Dr. Daniel Wachsmuth
- Tag der Promotion: 20.07.2016
- Erscheinungsjahr: 2016

Abstract

In this thesis hp -finite element methods are applied to linear quadratic optimal control problems subject to partial differential equations.

In particular two kind of model problems are considered: a boundary control problem and a distributed optimal control problem. Both problems are discretized with variational discretization due to Hinze, that means only the state and the adjoint are discretized, whereas the control is discretized implicitly via the projection formula.

Due to the projection formula, which separates the domain in active and inactive parts, a rough knowledge on the regularity of the solution is given. Since the parts with low regularity are at the interfaces between active and inactive set and corners of the domain, there h -refinement has to be applied. In all other parts of the domain, p -refinement can be applied. In the case of boundary control all interfaces between active and inactive sets are at the boundary. Suitable hp -refinements – as boundary concentrated refinement or vertex concentrated refinement – are already known. Here, a third suitable refinement, a Neumann boundary concentrated refinement with additional h -refinement in corners is proposed.

In case of distributed optimal control the interfaces between active and inactive sets have to be determined. Then, a h -refinement on the interface between these sets combined with either using a-priori information only or in combination with error estimators is suggested.

Both model problems are solved with a semismooth Newton method. For the optimal boundary control problem several numerical experiments in three dimensions are presented. For the distributed optimal control problem several two-dimensional examples are considered. For both problems the results are compared with uniform h -refinement. The numerical experiments show in both cases a decrease of the number of degrees of freedom compared to the L_2 -error. Furthermore, these examples demonstrate, that the proposed refinement strategies for the distributed optimal control problem work very well.

The final part of this thesis considers the efficient solution of the discretized optimization problems. Here the semismooth Newton method, where in each iteration step a linear system of algebraic equations has to be solved, is used. The algebraic equation system can be written as symmetric but indefinite problem and has a saddle point structure. Here three different iterative solvers with preconditioners which belong to Krylov subspace methods are investigated.

The main results are the h , p and α independent condition number in case of using the Schöberl-Zulehner PCG in combination with suitable preconditioners. For the MINRES at least h and p independent iteration numbers are possible. Furthermore a preconditioner for the GMRES is proposed. At the end, for all three Krylov subspace methods numerical examples are presented in order to confirm the theoretical results.

Zusammenfassung

In dieser Arbeit werden hp -Finite Elemente Methoden auf linear-quadratische Optimalsteuerungsprobleme mit Nebenbedingungen aus elliptischen partiellen Differentialgleichungen behandelt.

Es werden zwei Modellprobleme, ein Randsteuerungsproblem und ein verteilte Steuerungsproblem betrachtet und mit variationeller Diskretisierung nach Hinze diskretisiert. Das heißt, nur der Zustand und die Adjungierte, nicht aber die Steuerung, werden als Finite Elemente Funktion dargestellt. Die Steuerung wird über die Projektionsformel berechnet und ist daher im Allgemeinen keine Finite Elemente Funktion.

Die Projektionsformel teilt das Gebiet in aktive und inaktive Mengen. An den Schnittstellen dieser Mengen ist aufgrund der Projektionsformel die Regularität geringer. Darüberhinaus ist die Regularität in Ecken des Gebietes geringer. Darauf aufbauend können geeignete hp -Verfeinerungsstrategien, das heißt h -Verfeinerung in allen Elementen mit geringerer Regularität und p -Verfeinerung sonst, entwickelt werden.

Beim Randsteuerungsproblem wird auf bekannten hp -Verfeinerungsverfahren für diese Problemklasse aufgebaut und ein modifiziertes Verfeinerungskonzept, einer Randkonzentrierten Verfeinerung nur am Neumannrand mit zusätzlicher Eckenverfeinerung, vorgeschlagen.

Beim verteilten Steuerungsproblem werden zwei hp -Verfeinerungen vorgeschlagen. In beiden Fällen wird die niedrigere Regularität durch die Projektionsformel beachtet, in einer vorgeschlagenen Verfeinerung wird zusätzlich Regularitätsinformation aus Fehlerschätzern verwendet.

Beide Probleme werden mit der halbglatte Newtonmethode gelöst. Weiters werden numerische Beispiele präsentiert, um die vorgeschlagenen Verfeinerungen zu testen und mit uniformer h -Verfeinerung zu vergleichen. Besonders hervorgehoben werden sollen dabei die Beispiele im Dreidimensionalen, wo für verschiedene Beispiele eine Randkonzentrierte Verfeinerung auf Randsteuerungsbeispiele angewandt wird.

Im letzten Teil der Arbeit wird die effiziente Lösung von diskretisierten Optimalsteuerungsproblemen betrachtet. Bei der halbglatte Newtonmethode muss in jedem Iterationsschritt ein lineares System von algebraischen Gleichungen gelöst werden. Dieses System kann als symmetrisches aber indefinites Problem geschrieben werden und hat eine Sattelpunktstruktur. In dieser Arbeit werden drei verschiedene Krylov-Unterraumverfahren mit Vorkonditionieren untersucht. Die Hauptresultate dabei sind die Anwendung des Schöberl-Zulehner PCG, der auf eine h , p und α unabhängige Konditionszahl führt und die Anwendung von Blockdiagonalvorkonditionierern beim MINRES die zumindest h und p unabhängige Iterationszahlen ermöglicht. Weiters wird ein geeigneter Vorkonditionierer für den GMRES vorgestellt. Abschließend werden für alle drei Krylov-Unterraumverfahren numerische Ergebnisse präsentiert um die theoretischen Resultate zu belegen.

Acknowledgements

First of all, I want to thank Prof. Dr. Sven Beuchler for giving me the opportunity to work on the project ‘higher-order finite element methods for optimal control problems’ and for supervising my thesis. I am very grateful for his support and encouragement.

Second, I want to thank Prof. Dr. Daniel Wachsmuth for co-supervising, inspiring discussions and for co-examining my thesis.

Third, I want to acknowledge the support of my colleague Dr. Jan-Eric Wurst, who substantially helped to develop the C++ code, which is used for the numerical results in this thesis and with whom I enjoyed several constructive and valuable discussions on the C++ code.

Furthermore, I appreciate the help of Prof. Dr. Veronika Pillwein with evaluating the integral over the inactive set exactly, even though we did not use it in the end. Further thanks go to Prof. Dr. Walter Zulehner for enabling to stay some days in Linz and giving some hints in order to develop the saddle point chapter. Moreover, I would like to thank my colleagues at Bonn, especially Christian Kuske for proofreading my thesis. I also thank James Munns for hints to improve the English.

Further thanks go to my family and my friends for their multifaceted support. Especially, I would like to thank Thorsten, who double-checked the English in parts of the draft version of this thesis and Bernadette (my favourite proofreader), who improved the English in the final version.

Last but not least I would like to thank you the Austrian Science Fund FWF, which supported me under the grant P23484-N18.

List of symbols and abbreviations

General notation

$a.e.$	almost everywhere
d	dimension of the space
\mathcal{A}^*	hermitian version of matrix \mathcal{A}
$\ \cdot\ _2$	euclidean norm
$\text{dist}(x, y)$	distance between the points x and y
κ	condition number
$L_i(x)$	i -th Legendre polynomial
$\hat{L}_i(x)$	i -th integrated (scaled) Legendre polynomial
$\tilde{L}_i(x)$	i -th integrated (unscaled) Legendre polynomial
$\mathcal{L}(V, W)$	the set of linear and continuous operators from V to W
$\text{meas}(Z)$	the measure of the set Z
\bar{Z}	closure of the set Z
Π_k	space of polynomials with maximal polynomial degree k
V^*	dual space of space V
A^*	adjoint operator of A
Ω	domain
$\Gamma_{\mathcal{D}}$	Dirichlet boundary
$\Gamma_{\mathcal{N}}$	Neumann boundary
$L_p(\Omega)$	the Banach space of p -times Lebesgue-integrable functions
$L_\infty(\Omega)$	the Banach space of essentially bounded functions
$W_p^k(\Omega)$	Sobolev space of functions whose weak derivatives up to order k are in $L_p(\Omega)$
$H^k(\Omega)$	the Hilbert space $W_2^k(\Omega)$
$\tilde{B}_\beta^2(\Omega)$	countably normed spaces
δ_{ij}	Kronecker delta
Id	the identity mapping
#	number of

Finite elements

M_N	mass matrix
K_N	stiffness matrix
N	number of degrees of freedom
\hat{K}	reference element
K	a finite element
F_K	Jacobian matrix
J_K	determinant of the Jacobian matrix
h_K	mesh size of element K

h	mesh size
p	polynomial degree
u^*, u_N^*	continuous and discrete solution to an equation system

Optimal control

α	regularization parameter
y_d	desired state
y, u, q	state, control and adjoint (state)
$\mathfrak{A}, \mathfrak{I}$	active and inactive set
U_{ad}	admissible set of $U_{ad} := \{u \in L_2(U) : u_a \leq u \leq u_b \text{ a.e. in } U\}$
$P_{U_{ad}}$	projection onto the feasible set U_{ad}

Saddle point problem

C_M	preconditioner for mass matrix
C_Y	preconditioner for Y_N
\mathcal{P}_{cg}	preconditioner for Schöberl-Zulehner PCG
$\mathcal{P}_{\text{minres}}$	(diagonal) preconditioner for MINRES
$\mathcal{P}_{\text{gmres}}$	preconditioner for GMRES

Abbreviations

fem	finite element method
CG, PCG	conjugate gradient method, preconditioned conjugate gradient method
MINRES	minimal residual method
GMRES	generalized minimal residual method
ASM	additive Schwarz methods
pde	partial differential equation
bc-refinement	boundary concentrated refinement
neubdry-refinement	bc-refinement on Neumann edges and additional h -refinement in corners
nic-refinement	neighbour interface concentrated refinement
errest-refinement	nic-refinement combined with error estimators

Contents

Acknowledgements	ix
List of symbols and abbreviations	xi
Introduction	1
0.1 Model problem	1
0.2 Discretization - finite element methods	2
0.3 Optimization	2
0.3.1 Semismooth Newton method	3
0.3.2 Saddle point formulation	3
0.3.3 High-order fem for optimal control	4
0.4 Outline of the thesis	5
1 Preliminaries	7
1.1 Matrices	7
1.2 Krylov subspace methods	8
1.2.1 PCG	9
1.2.2 MINRES	10
1.2.3 GMRES	11
1.3 Perturbations in equation systems	12
1.4 Integrated Legendre polynomials	13
1.5 Functional analysis	14
1.6 Spaces	15
1.7 Sobolev spaces	16
2 Optimal control problems for pdes	19
2.1 Boundary control problem	21
2.2 Distributed control problem	22
3 The finite element method	25
3.1 Basics of finite element method	25
3.1.1 Finite element and triangulation	27
3.2 High-order finite element method	28
3.2.1 Reference element and basis functions	29
3.2.2 Mapping	30
3.2.3 Element matrices	31
3.2.4 Assembling	32
3.2.5 Hanging nodes and projector	32
3.2.6 Refinement strategies	38
3.2.7 Error estimates	41

3.3	Fast solvers	42
3.3.1	hp -preconditioners	48
3.3.2	Extension to hanging nodes	50
3.4	Numerical experiments	51
4	Optimal control problems with semismooth Newton	55
4.1	Discretization	55
4.2	Semismooth Newton method	58
4.3	Optimal boundary control problem	59
4.3.1	Two-dimensional case	60
4.3.2	Three-dimensional case	65
4.4	Distributed optimal control problem	77
4.4.1	Refinement strategies	78
4.4.2	Integration on interface elements	80
4.4.3	Suitable starting mesh	86
4.4.4	Numerical examples	87
5	Saddle point problem	99
5.1	Schöberl-Zulehner PCG	100
5.2	MINRES	104
5.2.1	Exact inverse as preconditioners	105
5.2.2	Block diagonal preconditioner	105
5.3	GMRES	106
5.4	Application to an optimal control problem	106
5.4.1	Discrete saddle point problem	112
5.4.2	Application of preconditioned CG to the discrete problem	113
5.4.3	Application of preconditioned MINRES to the discrete system	118
5.4.4	Application of GMRES to discrete system	121
5.5	Numerical experiments	121
5.5.1	Square	121
5.5.2	Hole	130
	Conclusion and Outlook	135
	Bibliography	137

Introduction

This thesis considers the application of hp -finite element methods to optimal control problems subject to partial differential equations. Such kind of problems appear in many cases, since the modeling of technical processes often leads to a description by partial differential equations, whose parameters have to be optimized. In many cases additional inequality constraints on certain parameters come into play. This is due to the fact that modulation of technical limitations, e.g. maximal temperatures, have to be considered. Possible applications are for example in fluid mechanics, heart medicine, vascular surgery or crystal growing.

The investigation of optimal control problems started in the 1970's, see e.g. [90, 105, 108, 158] and gained interest over the last decade due to the increase in computational power, which enables to calculate the numerical solution to more and more real life problems.

0.1 Model problem

For the convenience of the reader a simple optimal control problem is given in order to explain the main parts of this thesis.

Considered is a linear-quadratic distributed optimal control problem

$$\min_{y,u} J(y, u) = \min_{y,u} \left(\frac{1}{2} \|y - y_d\|_{L_2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L_2(\Omega)}^2 \right)$$

subject to the elliptic partial differential equation

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) && \text{for } x \in \Omega, \\ y(x) &= 0 && \text{for } x \in \Gamma = \partial\Omega, \end{aligned}$$

in the domain Ω and its boundary Γ . Moreover, there hold box constraints $u_a, u_b \in L_2(\Omega)$ on the *control* u , that means

$$u_a \leq u(x) \leq u_b \quad \text{almost everywhere (a.e.) in } \Omega.$$

y_d denotes the *desired state*, y the *state* and α is called *regularization parameter*. The last is a parameter for modeling the costs that means if the costs shall be low, α is set to a very high value. Then, the term

$$\frac{\alpha}{2} \|u\|_{L_2(\Omega)}^2$$

dominates the functional $J(y, u)$, which ensures that $\|u\|_{L_2(\Omega)}$ stays small. For no bounds on the cost, the regularization parameter is set to zero, which catches an explosion of $\|u\|_{L_2(\Omega)}$ in the minimizing functional $J(y, u)$.

In order to solve such problems usually a so-called *adjoint state* (short: adjoint) is introduced,

see e.g. [158]. Therewith, it is possible to rewrite the problem as an equation system and seek the solution (y^*, u^*, q^*) to the primal equation

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) && \text{for } x \in \Omega, \\ y(x) &= 0 && \text{for } x \in \Gamma = \partial\Omega, \end{aligned} \quad (0.1)$$

the adjoint equation

$$\begin{aligned} -\Delta q(x) + q(x) &= y(x) - y_d(x) && \text{for } x \in \Omega, \\ q(x) &= 0 && \text{for } x \in \Gamma = \partial\Omega, \end{aligned} \quad (0.2)$$

and the projection formula

$$u(x) = \max \left\{ u_a(x), \min \left\{ -\frac{1}{\alpha} q(x), u_b(x) \right\} \right\} \quad \text{for } x \in \Omega, \quad (0.3)$$

see e.g. [158] for more information. These infinite dimensional optimization problems cannot be solved by hand in general but with the power of computers. Thereby two main points have to be taken into account: choosing a suitable efficient discretization scheme for the partial differential equations and an appropriate optimization process.

0.2 Discretization - finite element methods

First, the choice of proper discretization schemes is considered.

In order to solve the problem numerically, the infinite dimensional problem is substituted by a finite dimensional approximation. Therewith the approximation properties of the finite dimensional space is crucial. Suitable discretizations schemes are for example finite differences, finite volumes or finite element methods. In this thesis the last one is considered.

Finite element methods (fem) go back to Courant [52]. The method introduced therein is based on papers by Ritz [128] and Galerkin [68], see also [69]. Since the finite element method only uses the variational formulation of the differential equation, it can be applied to general problems. Possible applications are in different engineering disciplines as for example in electrical, mechanical and civil engineering, see [9, 46, 87, 93, 99, 121, 123, 153].

In finite element methods there are two basic types of refinement: h -refinement and p -refinement. In h -fem an element is subdivided in order to get a better approximation, in p -fem the polynomial degree on an element is increased. Moreover, a combination of these two refinements, the so-called hp -fem, is possible. This is especially useful if parts with high and low regularity appear, since h -fem is superior for low and p -fem for high regularity. In particular hp -fem applied to elliptic boundary value problems leads to an exponential convergence rate for a sufficient refinement for a wider class of problems than p -refinement, whereas h -fem only results in an algebraic convergence rate. For more information see e.g. [36, 43] for h -refinement, [17, 18] for hp -refinement and e.g. [140] for p - and hp -refinement. For special hp -refinements as geometric refinement see [19, 140], for boundary concentrated refinement see [98].

0.3 Optimization

Second, an overview on appropriate optimization processes is given. Suitable methods to solve the problem are (projected) gradient methods, active set strategies [85], interior point

methods, (semismooth) Newton methods [92, 160] or rewriting the problems as saddle point problem ([139, 143]) and apply sufficient solution methods for them.

A further point which matters – and influences especially the mathematical tools to obtain the discretization error – is at which stage the discretization is done. That means, if *First discretize, then optimize* or *First optimize, then discretize* is taken. In this thesis the focus is on the first one. However (see [91] for a comparison) here the focus is not on the conventional approach, where the state y , the adjoint q and the control u are discretized a-priori, but *variational discretization* due to Hinze [88] is used. There, only the state y and the adjoint q are discretized, whereas the control u is discretized implicitly via the projection formula (0.3). The advantage when using variational discretization is, that the error for the control can be estimated by the discretization error of the adjoint. Nevertheless, the drawback is, that it gets necessary to integrate over parts of the element, since the control u is no finite element function in general.

Further approaches – beside variational discretization – and methods to obtain error estimates for optimal control problems can be found in [7, 6, 118, 129].

0.3.1 Semismooth Newton method

In this thesis the focus is on semismooth Newton methods (see e.g. [160] for an introduction) based on the projection formula. In case of distributed optimal control, the semismooth Newton method leads to the (inner) equation system

$$\left(M_{\mathfrak{J}} + \frac{1}{\alpha} M_{\mathfrak{J}} K_N^{-1} M_{\Omega} K_N^{-1} M_{\mathfrak{J}} \right) \vec{u} = \frac{1}{\alpha} M_{\mathfrak{J}} K_N^{-1} M_{\Omega} \left(\vec{y}_d - K_N^{-1} \vec{u}_{\mathfrak{A}} \right). \quad (0.4)$$

There, $M_{\mathfrak{J}}$ denotes the mass matrix on the inactive set \mathfrak{J} , that is the set where $u_a \leq u \leq u_b$ holds. The set, where the constraints u_a and u_b come into play, is called active set and denoted by \mathfrak{A} . M_{Ω} denotes the mass matrix and K_N the stiffness matrix.

Such kind of problems were already considered for h -fem, see [92] and in case of boundary control problems, see [32, 165]. Since the inner equation system has to be solved in each (outer) Newton step, it is crucial to apply the mass matrix and the inverse of the stiffness matrix fast.

However, in case of applying hp -fem instead of h -fem, the mass matrix is no longer well-conditioned. A further drawback is, that the results for the stiffness matrix, which can be inverted in quasi-optimal time for dimension $d = 2$ by using special refinements and suitable direct methods, cannot be extended to three dimensions. Moreover, due to the increase in dimension, the number of degrees of freedom increases. Therewith, an (application of) the inverse inside the (inner) equation system shall be avoided in three dimensions in order to get at least quasi-optimal costs for the overall computations.

0.3.2 Saddle point formulation

A possible way to avoid troubles occurring when applying semismooth Newton method to optimal control problems discretized with hp -fem, is to rewrite it in a saddle point formulation. In order to simplify the problem, in this thesis the box constraints are set to $u_a = -\infty$ and $u_b = \infty$ in the case of using the saddle point formulation.

For the theory of saddle point formulations see e.g. [44] and [27, 170] and references therein for solving them efficiently. For solving optimal control problems in a saddle point formulation

see e.g. [103, 138, 139, 143]. There, several preconditioners are applied in order to solve the saddle point formulation fast. However, these papers only use piecewise linear elements for discretization.

In case of hp -discretizations, suitable hp -preconditioners have to be chosen. There, it has to be considered, that not only good preconditioners for the stiffness matrix as in case of h -refinement, but also for the mass matrix are necessary, since the condition number of the mass matrix depends on the polynomial degree. Good preconditioners for hp -fem are a combination of preconditioners for h -fem (see e.g. [40, 41, 42, 74, 167, 169]) and preconditioners for p -fem (see e.g. [15, 125]). hp -fem preconditioners are for example presented in [29, 35, 63, 100, 101, 136]. A quite important class of preconditioners are based on additive Schwarz methods, introduced by Schwarz in 1870 [141] for two overlapping subdomains (see also [45, 79, 80, 127, 147, 148, 149, 155]). This class of preconditioners is chosen in this thesis.

It has to be mentioned, that there are already contributions using active constraints in a saddle point formulation, see e.g. [83]. However, there only the conventional approach in the discretization has been used. That means, all three variables, the state y , the adjoint q and the control u are fully discretized.

0.3.3 hp -fem for optimal control

Even if optimal control problems subject to elliptic partial differential equations are a well investigated topic, usually (adaptive) h -fem is applied, see [5, 7, 26, 49, 48, 53, 54, 55, 110, 117, 118] and there is few literature on the application of hp -fem to these kind of problems. hp -fem discretizations in combination with semismooth Newton methods are used in [31, 32, 163, 165] for different kinds of optimal boundary control problems. Except in [163] a special hp -discretization, the boundary concentrated fem is applied. There, in each refinement step the mesh is h -refinement in all boundary elements but p -refined at all elements in the interior of the domain. This refinement strategy presented in [98], has about the same number of degrees of freedom as the boundary finite elements (see e.g. [151]), that means about h^{1-d} instead of h^{-d} for the mesh size h on the boundary. Furthermore this refinement captures especially in the case of boundary control the parts with less-smoothness quite well.

A refinement merely based on the geometry of the domain and the projection formula is the so-called vertex-concentrated fem introduced in [163], which leads to exponential order of convergence.

In [164] a distributed optimal control problem is considered. However, the problem there is solved with the interior point method. Therewith, it has to be stated that there are very few contributions, where distributed optimal control problems are solved with hp -fem and to the knowledge of the author there are especially no contributions on the solution of distributed optimal control problems solved with semismooth Newton methods or as saddle point formulation in case of an hp -discretization.

The problem in applying the semismooth Newton method is, that in contrast to [165] it is not clear where the interface between the active and inactive set is, which makes the choice of a suitable hp -discretization more challenging. Furthermore, it is necessary to calculate M_J , i.e. to evaluate the integral over parts of an element.

When rewriting the problem as saddle point formulation, it is crucial to choose suitable preconditioners to solve the problem fast. That approach is especially advantageous in three dimensions, since there it is not possible to gain at least quasi-optimal complexity with the

semismooth Newton method. However, especially in the case of three dimensions the complexity is a quite crucial point due to the increase of degrees of freedom.

0.4 Outline of the thesis

This thesis considers the application of hp -finite element methods to optimal control problems subject to elliptic boundary value problems with constraints on the control. In most cases distributed optimal control problems are considered. The structure of this thesis is as follows:

- In chapter 1 some notation and general issues, especially a short summary of well-known theorems to (preconditioned) Krylov subspace methods, are given.
- In chapter 2 basics to optimal control problems, especially the existence and uniqueness of a solution are recalled. Furthermore, two model problems, an optimal boundary control problem and a distributed optimal control problem, are presented.
- Since hp -fem is applied to optimal control problems, in chapter 3, an introduction in finite element methods, especially the hp -fem, is given. Since developing an hp -finite element code was one of the main parts of this thesis, some basics for implementing hp element methods are pointed out. The focus is especially on the used basis functions and the handling of hanging nodes which appear in adaptive refinement in case of using square elements. As the application of fem usually leads to a large equation system, which shall be solved iteratively, suitable hp -preconditioners to decrease the number of iterations are recalled. Furthermore, numerical results in order to understand the behaviour of the given preconditioners are presented.
- In chapter 4, the optimal boundary control problems are discretized by variational discretization. Moreover, the existence and uniqueness of the discrete solution is investigated. For both model problems, suitable refinement strategies and if known – error estimates – are given. Both model problems are then solved with the semismooth Newton method. For the optimal boundary control problem, results in two dimensions are given in order to test the proposed refinement strategy. Furthermore, results in three dimensions with a bc-refinement, also applied in [31] for two-dimensional problems, are presented. For the distributed optimal control problem, two refinement strategies – one which only uses a-priori information and a further one which combines a-priori information with error estimators – are applied and several numerical examples are presented in subsection 4.4.4
- Chapter 5 concentrates on solving distributed optimal control problems in a saddle point formulation. There, for simplicity inactive constraints are assumed. In order to solve the saddle point formulation efficiently, especially in the case of hp -fem it is crucial to use suitable preconditioners. Moreover, the chosen Krylov subspace method plays an important role. Therewith, for different Krylov subspace methods suitable preconditioners are applied. There, the focus is on a modified preconditioned conjugate gradient method by Schöberl and Zulehner, see [139]. Also a preconditioned MINRES method and a preconditioned GMRES method is considered. Numerical results to confirm theoretical results are investigated in section 5.5 for all three Krylov subspace methods.

The obtained results show a significant reduction of the degrees of freedom in case of applying suitable hp -finite element methods instead of uniform h -fem. That is especially important, since a suitable discretizing strategy which keeps the number of degrees of freedom low, allows to solve larger problems, as storage and time in computers are a limited resource. Another highlight of this thesis are the obtained numerical results for an optimal boundary control problem solved in three dimensions with a hp -finite element discretization.

Further important results are the extension of the preconditioned conjugate gradient method by [139] to hp -finite elements. The results calculated with that method are especially independent of the discretization parameters h , p and the regularization parameter α in case of applying suitable hp -preconditioners. For bc-refinement an application of these preconditioners (in case of no hanging nodes) can even be performed in optimal complexity.

In case of applying the proposed preconditioned MINRES, it is shown that the results are independent of the discretization parameters h and p . Moreover, in case of bc-refinement (without hanging nodes) the application of the preconditioner can be performed in optimal complexity.

Although only two-dimensional results are given for the saddle point formulation, the theoretical estimates are not limited to three dimensions and especially not to quadrilateral elements. Therewith, these results provide an excellent background for further research and applications.

1 Preliminaries

First, some basic results are given and the corresponding notation is introduced. The starting point is a clarification of matrix notation issues and a short summary of Krylov subspace methods. Moreover, the so-called integrated Legendre polynomials, which will be used in subsection 3.2.1 to define the basis functions, are introduced. Then, some important issues on functional analysis and Sobolev spaces are recalled.

1.1 Matrices

Here, only some well-known basics on matrices are pointed out to clarify the notation. For more information see e.g. [72, 141].

Let $N, M \in \mathbb{N}$ and $A, \tilde{A} \in \mathbb{R}^{N \times N}$ be two symmetric matrices. The relation $A < \tilde{A}$ ($\tilde{A} > A$) denotes that $\tilde{A} - A$ is positive definite and $A \leq \tilde{A}$ ($\tilde{A} \geq A$) denotes that $\tilde{A} - A$ is positive semidefinite. If there are constants $c_1, c_2 > 0$, such that $c_1 \tilde{A} - A > 0$ and $c_2 \tilde{A} - A < 0$, the matrices \tilde{A} and A are called spectrally equivalent.

Let $B \in \mathbb{R}^{M \times N}$, $C \in \mathbb{R}^{N \times M}$, $D \in \mathbb{R}^{M \times M}$ be matrices and let A and the Schur complement $S = D - CA^{-1}B$ be nonsingular. Then, the matrix \mathcal{A} given by

$$\mathcal{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

is invertible and its inverse is (see e.g. [72, 141])

$$\mathcal{A}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BS^{-1}CA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{pmatrix}. \quad (1.1)$$

Let $\vec{v}, \vec{w} \in \mathbb{R}^N$ be two vectors and let $\mathbf{v}_i, \mathbf{w}_i$ be the i -th component of \vec{v}, \vec{w} , respectively. The Euclidean inner product of \vec{v}, \vec{w} is denoted by

$$\langle \vec{v}, \vec{w} \rangle = \sum_{i=1}^N \mathbf{v}_i \mathbf{w}_i$$

and its induced norm is

$$\|\vec{v}\|_2 = \sqrt{\langle \vec{v}, \vec{v} \rangle}.$$

The scalar product induced by a symmetric positive definite matrix A is written as

$$\langle A \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_A$$

and its corresponding norm is denoted by $\|\cdot\|_A = \sqrt{\langle A \cdot, \cdot \rangle_A}$. The Rayleigh coefficient for a quadratic real and symmetric matrix E is

$$\lambda_{\min}(E) \vec{x}^\top \vec{x} \leq \vec{x}^\top E \vec{x} \leq \lambda_{\max}(E) \vec{x}^\top \vec{x}, \quad (1.2)$$

where λ_{\min} and λ_{\max} denote the minimal and maximal eigenvalue of E , see e.g. [82]. Moreover, the definition for the Kronecker product is given.

Definition 1.1. Let $A \in \mathbb{R}^{N_A \times N_C}$, $B \in \mathbb{R}^{N_B \times N_D}$. The product

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1N_C}B \\ \vdots & \ddots & \vdots \\ a_{N_A1}B & \cdots & a_{N_A N_C}B \end{pmatrix} \in \mathbb{R}^{N_A N_B \times N_C N_D}$$

is called **Kronecker product**.

1.2 Krylov subspace methods

Krylov subspace methods are iterative methods based on a general projection process and are a famous tool to solve (large) equation systems, i.e. to solve

$$\mathcal{A}\vec{z} = \vec{g} \tag{1.3}$$

for $\vec{z} \in \mathbb{R}^N$, where $\mathcal{A} \in \mathbb{R}^{N \times N}$ and $\vec{g} \in \mathbb{R}^N$ are given. Let \vec{z}_* denote the exact solution to (1.3). For an introduction into Krylov subspace methods, see e.g. [72, 73, 95, 112, 131]. In here, three methods which yield (in infinite arithmetic) the exact solution after N steps, are considered.

The first one, is the conjugate gradient method (CG-method) developed in 1952 by Hestens and Stiefel, see [84]. The CG method is applicable to equation systems with symmetric and positive definite system matrix \mathcal{A} with respect to the scalar product used in CG. Usually, the standard scalar product $\langle \cdot, \cdot \rangle$ is taken.

The second one, is the MINRES, developed in 1975 by Paige and Saunders, see [124]. It can be applied to symmetric, but indefinite matrices \mathcal{A} and is – as the CG – based on the Lanczos method, see [104], a method to construct a basis for the Krylov subspace with a three-term recurrence.

As last method, the GMRES, developed by Saad and Schultz [132] in 1986, is recalled. Although the GMRES only needs a regular system matrix, it is usually not the first choice, since it needs a full term recurrence, the Arnoldi method (see e.g. [95]) for constructing the Krylov subspace basis. For a comparison of the cost of the three considered Krylov subspace methods, see table 1.1.

Usually a preconditioned system is solved, since a suitable preconditioner can decrease the number of iterations substantially, which is especially important when solving large equation systems. Due to the assumptions on the system matrix for different methods, in the case of

method	work	storage
CG	$\mathcal{O}(k(N))$	$\mathcal{O}(N)$
MINRES	$\mathcal{O}(kN)$	$\mathcal{O}(N)$
GMRES	$k^2 \frac{N}{2} \mathcal{O}(kN)$	$\mathcal{O}(N)$

Table 1.1: Costs for different subspace methods for the k -th iterate \vec{z}_k (see [82, 95]).

the CG or the MINRES a symmetric and positive definite preconditioner has to be chosen. Then, the symmetric and positive definite preconditioner, denoted by \mathcal{P} , can be decomposed in its Cholesky decomposition $\mathcal{P} = CC^\top$ and instead of the preconditioned system

$$\mathcal{P}^{-1}\mathcal{A}\vec{z} = \mathcal{P}^{-1}\vec{g}. \tag{1.4}$$

the equivalent system

$$C^{-1}\mathcal{A}C^{-\top}(C^{\top}\vec{z}) = C^{-1}\vec{g}, \quad (1.5)$$

is considered. For more information on different choices of preconditioners see e.g. [28, 71, 72, 73, 130, 131]. General known results on convergence, a pseudo-code of the algorithm and termination conditions are given in the corresponding subsections for each method.

1.2.1 PCG

The preconditioned conjugate gradient method is summarized in algorithm 1. Its convergence

Algorithm 1: preconditioned conjugate gradient method (PCG), see e.g. [95]

input : $\mathcal{A}, \mathcal{P}^{-1}, \vec{g}, \varepsilon$

output: solution \vec{u}_k

choose a suitable start vector \vec{u}_0

$$\vec{r}_0 = \mathcal{A}\vec{u}_0 - \vec{g}$$

$$\vec{w}_0 = \mathcal{P}^{-1}\vec{r}_0$$

$$\vec{q}_0 = \vec{w}_0$$

$$\rho_0 = \langle \vec{w}_0, \vec{r}_0 \rangle$$

for $k = 0, \dots, N$ **do**

$$\alpha_k = -\frac{\langle \vec{w}_k, \vec{r}_k \rangle}{\langle \mathcal{A}\vec{q}_k, \vec{q}_k \rangle}$$

$$\vec{u}_{k+1} = \vec{u}_k + \alpha_k \vec{q}_k$$

$$\vec{r}_{k+1} = \vec{r}_k - \alpha_k \mathcal{A}\vec{q}_k$$

$$\vec{w}_{k+1} = \mathcal{P}^{-1}\vec{r}_{k+1}$$

$$\beta_k = \frac{\langle \vec{w}_{k+1}, \vec{r}_{k+1} \rangle}{\langle \vec{w}_k, \vec{r}_k \rangle}$$

$$\vec{q}_{k+1} = \vec{w}_{k+1} + \beta_k \vec{q}_k$$

$$\rho_k = \langle \vec{w}_k, \vec{r}_k \rangle$$

if $\rho_k < \varepsilon^2 \rho_0$ **then**

└ stop

can be estimated by using the condition number

$$\kappa(\mathcal{A}) := \frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})}$$

with the following theorem:

Theorem 1.2. (see e.g. [95, 151]) Let $\mathcal{A} \in \mathbb{R}^{N \times N}$ be a symmetric and positive definite matrix, $\mathcal{P} \in \mathbb{R}^{N \times N}$ a suitable symmetric and positive definite preconditioner for \mathcal{A} and let \vec{z}_* denote the exact solution of (1.3). Then, the preconditioned conjugate gradient method converges for an arbitrary start value \vec{z}_0 to the exact solution \vec{z}_* and it holds

$$\|\vec{z}_k - \vec{z}_*\|_{\mathcal{A}} \leq \frac{2\rho^k}{1 + \rho^{2k}} \|\vec{z}_0 - \vec{z}_*\|_{\mathcal{A}} \quad \text{with } \rho = \frac{\sqrt{\kappa(\mathcal{P}^{-1}\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{P}^{-1}\mathcal{A})} + 1}.$$

A special variant of the PCG (see [139]) with different scalar product is considered in chapter 5.

1.2.2 MINRES

As in the PCG, for the MINRES (see algorithm 2) a full convergence theory is known.

Algorithm 2: preconditioned MINRES, see e.g. [73, 95]

input : $\mathcal{A}, \mathcal{P}_{\text{minres}}^{-1}, \vec{x}_0, \vec{g}, \varepsilon$
output: \vec{x}
 $\vec{v}_{old} = 0, \vec{v} = \vec{g} - \mathcal{A}\vec{x}, \vec{z} = \mathcal{P}_{\text{minres}}^{-1}\vec{v}, \vec{w}_{old} = 0, \vec{w} = 0$
 $\beta_{old} = 1, \beta = \sqrt{\langle v, z \rangle}, c_{old} = 1, c = 1, s_{old} = 0, s = 0, \eta = \beta, \eta_0 = \eta$
for $k = 1, 2, \dots$ **do**
 $\vec{v}_{normed} = \frac{1}{\beta}\vec{z}$
 $\vec{q} = \mathcal{A}\vec{v}_{normed}$
 $\alpha = \langle \vec{q}, \vec{v}_{normed} \rangle$
 $\vec{v}_{new} = \vec{q} - \frac{\beta}{\beta_{old}}\vec{v}_{old} - \frac{\alpha}{\beta}\vec{v}$
 $\vec{v}_{old} = \vec{v}, \vec{v} = \vec{v}_{new}$
 $\vec{z} = \mathcal{P}_{\text{minres}}^{-1}\vec{v}$
 $\beta_{new} = \sqrt{\langle \vec{v}, \vec{z} \rangle}$
 $\tilde{\rho}^{(1)} = c\alpha - s c_{old}\beta$
 $\rho^{(1)} = \sqrt{(\tilde{\rho}^{(1)})^2 + \beta_{new}^2}$
 $\rho^{(2)} = s\alpha + c c_{old}\beta$
 $\rho^{(3)} = s_{old}\beta$
 $c_{new} = \frac{\tilde{\rho}^{(1)}}{\rho^{(1)}}, c_{old} = c, c = c_{new}$
 $s_{new} = \frac{\beta_{new}}{\rho^{(1)}}, s_{old} = s, s = s_{new}$
 $w^{new} = \frac{1}{\rho^{(1)}}(\vec{v}_{normed} - \rho^{(3)}\vec{w}_{old} - \rho^{(2)}\vec{w})$
 $\vec{x} = \vec{x} + c_{new}\eta w^{new}$
 $\vec{w}_{old} = \vec{w}, \vec{w} = \vec{w}^{new}, \beta_{old} = \beta, \beta = \beta_{new}$
 $\eta = -s_{new}\eta$
 if $\left| \frac{\eta}{\eta_0} \right| \leq \varepsilon$ **then**
 └ stop

The residuum for the preconditioned MINRES can be estimated by

$$\frac{\|\vec{r}_k\|_{\mathcal{P}^{-1}}}{\|\vec{r}_0\|_{\mathcal{P}^{-1}}} \leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_{\lambda} |p_k(\lambda)|, \quad (1.6)$$

where λ is an eigenvalue of $\mathcal{P}^{-1}\mathcal{A}$ (see e.g. [64]) and Π_k denotes the space of polynomials with maximal polynomial degree k . If the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are known and its negative eigenvalues lie in $[-a, -b]$ and its positive eigenvalues in $[c, d]$ for $a, b, c, d > 0$, then the following theorem holds:

Theorem 1.3. (see e.g. [64]) *After $2k$ steps of the minimum residual method, the iteration residual $\vec{r}_{2k} = \vec{g} - \vec{A}\vec{z}_0$ satisfies the bound*

$$\|\vec{r}_{2k}\|_{\mathcal{P}^{-1}} \leq \left(2 \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right)^k \|\vec{r}_0\|_{\mathcal{P}^{-1}}.$$

Remark 1.4. It has to be mentioned, that “stair-casing”, i.e.

$$\|\vec{r}_{2k+1}\|_{\mathcal{P}^{-1}} = \|\vec{r}_{2k}\|_{\mathcal{P}^{-1}}$$

cannot be avoided in case of theorem 1.3.

1.2.3 GMRES

Although for the application of the GMRES the matrix needs only to be regular (see algorithm 3), compared with the PCG and the MINRES, there are several disadvantages (see also table 1.1).

Algorithm 3: preconditioned GMRES with nonsingular preconditioner $\mathcal{P}_{\text{gmres}}^{-1}$, see e.g. [95]

input : $\mathcal{A}, \mathcal{P}_{\text{gmres}}^{-1}, \vec{z}_0, \vec{g}, \varepsilon$
output: \vec{z}_k
 $\vec{q}_0 = \mathcal{P}_{\text{gmres}}^{-1}(\vec{g} - \mathcal{A}\vec{z}_0)$
 $\beta = \|\vec{q}_0\|, \vec{v}_1 = \vec{q}_0/\beta, t_1 = \beta$
for $k = 1, 2, \dots$ **do**
 $\vec{w}_k = \mathcal{P}_{\text{gmres}}^{-1}\mathcal{A}\vec{v}_k$
 for $i = 1, \dots, k$ **do**
 $h_{ik} = \vec{v}_i^\top \vec{w}_k$
 $\vec{w}_k = \vec{w}_k - h_{ik}\vec{v}_i$
 $h_{k+1,k} = \|\vec{w}_k\|$
 $\vec{v}_{k+1} = \vec{w}_k/h_{k+1,k}$
 for $i = 1, \dots, k-1$ **do**
 $\begin{pmatrix} h_{ik} \\ h_{i+1,k} \end{pmatrix} := \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix} \begin{pmatrix} h_{ik} \\ h_{i+1,k} \end{pmatrix}$
 $\tau = |h_{kk}| + |h_{k+1,k}|$
 $\nu = \tau \cdot \sqrt{(h_{kk}/\tau)^2 + (h_{k+1,k}/\tau)^2}$
 $c_k = h_{kk}/\nu, s_k = h_{k+1,k}/\nu$
 $h_{kk} = \nu, h_{k+1,k} = 0, t_{k+1} = -s_k t_k, t_k = c_k t_k$
 if $|z_{k+1}|/\beta \leq \varepsilon$ **then**
 | stop
 $\vec{y}_k = t_k/h_{kk}$
 for $i = k-1, \dots, 1$ **do**
 $y_i = (t_i - \sum_{j=i+1}^k h_{ij}y_j)/h_{ii}$
 $\vec{z}_k = \vec{z}_0 + \sum_{i=1}^k y_i \vec{v}_i$

The most important one is, that there is no convergence analysis based solely on the matrix \mathcal{A} if it is non normal, that is $\mathcal{A}^* \mathcal{A} \neq \mathcal{A} \mathcal{A}^*$.

Theorem 1.5. (see e.g. [131, Proposition 6.32]) Assume that \mathcal{A} is diagonalizable and let $\mathcal{A} = \mathcal{X} \mathcal{D} \mathcal{X}^{-1}$ where $\mathcal{D} = (\lambda_1, \lambda_2, \dots, \lambda_N)$ is the diagonal matrix of eigenvalues. Then, the relative residual achieved in the k -th step can be estimated by

$$\frac{\|\vec{r}_k\|_2}{\|\vec{r}_0\|_2} \leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_{i=1, \dots, k} |p_k(\lambda_i)| \|\mathcal{X}\|_2 \|\mathcal{X}^{-1}\|_2.$$

Remark 1.6. *If \mathcal{A} is a non normal matrix, the norms $\|\mathcal{X}\|_2\|\mathcal{X}^{-1}\|_2$ can get arbitrarily large (see e.g. [27] and references therein for further information).*

Remark 1.7. *In the case of a normal matrix \mathcal{A} it holds $\|\mathcal{X}\|_2\|\mathcal{X}^{-1}\|_2 = 1$.*

For results on estimating the min max problem if \mathcal{A} is a non symmetric and indefinite matrix see [65, 106, 107, 115, 131, 154]. If the matrix \mathcal{A} is normal, symmetric and positive definite, a solution to the min max problem is known, see e.g. [112].

Remark 1.8. *To decrease the costs in storage – which can be a difficult problem when using large equation systems – there are modifications of the GMRES which need less storage, for example by restarting the method. However, then the property of finding the exact solution in N steps gets lost, see e.g. [73, 95, 131, 144].*

In GMRES both choices, applying the preconditioner to the right or the left side, are possible. Here, preconditioning on the left side is used, i.e.

$$\mathcal{P}_{\text{gmres}}^{-1}\mathcal{A}\vec{z} = \mathcal{P}_{\text{gmres}}^{-1}\vec{g}. \quad (1.7)$$

For differences in left and right preconditioning see e.g. [95, 131].

1.3 Perturbations in equation systems

In many cases the matrix \mathcal{A} and the right-hand side \vec{g} are perturbed by an error. Then, in fact the equations system

$$\overline{\mathcal{A}}\vec{z} = \vec{g} \quad (1.8)$$

instead of (1.3) is solved. Therefore it is important to handle the perturbations in order to make sure that perturbed solution \vec{z}_* is close to the exact solution \vec{z}_* .

The following theorem gives an estimate on that topic.

Theorem 1.9. *([134]) Let there be a vector norm and a corresponding multiplicative matrix norm, which fulfill the following:*

(a) *If $\mathcal{A} \in \mathbb{R}^{N \times N}$ is invertible and the perturbation $\Delta\mathcal{A} \in \mathbb{R}^{N \times N}$ fulfills*

$$\|\Delta\mathcal{A}\| < \frac{1}{\|\mathcal{A}^{-1}\|}.$$

Then, also $\overline{\mathcal{A}} = \mathcal{A} + \Delta\mathcal{A}$ is invertible.

(b) *Let additionally $\vec{g} \in \mathbb{R}^N$, $\vec{g} \neq 0$, $\Delta\vec{g} \in \mathbb{R}^N$ and $\vec{g} = \vec{g} + \Delta\vec{g}$ hold and $\vec{z}_* \in \mathbb{R}^N$ is the solution to (1.3). Let \vec{z}_* be the solution to the perturbed equation system (1.8). Then it holds*

$$\varepsilon_x \leq \frac{\kappa(\mathcal{A})}{1 - \kappa(\mathcal{A})\varepsilon_{\mathcal{A}}}(\varepsilon_{\mathcal{A}} + \varepsilon_g)$$

with $\varepsilon_{\mathcal{A}} = \frac{\|\Delta\mathcal{A}\|}{\|\mathcal{A}\|}$ and $\varepsilon_g = \frac{\|\Delta\vec{g}\|}{\|\vec{g}\|}$.

1.4 Integrated Legendre polynomials

The integrated Legendre polynomials are defined via the Legendre polynomials, see e.g. [34, 101, 156] and references therein. The i -th Legendre polynomial is given by

$$L_i(x) = \frac{1}{2^i i!} \frac{d^i}{dx^i} (x^2 - 1)^i \quad \text{for } i \geq 2.$$

The first two integrated Legendre polynomials are set to

$$\hat{L}_0(x) = \frac{1-x}{2} \quad \text{and} \quad \hat{L}_1(x) = \frac{1+x}{2}.$$

With

$$\gamma_i := \sqrt{\frac{(2i-3)(2i-1)(2i+1)}{4}}$$

the i -th integrated Legendre polynomial is defined by

$$\hat{L}_i(x) = \gamma_i \int_{-1}^x L_{i-1}(s) ds \quad \text{for } i \geq 2.$$

Remark 1.10. *The scaling factor γ_i ensures, that the norm of the integrated Legendre polynomials \hat{L}_i is one.*

Remark 1.11. *In subsection 3.2.5 the unscaled integrated Legendre polynomials, denoted by $\tilde{L}_i(x)$ and given by*

$$\begin{aligned} \tilde{L}_0(x) &= \hat{L}_0(x), \\ \tilde{L}_1(x) &= \hat{L}_1(x), \\ \tilde{L}_i(x) &= \frac{1}{\gamma_i} \hat{L}_i(x), \end{aligned}$$

are applied. Since most of the time scaled integrated Legendre polynomials are used, the notation of integrated Legendre polynomials refers always to the scaled ones.

The following lemma recaps some properties of the Legendre and integrated Legendre polynomials.

Lemma 1.12. *(see e.g. [156]) For the Legendre polynomials hold the recurrence relation*

$$(i+1)L_{i+1}(x) = (2i+1)xL_i(x) - iL_{i-1}(x) \quad \text{for } i \geq 1, \quad (1.9)$$

and the orthogonality relation

$$\int_{-1}^1 L_i(x)L_j(x) dx = \delta_{ij} \frac{2}{2i+1} \quad \text{for } i \geq 0.$$

Furthermore, it is

$$(2i+1)L_i(x) = \frac{d}{dx} (L_{i+1}(x) - L_{i-1}(x)).$$

Between the Legendre polynomials and the integrated Legendre polynomials hold the relations

$$\frac{d}{dx} \hat{L}_i(x) = \gamma_i L_{i-1}(x) \quad \text{for } i \geq 2, \quad (1.10)$$

$$\hat{L}_i(x) = \sqrt{\frac{(2i+1)(2i-3)}{4(2i-1)}} (L_i(x) - L_{i-2}(x)) \quad \text{for } i \geq 2. \quad (1.11)$$

Moreover, the integrated Legendre polynomials fulfill

$$\hat{L}_i(1) = 0 \quad \text{for } i \geq 2, \quad (1.12)$$

$$\hat{L}_i(-1) = 0 \quad \text{for } i \geq 2. \quad (1.13)$$

Especially the relations (1.9), (1.10) and (1.11) are useful for a fast point evaluation of the integrated Legendre polynomials.

1.5 Functional analysis

This section recalls some well-known parts of functional analysis. For more information on this topic, see e.g. [4, 151, 158] and the references therein.

A space V is called Banach space with the norm $\|\cdot\|_V$ if every Cauchy sequence converges in V with respect to the norm $\|\cdot\|_V$. Here, mainly Hilbert spaces, i.e. a Banach space whose norm is induced by a scalar product $\|\cdot\|_V = \langle \cdot, \cdot \rangle_V$, are used.

Let V, W denote Banach spaces. An operator $A : V \rightarrow W$ is then called *bounded* if

$$\exists c_A \geq 0 : \|Av\|_W \leq c_A \|v\|_V.$$

Definition 1.13. Let $A : V \rightarrow W$ is a linear and bounded operator. Then

$$\|A\| := \sup_{\|v\|_V=1} \|Av\|_W$$

is finite and called norm of A . Possible notations are $\|A\|$ or $\|A\|_{\mathcal{L}(V,W)}$ if it is necessary to clarify the spaces V, W .

The set of all linear and bounded operators is denoted by $\mathcal{L}(V, W)$ and is a Banach space if W is complete. The *dual space* of V is defined by

$$V^* = \mathcal{L}(V, \mathbb{R}).$$

Due to the completeness of \mathbb{R} , the dual space V^* is a Banach space. Moreover, it is reflexive, if and only if it holds $(V^*)^* = V$.

An important theorem which shall also be given is:

Theorem 1.14 (Riesz mapping theorem, see e.g. [158]). Let $\{V, \langle \cdot, \cdot \rangle_V\}$ be a Hilbert space. Then, for an arbitrary bounded and continuous functional $F \in V^*$, the element $f \in V$ with $\|F\|_{V^*} = \|f\|_V$, which can be denoted as

$$F(v) = \langle f, v \rangle_V$$

is uniquely defined.

Let V, W be Hilbert spaces and the operator $A : V \rightarrow W$ be linear and bounded. Its adjoint operator A^* is given by

$$A^* : W^* \rightarrow V^*$$

and it holds

$$\langle Av, w \rangle_W = \langle v, A^*w \rangle_V.$$

At the end of this section, a quite important theorem to investigate the unique solvability of boundary value problems is given.

Theorem 1.15. (*Lemma of Lax-Milgram, see e.g. [4]*) *Let V denote a Hilbert space. Furthermore, let the linear operator $A : V \rightarrow V^*$ be bounded and V -elliptic, i.e.*

$$\exists c_A > 0 : \quad \langle Av, v \rangle \geq c_A \|v\|_V^2 \quad \text{for all } v \in V.$$

Then, the operator equation

$$Av = f,$$

possesses for each $f \in V^$ a unique solution, and it holds*

$$\|v\|_V \leq \frac{1}{c_A} \|f\|_{V^*}.$$

1.6 L_p spaces

In this section some basic definitions of the well-known L_p spaces are recalled. Further information can e.g. be found in [1, 151, 158].

First, the notation domain is clarified. A domain Ω , is an open, bounded and connected subset of \mathbb{R}^d , where d denotes the dimension of the space. Furthermore, Ω is assumed to be a Lipschitz domain. Its boundary is denoted by $\Gamma = \partial\Omega$.

Next, the Banach spaces $L_p(\Omega)$ are introduced (see e.g. [1]). The space $L_p(\Omega)$ for $p \in \mathbb{N}$ with $1 \leq p < \infty$ denotes the space of measurable functions with finite norm

$$\|v\|_{L_p(\Omega)} := \left(\int_{\Omega} |v(x)|^p dx \right)^{1/p}.$$

In fact $L_p(\Omega)$ denotes the space of equivalence classes of all defined measurable functions in Ω whose p power is integrable. For $p = \infty$, it holds that $L_{\infty}(\Omega)$ is the space of all (equivalence classes of) almost everywhere uniformly bounded and measurable functions v associated with the norm

$$\|v\|_{L_{\infty}(\Omega)} = \operatorname{ess\,sup}_{x \in \Omega} |v(x)| := \inf_{|L|=0} \left(\sup_{x \in \Omega \setminus L} |v(x)| \right),$$

where $\operatorname{ess\,sup}$ is the essential or real maximum or supremum of a function.

Remark 1.16. The space $L_2(\Omega)$ equipped with the scalar product in $L_2(\Omega)$

$$\langle v, w \rangle_{L_2(\Omega)} := \int_{\Omega} v(x)w(x) \, dx$$

and the induced norm

$$\langle v, v \rangle_{L_2(\Omega)} = \|v\|_{L_2(\Omega)}^2 \quad \forall v \in L_2(\Omega),$$

is a Hilbert space.

For $1 \leq p < \infty$ and q with

$$\frac{1}{p} + \frac{1}{q} = 1,$$

the space $L_q(\Omega)$ is the dual space to $L_p(\Omega)$ with the norm

$$\|v\|_{L_q(\Omega)} := \sup_{0 \neq w \in L_p(\Omega)} \frac{|\langle w, v \rangle_{\Omega}|}{\|w\|_{L_p(\Omega)}} \quad \text{for } 1 \leq p < \infty$$

and the duality pairing

$$\langle w, v \rangle_{\Omega} = \int_{\Omega} w(x)v(x) \, dx.$$

Furthermore, the space $L_p(\Gamma)$ for the boundary $\Gamma = \partial\Omega$ is given. It is the space of all functions in \mathbb{R} with finite norm

$$\|v\|_{L_p(\Gamma)} := \left(\int_{\Gamma} |v(x)|^p \, ds_x \right)^{1/p}.$$

The corresponding scalar product for $L_2(\Gamma)$ is denoted by $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$. The set

$$\text{supp } v := \overline{\{x \in \Omega : v(x) \neq 0\}}$$

is called support of v . It is the smallest closed set, on which outside v vanish identically.

1.7 Sobolev spaces

Before introducing Sobolev spaces, the definition of the partial derivation of a function $v(x_1, \dots, x_d)$ for $d \in \mathbb{N}$ is given.

Definition 1.17. A vector $\varrho = (\varrho_1, \varrho_2, \dots, \varrho_d)$, $\varrho_i \in \mathbb{N}_0$ with the absolute value $|\varrho| = \varrho_1 + \varrho_2 + \dots + \varrho_d$ is called multiindex. For a suitably differentiable real function $v(x)$, the weak partial derivative of order ϱ is denoted by

$$D^{\varrho}v(x) := \left(\frac{\partial}{\partial x_1} \right)^{\varrho_1} \left(\frac{\partial}{\partial x_2} \right)^{\varrho_2} \cdots \left(\frac{\partial}{\partial x_d} \right)^{\varrho_d} v(x_1, x_2, \dots, x_d).$$

Here, some basics of Sobolev spaces are recalled, for details see e.g. [1].

Definition 1.18. Let $1 \leq p < \infty$, $p, k \in \mathbb{N}$. The space $W_p^k(\Omega)$ is defined as linear space of all $v \in L_p(\Omega)$, whose weak derivate $D^\varrho v$ with $|\varrho| \leq k$ exists and is contained in $L_p(\Omega)$ with the finite norm

$$\|v\|_{W_p^k(\Omega)} = \left(\sum_{|\varrho| \leq k} \int_{\Omega} |D^\varrho v(x)|^p dx \right)^{1/p}.$$

The definition of $W_\infty^k(\Omega)$ for $p = \infty$ is introduced analogously with the norm

$$\|v\|_{W_\infty^k(\Omega)} = \max_{|\varrho| \leq k} \|D^\varrho v\|_{L_\infty(\Omega)}.$$

The spaces $W_p^k(\Omega)$ are Banach spaces and are called Sobolev spaces. A special case is $k = 2$, where

$$H^k(\Omega) := W_2^k(\Omega).$$

The space $H^1(\Omega)$ is very important in here, therefore its norm definition is recalled

$$\|v\|_{H^1(\Omega)} = \left(\int_{\Omega} (v^2 + |\nabla v|^2) dx \right)^{1/2}.$$

By introducing the scalar product

$$\langle w, v \rangle_{H^1(\Omega)} = \int_{\Omega} wv dx + \int_{\Omega} \nabla w \cdot \nabla v dx,$$

the space $H^1(\Omega)$ becomes a Hilbert space. Furthermore, a specification of this space is pointed out, assumed that $\Gamma = \overline{\Gamma_{\mathcal{D}}} \cup \Gamma_{\mathcal{N}}$ and $\Gamma_{\mathcal{D}} \cap \Gamma_{\mathcal{N}} = \emptyset$ for the Dirichlet boundary $\Gamma_{\mathcal{D}}$ and the Neumann boundary $\Gamma_{\mathcal{N}}$. Then, the space $H_{\Gamma_{\mathcal{D}}}^1(\Omega)$ is defined by

$$H_{\Gamma_{\mathcal{D}}}^1(\Omega) := \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_{\mathcal{D}}\}.$$

The Sobolev-Slobodeckij space $W_p^s(\Omega)$ with real values $s = k + \epsilon > 0$, for $k \in \mathbb{N}$, $\epsilon \in (0, 1)$ is the linear space $v \in L_p(\Omega)$ of functions with finite norm

$$\|v\|_{W_p^s(\Omega)}^p := \|v\|_{W_p^k(\Omega)}^p + \int_{\Omega} \int_{\Omega} \sum_{|\varrho|=k} \frac{|D^\varrho v(x) - D^\varrho v(y)|^p}{|x - y|^{2+\epsilon p}} dx dy.$$

Furthermore, the space $H^s(\Gamma)$, $s \in (0, 1)$ with the norm

$$\|v\|_{H^s(\Gamma)} := \left(\|v\|_{L_2(\Gamma)}^2 + \int_{\Gamma} \int_{\Gamma} \frac{(v(x) - v(y))^2}{|x - y|^{d-1+2s}} ds_x ds_y \right)^{1/2}$$

is given.

Finally, the countably normed spaces \tilde{B}_β^2 , see [98], akin to [17, 18], are recalled. For $\beta \in [0, 1)$ they are based on the space H_β^2 , which is the completion of $C^\infty(\overline{\Omega})$ with finite norm

$$\|v\|_{H_\beta^2(\Omega)}^2 := \|v\|_{H^1(\Omega)}^2 + \|r^\beta \nabla^2 v\|_{L_2(\Omega)}^2$$

where $r(x) := \text{dist}(x, \partial\Omega)$ denotes the distance to the boundary. For $c_B, \gamma > 0, \beta \in [0, 1)$ the space \tilde{B}_β^2 is then defined by

$$\tilde{B}_\beta^2(c_B, \gamma) = \{v \in H_\beta^2(\Omega) \mid \|v\|_{H_\beta^2(\Omega)} \leq c_B, \|r^{\beta+p} \nabla^{p+2} v\|_{L_2(\Omega)} \leq c_B \gamma^p p! \quad \forall p \in \mathbb{N}\}.$$

Countably normed spaces allow the controlling of derivatives near the boundary of the domain. According to [18, Lemma 2.4], functions in these spaces are analytic away from the zeros of the weight function.

2 Optimal control problems for pdes

In this chapter linear quadratic optimal control problems subject to an elliptic partial differential equation are considered, i.e. problems of the type

$$\begin{aligned} \min_{y,u} J(y,u) &:= \min_{y,u} \left(\frac{1}{2} \|y - y_d\|_{L_2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L_2(U)}^2 \right) \\ &\text{subject to} \\ Ay &= Bu + f \quad \text{on } \Omega \\ u &\in U_{ad} := \{u \in L_2(U) \mid u_a \leq u \leq u_b \text{ a.e. in } U\} \end{aligned} \tag{2.1}$$

where the set $U \subset \bar{\Omega} \subset \mathbb{R}^d$ with $d = 1, 2, 3$. U_{ad} denotes the set of admissible controls and is a non empty, convex and closed subset of $L_2(U)$. Each $u \in U_{ad}$ is called an admissible control.

$$Ay = Bu + f \tag{2.2}$$

is a constraint for the state y . If the assumptions of lemma 1.15 are fulfilled, for each admissible control $u \in U_{ad}$, there exists a unique weak solution y to (2.2). The state y can therefore be written as $y = y(u)$. y_d denotes the desired state, α the regularization parameter. In this thesis two cases, Neumann boundary control, i.e. $U = \Gamma_{\mathcal{N}}$ and distributed control $U = \Omega$ are considered.

At this point only the most important theoretical results for these kind of optimal control problems are recalled. For further information and an overall introduction see [90, 158] and the references therein.

First, some general issues and notation are clarified, second the existence and uniqueness of solutions is investigated. Therewith, usually a so-called adjoint state, denoted by q is introduced. It can be calculated by using the Lagrange formulation (see e.g. [158]), and enables the rewriting of the problem as an inequality system. By using the adjoint, the control u can be described, whose regularity is considered afterwards.

First, general assumptions are stated.

Assumption 2.1. *The domain Ω is assumed to be open and bounded with polygonal boundary $\partial\Omega = \Gamma$. Let Y, Z denote Banach spaces over Ω . The differential operator $A \in \mathcal{L}(Y, Z)$ is assumed to be elliptic and bounded and the operator $B \in \mathcal{L}(L_2(U), Z)$. For the desired state y_d holds $y_d \in L_2(\Omega)$, the regularization parameter fulfills $\alpha > 0$ and $f \in L_2(\Omega)$. Furthermore, it has to hold that $u_a, u_b \in L_2(U)$ and $u_a \leq u_b$ almost everywhere.*

Second, a general statement to classify an optimal control, is given.

Definition 2.2. *A control $u^* \in U_{ad}$ and its corresponding state $y^* = y(u^*)$ is called optimal control and state respectively, if it holds*

$$J(y^*, u^*) \leq J(y(u), u) \quad \forall u \in U_{ad}.$$

Next, conditions for the existence of a unique solution are given.

Theorem 2.3. (see e.g. [158]) *Under assumption 2.1, there exists a solution to the optimal control problem (2.1). If $A^{-1}B$ is injective, there exists a unique solution.*

By introducing the adjoint q , the unique solution can be yielded under the following conditions:

Theorem 2.4. (see e.g. [158]) *Let assumption 2.1 holds. Then, the optimal control problem (2.1) has a unique solution (y^*, u^*, q^*) if and only if*

$$Ay^* = Bu^* + f, \quad (2.3)$$

$$A^*q^* = y^* - y_d, \quad (2.4)$$

$$\langle B^*q + \alpha u^*, u - u^* \rangle_U \geq 0 \quad \text{for all } u \in U_{ad}. \quad (2.5)$$

Next, the variational inequality (2.5) is reformulated.

Theorem 2.5. (see e.g. [90, 158]) *Let $P_{U_{ad}} : L_2(U) \rightarrow U_{ad}$ be the pointwise projection*

$$u(x) = \max\{u_a(x), \min\{u(x), u_b(x)\}\} \quad \text{for } x \in U$$

with the obvious modifications if less than two bounds are present. If $\alpha > 0$, then the variational inequality (2.5) is equivalent to the projection formula

$$u^* = P_{U_{ad}} \left(-\frac{1}{\alpha} B^*q^* \right). \quad (2.6)$$

With the help of the projection formula, the domain Ω is separated in different sets.

Definition 2.6. *The set*

$$\mathfrak{A} := \{x \in U : u^*(x) = u_a(x) \vee u^*(x) = u_b(x)\}$$

is called active set. The complement $\mathfrak{A}^C = U \setminus \mathfrak{A}$ is called inactive set and denoted by \mathfrak{I} . Accordingly, the points of \mathfrak{A} and \mathfrak{I} are called active or inactive points, respectively.

Remark 2.7. *The active set \mathfrak{A} is the union of the two sets*

$$\mathfrak{A}_a := \{x \in U : u^*(x) = u_a(x)\},$$

$$\mathfrak{A}_b := \{x \in U : u^*(x) = u_b(x)\}.$$

Furthermore, a statement on the regularity induced by the projection is given.

Theorem 2.8. (see e.g. [6, 102]) *Assume that $v \in W_p^\epsilon(U)$ along with $u_a, u_b \in W_p^\epsilon(U)$ and $\epsilon \in [0, 1]$. Then it holds $u := P_{U_{ad}}(v) \in W_p^\epsilon(U)$ for $U = \Omega, \Gamma_{\mathcal{N}}$.*

The theorem above gives a clue of how the regularity of the control is influenced by the regularity of the adjoint.

Next, the two considered optimal control problems are given. First, a model problem for Neumann boundary control, i.e. $U = \Gamma_{\mathcal{N}}$ is considered. In this case it holds

$$Bu(\cdot) = \int_{\Gamma_{\mathcal{N}}} u \gamma_0(\cdot) dx : L_2(\partial\Omega) \rightarrow \left(H_{\Gamma_{\mathcal{D}}}^1(\Omega) \right)^*$$

where γ_0 denotes the trace operator.

Second, a distributed optimal control problem, i.e. $U = \Omega$ is given. There, the operator B is

$$B = \text{Id} : L_2(\Omega) \rightarrow \left(H_{\Gamma_{\mathcal{D}}}^1(\Omega) \right)^*.$$

2.1 Boundary control problem

In the case of Neumann boundary control the following model problem is considered:

$$\min_{y,u} J(y,u) = \min_{y,u} \left(\frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 dx + \frac{\alpha}{2} \int_{\Gamma_{\mathcal{N}}} u(x)^2 dx \right)$$

subject to the elliptic boundary value problem

$$\begin{aligned} -\nabla \cdot (D(x)\nabla y(x)) + c(x)y(x) &= f(x) && \text{in } \Omega, \\ y(x) &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ D(x)\frac{\partial y}{\partial n}(x) &= u(x) && \text{on } \Gamma_{\mathcal{N}}, \end{aligned} \tag{2.7}$$

with box constraints on the control

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{a.e. in } \Gamma_{\mathcal{N}} \neq \emptyset.$$

The set of admissible controls u is denoted by

$$U_{ad} := \{u \in L_2(\Gamma_{\mathcal{N}}) : u_a \leq u \leq u_b \quad \text{a.e. on } \Gamma_{\mathcal{N}}\}.$$

To ensure unique solvability, the following assumptions are made (see [31]):

Assumption 2.9. *Let $D(x)$ be symmetric and positive definite in $\bar{\Omega}$, i.e. there exists a $D_0 > 0$ such that for all $x \in \bar{\Omega}$ it holds $\xi^\top D(x)\xi > D_0|\xi|^2$ for arbitrary $\xi \in \mathbb{R}^2$. Let $c(x) \geq 0$ for all $x \in \bar{\Omega}$ and $c(x) \geq c_0 > 0$ if $\text{meas}(\Gamma_{\mathcal{D}}) = 0$. In addition, it holds $\alpha > 0$, $u_a, u_b \in H^{1/2}(\Gamma_{\mathcal{N}})$ with $u_a \leq u_b$ a.e. on $\Gamma_{\mathcal{N}}$, and $f, y_d \in L_2(\Omega)$.*

To yield the desired regularity, additional assumptions on D and c and assumptions on the primal equation are made.

Assumption 2.10. *Let the functions D, c be analytic in $\bar{\Omega}$ and satisfy*

$$\|\nabla^p D\|_{L_\infty(\Omega)} + \|\nabla^p c\|_{L_\infty(\Omega)} \leq c_d \gamma_d^p p! \quad \forall p \in \mathbb{N}_0$$

for $c_d, \gamma_d > 0$.

There exists a constant $c_1 > 0$ such that for $f \in L_2(\Omega)$ and $u \in L_2(\Gamma_{\mathcal{N}})$ the solution to (2.7) is in $H^{3/2}(\Omega)$ and satisfies

$$\|y\|_{H^{3/2}(\Omega)} \leq c_1 \left(\|f\|_{L_2(\Omega)} + \|u\|_{L_2(\Gamma_{\mathcal{N}})} \right).$$

Additionally, there is a $\delta \in [1/2, 1]$ such that for $f \in L_2(\Omega)$ and $u \in H^{1/2}(\Gamma_{\mathcal{N}})$ the solution y to (2.7) is in $H^{1+\delta}(\Omega)$ and satisfies

$$\|y\|_{H^{1+\delta}(\Omega)} \leq c_1 \left(\|f\|_{L_2(\Omega)} + \|u\|_{H^{1/2}(\Gamma_{\mathcal{N}})} \right).$$

Under the assumption 2.9 and assumption 2.10, for a right hand side $f \in L_2(\Omega)$, and a desired state $y_d \in L_2(\Omega)$, the application of theorem 2.4 gives a unique solution $(y^*, u^*, q^*) \in$

$H^{1+\delta}(\Omega) \times H^{1/2}(\Gamma_{\mathcal{N}}) \times H^{1+\delta}(\Omega)$ (see [31]) for $\delta \in (0, 1]$. Then, by introducing the bilinear form $a(\cdot, \cdot)$ and a functional $\langle \cdot, \cdot \rangle_{\Omega}$, i.e.

$$a(y, v) = \int_{\Omega} D(x) \nabla y(x) \cdot \nabla v(x) \, dx + \int_{\Omega} c(x) y(x) v(x) \, dx, \quad (2.8)$$

$$\langle f, v \rangle_{\Omega} = \int_{\Omega} f(x) v(x) \, dx \quad (2.9)$$

and by using the equivalence of the variational inequality of (2.5) to

$$u^*(x) = P_{U_{ad}} \left(-\frac{1}{\alpha} q^*|_{\Gamma_{\mathcal{N}}}(x) \right) \quad \text{a.e. on } \Gamma_{\mathcal{N}}, \quad (2.10)$$

where $P_{U_{ad}}$ denotes the L_2 -projection onto the convex set U_{ad} , it follows:

Theorem 2.11. (see e.g. [158]) *Let assumption 2.9 and assumption 2.10 hold. There exists a unique solution $(y^*, u^*, q^*) \in H^{1+\delta}(\Omega) \times H^{1/2}(\Gamma_{\mathcal{N}}) \times H^{1+\delta}(\Omega)$ for $\delta \in (0, 1]$ to*

$$\begin{aligned} a(y^*, v) &= \langle f, v \rangle_{\Omega} + \langle u, v \rangle_{\Gamma_{\mathcal{N}}} & \forall v \in H_{\Gamma_{\mathcal{D}}}^1(\Omega), \\ a(v, q^*) &= \langle y - y_d, v \rangle_{\Omega} & \forall v \in H_{\Gamma_{\mathcal{D}}}^1(\Omega), \\ u^* &= P_{U_{ad}} \left(-\frac{1}{\alpha} q^*|_{\Gamma_{\mathcal{N}}}(x) \right). \end{aligned}$$

2.2 Distributed control problem

The distributed optimal control model problem, considered in this thesis, is given by

$$\min_{y, u} J(y, u) = \min_{y, u} \left(\frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 \, dx + \frac{\alpha}{2} \int_{\Omega} u(x)^2 \, dx \right) \quad (2.11)$$

subject to the elliptic boundary value problem

$$\begin{aligned} -\nabla \cdot (D(x) \nabla y(x)) + c(x) y(x) &= u(x) + f(x) & \text{in } \Omega, \\ y(x) &= 0 & \text{on } \Gamma_{\mathcal{D}}, \\ D(x) \frac{\partial y}{\partial n}(x) &= 0 & \text{on } \Gamma_{\mathcal{N}}. \end{aligned} \quad (2.12)$$

Additionally, the inequality constraints

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{a.e. in } \Omega$$

hold. To ensure unique solvability and smoothness of the boundary value problem (2.12) the following assumptions are made.

Assumption 2.12. *For the coefficients of the differential equation, there holds $D(x) \geq D_0 > 0$, $c(x) > 0$ and if $\text{meas}(\Gamma_{\mathcal{D}}) = 0$ it holds $c(x) \geq c_0 > 0$. Furthermore, it is assumed that $\alpha > 0$, $y_d, f \in L_2(\Omega)$, $u_a, u_b \in L_2(\Omega)$ with $u_a \leq u_b$ a.e. in Ω .*

Analogue to the assumptions in section 2.1, the following assumptions on the regularity are made.

Assumption 2.13. *It is assumed that the coefficients of the differential equation $D, c : \bar{\Omega} \rightarrow \mathbb{R}$ are bounded and analytic.*

Again, by using the assumption 2.12, it can be shown, that there exists a unique solution (y^*, u^*, q^*) . If the domain and the coefficients D, c are smooth enough, the regularity $u^* \in L_2(\Omega)$, $y^*, q^* \in H^2(\Omega)$ is possible. For more information on regularity of boundary value problems see [75].

Next, a theorem on the unique solvability is given.

Theorem 2.14. *It is assumed that assumption 2.12 holds. There exists a unique solution $(y^*, u^*, q^*) \in H^{1+\delta}(\Omega) \times L_2(\Omega) \times H^{1+\delta}(\Omega)$ to*

$$\begin{aligned} a(y^*, v) &= \langle u, v \rangle_{\Omega} + \langle f, v \rangle_{\Omega} & \forall v \in H_{\Gamma_D}^1(\Omega), \\ a(v, q^*) &= \langle y - y_d, v \rangle_{\Omega} & \forall v \in H_{\Gamma_D}^1(\Omega), \\ u^* &= P_{U_{ad}} \left(-\frac{1}{\alpha} q^*|_{\Omega}(x) \right). \end{aligned}$$

Here, the regularity shall not be investigated any further. Nevertheless, it is assumed that assumption 2.13 holds, due to the expectations to get a higher regularity in the interior of the domain. The projection formula (2.6) implies

$$u^*|_{\Omega_j}(x) = -\frac{1}{\alpha} q^*|_{\Omega_j}(x) \quad \forall x \in \Omega_j$$

where $\Omega_j \subset \mathfrak{J}$. Therewith, the regularity of u^* is partially higher. This is due to the fact that the regularity moves from the adjoint to the state, in the interior of Ω_j . Nevertheless, the boundary $\partial\Omega_j$ has to be considered too. However, the shape of the boundary Ω_j is unknown. For determining the regularity, the first step would be to get to know the shape of Ω_j , which is left as an open problem to further research.

3 The finite element method

In this chapter the most important basics of the finite element method (fem) are recalled. For simplicity, the model problem

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u(x)) + c(x)u(x) &= f(x) && \text{in } \Omega \\ u(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \\ D(x) \cdot \frac{\partial u}{\partial n}(x) &= 0 && \text{on } \Gamma_{\mathcal{N}} \end{aligned} \quad (3.1)$$

for an open and bounded Lipschitz domain Ω with boundary $\Gamma = \overline{\Gamma_{\mathcal{D}}} \cup \overline{\Gamma_{\mathcal{N}}}$, $\Gamma_{\mathcal{N}} \cap \Gamma_{\mathcal{D}} \neq \emptyset$ and $\text{meas}(\Gamma_{\mathcal{D}}) \neq 0$ if $c(x) = 0$, is considered. The weak form of problem (3.1) is denoted by

$$\int_{\Omega} D(x)\nabla u(x) \cdot \nabla v(x) \, dx + \int_{\Omega} c(x)u(x)v(x) \, dx = \int_{\Omega} f(x)v(x) \, dx \quad (3.2)$$

for an arbitrary test function $v(x)$. In order to fulfill the boundary conditions a suitable function space for $u(x)$ is the space $H_{\Gamma_{\mathcal{D}}}^1(\Omega)$, see e.g. [51, 151]. To solve the variational formulation of (3.2) approximately, some general remarks and important theorems to solve such kind of problems are recalled. Furthermore, different well-known refinement strategies that can later be applied to the two optimal control model problems are given.

In fem usually large equation systems are yielded. These equation systems shall be solved with iterative methods. Therefore, different preconditioners to lower the number of iterations are presented – and if necessary – adapted (see section 3.3). Moreover, some numerical results to understand the behavior of these preconditioners are given.

3.1 Basics of finite element method

First, following [36, 43, 44, 51, 152], a short overview on the most important basics of the finite element method are given. By using the bilinear form

$$a(u, v) = \int_{\Omega} D(x)\nabla u(x) \cdot \nabla v(x) \, dx + \int_{\Omega} c(x)u(x)v(x) \, dx$$

and the linear form $F(v) = \langle f, v \rangle_{\Omega}$, the model problem (3.1) can be written as abstract variational problem

$$\text{find } u \in V : \quad a(u, v) = F(v) \quad \forall v \in V. \quad (3.3)$$

For a unique solution to (3.3), the following assumptions are made:

Assumption 3.1.

1. V is a Hilbert space with scalar-product $\langle \cdot, \cdot \rangle_V$. The corresponding norm is denoted by $\| \cdot \|_V$.

2. The bilinear form $a : V \times V \rightarrow \mathbb{R}$ is

a) bounded on V :

$$\exists \mathbf{a} \geq 0 : \quad |a(u, v)| \leq \mathbf{a} \|u\|_V \|v\|_V \quad \forall u, v \in V,$$

b) V -elliptic (coerzive):

$$\exists \mathbf{b} > 0 : \quad \mathbf{b} \|u\|_V^2 \leq a(u, u) \quad \forall u \in V.$$

3. The linear functional $F : V \rightarrow \mathbb{R}$ is bounded, i.e.

$$\exists c_F \geq 0 : \quad |F(v)| \leq c_F \|v\|_V \quad \forall v \in V,$$

e.g. $F \in V^*$.

The uniqueness and existence of solutions to problem (3.3) are yielded by applying the lemma of Lax-Milgram 1.15.

Theorem 3.2. (see e.g. [51]) *The variational problem (3.3) possesses a unique solution under assumption 3.1. The solution $u \in V$ depends continuously on the data*

$$\|u\|_V \leq \frac{1}{\mathbf{b}} \|F\|_{V^*}.$$

It has to be mentioned, that the space V equipped with the scalar product $a(\cdot, \cdot)$ is again a Hilbert space. The induced norm is called energy norm and denoted by $\|\cdot\|_A$. Usually, variational problems as problem (3.3) cannot be solved analytically. Therefore numerical approximations as the finite element method are used. In this thesis only conforming finite element methods are considered. For non-conforming fem see e.g. [11, 66], for non-conforming methods as discontinuous Galerkin methods see e.g. [12, 13, 56, 59]. In conforming finite element methods the infinite-dimensional space V is approximated by a N dimensional space $V^{(N)} \subset V$. Then, instead of solving the variational problem (3.3), the Galerkin approximation

$$\text{find } u_N \in V^{(N)} : \quad a(u_N, v_N) = F(v_N) \quad \forall v_N \in V^{(N)} \quad (3.4)$$

is solved. The existence and uniqueness can again be shown by applying the lemma of Lax-Milgram 1.15.

Corollary 3.3. (see e.g. [151]) *Let $V^{(N)}$ be closed and a subset of V . Furthermore, the assumptions 3.1 hold for the discrete problem (3.4). Then (3.4) admits a unique solution u_N^* .*

Remark 3.4. *There holds the Galerkin orthogonality*

$$a(u^* - u_N^*, v_N) = 0 \quad \forall v_N \in V^{(N)}.$$

However, beside the existence and uniqueness of the discrete solution, the approximation to the solution is of interest. Therefore, Cea's lemma is used:

Theorem 3.5 (Cea's lemma). (see e.g. [36]) *Let u be the solution to (3.3) and u_N^* the discrete solution to (3.4). Then it holds*

$$\|u^* - u_N^*\|_V \leq \frac{\mathbf{a}}{\mathbf{b}} \inf_{v_N \in V^{(N)}} \|u^* - v_N\|_V.$$

Remark 3.6. *The discretization error is quasi-optimal, i.e.*

$$\inf_{v_N \in V^{(N)}} \|u^* - v_N\|_V \leq \|u^* - u_N^*\|_V \leq \frac{\mathfrak{a}}{\mathfrak{b}} \inf_{v_N \in V^{(N)}} \|u^* - v_N\|_V.$$

As approximation space, a family of conforming subspaces $V^{(N)} \subset V^{(N+1)}$ and $V^{(N)} \subset V$ of V with

$$\overline{\bigcup_{N=1}^{\infty} V^{(N)}} = V$$

is used. Equipping the finite dimensional space $V^{(N)}$ with a basis $\{\varphi_i(x)\}_{1 \leq i \leq N}$, the approximation of the infinite dimensional space V by the finite dimensional spaces $V^{(N)}$ leads to a linear equation system:

$$\vec{u} \in \mathbb{R}^N : \quad A\vec{u} = \vec{f}, \quad (3.5)$$

where functions in $V^{(N)}$ can be written as

$$u_N(x) = \sum_{i=1}^N u_i \varphi_i(x)$$

and $\vec{u} = (u_1, \dots, u_N)$. The entries of the matrix $A \in \mathbb{R}^{N \times N}$ and the right hand side $\vec{f} \in \mathbb{R}^N$ are given by

$$\begin{aligned} A_{ij} &= a(\varphi_j, \varphi_i), \\ f_i &= F(\varphi_i). \end{aligned}$$

By choosing a suitable basis, the matrix A is sparse. Preferably the matrix only has $\mathcal{O}(N)$ entries, see [36, 51, 93, 96]. The basis functions φ_i used in this thesis are introduced in section 3.2.1.

Remark 3.7. *There are two properties which are passed from the continuous problem to the discrete problem. The positive definiteness of the system matrix A follows by the ellipticity of the bilinear form $a(\cdot, \cdot)$. Furthermore, a symmetric bilinear form $a(\cdot, \cdot)$ leads to a symmetric matrix A .*

In the following section some important general results on finite element spaces and convergence are given. Moreover, the triangulation, which is used in the following subsections, is specified.

3.1.1 Finite element and triangulation

The focus in this thesis is on quadrilateral elements. More information on finite element definitions – also for triangular elements – can be found in [36, 43, 57, 58, 93, 151].

Before specifying the discretization, some notation is clarified, see e.g. [3, 23, 51, 150].

- An **element** K is an open quadrilateral set with $K \subset \Omega \subset \mathbb{R}^2$.
- Each open side of the quadrilateral is called **edge**.

- The endpoints of the edge are called **vertices** or **nodes**.

Definition 3.8. A vertex of an element is called **regular node** if and only if it is a vertex to each neighbouring element. Vertices which are not regular are called **irregular** or **hanging node**.

Remark 3.9. In three dimensions not only hanging nodes but also hanging edges occur.

Remark 3.10. For an efficient triangular local refinement, it is not necessary to allow hanging nodes. Instead, red-green refinement see e.g. [140] can be used.

Definition 3.11. A **k -irregular triangulation** τ_h of the domain $\Omega \subset \mathbb{R}^2$ is a collection of open, convex and nonempty elements K , such that

- $\bar{\Omega} = \bigcup_{K \in \tau_h} \bar{K}$
- $\bar{K}_i \cap \bar{K}_j$ is either empty, a vertex or an edge (in three dimensions even a face),
- each edge contains at most k irregular nodes.

In this thesis, the word triangulation is interchangeably used with mesh. Furthermore, only 1-irregular triangulations are used.

Each element in the mesh is associated with a size h and a polynomial degree p . In order to get a better approximation, the mesh is refined either by h , p or hp refinement (see Figure 3.1). This leads to three kinds of fem: h -fem, p -fem and hp -fem. h -refinement means, that elements are divided in order to get a better approximation to the solution. Whereas in uniform h -fem all elements are h -refined, in adaptive h -fem elements can be refined based on a-posteriori information by using error estimators (see e.g. [2, 162]). p -refinement increases the polynomial degree on the elements whilst hp -refinement is a combination of both strategies (see e.g. [140, 152]). While h -refinement is superior to p -refinement on elements where the solution or the domain is not smooth, p -refinement leads to better results on smooth parts, see e.g. [17, 18, 19, 20, 140].

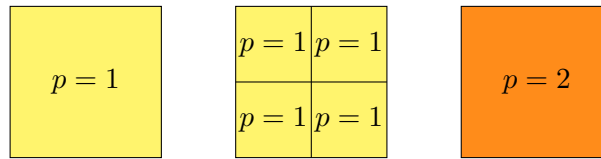


Figure 3.1: element, h -refinement, p -refinement

3.2 hp -finite element method

The advantage of using hp -fem is – assuming that the solution is smooth enough – that it is possible to obtain an exponential order of convergence. Although this is also possible in p -fem, the requirements on the smoothness are usually too high for non-academic examples, see e.g. [19, 140]. For h -fem the convergence rate is only algebraic. Moreover, hp -fem is also advantageous if it is applied to problems with only algebraic rate of convergence, since a suitable hp -refinement can substantially reduce the number of degrees of freedom (see e.g. [98]).

In this thesis the implementation of a hp -finite element code was one of the main parts.

Therewith, the main aspects of the implementation, some remarks and difficulties when implementing hp -finite element code will be given. An overall introduction in implementing hp -fem is given in [57, 58]. First, the reference element \hat{K} , one of the main ingredients of fem, is given.

3.2.1 Reference element and basis functions

Again, the focus is on the two-dimensional case. The extension to three dimensions is straightforward, see [33]. In three dimensions, the already implemented basis functions of the 3D Fortran code¹ were used.

As reference element, the open square $\hat{K} = (-1, 1)^2$ is chosen. The numbering of the nodes (vertices), the edges and the orientation of the edges is given in figure 3.2.

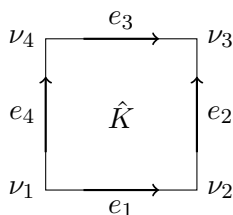


Figure 3.2: reference element \hat{K}

When selecting suitable basis functions for finite elements, several points have to be considered. The basis functions have to enable a fast computation, the yielded system matrix has to be sparse and has to have a suitable condition number. Therefore, hierarchical basis functions based on a family of orthogonal polynomials are preferred on quadrilateral elements. The advantage of such a choice is the sparsity and hence the fast implementation, since not the whole basis has to be recomputed if the polynomial degree is increased. For more information see e.g. [25, 34, 36, 96]. For a hierarchical basis usually three kinds of basis functions are distinguished, vertex (node or hat) basis functions, edge bubbles and element bubbles, see figure 3.3. In this thesis integrated Legendre polynomials, see section 1.4, are used to construct

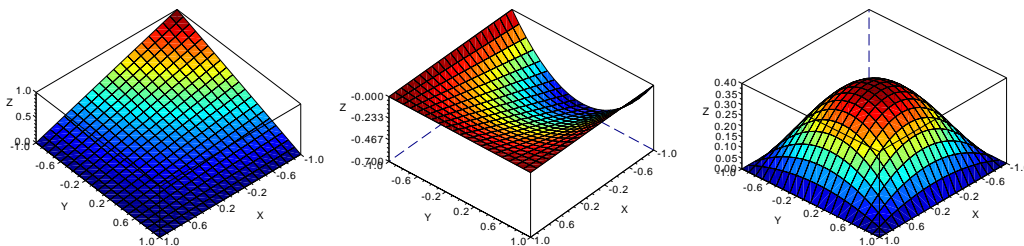


Figure 3.3: on one element: vertexbasisfunction, edge- and element bubble, respectively.

the basis, see e.g. [16, 35, 101, 109, 120, 125]. Advantageously the Legendre polynomials can be efficiently computed by using orthogonality relations and recurrence formulas, see lemma 1.12. For square elements (or cubes) it is even possible to exploit the tensor product structure

¹see [33]

to compute the element matrices. The basis functions for the reference element 3.2 are given in the following. The nodal or hat basis functions are given by

$$\begin{aligned}\varphi_1^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 - \hat{x}_1)(1 - \hat{x}_2), & \varphi_2^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 + \hat{x}_1)(1 - \hat{x}_2), \\ \varphi_3^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 + \hat{x}_1)(1 + \hat{x}_2), & \varphi_4^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 - \hat{x}_1)(1 + \hat{x}_2).\end{aligned}$$

The edge bubbles are given by

$$\begin{aligned}\varphi_i^{e_1}(\hat{x}_1, \hat{x}_2) &= \hat{L}_i(\hat{x}_1) \frac{1 - \hat{x}_2}{2} \quad i \geq 2, & \varphi_i^{e_2}(\hat{x}_1, \hat{x}_2) &= \frac{1 + \hat{x}_1}{2} \hat{L}_i(\hat{x}_2) \quad i \geq 2, \\ \varphi_i^{e_3}(\hat{x}_1, \hat{x}_2) &= \hat{L}_i(\hat{x}_1) \frac{1 + \hat{x}_2}{2} \quad i \geq 2, & \varphi_i^{e_4}(\hat{x}_1, \hat{x}_2) &= \frac{1 - \hat{x}_1}{2} \hat{L}_i(\hat{x}_2) \quad i \geq 2,\end{aligned}$$

with $1 - p_{e_k}$ edge bubbles on each edge where p_{e_k} denotes the polynomial degree on edge e_k for $k = 1, \dots, 4$. Furthermore, let denote p_C , the polynomial degree on the element. There are $(p_C - 1)^2$ element bubbles on each element given by the relation

$$\varphi_{ij}^C(\hat{x}_1, \hat{x}_2) = \hat{L}_i(\hat{x}_1) \hat{L}_j(\hat{x}_2) \quad i, j \geq 2.$$

Remark 3.12. *In three dimensions the basis is constructed analogously. There are not only node basis functions, edge bubbles, element bubbles but also face bubbles occur.*

The given basis functions (see figure 3.3 for examples) are defined on the reference element \hat{K} and are suitably extended outside the element. There, an extension by zero is used if possible. The nodal functions are the usual hat functions. The support for an edge bubble are the two elements which share this edge. On all other elements they can be extended by zero due to (1.12) and (1.13). The element bubbles can continuously be extended by zero to the neighbour elements due to (1.12) and (1.13).

Remark 3.13. *It has to be stated that the orientation of the edges is given by the edge bubbles.*

3.2.2 Mapping

A key issue in the design, the analysis and the numerical realization of fem is the mapping of the reference element (see e.g. [57, 152]). The trick is that each physical (or global) element K is defined as a transformation of the reference element \hat{K} . This construction has several advantages, especially the numerical integration and differentiation only need to be performed on the reference element. The transformation from the reference element \hat{K} to the global element K is denoted by

$$\phi_K : \hat{K} \rightarrow K.$$

Suitable mappings are continuously differentiable, one-to-one and onto mappings. That means, if \hat{x} denotes a coordinate system on \hat{K} , $x = \phi_K(\hat{x})$ is the corresponding coordinate system on K . In this thesis isoparametric mappings are used, see e.g. [152]. Since only non-curved elements are allowed, a bilinear mapping is sufficient. It is given by

$$\begin{aligned}x_1 &= \sum_{i=1}^4 \varphi_i^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) X_i, \\ x_2 &= \sum_{i=1}^4 \varphi_i^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) Y_i.\end{aligned}$$

where (X_i, Y_i) for $i = 1, \dots, 4$ are the coordinates of the global element. Therewith, the Jacobian matrix F_K is denoted by

$$F_K = \begin{pmatrix} \frac{\partial x_1}{\partial \hat{x}_1} & \frac{\partial x_2}{\partial \hat{x}_1} \\ \frac{\partial x_1}{\partial \hat{x}_2} & \frac{\partial x_2}{\partial \hat{x}_2} \end{pmatrix}$$

and the determinant is given by $J_K = \det(F_K)$. In the case of affine-linear elements, the determinant of the transformation is constant and especially independent of x .

Remark 3.14. *It has to be mentioned that the orientation of the edges comes into play in the mapping from the reference element \hat{K} to the global element K for quadrilateral elements, since in general it is not possible to find mappings with correct orientation a-priori.*

Remark 3.15. *For triangular elements the orientation of a global element K coincide with the orientation of the reference element \hat{K} if barycentric coordinates are used to define the basis functions, or if the mapping is suitably chosen.*

In the following, further standard definitions on mappings are given.

Definition 3.16. *Let τ_h be a triangulation of Ω and h_K the diameter of the element $K \in \tau_h$. The triangulation τ_h is called **γ -shape regular**, if there is a constant γ , such that*

$$h_K^{-1} \|F_K\|_{L^\infty(K)} + h_K \|(F_K)^{-1}\|_{L^\infty(K)} \leq \gamma \quad \forall K \in \tau_h.$$

Due to the structure of the considered boundary value problems, it makes sense to use H^1 -conforming spaces. For other problem classes this may be different, see e.g. [168]. Therefore, the polynomial space on an interval and a square are defined by

$$\begin{aligned} I_p &:= \{x^i\}_{i=0, \dots, p}, \\ Q_p &:= \{x_1^i x_2^j\}_{0 \leq i, j \leq p}. \end{aligned}$$

Then, the space for the finite elements for a given triangulation τ_h , where all edges of a given element K are collected in the set E_K , is defined by

$$V_{hp} := \{v \in H_{\Gamma_D}^1(\Omega) : v|_K \circ F_K \in Q_{p_K} \quad v|_{e_K} \circ F_K \in I_{p_K} \text{ for all } K \in \tau_h \text{ and all } e_K \in E_K\}.$$

3.2.3 Element matrices

The introduced basis functions and the mapping enable to set up of the element matrices and the corresponding right-hand side on the element level. The ordering of the basis functions is

$$\begin{pmatrix} \varphi_{\mathcal{V}\mathcal{V}} & \varphi_{\mathcal{V}E} & \varphi_{\mathcal{V}C} \\ \varphi_{E\mathcal{V}} & \varphi_{EE} & \varphi_{EC} \\ \varphi_{C\mathcal{V}} & \varphi_{CE} & \varphi_{CC} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \varphi_{\mathcal{V}} \\ \varphi_E \\ \varphi_C \end{pmatrix}$$

where \mathcal{V} denotes vertex based parts, E edge based parts and C the element based part of the matrix and vectors. For affine-linear elements there are two possibilities to set up these element matrices. Either one-dimensional matrices or suitable quadrature rules as the Gauss-integration are used. In both cases the tensor product structure is used. In the first case, the Kronecker product is applied to the one-dimensional matrices. In the second one, a Gauss quadrature, i.e. Gauss-Legendre quadrature rule, is used to calculate the integrals. Then, in order to make integration efficiently, the runtime of the integration is decreased by using the tensor-product structure see e.g. [34, 61, 113, 122].

Remark 3.17. *If the orientation of the reference element \hat{K} does not coincide with the orientation of the element K , the corresponding entries have to be multiplied with -1 in the element matrix for odd polynomial degrees. Due to efficiency reasons, the usual approach is not to multiply the corresponding matrix entries, but the corresponding entries of the vector multiplied with the matrix. This procedure is especially practicable, if the multiplication is done on the element level with an assembling of the corresponding element vectors as it is used in the implemented code for this thesis.*

Remark 3.18. *In three dimensions the orientation of edges and faces has to be adjusted.*

3.2.4 Assembling

The assembling of the matrices is – as the choice of the basis functions – also more difficult in hp -fem than in h -fem. For assembling the global matrices and vectors from the element matrices and element vectors, each local degree of freedom on the element level is associated with a global degree of freedom in the global mesh. For an overall introduction see e.g. [57, 58]. Fast assembling techniques are given in [61, 113].

A further important – and quite nasty – task in the used triangulation, is the handling of irregular nodes. Since a conforming solution is calculated, hanging nodes as the corresponding edges, are in fact no real degrees of freedom. In order to get a conforming discretization, the value on the hanging nodes and the corresponding edges has to be fixed by the bigger neighbouring edges/nodes. To illustrate the problem, the simple case of $p = 1$ is considered in detail.

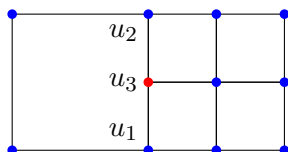


Figure 3.4: regular nodes (blue) and hanging node (red)

Figure 3.4 shows a mesh with a hanging node (red). In order to get a conforming solution on the mesh for polynomial degree $p = 1$, the value u_3 in the solution vector has to be adjusted by

$$u_3 = \frac{1}{2} (u_1 + u_2)$$

since the hanging node is exactly in the middle of u_1 and u_2 . For higher order basis functions this approach has to be extended. This will be done in the following subsection.

3.2.5 Hanging nodes and projector

The idea to get a conforming solution although hanging nodes appear is to use a projector, see e.g. [3, 116, 150, 161] for literature on this topic.

These papers show, that the most promising ansatz when dealing with hanging nodes is the use of local applicable projectors P_{loc} . This ansatz saves time, storage and avoids a global matrix-matrix multiplication, see [161]. In this section a locally applicable projector is derived. For deriving the structure and the numbers with respect to the used basis functions [126] was used. There, the given results are derived for a similar basis.

Before deriving the used (global) projector, a suitable algorithm to apply the projector locally is given in algorithm 4.

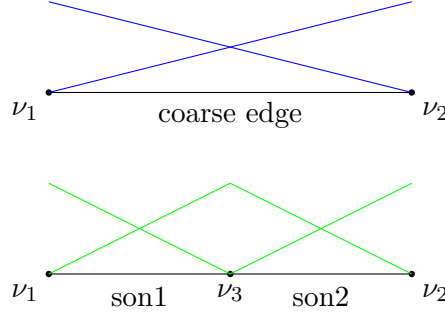


Figure 3.5: notation for refinement of edges

Figure 3.5 shows a coarse edge with its two sons. The basis functions (for polynomial degree $p = 1$) on the coarse edge are drawn in blue, the basis functions on the refined edges in green.

Algorithm 4: global application of projector

input : vector \vec{u}

output: vector \vec{w}

for $k = 1, \dots, \#edges$ **do**

if edge k has hanging node **then**

 determine the two sons of edge k (called son1 and son2)

 determine all nodes and edges connected to edge k and to its two sons

 apply local projector P_{loc} to them, i.e. $\vec{w}|_{e_k} = P_{loc}\vec{u}|_{e_k}$

Remark 3.19. *The order of son1 and son2 is fixed by the orientation of the edge. To avoid problems with the orientation (in the implementation), it is recommended to use the orientation of the coarse edge also for its sons.*

Next, the structure of the local projector P_{loc} is given. As an example, the projector for polynomial degree $p = 4$ is given.

$$P_{loc}^\top = \left(\begin{array}{ccc|ccc|ccc|ccc} 1 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \mathfrak{B}_{2,1}^{(12)} & \mathfrak{B}_{2,2}^{(1)} & 0 & 0 & \mathfrak{B}_{2,2}^{(2)} & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \mathfrak{B}_{3,1}^{(12)} & \mathfrak{B}_{3,2}^{(1)} & \mathfrak{B}_{3,3}^{(1)} & 0 & \mathfrak{B}_{3,2}^{(2)} & \mathfrak{B}_{3,3}^{(2)} & 0 & 0 & 1 & 0 \\ 0 & 0 & \mathfrak{B}_{4,1}^{(12)} & \mathfrak{B}_{4,2}^{(1)} & \mathfrak{B}_{4,3}^{(1)} & \mathfrak{B}_{4,4}^{(1)} & \mathfrak{B}_{4,2}^{(2)} & \mathfrak{B}_{4,3}^{(2)} & \mathfrak{B}_{4,4}^{(2)} & 0 & 0 & 1 \end{array} \right) \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ e_{son1,2} \\ e_{son1,3} \\ e_{son1,4} \\ e_{son2,2} \\ e_{son2,3} \\ e_{son2,4} \\ e_{coarse,2} \\ e_{coarse,3} \\ e_{coarse,4} \end{pmatrix}$$

The small matrix \mathfrak{B} (for arbitrary p) which is given through

$$\left(\begin{array}{c|cccc|cccc} \mathfrak{B}_{2,1}^{(12)} & \mathfrak{B}_{2,2}^{(1)} & 0 & \dots & 0 & \mathfrak{B}_{2,2}^{(2)} & 0 & \dots & 0 \\ \mathfrak{B}_{3,1}^{(12)} & \mathfrak{B}_{3,2}^{(1)} & \mathfrak{B}_{3,3}^{(1)} & \dots & 0 & \mathfrak{B}_{3,2}^{(2)} & \mathfrak{B}_{3,3}^{(2)} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathfrak{B}_{p,1}^{(12)} & \mathfrak{B}_{p,2}^{(1)} & \mathfrak{B}_{p,3}^{(1)} & \dots & \mathfrak{B}_{p,p}^{(1)} & \mathfrak{B}_{p,2}^{(2)} & \mathfrak{B}_{p,3}^{(2)} & \dots & \mathfrak{B}_{p,p}^{(2)} \end{array} \right)$$

can be calculated with a linear system of equations which depend on the basis functions. In here, the basis is given by the integrated Legendre polynomials, introduced in subsection 3.2.1. Since the coefficients $\mathfrak{B}_{i,j}$ are known for the unscaled integrated Legendre polynomials $\tilde{L}_i(x)$, these results, computed with the aid of computer algebra systems in [126], are used. Then, knowing the matrix \mathfrak{B} which has the same structure as the matrix \mathfrak{B} , the coefficients of \mathfrak{B} are determined by the coefficients of $\tilde{\mathfrak{B}}$. For the determination of the coefficients of $\tilde{\mathfrak{B}}$ the integrated Legendre polynomials are not only represented on the coarse edge $(-1, 1)$ but also on its two sons $(-1, 0)$ and $(0, 1)$. The first two integrated Legendre polynomials, the hat functions, are given by definition through

$$\begin{aligned} \hat{L}_0(x) &= \tilde{L}_0(x) = \frac{1-x}{2} & -1 \leq x \leq 1, \\ \hat{L}_1(x) &= \tilde{L}_1(x) = \frac{1+x}{2} & -1 \leq x \leq 1. \end{aligned}$$

On the two sons, the subintervals $I^{(1)} = [-1, 0]$ and $I^{(2)} = [0, 1]$ it holds

$$\begin{aligned} \hat{L}_1^{(1)}(x) &= \tilde{L}_1^{(1)}(x) = \tilde{L}_1(2x+1) & -1 \leq x \leq 0, \\ \frac{1}{\gamma_i} \hat{L}_i^{(1)}(x) &= \tilde{L}_i^{(1)}(x) = \tilde{L}_i(2x+1) & -1 \leq x \leq 0 \quad i \geq 2, \\ \hat{L}_1^{(2)}(x) &= \tilde{L}_1^{(2)}(x) = \tilde{L}_1(2x-1) & 0 \leq x \leq 1, \\ \frac{1}{\gamma_i} \hat{L}_i^{(2)}(x) &= \tilde{L}_i^{(2)}(x) = \tilde{L}_i(2x-1) & 0 \leq x \leq 1 \quad i \geq 2, \end{aligned}$$

and furthermore it is

$$\begin{aligned} \tilde{L}_0^{(1)}(x) &= \hat{L}_0^{(1)}(x) = \hat{L}_0(2x+1), \\ \tilde{L}_0^{(2)}(x) &= \hat{L}_0^{(2)}(x) = \hat{L}_0(2x-1). \end{aligned}$$

Therewith, one can define

$$\begin{aligned} \tilde{L}_0(x) &= \hat{L}_0(x) := \hat{L}_0^{(1)}(x) + \frac{1}{2} \left(\hat{L}_1^{(1)}(x) + \hat{L}_1^{(2)}(x) \right), \\ \tilde{L}_1(x) &= \hat{L}_1(x) := \hat{L}_0^{(2)}(x) + \frac{1}{2} \left(\hat{L}_1^{(1)}(x) + \hat{L}_1^{(2)}(x) \right). \end{aligned}$$

The matrix $\tilde{\mathfrak{B}}$ is, as the matrix \mathfrak{B} , a $(p-1) \times 2p$ matrix and the coefficients can be calculated by the following linear system of equations:

$$\tilde{L}_i(x) = \tilde{\mathfrak{B}}_{i,1}^{(1)} \tilde{L}_1^{(1)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(1)} \tilde{L}_j^{(1)}(x) + \tilde{\mathfrak{B}}_{i,1}^{(2)} \tilde{L}_1^{(2)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(2)} \tilde{L}_j^{(2)}(x) \quad i \geq 2.$$

Remark 3.20. *There it is always $\tilde{\mathfrak{B}}_{i,1}^{(1)} = \tilde{\mathfrak{B}}_{i,1}^{(2)}$. Therefore, it is set*

$$\tilde{\mathfrak{B}}_{i,1}^{(12)} = \tilde{\mathfrak{B}}_{i,1}^{(1)} = \tilde{\mathfrak{B}}_{i,1}^{(2)}$$

and the corresponding basis function $\tilde{L}_1^{(12)} = \hat{L}_1^{(12)}$ is given by

$$\tilde{L}_1^{(12)}(x) = \begin{cases} \tilde{L}_1^{(1)}(x) & -1 \leq x < 0, \\ \tilde{L}_1^{(2)}(x) & 0 \leq x \leq 1. \end{cases}$$

So in fact one solves the linear system of equations:

$$\tilde{L}_i(x) = \tilde{\mathfrak{B}}_{i,1}^{(12)} \tilde{L}_1^{(12)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(1)} \tilde{L}_j^{(1)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(2)} \tilde{L}_j^{(2)}(x) \quad i \geq 2. \quad (3.6)$$

The coefficients are calculated by equation of coefficients over both subintervals with the aid of computer algebra, whereas one has to consider that

$$\begin{aligned} \tilde{L}_j^{(2)}(x) &= 0 & \text{for all } j & \quad -1 \leq x \leq 0 \\ \tilde{L}_j^{(1)}(x) &= 0 & \text{for all } j & \quad 0 \leq x \leq 1 \end{aligned}$$

i.e. the two linear equation systems

$$\begin{aligned} \tilde{L}_i(2x+1) &= \tilde{\mathfrak{B}}_{i,1}^{(12)} \tilde{L}_i^{(12)}(2x+1) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(1)} \tilde{L}_j^{(1)}(2x+1) & -1 \leq x \leq 0 \\ \tilde{L}_i(2x-1) &= \tilde{\mathfrak{B}}_{i,1}^{(12)} \tilde{L}_i^{(12)}(2x-1) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(2)} \tilde{L}_j^{(2)}(2x-1) & 0 \leq x \leq 1, \end{aligned}$$

have to be solved. The coefficients and their calculation can be found in algorithm 5. There

$$(a)_n := a(a+1) \cdot \dots \cdot (a+n-1)$$

is used. Since (3.6) is equivalent to

$$\begin{aligned} \gamma_i \tilde{L}_i(x) &= \tilde{\mathfrak{B}}_{i,1}^{(12)} \gamma_i \tilde{L}_1^{(12)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(1)} \frac{\gamma_i}{\gamma_j} \tilde{L}_j^{(1)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(2)} \frac{\gamma_i}{\gamma_j} \tilde{L}_j^{(2)}(x) & i \geq 2, \\ \hat{L}_i(x) &= \tilde{\mathfrak{B}}_{i,1}^{(12)} \gamma_i \hat{L}_1^{(12)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(1)} \frac{\gamma_i}{\gamma_j} \hat{L}_j^{(1)}(x) + \sum_{j=2}^i \tilde{\mathfrak{B}}_{i,j}^{(2)} \frac{\gamma_i}{\gamma_j} \hat{L}_j^{(2)}(x) & i \geq 2, \end{aligned}$$

Algorithm 5: calculation of \mathfrak{B}

input : polynomial degree p **output:** \mathfrak{B} set $\tilde{\mathfrak{B}} = 0$ **for** $i = 1, \dots, \lfloor \frac{p}{2} \rfloor$ **do**

$$\tilde{\mathfrak{B}}_{2i,1}^{(12)} = (-1)^{i+1} \frac{(-\frac{1}{2})^i}{i!}$$

for $i = 1, \dots, p$ **do**

$$\tilde{\mathfrak{B}}_{i,i}^{(1)} = \tilde{\mathfrak{B}}_{i,i}^{(2)} = 2^{-i}$$

for $i = 2, \dots, \lfloor \frac{p}{2} \rfloor$ **do**

$$\tilde{\mathfrak{B}}_{2i,2}^{(1)} = \frac{3}{2} \tilde{\mathfrak{B}}_{2i,1}^{(12)}$$

$$\tilde{\mathfrak{B}}_{2i,2}^{(2)} = \tilde{\mathfrak{B}}_{2i,2}^{(1)}$$

$$\tilde{\mathfrak{B}}_{2i+1,2}^{(1)} = (-1)^i \frac{(-\frac{3}{2})_{i+1}}{(i+1)!}$$

$$\tilde{\mathfrak{B}}_{2i+1,2}^{(2)} = -\tilde{\mathfrak{B}}_{2i+1,2}^{(1)}$$

for $j = 2, \dots, p-1$ **do****for** $i = 1, \dots, j+1$ **do**

$$\tilde{\mathfrak{B}}_{i+2,j+1}^{(1)} = -\tilde{\mathfrak{B}}_{i,j+1}^{(1)} + \frac{1}{2} \tilde{\mathfrak{B}}_{i+1,j}^{(1)} - \tilde{\mathfrak{B}}_{i+1,j+1}^{(1)} + \frac{1}{2} \tilde{\mathfrak{B}}_{i+1,j+2}^{(1)}$$

for $i = 1, \dots, \lfloor \frac{p}{2} \rfloor$ **do****for** $j = 2, \dots, \lfloor \frac{p}{2} \rfloor$ **do**

$$\tilde{\mathfrak{B}}_{2i,j}^{(2)} = (-1)^{j+1} \tilde{\mathfrak{B}}_{2i,j}^{(1)}$$

$$\tilde{\mathfrak{B}}_{2i+1,j}^{(2)} = (-1)^j \tilde{\mathfrak{B}}_{2i+1,j}^{(1)}$$

the coefficients of \mathfrak{B} are then given by

$$\begin{aligned}
\mathfrak{B}_{2i,1}^{(12)} &= \tilde{\mathfrak{B}}_{2i,1}^{(12)} & i \geq 1, \\
\mathfrak{B}_{2i+1,1}^{(12)} &= 0 & i \geq 1, \\
\mathfrak{B}_{i,i}^{(1)} &= \mathfrak{B}_{i,i}^{(2)} = \tilde{B}_{i,i}^{(2)} & i \geq 2, \\
\mathfrak{B}_{2i,2}^{(1)} &= \frac{\gamma_{2i}}{\gamma_2} \tilde{\mathfrak{B}}_{2i,2}^{(1)} & i > 1, \\
\mathfrak{B}_{2i+1,2}^{(1)} &= \frac{\gamma_{2i+1}}{\gamma_2} \tilde{\mathfrak{B}}_{2i+1,2}^{(1)} & i \geq 1, \\
\mathfrak{B}_{2i+1,2}^{(2)} &= \frac{\gamma_{2i+1}}{\gamma_2} \tilde{\mathfrak{B}}_{2i+1,2}^{(2)} & i \geq 1, \\
\mathfrak{B}_{2i,j}^{(2)} &= \frac{\gamma_{2i}}{\gamma_j} \tilde{\mathfrak{B}}_{2i,j}^{(2)} & i \geq 1, j \geq 2, \\
\mathfrak{B}_{2i+1,j}^{(2)} &= \frac{\gamma_{2i+1}}{\gamma_j} \tilde{\mathfrak{B}}_{2i+1,j}^{(2)} & i \geq 1, j \geq 2,
\end{aligned}$$

and $\mathfrak{B}_{i+2,j+1}^{(1)} = \frac{\gamma_{i+2}}{\gamma_{j+1}} \tilde{\mathfrak{B}}_{i+2,j+1}^{(1)}$, $\mathfrak{B}_{i+2,j+1}^{(2)} = \frac{\gamma_{i+2}}{\gamma_{j+1}} \tilde{\mathfrak{B}}_{i+2,j+1}^{(2)}$ and $\mathfrak{B}_{i,j}^{(k)} = 0$ for $k = 1, 2$ and $j > i$. The developed projector can now be used to solve linear systems of equations. Next, details on using the projector in the preconditioned conjugate gradient method (PCG) are given.

3.2.5.1 Projected PCG

For solving a linear equation system

$$A\vec{u} = \vec{f}$$

with $A \in \mathbb{R}^{N \times N}$, $\vec{u}, \vec{f} \in \mathbb{R}^N$ with a symmetric and positive definite matrix A and a positive definite and symmetric preconditioner C , a modification of algorithm 1 is used. The projected PCG is proposed in [116] and substitutes each preconditioning step

$$\vec{w} = C^{-1}A\vec{r}$$

by

$$\vec{w} = PC^{-1}P^\top A\vec{r}$$

in algorithm 1. For general remarks on the PCG see also subsection 1.2.1.

Remark 3.21. *The global application of the projector, i.e. the application of P is in fact a call of algorithm 4.*

Remark 3.22. *For a local application of the projector only the matrix \mathfrak{B} for the highest polynomial degree in the whole mesh needs to be stored.*

3.2.5.2 Dirichlet conditions

In the case of boundary value problems with Dirichlet conditions, the so called OXER technique is used for iterative methods, see [93, 99]. There, the diagonal of the assembled matrix is multiplied with a huge number, e.g. 10^{40} for each degree of freedom on a Dirichlet edge.

The corresponding entries on the right-hand side are multiplied with the same huge number. To enforce Dirichlet conditions although for meshes with hanging nodes, the projector has to be adjusted. For a diagonal preconditioning matrix $D = \text{diag}(A)$, the matrix \tilde{D} given by

$$\tilde{D}_{ii} = \begin{cases} 0 & \text{if Dirichlet conditions holds on edge or node} \\ 1 & \text{otherwise} \end{cases}$$

can be used by replacing the projected preconditioning step in algorithm 1 by

$$\tilde{D}PD^{-1}P^\top\tilde{D}\vec{r}.$$

3.2.5.3 Mapping from coarse to fine mesh

The derived projector is also used to project a calculated solution on a coarser mesh to a finer one. This procedure is especially important in order to start with a suitable solution in the semismooth Newton method. For simplicity it is assumed, that in each refinement step only h -, p -refinement or no refinement is performed. Then, for a general mapping $M_{K,coarse}^{fine}$ two cases are distinguished: h -refinement and p -refinement. In h -refinement each coarse element leads to four refined elements. The transformation of the values (since the polynomial degree can already be higher than one) of the coarse basis functions to the refined ones is done by the Kronecker product of the projector. In the case of p -refinement the situation is easier, since only the new entries of the basis functions have to be filled with zero. Algorithm 6 shows the implementation. With this algorithm it is even possible to map from the coarse mesh to the

Algorithm 6: mapping from coarse to fine mesh

input : vector on coarse mesh \vec{u}^{coarse} , refined mesh τ_h

output: vector on fine mesh \vec{u}^{fine}

for $k = 1, \dots, \#elements$ (coarse mesh) **do**

if element k was h -refined **then**

 | $\vec{u}_{el}^{fine} = (P \otimes P)M_{K,coarse}^{fine}\vec{u}_{el}^{coarse}$

else if element k was p -refined **then**

 | $\vec{u}_{el}^{fine} = M_{K,coarse}^{fine}\vec{u}_{el}^{coarse}$

else

 | $\vec{u}_{el}^{fine} = \vec{u}_{el}^{coarse}$

finest one, if it is applied after each refinement. \vec{u}_{el} denotes the element-vector on the fine or coarse mesh respectively.

3.2.6 Refinement strategies

In this subsection the basic strategies for refining the mesh used in this thesis are given. For more information see e.g. [57, 140]. A very important condition in hp -fem, is the so called minimum degree condition given in the following definition.

Definition 3.23. Let τ_h be a shape-regular, 1-irregular triangulation. The collection of polynomial degrees p_K on an element $K \in \tau_h$ is called polynomial degree vector $p := (p_K)_{K \in \tau_h}$.

For each edge e_K with $e_K \in \overline{K}$, the **minimum degree condition**

$$p_{e_K} := \min\{p_{K'} : e_K \cap \overline{K'} \neq \emptyset, K' \in \tau_h\}$$

has to hold.

This condition ensures, that a unique and not too high polynomial degree on each edge is chosen.

Remark 3.24. In order to enforce the minimum degree condition in an implementation, each element and all of its edges are set to the chosen polynomial degree. In a second step, the polynomial degrees on the edges are adjusted. Therefore for each edge the polynomial degree of its neighbouring elements is compared, the minimum of these gives the polynomial degree on the edge.

For deciding if h - or p -refinement in a given element is superior, knowledge on the smoothness is necessary, see e.g [17, 18, 77, 78]. The smoothness is influenced by the domain and the solution. In case of the domain it depends on the boundary, especially on angles of the corners of the boundary. Therewith, due to geometry, a suitable a-priori refinement is h -refinement in all corners of the domain, or even to do h -refinement on the whole boundary. The smoothness of the solution is harder to guess a-priori. It might follow from the structure of the problem, as it is in the case of the considered optimal control problems. However, also error estimates, which decide if h - or p -refinement has to be performed, can be used in order to generate a mesh automatically. In the numerical examples different kinds of hp -refinement are used. In this section two strategies, the boundary concentrated fem and an error estimated refinement are pointed out.

3.2.6.1 Boundary concentrated fem

The boundary concentrated fem (bc-fem) goes back to [98]. To clarify the refinement, two definitions are necessary. The first one explains the h -refinement.

Definition 3.25. Let τ_h be a shape-regular and 1-irregular triangulation and denote $h := \min_{\overline{K} \cap \partial\Omega \neq \emptyset} \{h_K\} < 1$ a measure for the mesh size on the boundary. If there exist constants $c_1, c_2 > 0$ such that $K \in \tau_h$:

1. if $\overline{K} \cap \partial\Omega \neq \emptyset$, then $h \leq h_k \leq c_2 h$
2. if $\overline{K} \cap \partial\Omega = \emptyset$, then $c_1 \inf_{x \in K} \text{dist}(x, \partial\Omega) \leq h_K \leq c_2 \sup_{x \in K} \text{dist}(x, \partial\Omega)$.

the mesh is called **geometric mesh**.

The second one clarifies the choice of the polynomial degree:

Definition 3.26. Let τ_h be a geometric mesh with mesh size h . Furthermore, the polynomial degree vector $\mathbf{p} = (p_K)_{K \in \tau_h}$ is said to be **linear** with slope $\zeta > 0$, if there exist constants $c_1, c_2 > 0$ such that

$$1 + \zeta c_1 \log\left(\frac{h_K}{h}\right) \leq p_K \leq 1 + \zeta c_2 \log\left(\frac{h_K}{h}\right).$$

A geometric mesh with linear polynomial degree vector is called boundary concentrated mesh. The corresponding space is denoted by $V_{bc}(\Omega)$. The application of the boundary concentrated fem is especially recommended if there is high smoothness in the interior and low on the boundary. An application of this method to suitable problems leads to a strong reduction of degrees of freedom. For further information see subsection 3.2.7.

3.2.6.2 Error estimators

Another possibility to suitably refine the mesh is the use of error estimators. There the idea is to find elements with large errors and refine them. More information can be found e.g. in [2, 36, 162].

One of the most difficult parts when applying error estimators in hp -fem, is the decision if h - or p -refinement has to be done. In pure h -fem only the question if refinement is necessary or not has to be answered. However, when using hp -fem, for each element with too big error, it has to be decided if h - or p -refinement has to be chosen. For an overview on different hp error estimators see [60]. In this thesis only the error estimator by Melenk and Wohlmuth [114], see algorithm 7 is used.

There, the error of each element K in the error estimator is estimated by η_K . If the error is low enough, the element remains unrefined. To estimate if an error is low enough, the mean value

$$\bar{\eta} := \frac{1}{\#\tau_h} \sum_{K \in \tau_h} \eta_K^2.$$

is used, where $\#\tau_h$ denotes the number of elements. If $\bar{\eta}$ is not below a given tolerance, a decision whether using h - or p -refinement has to be made. Therefore, the estimated error η_K is compared with the predicted error $\eta_K^{(prec)}$. For the predicted error $\eta_K^{(prec)}$ analyticity is assumed. A comparison between these two errors therefore yields an indirect statement on the local regularity of the solution. σ , γ_h , γ_p and γ_n are given parameters. As in the

Algorithm 7: hp -adaptive algorithm for refinement based on error estimators, see [114]

input : mesh τ_h
output: refined mesh τ_h

if $\eta_K^2 > \sigma \bar{\eta}^2$ **then**
 mark element for refinement
 if $\eta_K^2 > (\eta_K^{(prec)})^2$ **then**
 perform h -refinement
 set $(\eta_K^{(prec)})^2 := \frac{1}{4} \gamma_h \left(\frac{1}{2}\right)^{2p_K} \eta_K^2$ on each son element of K
 else
 perform p -refinement
 $p_K := p_K + 1$
 set $(\eta_K^{(pred)})^2 := \gamma_p \eta_K^2$
else
 no refinement
 set $(\eta_K^{(prec)})^2 := \gamma_n (\eta_K^{(prec)})^2$

numerical experiments in [114] these parameters are chosen to be

$$\sigma = 0.75, \quad \gamma_h = 4, \quad \gamma_p = 0.4, \quad \gamma_n = 1,$$

in numerical experiments in this thesis. Furthermore, in the initial triangulation $\eta_K^{(pred)} := 0$ is set to get an h -refinement in the first step. To enforce p -refinement $\eta_K^{(pred)} := \infty$ has to be

chosen.

Next, the question how fast the error decreases by applying the introduced refinement strategies, is answered.

3.2.7 Error estimates

In this subsection error estimates for uniform h -refinement, uniform p -refinement and bc-refinement are presented. For more literature on error estimates see e.g. [21, 140]. Error estimates can be yielded by investigating the approximation properties of the chosen discrete space. An application of Cea's Lemma, see theorem 3.5, then yields the error estimates. First, general estimates for h - and p -refinement are given.

Theorem 3.27. (see e.g. [21]) *Let τ_h be shape-regular with mesh size h , let the polynomial degree p on the whole mesh be constant and let $k \geq 1$. Furthermore, let the solution u^* be in $H^k(\Omega)$. Then it holds*

$$\|u^* - u_N^*\|_{H^1(\Omega)} \leq \tilde{c} h^{\mu-1} p^{-(k-1)} \|u^*\|_{H^k(\Omega)},$$

where $\mu = \min(k, p+1)$ and the constant $\tilde{c} > 0$ does not depend on h or p .

Moreover, there is an estimate in the $L_2(\Omega)$ -norm which can usually be yielded by duality arguments.

Theorem 3.28. (see e.g. [21]) *Let τ_h be shape-regular with mesh size h , the polynomial degree p is constant on the whole mesh and let $k \geq 2$. Furthermore, the solution u^* is in $H^k(\Omega)$. Then it holds*

$$\|u^* - u_N^*\|_{L_2(\Omega)} \leq \tilde{c} h^{\mu} p^{-k} \|u^*\|_{H^k(\Omega)},$$

where $\mu = \min(k, p+1)$ and the constant $\tilde{c} > 0$ does not depend on h or p .

Remark 3.29. *In the case of $\delta \in (0, 1)$ and a solution $u^* \in H^{1+\delta}(\Omega)$ the estimates*

$$\begin{aligned} \|u^* - u_N^*\|_{H^1(\Omega)} &\leq \tilde{c} h^{\delta} \|u\|_{H^{1+\delta}(\Omega)} \\ \|u^* - u_N^*\|_{L_2(\Omega)} &\leq \tilde{c} h^{2\delta} \|u\|_{H^{1+\delta}(\Omega)} \end{aligned}$$

can be yielded by using interpolation theory between Sobolev spaces, see [37, 157].

Therewith for uniform h -refinement with constant polynomial degree $p = 1$, under the assumption that $u^* \in H^2(\Omega)$, there holds

$$\|u^* - u_N^*\|_{H^1(\Omega)} \leq \tilde{c} h \|u\|_{H^2(\Omega)}, \quad (3.7)$$

$$\|u^* - u_N^*\|_{L_2(\Omega)} \leq \tilde{c} h^2 \|u\|_{H^2(\Omega)}. \quad (3.8)$$

However, for uniform p -refinement – assuming that $k > p+1$ and $u^* \in H^k(\Omega)$ – the error can be estimated by

$$\|u^* - u_N^*\|_{H^k(\Omega)} \leq \tilde{c} p^{-(k-1)} \|u^*\|_{H^k(\Omega)}, \quad (3.9)$$

$$\|u^* - u_N^*\|_{L_2(\Omega)} \leq \tilde{c} p^{-k} \|u^*\|_{H^k(\Omega)}. \quad (3.10)$$

Remark 3.30. *The given estimates also hold element wise. By using a sufficient hp -refinement and due to the fact that in general the solution $u^* \in H^k(\Omega)$ holds only for $k \leq 2$, better error estimates than (3.7) can be yielded, see e.g. [140].*

Next, error estimates for bc-fem are given. Therewith, some important results for estimating the number of degrees of freedom are recalled.

Theorem 3.31. *(see [98, Proposition 2.7]) Let τ_h be a geometric mesh with boundary mesh size h , and let \mathbf{p} denote the linear degree vector with slope $\zeta > 0$. Then, there exists a $\tilde{c} > 0$ depending on Ω , the shape-regularity constant γ and the constants in definition 3.25 and definition 3.26 such that*

$$\begin{aligned} \sum_{K \in \tau_h} 1 &\leq \tilde{c}h^{-1}, \\ \dim(V_{bc}(\Omega)) &\sim \sum_{K \in \tau_h} p_K^2 \leq \tilde{c}h^{-1}, \\ \max_{K \in \tau_h} p_K &\leq \tilde{c}|\log h|. \end{aligned}$$

Theorem 3.32. *(see [98]) Let τ_h be a geometric mesh with boundary mesh size h and \mathbf{p} a linear degree vector on τ_h with slope $\zeta > 0$. Furthermore, let $u^* \in H^{1+\delta}(\Omega)$, $\delta \in (0, 1)$, be the solution to (3.3), where the right-hand-side f is analytic in Ω . Then the finite element solution u_N^* given by (3.4) satisfies*

$$\|u^* - u_N^*\|_{H^1(\Omega)} \leq \tilde{c}h^\delta,$$

if the slope ζ is chosen sufficiently large.

These results show, that the number of degrees of freedom corresponds to a discretization with the boundary element method (see e.g. [133, 135]). That means $h \sim N^{-1}$ (see [98]). However, the bc-fem is applicable on a broader field, since the boundary element method can only be applied if the fundamental solution is known.

Now, having everything at hand to get an equation system, the question is, how to solve it. In this thesis, most equation systems are solved with iterative methods. Therewith suitable preconditioners are given in the next section.

3.3 Fast solvers

The system of algebraic equations (3.5) is solved iteratively. For a fast convergence of the iterative method, the condition number of the system matrix A is crucial (see chapter 1). Therefore the system is preconditioned, since choosing suitable preconditioners keep the iteration numbers low. Furthermore, this behaviour saves time to solve the equation system, see e.g. [112, 131, 171] for a general introduction.

In hp -fem a preconditioner is necessary, since the condition number of the mass matrix is p -dependent, whereas the condition number of the stiffness matrix is h and p dependent (see table 3.1 and [109]).

The preconditioners in this thesis are based on additive Schwarz methods (ASM). First, a general introduction on the Schwarz methods is given. Second, for two different Schwarz methods, the BPX, see [42] and for a special domain decomposition method, see [125], more

refinement	mass matrix	stiffness matrix
h -fem	$\mathcal{O}(1)$	$\mathcal{O}(h^{-2})$
p -fem	$\mathcal{O}(p^{2d})$	$\mathcal{O}(p^{2(d-1)})$

Table 3.1: condition number for basis functions in subsection 3.2.1 for dimension d , see [109] for p -fem and e.g. [76] for h -fem

details are presented. A general introduction to domain decomposition methods can e.g. be found in [45, 79, 80, 127, 147, 148, 149, 155]. The Schwarz method considers the problem:

$$\text{find } u_N \in V^{(N)} : a(u_N, v_N) = f(v_N) \quad \forall v_N \in V^{(N)}, \quad (3.11)$$

where u^* denotes the exact solution of (3.11). In this thesis, only the case of a bounded and elliptic bilinear form $a(\cdot, \cdot)$ is considered. In the Schwarz methods a decomposition in $n + 1$ subspaces such that

$$V^{(N)} = V_0 + V_1 + \dots + V_n \quad \text{with } N := \dim V \quad N_i := \dim V_i \quad (3.12)$$

is defined. Moreover, a mapping $\mathbf{T}_i : V^{(N)} \rightarrow V_i$ is defined by

$$a(\mathbf{T}_i u, v) = a(u, v) \quad \forall v \in V_i, \quad u \in V^{(N)} \quad (3.13)$$

and

$$\mathbf{T} = \mathbf{T}_0 + \mathbf{T}_1 + \dots + \mathbf{T}_n.$$

Then, the application of the additive Schwarz projector \mathbf{T} on the error $e_k = u_k - u^*$ leads to

$$\begin{aligned} w_k &= \mathbf{T}e_k \\ &= \mathbf{T}(u_k - u^*) \\ &= \mathbf{T}\mathbf{A}^{-1}\mathbf{A}(u_k - u^*) \\ &= \mathbf{T}\mathbf{A}^{-1}r_k, \end{aligned}$$

where \mathbf{A} denotes the corresponding operator to the matrix A . The next task is to get a matrix representation of the preconditioner $\mathbf{C}^{-1} = \mathbf{T}\mathbf{A}^{-1}$. Let $[\varphi_j]_{j=1}^N$ be a basis for V and $[\varphi_j^i]_{j=1}^{N_i}$ a basis for V_i . For a matrix representation of \mathbf{C}^{-1} the operators \mathbf{T}_i have to be represented in the standard basis. Since there exist matrices $W_i \in \mathbb{R}^{N \times N_i}$ such that

$$u_i = \mathbf{T}_i u = [\varphi_j]_{j=1}^N W_i \vec{u}_i \quad (3.14)$$

and

$$\begin{aligned} v_i &= [\varphi_j]_{j=1}^N W_i \vec{v}_i \\ u &= [\varphi_j]_{j=1}^N \vec{u} \end{aligned}$$

with (3.13) one yields

$$a([\varphi_j]_{j=1}^N W_i \vec{u}_i, [\varphi_j]_{j=1}^N W_i \vec{v}) = a([\varphi_j]_{j=1}^N \vec{u}, [\varphi_j]_{j=1}^N W_i \vec{v}) \quad \forall \vec{v} \in \mathbb{R}^{N_i}.$$

By writing the bilinear form as an application of a matrix and by equivalence relations, it follows

$$\begin{aligned}\bar{u}_i^\top W_i^\top A W_i \bar{v} &= \bar{u}^\top A W_i \bar{v} \quad \forall \bar{v} \in \mathbb{R}^{N_i} \\ W_i^\top A W_i \bar{u}_i &= W_i^\top A \bar{u} \\ \bar{u}_i &= \left(W_i^\top A W_i \right)^{-1} W_i^\top A \bar{u} \quad \forall \bar{v} \in \mathbb{R}^{N_i}.\end{aligned}$$

With (3.14), the matrix representation of \mathbf{T}_i in the standard basis is given by

$$[T_i] = W_i \left(W_i^\top A W_i \right)^{-1} W_i^\top A. \quad (3.15)$$

Then, a matrix representation of the preconditioner \mathbf{C}^{-1} can be derived by

$$\begin{aligned}\mathbf{C}^{-1} &= \mathbf{T} \mathbf{A}^{-1} \\ \mathbf{C}^{-1} &= \sum_{i=0}^n \mathbf{T}_i \mathbf{A}^{-1} \\ C^{-1} &= \sum_{i=0}^n W_i \left(W_i^\top A W_i \right)^{-1} W_i^\top A A^{-1}\end{aligned}$$

and it follows

$$C^{-1} = \sum_{i=0}^n W_i \left(W_i^\top A W_i \right)^{-1} W_i^\top. \quad (3.16)$$

The preconditioners used later on can be derived with this ansatz (see [42, 79, 80, 125]).

BPX. As first choice, a multilevel preconditioner by Bramble, Pasciak and Xu, see [42, 74, 169], called BPX, is given. Since the BPX is a preconditioner for h -fem, the concentration is on this case. In the BPX not only the information of the actual mesh, but also the meshes of coarser triangulations are used. For simplicity it is assumed to have a uniform mesh refinement with $p = 1$ everywhere. The coarsest mesh is denoted by $\tau_h^{(0)}$. The refinement leads to a sequence of meshes $\tau_h^{(l)}$ for $l = 0, \dots, L$. The space $V^{(N)}$ is then given by

$$V^{(N)} = V^{(L)} = \text{span} \{ \varphi_i^{(L)} \}_{i=1}^{N_L}$$

where $V^{(l)}$ denotes the space on level l and $V_i^{(l)} = \text{span} \{ \varphi_i^{(l)} \}$ the basis functions on level l for $l = 0, \dots, L$. This leads to the ASM-splitting

$$V^{(L)} = \sum_{l=0}^L \sum_{i=0}^{N_l} V_i^{(l)}. \quad (3.17)$$

A matrix representation of the BPX is given by

$$C_{BPX}^{-1} = \sum_{l=0}^L I_l^L D_l^{-1} I_l^L. \quad (3.18)$$

There, $D_l^{-1} = \text{diag}(A^{(l)})$, $I_l^{l+1} \in \mathbb{R}^{N_{l+1} \times N_l}$ is the finite element interpolation matrix. $I_{l+1}^l = (I_l^{l+1})^\top$ is the finite element restriction matrix and

$$I_j^l := I_{l-1}^l I_{l-2}^{l-1} \cdot \dots \cdot I_j^{j+1} \quad j < l. \quad (3.19)$$

The appearance of the interpolation matrix depends on the elements. The situation is demonstrated for the case of a refined edge.

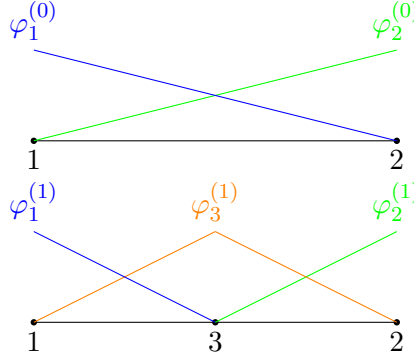


Figure 3.6: basis functions in level 0 and level 1

Figure 3.6, indicates, that it has to hold

$$\begin{aligned} \varphi_1^{(0)} &= \varphi_1^{(1)} + \frac{1}{2}\varphi_3^{(1)} \\ \varphi_2^{(0)} &= \varphi_2^{(1)} + \frac{1}{2}\varphi_3^{(1)}. \end{aligned}$$

Therewith, the finite element interpolation matrix is given by

$$I_0^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}.$$

Next, the case for a square element, see Figure 3.7 for the numbering, is considered.

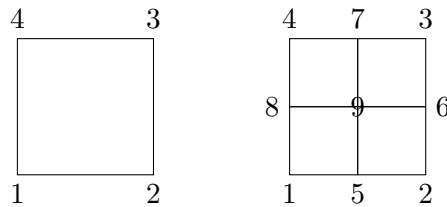


Figure 3.7: Numbering of nodes

Since the nodes 5, ..., 8 are sons of edges, the construction of its interpolation is analogue to the case above. However, the situation for node 9 is different, since node 9 has four fathers, the nodes 1, 2, 3, 4. This has to be considered when setting up the interpolation matrix, which

is in this case given by

$$I_0^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

Therewith, the preconditioner matrix (3.18) can be calculated.

Remark 3.33. *For an efficient evaluation the multiplications of $I_l^L \vec{w}$ and $I_L^l \vec{w}$ are implemented by using (3.19).*

Finally, a result on the condition number is given.

Theorem 3.34. [169] *Let the ASM splitting (3.17) and continuous, piecewise linear elements for the triangulations $\tau_h^{(l)}$ be used. Then, the condition number of the preconditioned system can be estimated by*

$$\kappa(\mathbf{T}) \leq \tilde{c},$$

with a constant $\tilde{c} > 0$ independent of the mesh size h (and the number of levels L).

Remark 3.35. *Beside additive Schwarz methods also multigrid methods, see e.g. [22, 81, 94, 111] can be used for preconditioning the h -part.*

Pavarino preconditioner. Next, a preconditioner for the p -version of fem is given. It was introduced by Pavarino in [125] for tensor product meshes, see also [136, 137] for the triangular and tetrahedral case. This paragraph recalls the most important aspects of this preconditioner. Let Q_p be the set of polynomials of degree less or equal than p in each variable, i.e.

$$Q_p := \{x_1^i x_2^j : 1 \leq i, j \leq p\}.$$

The elliptic boundary value problem (3.11) (it is assumed to have homogeneous Dirichlet boundary conditions) is discretized with continuous, piecewise finite polynomial elements of degree p . Therewith, the space V^p is given by

$$V^p = \{v \in H_0^1(\Omega) : v|_K \circ F_K \in Q_{pK}, i = 1, \dots, N\}.$$

The discrete boundary value problem takes the form

$$\text{find } u \in V_{\mathcal{D}}^p : \quad a(u, v) = F(v) \quad \forall v \in V_{\mathcal{D}}^p$$

with $V_{\mathcal{D}}^p := \{v \in V^p : v = 0 \text{ on } \Gamma_{\mathcal{D}}\}$. Furthermore, let N be the number of interior nodes. According to (3.12), the additive Schwarz method splitting

$$V_{\mathcal{D}}^p = V_0^p + V_{\nu_1}^p + \dots + V_{\nu_n}^p \tag{3.20}$$

is used. There, the first space V_0^p serves as coarse space. In fact the space V_0^p is chosen to be the space of continuous and piecewise linear functions of the mesh, i.e. $V_0^p = V_{\mathcal{D}}^1$. The spaces $V_{\nu_i}^p$ are defined by

$$V_{\nu_i}^p = V^p \cap H_0^1(\Omega_{\nu_i}),$$

where Ω_{ν_i} denotes the vertex patch of the i -th node, i.e.

$$\Omega_{\nu_i} = \{\cup \bar{K} \in \tau_h : \nu_i \in \bar{K}\}.$$

A vertex patch is specified in Figure 3.8. Furthermore, the index set of Ω_i given by

$$J(\nu_i) = [j_1^{\nu_i}, \dots, j_{n_{\nu_i}}^{\nu_i}]$$

containing all basis functions which live on $\text{supp}(\varphi_j) \subset \Omega_{\nu_i}$, i.e. all basis functions living completely on Ω_{ν_i} . That means in fact that there are (homogeneous) Dirichlet boundary conditions for each patch where ν_i is not a node on the boundary.

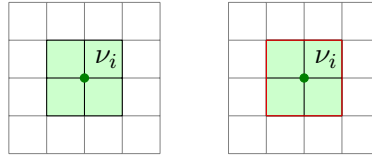


Figure 3.8: Patch Ω_{ν_i} (green) for a node ν_i and the domain where the basis functions $J(\nu_i)$ lives (green), respectively.

Since the preconditioner by Pavarino is later applied to problems with homogeneous Neumann boundary conditions, it has to be extended. According to [125] for Neumann boundary conditions the corresponding subspaces have to be included in order to keep a constant condition number.

Then, the summation in (3.20) runs over all non Dirichlet nodes, see figure 3.9.

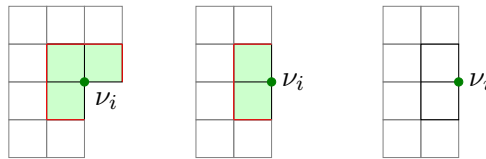


Figure 3.9: domain which is spanned by the basis function $J(\nu_i)$ for a Neumann node in an L-shape (left), a Neumann node (middle) and a Dirichlet node (right), where all corresponding domains are marked green

Theorem 3.36. [125, Theorem 1] *The operator \mathbf{T} of the additive splitting (3.20) defined by the spaces $V_{\nu_i}^p$ satisfies the estimate*

$$\kappa(\mathbf{T}) \leq \tilde{c},$$

where $\tilde{c} > 0$ is independent of the polynomial degree p , the number of subspaces n and the minimal mesh size h .

To enable the rewriting of the preconditioner by Pavarino in matrix notation later on, some additional notation is introduced. For a given mesh τ_h it holds

$$\begin{aligned}\mathcal{V} &:= \{\nu_i \text{ is a node of } \tau_h\} \\ \mathcal{V}_N &:= \{\nu_i \in \mathcal{V} \text{ and without Dirichlet boundary conditions on } \nu_i\} \\ \mathcal{V}_C &:= \mathcal{V} \setminus \mathcal{V}_N.\end{aligned}$$

Moreover, the fe restriction matrix $P_{\nu_i} \in \mathbb{R}^{n \times n_{\nu_i}}$ is defined by

$$(P_{\nu_i})_{lj} = \begin{cases} 1 & \text{if } l = j = j'_k, 1 \leq k \leq n_{\nu_i} \\ 0 & \text{else} \end{cases}$$

i.e. with the help of this map one can extend a mapping from Ω_{ν_i} to Ω or restrict it from Ω to Ω_{ν_i} with $P_{\nu_i}^\top$.

3.3.1 hp -preconditioners

All given preconditioners are based on the preconditioner by Pavarino and a suitable choice of the space V_0^p , where the first two choices for the space V_0^p were already proposed by Pavarino in [125].

First, a hp -preconditioner based on the additive Schwarz splitting (3.20), where V_0^p is the space of continuous and (bi-)linear basis functions is considered. This gives

$$C_P^{-1} = P_0^\top (A_{p=1})^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top A_{\nu_i}^{-1} P_{\nu_i}, \quad (3.21)$$

where A_{ν_i} denotes the stiffness matrix on the patch to node ν_i and P_0 the fe restriction matrix onto the mesh with polynomial degree $p = 1$ everywhere. $A_{p=1}$ denotes the assembled stiffness matrix for polynomial degree $p = 1$ on the whole mesh.

Remark 3.37. *Usually, the inverse matrix of $A_{p=1}$ is not calculated directly, only the action of $A_{p=1}^{-1} \vec{r}$ is available. A suitable method therefore is, for example, sparse LU decomposition.*

Theorem 3.38. *(see e.g. [125]) The preconditioner C_P^{-1} given in (3.21) satisfies*

$$\kappa(C_P^{-1} A) \leq \tilde{c}.$$

The constant $\tilde{c} > 0$ is independent of h and p , i.e. the condition number is bounded by a constant for uniform refinement. The costs for applying $\vec{w} = C_P^{-1} \vec{r}$ are $\mathcal{O}(N^2)$ for $d = 2$ and $\mathcal{O}(N^{7/3})$ for $d = 3$.

Remark 3.39. *The high costs are caused by the fact, that in each application of $C_P^{-1} \vec{r}$ the equation system*

$$A_{p=1} \vec{u} = \vec{r}$$

has to be solved. In some cases, the costs can be reduced by nested dissection, see e.g. [70].

Remark 3.40. *In the case of a bc-refinement, the costs for applying the preconditioner can be reduced to $\mathcal{O}(N \log^8 N)$, see [97] for $d = 2$. However, in $d = 3$ it becomes too expensive.*

In the sense of additive Schwarz methods with inexact subproblem solvers [79] the application of $A_{p=1}^{-1}\vec{r}$ can be replaced by a multilevel preconditioner, e.g. multigrid or BPX. Here, the second choice is used, which leads to the preconditioner

$$C_{BPXP}^{-1} = P_0^\top C_{p=1, BPX}^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top A_{\nu_i}^{-1} P_{\nu_i}, \quad (3.22)$$

where $C_{p=1, BPX}^{-1}$ denotes an application of the BPX on the stiffness matrix A for constant polynomial degree $p = 1$ on the whole mesh. Next, two cases, the case of moderate polynomial degrees and very high ones are considered in the next two theorems.

Theorem 3.41. (see [63]) Let C_{BPXP}^{-1} be defined by (3.22). Then,

$$\kappa(C_{BPXP}^{-1}A) \leq \tilde{c},$$

where the constant $\tilde{c} > 0$ is independent of the mesh size h and the polynomial degree p . In the case of bc-fem, the costs for applying $\vec{w} = C_{BPXP}^{-1}\vec{r}$ are $\mathcal{O}(N)$.

In the case of high polynomial degrees, i.e. polynomial degrees above ten, the application of $A_{\nu_i}^{-1}\vec{r}$ by sparse LU is too expensive (see e.g. [29]). Therefore, $A_{\nu_i}^{-1}$ is replaced by a suitable preconditioner. Then, a suitable hp -preconditioner is given by

$$C_{BPXP_2}^{-1} = P_0^\top C_{p=1, BPX}^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top C_{\nu_i}^{-1} P_{\nu_i}, \quad (3.23)$$

where $C_{\nu_i}^{-1}$ is a suitable preconditioner for the matrix on the node patch ν_i .

Theorem 3.42. (see [29]) Let $C_{\nu_i}^{-1}$ be the preconditioner using wavelet methods of [29]. Then, for the condition number it holds

$$\kappa(C_{BPXP_2}^{-1}A) \leq \tilde{c}(\log(p) \log^\chi(\log(p)))^3$$

for any $\chi > 1$ with a constant $\tilde{c} > 0$. The action $\vec{w} = C_{BPXP_2}^{-1}\vec{r}$ requires then $\mathcal{O}(N)$ operations.

Remark 3.43. In the case of bc-refinement a similar preconditioner but based on the structure of the mesh, see [63], can be applied for the stiffness matrix.

A similar construction as (3.22) can be done for the mass matrix, see [35]. Since the mass matrix for h -fem can efficiently be preconditioned by its diagonal, a suitable preconditioner is given by

$$C_M^{-1} = P_0^\top (\text{diag}(M)_{p=1})^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top M_{\nu_i}^{-1} P_{\nu_i}. \quad (3.24)$$

It has to be stated that this choice is only a good preconditioner for the mass matrix but not for the stiffness matrix, since the mass matrix is well conditioned for fixed polynomial degree p . There holds:

Theorem 3.44. ([35]) Let M be the mass matrix and C_M^{-1} defined by (3.24), then the condition number

$$\kappa(C_M^{-1}M) \leq \tilde{c},$$

is constant with $\tilde{c} > 0$ independent of h and p . The work for applying $\vec{w} = C_M^{-1}\vec{r}$ is $\mathcal{O}(N)$.

Remark 3.45. The matrices M and C_M are even spectrally equivalent, i.e. it holds

$$M \sim C_M.$$

As already stated, in this thesis quadrilateral elements with hanging nodes are used. For general results on handling hanging nodes, see e.g. [3, 116, 150, 161].

3.3.2 Extension to hanging nodes

In all the cases considered in this section, the projector P introduced in section 3.2.5 is applied in order to get a conform solution. To simplify the considerations, the preconditioners are separated in its different parts.

In the case of the diagonal as preconditioner or the inversion for $p = 1$ with sparse LU, the projector is applied according to [3]. There, the global projectors P and P^\top introduced in subsection 3.2.5 are used. This leads to

$$\begin{aligned} C_{\text{diag},p=1}^{-1} &= P^\top P_0^\top (\text{diag}(A))^{-1} P_0 P \\ C_{p=1}^{-1} &= P^\top P_0^\top (A_{p=1})^{-1} P_0 P. \end{aligned}$$

In the case of BPX it is assumed to have only linear, piecewise finite elements, i.e. polynomial degree $p = 1$ on the whole mesh. Following [116] in order to preserve the multilevel structure, the best strategy is to consider all nodes, including all hanging ones, as (real) degrees of freedom. In order to get a conforming solution at the end, the preconditioner is chosen as

$$C_{BPX}^{-1} = P^\top \left(\sum_{l=0}^L I_l^L D_l^{-1} I_l^L \right) P. \quad (3.25)$$

In the case of the preconditioner of Pavarino, the situation is a bit more complicated. The part for $p = 1$ is already considered in (3.25). Applying the projector for the patches in the same way as in the BPX, i.e. treating each entry of hanging nodes and edges, lead to problems when inverting the patch matrix. The reason therefore is, that in this case not all edges which are necessary to enforce the conformity conditions appear in the vertex patch. Therewith, the preconditioner of Pavarino is applied to the conform mesh, i.e. only patches of regular nodes occur. Figure 3.10 shows the construction of patches in order to get patch matrices which enable the enforcing of conformity conditions. Furthermore, this construction preserves the partition of unity that is necessary in the proof of the upper eigenvalue [125, Theorem 1].

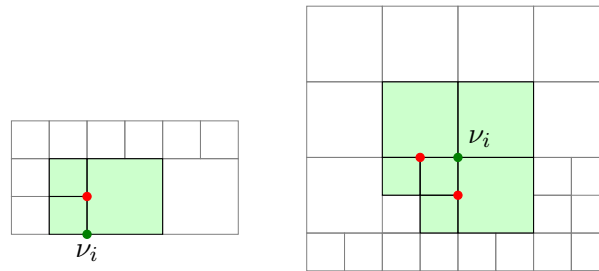


Figure 3.10: patches for nodes in green and hanging nodes are plotted in red

Since in each patch, only some additional hanging nodes appear (see Figure 3.10) which are removed by enforcing the conformity on the patch, the partition of unity is kept. Furthermore, since for each coarse edge their son edges are also contained in the plot, it is possible to enforce the conformity on the patch. For notation issues, the set

$$\mathcal{V}_H := \{ \nu_i \in \mathcal{V}_C : \nu_i \text{ no hanging node} \}$$

is introduced. Therewith, the preconditioner is given

$$C_{PP}^{-1} = P^\top \left(\sum_{\nu_i \in \mathcal{V}_H} P_{\nu_i}^\top A_{\nu_i}^{-1} P_{\nu_i} \right) P,$$

where P_{ν_i} and $P_{\nu_i}^\top$ enable the mapping on the conform patch space and $A_{\nu_i}^{-1}$ is the conform patch matrix, inverted by sparse LU.

A combination of the different parts of the preconditioners gives the three mainly considered preconditioners in this thesis.

3.4 Numerical experiments

In this section the behaviour of the preconditioners is investigated since the introduced preconditioners are later on applied to more complex problems, see subsection 4.4.4 and section 5.5. The goal now is to show the h - and p -independent behaviour of the introduced preconditioners. Therewith the condition number, given by

$$\kappa(C^{-1}A) = \frac{\lambda_{\max}(C^{-1}A)}{\lambda_{\min}(C^{-1}A)}$$

has to be constant. The approximation of the minimal and maximal eigenvalues is done by the inverse vectoriteration and the vectoriteration, respectively (see e.g. [72]).

For the numerical experiments, the model problem

$$\begin{aligned} -\Delta u(x) + u(x) &= f(x) && \text{in } \Omega \\ \frac{\partial u}{\partial n}(x) &= 0 && \text{on } \partial\Omega \end{aligned} \quad (3.26)$$

in the unit square $\Omega = (-1, 1)^2$ is considered. The right-hand-side is chosen, such that the solution is

$$u(x) = e^{x_1^3/3-x_1} e^{x_2^3/3-x_2}.$$

The mass matrix M_N is given by the entries

$$(M_N)_{ji} = \int_{\Omega} \varphi_i(x) \varphi_j(x) \, dx \quad i, j = 1, \dots, N$$

and the stiffness matrix by

$$(K_N)_{ji} = \int_{\Omega} \nabla \varphi_i(x) \cdot \nabla \varphi_j(x) \, dx + \int_{\Omega} \varphi_i(x) \varphi_j(x) \, dx \quad i, j = 1, \dots, N.$$

The entries of the right-hand-side \vec{f} are given by

$$f_i = \int_{\Omega} f(x) \varphi_i(x) \, dx.$$

In order to obtain the iteration numbers for the stiffness matrix, the model problem (3.26) is solved. That means the solution to the (preconditioned) equation system

$$C^{-1}K_N \vec{u} = C^{-1} \vec{f}, \quad (3.27)$$

for several refinements and suitable preconditioners C^{-1} is calculated with the preconditioned CG method. In fact the BPX preconditioner and the hp -preconditioners (3.22) and (3.21) – all introduced in section 3.3 – are applied.

To get iteration numbers in case of applying the preconditioner C_M^{-1} given in (3.24), the equation system

$$C_M^{-1}M_N\vec{y} = C_M^{-1}\vec{f}, \quad (3.28)$$

is solved.

First, the BPX preconditioner C_{BPX}^{-1} is considered. Its minimal and maximal eigenvalue for uniform h -refinement and bc-refinement with constant polynomial degree $p = 1$ are given in table 3.11.

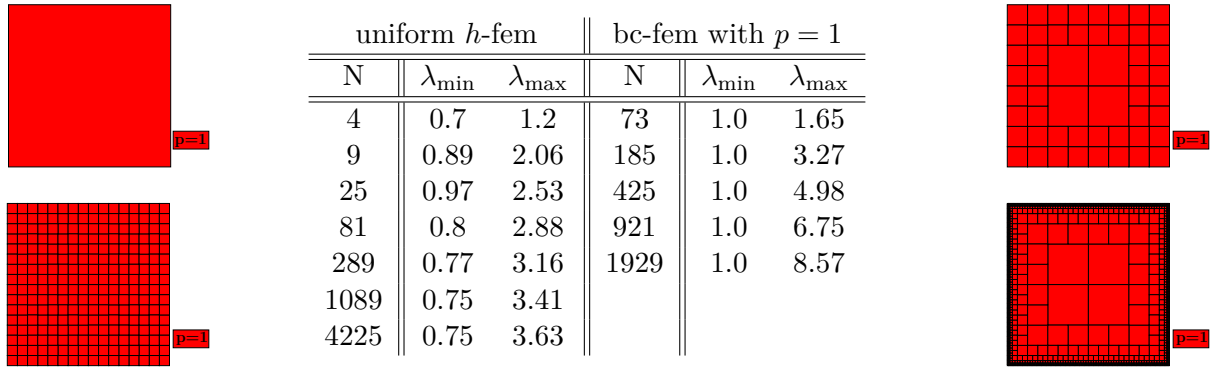


Figure 3.11: eigenvalues for $C_{BPX}^{-1}K_N$

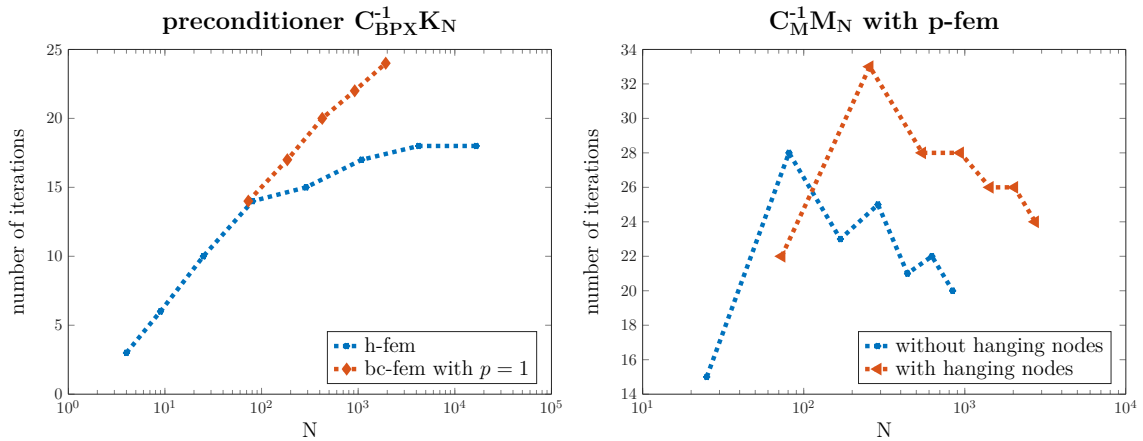
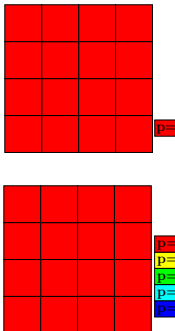


Figure 3.12: iteration numbers versus degrees of freedom for $C_{BPX}^{-1}K_N$ and $C_M^{-1}M_N$

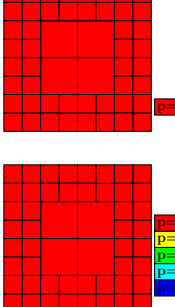
The iteration numbers for solving the model problem for both refinements are given in figure 3.12. There, a relative stopping criterion with accuracy 10^{-8} is used. The results show, that the condition number for the BPX is constant in the case of uniform h -refinement. However, this is not true if hanging nodes appear. This can especially be observed in the increasing iteration numbers for bc-refinement with constant polynomial degree $p = 1$ in figure 3.12 and the increasing maximal eigenvalue in table 3.11.

nrDofs	$C_M^{-1}M_N$		$C_P^{-1}K_N$		$C_{BPXP}^{-1}K_N$	
	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}
25	0.50	4.5	1.1	2.47	0.18	2.94
81	0.47	5.84	0.95	4.01	0.18	4.0
169	0.58	5.94	1.1	4.0	0.18	4.0
289	0.65	6.09	0.99	4.0	0.18	4.0
441	0.70	6.12	1.0	4.0	0.18	4.0
625	0.74	6.17	1.0	4.0	0.18	4.0
841	0.77	6.18	1.0	4.0	0.18	4.0


Figure 3.13: eigenvalues for uniform p -refinement

Next, the minimal and maximal eigenvalue for different hp -preconditioners are approximated, see table 3.13 for uniform p -fem. To investigate the influence of hanging nodes, in table 3.14 as starting mesh a bc-mesh with constant polynomial degree $p = 1$ is used and uniform p -fem applied. Both tables show the expected results, i.e. constant eigenvalues (or at least eigenvalues which can be bounded by a constant).

N	$C_M^{-1}M_N$		$C_P^{-1}K_N$		$C_{BPXP}^{-1}K_N$	
	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}
73	1.0	4.72	1.0	2.57	1.0	3.2
257	1.0	6.06	1.0	4.0	1.0	4.0
545	1.0	6.19	1.0	4.0	1.0	4.0
937	1.0	6.34	1.0	4.0	1.0	4.0
1433	1.0	6.37	1.0	4.0	1.0	4.0


Figure 3.14: eigenvalues for uniform p -refinement with bc-fem starting mesh

The iteration numbers for applying the preconditioners C_P^{-1} and C_{BPXP}^{-1} to the equation system (3.27) are given in figure 3.15. In case of the mass matrix, the iteration numbers for solving the equation system (3.28) can be found in figure 3.12. In all cases the preconditioned CG with relative termination condition and accuracy of 10^{-10} is used. It has to be stated, that the numerical experiments confirm the theoretical results.

Moreover, in figure 3.15 the iteration numbers with respect to the number of degrees of freedom for different preconditioners and bc-refinement (with increasing polynomial degree) are given (see figure 3.16). These results show that the iteration numbers (obtained with an relative termination condition of 10^{-10}) for the mass matrix preconditioner C_M^{-1} and the stiffness matrix preconditioner C_P^{-1} lead both to constant iteration numbers. In the case of C_{BPXP}^{-1} the situation is different, since both, the extremal eigenvalues and the iteration numbers seems to grow logarithmically. The reason for this behaviour is, that the eigenvalues and therefore the iteration numbers for bc-refinement with constant polynomial degree are not constant due to the hanging nodes, see table 3.11 and figure 3.12.

Remark 3.46. *By using suitable hp -triangulations without hanging nodes, i.e. triangular ele-*

ments with red-green refinement, a constant condition number and therefore constant iteration numbers are yielded.

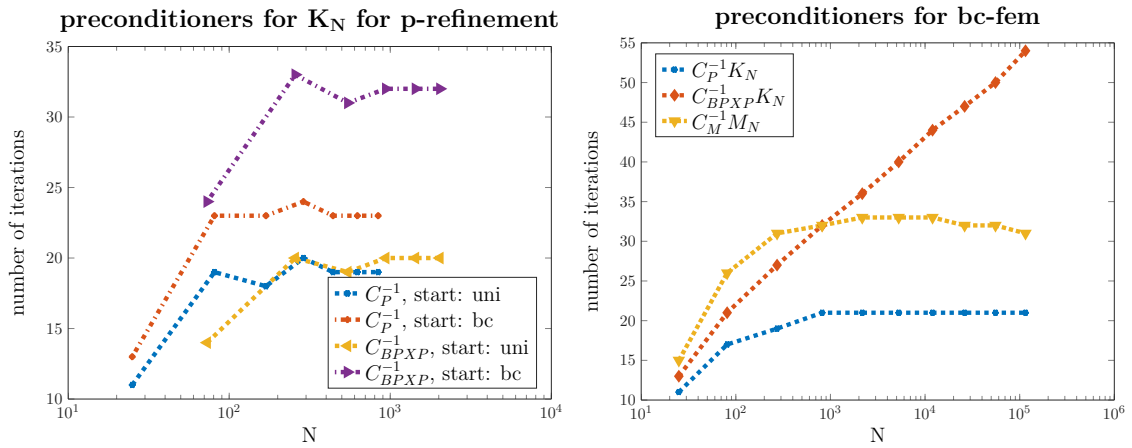


Figure 3.15: number of degrees of freedom versus iteration numbers for different preconditioners.

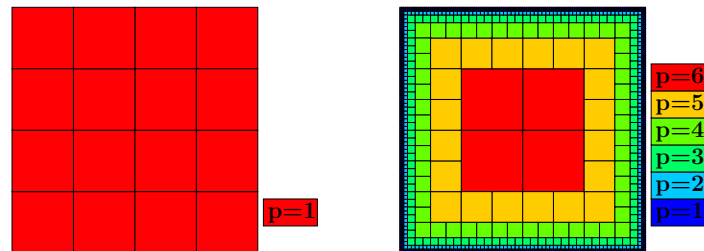


Figure 3.16: starting mesh and mesh after five refinements for bc-fem

4 Optimal control problems with semismooth Newton

In this chapter the optimal control problems of section 2 are discretized by using the variational discretization concept by Hinze [88] and hp -fem. For solving the equation system, a semismooth Newton method (see e.g. [160]) has been chosen.

First, a boundary control problem is discretized with bc-fem as in [31]. Second, a suitable hp -refinement for a distributed optimal control problem is presented and furthermore, solved with a semismooth Newton method. To the best knowledge of the author this is new, since in general, such problems are solved with h -fem, see e.g. [6, 88, 89, 91]. There are even extensions to solve such problems with higher but fixed polynomial degree, see [10, 50, 142]. Only in [164] hp -fem is used to solve a distributed optimal control problem. However, in that case interior point methods are applied to solve the problem, which is different from the approach presented in here. In this thesis, the focus is on choosing suitable hp -refinement strategies and setting up the matrices on inactive and active sets of the domain for applying the semismooth Newton method. This is investigated in section 4.4. First, a general introduction in solving optimal control problems and a summary of known results is given.

4.1 Discretization

When solving optimal control problems, there are two different approaches ([158]), the *First optimize, then discretize* ansatz and the *First discretize, then optimize* ansatz. In this thesis, the second one is used. The conventional ansatz in *First discretize, then optimize*, is to discretize all three variables, the state, the adjoint and the control by finite elements. However, in the discretization by *variational discretization* due to Hinze [88], only the state and the adjoint are discretized by finite elements, whereas the control is discretized implicitly via the projection formula. This has the advantage, that the error of the control can be estimated by the error of the adjoint, which is not possible in the case of the conventional ansatz. For a comparison between the conventional ansatz and variational discretization see [91]. Before investigating the advantages and problems appearing in the case of variational discretization, the existence and uniqueness of the discrete system is considered.

Theorem 4.1. (see e.g. [158]) *Let assumption 2.1 be satisfied and let (y^*, u^*, q^*) be the exact solution of (2.1). The discrete system of (2.1) is given by*

$$\begin{aligned} Ay_N^* &= Bu_N^* + f \\ A^* q_N^* &= y_N^* - y_d \\ u_N^* &= P_{U_{ad}} \left(-\frac{1}{\alpha} B^* q_N^* \right) \end{aligned} \tag{4.1}$$

possesses the discrete unique solution (y_N^*, u_N^*, q_N^*) with the discrete projection

$$u_N^* = \max \left\{ u_a(x), \min \left\{ -\frac{1}{\alpha} B^* q_N^*, u_b(x) \right\} \right\}.$$

Remark 4.2. The projection formula in (4.1) is equivalent to the variational formulation

$$\langle B^* q_N^* + \alpha u_N^*, u - u_N^* \rangle_U \geq 0 \quad \forall u \in U_{ad},$$

see chapter 2.

Remark 4.3. In variational discretization only the state and the adjoint are discretized. Therefore y_N^* and q_N^* are finite element functions. The control u_N^* is discretized implicitly via the projection formulation and is in general not a finite element function. Therewith u_N^* may have kinks that are not along the mesh elements.

Next, a general error estimate is given. Let y^N and q^N be the finite element solutions to the primal and dual equation, that is

$$a(y^N, v_N) = \langle f, v_N \rangle_\Omega + \langle B u^*, v_N \rangle_U \quad (4.2)$$

$$a(v_N, q^N) = \langle y^* - y_d, v_N \rangle_\Omega. \quad (4.3)$$

The equation system (4.1) implies

$$a(y_N^*, v_N) = \langle f, v_N \rangle_\Omega + \langle B u_N^*, v_N \rangle_U. \quad (4.4)$$

$$a(v_N, q_N^*) = \langle y_N^* - y_d, v_N \rangle_\Omega. \quad (4.5)$$

The continuous and discrete variational inequality is given by

$$\langle B^* q^* + \alpha u^*, u - u^* \rangle_U \geq 0 \quad \forall u \in U_{ad} \quad (4.6)$$

$$\langle B^* q_N^* + \alpha u_N^*, u - u_N^* \rangle_U \geq 0 \quad \forall u \in U_{ad}, \quad (4.7)$$

respectively. With simple arguments, an error estimate of the variational discretization is given.

Theorem 4.4. ([89]) For the solutions (y^*, u^*, q^*) to (2.1) and its discretized version (4.1) with the solution (y_N^*, u_N^*, q_N^*) , there holds

$$\alpha \|u_N^* - u^*\|_{L_2(U)}^2 + \|y^* - y_N^*\|_{L_2(\Omega)}^2 \leq \frac{1}{\alpha} \|q^* - q^N\|_{L_2(U)}^2 + \|y^* - y_N^*\|_{L_2(\Omega)}^2.$$

Proof. The proof is given in [89] and is repeated here. By using $u = u_N^*$ and $u = u^*$ in (4.6) and (4.7), respectively, adding both terms yields

$$\langle B^*(q_N^* - q^*) + \alpha(u_N^* - u^*), u^* - u_N^* \rangle_U \geq 0.$$

This is equivalent to

$$\alpha \|u^* - u_N^*\|_{L_2(\Omega)}^2 \leq \langle B(u^* - u_N^*), q_N^* - q^N \rangle_U + \langle B(u^* - u_N^*), q^N - q^* \rangle_U. \quad (4.8)$$

The first term in (4.8) can be rewritten with (4.2) and (4.4), the second one can be estimated by using Cauchy Schwarz, which gives

$$\begin{aligned} \alpha \|u^* - u_N^*\|_{L_2(\Omega)}^2 &\leq a(y^N - y_N^*, q_N^* - q^N) + \langle u^* - u_N^*, B^*(q^N - q^*) \rangle_U \\ &\leq a(y^N - y_N^*, q_N^* - q^N) + \|u^* - u_N^*\|_{L_2(U)} \|B^*(q^N - q^*)\|_{L_2(U)}. \end{aligned} \quad (4.9)$$

Then, both terms are estimated separately. For the first term in (4.9) it holds

$$\begin{aligned} a(y^N - y_N^*, q_N^* - q^N) &\stackrel{(4.3),(4.5)}{=} \langle y_N^* - y^*, y^N - y_N^* \rangle_\Omega \\ &= \int_\Omega (y^* - y_N^*)(y_N^* - y^N) \, dx \\ &\leq -\frac{1}{2} \|y^* - y_N^*\|_{L_2(\Omega)}^2 + \|y^* - y^N\|_{L_2(\Omega)}^2, \end{aligned} \quad (4.10)$$

where in the last step the inequality

$$2(a-b)(b-c) \leq -(a-b)^2 + (a-c)^2$$

is applied. For the second term in (4.9) Young's inequality

$$ab \leq \frac{1}{2\varepsilon} a^2 + \frac{\varepsilon}{2} b^2$$

is used, which yields

$$\|u^* - u_N^*\|_{L_2(U)} \|q^N - q^*\|_{L_2(U)} \leq \frac{\alpha}{2} \|u^* - u_N^*\|_{L_2(U)}^2 + \frac{1}{2\alpha} \|q^N - q^*\|_{L_2(U)}^2 \quad (4.11)$$

since B^* is either the trace operator or identity. By inserting (4.10) and (4.11) in (4.9) and a multiplication by two, the desired estimate

$$\alpha \|u^* - u_N^*\|_{L_2(U)}^2 + \|y^* - y_N^*\|_{L_2(\Omega)}^2 \leq \frac{1}{\alpha} \|q^* - q^N\|_{L_2(U)}^2 + \|y^* - y^N\|_{L_2(\Omega)}^2$$

is yielded. □

Remark 4.5. For uniform h -fem and by using the error estimates in subsection 3.2.7 a comparison (see [91]) of a piecewise constant control u_N^* , a continuous and piecewise linear control u_N^* and variational discretization with piecewise linear adjoint q_N^* yields

$$\alpha \|u^* - u_N^*\|_U + \|y^* - y_N^*\|_{L_2(\Omega)} \leq \begin{cases} ch & \text{for piecewise constant } u_N^* \\ ch^{3/2} & \text{for continuous and piecewise linear } u_N^* \\ ch^2 & \text{for variational discretization} \end{cases}$$

This comparison shows the main advantage of variational discretization, i.e. a higher convergence rate for the control u^* , which follows by theorem 4.4. However, it has to be mentioned, that for piecewise constant control u_N^* suitable postprocessing [117, 129] yields the same results as variational control.

Better error estimates are possible for special discretizations, see section 4.3 and section 4.4 for more detailed results. Next, a method to solve the problem, is presented. A short overview and further references on solving optimal control problems in general are given in [158]. Possible choices are gradient-based methods, the primal-dual active set method or Newton like methods.

4.2 Semismooth Newton method

In this chapter, the optimal control problems are solved with the semismooth Newton method. An introduction into it can be found in [160]. The semismooth Newton method is locally super-linear convergent (see [160]) and under slightly stronger assumptions, the convergence rate is $\rho > 1$. Although the solution to an inequality constrained problem has to be calculated, only one linear equation has to be solved per iteration. Therefore, the costs per iteration are comparable to the Newton method for smooth operators. The semismooth Newton method can furthermore be interpreted as primal-dual active set method, see [85]. For a step-to-step application of the semismooth Newton method see e.g. [92, 160], further information is given in [32, 86, 89, 159]. In here only the most important basics are pointed out.

For solving the discrete optimal control problem (4.1) eliminating y_N^* from the first equation, eliminating q_N^* from the second one and inserting it in the last one yields

$$u = P_{U_{ad}} \left(-\frac{1}{\alpha} B^* \left((A^*)^{-1} (A^{-1} B u + f - y_d) \right) \right).$$

Therewith, the semismooth Newton algorithm is applied to the equation

$$G(u) := u - P_{U_{ad}} \left(-\frac{1}{\alpha} B^* q_N(u) \right) = 0 \quad \text{in } U$$

for given $u \in U$. The discrete state $y_N(u)$ and the adjoint state q_N have to fulfill the primal (4.4) and dual problem (4.5). Due to the projection formula

$$u_N^* = P_{U_{ad}} \left(-\frac{1}{\alpha} B^* q_N^* \right) \tag{4.12}$$

this setting admits the unique solution $u_N^* \in U_{ad}$. According to [89, 160] the mapping $G : L_2(U) \rightarrow L_2(U)$ is semismooth in the sense, that

$$\sup_{\mathfrak{M} \in \partial G(u+s)} \|G(u+s) - G(u) - \mathfrak{M}s\|_{L_2(U)} = \mathcal{O}(\|s\|_{L_2(U)}) \quad \text{as } \|s\|_{L_2(U)} \rightarrow 0.$$

The generalized differential is there given by

$$\partial G(u) := \left\{ I + D(u) \left(\frac{1}{\alpha} B^* q'_N(u) \right) \right\}$$

with

$$D(u)(x) = \begin{cases} 0 & \text{if } -\frac{1}{\alpha} B^* q_N(u)(x) \geq u_b \\ 1 & \text{if } -\frac{1}{\alpha} B^* q_N(u)(x) \in (u_a, u_b) \\ 0 & \text{if } -\frac{1}{\alpha} B^* q_N(u)(x) \leq u_a \end{cases}.$$

Let $g \equiv g(u)$ denote the indicator function of the inactive set

$$\mathfrak{I}(u) := \{x \in U : -\frac{1}{\alpha} B^* q_N(u)(x) \in (u_a, u_b)\}$$

and with

$$q'_N(u) = (A^*)^{-1} A^{-1} B$$

it is set

$$G'(u) := I + \frac{1}{\alpha} g(u) B^* (A^*)^{-1} A^{-1} B \in \partial G(u).$$

Due to [86] $G'(u)$ is bounded invertible. Furthermore, it can be shown that mesh independence holds under appropriate conditions, see [92]. Mesh independence in here means that some kind of convergence also holds for the behaviour of the discrete algorithm towards the behaviour of the algorithm in the infinite-dimensional space.

The semismooth Newton method is given by algorithm 8.

Algorithm 8: Semismooth Newton algorithm, see [89]

input : starting value $u \in U$

output: iterate $u^{new} \in U$

while $G(u) \neq 0$ **do**

 solve $G'(u)u^{new} = G'(u)u - G(u)$ for u^{new}
 set $u = u^{new}$.

Remark 4.6. *Algorithm 8 is given in the infinite-dimensional space U . Nevertheless, it can be shown that it is numerically implementable, see [88].*

Remark 4.7. *In general not the given but a modified termination condition is used in algorithm 8, see e.g. [160].*

Further remarks and drawbacks when implementing the algorithm are given in section 4.3 and section 4.4.

Next, the two model problems are considered separately. First, a suitable hp -discretization concept is presented for each method. Second, the semismooth Newton method is applied. Finally, numerical examples are presented.

4.3 Optimal boundary control problem

The focus in this section is to present numerical results for boundary control with a partial differential equation in \mathbb{R}^3 . Moreover, the results of [31], also presented in [166] are recalled and a further refinement strategy for these problems is presented.

The optimal boundary control problem stated in section 2.1 leads to the discrete equation system

$$a(y_N^*, v_N) = \langle f, v_N \rangle_\Omega + \langle u_N^*, v_N \rangle_{\Gamma_N} \quad \forall v_N \in V_{hp} \quad (4.13)$$

$$a(v_N, q_N^*) = \langle y_N^* - y_d, v_N \rangle_\Omega \quad \forall v_N \in V_{hp} \quad (4.14)$$

$$u_N^* = P_{[u_a, u_b]} \left(-\frac{1}{\alpha} q_N^*|_{\Gamma_N} \right) \quad (4.15)$$

i.e. $U = L_2(\Gamma_N)$.

4.3.1 Two-dimensional case

Theoretical results for \mathbb{R}^2 are given in [31, 163, 166]. Here, the most important theoretical statements on applying bc-fem are recalled. Furthermore, some advantages and disadvantages when using the vertex concentrated fem (vc-fem) in [163] are given and a combination of the bc-fem and vc-fem is presented in the numerical examples.

First the concentration is on the bc-fem. The first task is to give an estimate for the state and the control under the assumptions given in section 2.1.

Theorem 4.8. [31] *Let assumption 2.9 and assumption 2.10 hold and $f, y_d \in B_{1-\delta}^0(c_f, \gamma_f)$ with $c_f, \gamma_f > 0$. Let τ be a geometric mesh on Ω with mesh size h , \mathbf{p} a linear degree vector with sufficiently large slope ζ . Let (u^*, y^*, q^*) and (u_N^*, y_N^*, q_N^*) be the solutions to the optimal boundary control problem and its discretized version with the corresponding states and adjoint states. The solution to the boundary value problem (4.13) shall be $H^{1+\delta}$ -regular with $\delta \in (0, 1)$, that means $y^*, q^* \in H^{1+\delta}(\Omega)$. Then there exists a constant $\tilde{c} > 0$ independent of h and it holds*

$$\|u^* - u_N^*\|_{L_2(\Gamma_N)} + \|y^* - y_N^*\|_{L_2(\Omega)} \leq \tilde{c}h^\delta.$$

Although the proof is not recalled here, the most important basics of the proof are cited to point out the discrepancy between the theoretical and numerical results. The results above can be yielded by applying the theorem of 4.4 and estimate $\|q^* - q^N\|_{L_2(\Gamma_N)}^2$ and $\|y^* - y^N\|_{L_2(\Omega)}^2$. According to [31, Lemma 3.6, Lemma 3.11] (see also [166]) it holds

$$\begin{aligned} \|q^* - q^N\|_{L_2(\Gamma_N)} &\leq \tilde{c}h^{\delta+1/2}, \\ \|y^* - y^N\|_{L_2(\Omega)} &\leq \tilde{c}h^\delta. \end{aligned}$$

Furthermore, it holds

$$\|q^* - q^N\|_{H^1(\Gamma_N)} \leq \tilde{c}h^\delta \quad \text{and} \quad \|y^* - y^N\|_{H^1(\Omega)} \leq \tilde{c}h^\delta.$$

The considerations above already indicate that the yield estimate might not be sharp. Numerical experiments confirm that since they show an error reduction of $h^{2\delta}$ (see [31, 166]). However, better estimates are not available yet and furthermore possibly hard to obtain, since the Aubine-Nitsche trick does not work for bc-fem because no error estimate of the type $\|y - y_N\|_{H^1(\Omega)} \leq \tilde{c}h^\delta \|f\|_{L_2(\Omega)}$ is available for solutions of the elliptic partial differential equation (4.13) with right-hand side f and $u = 0$. The best currently available L_2 -estimate was proven by Eibner and Melenk in [62]. They show that for every compact $\Omega' \subset\subset \Omega$ there exists $\delta' \in [0, \delta]$ such that for all elements $K \subset\subset \Omega'$ the error estimate $\|y - y_N\|_{L_2(K)} \leq \tilde{c}h^{\delta+\delta'}$ holds. However, δ' depends on Ω' , and it is unclear under which conditions $\delta = \delta'$ can be proven.

Remark 4.9. *Due to theorem 3.31, i.e. $h \sim N^{-1}$, for bc-refinement, theorem 4.8 yields*

$$\sqrt{\alpha} \|u^* - u_h^*\|_{L_2(\Gamma_N)} + \|y^* - y_h^*\|_{L_2(\Omega)} \leq \tilde{c}N^{-\delta}$$

if the problem is $H^{1+\delta}$ -regular. In the case of uniform h -refinement, the approximation space grows as $N \sim h^{-2}$, which would lead to (combined with an estimate of [48])

$$\sqrt{\alpha} \|u^* - u_h^*\|_{L_2(\Gamma_N)} + \|y^* - y_h^*\|_{L_2(\Omega)} \leq \tilde{c}N^{-\frac{3}{4}\delta}$$

for an $H^{1+\delta}$ -regular problem (see [31]). For an H^2 -regular problem, therewith it follows $\mathcal{O}(N^{-1})$ for bc-fem and $\mathcal{O}(N^{-3/4})$ for uniform h-fem. That means for a N being large enough the discretization by bc-fem gives a smaller error than for uniform h-fem. However, it has to be mentioned that using grading meshes, similar results for h-fem can be yielded, see [6].

In [166] also another refinement strategy, the so called vertex concentrated finite element method (vc-fem) is used to solve the optimal boundary control problem. In the case of vc-refinement, all nodes in corners and all elements where the active and the inactive set meet, are h -refined, all other elements are p -refined. This refinement strategy is again based on the projection formulation, but avoids h -refinement on elements with enough smoothness, i.e. elements on which the control u_N is active or inactive on the whole element.

The advantage when using the vertex concentrated fem is that one yields an exponential order of convergence compared to the algebraic one for bc-fem, see [163, 166]. The drawback is that it is assumed to know all switching nodes at the very beginning for obtaining numerical results. Of course in practice these switching points are unknown. However, a suitable starting mesh and further h -refinement of all neighbour elements of the expected switching points lead to good results and show that the proved exponential convergence rate can be yielded ([166, Figure 5.7, Figure 5.8]). Nevertheless, in the case of an oscillating control u_N with active constraints u_a and u_b , problems to find a not too fine suitable mesh are expected.

In the case of a not vanishing Dirichlet boundary, i.e. $\text{meas}(\Gamma_{\mathcal{D}}) \neq \emptyset$, a modified bc-fem, the Neumann bc-fem is proposed. In Neumann bc-fem only elements in corners and elements on the Neumann boundary are h -refined, all other elements are p -refined. This reduces the number of degrees of freedom compared to bc-fem and furthermore, avoids problems to estimate the switching points. However, it has to be stated, that for Neumann bc-fem the same convergence rate as for bc-fem is expected. This is due to the fact that an exponential order of convergence can only be yielded if h -refinement is done in a limited number of points, see [140].

Remark 4.10. *In the case of pure Neumann boundary, the Neumann bc-fem corresponds to bc-fem.*

After choosing a suitable refinement strategy, the boundary control problem is solved with the semismooth Newton method. This leads to the equation system

$$\left(M_{\Gamma_{\mathcal{N}},\mathcal{J}} + \frac{1}{\alpha} M_{\Gamma_{\mathcal{N}},\mathcal{J}}^{\top} K^{-1} M_{\Gamma_{\mathcal{N}}} K^{-1} M_{\Gamma_{\mathcal{N}},\mathcal{J}} \right) \vec{u} = -\frac{1}{\alpha} M_{\Gamma_{\mathcal{N}},\mathcal{J}}^{\top} K^{-1} M_{\Gamma_{\mathcal{N}}} K^{-1} (\vec{f} - \vec{y}_d), \quad (4.16)$$

which has to be solved in each Newton step, where

$$\mathfrak{f}_j = \int_{\Omega} f(x) \varphi_j(x) \, dx + \int_{\mathfrak{A}_a(u_N)} u_a \varphi_j(x) \, dx + \int_{\mathfrak{A}_b(u_N)} u_b \varphi_j(x) \, dx.$$

$M_{\Gamma_{\mathcal{N}}}$ denotes the mass matrix on the boundary and $M_{\Gamma_{\mathcal{N}},\mathcal{J}}$ the mass matrix on the boundary over the inactive set.

4.3.1.1 Numerical experiments

Next a simple example – named example 4.3.1.1 is considered. For further results to bc- and vc-fem see [31, 163].

Here a problem with oscillating adjoint is considered. It can be described by the boundary value problem for the state

$$\begin{aligned} -\Delta y(x) &= f(x) && \text{in } \Omega \\ \frac{\partial y}{\partial n}(x) &= u(x) + e_y(x) && \text{on } \Gamma_{\mathcal{N}} \\ y(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \end{aligned}$$

the boundary value problem for the adjoint

$$\begin{aligned} -\Delta q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ \frac{\partial q}{\partial n}(x) &= e_q(x) && \text{in } \Gamma_{\mathcal{N}} \\ q(x) &= 0 && \text{in } \Gamma_{\mathcal{D}} \end{aligned}$$

and the projection formula

$$u(x) = P_{[-0.5, 0.5]}(-q(x)|_{\Gamma_{\mathcal{N}}}).$$

The domain is given by $\Omega = (0, 1)^2$, the Neumann boundary is $\Gamma_{\mathcal{N}} = \{x_1 = 1\} \cup \{x_2 = 1\}$, the Dirichlet boundary is $\Gamma_{\mathcal{D}} = \{x_1 = 0\} \cup \{x_2 = 0\}$. The exact solution to this optimal boundary control problem is

$$\begin{aligned} y(x) &= x_1 x_2 e^{x_1 + x_2}, \\ q(x) &= -x_1 x_2^2 \sin(15\pi x_1) \cos(15\pi x_2), \end{aligned}$$

the data is chosen accordingly.

Remark 4.11. *The inhomogeneities $e_y(x), e_q(x)$ are introduced to construct a test example with known analytical solution. The theoretical estimates are not affected by these inhomogeneities (see [31]).*

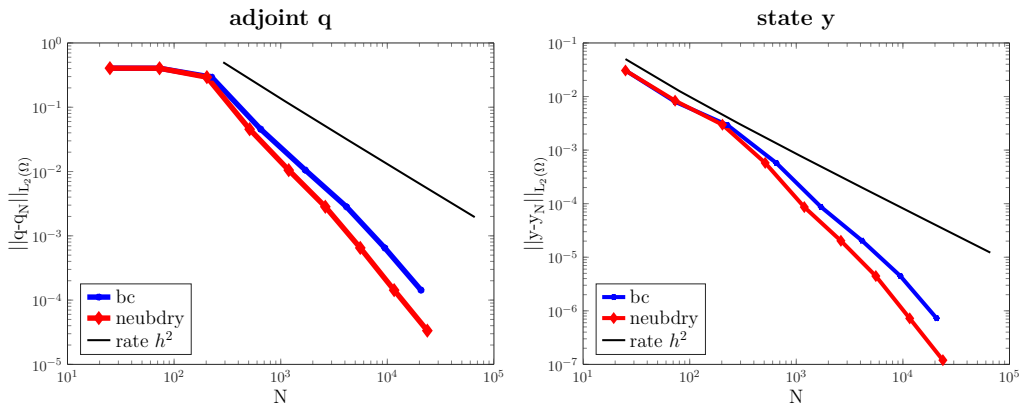


Figure 4.1: comparison of bc-fem and neubdry-fem for example 4.3.1.1

First, the difference between bc-fem and bc-fem only on the Neumann boundary (and additional h refinement on corners) – called neubdry-fem – is considered. The L_2 error in

dependence of the degrees of freedom for both refinements is given in figure 4.1, the corresponding meshes are given in figure 4.2. The expected order of convergence for uniform h -fem is two. The black line in figure 4.1 shows that rate. Furthermore, it can be observed that both convergence rates – the one for bc-refinement and the one for neubdry-refinement – are greater than two. Although neubdry-refinement leads to a decrease in the number of degrees of freedom, the convergence rate stays the same.

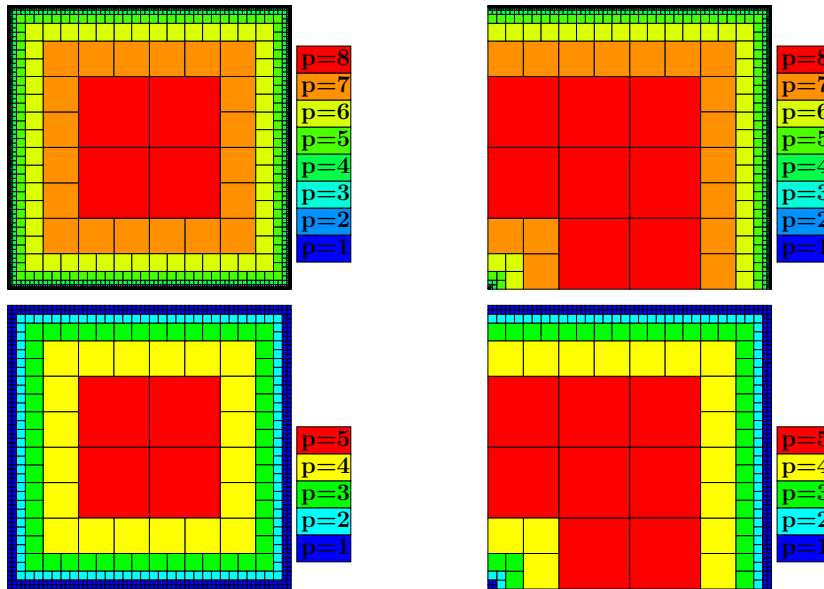


Figure 4.2: left: bc-refinement, right: neubdry-refinement, after five and eight refinements

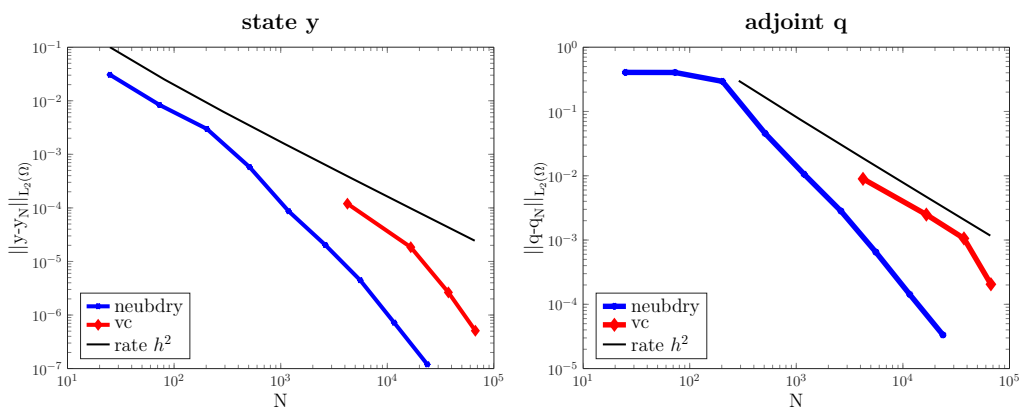


Figure 4.3: comparison of neubdry-fem and vc-fem for example 4.3.1.1

Furthermore, vc-fem and neubdry-fem are compared in figure 4.3. Due to the highly oscillating adjoint q and the need to start with a suitable mesh in vc-fem, the starting mesh (obtained with uniform h -refinement) for vc-fem is quite fine (see figure 4.5). This explains why neubdry-refinement is favourable in this case although with neubdry-refinement only an algebraic convergence rate is yielded, whereas in vc-fem an exponential one is possible.

One possibility to decrease the number of degrees of freedom for vc-fem is, to use neubdry-

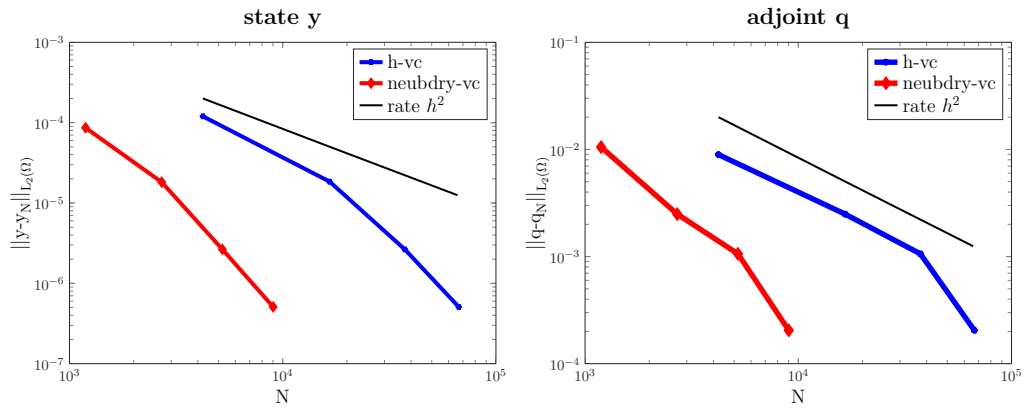
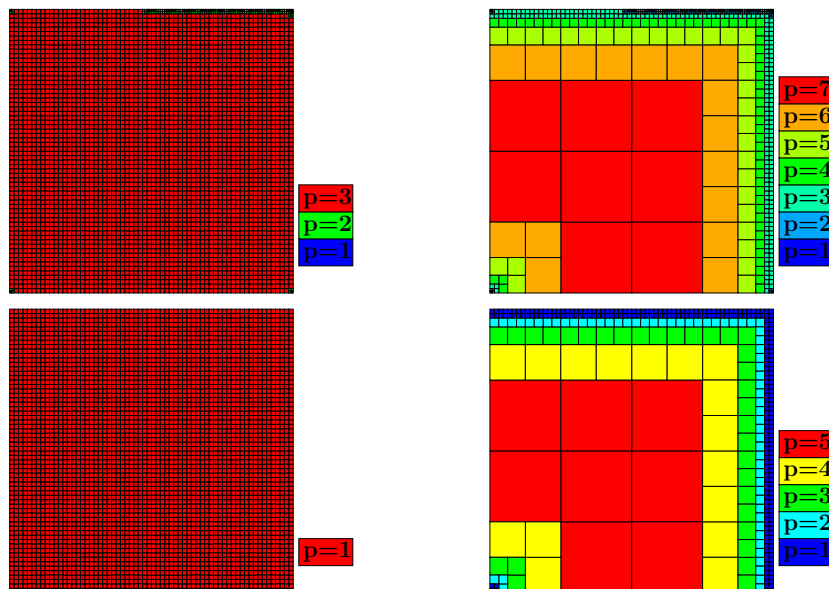


Figure 4.4: influence of different start meshes for vc-fem

Figure 4.5: left: start mesh uniform h -refinement, then vc-fem, right: start mesh neubdry refinement, then vc-fem

refinement instead of uniform h -refinement to find a suitable starting mesh. This choice decreases the number of degrees of freedom significantly, see figure 4.4 for the L_2 error and figure 4.5 for the corresponding meshes.

4.3.2 Three-dimensional case

Since theoretical results are not available now, this subsection concentrates on numerical results.

Remark 4.12. *For the semismooth Newton method (4.16) a suitable starting vector \vec{u}_0 is used. In the given results this starting vector is chosen – except for the first mesh – by a projection of the solution to the former mesh to the considered one.*

Furthermore, in all examples a preconditioner to solve (4.16) is used. Since the mass matrix $M_{\Gamma_{\mathcal{N}}}$ dominates this equation, a suitable preconditioner for the mass matrix is taken. In the case of inactive constraints, as preconditioner $\text{diag}(M_{\Gamma_{\mathcal{N}}})$ is used because there are only low order basis functions on the boundary, due to the application of bc-fem. Therewith $M_{\Gamma_{\mathcal{N}}}$ is well-conditioned and the diagonal as preconditioner is sufficient. In the case of inactive constraints, the preconditioner is modified, and

$$\text{diag}(M_{\Gamma_{\mathcal{N}},\mathcal{I}})$$

is chosen. This is done in order to bring the element size for partly active and inactive elements into play.

4.3.2.1 Example: Cube

The first numerical example is a cube with known analytical solution. The domain of the cube is given by $\Omega = (-1, 1) \times (-1, 1) \times (-1, 1)$. The Neumann boundary is on the faces

$$\begin{aligned} &(-1, 1) \times (-1, 1) \text{ and } x_3 = -1, \\ &(-1, 1) \times (-1, 1) \text{ and } x_3 = 1. \end{aligned}$$

On the other faces homogeneous Dirichlet boundary hold. The optimal control problem is given by

$$\min J(y, u) := \frac{1}{2} \|y - y_d\|_{L_2(\Omega)}^2 + \frac{1}{2} \|u\|_{L_2(\Gamma_{\mathcal{N}})}^2 + \int_{\Gamma_{\mathcal{N}}} e_q y \, ds_x$$

subject to the constraints

$$\begin{aligned} -\Delta y(x) &= f(x) && \text{in } \Omega \\ y(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \\ \frac{\partial y}{\partial n}(x) &= u(x) + e_y(x) && \text{on } \Gamma_{\mathcal{N}} \end{aligned} \tag{4.17}$$

and the box constraints on the control

$$u_a \leq u(x) \leq u_b \quad \text{a.e. on } \Gamma_{\mathcal{N}}.$$

There, the desired state y_d , the inhomogenities e_y, e_q , the right-hand-side f and the box constraints u_a, u_b are given, the state y and the control u have to be calculated.

According to theorem 2.4 and by using the equivalence in theorem 2.5, this is equivalent to calculate the solution to the primal equation (4.17), the dual equation

$$\begin{aligned} -\Delta q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ q(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \\ \frac{\partial q}{\partial n}(x) &= e_q(x) && \text{on } \Gamma_{\mathcal{N}} \end{aligned}$$

and the projection formula

$$u(x) = P_{[-2.3, -0.5]}(-q(x)|_{\Gamma_{\mathcal{N}}}).$$

Remark 4.13. *The inhomogenities $e_q(x), e_y(x) \in H^{1/2}(\Gamma_{\mathcal{N}})$ are introduced to construct a test example (see [158]) with known solution.*

The state and the adjoint are

$$\begin{aligned} y(x) &= \left(\frac{\pi}{2} - 1\right) \cos\left(x_1 \frac{\pi}{2}\right) \cos\left(x_2 \frac{\pi}{2}\right) e^{x_3}, \\ q(x) &= \cos\left(x_1 \frac{\pi}{2}\right) \cos\left(x_2 \frac{\pi}{2}\right) e^{x_3}. \end{aligned}$$

First, a comparison between uniform h - and bc-refinement is made, see figure 4.6. Since the coarse mesh consists of only one cube, the first three refinements are identical. That is caused by the fact that several uniform refinements are necessary to get higher polynomial degrees in bc-fem, because in here, in each step only h - or p -refinement is performed. Figure 4.6 shows, that bc-fem is superior to uniform h -fem with respect to the number of degrees of freedom. Furthermore, it can be observed that the uniform h -fem has the expected quadratic convergence rate.

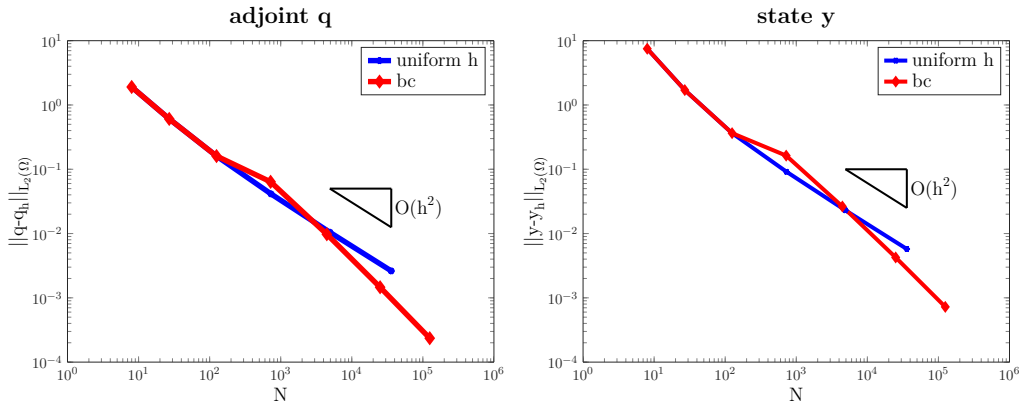


Figure 4.6: L_2 versus number of degrees of freedom for h - and bc-refinement for example 4.3.2.1

Remark 4.14. *For a further decrease of the number of degrees of freedoms, analogue refinement strategies as in two dimension, i.e. vc -refinement or $neubdry$ -refinement could be applied.*

Moreover, the adjoint q and the state y are plotted for different meshes. In figure 4.7 the adjoint is plotted for 25 031 degrees of freedom, i.e. the mesh after 5 refinement steps. In figure 4.9 again the adjoint but with a further bc-refinement step, is plotted. The results for the corresponding state y for these two meshes is plotted in figure 4.8 and figure 4.10. The plots show, that on most faces (except two faces of the coarse cube), there is homogeneous Dirichlet boundary, which explains, why there is zero on nearly all boundary faces.

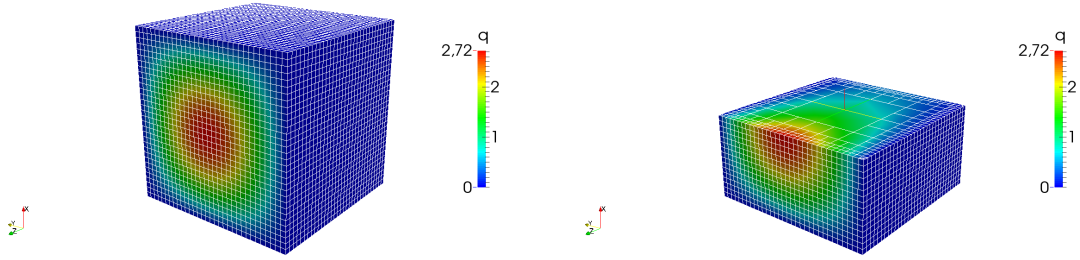


Figure 4.7: adjoint (on whole and truncated domain) for 25 031 degrees of freedom for example 4.3.2.1

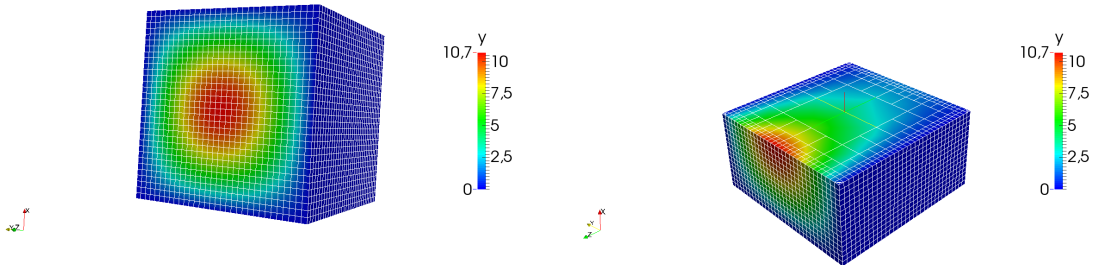


Figure 4.8: state (on whole and truncated domain) for 25 031 degrees of freedom for example 4.3.2.1

The polynomial distribution of the elements can be found in figure 4.11. It shows the usual bc-refinement, i.e. in each refinement step all elements on the boundary are h -refined whereas all non-boundary elements are p -refined.

Figure 4.12 shows, that there are active parts for the set \mathfrak{A}_a and \mathfrak{A}_b . Due to the constraints $u_a = -2.3$ and $u_b = -0.5$, the projection formula cuts the adjoint q at 2.3 for the upper bound and 0.5 for the lower bound. All these parts are active sets, therefore for the control these values are adjusted (and set to u_a or u_b , depending on the kind of active set) by the projection formula.

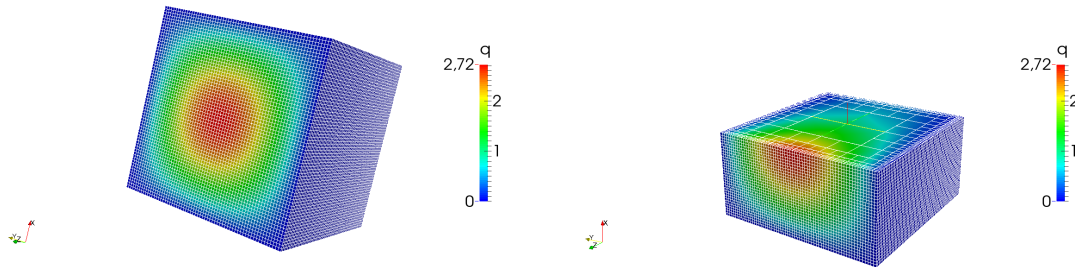


Figure 4.9: adjoint (on whole and truncated domain) for 124 921 degrees of freedom for example 4.3.2.1

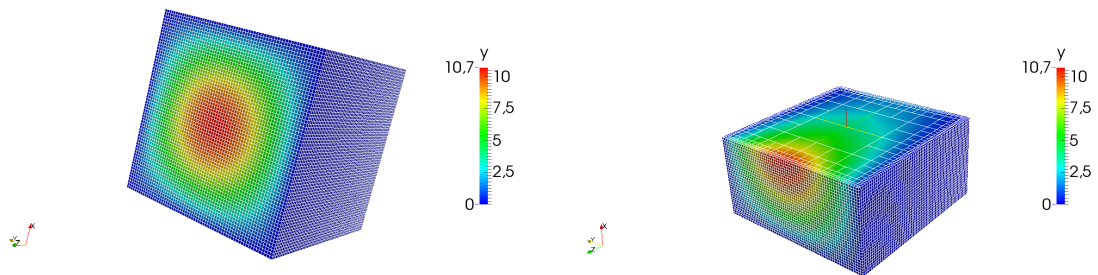


Figure 4.10: state (on whole and truncated domain) for 124 921 degrees of freedom for example 4.3.2.1

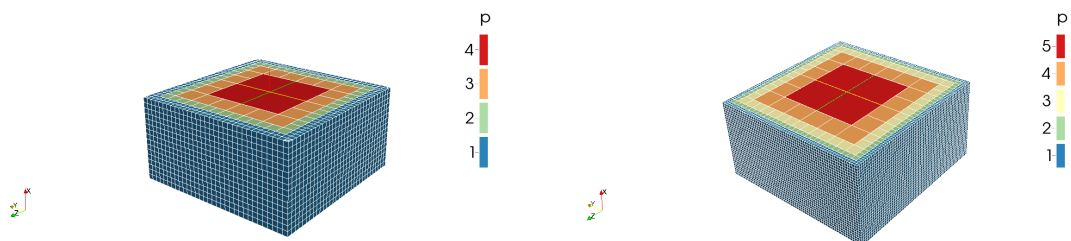
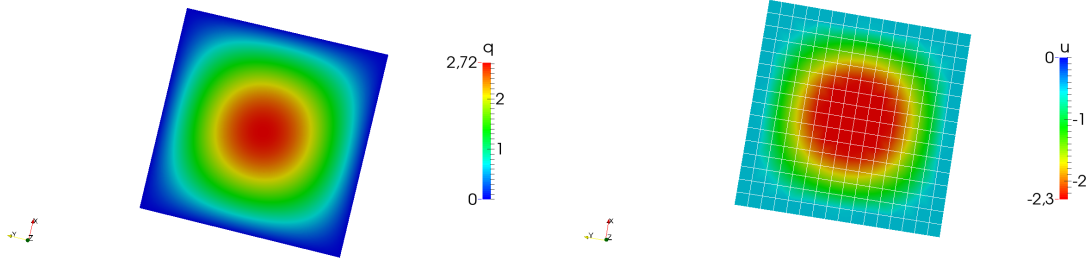


Figure 4.11: polynomial degree distribution (intersection in the middle of the cube) for 25 031 and 124 921 degrees of freedom for example 4.3.2.1, respectively

Figure 4.12: adjoint q and control u on face with constraints

4.3.2.2 Example: Steam

As first example with unknown solution a steam is considered. There the domain is given by $(0, 2) \times (0, 1) \times (0, 4)$. The differential equation for the state y is

$$\begin{aligned} -\Delta y(x) &= f(x) && \text{in } \Omega \\ y(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \\ \frac{\partial y}{\partial n}(x) &= u(x) && \text{on } \Gamma_{\mathcal{N}} \end{aligned}$$

the dual one

$$\begin{aligned} -\Delta q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ q(x) &= 0 && \text{on } \Gamma_{\mathcal{D}} \\ \frac{\partial q}{\partial n}(x) &= 0 && \text{on } \Gamma_{\mathcal{N}}. \end{aligned}$$

The projection formula is given by

$$u(x) = P_{[0,1,1,0]}(-q(x)|_{\Gamma_{\mathcal{N}}}).$$

The data is chosen by

$$\begin{aligned} f(x) &= 0, \\ y_d(x) &= 2e^{x_1+x_2} + 3e^{x_2+x_3}. \end{aligned}$$

Again, the state y and the adjoint q are plotted for two meshes. The adjoint is plotted in figure 4.13 for 28 327 and in figure 4.15 for 143 911 degrees of freedom. The results for the state are given in figure 4.14 and figure 4.16. The plots show that the adjoint is between $[-3.07, 0]$, whereas the state is between $[0, 0.34]$. As in the example before, there are again two faces of the steam on which a Neumann control holds. For the Dirichlet boundary, there hold homogeneous boundary conditions, therewith the values are zero there. In figure 4.14 and figure 4.16 it can be observed that the state $y(x)$ is quite small on most parts of the domain. For the adjoint $q(x)$ the situation is different, see figure 4.13 and figure 4.15.

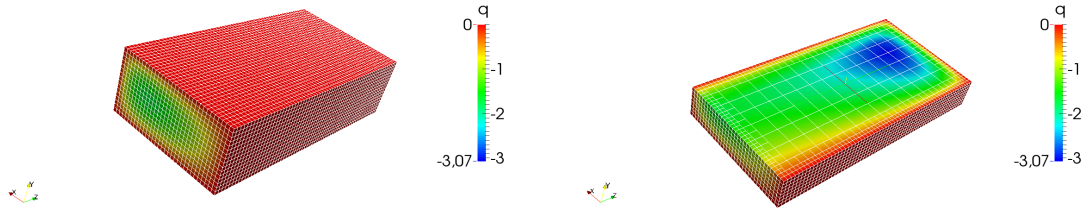


Figure 4.13: adjoint (on whole and truncated domain) for 28 327 degrees of freedom for example 4.3.2.2

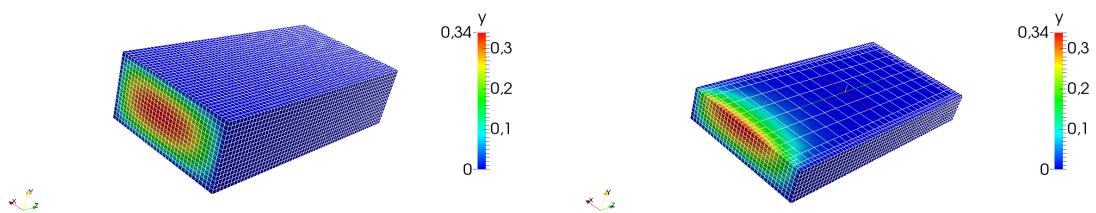


Figure 4.14: state (on whole and truncated domain) for 28 327 degrees of freedom for example 4.3.2.2

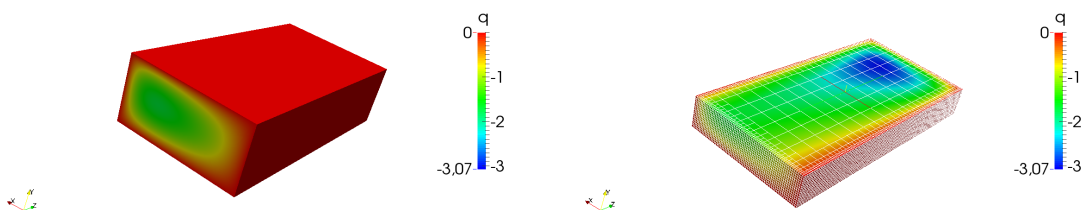


Figure 4.15: adjoint (on whole and truncated domain) for 143 911 degrees of freedom for example 4.3.2.2

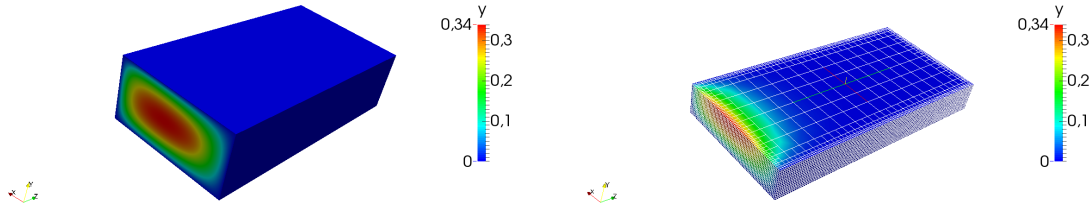


Figure 4.16: state (on whole and truncated domain) for 143 911 degrees of freedom for example 4.3.2.2

The polynomial distribution for these two meshes is plotted in figure 4.17. As in example 4.3.2.1. bc-refinement is used for the refinement and figure 4.17 shows the polynomial degree distribution of the elements. The polynomial degree of the corresponding edges and faces is determined by the minimum degree condition (see definition 3.23). For example, a face between two elements – one with polynomial degree 2 and the other one with polynomial degree 3 – has polynomial degree 2.

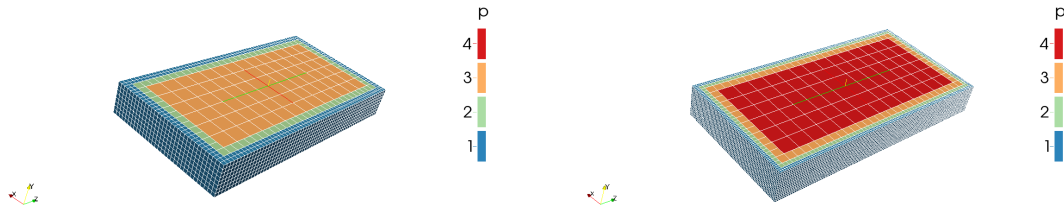


Figure 4.17: polynomial degree distribution (truncated domain) for 28 327 and 143 911 degrees of freedom for example 4.3.2.2, respectively

In figure 4.18 by using the projection formula, and $u_a = 0.1$, $u_b = 1.0$, a comparison between the adjoint q and the control u shows, that there are both, active parts \mathfrak{A}_a and active parts \mathfrak{A}_b . In the plot of the adjoint, all elements with values between $(-1.0, -0.1)$ are inactive, all other elements are fully active or contain at least active parts.

4.3.2.3 Example: Three feet

The next example fulfills the same equations, then the steam-example, but with a different desired state and a different domain. As domain four cubes put together in order to get a three feet shape, are used (see figure 4.19). Except on the L-face on the boundary (bottom on the left plot in figure 4.3.2.3) on the faces hold homogeneous Dirichlet boundary. On the L-face, there is the Neumann boundary where the control comes into play. The right-hand-side

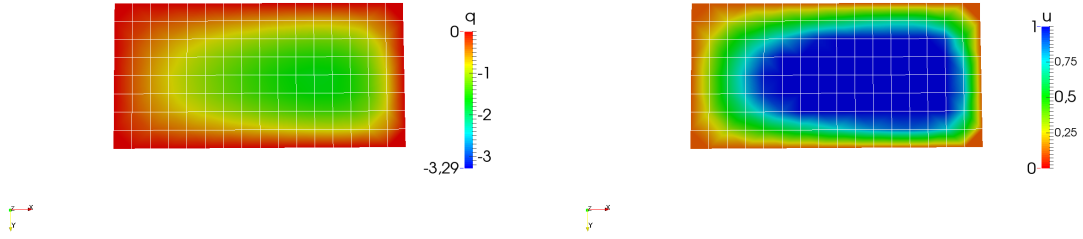
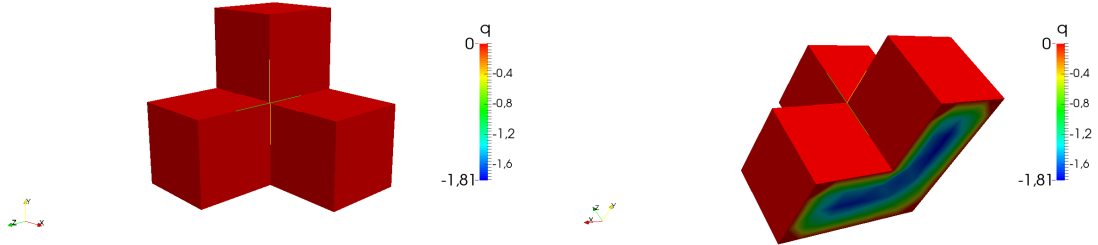
Figure 4.18: adjoint q and control u for face with interface elements, example 4.3.2.2

Figure 4.19: domain (plotted is the adjoint) for example 4.3.2.3

of the state equation and the desired state are given by

$$\begin{aligned} f(x) &= 0 \\ y_d(x) &= 2e^{x_1+x_2} + 3e^{x_2+x_3} \end{aligned}$$

and the constraints for the control are set to $u_a = 0.1$ and $u_b = 1.2$.

In figure 4.20 the adjoint for 2610 degrees of freedom, which consists of 1607 elements, is plotted. The corresponding state is given in figure 4.22. On the most faces at the outside there hold homogeneous Dirichlet boundary conditions, therewith the values are zero. The Neumann boundary on which the control acts, has the shape of an L. Moreover, figure 4.20 shows, that there are elements with active sets, since only values between $[-1.2, -0.1]$ lead to inactive elements due to the projection formula and the constraints $u_a = 0.1$ and $u_b = 1.2$ for the used regularization parameter $\alpha = 1.0$.

A closer look on the polynomial degree distribution for the mesh with 1607 elements is given in figure 4.21. Since the polynomial degree is one on the whole boundary due to the boundary concentrated refinement, the parts of interest are in the interior of the domain. Therewith, the domain in figure 4.21 shows two different truncations of the domain in order to get a better knowledge of the polynomial distribution in the interior. It shows, that the highest polynomial degree for these results is two.

The results for 15857 degrees of freedom can be found in figure 4.23, figure 4.24 and figure 4.25. The corresponding mesh has 8299 elements. In figure 4.23 it can be observed, that both constraints are active, since the adjoint is smaller than -1.2 and bigger than 0.1 .

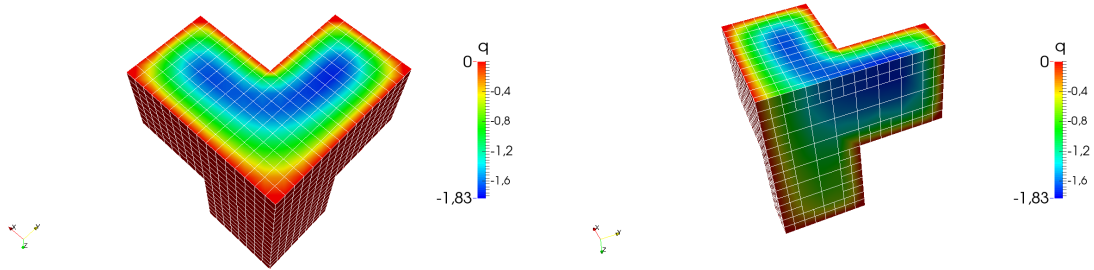


Figure 4.20: adjoint (on whole and truncated domain) for example 4.3.2.3 for 2610 degrees of freedom

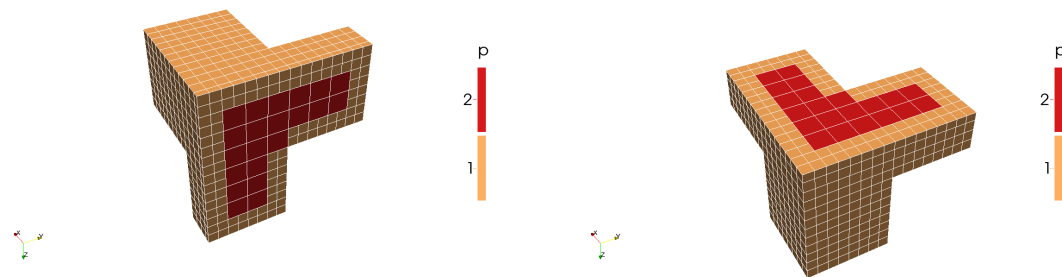


Figure 4.21: polynomial degree distribution for example 4.3.2.3 for 2610 degrees of freedom

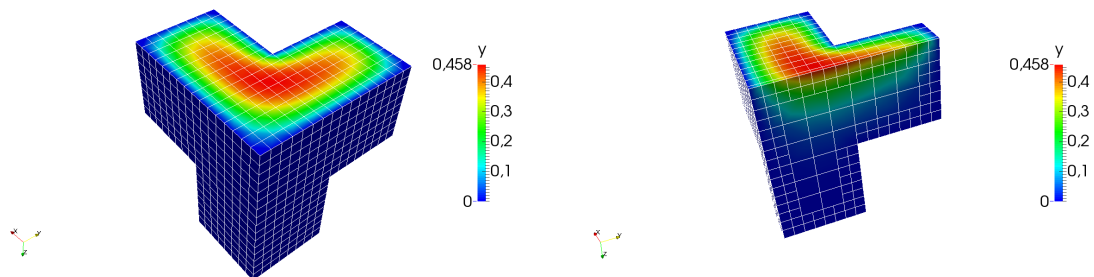


Figure 4.22: state (on whole and truncated domain) for example 4.3.2.3 for 2610 degrees of freedom

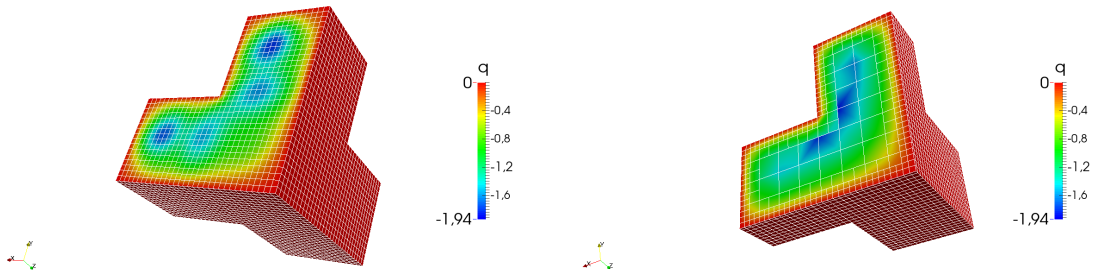


Figure 4.23: adjoint (on whole and truncated domain) for example 4.3.2.3 for 15 857 degrees of freedom

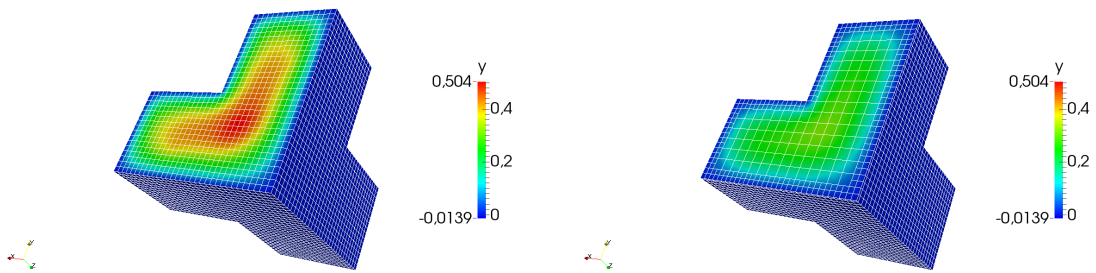


Figure 4.24: state (on whole and truncated domain) for example 4.3.2.3 for 15 875 degrees of freedom

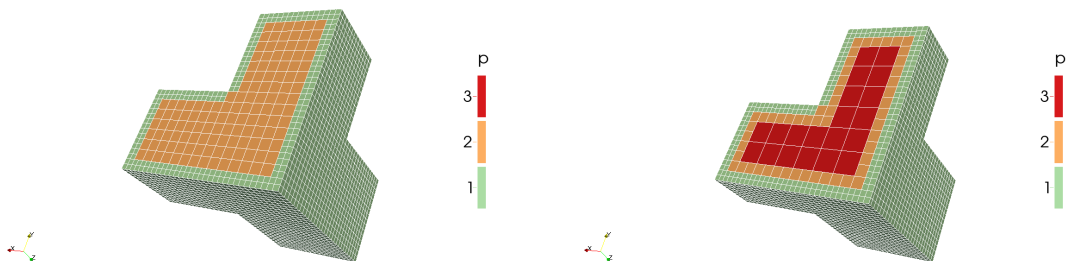


Figure 4.25: polynomial degree for example 4.3.2.3 for 15 875 degrees of freedom

Remark 4.15. In figure 4.23 the discontinuities occur only in the plot, since only the values in the nodes are plotted, which can lead to discontinuities in the plot if hanging nodes appear. In fact the adjoint is continuous.

To give a better understanding of the polynomial degree distribution two blocks are cut out of the domain, see figure 4.25. There it can be seen, that on the boundary the polynomial degree is 1, but in the interior the polynomial degree gets higher. The highest polynomial degree on this mesh is 3.

Remark 4.16. For the considered numerical examples face-based refinement has been chosen, i.e. all elements which have a face on the boundary are h -refined, all other elements are p -refined. That is the reason why the element in the L -corner is p - but not h -refined. An edge-based refinement would lead to different results.

4.3.2.4 Example: Cube with holes

The last example fulfills again the same equations but again the desired state, the right-hand-side and the domain are different. The domain is plotted in figure 4.26.

The right-hand-side for the state and the desired state are given by

$$\begin{aligned} f(x) &= 0 \\ y_d(x) &= 10 \sin(\pi x_1) + 5 \cos(\pi x_2 x_3) \end{aligned}$$

the constraints for the control are $u_a = -0.7$ and $u_b = 0.9$.

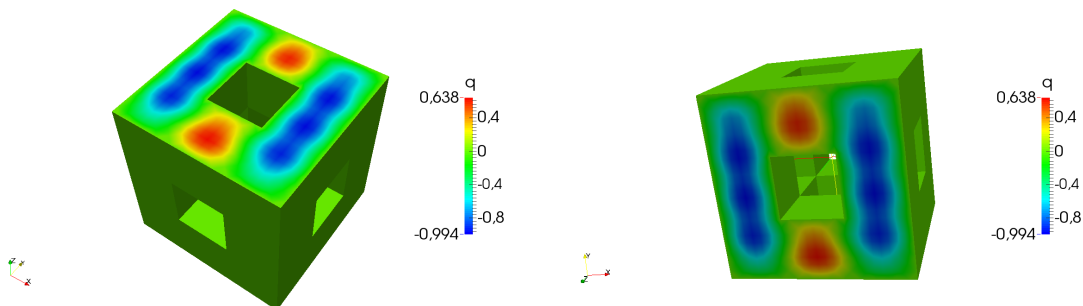


Figure 4.26: domain (plotted is the adjoint) for example 4.3.2.4

In figure 4.27, the adjoint q is plotted. It shows that only the faces on top of the domain have Neumann boundary with controls. All other faces have homogeneous Dirichlet boundary. Since the constraints are $u_a = -0.7$ and $u_b = 0.9$, it follows that the adjoint is active for values bigger than 0.7 and smaller than -0.9 . Therewith, figure 4.27 shows that only the upper constraint is active, the constraint u_a is inactive.

Furthermore, in figure 4.27 some values of the adjoint in the interior are given, since in the figure on the left hand side a slice of the top of the adjoint is cut. Therewith, it is possible to get a glance at some values in the interior. Moreover, the size of the elements – which gets bigger in the interior due to higher polynomial degrees – can be observed.

The state is given in figure 4.28. As in the plot for the adjoint, in the left part of figure 4.28 a slice is cut from the state in order to see the mesh-size and the values of the state there.

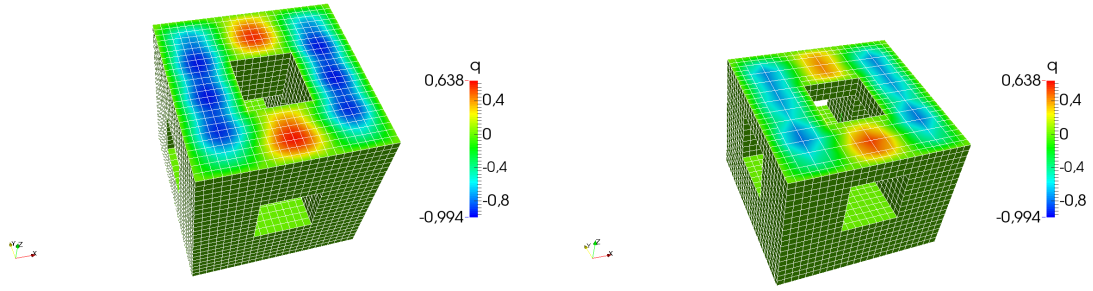


Figure 4.27: adjoint (on whole and truncated domain) for example 4.3.2.4

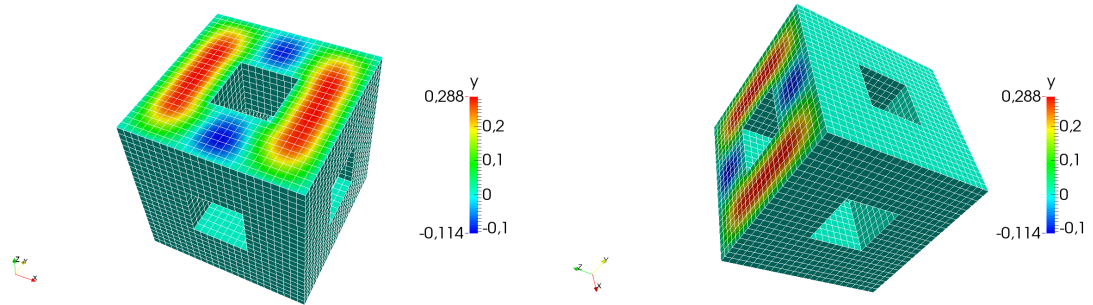


Figure 4.28: state y (on whole domain) for example 4.3.2.4

The polynomial degree distribution of the elements is given in figure 4.29. For a better understanding the domain is cut in order to plot the polynomial degree inside the domain. On the boundary, the polynomial degree of all elements with a boundary face have polynomial degree 1. Elements in the interior have polynomial degree 2, which is the highest polynomial degree of the considered mesh.

Remark 4.17. *Due to the bc-fem, the number of elements with low polynomial degree is quite high. In order to decrease the number of degrees of freedom, Neumann boundary concentrated refinement, as in subsection 4.3.1 could be used. A further possibility is to extend the idea of vertex concentrated refinement considered in subsection 4.3.1 to three dimensions. Then, each element with edges that have active and inactive parts is h -refined, all others are p -refined.*

Remark 4.18. *In order to check if an element is active or inactive for boundary control in three dimensions and polynomial degree $p = 1$, the values of the nodes have to be checked.*

Remark 4.19. *In case of using tetrahedral elements hanging nodes can be avoided by using red-green refinement.*

Next, distributed optimal control problems are considered.

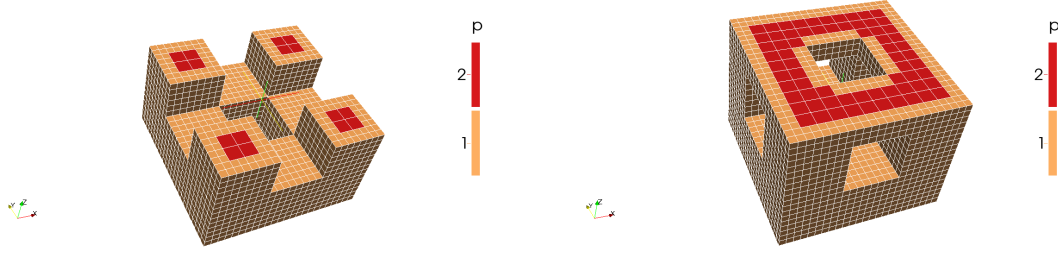


Figure 4.29: polynomial degree distribution (both are truncated versions of the domain) for example 4.3.2.4

4.4 Distributed optimal control problem

In this section the distributed optimal control problem introduced in section 2.2, is considered. Its discretized version is given by

$$\begin{aligned}
 a(y_N^*, v_N) &= \langle u_N^*, v_N \rangle_\Omega & \forall v_N \in V_{hp} \\
 a(v_N, q_N^*) &= \langle y_N^* - y_d, v_N \rangle_\Omega & \forall v_N \in V_{hp} \\
 u_N^* &= P_{U_{ad}} \left(-\frac{1}{\alpha} q_h^*|_\Omega(x) \right)
 \end{aligned} \tag{4.18}$$

where (u_N^*, y_N^*, q_N^*) is the discrete solution to the distributed optimal control problem. The problem is – as in the case of boundary control – again discretized by variational discretization by Hinze [88]. Therewith, only the discrete state y_N^* and the discrete adjoint q_N^* are finite element functions. The discrete control u_N^* is only given by the projection formula (4.18). For h -fem such kind of problems are considered in [6, 92], further results on this topic can for example be found in [8, 47, 117]. An application of hp -fem to the considered problem is already given in [164]. However, there interior points methods are used in order to solve the problem. In this thesis the problem is solved with the semismooth Newton method, where the hp -discretization of the problem is based on its structure, see subsection 4.4.1. Due to the application of variational discretization and the use of semismooth Newton methods, an integration only over parts of the reference element is necessary. This topic is investigated in subsection 4.4.2. Numerical experiments are given in two dimensions in subsection 4.4.4. An extension to three dimensions is possible.

First, some remarks on regularity are made. The main difficulty here is, that in distributed optimal control problems the interface between the active and inactive sets, which comes into play by the projection formula (4.18) is unknown. Moreover, to the best knowledge of the author, there are no results on the structure of the active and inactive sets and the regularity of the control u which can be used for hp -fem. Therewith, only known theoretical results for h -fem are given, since here less regularity is needed in order to prove them.

Remark 4.20. *In the case of interfaces which separate the domain in Lipschitz domains, an extension of [165] is possible.*

In the case of uniform h -fem with piecewise linear elements, the following result is known (see remark 4.5):

Corollary 4.21. (see e.g. [91]) Let assumption 2.12 hold, then, the error of the distributed optimal control problem introduced in section 2.2 can be estimated by

$$\alpha \|u_N^* - u^*\|_{L_2(\Omega)} + \|y^* - y_N^*\|_{L_2(\Omega)} \leq \tilde{c} \left(\frac{1}{\alpha} + 1 \right) h^2,$$

in the case of a uniform h -fem discretization for a constant $\tilde{c} > 0$

4.4.1 Refinement strategies

In this subsection two hp -refinement strategies are presented, suitably for refining the considered optimal control problem.

It is known a-priori that parts with less smoothness are corner regions and the interface between the active and inactive sets due to the projection formula (4.18). For corners the regularity of the solution depends on the angle of the corner, see e.g. [140]. In this thesis, the concentration is on the lack of smoothness due to the interface between active and inactive sets. As already stated, h -refinement is favourable if there is less smoothness. Therewith, the interface between active and inactive sets is h -refined.

A possible fully a-priori hp -refinement is given in figure 4.30. A similar refinement – named ic-fem – is given in [30]. Moreover, the given a-priori strategy is in fact an extension of the vc-fem and inspired by bc-fem. In [165] ic-fem is already used to solve an optimal control problem with non-moving and a-priori known interface. However, in the case of distributed optimal control, the interface can move, since it is given by the projection formula.

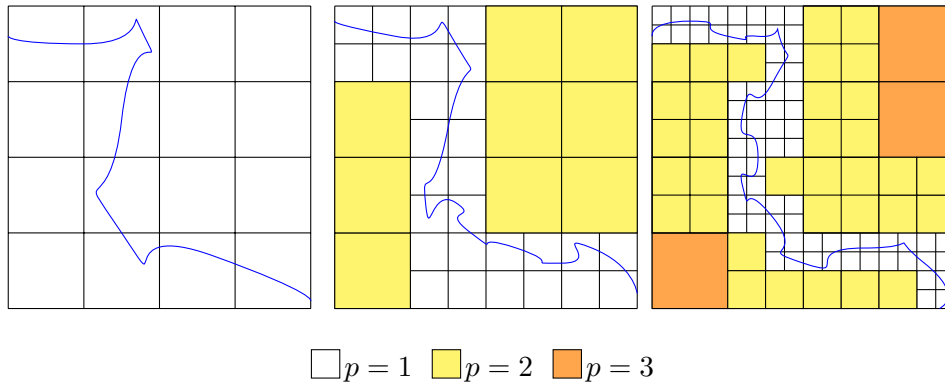


Figure 4.30: possible refinement due to projection formula (4.18)

In fact, this leads to the same problem as in vc-refinement for the optimal boundary control problem (compare subsection 4.3.1): how to refine the mesh that the interface does not jump out of elements with low polynomial degree? Therewith, an extended strategy, i.e. the h -refinement of all interface-neighbour elements is proposed, in order to ensure, that the whole interface is contained in h -refined elements. Next, an a-priori hp -refinement, based on the considered facts is given in algorithm 9. Since it is an extension of ic-fem, it is named nic-fem, meaning neighbour ic-fem. Three different kinds of elements are distinguished.

1. active elements : elements with $u_N|_K = u_a$ or $u_N|_K = u_b$.
2. inactive elements : elements with $u_N|_K = -\frac{1}{\alpha}q_N|_K$.

3. interface elements : elements which have both, active and inactive parts.

Remark 4.22. *Algorithm 9 concentrates only on the necessary h -refinement due to the projection formula (4.18). In numeric examples of course the structure of the domain and its boundary conditions are taken into account. In all such elements, where a-priori less smoothness is known, additional h -refinement is made.*

Algorithm 9: a-priori refinement strategy, called nic-fem

input : suitable mesh τ_0
output: refined mesh τ

for $k = 0, \dots, \#elements\ of\ \tau_0$ **do**
 | find all interface elements and collect them in $\cup K_X$
 | find all neighbour elements and collect them in $\cup K_N$
 | find all corner elements and collect them in $\cup K_C$
if element $K_j \in (\cup K_X) \cup (\cup K_N) \cup (\cup K_C)$ **then**
 | do h refinement
else
 | do p refinement

Furthermore, a second, similar strategy but based on error estimators, as introduced in subsection 3.2.6.2 is proposed. The proposed strategy is a modification of algorithm 7. In this new refinement strategy, named errest-refinement, on each interface element (and again its neighbour elements) the predicted error is set to be zero. This enforces h -refinement in all interface elements and its neighbours. The proposed strategy is given in algorithm 10.

Algorithm 10: error estimator based nic-refinement, called errest-refinement

input : suitable mesh τ_0
output: refined mesh τ

for $k = 0, \dots, \#elements\ of\ \tau_0$ **do**
 | find all interface elements and collect them in $\cup K_X$
 | find all neighbour elements and collect them in $\cup K_N$
for element $K_j \in (\cup K_X) \cup (\cup K_N)$ **do**
 | set $\eta_{K_j}^{(prec)} = 0$
 | set $\tilde{\eta}_{K_j}^2 = \sigma \bar{\eta}^2 + 1$
call algorithm 7 and use $\tilde{\eta}_K^2$ if available instead of η_K^2 , but do not change the calculation of $\bar{\eta}$

Remark 4.23. *For the calculation of the error estimates either the primal problem for the state y or the dual problem for the adjoint q can be used. The choice for one of it may depend on the expected smoothness of these unknowns. A further possibility would be the application of the error estimator of both of it and combine them suitably.*

Next, the discretized optimal problem is solved with the semismooth Newton method. Since a main part of this thesis was the implementation of a code therefore, the most important points are pointed out.

4.4.2 Integration on interface elements

In order to solve the optimal control problem the semismooth Newton method, algorithm 8 is used. There, in each Newton step the equation system

$$\left(M_{\mathcal{J}} + \frac{1}{\alpha} M_{\mathcal{J}} K_N^{-1} M_{\Omega} K_N^{-1} M_{\mathcal{J}} \right) \vec{u} = \frac{1}{\alpha} M_{\mathcal{J}} K_N^{-1} M_{\Omega} \vec{y}_d - \frac{1}{\alpha} M_{\mathcal{J}} K_N^{-1} M_{\Omega} K^{-1} \vec{f} \quad (4.19)$$

has to be solved. K_N denotes the stiffness matrix and M the mass matrix. Furthermore, it is

$$\begin{aligned} f_j &= \int_{\Omega} f(x) \varphi_j(x) dx + \int_{\mathfrak{A}_a} u_a \varphi_j(x) dx + \int_{\mathfrak{A}_b} u_b \varphi_j(x) dx, \\ (M_{\mathcal{J}})_{ji} &= \int_{\mathcal{J}} \varphi_i(x) \varphi_j(x) dx, \end{aligned} \quad (4.20)$$

where φ_j denotes a basis function, $j = 1, \dots, N$. One important point in order to solve the equation system (4.19) is to integrate only the active or the inactive set of the interface elements, i.e. to calculate

$$\begin{aligned} \int_{K \cap \mathfrak{A}_a} \varphi_i(x) \varphi_j(x) dx & \quad \text{for } K \in \tau_h, \\ \int_{K \cap \mathfrak{A}_b} \varphi_i(x) \varphi_j(x) dx & \quad \text{for } K \in \tau_h, \\ \int_{K \cap \mathcal{J}} \varphi_i(x) \varphi_j(x) dx & \quad \text{for } K \in \tau_h. \end{aligned}$$

Assumption 4.24. *For simplicity it is assumed that on each interface element there is either an intersection with \mathfrak{A}_a or with \mathfrak{A}_b but not with both.*

Remark 4.25. *Assumption 4.24, can be encountered without the loss of generality by starting with a suitable mesh, since bang-bang control is not considered in this chapter.*

First, the shape of the interface between active and inactive set for polynomial degree $p = 1$ in case of a discretization with quadrilateral elements is investigated. For simplicity it is assumed to have affine-linear mappings, i.e. elements with constant determinant. Then it is sufficient to consider only the integration on the reference element \hat{K} . Moreover it is assumed to have constant bounds u_a, u_b .

Lemma 4.26. *Let u_N^* be the control of the discretized optimal control problem introduced in section 4.4 and let the polynomial degree be $p = 1$ on each interface element. Furthermore, let the constrains u_a and u_b be constant. Then the shape of the interface between the active or the inactive set is either a line, a hyperbola or non-existent.*

Proof. Since it is assumed to only have affine-linear mappings, it is sufficient to consider only the reference element. Each interface element has polynomial degree $p = 1$, therefore on the reference element it holds

$$\begin{aligned} u_N|_{\hat{K}}(\hat{x}) &= \sum_{i=1}^4 \varphi_i^{\mathcal{V}}(\hat{x}_1, \hat{x}_2) \mathbf{u}_i \\ &= \frac{1}{4} ((\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 + \mathbf{u}_4) + (-\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 - \mathbf{u}_4) \hat{x}_1 \\ &\quad + (-\mathbf{u}_1 - \mathbf{u}_2 + \mathbf{u}_3 + \mathbf{u}_4) \hat{x}_2 + (\mathbf{u}_1 - \mathbf{u}_2 + \mathbf{u}_3 - \mathbf{u}_4) \hat{x}_1 \hat{x}_2) \end{aligned}$$

where u_i are the values in the nodes of \hat{K} . To find the shape of the interface, it has to be checked if

$$u_a \leq u_N|_{\hat{K}}(\hat{x}) \leq u_b.$$

for all values of \hat{x} . This is analogue to check

$$\begin{aligned} a(\hat{x}) &= u_N|_{\hat{K}}(\hat{x}) - u_a \geq 0, \\ b(\hat{x}) &= u_b - u_N|_{\hat{K}}(\hat{x}) \geq 0, \end{aligned}$$

for the given bounds u_a, u_b . Without loss of generality in here only $a(\hat{x})$ is considered, the bound $b(\hat{x})$ is analogue. Since the bounds u_a and u_b are constant, $a(\hat{x})$ has the same form as $u_N|_{\hat{K}}(\hat{x})$, in fact it is given by

$$\begin{aligned} a(\hat{x}) &= \frac{1}{4}(a_1 + a_2 + a_3 + a_4) + (-a_1 + a_2 + a_3 - a_4)\hat{x}_1 \\ &\quad + (-a_1 - a_2 + a_3 + a_4)\hat{x}_2 + (a_1 - a_2 + a_3 - a_4)\hat{x}_1\hat{x}_2, \end{aligned}$$

where the coefficients a_i are given by $a_i = u_i - u_a$. To determine the curve $a(\hat{x}) = 0$ quadrics and the principal axis theorem are used. The goal now is to rewrite $a(\hat{x})$ for determining its form. First, $a(\hat{x})$ shall be written in the form

$$\hat{x}^\top A \hat{x} + v^\top \hat{x} = \mathbf{c},$$

where A is a symmetric matrix and $\hat{x} = (\hat{x}_1, \hat{x}_2)^\top$. By equating of coefficients, for the coefficients of the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and $v = (v_1, v_2)^\top$, it follows

$$a_{11} = 0, \quad a_{22} = 0, \quad a_{12} = a_{21} = \frac{1}{8}(a_1 - a_2 + a_3 - a_4),$$

and furthermore

$$v_1 = \frac{1}{4}(-a_1 + a_2 + a_3 - a_4), \quad v_2 = \frac{1}{4}(-a_1 - a_2 + a_3 + a_4), \quad \mathbf{c} = \frac{1}{4}(a_1 + a_2 + a_3 + a_4).$$

This matrix shall now be transformed in the form

$$y^\top B^\top A B y + v^\top B y = \mathbf{c} \quad (4.21)$$

to enable the determination of the curve type. For the transformation B is constructed to be orthogonal, i.e. $B^\top B = I$. In fact the matrix B will consist of the eigenvectors of A and the diagonal matrix $D = B^\top A B$, which enables the determination of the curve, consists of the eigenvalues of A . Next the eigenvalues and eigenvectors of B are determined. The eigenvalues of A are

$$\lambda_{1,2} = \pm \frac{1}{2}a_{12},$$

and $\hat{v}_1 = (1, 1)$ and $\hat{v}_2 = (-1, 1)$ are eigenvectors. Therefore it is set

$$B = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad D = \frac{1}{8} \begin{pmatrix} a_{12} & 0 \\ 0 & -a_{12} \end{pmatrix}.$$

Inserting in (4.21) leads to

$$\frac{1}{8}a_{12}y_1^2 - \frac{1}{8}a_{12}y_2^2 + \frac{1}{4\sqrt{2}}(a_1 + a_2)y_1 + \frac{1}{4\sqrt{2}}(-a_1 + a_2)y_2 = \mathbf{c}.$$

Afterwards, the curve is put into the origin, which leads to

$$\left(\frac{1}{8}\sqrt{a_{12}}y_1 + \frac{1}{4\sqrt{a_{12}}}(a_1 + a_2) \right)^2 - \left(\frac{1}{8}\sqrt{a_{12}}y_2 - \frac{1}{4\sqrt{a_{12}}}(-a_1 + a_2) \right)^2 = \mathbf{c} - \frac{a_1 a_2}{a_{12}} = \tilde{\mathbf{c}} \quad (4.22)$$

if $a_{12} > 0$. The case $a_{12} < 0$ is analogue, for $a_{12} = 0$ the curve reduces to

$$(a_1 + a_2)y_1 + (-a_1 + a_2)y_2 = 4\sqrt{2}\mathbf{c},$$

i.e. in this case the curve which separated the active and inactive set is a line. The translation

$$\begin{aligned} z_1 &= \frac{1}{\sqrt{8}}\sqrt{a_{12}}y_1 + \frac{1}{4\sqrt{a_{12}}}(a_1 + a_2), \\ z_2 &= \frac{1}{\sqrt{8}}\sqrt{a_{12}}y_2 - \frac{1}{4\sqrt{a_{12}}}(-a_1 + a_2), \end{aligned}$$

leads to

$$z_1^2 - z_2^2 = \tilde{\mathbf{c}},$$

i.e. the curve is now transformed to the first main diagonal form and therewith it follows that for $a(\hat{x}) = 0$ one gets a hyperbola if $a_{12} \neq 0$. In the case of $a_{12} = 0$ it is a line. By mapping the reference element to an arbitrary one, the results hold for all elements with affine-linear mapping to the reference square \hat{K} . \square

Due to the assumptions of theorem 4.26 on each element only active bounds u_a or active bounds u_b can occur, see figure 4.31 for examples of different shapes of the interface.

Remark 4.27. *The results can be extended to two different active bounds on one element.*

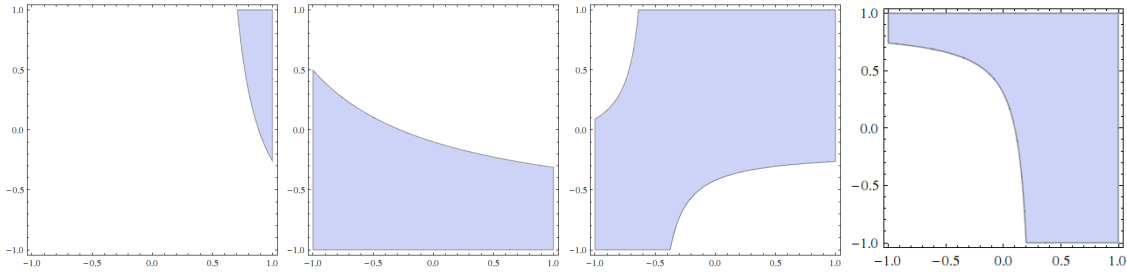


Figure 4.31: different shapes of active and inactive set in interface elements

The next question is how to integrate on these hyperbola parts. Theoretically it is possible to consider different cases and to find a closed form for each of it. However, since 73 cases occur [126], this method is not practical. Therefore the integration is done numerically.

Remark 4.28. For the practical realization it is important to have a low polynomial degree, i.e. $p = 1, 2$ on interface elements, since there an integration only over the active and inactive parts of the element has to be performed. Furthermore, it has to be ensured, that the interface does not move into elements with higher polynomial degree, because then it is more costly to check if it is an interface element or not.

Remark 4.29. For avoiding problems with the integration over parts of an element, triangular elements can be used. There the integral for $p = 1$ can be calculated exactly, since the triangle is separated in another triangle and a quadrangle due to the linear basis functions. In [142] there are even formulas to evaluate the integral for $p = 2$ exactly.

The drawback in numeric integration is that an additional error occurs. Then, in fact instead the (inner) equation system (4.19), in each Newton step the system

$$\left(\tilde{M}_{\mathcal{I}} + \frac{1}{\alpha} \tilde{M}_{\mathcal{I}} K_N^{-1} M_{\Omega} K_N^{-1} \tilde{M}_{\mathcal{I}} \right) \underline{\vec{u}} = \frac{1}{\alpha} \tilde{M}_{\mathcal{I}} K_N^{-1} M_{\Omega} (\vec{y}_d - K_N^{-1} \vec{f}) \quad (4.23)$$

is solved. For simplicity it is set

$$\begin{aligned} \mathfrak{M} &= \left(M_{\mathcal{I}} + \frac{1}{\alpha} M_{\mathcal{I}} K_N^{-1} M_{\Omega} K_N^{-1} M_{\mathcal{I}} \right), \\ \tilde{\mathfrak{M}} &= \left(\tilde{M}_{\mathcal{I}} + \frac{1}{\alpha} \tilde{M}_{\mathcal{I}} K_N^{-1} M_{\Omega} K_N^{-1} \tilde{M}_{\mathcal{I}} \right) \end{aligned}$$

and the right-hand sides are set to

$$\begin{aligned} \vec{g} &= \frac{1}{\alpha} M_{\mathcal{I}} K_N^{-1} M_{\Omega} (\vec{y}_d - K_N^{-1} \vec{f}), \\ \underline{\vec{g}} &= \frac{1}{\alpha} \tilde{M}_{\mathcal{I}} K_N^{-1} M_{\Omega} (\vec{y}_d - K_N^{-1} \vec{f}). \end{aligned}$$

In order to keep the overall convergence rate, the error between the exact solution \vec{u} and the perturbed solution $\underline{\vec{u}}$ has to be small. A suitable theorem therefore, is theorem 1.9, which is applied in the following.

Theorem 4.30. Let \vec{u}_* be the solution to (4.19) and $\underline{\vec{u}}_*$ the solution to the perturbed system (4.23). Moreover, let assumption

$$\|M_{\mathcal{I}} - \tilde{M}_{\mathcal{I}}\|_2 \leq \varepsilon \|M_{\mathcal{I}}\|_2 \quad \text{with } 0 < \varepsilon < 1 \quad (4.24)$$

be fulfilled. Then, for the error $\varepsilon_u := \frac{\|\vec{u}_* - \underline{\vec{u}}_*\|_2}{\|\vec{u}_*\|_2}$ it holds

$$\varepsilon_u \leq \frac{\kappa(\mathfrak{M})}{1 - \kappa(\mathfrak{M})\varepsilon_{\mathfrak{M}}} (\varepsilon_{\mathfrak{M}} + \varepsilon_g) \quad (4.25)$$

with

$$\begin{aligned} \varepsilon_g &= \varepsilon \cdot \kappa(M_{\mathcal{I}}) \\ \varepsilon_{\mathfrak{M}} &= \varepsilon \left(1 + \kappa(M_{\mathcal{I}}) + (1 + \varepsilon) (\kappa(M_{\mathcal{I}}))^2 \right). \end{aligned}$$

Proof. With assumption (4.24) an upper and lower bound for the norm $\|\tilde{M}_{\mathcal{J}}\|_2$ can be yielded. The upper one is derived by

$$\begin{aligned}\|\tilde{M}_{\mathcal{J}}\|_2 &= \|\tilde{M}_{\mathcal{J}} + M_{\mathcal{J}} - M_{\mathcal{J}}\|_2 \leq \|\tilde{M}_{\mathcal{J}} - M_{\mathcal{J}}\|_2 + \|M_{\mathcal{J}}\|_2 \\ &\leq \varepsilon \|M_{\mathcal{J}}\|_2 + \|M_{\mathcal{J}}\|_2,\end{aligned}$$

which yields

$$\|\tilde{M}_{\mathcal{J}}\|_2 \leq (1 + \varepsilon) \|M_{\mathcal{J}}\|_2. \quad (4.26)$$

The lower can be shown by

$$\begin{aligned}\|M_{\mathcal{J}}\|_2 &= \|M_{\mathcal{J}} + \tilde{M}_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}\|_2 \leq \|M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}\|_2 + \|\tilde{M}_{\mathcal{J}}\|_2 \\ &\leq \varepsilon \|M_{\mathcal{J}}\|_2 + \|\tilde{M}_{\mathcal{J}}\|_2,\end{aligned}$$

which leads to

$$\|\tilde{M}_{\mathcal{J}}\|_2 \geq (1 - \varepsilon) \|M_{\mathcal{J}}\|_2. \quad (4.27)$$

Then, the right hand side estimate in order to obtain ε_g is

$$\begin{aligned}\varepsilon_g &= \frac{\|\frac{1}{\alpha}(M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}})(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2}{\|\frac{1}{\alpha}M_{\mathcal{J}}(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2} \\ &\leq \frac{\|M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}\|_2 \|M_{\mathcal{J}}^{-1}M_{\mathcal{J}}(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2}{\|M_{\mathcal{J}}(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2} \\ &\leq \frac{\|M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}\|_2 \|M_{\mathcal{J}}^{-1}\|_2 \|M_{\mathcal{J}}(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2}{\|M_{\mathcal{J}}(K_N^{-1}M_{\Omega}(\tilde{y}_d - K_N^{-1}\tilde{f}))\|_2} \\ &\stackrel{(4.24)}{\leq} \varepsilon \|M_{\mathcal{J}}\| \|M_{\mathcal{J}}^{-1}\| = \varepsilon \cdot \kappa(M_{\mathcal{J}}).\end{aligned}$$

In order to estimate $\mathfrak{M} - \tilde{\mathfrak{M}}$, it is rewritten by

$$\begin{aligned}\mathfrak{M} - \tilde{\mathfrak{M}} &= (M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}) + \frac{1}{\alpha}M_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}} - \frac{1}{\alpha}\tilde{M}_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}\tilde{M}_{\mathcal{J}} \\ &= (M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}) + \frac{1}{\alpha}M_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}} - \frac{1}{\alpha}\tilde{M}_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}\tilde{M}_{\mathcal{J}} \\ &\quad + \frac{1}{\alpha}\tilde{M}_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}} - \frac{1}{\alpha}\tilde{M}_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}} \\ &= (M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}}) + \frac{1}{\alpha}(M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}})K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}} + \frac{1}{\alpha}\tilde{M}_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}(M_{\mathcal{J}} - \tilde{M}_{\mathcal{J}})\end{aligned}$$

With two further inequalities, i.e.

$$\|M_{\mathcal{J}}\|_2 \leq \|M_{\mathcal{J}} + \frac{1}{\alpha}M_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}}\|_2 \quad (4.28)$$

$$\|\frac{1}{\alpha}M_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}}\|_2 \leq \|M_{\mathcal{J}} + \frac{1}{\alpha}M_{\mathcal{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathcal{J}}\|_2 \quad (4.29)$$

the overall bound for $\varepsilon_{\mathfrak{M}}$ can then be yielded by using the triangle inequality

$$\begin{aligned} \varepsilon_{\mathfrak{M}} &= \frac{\|(M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}}) + \frac{1}{\alpha}(M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}})K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}} + \frac{1}{\alpha}\tilde{M}_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}(M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}})\|_2}{\|M_{\mathfrak{J}} + \frac{1}{\alpha}M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} \\ &\leq \frac{\|M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}}\|_2 + \|\frac{1}{\alpha}(M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}})K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2 + \|\frac{1}{\alpha}\tilde{M}_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}(M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}})\|_2}{\|M_{\mathfrak{J}} + \frac{1}{\alpha}M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} \end{aligned}$$

By using the estimates (4.28) and (4.29) to estimate the denominator, it follows

$$\begin{aligned} \varepsilon_{\mathfrak{M}} &\leq \frac{\|M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}\|_2} + \frac{\|M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}}\|_2\|K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} + \frac{\|\tilde{M}_{\mathfrak{J}}\|_2\|K_N^{-1}M_{\Omega}K_N^{-1}\|_2\|M_{\mathfrak{J}} - \tilde{M}_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} \\ &\leq \frac{\varepsilon\|M_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}\|_2} + \frac{\varepsilon \cdot \kappa(M_{\mathfrak{J}})\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} \\ &\quad + \frac{\varepsilon \cdot \kappa(M_{\mathfrak{J}})\|\tilde{M}_{\mathfrak{J}}\|_2\|M_{\mathfrak{J}}^{-1}\|_2\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2}{\|M_{\mathfrak{J}}K_N^{-1}M_{\Omega}K_N^{-1}M_{\mathfrak{J}}\|_2} \\ &\leq \varepsilon + \varepsilon \cdot \kappa(M_{\mathfrak{J}}) + \varepsilon \cdot \kappa(M_{\mathfrak{J}})\|\tilde{M}_{\mathfrak{J}}\|_2\|M_{\mathfrak{J}}^{-1}\|_2 \\ &\leq \varepsilon(1 + \kappa(M_{\mathfrak{J}}) + (1 + \varepsilon)(\kappa(M_{\mathfrak{J}}))). \end{aligned}$$

□

Theorem 4.30 states that the error of the disturbed inner equation system in the semismooth Newton method depends on the parameter ε and the condition number $\kappa(M_{\mathfrak{J}}) = \|M_{\mathfrak{J}}\|_2\|M_{\mathfrak{J}}^{-1}\|_2$. Therewith, the minimal and maximal eigenvalue of $M_{\mathfrak{J}}$ are necessary to estimate the error. That is expected to cause problems in case of very small parts of the inactive set \mathfrak{J} in elements, since then, the minimal eigenvalue tends to zero.

Nevertheless, not only the error of the (inner) equation system (4.19) has to be considered but also its effect to the (outer) semismooth Newton method. Possible choices are to consider the semismooth Newton method as inexact semismooth Newton method (see e.g. [160]).

However, first a numeric integration scheme is chosen. In this thesis the generalized mid point rule (see e.g. [82]) is used. By rewriting the integral over the inactive set on the reference element \hat{K} ,

$$\int_{\mathfrak{J} \cap \hat{K}} \mathfrak{g}(\hat{x}_1, \hat{x}_2) d\hat{x} = \int_{-1}^1 \int_{-1}^1 \mathfrak{g}(\hat{x}_1, \hat{x}_2) \chi_{\mathfrak{J}}(\hat{x}_1, \hat{x}_2) d\hat{x}$$

with the characteristic set

$$\chi_{\mathfrak{J}}(\hat{x}_1, \hat{x}_2) = \begin{cases} 1 & (\hat{x}_1, \hat{x}_2) \in \mathfrak{J} \\ 0 & \text{else} \end{cases}$$

the numeric integration can be performed by computing

$$\int_{-1}^1 \int_{-1}^1 \mathfrak{g}(\hat{x}_1, \hat{x}_2) \chi_{\mathfrak{J}}(\hat{x}_1, \hat{x}_2) d\hat{x} \approx \mathfrak{h}_n^2 \sum_{i,j=1}^n \mathfrak{g}(\mathfrak{x}_i, \mathfrak{x}_j) \chi_{\mathfrak{J}}(\mathfrak{x}_i, \mathfrak{x}_j).$$

There, $\mathfrak{h}_n = \frac{2}{n}$ denotes the mesh-width of the numeric integration and $\mathfrak{x}_i = -1 + \mathfrak{h}_n \left(i - \frac{1}{2}\right)$ for $i = 1, \dots, n$ are the integration points. n is a parameter in order to make the uniform grid for the numeric integration finer or coarser. The mesh-width \mathfrak{h}_n of the numeric integration is especially independent of the mesh size of the element in the mesh.

In order to get general error estimates, first the error of that numeric integration has to be estimated. However, due to the non-continuity of the integrand, the usual error estimates of numeric integration with the generalized mid point rule are not applicable. Estimates which use the supremum of $\mathbf{g}(\hat{x}_1, \hat{x}_2)\chi_{\mathfrak{J}}(\hat{x}_1, \hat{x}_2)$ in combination with the mesh-width h_n , seems to be too coarse, especially since $\mathbf{g}(\hat{x}_1, \hat{x}_2)\chi_{\mathfrak{J}}(\hat{x}_1, \hat{x}_2)$ is unknown and might get very large.

Remark 4.31. *Further possibilities to estimate the overall error are using general approaches to estimate the integration error by using Strang's Lemma (see e.g. [51, 76]). There, usually the Bramble-Hilbert lemma is used. In order to apply that lemma, the numeric integration has to be exact for a given polynomial degree p . However, in case of curves as in figure 4.31, this is not possible for the generalized mid point rule. A possible integration rule which is exact for constant functions ($p = 0$) is*

$$\int_{\mathfrak{J}} \mathbf{g}(\hat{x}_1, \hat{x}_2) dx \approx \sum_{i=1}^n \omega_i \mathbf{g}(\hat{x}_{1i}, \hat{x}_{2i}) \quad (4.30)$$

with the weights $\omega_i = \frac{\text{area}(\mathfrak{J})}{n}$ and suitable integrations points $(\hat{x}_{1i}, \hat{x}_{2i}) \in \mathfrak{J}$ for $i = 1, \dots, n$. Again n is a parameter in order to make the integration routine better. Nevertheless, in case of applying the integration rule (4.30) the best expected convergence rate is h , which is less than the yielded one for the unperturbed system with convergence rate two.

Remark 4.32. *If the perturbations are small enough, it might even be possible to apply [85, theorem 3.4]. Nevertheless, also in this case the error which occurs due to the numeric integration has to be estimated.*

Remark 4.33. *Even if an error estimate is missing – in the yield numerical examples the generalized mid point rule is used. The given results show, that it works quite well and especially does not decrease the overall convergence rate.*

4.4.3 Suitable starting mesh

Another important part when using the Newton algorithm is to choose a suitable starting mesh. This is especially important since it is assumed to have only the interface between the active set \mathfrak{A}_a and the interface \mathfrak{J} or the active set \mathfrak{A}_b and the interface \mathfrak{J} on one element. Therewith, before starting the Newton algorithm such a mesh has to be found, see Algorithm 11. After having a suitable starting mesh, the semismooth Newton algorithm (algorithm 8)

Algorithm 11: calculate starting mesh

input : starting value $u_0 \in U$

output: adjoint q

compute adjoint q for u_0 (for calculating active and inactive sets)

while $\exists K : u_{K_1} = u_a$ and $u_{K_2} = u_b$ for $K_1, K_2 \subset K$ **do**

find all elements K with $u_{K_1} = u_a$ and $u_{K_2} = u_b$ where $K_1, K_2 \subset K$

do h -refinement for these elements

map q with algorithm 6 to new mesh and set constraints $\rightarrow u_{new}$

compute adjoint q for u_{new}

is used to solve the distributed optimal control problem. For constant constraints u_a and u_b , as starting vector \vec{u}_0 in the first step the i - component of the vector

$$(u_0)_i = \frac{u_a + u_b}{2}$$

and in all further refinement steps, the solution of the former mesh is used. This is realized by projecting the adjoint after solving the system and refining the mesh to the new mesh by algorithm 6.

Remark 4.34. *The given procedure with the starting mesh only makes sense if the adjoint is not oscillating. In such a case the mesh is already quite fine, therewith uniform h-fem or possibly adaptive h-fem is proposed.*

4.4.4 Numerical examples

In this section several distributed optimal control problems are solved to show that the proposed hp -refinement strategies are working, but also to confirm the expectation that hp -refinement reduces the number of degrees of freedom.

Although, there are no theoretical estimates on this topic to the best knowledge of the author, a reduction of degrees of freedom will simplify and accelerate the simulation of distributed optimal control problems significantly.

Remark 4.35. *In the plots of the solution only the values at the nodes are used for plotting. However, for the calculation of the solution, hp -refinement is used as it is for the calculation of the L_2 -error.*

4.4.4.1 Example Square

The first example is taken from [158], see also [166] where the interior point method instead of the semismooth Newton method is used to solve the problem. It is given by

$$\min J(y, u) := \frac{1}{2} \|y - y_d\|_{L_2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L_2(\Omega)}^2 + \int_{\Gamma} e_q y \, ds_x$$

subject to

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) + f(x) && \text{in } \Omega \\ \frac{\partial y}{\partial n}(x) &= 0 && \text{on } \Gamma = \partial\Omega \end{aligned} \tag{4.31}$$

and the box constraints on the control

$$u_a \leq u(x) \leq u_b \quad \text{a.e. in } \Omega.$$

The desired state y_d , the inhomogenities e_y , e_q , the right-hand-side f , the box constraints u_a, u_b and the regularization parameter α are given. In the following examples the regularization parameter α is set to 1.

By introducing the adjoint state q given by

$$\begin{aligned} -\Delta q(x) + q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ \frac{\partial q}{\partial n}(x) &= e_q(x) && \text{on } \Gamma, \end{aligned} \tag{4.32}$$

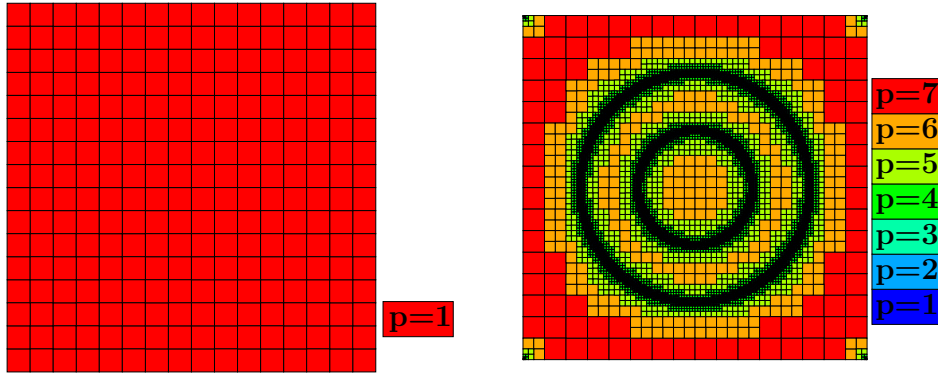


Figure 4.32: initial mesh and nic-refined mesh for example 4.4.4.1

in order to calculate the solution to the problem, the equation system of the primal equation (4.31), the dual equation (4.32) and the projection formula

$$u(x) = -P_{[0,0,1,0]}(q(x))$$

has to be solved (according to theorem 2.4 and theorem 2.5).

As solution to the problem

$$\begin{aligned} r(x) &= (x_1 - 0.5)^2 + (x_2 - 0.5)^2, \\ q(x) &= -12r(x) + \frac{1}{3}, \\ y(x) &= 1.0, \end{aligned}$$

is used with the domain $(0, 1)^2$. The data is therewith implicitly given. Since the polynomial degree is only two, this example is mainly considered in order to give a first glance on the suggested refinement and enable furthermore a direct comparison with the results in [158, 166]. The initial mesh obtained with algorithm 11 from a mesh with nine nodes is given in figure 4.32. There, the polynomial distribution after some nic-refinement steps is given. A comparison with the results in [166, figure 6.1] where interior point methods are used for the refinement, shows similar results.

In figure 4.33 the L_2 error with respect to the mesh size h and the number of degrees of freedom N is given. The expected convergence rate given by the theory is two for a uniform h -refinement. The right plot in figure 4.33 shows that for both refinements a convergence rate of two is yielded, i.e. the convergence rate from theory is yielded. However, a comparison of the L_2 error with respect to the number of degrees of freedom (right plot in figure 4.33) shows, that the suggested application of a suitable hp -refinement, the nic-refinement, leads to better results, i.e. in the case of an L_2 error about $4.0e - 5$, uniform h -refinement needs about 300 000 degrees of freedom, whereas nic-refinement only needs about 40 000 degrees of freedom.

In figure 4.34 the state and the adjoint are plotted for nic-refinement, in figure 4.35 the control u is given. These plots show, that the state y is constant, i.e. $y(x) = 1.0$. A comparison between the adjoint in figure 4.34 and in figure 4.35 for the control u shows, on the one side, that there are both active and inactive parts for the control u . Moreover, it can be seen, that the control is in fact the truncated adjoint q if the values of the adjoint are not in between the

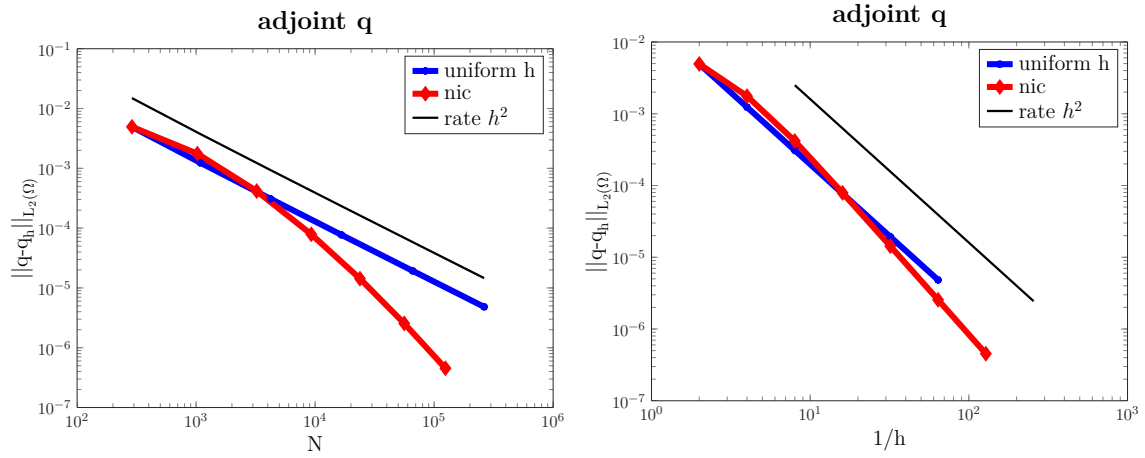


Figure 4.33: comparison of uniform h - and nic-refinement for example (4.4.4.1)

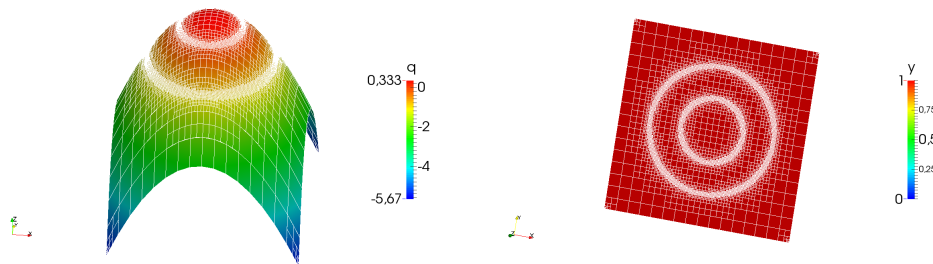


Figure 4.34: adjoint q and state y for example 4.4.4.1

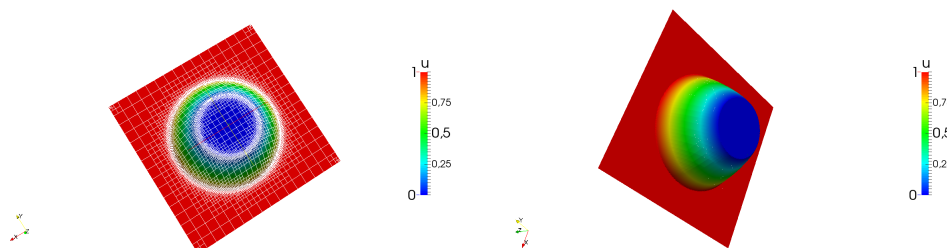


Figure 4.35: control u for example 4.4.4.1

box constraints. A glance at the refinement, i.e. the distribution between elements with low polynomial degree (and small mesh size), and high polynomial degree (and big mesh size), shows, that the interface elements, i.e. the elements that have active and inactive parts, are exactly at the set of the cuts. This shows that the refinement works as expected.

4.4.4.2 Example Sinus

As second example, an example with known analytical solution but higher polynomial degree in the solution, is considered. Again, the domain is a square, however now it is chosen to be $(-1, 1)^2$. It is assumed to have homogeneous Dirichlet boundary on the whole boundary.

The problem is given by

$$\min J(y, u) := \frac{1}{2} \|y - y_d\|_{L_2(\Omega)}^2 + \frac{1}{2} \|u\|_{L_2(\Omega)}^2$$

subject to the primal equation

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) + f(x) && \text{in } \Omega \\ y(x) &= 0 && \text{on } \Gamma = \partial\Omega \end{aligned} \quad (4.33)$$

Additionally for the control there hold the box constraints

$$-0.7 \leq u \leq 0.75 \quad \text{a.e. in } \Omega$$

According to theorem 2.4 and theorem 2.5, the problem can even be described by the primal equation (4.33), the adjoint equation

$$\begin{aligned} -\Delta q(x) + q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ q(x) &= 0 && \text{on } \Gamma \end{aligned}$$

and the projection formula

$$u(x) = P_{[-0.7, 0.75]}(-q(x)).$$

The solution is given by

$$\begin{aligned} y(x) &= (x_1^2 - 1)(x_2^2 - 1) \sin(\pi x_1) \cos(\pi x_2), \\ q(x) &= (x_1^2 - 1)(x_2^2 - 1) \sin(\pi x_1). \end{aligned}$$

First, the obtained convergence rate is considered. In figure 4.36 it can be observed, that both – uniform h - and nic-refinement – yield the same rate of convergence, i.e. a convergence rate of two. Therewith the theoretical results for uniform h -fem are confirmed.

However, a comparison of the L_2 error versus the number of degrees of freedom for uniform h -refinement and nic-refinement is given in figure 4.37 and shows the advantage of the nic-refinement. Even though the convergence rate is two in both cases, the choice of the nic-refinement leads to a substantially decrease of the number of degrees of freedom and faster convergence.

Second, a comparison between the two proposed refinements, the nic-refinement and the refinement with additional use of error estimators, called errest-refinement (see algorithm 10), is used. The results are given in figure 4.38, where the expected convergence rate for

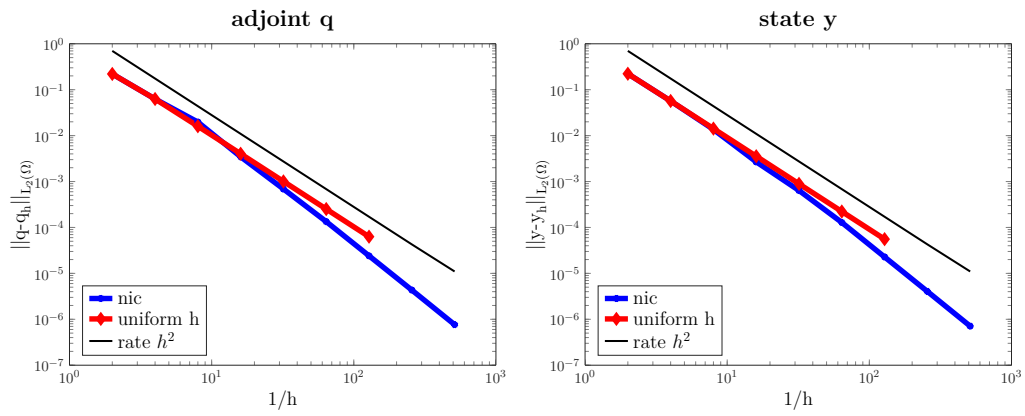


Figure 4.36: comparison of h - and nic-refinement for example 4.4.4.2

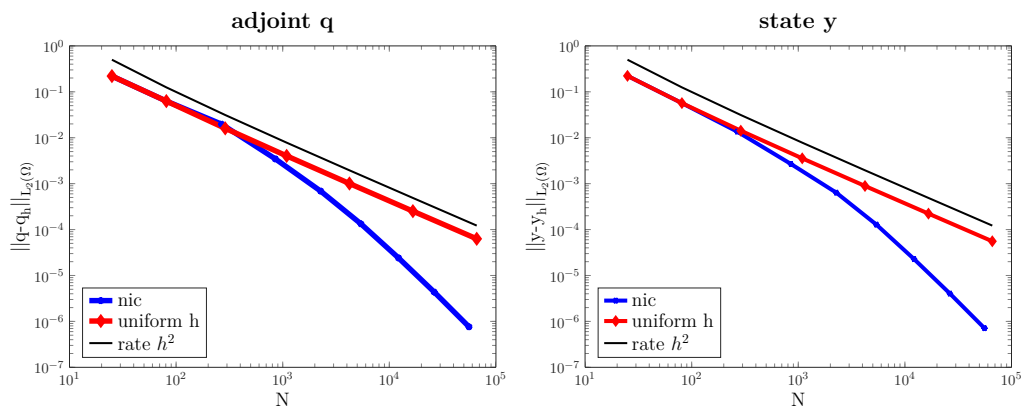


Figure 4.37: comparison of h - and nic-refinement for example 4.4.4.2

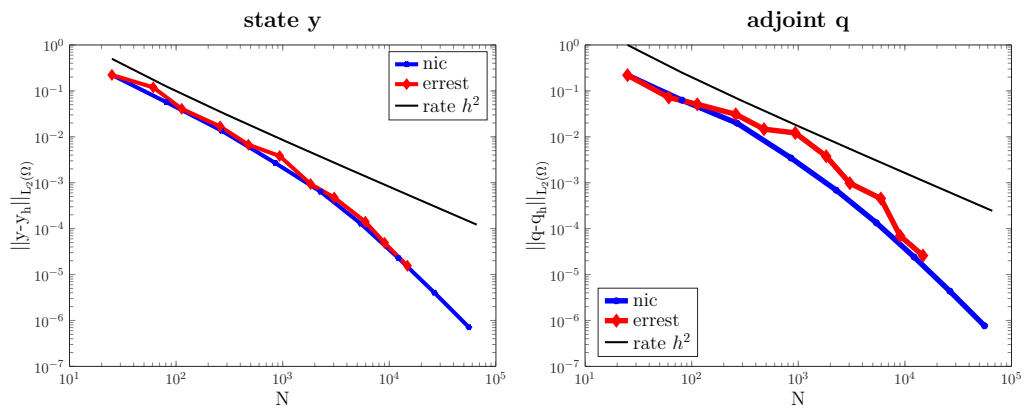


Figure 4.38: comparison of nic- and errest- refinement for 4.4.4.2

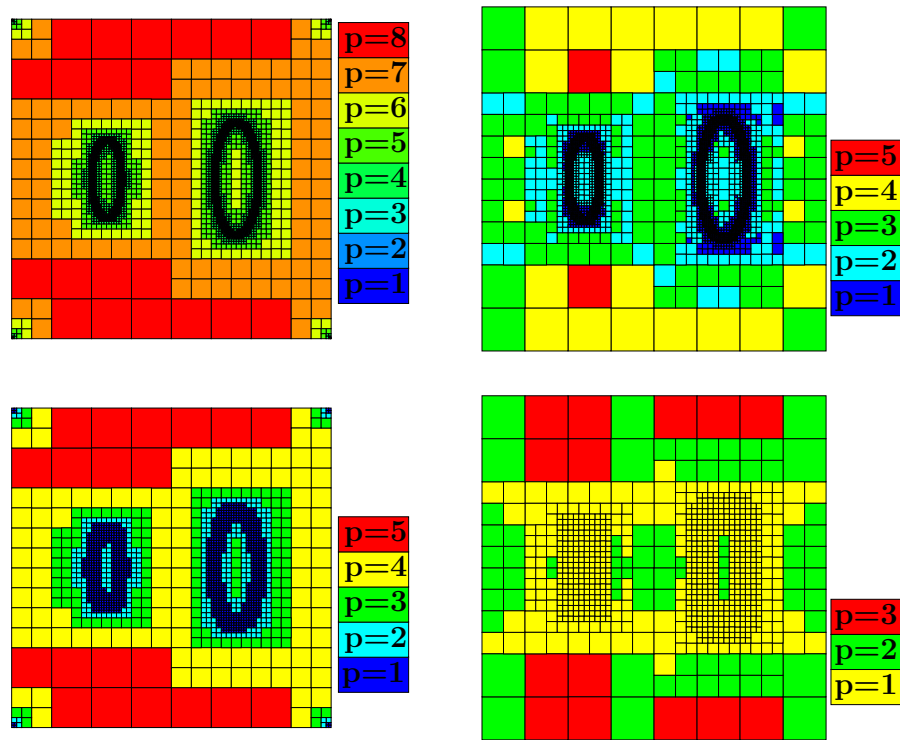


Figure 4.39: meshes yielded with nic-refinement (left), meshes yielded with errest-refinement (right) for example 4.4.4.2

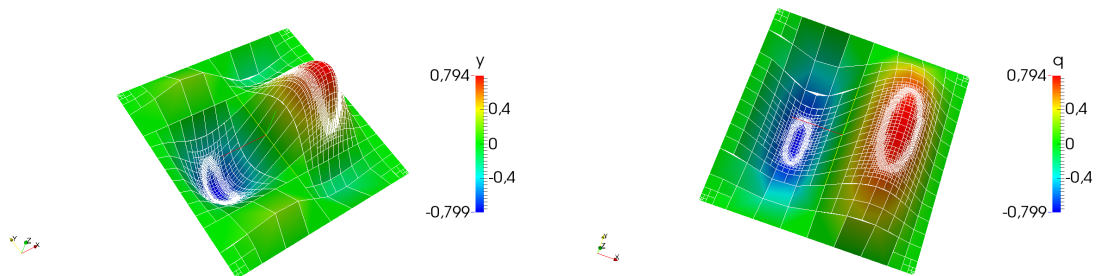


Figure 4.40: state y and adjoint q with nic-refinement for example 4.4.4.2

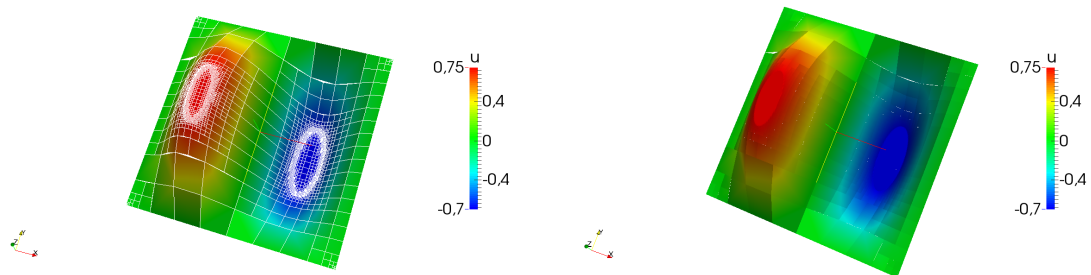


Figure 4.41: control u with nic-refinement for example 4.4.4.2

uniform h -refinement is plotted in black. It has to be stated that both refinements lead to better results than a uniform h -refinement would yield. Furthermore, it can be observed that both refinements yield similar results, but the L_2 error with respect to N decreases faster for the state y . This effect occurs, since the primal equation, i.e. the state y , is used in algorithm 10.

Remark 4.36. *The choice of using error estimators for the state y instead of the adjoint q in here, is motivated by expecting to get broader information on the regularity of the whole problem, since the adjoint is already used in the projection formula. Different choices and especially the combination of using error estimators for the state and the adjoint are possible.*

The polynomial distribution of the meshes for nic- and errest-refinement are given in figure 4.39. These polynomial distributions (in fact the polynomial distribution of the elements, since the polynomial distribution of the edges is determined by the minimal degree condition) show that for both refinements, the elements' size near the interface decreases, whereas the polynomial degree increases if the distance to the interface gets bigger. The differences between the two refinements is mainly, that in case of nic-refinement all corner elements are h -refined, which is not the case for errest-refinement. The reason therefore is, that there occur no singularities in the corners.

Remark 4.37. *In the examples in here a difference of two for the polynomial degrees in the errest-refinement is allowed.*

The solution – calculated with nic-refinement – can be found in figure 4.40, the corresponding control u is given in figure 4.41. There, it can be observed that again both constraints are active. The set \mathfrak{A}_a is located at the bigger curve – similar to an ellipse – which is plotted in blue in figure 4.41. The smaller ellipse – plotted in red in figure 4.41 shows the set \mathfrak{A}_b . Outside of these curves, the box constraints are inactive, i.e. are contained in the set \mathfrak{J} .

4.4.4.3 Example Hole

Next, an example with more complex geometry and unknown solution is given (see figure 4.42). The domain is a subset of the square $(0, 3)^2$, i.e. that set without the set $\{(x_1, x_2) \mid 1 \leq x_i \leq 2, i = 1, 2\}$. The primal and the dual problem are given by

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) && \text{in } \Omega \\ y(x) &= 0 && \text{on } \Gamma = \partial\Omega \\ -\Delta q(x) + q(x) &= y(x) - y_d(x) && \text{in } \Omega \\ q(x) &= 0 && \text{on } \Gamma. \end{aligned}$$

The control is given by the projection formula

$$u(x) = P_{[-0.3, 0.95]}(-q(x))$$

and the desired state is given by

$$y_d(x) = 10 \sin(\pi x_1) + 5 \cos(\pi x_2^2).$$

The corresponding meshes are given in figure 4.42 for nic- and errest-refinement. As in the example before, these meshes, which show the polynomial distribution of the elements, are

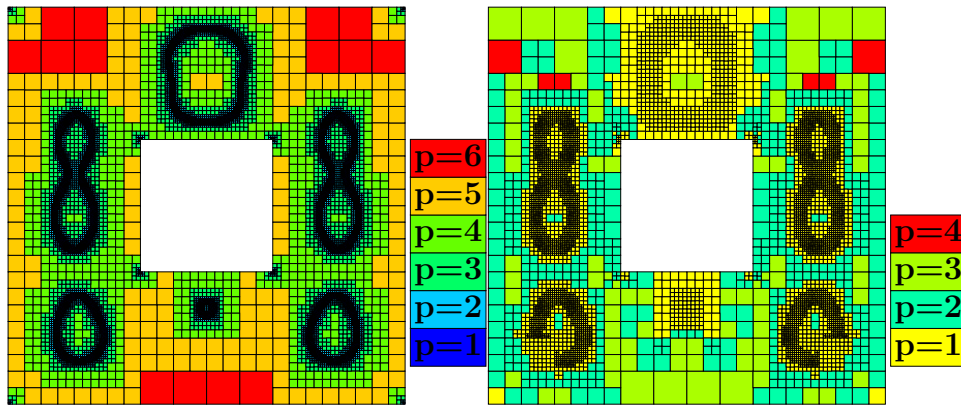


Figure 4.42: meshes for nic- and errest- refinement for example 4.4.4.3

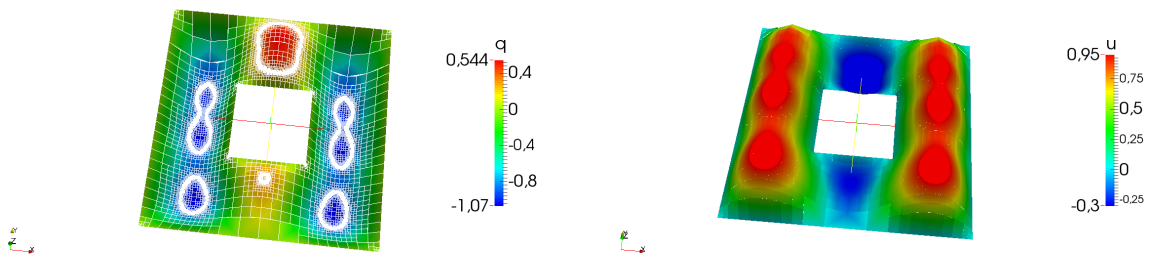


Figure 4.43: adjoint q (left) and control u (right) for nic-refinement for example 4.4.4.3

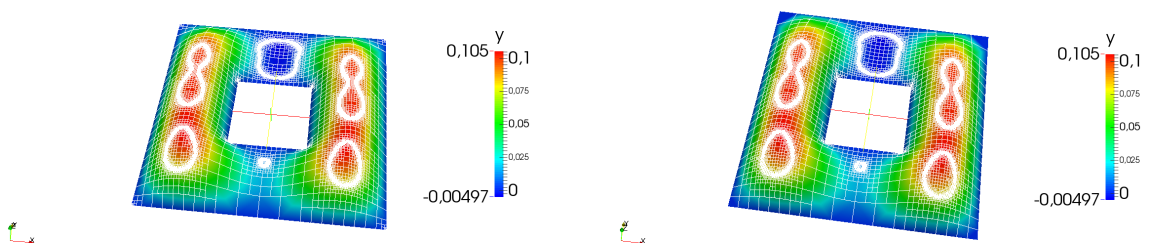


Figure 4.44: state y for nic-refinement (left) and errest-refinement (right) for example 4.4.4.3

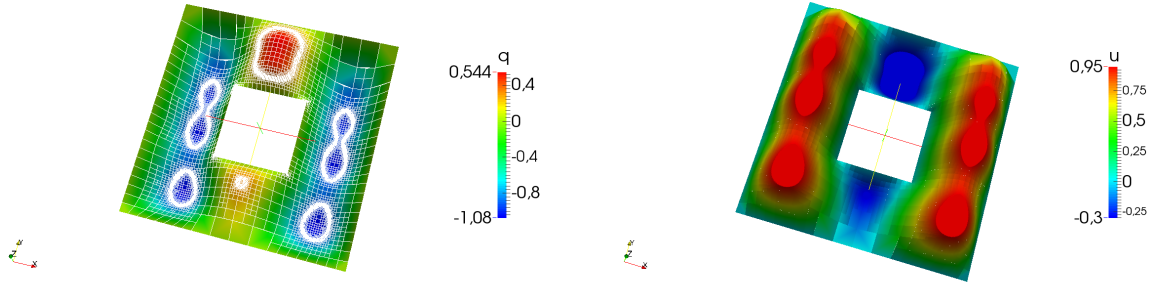


Figure 4.45: adjoint q (left) and control u (right) for errest-refinement for example 4.4.4.3

quite similar. It can be observed, that the interfaces between the active and inactive set are in the same places. In the case of nic-refinement in all corners h -refinement is performed, whereas in errest-refinement only L -shape corners are h -refined.

For both refinements the solution is calculated. Figure 4.43 shows the adjoint q and the control u for the nic-refinement, in figure 4.44 the corresponding state y is plotted. The state for errest-refinement is plotted there too, the adjoint and the control for errest-refinement can be found in figure 4.45.

For a description of the results, the concentration is on the nic-refinement (interpretation of errest-refinement leads to analog considerations). As expected the control u is between -0.3 and 0.95 . The round curves and the 8-shapes show the interface between the active and inactive sets. The active sets \mathfrak{A}_a are in blue (for the control u), the active sets \mathfrak{A}_b are the truncated red hills in figure 4.43. Again the mesh size (and the polynomial degree) near the interface is small, whereas it gets bigger if the distance to it increases. Moreover, the polynomial degree increases for elements who are far from the interface (i.e. on active or inactive elements) and for elements which do not live near corners of the domain.

4.4.4.4 Example Double L

Last, a further example with unknown solution is considered. Its geometry is given in figure 4.46. The primal and the dual problem are given by

$$\begin{aligned}
 -\Delta y(x) + y(x) &= u(x) && \text{in } \Omega \\
 y(x) &= 0 && \text{on } \Gamma = \partial\Omega \\
 -\Delta q(x) + q(x) &= y(x) - y_d(x) && \text{in } \Omega \\
 q(x) &= 0 && \text{on } \Gamma.
 \end{aligned}$$

The control is given by the projection formula

$$u(x) = P_{[0.5, 7.7]}(-q(x))$$

and the desired state is given by

$$y_d(x) = x_1^2 + (x_2 + 1)^3(x_1 + 1) + \frac{1}{3}e^{x_1+x_2}.$$

Again, the example is discretized and calculated with nic- and errest-refinement. The corresponding meshes are given in figure 4.46. These refinements show that again for both

refinements the very small elements with low polynomial degree lie at the interface (compare with figure 4.47). As in the plots before, in case of nic-refinement all corners are h -refined, whereas this is not the case for errest-refinement.

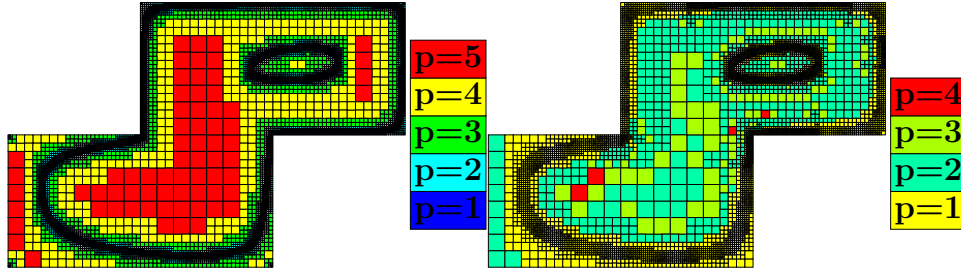


Figure 4.46: mesh with nic- and errest- refinement for example 4.4.4.4

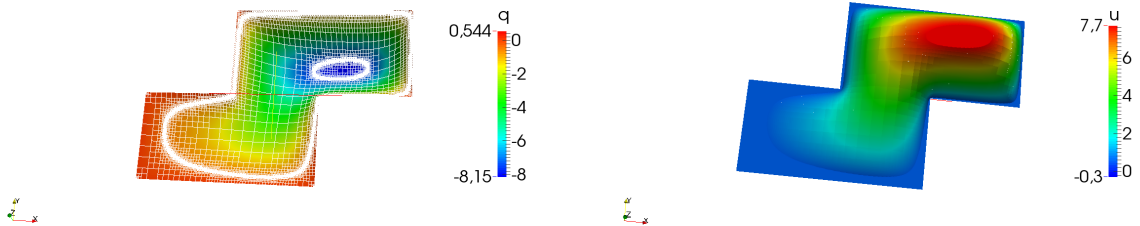


Figure 4.47: adjoint q (left) and control u (right) for nic-refinement for example 4.4.4.4

The adjoint and the control for nic-refinement can be found in figure 4.47, the corresponding state is plotted in figure 4.48. Here, the state for errest-refinement is also given. The control and the adjoint for errest-refinement are denoted in figure 4.49.

These plots show that the adjoint q is between $[-8.15, 0.544]$ and the state y is in between $[-0.00497, 0.854]$. Moreover, the plot for the control shows that there are both active parts \mathfrak{A}_a and active parts \mathfrak{A}_b . The set \mathfrak{A}_a is mainly on the boundary (in the plot it is blue) whereas the active set \mathfrak{A}_b is plotted in red.

Remark 4.38. *The numerical results show that the proposed refinement strategies work quite well and lead indeed to a faster convergence than uniform h -refinement. For the presented results no huge difference between nic-refinement and errest-refinement can be observed. It is expected that nic-refinement leads to better results in case of very smooth adjoint and state, whereas errest-refinement might be better otherwise.*

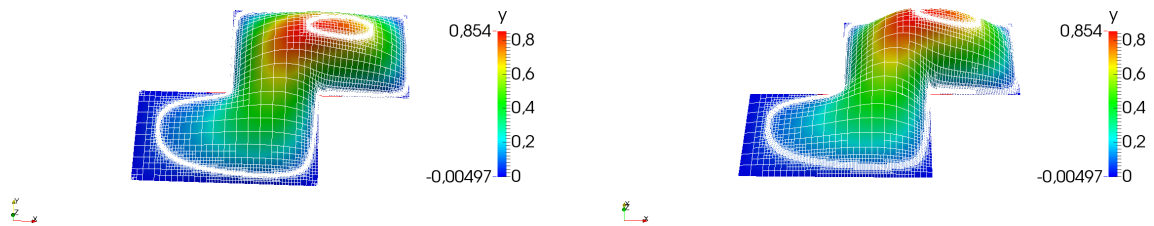


Figure 4.48: state y for nic-refinement (left) and errest-refinement (right) for example 4.4.4.4

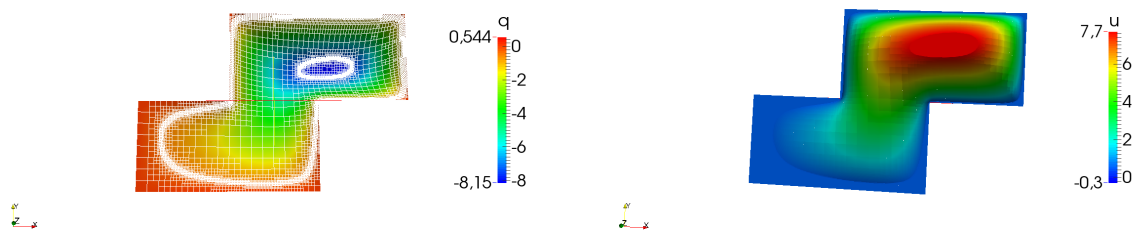


Figure 4.49: adjoint q (left) and control u (right) for errest-refinement for example 4.4.4.4

5 Saddle point problem

The discretization of the distributed optimal control problem (or the optimal boundary control problem) yields a system of (nonlinear) algebraic equations. Using semismooth Newton, a system of linear algebraic equations of the form (4.16) or (4.19) has to be solved in each Newton step. This is performed by a conjugate gradient method or an alternative iterative method. Note that each iteration requires a matrix-vector-multiplication with the matrix in (4.16) or (4.19). This involves multiplications with the mass matrix and the inverse of the stiffness matrix. In two dimensions the fast inversion of the stiffness matrix can be performed by direct methods in almost optimal arithmetical complexity, see [97], where such a solver was developed for boundary concentrated fem. However, these direct solvers are too expensive in three dimensions. Therewith, the focus is now on an alternative solution method, which is based on a rewriting of the problem in a saddle point formulation.

This chapter is structured as follows: First, the general saddle point formulation considered later on is given. Since three preconditioned Krylov subspace methods are used to solve the problem, a short introduction and especially suitable preconditioners for the saddle point formulation are given. Afterwards, these general considerations are applied to a simplified optimal control problem for all three methods. Furthermore, estimates on the condition number and the dependence on problem dependent parameters and discretization parameters – if possible – are investigated. Finally, numerical examples are presented in order to confirm theoretical results.

The linear equation system in saddle point formulation

$$\begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} \vec{x} \\ \vec{q} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{s} \end{pmatrix} \quad (5.1)$$

is considered under the following assumptions:

Assumption 5.1. *Let $A \in \mathbb{R}^{N_A \times N_A}$ be a symmetric and positive semidefinite matrix, $B \in \mathbb{R}^{N_B \times N_A}$ has full rank $N_B \leq N_A$. Moreover, it is set $N = N_A + N_B$ and it holds*

$$\langle A\vec{x}, \vec{x} \rangle > 0 \quad \text{for all } \vec{x} \in \ker B \text{ with } \vec{x} \neq \vec{0}.$$

Such systems also result from the discretization of mixed variational problems for systems of partial differential equations (see e.g. [44]). In particular such problems arise from the discretization of optimization problems with partial differential equation constraints.

The equation system (5.1) under the assumptions 5.1 can be interpreted as Karush-Kuhn-Tucker (KKT) conditions of the optimization problem

$$\min_{\vec{x}} \frac{1}{2} \langle A\vec{x}, \vec{x} \rangle - \langle \vec{f}, \vec{x} \rangle \quad \text{subject to the constraints} \quad B\vec{x} = \vec{s}$$

with associated Lagrangian parameter \vec{q} (see e.g. [67]).

For simplicity, it is set

$$\mathcal{A} = \begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix},$$

i.e. the equation system (5.1) can be rewritten as

$$\mathcal{A}\vec{z} = \vec{g}. \quad (5.2)$$

There, $\mathcal{A} \in \mathbb{R}^{N \times N}$ and \vec{z}_* denotes the exact solution.

Next, suitable preconditioners for the Krylov subspace methods introduced in section 1.2 are investigated. For a general overview of the solution to saddle point problems, see [27] and the references therein.

5.1 Schöberl-Zulehner PCG

First, a PCG method is applied to the considered saddle point problem (5.1). It has to be stated that usually – if possible – the PCG method is preferred. There are several methods to apply a modified CG when dealing with non-symmetric or non-positive definite matrices, see e.g. [38, 39, 95]. In this thesis, the modified PCG method by Schöberl and Zulehner [139] is used, since a direct application of the CG is impossible due to the indefiniteness of \mathcal{A} . The Schöberl-Zulehner PCG method is based on choosing a suitable preconditioner in order to apply the CG method with respect to a non-standard scalar product. As an introduction, the most important facts of the PCG method by Schöberl and Zulehner, see [139], are recalled. In this section the preconditioner for \mathcal{A} in (5.2), denoted by \mathcal{P}_{cg} , is chosen as

$$\mathcal{P}_{cg} = \begin{pmatrix} \hat{A} & B^\top \\ B & B\hat{A}^{-1}B^\top - \hat{S} \end{pmatrix},$$

where \hat{A} and \hat{S} are symmetric and positive definite matrices with respect to the standard scalar product. \mathcal{P}_{cg} is a well-known class of preconditioners, see e.g. [24]. Moreover, the following theorem holds:

Theorem 5.2. [139, Theorem 2.1] *Let assumption 5.1 and the relations $\hat{A} > 0$ and $\hat{S} > 0$ hold.*

1. *If*

$$\hat{A} \geq A \text{ and } \hat{S} \leq B\hat{A}^{-1}B^\top, \quad (5.3)$$

then all eigenvalues of $\mathcal{P}_{cg}^{-1}\mathcal{A}$ are real and positive.

2. *If*

$$\hat{A} > A \text{ and } \hat{S} < B\hat{A}^{-1}B^\top, \quad (5.4)$$

then $\mathcal{P}_{cg}^{-1}\mathcal{A}$ is symmetric and positive definite with respect to the scalar product

$$\left\langle \begin{pmatrix} \vec{x} \\ \vec{p} \end{pmatrix}, \begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix} \right\rangle_{\mathcal{D}} = \langle (\hat{A} - A)\vec{x}, \vec{w} \rangle + \langle (B\hat{A}^{-1}B^\top - \hat{S})\vec{p}, \vec{q} \rangle. \quad (5.5)$$

The next task is to investigate the condition number, i.e. the quality of the preconditioner. The preconditioner \mathcal{P}_{cg} consists of two preconditioners \hat{A} and \hat{S} . \hat{A} has to be an approximation of A , whereas \hat{S} approximates the so-called Schur complement $B\hat{A}^{-1}B^\top$. Suitable assumptions on how to choose these approximations are given in the next theorem. It gives an estimate of the minimal and maximal eigenvalue of the preconditioned system. Therewith, the condition number κ can be estimated later on.

Theorem 5.3. [139, Theorem 2.2] *Let assumption 5.1 be fulfilled. Furthermore, it is assumed that the relations $\hat{A} > 0$ and $\hat{S} > 0$ with*

$$\langle A\vec{w}, \vec{w} \rangle \geq \nu_1 \langle \hat{A}\vec{w}, \vec{w} \rangle \quad \text{for all } \vec{w} \in \ker B \text{ and } \hat{A} \geq A,$$

and

$$\hat{S} \leq B\hat{A}^{-1}B^\top \leq \nu_2\hat{S},$$

with constants ν_1, ν_2 and $0 < \nu_1 \leq 1$ and $\nu_2 \geq 1$ hold. Then for the maximal and minimal eigenvalue it holds

$$\lambda_{\max}(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}) \leq \nu_2 \left(1 + \sqrt{1 - \frac{1}{\nu_2}} \right)$$

and

$$\lambda_{\min}(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}) \geq \nu_1 \left[\frac{2}{\sqrt{1 - \frac{1}{\nu_2}} + \sqrt{5 - \frac{1}{\nu_2}}} \right]^2 > 0.$$

Due to theorem 1.2 the error of the k -th iteration can be estimated by

$$\frac{\|\vec{e}_k\|_{\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}}}{\|\vec{e}_0\|_{\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}}} \leq \frac{2\rho^k}{1 + \rho^{2k}} \quad (5.6)$$

with $\vec{e}_k = \vec{z}_k - \vec{z}_*$ and

$$\rho = \frac{\sqrt{\kappa(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A})} + 1}.$$

There $\kappa(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A})$ denotes the condition number

$$\kappa(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}) = \frac{\lambda_{\max}(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A})}{\lambda_{\min}(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A})}.$$

Due to (5.6) an application of theorem 5.3 yields an upper bound for the condition number (see [139])

$$\kappa(\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}) \leq \frac{\nu_2}{\nu_1} \left(1 + \sqrt{1 - \frac{1}{\nu_2}} \right) \left[\frac{\sqrt{1 - \frac{1}{\nu_2}} + \sqrt{5 - \frac{1}{\nu_2}}}{2} \right]^2 = \kappa(\nu_1, \nu_2). \quad (5.7)$$

That indicates a condition number depending on the constants ν_1 and ν_2 only.

Remark 5.4. *The closer the constants ν_1 and ν_2 are to one, the better the preconditioner \mathcal{P}_{cg} is expected to be.*

The next task is to investigate the conditions for the choice of suitable preconditioners. The combination of theorem 5.3 and theorem 5.2 allow the application of the CG to the preconditioned system

$$\mathcal{P}_{\text{cg}}^{-1} \mathcal{A} \begin{pmatrix} \vec{x} \\ \vec{q} \end{pmatrix} = \mathcal{P}_{\text{cg}}^{-1} \begin{pmatrix} \vec{f} \\ \vec{s} \end{pmatrix},$$

assuming

$$\langle A\vec{x}, \vec{x} \rangle \geq \nu_1 \langle \hat{A}\vec{x}, \vec{x} \rangle \quad \text{for all } \vec{x} \in \ker B \text{ and } \hat{A} > A \quad (5.8)$$

and

$$\hat{S} < B\hat{A}^{-1}B^\top \leq \nu_2 \hat{S}, \quad (5.9)$$

since the scalar product (5.5) is well defined under these assumptions. To ensure that the preconditioners \hat{A} and \hat{S} fulfill the conditions for applying the CG method, the construction of \hat{A} and \hat{S} is separated in two parts. First, two preliminary candidates \hat{A}_0 and \hat{S}_0 are chosen, which approximate A and $B\hat{A}_0^{-1}B^\top$ respectively. Second, the chosen candidates are scaled properly, i.e.

$$\hat{A} = \frac{1}{\sigma} \hat{A}_0$$

and

$$\hat{S} = \frac{\sigma}{\tau} \hat{S}_0,$$

where the positive constants σ and τ have to be chosen such that (5.4) holds, i.e.

$$\frac{1}{\sigma} \hat{A}_0 > A \quad \text{and} \quad \frac{1}{\tau} \hat{S}_0 < B\hat{A}_0^{-1}B^\top. \quad (5.10)$$

This shows, that a suitable choice of the parameters σ and τ requires at least some knowledge of the eigenvalues. To derive these estimates, the Rayleigh-coefficient (1.2) is used. For the choice of σ one has by $\hat{A} > A$ the relation

$$\frac{1}{\sigma} \hat{A}_0 > A$$

which is equivalent to

$$\frac{1}{\sigma} \vec{x}^\top \vec{x} > \vec{x}^\top \hat{A}_0^{-1} A \vec{x}.$$

By using the Rayleigh-coefficient, one can deduce

$$\vec{x}^\top \hat{A}_0^{-1} A \vec{x} \leq \lambda_{\max}(\hat{A}_0^{-1} A) \vec{x}^\top \vec{x}$$

and choose

$$\frac{1}{\sigma} = \lambda_{\max}(\hat{A}_0^{-1} A) + \tilde{\varepsilon} \quad (5.11)$$

Algorithm 12: PCG method for saddle point problem, see [139]

input : $\mathcal{A}, \mathcal{P}_{\text{cg}}^{-1}, \vec{g}, \vec{z}_0$
output: \vec{z}_{k+1}
for $k = 0, 1, 2, \dots$ **do**
 if $k = 0$ **then**
 $\vec{p}_k = \vec{r}_0$
 else
 $\beta_{k-1} = -\frac{\langle \mathcal{P}_{\text{cg}}^{-1} \mathcal{A} \vec{r}_k, \vec{p}_{k-1} \rangle_{\mathcal{D}}}{\langle \mathcal{P}_{\text{cg}}^{-1} \mathcal{A} \vec{p}_{k-1}, \vec{p}_{k-1} \rangle_{\mathcal{D}}}$
 $\vec{p}_k = \vec{r}_k + \beta_{k-1} \vec{p}_{k-1}$
 $\alpha_k = \frac{\langle \mathcal{P}_{\text{cg}}^{-1} (\vec{g} - \mathcal{A} \vec{z}_k), \vec{p}_k \rangle_{\mathcal{D}}}{\langle \mathcal{P}_{\text{cg}}^{-1} \mathcal{A} \vec{p}_k, \vec{p}_k \rangle_{\mathcal{D}}}$
 $\vec{z}_{k+1} = \vec{z}_k + \alpha_k \vec{p}_k$
 $\vec{r}_{k+1} = \vec{r}_k - \alpha_k \mathcal{P}_{\text{cg}}^{-1} \mathcal{A} \vec{p}_k$

with $\tilde{\varepsilon} > 0$. For the choice of τ one has

$$\frac{1}{\tau} = \lambda_{\min}(\hat{S}_0^{-1} B \hat{A}_0^{-1} B^{\top}) - \tilde{\varepsilon} \quad (5.12)$$

by the same arguments. The values ν_1 and ν_2 are only necessary for analysis but not for the construction of the preconditioners.

It has to be mentioned, that the Schöberl-Zulehner PCG method requires the evaluation of the non-standard scalar product

$$\left\langle \begin{pmatrix} \vec{x} \\ \vec{p} \end{pmatrix}, \begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix} \right\rangle_{\mathcal{D}} = \langle (\hat{A} - A) \vec{x}, \vec{w} \rangle + \langle (B \hat{A}^{-1} B^{\top} - \hat{S}) \vec{p}, \vec{q} \rangle.$$

Therewith, a straightforward implementation of the Schöberl-Zulehner PCG method would lead to

$$\mathcal{D} \begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix} \text{ with } \mathcal{D} = \begin{pmatrix} \hat{A} - A & 0 \\ 0 & B \hat{A}^{-1} B^{\top} - \hat{S} \end{pmatrix} = \mathcal{P}_{\text{cg}} - \mathcal{A},$$

i.e. matrix-vector multiplications with \hat{A} , \hat{A}^{-1} and \hat{S} have to be available. This would be very cost-intensive. A closer look onto the Schöberl-Zulehner PCG method shows that the multiplication with \mathcal{D} is only required for vectors of the form

$$\begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix} = \mathcal{P}_{\text{cg}}^{-1} \begin{pmatrix} \vec{v} \\ \vec{t} \end{pmatrix}.$$

Since it holds

$$\mathcal{D} \begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix} = \mathcal{D} \mathcal{P}_{\text{cg}}^{-1} \begin{pmatrix} \vec{v} \\ \vec{t} \end{pmatrix} = (\mathcal{P}_{\text{cg}} - \mathcal{A}) \mathcal{P}_{\text{cg}}^{-1} \begin{pmatrix} \vec{v} \\ \vec{t} \end{pmatrix} = \begin{pmatrix} \vec{v} \\ \vec{t} \end{pmatrix} - \mathcal{A} \begin{pmatrix} \vec{w} \\ \vec{q} \end{pmatrix},$$

the evaluation of the scalar product involves only multiplications with \mathcal{A} . Furthermore, an application of $\mathcal{P}_{\text{cg}}^{-1}$ requires only matrix-vector multiplications with \hat{A}^{-1} and \hat{S}^{-1} , due to

(1.1). Using (5.6), the quality of the approximation is measured by the energy norm, i.e.

$$\frac{\|\vec{e}_k\|_{\mathcal{DP}_{\text{cg}}^{-1}\mathcal{A}}}{\|\vec{e}_0\|_{\mathcal{DP}_{\text{cg}}^{-1}\mathcal{A}}} \leq \varepsilon, \quad (5.13)$$

with a given tolerance ε . Nevertheless, since $\vec{e}_k = \vec{z}_k - \vec{z}_*$ the calculation of (5.13) is difficult in practice since the exact solution \vec{z}_* is usually unknown. However, it can be used for testing the algorithm with a known solution. In these cases the solution \vec{z}_* is usually constructed such that $\vec{z}_* = \vec{0}$ holds. A more practicable termination condition is the use of the residuum \vec{r}_k and the calculation of

$$\frac{\|\vec{r}_k\|_2}{\|\vec{r}_0\|_2} \leq \varepsilon. \quad (5.14)$$

Numerical experiments in [139] indicate a similar behavior of the residual in the Euclidean norm (5.14) although this is not predicted by theory.

5.2 MINRES

The second Krylov subspace method which is applied to the saddle point problem (5.1) is the MINRES, see subsection 1.2.2. Since the system matrix \mathcal{A} is symmetric but indefinite, the requirements for applying the MINRES are fulfilled.

The goal now is to find a suitable preconditioner to keep the iteration numbers of the preconditioned saddle point problem to a minimum. A possible choice of a suitable preconditioner \mathcal{P} for \mathcal{A} is given in the next theorem.

Theorem 5.5. ([119, Proposition 1, Remark 1]) *If*

$$\mathcal{A} = \begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix}$$

is preconditioned by

$$\mathcal{P} = \begin{pmatrix} A & 0 \\ 0 & BA^{-1}B^\top \end{pmatrix}, \quad (5.15)$$

then the preconditioned system matrix $T = \mathcal{P}^{-1}\mathcal{A}$, has the eigenvalues

$$\begin{aligned} \lambda_1 &= \frac{1}{2}(1 - \sqrt{5}), \\ \lambda_2 &= 1, \\ \lambda_3 &= \frac{1}{2}(1 + \sqrt{5}). \end{aligned}$$

Remark 5.6. ([119]) *Theorem 5.5 even holds if \mathcal{A} is preconditioned by $\mathcal{A}\mathcal{P}^{-1}$ or by $\mathcal{P}_1^{-1}\mathcal{A}\mathcal{P}_2^{-1}$ where $\mathcal{P} = \mathcal{P}_1\mathcal{P}_2$.*

Remark 5.7. *Due to theorem 5.5 there exist only three eigenvalues and the preconditioned MINRES terminates after three iterations because then (1.6) is zero.*

In general there are two possibilities to use these theoretical results.

5.2.1 Exact inverse as preconditioners

The first possibility is, to use the preconditioner (5.15) directly, i.e.

$$\mathcal{P}_{\text{minres,e}}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & (BA^{-1}B^\top)^{-1} \end{pmatrix}. \quad (5.16)$$

There, system solves of A and $BA^{-1}B^\top$ are necessary in order to apply $\mathcal{P}_{\text{minres,e}}^{-1}$ (see algorithm 13). These system solves can be performed by an inner iterative method, for example by suitable preconditioned Krylov subspace methods. In this case, the eigenvalues of the system are known, see theorem 5.5 and the outer iteration takes three iterations (see remark 5.7).

5.2.2 Block diagonal preconditioner

To avoid the system solves of A and especially $BA^{-1}B^\top$, these matrices can be substituted by suitable preconditioners. In this case the block diagonal preconditioner

$$\mathcal{P}_{\text{minres,d}} = \begin{pmatrix} \hat{A} & 0 \\ 0 & \hat{S} \end{pmatrix}, \quad (5.17)$$

is used. There, it remains to find symmetric and positive definite preconditioners \hat{A} and \hat{S} . However, since remark 5.7 only holds for the exact inverse matrices as preconditioners, in the considered case the convergence theory does not hold any longer. Nevertheless, according to [14], the following holds

Theorem 5.8. (*[14, Corollary 2]*) *Let \hat{A} and \hat{S} be symmetric and positive definite preconditioners to A and S , respectively. Then, the eigenvalues $\mathcal{P}_{\text{minres,d}}^{-1}\mathcal{A}$ are contained in the intervals*

$$\left[-\lambda_{\max}(\hat{S}^{-1}S), \frac{-\lambda_{\min}(\hat{S}^{-1}S)}{1 + \frac{1}{\lambda_{\min}(\hat{A}^{-1}A)}} \right] \cup \left[\lambda_{\min}(\hat{A}^{-1}A), \lambda_{\max}(\hat{A}^{-1}A) + \lambda_{\max}(\hat{S}^{-1}S) \right].$$

The theorem above states, that the eigenvalues of the preconditioned system only depend on the choice of the preconditioners \hat{A} and \hat{S} . By applying theorem 1.3 and by using suitable preconditioners, discretization parameter independent results can be yielded.

In the case of using the diagonal block preconditioner $\mathcal{P}_{\text{minres,d}}$, algorithm 2 can be applied directly with $\mathcal{P}_{\text{minres}} = \mathcal{P}_{\text{minres,d}}$. If the exact inverse matrices are taken as preconditioners, i.e. $\mathcal{P}_{\text{minres,e}}$, algorithm 13 is used.

Algorithm 13: application of $\mathcal{P}_{\text{minres,e}}^{-1}$

input : $\vec{x} = (\vec{w}, \vec{z})$

output: $\vec{x}_{\text{new}} = (\vec{w}_{\text{new}}, \vec{z}_{\text{new}})$

in each call of $\mathcal{P}_{\text{minres,e}}$ of algorithm 2 do

 solve $\hat{A}^{-1}A\vec{w}_{\text{new}} = \vec{w}$ with suitable iterative method

 solve $\hat{S}^{-1}(BA^{-1}B^\top)\vec{z}_{\text{new}} = \vec{z}$ with suitable iterative method

5.3 GMRES

The third method applied to the saddle point problem (5.1) is the GMRES. Although the matrix only needs to be regular for an application of the GMRES, the drawback is that there is no convergence theory if the matrix \mathcal{A} is only regular (see subsection 1.2.3).

Nevertheless, a preconditioner is chosen for the GMRES. Inspired by the preconditioner for the PCG method,

$$\mathcal{P}_{\text{gmres}} = \begin{pmatrix} \hat{A} & B^\top \\ B & B\hat{A}^{-1}B^\top - \hat{S} \end{pmatrix} \quad (5.18)$$

is used. There, \hat{A} is a suitable preconditioner for A and \hat{S} a suitable preconditioner for the Schur complement $B\hat{A}^{-1}B^\top$. Due to (1.1), the inverse of $\mathcal{P}_{\text{gmres}}$ is then given by

$$\mathcal{P}_{\text{gmres}}^{-1} = \begin{pmatrix} \hat{A}^{-1} - \hat{A}^{-1}B^\top\hat{S}B\hat{A}^{-1} & \hat{A}^{-1}B^\top\hat{S} \\ \hat{S}^{-1}B\hat{A}^{-1} & -\hat{S}^{-1} \end{pmatrix},$$

since the Schur complement of $\mathcal{P}_{\text{gmres}}$ is

$$B\hat{A}^{-1}B^\top - B\hat{A}^{-1}B^\top - \hat{S} = -\hat{S}.$$

Therewith, the application of $\mathcal{P}_{\text{gmres}}$ only needs the inversion of \hat{S} and \hat{A} . Here, for the GMRES left preconditioning is used, i.e.

$$\mathcal{P}_{\text{gmres}}^{-1}\mathcal{A}\vec{z} = \mathcal{P}_{\text{gmres}}^{-1}\vec{g}. \quad (5.19)$$

Remark 5.9. *Although the matrix $\mathcal{P}_{\text{gmres}}^{-1}$ and \mathcal{A} are symmetric, the product $\mathcal{P}_{\text{gmres}}^{-1}\mathcal{A}$ is not symmetric. Therewith, the matrix $\mathcal{P}_{\text{gmres}}^{-1}\mathcal{A}$ is non-normal.*

If the preconditioner $\mathcal{P}_{\text{cg}}^{-1}$ instead of $\mathcal{P}_{\text{gmres}}^{-1}$ is used, an estimate on the convergence rate can be given by [112, Corollary 4.79], since then a positive definite system matrix is yielded (due to theorem 5.2 the eigenvalues are real and positive in this case). Since the calculation of σ and τ shall be avoided, this choice is not used for the GMRES.

In the next subsections, the Krylov subspace methods are applied to optimal control problems. First, a suitable problem is given and the assumption for existence and uniqueness in the saddle point formulation are checked. For similar results and further literature, see e.g. [83, 130, 139, 145, 146].

5.4 Application to an optimal control problem

Considered is the optimal control problem

$$\min_{y,u} J(y,u) = \min_{y,u} \left(\frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 dx + \frac{\alpha}{2} \int_{\Omega} u^2(x) dx \right)$$

subject to

$$\begin{aligned} -\nabla \cdot D(x) (\nabla y(x)) + c(x)y(x) &= u(x) + f(x) && \text{in } \Omega, \\ y(x) &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ D(x) \frac{\partial y}{\partial n}(x) &= 0 && \text{on } \Gamma_{\mathcal{N}}, \end{aligned} \quad (5.20)$$

where the following assumptions hold:

Assumption 5.10. *The domain $\Omega \subset \mathbb{R}^d$ is open, bounded and has a polygonal boundary $\partial\Omega := \Gamma = \overline{\Gamma_{\mathcal{D}}} \cup \overline{\Gamma_{\mathcal{N}}}$ and $\Gamma_{\mathcal{N}} \cap \Gamma_{\mathcal{D}} = \emptyset$. For the coefficients D and c it holds that $D, c \in L_{\infty}(\Omega)$, $D > 0, c \geq 0$, and if $\text{meas}(\Gamma_{\mathcal{D}}) = \emptyset$ it holds $c > 0$. Moreover D and c are chosen such that the differential operator is uniformly elliptic. For the regularization parameter α it holds $\alpha > 0$, the desired state $y_d \in L_2(\Omega)$ and $f \in L_2(\Omega)$.*

To solve the optimal control problem the adjoint state q is introduced. The adjoint state q is the weak solution to

$$\begin{aligned} -\nabla \cdot D(x) (\nabla q(x)) + c(x)q(x) &= y_d(x) - y(x) && \text{in } \Omega, \\ q(x) &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ D(x) \frac{\partial q}{\partial n}(x) &= 0 && \text{on } \Gamma_{\mathcal{N}}. \end{aligned} \quad (5.21)$$

Since there are no bounds on the control, the projection formulation is given by

$$u(x) = \frac{1}{\alpha} q(x) \quad \text{in } \Omega, \quad (5.22)$$

see e.g. [158]. The goal is to write the problem in a saddle point formulation. This section follows the ansatz in [139]. There a α independent preconditioned CG-version is developed for the considered optimal control problem with $D = c = 1$. It is solved with a uniform h -fem discretization and polynomial degree $p = 1$. An extension to coefficients D, c under the assumption 5.10 is straightforward and is given here. The main result is the extension of the solver [139] for hp -fem.

Remark 5.11. *For the convenience of the reader in further $\Gamma = \Gamma_{\mathcal{N}}$ is assumed. The case of homogeneous Dirichlet boundary or mixed boundary conditions can be proven by the same arguments.*

For rewriting the problem, let X, Q denote real Hilbert spaces, where the space $X = Y \times U$ with $Y = H^1(\Omega)$, $U = L_2(\Omega)$ and $Q = H^1(\Omega)$. For $z = (y, u)$, $\tilde{z} = (\tilde{y}, \tilde{u})$ the optimal control problem can be formulated as

$$\begin{aligned} a(z, \tilde{z}) + b(\tilde{z}, q) &= \langle F, \tilde{z} \rangle_{\Omega} && \text{for all } \tilde{z} \in X \\ b(z, \tilde{q}) &= \langle G, \tilde{q} \rangle_{\Omega} && \text{for all } \tilde{q} \in Q \end{aligned} \quad (5.23)$$

with the bilinear forms

$$a(z, \tilde{z}) = \int_{\Omega} y \tilde{y} \, dx + \alpha \int_{\Omega} u \tilde{u} \, dx \quad (5.24)$$

$$b(z, \tilde{q}) = \int_{\Omega} D \nabla y \cdot \nabla \tilde{q} \, dx + \int_{\Omega} c y \tilde{q} \, dx - \int_{\Omega} u \tilde{q} \, dx \quad (5.25)$$

and the linear forms

$$\begin{aligned} \langle F, \tilde{z} \rangle_{\Omega} &= \int_{\Omega} y_d \tilde{y} \, dx, \\ \langle G, \tilde{q} \rangle_{\Omega} &= \int_{\Omega} f \tilde{q} \, dx. \end{aligned}$$

There, the second equation in (5.23) represents the variational formulation of (5.20), where the first equation in (5.23) represents the sum of (5.21) and (5.22). For the existence and uniqueness of solutions to (5.23) Brezis theorem can be applied.

Theorem 5.12. (Brezzis theorem, see e.g. [44]) Let X, Q denote real Hilbert spaces, $a : X \times X \rightarrow \mathbb{R}$, $b : X \times Q \rightarrow \mathbb{R}$ are bilinear forms, $F : Q \rightarrow \mathbb{R}$ is a continuous linear functional. Furthermore, it is assumed that:

1. The bilinear form $a(\cdot, \cdot)$ is bounded:

$$a(z, w) \leq a_0 \|z\|_X \|\tilde{z}\|_X \quad \forall z, \tilde{z} \in X.$$

2. The bilinear form $a(\cdot, \cdot)$ is coercive on $\ker B = \{\tilde{z} \in X : b(\tilde{z}, \tilde{q}) = 0 \quad \forall \tilde{q} \in Q\}$, i.e. there exists a constant $a_1 > 0$ such that

$$a(\tilde{z}, \tilde{z}) \geq a_1 \|\tilde{z}\|_X^2 \quad \forall \tilde{z} \in \ker B.$$

3. The bilinear form $b(\cdot, \cdot)$ is bounded:

$$\exists b_0 > 0 : \quad \sup_{0 \neq z \in X} \frac{b(z, q)}{\|z\|_X} \leq b_0 \|q\|_Q \quad \forall q \in Q.$$

4. The bilinear form $b(\cdot, \cdot)$ satisfies the inf-sup condition: There exists a constant $b_1 > 0$ such that

$$\sup_{0 \neq w \in X} \frac{b(w, q)}{\|w\|_X} \geq b_1 \|q\|_Q \quad \forall q \in Q.$$

Then, for $(z, q) \in X \times Q$ such that

$$\begin{aligned} a(z, \tilde{z}) + b(\tilde{z}, q) &= \langle F, \tilde{z} \rangle & \forall \tilde{z} \in X \\ b(z, \tilde{q}) &= \langle G, \tilde{q} \rangle & \forall \tilde{q} \in Q \end{aligned}$$

exists a unique solution and the a-priori estimates

$$\begin{aligned} \|z\|_X &\leq \frac{1}{a_1} \|F\|_{X^*} + \frac{1}{b_1} \left(1 + \frac{a_0}{a_1}\right) \|G\|_{Q^*} \\ \|q\|_Q &\leq \frac{1}{b_1} \left(1 + \frac{a_0}{a_1}\right) \|F\|_{X^*} + \frac{a_0}{b_1^2} \left(1 + \frac{a_0}{a_1}\right) \|G\|_{Q^*} \end{aligned}$$

hold.

With Brezzis theorem 5.12 the unique solvability of the saddle point problem (5.23) is shown. As in [139] non-standard scalar products are introduced to obtain α -independent constants a_0, a_1, b_0 and b_1 . Therefore let

$$\begin{aligned} \langle u, \tilde{u} \rangle_U &= \alpha \langle u, \tilde{u} \rangle_\Omega \\ \langle y, \tilde{y} \rangle_Y &= \langle y, \tilde{y} \rangle_\Omega + \sqrt{\alpha} \langle D \nabla y, \nabla \tilde{y} \rangle_\Omega + \sqrt{\alpha} \langle cy, \tilde{y} \rangle_\Omega \\ \langle q, \tilde{q} \rangle_Q &= \frac{1}{\alpha} \langle q, \tilde{q} \rangle_\Omega + \frac{1}{\sqrt{\alpha}} \langle D \nabla q, \nabla \tilde{q} \rangle_\Omega + \frac{1}{\sqrt{\alpha}} \langle cq, \tilde{q} \rangle_\Omega \end{aligned}$$

denote scalar products in U, Y and Q . The scalar product in the space X is given by

$$\langle z, \tilde{z} \rangle_X = \langle u, \tilde{u} \rangle_U + \langle y, \tilde{y} \rangle_Y \quad \text{for } z = (y, u), \tilde{z} = (\tilde{y}, \tilde{u}) \in X.$$

Then, the energy norms are defined as

$$\begin{aligned} \|z\|_X^2 &= \langle z, z \rangle_X, \\ \|q\|_Q^2 &= \langle q, q \rangle_Q. \end{aligned}$$

Remark 5.13. For fixed $\alpha > 0$ the introduced norms are equivalent to the usual $H^1(\Omega)$ norm.

Next, the assumptions of Brezis theorem 5.12 are checked.

Lemma 5.14. Let the bilinear forms

$$\begin{aligned} a(z, \tilde{z}) &= \int_{\Omega} y\tilde{y} \, dx + \alpha \int_{\Omega} u\tilde{u} \, dx \\ b(z, \tilde{q}) &= \int_{\Omega} D\nabla y \cdot \nabla \tilde{q} \, dx + \int_{\Omega} cy\tilde{q} \, dx - \int_{\Omega} u\tilde{q} \, dx \end{aligned}$$

be given. Then it holds:

1. The bilinear form $a(\cdot, \cdot)$ is bounded, i.e.

$$a(z, \tilde{z}) \leq \|z\|_X \|\tilde{z}\|_X \quad \forall z, \tilde{z} \in X.$$

2. The bilinear form $a(\cdot, \cdot)$ is coercive on $\ker B$, i.e.

$$a(z, \tilde{z}) \geq \frac{2}{3} \|\tilde{z}\|_X^2 \quad \forall \tilde{z} \in \ker B.$$

3. The bilinear form $b(\cdot, \cdot)$ is bounded, i.e.

$$\sup_{0 \neq \tilde{z} \in X} \frac{b(\tilde{z}, q)}{\|\tilde{z}\|_X} \leq \|q\|_Q \quad \forall q \in Q.$$

4. The bilinear form $b(\cdot, \cdot)$ fulfills the inf-sup-condition

$$\sup_{0 \neq \tilde{z} \in X} \frac{b(\tilde{z}, q)}{\|\tilde{z}\|_X} \geq \sqrt{\frac{3}{4}} \|q\|_Q \quad \forall q \in Q.$$

Proof. The proof is separated in four parts and follows the proof [139, Lemma 4.1]. There the case $D = c = 1$ is proven. Furthermore the proof uses ideas of [83, Lemma 3.1].

Boundedness of $a(\cdot, \cdot)$. There Cauchy-Schwarz in \mathbb{R}^2 , i.e.

$$ab + \alpha cd \leq (a^2 + \alpha c^2)^{1/2} (b^2 + \alpha d^2)^{1/2}, \quad (5.26)$$

is used. For proving the first inequality, one starts with the bilinear form, apply the triangle inequality and Cauchy-Schwarz inequality and yields

$$\begin{aligned} |a(z, \tilde{z})| &= \left| \int_{\Omega} y\tilde{y} \, dx + \int_{\Omega} \alpha u\tilde{u} \, dx \right| \\ &\leq \left| \int_{\Omega} y\tilde{y} \, dx \right| + \alpha \left| \int_{\Omega} u\tilde{u} \, dx \right| \\ &\leq \|y\|_{L_2(\Omega)} \|\tilde{y}\|_{L_2(\Omega)} + \alpha \|u\|_{L_2(\Omega)} \|\tilde{u}\|_{L_2(\Omega)}. \end{aligned}$$

Application of (5.26) leads to

$$\begin{aligned} |a(z, \tilde{z})| &\leq \left(\underbrace{\|y\|_{L_2(\Omega)}^2}_{\leq \|y\|_Y^2} + \underbrace{\alpha \|u\|_{L_2(\Omega)}^2}_{\leq \alpha \|u\|_U^2} \right)^{1/2} \left(\underbrace{\|\tilde{y}\|_{L_2(\Omega)}^2}_{\leq \|\tilde{y}\|_Y^2} + \underbrace{\alpha \|\tilde{u}\|_{L_2(\Omega)}^2}_{\leq \alpha \|\tilde{u}\|_U^2} \right)^{1/2} \\ &\leq \|z\|_X \|\tilde{z}\|_X, \end{aligned}$$

i.e. the desired estimate.

For proving the **coercivity on ker B of the bilinear form** $a(\cdot, \cdot)$ one considers $b(\tilde{z}, \tilde{q}) = 0$, with $\tilde{z} = (\tilde{y}, \tilde{u})$ i.e.

$$\begin{aligned} \int_{\Omega} D\nabla\tilde{y} \cdot \nabla\tilde{q} \, dx + \int_{\Omega} c\tilde{y}\tilde{q} \, dx &= \int_{\Omega} \tilde{u}\tilde{q} \, dx \\ &\leq \|\tilde{u}\|_{L_2(\Omega)} \|\tilde{q}\|_{L_2(\Omega)} \end{aligned} \quad (5.27)$$

and applies the estimate above on the rewritten norm

$$\begin{aligned} \|\tilde{z}\|_X^2 &= \|\tilde{y}\|_Y^2 + \|\tilde{u}\|_U^2 \\ &= \|\tilde{y}\|_{L_2(\Omega)}^2 + \sqrt{\alpha} (\langle D\nabla\tilde{y}, \nabla\tilde{y} \rangle_{\Omega} + \langle c\tilde{y}, \tilde{y} \rangle_{\Omega}) + \alpha \|\tilde{u}\|_{L_2(\Omega)}^2 \\ &\stackrel{(5.27) \text{ with } \tilde{q}=\tilde{y}}{\leq} \|\tilde{y}\|_{L_2(\Omega)}^2 + \sqrt{\alpha} \|\tilde{u}\|_{L_2(\Omega)} \|\tilde{y}\|_{L_2(\Omega)} + \alpha \|\tilde{u}\|_{L_2(\Omega)}^2. \end{aligned}$$

To fulfill the desired estimate, a suitable constant a_1 to fulfill

$$a(\tilde{z}, \tilde{z}) \geq a_1 \|\tilde{z}\|_X^2 \quad \forall \tilde{z} \in X$$

is needed. To obtain the desired estimate, equivalent transformations are performed. Since it is

$$a(\tilde{z}, \tilde{z}) = \|\tilde{y}\|_{L_2(\Omega)}^2 + \alpha \|\tilde{u}\|_{L_2(\Omega)}^2,$$

setting

$$\begin{aligned} \chi &= \sqrt{\alpha} \|\tilde{u}\|_{L_2(\Omega)} \\ \zeta &= \|\tilde{y}\|_{L_2(\Omega)}, \end{aligned}$$

yields that

$$\chi^2 + \zeta^2 \geq \frac{2}{3} (\chi^2 + \chi\zeta + \zeta^2)$$

has to be proved. This is fulfilled, since it is equivalent to

$$\frac{1}{3} (\chi - \zeta)^2 \geq 0.$$

Therewith it holds

$$a(\tilde{z}, \tilde{z}) \geq \frac{2}{3} \|\tilde{z}\|_X^2 \quad \forall \tilde{z} \in X,$$

i.e. the ker B coercivity of $a(\cdot, \cdot)$ holds.

For the **boundedness of the bilinear form** $b(\cdot, \cdot)$ an application of Cauchy-Schwarz leads to

$$\begin{aligned} b(z, q) &= \int_{\Omega} D\nabla y \cdot \nabla q \, dx + \int_{\Omega} cyq \, dx - \int_{\Omega} uq \, dx \\ &\leq \left(\int_{\Omega} D\nabla y \cdot \nabla y \, dx \right)^{1/2} \left(\int_{\Omega} D\nabla q \cdot \nabla q \, dx \right)^{1/2} + \left(\int_{\Omega} cyy \, dx \right)^{1/2} \left(\int_{\Omega} cq q \, dx \right)^{1/2} \\ &\quad + \|u\|_{L_2(\Omega)} \|q\|_{L_2(\Omega)}. \end{aligned}$$

Using Cauchy-Schwarz in \mathbb{R}^3 , i.e.

$$ab + cd + ef \leq \left(\alpha e^2 + \sqrt{\alpha} a^2 + \sqrt{\alpha} c^2 \right)^{1/2} \left(\frac{1}{\alpha} f^2 + \frac{1}{\sqrt{\alpha}} b^2 + \frac{1}{\sqrt{\alpha}} d^2 \right)^{1/2},$$

this yields

$$\begin{aligned} b(z, q) &\leq \left(\alpha \|u\|_{L_2(\Omega)}^2 + \sqrt{\alpha} \langle D\nabla y, \nabla y \rangle_{\Omega} + \sqrt{\alpha} \langle cy, y \rangle_{\Omega} \right)^{1/2} \\ &\quad \left(\frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2 + \frac{1}{\sqrt{\alpha}} \langle D\nabla q, \nabla q \rangle_{\Omega} + \frac{1}{\sqrt{\alpha}} \langle cq, q \rangle_{\Omega} \right)^{1/2} \\ &\leq \|z\|_X \|q\|_Q. \end{aligned}$$

Since it holds for all $z \in X, q \in Q$ it follows

$$\sup_{0 \neq z \in X} \frac{b(z, q)}{\|z\|_X} \leq \|q\|_Q.$$

The **inf-sup condition** is based on

$$\frac{(a-b)^2}{c^2 + d^2} \leq \frac{a^2}{c^2} + \frac{b^2}{d^2} \quad \text{with equality for } ac^2 = -bd^2.$$

Considering the supremum, it is

$$\begin{aligned} \sup_{0 \neq \tilde{z} \in X} \frac{b(\tilde{z}, q)}{\|\tilde{z}\|_X^2} &= \sup_{0 \neq (\tilde{y}, \tilde{u}) \in Y \times U} \frac{(\langle D\nabla \tilde{y}, \nabla q \rangle_{\Omega} + \langle c\tilde{y}, q \rangle_{\Omega} - \langle \tilde{u}, q \rangle_{\Omega})^2}{\|\tilde{y}\|_Y^2 + \|\tilde{u}\|_U^2} \\ &= \sup_{0 \neq \tilde{y} \in Y} \frac{(\langle D\nabla \tilde{y}, \nabla q \rangle_{\Omega} + \langle c\tilde{y}, q \rangle_{\Omega})^2}{\|\tilde{y}\|_Y^2} + \sup_{0 \neq \tilde{u} \in U} \frac{\langle \tilde{u}, q \rangle_{\Omega}^2}{\|\tilde{u}\|_U^2}, \end{aligned}$$

where \tilde{y}, \tilde{u} are chosen such that

$$(\langle D\nabla \tilde{y}, \nabla q \rangle_{\Omega} + \langle c\tilde{y}, q \rangle_{\Omega}) \|\tilde{y}\|_Y^2 = -\langle \tilde{u}, q \rangle_{\Omega} \|\tilde{u}\|_U^2.$$

The second term can be reformulated by using Cauchy Schwarz and a special choice of $\tilde{u} = q$, i.e.

$$\sup_{0 \neq \tilde{u} \in U} \frac{\langle \tilde{u}, q \rangle_{\Omega}^2}{\|\tilde{u}\|_U^2} \geq \frac{\langle q, q \rangle_{\Omega}^2}{\alpha \|q\|_{L_2(\Omega)}^2} \geq \frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2.$$

Considering again the whole term and using the choice of $\tilde{y} = q$, it follows

$$\sup_{0 \neq \tilde{z} \in X} \frac{b(\tilde{z}, q)}{\|\tilde{z}\|_X^2} \geq \frac{(\langle D\nabla q, \nabla q \rangle_{\Omega} + \langle cq, q \rangle_{\Omega})^2}{\|q\|_Y^2} + \frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2.$$

To derive the inf-sup condition

$$\sup_{0 \neq \tilde{z} \in X} \frac{b(\tilde{z}, q)}{\|\tilde{z}\|_X} \geq b_1 \|q\|_Q,$$

equivalent formulations are used to find a suitable value b_1 . Using the results from above and setting $d^2 = \langle D\nabla q, \nabla q \rangle_\Omega + \langle cq, q \rangle_\Omega$ for simplicity, one considers

$$\begin{aligned} & \frac{d^4}{\|q\|_Y^2} + \frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2 \geq b_1^2 \|q\|_Q^2, \\ \Leftrightarrow & d^4 + \frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2 \|q\|_Y^2 \geq b_1^2 \|q\|_Q^2 \|q\|_Y^2, \\ \Leftrightarrow & d^4 + \frac{1}{\alpha} \left(\|q\|_{L_2(\Omega)} + \sqrt{\alpha} d^2 \right) \geq b_1^2 \left(\frac{1}{\alpha} \|q\|_{L_2(\Omega)}^2 + \frac{1}{\sqrt{\alpha}} d^2 \right) \left(\|q\|_{L_2(\Omega)} + \sqrt{\alpha} d^2 \right). \end{aligned}$$

Therewith, it holds the equivalent equation

$$(1 - b_1^2) d^4 + \frac{1}{\alpha} (1 - b_1^2) \|q\|_{L_2(\Omega)}^4 + \frac{1}{\sqrt{\alpha}} (1 - 2b_1^2) \|q\|_{L_2(\Omega)}^2 d^2 \geq 0.$$

Choosing $b_1^2 = \frac{3}{4}$ it follows

$$\frac{1}{4} + \frac{1}{4\alpha} \|q\|_{L_2(\Omega)}^4 + \frac{1}{2\sqrt{\alpha}} \|q\|_{L_2(\Omega)}^2 d^2 = \frac{1}{4} \left(d^2 - \frac{1}{\sqrt{\alpha}} \|q\|_{L_2(\Omega)}^2 \right)^2 \geq 0,$$

i.e. the desired estimate with $b_1 = \sqrt{\frac{3}{4}}$. \square

Since the assumptions of Brezis theorem are fulfilled, the existence and uniqueness of the solution follows.

Theorem 5.15. *The saddle point problem (5.23) with the non-standard norms $\|z\|_X, \|q\|_Q$ has a unique solution.*

Proof. As proven in lemma 5.14, the assumptions for Brezis theorem 5.12 hold with the α -independent constants $a_0 = 1, a_1 = \frac{2}{3}, b_0 = 1$ and $b_1 = \sqrt{\frac{3}{4}}$. \square

The next step is to discretize the saddle point problem (5.23). For the discretization, hp finite elements are used.

5.4.1 Discrete saddle point problem

The discrete formulation of the saddle point problem is given by

$$\begin{aligned} a(z_N, \tilde{z}_N) + b(\tilde{z}_N, q_N) &= \langle F, \tilde{z}_N \rangle_\Omega \quad \forall \tilde{z}_N \in X_N \\ b(z_N, \tilde{q}_N) &= \langle G, \tilde{q}_N \rangle_\Omega \quad \forall \tilde{q}_N \in Q_h \end{aligned} \tag{5.28}$$

with the finite dimensional spaces $X_N \subset X, Q_N \subset Q$. The discrete bilinear forms are

$$\begin{aligned} a(z_N, \tilde{z}_N) &= \int_\Omega y_N \tilde{y}_N \, dx + \alpha \int_\Omega u_N \tilde{u}_N \, dx \\ b(z_N, \tilde{q}_N) &= \int_\Omega D\nabla y_N \cdot \nabla \tilde{q}_N \, dx + \int_\Omega c y_N \tilde{q}_N \, dx - \int_\Omega u_N \tilde{q}_N \, dx \end{aligned}$$

and the linear forms

$$\begin{aligned} \langle F, \tilde{z}_N \rangle_\Omega &= \int_\Omega y_d \tilde{y}_N \, dx, \\ \langle G, \tilde{q}_N \rangle_\Omega &= \int_\Omega f \tilde{q}_N \, dx. \end{aligned}$$

In matrix notation the discrete saddle point problem is given by

$$\begin{aligned} A_N \vec{z}_N + B_N^\top \vec{q}_N &= \vec{f}_N \\ B_N \vec{z}_N &= \vec{s}_N \end{aligned} \quad (5.29)$$

with

$$A_N = \begin{pmatrix} M_N & 0 \\ 0 & \alpha M_N \end{pmatrix} \quad \text{and} \quad B_N = \begin{pmatrix} K_N & -M_N \end{pmatrix}$$

where M_N denotes the mass matrix representing the $L_2(\Omega)$ inner product and K_N denotes the stiffness matrix of the state equation, i.e.

$$K_N = \langle D\nabla y_N, \nabla y_N \rangle_\Omega + \langle c y_N, y_N \rangle_\Omega$$

For existence and uniqueness of the solution Brezis theorem shall be applied. Since $X_N = Y_N \times U_N$ and by choosing $Y_N = Q_N \subset U_N$, a closer look on the proof of the assumptions of Brezis theorem shows that the four assumptions are also fulfilled in the discrete case.

Theorem 5.16. *Assume that $X_N \subset X$, $Q_N \subset Q$ and $X_N = Y_N \times U_N$ with $Y_N = Q_N \subset U_N$. Then the discrete saddle point problem (5.28) possesses a unique solution.*

Furthermore, the constants in Brezis theorem are independent of mesh parameters and the regularization parameter α . Therewith, if possible, the preconditioners are chosen such that the iteration numbers, needed to solve the saddle point problem (5.29), are independent of these parameters.

Remark 5.17. *The convergence rates which include BPX preconditioners in further, are all given for triangulations without hanging nodes, i.e. by using locally refined triangular elements or uniform refined meshes. In case of using the BPX with hanging nodes, a logarithmic dependence comes into play.*

5.4.2 Application of preconditioned CG to the discrete problem

The main task in this section is to find suitable preconditioners for getting α , h and p independent iteration numbers when applying the preconditioned CG method from Schöberl and Zulehner, see [139]. As stated in section 5.1 the preconditioner

$$\mathcal{P}_{\text{cg}} = \begin{pmatrix} \hat{A}_N & B_N^\top \\ B_N & B_N \hat{A}_N^{-1} B_N^\top - \hat{S}_N \end{pmatrix}$$

is used. To get h and p independence, the preconditioners \hat{A}_N and \hat{S}_N have to support h , and p independence. The α independence can be yielded by using non-standard scalar products as introduced in section 5.4, i.e.

$$\langle z, \tilde{z} \rangle_X = \langle y, \tilde{y} \rangle_Y + \langle u, \tilde{u} \rangle_U \quad \text{and} \quad \langle q, \tilde{q} \rangle_Q.$$

In the discrete space, the scalar products $\langle z, \tilde{z} \rangle_X$ and $\langle q, \tilde{q} \rangle_Q$ are bilinear forms on $X_N \subset X$ and $Q_N \subset Q$. The associated matrices representing these scalar products are denoted by \underline{X}_N and \underline{Q}_N , i.e.

$$\langle z_N, w_N \rangle_X = \langle \underline{X}_N \vec{z}_N, \vec{w}_N \rangle, \quad \langle q_N, p_N \rangle_Q = \langle \underline{Q}_N \vec{q}_N, \vec{p}_N \rangle$$

with

$$\underline{X}_N = \begin{pmatrix} M_N + \sqrt{\alpha}K_N & 0 \\ 0 & \alpha M_N \end{pmatrix} \quad \text{and} \quad \underline{Q}_N = \left(\frac{1}{\alpha}M_N + \frac{1}{\sqrt{\alpha}}K_N \right).$$

As in [139] the preconditioners \hat{A}_N and \hat{S}_N are chosen to be

$$\hat{A}_N = \frac{1}{\sigma} \hat{X}_N, \quad \hat{S}_N = \frac{\sigma}{\tau} \hat{Q}_N \quad (5.30)$$

for some real parameters $\sigma > 0$, $\tau > 0$, which have to be determined. Furthermore, it is assumed that the quality of the preconditioners can be described by the spectral estimates

$$(1 - q_X) \hat{X}_N \leq \underline{X}_N \leq \hat{X}_N \quad \text{and} \quad (1 - q_Q) \hat{Q}_N \leq \underline{Q}_N \leq \hat{Q}_N \quad (5.31)$$

with constants $q_X, q_Q \in [0, 1)$. With this choice, the following lemma holds.

Lemma 5.18. *[139, Lemma 3.1] Assume that the assumptions for Brezzis theorem are fulfilled for (5.28). Furthermore, it holds (5.30) and (5.31). Then, the conditions (5.8) and (5.9) are satisfied with*

$$\nu_1 = \sigma(1 - q_X)a_0 \quad \text{and} \quad \nu_2 = \tau b_0^2,$$

if the parameters σ and τ are chosen such that

$$\sigma < \frac{1}{a_0} \quad \text{and} \quad \tau > \frac{1}{(1 - q_X)(1 - q_Q)b_1^2}.$$

Lemma 5.18 indicates, that by choosing suitable preconditioners \hat{A}_N and \hat{S}_N , convergence rates independent of the discretization parameters and the regularization parameter α are yielded. The concrete choice of the preconditioners follows the two-step construction described in section 5.1. To get α independent iterations numbers, it is necessary to incorporate the regularization parameter α into the preconditioners. Therefore, one introduces

$$\hat{A}_{N_0} = \begin{pmatrix} \hat{Y}_N & 0 \\ 0 & \alpha \hat{M}_N \end{pmatrix} \quad \text{and} \quad \hat{S}_{N_0} = \frac{1}{\alpha} \hat{Y}_N,$$

where \hat{Y}_N is a suitable preconditioner for $Y_N = \sqrt{\alpha}K_N + M_N$ and \hat{M}_N a suitable preconditioner for M_N . This leads to

$$\hat{A}_N = \frac{1}{\sigma} \hat{A}_{N_0} = \frac{1}{\sigma} \begin{pmatrix} \hat{Y}_N & 0 \\ 0 & \alpha \hat{M}_N \end{pmatrix} \quad \text{and} \quad \hat{S}_N = \frac{\sigma}{\tau} \frac{1}{\alpha} \hat{Y}_N$$

with suitably chosen real parameters $\sigma > 0$, $\tau > 0$. For the well-possessedness of the CG method it is sufficient to assume

$$(1 - q_X) \hat{Y}_N \leq Y_N \leq \hat{Y}_N \quad \text{and} \quad (1 - q_X) \hat{M}_N \leq M_N \leq \hat{M}_N \quad (5.32)$$

for $q_X \in [0, 1)$, where the constant one on the upper bound is important. The factor q_X describes the quality of the preconditioners \hat{Y}_N and \hat{M}_N . Lemma 5.18 indicates, that the CG method applied to the discrete problem is well defined, when choosing

$$\nu_1 = \sigma(1 - q_X) \frac{2}{3} \quad \text{and} \quad \nu_2 = \tau. \quad (5.33)$$

Then, the parameters σ and τ satisfy

$$\sigma < 1 \text{ and } \tau > \frac{4}{3(1 - q_X)^2}.$$

Remark 5.19. *This indicates that by choosing h and p independent preconditioners, one yields h , p and α independent results if q_X is constant and independent of h and p . Nevertheless, one has to consider, that the calculated values a_0, a_1, b_0, b_1 are calculated for the discrete problem with exact inverse as preconditioners. By using approximations the constants can change, nevertheless the approximations \hat{A}_N, \hat{S}_N can be chosen to be h and p independent.*

In order to reduce the number of iterations, a well balanced choice of σ and τ is quite important. Therefore, both parameters are chosen according to (5.11) and (5.12).

Remark 5.20. *In the paper of Schöberl and Zulehner [139] h -fem is used for discretization. As the mass matrix M_N is well conditioned for polynomial degree $p = 1$, a simple preconditioner, e.g. a few steps of a symmetric Gauss-Seidel iterations, are a good preconditioner \hat{M}_N . As preconditioner for \hat{Y}_N they chose a standard multigrid preconditioner for the elliptic differential operator represented by the bilinear form*

$$\sqrt{\alpha} \langle D \nabla y, \nabla q \rangle_{\Omega} + (\sqrt{\alpha} + 1) \langle y, q \rangle_{\Omega}.$$

Another possible choice for the preconditioner \hat{Y}_N would be the BPX-preconditioner.

For an hp finite element discretization the situation is different. Since neither the mass matrix nor the stiffness matrix is well conditioned (see section 3.3), one of the major tasks is to use preconditioners suited for this problem. In here the preconditioners introduced in section 3.3 are used. For the mass matrix the preconditioner C_M^{-1} given by (3.24) is used. For preconditioning matrix Y_N either (3.21) or (3.22), both by using the system matrix Y_N , are taken. The first one is denoted by C_{YBPXP}^{-1} , the second one C_{YPE}^{-1} . In further C_Y^{-1} states that any of these two preconditioners is chosen.

Since the choice $\hat{Y}_N = C_Y^{-1}$ and $\hat{M}_N = C_M^{-1}$ would violate condition (5.32), suitable scaling factors have to be found. First, a scaling factor for the preconditioner for the mass matrix is calculated. To calculate them, the spectral equivalence in remark 3.45 is used. It induces the inequality

$$\underline{c}_M C_M \leq M_N \leq \overline{c}_M C_M.$$

The goal now is to determine the constants \underline{c}_M and \overline{c}_M . The lower bound is equivalent to

$$\begin{aligned} \underline{c}_M &= \min_{\vec{z} \in \mathbb{R}^n} \frac{\langle M_N \vec{z}, \vec{z} \rangle}{\langle C_M \vec{z}, \vec{z} \rangle} \\ &= \min_{\vec{y} = C_M^{-1/2} \vec{z}} \min_{\vec{y} \in \mathbb{R}^n} \frac{\langle M_N C_M^{-1/2} \vec{y}, C_M^{-1/2} \vec{y} \rangle}{\langle C_M C_M^{-1/2} \vec{y}, C_M^{-1/2} \vec{y} \rangle} \\ &= \min_{\vec{y} \in \mathbb{R}^n} \frac{\langle C_M^{-1/2} M_N C_M^{-1/2} \vec{y}, \vec{y} \rangle}{\langle \vec{y}, \vec{y} \rangle} \\ &= \lambda_{\min}(C_M^{-1} M_N), \end{aligned}$$

where in the last step the Rayleigh-coefficient can be used, since the matrices $C_M^{-1/2}M_N C_M^{-1/2}$ and $C_M^{-1}M_N$ have the same eigenvalues. The upper bound follows analogously. Therefore it holds

$$\bar{c}_M = \lambda_{\max}(C_M^{-1}M_N).$$

Since the constant condition number in theorem 3.38 and theorem 3.41 induces spectral equivalence, i.e. $C_{YPE} \sim Y$ and $C_{YBPXP} \sim Y$, the estimates for C_Y can be obtained analogously. Therewith, the preconditioners are chosen by

$$\hat{Y}_N = \lambda_{\max}(C_Y^{-1}Y_N)C_Y \quad (5.34)$$

$$\hat{M}_N = \lambda_{\max}(C_M^{-1}M_N)C_M \quad (5.35)$$

in order to get the constant one at the upper bound of (5.32).

Remark 5.21. *In general, the determination of the constants $\lambda_{\max}(C_Y^{-1}Y_N)$ and $\lambda_{\max}(C_M^{-1}M_N)$ is of course avoided on each mesh. In fact only a rough estimate of the maximal eigenvalue is necessary in order to satisfy the condition (5.31).*

Suitable constants σ and τ can be determined by calculating the maximal and minimal eigenvalue of the eigenvalue problems

$$\begin{aligned} \hat{A}_{N_0}^{-1}A\vec{z} &= \lambda_1\vec{z} \\ \hat{S}_{N_0}^{-1}B\hat{A}_{N_0}^{-1}B^\top\vec{z} &= \lambda_2\vec{z} \end{aligned}$$

respectively. Summarizing, for moderate polynomial degree p and meshes without hanging nodes the main result of this section is:

Theorem 5.22. *The application of the Schöberl-Zulehner PCG [139] with preconditioners*

$$\begin{aligned} \hat{Y}_N &= c_1 C_{YBPXP} \\ \hat{M}_N &= c_2 C_M \end{aligned}$$

to the optimal control problem (5.29) discretized with hp-finite elements, leads then to the condition number $\kappa(\sigma, \tau, q_X)$ given in (5.36) if suitable constants $c_1, c_2 > 0$ are used. The condition number is especially independent of h , p and α .

Proof. The proof is analogue to the proof for h -fem in [139], i.e. the estimates (5.7) and (5.33) are used to get the condition number

$$\kappa = \kappa(\sigma, \tau, q_X) = -\frac{3(\sqrt{-1+\tau} + \sqrt{\tau})(\sqrt{-1+\tau} + \sqrt{-1+5\tau})^2}{8(-1+q_X)\sigma\sqrt{\tau}}. \quad (5.36)$$

□

Remark 5.23. *The given estimate for the condition number implies that the quality of the preconditioners q_X and a suitable choice of the parameters σ and τ are very important in order to get low iteration numbers.*

Next, the costs for an application of $\mathcal{P}_{\text{cg}}^{-1}\mathcal{A}$ are estimated. Therewith, it is assumed to have bc-refinement and a moderate polynomial degree p .

Theorem 5.24. *Let the assumptions of theorem 5.22 be satisfied and let bc-refinement be performed. Then, each action of $\mathcal{P}_{cg}^{-1}\mathcal{A}\vec{r}$ in the preconditioned CG costs $\mathcal{O}(N)$.*

Proof. Due to (1.1), \mathcal{P}_{cg}^{-1} is given by

$$\mathcal{P}_{cg}^{-1} = \begin{pmatrix} \hat{A}_N^{-1} - \hat{A}_N^{-1}B_N^\top S_N^{-1}B_N\hat{A}_N^{-1} & \hat{A}_N^{-1}B_N^\top \hat{S}_N^{-1} \\ \hat{S}_N^{-1}B_N\hat{A}_N^{-1} & -\hat{S}_N^{-1} \end{pmatrix} \quad (5.37)$$

and the matrices

$$\hat{A}_N^{-1} = \sigma \begin{pmatrix} \hat{Y}_N^{-1} & 0 \\ 0 & \frac{1}{\alpha}\hat{M}_N^{-1} \end{pmatrix} \quad \text{and} \quad \hat{S}_N^{-1} = \frac{\tau}{\sigma}\alpha \left(\hat{Y}_N^{-1} \right).$$

There are N degrees of freedom for the state, the adjoint and the control. Since the application of C_Y^{-1} costs $\mathcal{O}(N)$ due to theorem 3.41 and the application of C_M^{-1} too (due to theorem 3.44), the costs for the application of \hat{A}_N^{-1} and \hat{S}_N^{-1} are $\mathcal{O}(N)$. Due to the structure of B_N , it is sparse too and a matrix-vector multiplication can be performed in $\mathcal{O}(N)$.

An overall application of \mathcal{P}_{cg}^{-1} costs four matrix-vector multiplications with \hat{S}_N^{-1} , two matrix-vector multiplications with B_N , two with B_N^\top and five matrix-vector multiplications with \hat{A}_N^{-1} , due to (5.37). Since each matrix-vector multiplication can be performed in $\mathcal{O}(N)$, and the matrix \mathcal{A} is sparse, an overall estimate of the costs give $\mathcal{O}(N)$.

Due to algorithm 12, the solution to the optimal control problem by applying the preconditioned CG can be yielded in $\kappa(\sigma, \tau, q_X)\mathcal{O}(N)$ effort. \square

Remark 5.25. *For bc-fem, an application of C_{YPE} instead of C_{YBPXP} yields again a constant condition number. However, the effort for the application of $\mathcal{P}_{cg}^{-1}\vec{r}$ in $d = 2$ is $\mathcal{O}(N \log^8 N)$ whereas for $d = 3$ quasi-optimal complexity is not possible any longer if C_{YPE} is used.*

Remark 5.26. *For a bc-refinement and by using the preconditioners proposed in [63], even in three dimensions the cost $\mathcal{O}(N)$ can be yielded.*

Next, the case of a very high polynomial degree p is considered. Then, a preconditioner as (3.23) can be used. Due to the non-constant condition number of the preconditioner (3.23), a $\log(p)$ dependence in q_X , σ and τ comes into play by analogue considerations as above.

Remark 5.27. *By considering the estimates in [29], suitable choices for c_M , c_Y , σ and τ can be found. An application of the preconditioner then can be performed in quasi-optimal time. However, the drawback then is, that after each p -refinement, the values c_M , c_Y , σ and τ have to be adjusted, since they depend on $\log(p)$.*

The preconditioned CG method is quite promising, however, the drawbacks are the necessary estimation of the maximal eigenvalues of $C_Y^{-1}Y_N$ and $C_M^{-1}M_N$ and the correct choice of σ and τ . If these constants are not chosen suitably, the preconditioned CG method fails, due to the loss of positive definiteness. To keep the costs low, the estimation of the four constants is done on the coarsest grid. Furthermore, for the calculation of σ and τ , a safeguard strategy – following Herzog and Sachs [83] – can be used. They multiplied σ by $\frac{1}{\sqrt{2}}$ and τ by $\sqrt{2}$ if the positive definiteness gets lost.

These considerations indicate that the determination of the four necessary constants for applying the preconditioned CG is nasty. Furthermore it can yield an unfavourable scaling of the preconditioners for \hat{Y}_N and \hat{M}_N , which can increase the number of iterations. Therewith, other Krylov subspace methods are considered.

5.4.3 Application of preconditioned MINRES to the discrete system

Another way to solve the saddle point problem (5.29) is to apply the MINRES, which is possible due to the symmetry of the system. In this subsection two cases are considered.

5.4.3.1 Exact inverses as preconditioners

In the first case, the exact inverses are taken as preconditioners, i.e.

$$\mathcal{P}_{\text{minres,e}} = \begin{pmatrix} A_N & 0 \\ 0 & B_N A_N^{-1} B_N^\top \end{pmatrix}.$$

Therewith, the theoretical estimates given in subsection 5.2, and especially remark 5.7 hold, i.e. only three iterations with the preconditioned MINRES are necessary in order to solve the system. However, the drawback is that inner solves of A_N and $B_N A_N^{-1} B_N^\top$ are necessary. This can be realized by an inner iterative method, see e.g. algorithm 13. Since the matrices A_N and $B_N A_N^{-1} B_N^\top$ are symmetric and positive definite, a preconditioned CG method with preconditioners as introduced in section 3.3 can be used. As preconditioners C_M^{-1} and C_{YBPXP}^{-1} shall be used.

Then, low and high polynomial degrees can again be distinguished. First, low polynomial degree and bc-refinement is considered.

Theorem 5.28. *Let the optimal control problem given by (5.29) be discretized with bc-fem and solved with a preconditioned MINRES. Furthermore, let $\mathcal{P}_{\text{minres,e}}$ be the preconditioner and let algorithm 13 be used in order to perform a multiplication with $\mathcal{P}_{\text{minres,e}}$. Then, the condition number $\kappa_{\mathcal{P}_{\text{minres,e}}}$ is constant and an application of $\mathcal{P}_{\text{minres,e}}$ costs $\mathcal{O}(N)$.*

Proof. Since the exact inverse is used as preconditioner, by applying theorem 5.5 and using the remark 5.7, the condition number is constant if algorithm 13 is performed with a suitable precision. Due to the estimates in section 3.3 and with the choice of

$$\hat{A}_{N_0} = \begin{pmatrix} \hat{Y}_N^{-1} & 0 \\ 0 & \frac{1}{\alpha} \hat{M}_N^{-1} \end{pmatrix} \text{ and } \hat{S}_{N_0}^{-1} = \left(\hat{Y}_N^{-1} \right)$$

the equation systems

$$\begin{aligned} \hat{A}_{N_0} A \vec{w}_{\text{new}} &= \vec{w} \\ \hat{S}_{N_0}^{-1} (B_N A_N^{-1} B_N^\top) \vec{z}_{\text{new}} &= \vec{z}, \end{aligned}$$

can be solved in $\mathcal{O}(N)$. Since only matrix-vector-multiplication is used and the overall preconditioned MINRES has three eigenvalues (see theorem 5.5), the costs for solving the equation system are $\mathcal{O}(N)$. \square

Remark 5.29. *In case of an arbitrary hp-mesh, the equation system can be solved in quasi-optimal complexity.*

Remark 5.30. *The drawback in that case is that the inner iterations can increase the overall iterations greatly, since they are necessary in each application of the preconditioner $\mathcal{P}_{\text{minres,e}}$.*

For higher polynomial degrees the preconditioners have to be of course modified. Therewith the work for applying the preconditioner increases.

Remark 5.31. *In the case of using $C_Y = C_{YBPXP2}$, i.e. the preconditioner (3.23) or the choice $C_Y = C_{YPE}$, the costs for an application of preconditioner $\mathcal{P}_{\text{minres,e}}$ are quasi-optimal.*

5.4.3.2 Diagonal preconditioner

Next, not the exact inverse, but an approximation of it, is used, i.e.

$$\mathcal{P}_{\text{minres,d}} = \begin{pmatrix} \hat{A}_N & 0 \\ 0 & \hat{S}_N \end{pmatrix}.$$

\hat{A}_N and \hat{S}_N are similar to the choices in the PCG method, i.e.

$$\hat{A}_N = \begin{pmatrix} \hat{Y}_N & 0 \\ 0 & \alpha \hat{M}_N \end{pmatrix} \quad \text{and} \quad \hat{S}_N = \frac{1}{\alpha} \hat{Y}_N. \quad (5.38)$$

According to subsection 5.2 in this case theorem 5.8 can be applied, which yields:

Theorem 5.32. *The eigenvalues of the preconditioned system $\mathcal{P}_{\text{minres,d}}^{-1} \mathcal{A}$ for $\hat{Y}_N = C_Y$ and $\hat{M}_N = C_M$ are contained in the intervals*

$$[-\Upsilon_1, -\Upsilon_2] \cup [\Upsilon_3, \Upsilon_4]$$

with $\Upsilon_i \geq 0$ for $i = 1, \dots, 4$ and Υ_i independent of h and p .

Proof. According to theorem 5.8 the parameters Υ_i depend only on the minimal and maximal eigenvalues of

$$\hat{A}_N^{-1} A_N = \begin{pmatrix} C_{YPE}^{-1} & 0 \\ 0 & \frac{1}{\alpha} C_M^{-1} \end{pmatrix} \begin{pmatrix} Y_N & 0 \\ 0 & \alpha M_N \end{pmatrix}$$

and the eigenvalues of $\hat{S}_N^{-1} S_N = \alpha C_{YPE}^{-1} Y_N$, which are constant due to section 3.3. For simplicity it is set

$$\begin{aligned} \lambda_{\max}(\hat{S}_N^{-1} S_N) &= \overline{c_S} & \lambda_{\min}(\hat{S}_N^{-1} S_N) &= \underline{c_S} \\ \lambda_{\max}(\hat{A}_N^{-1} A_N) &= \overline{c_A} & \lambda_{\min}(\hat{A}_N^{-1} A_N) &= \underline{c_A}, \end{aligned}$$

which yields

$$\begin{aligned} \Upsilon_1 &= \overline{c_S} \\ \Upsilon_2 &= \frac{\underline{c_S}}{1 + \frac{1}{\underline{c_A}}} \\ \Upsilon_3 &= \underline{c_A} \\ \Upsilon_4 &= \overline{c_A} + \overline{c_S}. \end{aligned}$$

Since these values are especially independent of the mesh size h and the polynomial degree p , Υ_i are independent of h and p , they are therefore independent of the mesh. \square

Remark 5.33. *Theorem 5.32 indicates h and p independent iterations numbers due to theorem 1.3. In case of using the choice $C_Y = C_{YBPXP}$ with hanging nodes or $C_Y = C_{YBPXP2}$, the Υ_i have a logarithmic dependence on h or an almost logarithmic dependence on p (see [29])*

for the eigenvalues of C_{YBPXP2}). By applying theorem 1.3 an estimate on the necessary iteration numbers k to reach a given tolerance ε can be calculated. For the choice $C_Y = C_{YBPXP}$ one yields

$$k = \log\left(\frac{2}{\varepsilon}\right) \mathcal{O}(\log(h)). \quad (5.39)$$

That means the iteration numbers increase logarithmically in the mesh size h . For the choice $C_Y = C_{YBPXP2}$ it follows

$$k = \log\left(\frac{2}{\varepsilon}\right) \mathcal{O}\left((\log p \log^\chi \log p)^{7/2}\right)$$

for a given tolerance ε and $\chi > 1$. That indicates an almost logarithmically increase of the iteration numbers k .

Remark 5.34. The difference to the application of the Schöberl-Zulehner PCG is that in that case an α independence cannot be expected.

In the case of low polynomial degree and bc-refinement, the following theorem holds:

Theorem 5.35. Let the optimal control problem (5.29) be solved in a saddle point formulation and discretized with bc-fem. Then, an application of $\mathcal{P}_{\text{minres,d}}^{-1}\vec{r}$ costs $\mathcal{O}(N)$.

Proof. The preconditioner is given by

$$\mathcal{P}_{\text{minres,d}}^{-1} = \begin{pmatrix} C_Y^{-1} & 0 & 0 \\ 0 & \frac{1}{\alpha}C_M^{-1} & 0 \\ 0 & 0 & C_Y^{-1} \end{pmatrix}.$$

Due to theorem 3.41, the costs for applying C_Y^{-1} are $\mathcal{O}(N)$, due to theorem 3.44 the costs for applying C_M^{-1} are $\mathcal{O}(N)$ too. Since in each application of the preconditioner, C_Y^{-1} has to be applied twice, whereas C_M^{-1} is applied once, the overall costs for one multiplication are $\mathcal{O}(N)$. \square

Remark 5.36. By combining theorem 5.32 and assuming bc-refinement (without hanging nodes), the overall costs for an application of $\mathcal{P}_{\text{minres,d}}^{-1}\vec{r}$ are $\mathcal{O}(N)$ if $C_Y = C_{YBPXP}$. In case of using C_{YPE} for two dimensions the costs are $\mathcal{O}(N \log^8 N)$.

Remark 5.37. In case of using $C_Y = C_{YBPXP2}$ an almost $\log(p)$ dependence comes into play for each action $\mathcal{P}_{\text{minres,d}}^{-1}\vec{r}$ due to theorem 3.42.

Remark 5.38. If a general hp-refinement is used for the discretization, the application of $C_{YBPX}^{-1}\vec{r}$ is quasi-optimal. Therewith, the application of $\mathcal{P}_{\text{minres,d}}^{-1}\vec{r}$ is quasi-optimal too.

An ansatz to yield α independent convergence rates with the MINRES, is presented in [172]. There, the saddle system (5.29) for $D = c = 1$ is considered. Since there are no bounds on the control u , the system can be written in the form

$$\begin{pmatrix} M_N & K_N \\ K_N & -\frac{1}{\alpha}M_N \end{pmatrix} \begin{pmatrix} \vec{y} \\ \vec{q} \end{pmatrix} = \begin{pmatrix} \vec{y}_d \\ \vec{0} \end{pmatrix}.$$

Then, block-diagonal preconditioners by the operator interpolation technique by Zulehner [172] are constructed. This leads to a preconditioner

$$\mathcal{P}_Z = \begin{pmatrix} M_N + \alpha^{\frac{1}{2}}K_N & 0 \\ 0 & \alpha^{-1}M_N + \alpha^{-\frac{1}{2}}K_N \end{pmatrix}$$

which yields α independent iteration numbers. In practice the block diagonal entries, both of the form

$$\mathfrak{m}M_N + K_N,$$

are usually replaced by efficient preconditioners, for example by multigrid or multilevel preconditioners. Even though in this case the MINRES can be applied, the theory developed in former sections does not hold in this case, since the system is different. For further results see [172].

5.4.4 Application of GMRES to discrete system

As stated in section 5.3 a preconditioner for the GMRES only needs to be regular, therewith

$$\mathcal{P}_{\text{gmres}} = \begin{pmatrix} \hat{A}_N & B_N^\top \\ B_N & B_N \hat{A}_N^{-1} B_N^\top - \hat{S}_N \end{pmatrix}$$

is used, where

$$\hat{A}_N = \begin{pmatrix} \hat{Y}_N & 0 \\ 0 & \hat{\alpha}M_N \end{pmatrix} \quad \text{and} \quad \hat{S}_N = \frac{1}{\alpha}\hat{Y}_N$$

with the choice $\hat{Y}_N = C_Y$ and $\hat{M}_N = C_M$ as defined in (3.22) and (3.24). Again, as in the MINRES, no constants σ and τ and no constants for scaling the preconditioner are necessary. However, there is a big drawback in theory, since no estimate on the condition number can be given due to missing theoretical estimates. Therewith, no statements on parameter independence can be given.

However, at least in the case of the exact inverse matrices $\hat{Y}_N = Y_N$ and $\hat{M}_N = M_N$, the numerical experiments indicate α independence, see subsection 5.5.1.

5.5 Numerical experiments

To confirm the theoretical results, several numerical experiments are given in this section.

5.5.1 Square

As first example, a simple example on the square $(-1, 1)^2$ with known solution, is considered. The primal equation is given by

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) + f(x) && \text{in } \Omega \\ \frac{\partial y}{\partial n}(x) &= 0 && \text{on } \Gamma \end{aligned}$$

and the adjoint equation by

$$\begin{aligned} -\Delta q(x) + q(x) &= y_d(x) - y(x) && \text{in } \Omega \\ \frac{\partial q}{\partial n}(x) &= 0 && \text{on } \Gamma. \end{aligned}$$

Since no active constraints are used, the projection formula is

$$u(x) = \frac{1}{\alpha} q(x) \quad \text{in } \Omega.$$

The solution to this optimal control problem is given by

$$\begin{aligned} y(x) &= e^{\frac{1}{3}x_1^3 - x_1} e^{\frac{1}{3}x_2^3 - x_2} \\ q(x) &= -y(x). \end{aligned}$$

For the convenience of the reader, the preconditioners introduced in section 3.3 are recalled. The preconditioner for the mass matrix M_N , introduced by (3.24) is denoted by

$$C_M^{-1} = P_0^\top (\text{diag}(M)_{p=1})^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top M_{\nu_i}^{-1} P_{\nu_i}. \quad (5.40)$$

C_Y denotes a preconditioner for

$$Y_N = \sqrt{\alpha} K_N + (\sqrt{\alpha} + 1) M_N.$$

Here, two choices are used. The first one is (3.21) used for Y_N and denoted by

$$C_{YPE}^{-1} = P_0^\top (Y_{p=1})^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top Y_{\nu_i}^{-1} P_{\nu_i}. \quad (5.41)$$

The second one uses a BPX preconditioner for the h -part (see (3.22)) and is given by

$$C_{YBPXP}^{-1} = P_0^\top C_{p=1, BPX}^{-1} P_0 + \sum_{\nu_i \in \mathcal{V}_C} P_{\nu_i}^\top Y_{\nu_i}^{-1} P_{\nu_i}, \quad (5.42)$$

where $C_{p=1, BPX}$ denotes the BPX preconditioner for the $p = 1$ part of Y_N .

Remark 5.39. *The following examples are calculated with relative tolerance criterion depending on the refinement. For uniform h -refinement the tolerance is 10^{-6} , for uniform p -refinement it is 10^{-12} and for uniform hp -refinement it is chosen to be 10^{-10} .*

5.5.1.1 Schöberl-Zulehner PCG

First, numerical results for the Schöberl-Zulehner PCG are presented. To confirm the constant condition number given in (5.36), the parameters σ , τ and the values $c_{1Y} = \lambda_{\max}(C_Y^{-1} Y_N)$ and $c_{1M} = \lambda_{\max}(C_M^{-1} M_N)$ are estimated by solving eigenvalue problems for different values of α and different discretization parameters h and p .

The estimation of the parameters is performed for the two choices: $\hat{Y}_N = C_{YPE}$, see (5.41), and $\hat{Y}_N = C_{YBPXP}$, (5.42). A look on the numerical results in chapter 3 expected the choice $\hat{Y}_N = C_{YPE}$ to be better in case of iterations numbers but worse if the costs are considered, see section 3.3.1.

The estimation of the parameters for uniform h -refinement can be found in figure 5.1 and figure 5.3. It has to be considered, that only in figure 5.1 the conditions (5.34) and (5.35) are taken into account, due to the slow convergence of the BPX, which influences the determination of $c_{1Y} = \lambda_{\max}(C_Y^{-1}Y_N)$ and therefore possibly perturbs the estimates of σ and τ . However, this does not change the overall behaviour of the parameters σ and τ , since the conditions (5.34) and (5.35) are only scaling conditions in order to ensure the positive definiteness of the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$.

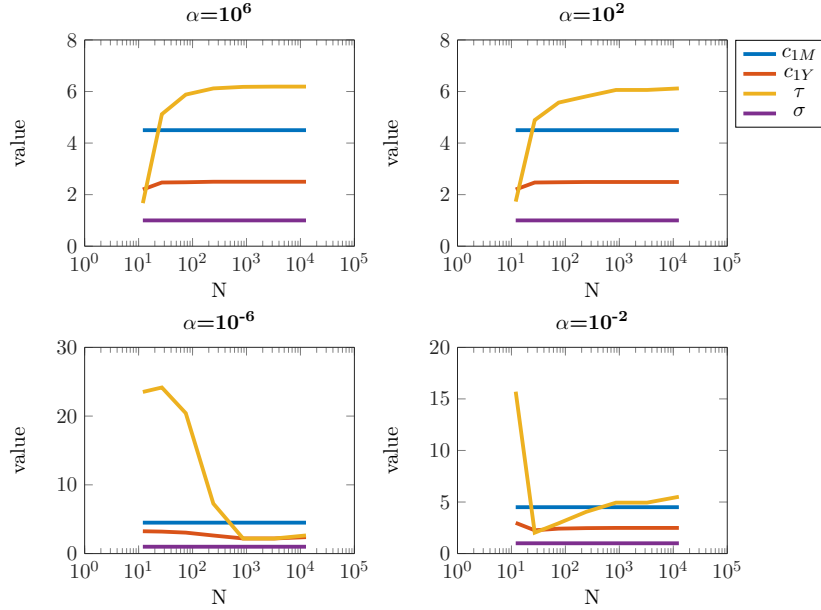


Figure 5.1: estimated parameters for different α and uniform h -refinement for $\hat{Y}_N = C_{YPE}$ and $\hat{M}_N = C_M$

A comparison of the estimation of the maximal eigenvalues, i.e. c_{1M} and c_{1Y} in figure 5.1 and in figure 5.3 shows, that – as expected – the value c_{1M} is the same for both choices of \hat{Y}_N . The value c_{1Y} however is different, in case of $\hat{Y}_N = C_{YPE}$ it is between $[2.2, 3.25]$, in case of $\hat{Y}_N = C_{YBPXP}$ it is – for the given meshes – in the interval $[1.2, 11.25]$. Due to the semilogarithmic scale in these plots it is clear that even c_{1Y} is constant if the mesh is chosen sufficiently fine.

Furthermore, it is observed that c_{1M} and c_{1Y} change if uniform p -refinement is used. Whereas for uniform h -refinement the values c_{1M} are exactly 4.5, for uniform p -refinement it increases up to 6.2 for the highest polynomial degree. For c_{1Y} the situation is similar, but here c_{1Y} is in a smaller range ($[1.45, 5.93]$) as for uniform h -refinement.

For uniform p -refinement some kind of oscillating can be observed for the parameter τ (see figure 5.2 and figure 5.4). This behaviour occurs especially in case of the preconditioner choice $\hat{Y}_N = C_{PE}$, for the choice $\hat{Y}_N = C_{YBPXP}$ it can only be seen for $\alpha = 10^{-2}$. The oscillating in estimating the bound for τ is due to the differences between odd and even polynomial degrees and a typical behaviour, which can also be observed in figure 3.15 in section 3.4).

Next, the best possible iteration numbers for uniform h -fem and different values of the regularization parameter α given in figure 5.5. There, the choice $\hat{Y}_N = Y_N$ and $\hat{M}_N = M_N$ are realized, i.e. the best possible preconditioner is used. Of course this choice is too expensive

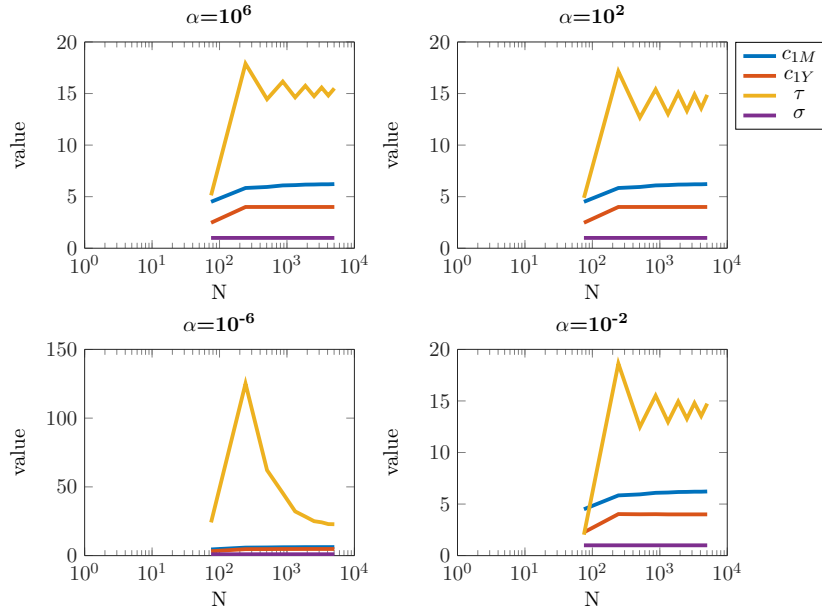


Figure 5.2: estimated parameters for different α and uniform p -refinement for $\hat{Y}_N = C_{YPE}$ and $\hat{M}_N = C_M$

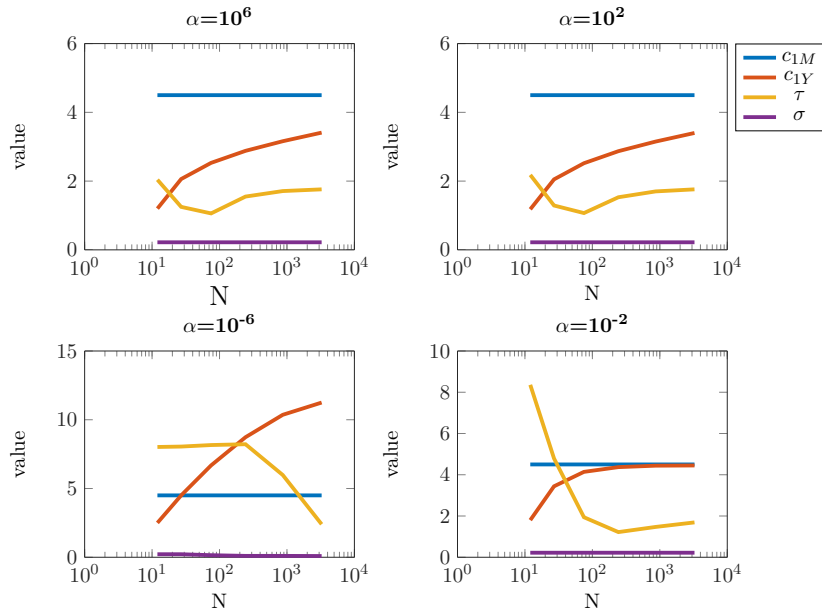


Figure 5.3: unscaled estimated parameters for different α and uniform h -refinement for $\hat{Y}_N = C_{YBXP}$ and $\hat{M}_N = C_M$

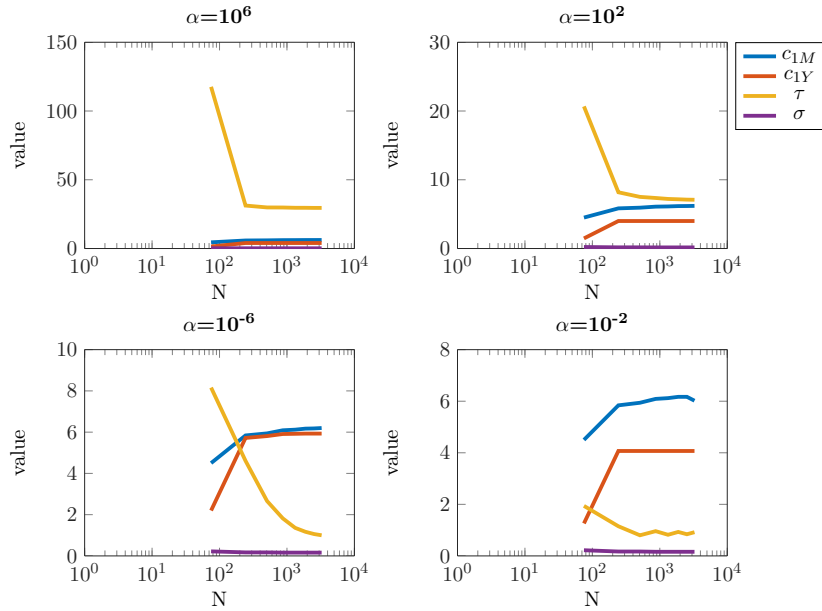


Figure 5.4: unscaled estimated parameters for different α and uniform p -refinement for $\hat{Y}_N = C_{YBPXP}$ and $\hat{M}_N = C_M$

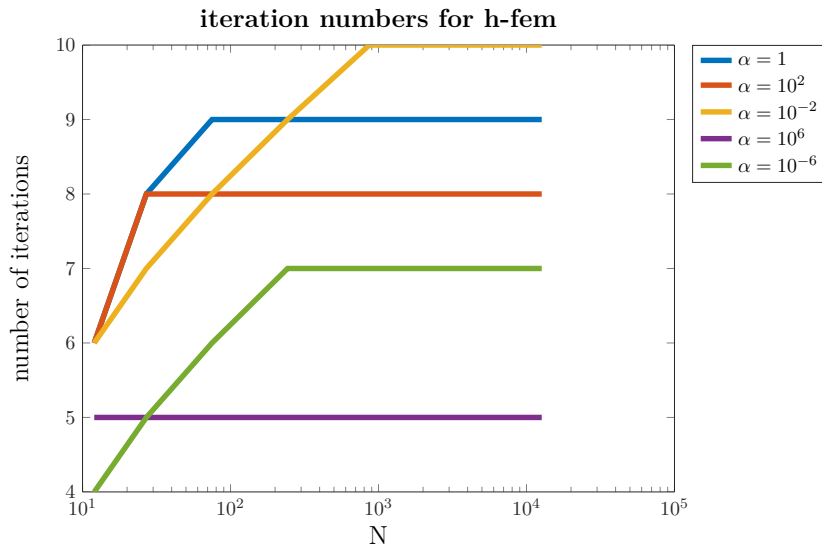


Figure 5.5: iterations number with exact inverses as preconditioners

in practice. However, it yields a good clue in order to evaluate the used preconditioners. The obtained results confirm the theory, i.e. the α -independence of the considered method. As termination condition the residual (see (5.14)) is used, i.e. a different one as indicated by theory (which would be (5.13)). The reason therefore is, that it is not possible to evaluate $\|\vec{e}_k\|_{\mathcal{DP}_{cg}^{-1}\mathcal{A}} = \|\vec{z}_k - \vec{z}_*\|_{\mathcal{DP}_{cg}^{-1}\mathcal{A}}$ if the calculated solution is unknown. Nevertheless, results in [139] show similar results for the residual based termination condition even this is not confirmed by theory.

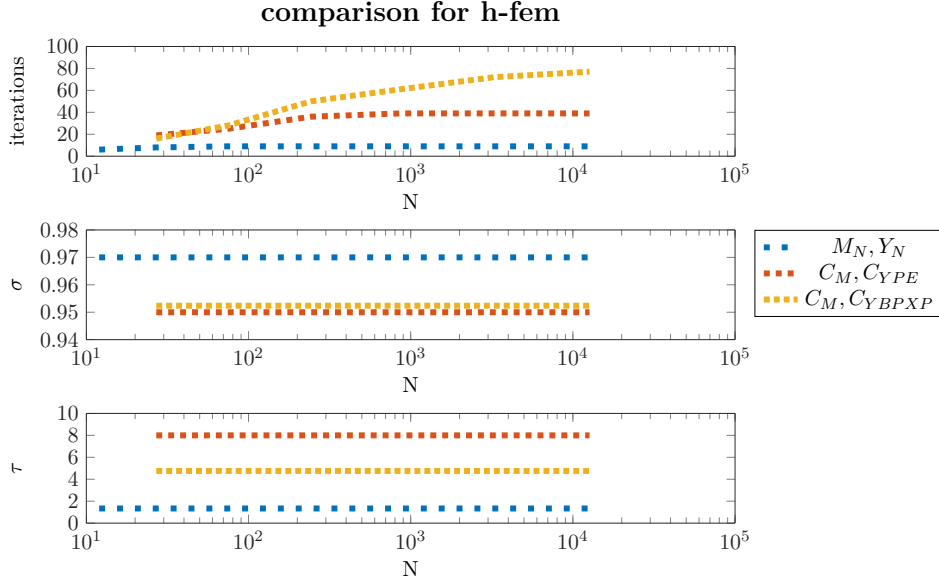


Figure 5.6: comparison of different preconditioner for uniform h -refinement

In figure 5.6 a comparison of different preconditioners is given for uniform h -fem and the regularization parameter $\alpha = 1$. As preconditioners the choices $\hat{M}_N = M_N, \hat{Y}_N = Y_N$, i.e. the best possible preconditioner and the choices $\hat{M}_N = C_M$ with $\hat{Y}_N = C_{YPE}$ and $\hat{Y}_N = C_{YBXP}$ are considered. The best possible iteration numbers are 9, in case of using C_{YPE} the iteration numbers increase up to 39, whereas if C_{YBXP} is taken, the iteration numbers are about 77 for the finest mesh. The parameters σ are constants for all different preconditioners, although they do not have the same values for each choice of the preconditioner.

In figure 5.7 a comparison of different preconditioner for uniform p -fem with the regularization parameter $\alpha = 1$ is presented. It can be observed, that the parameters σ and τ are constant, whereas the condition numbers for different choices of preconditioners seem to oscillate a bit. This is caused by the differences in the behavior of the preconditioner for even and odd polynomial degrees.

In contrast to figure 5.6, the iteration numbers in figure 5.7 get much higher if the preconditioners C_{YP} or C_{YBXP} are used. The reason therefore is, that the matrices for p -refinement are worse conditioned than the one for h -fem and that the parameters σ and τ are further away from 1 than in figure 5.6. Moreover, the difference can already be observed in case of the best possible iteration numbers, which are 9 for uniform h -refinement but already 19 for p -refinement. Nevertheless, a better choice of σ and τ is expected to decrease the number of iterations.

Remark 5.40. *Although a rough value of the parameters σ and τ can be obtained by using*

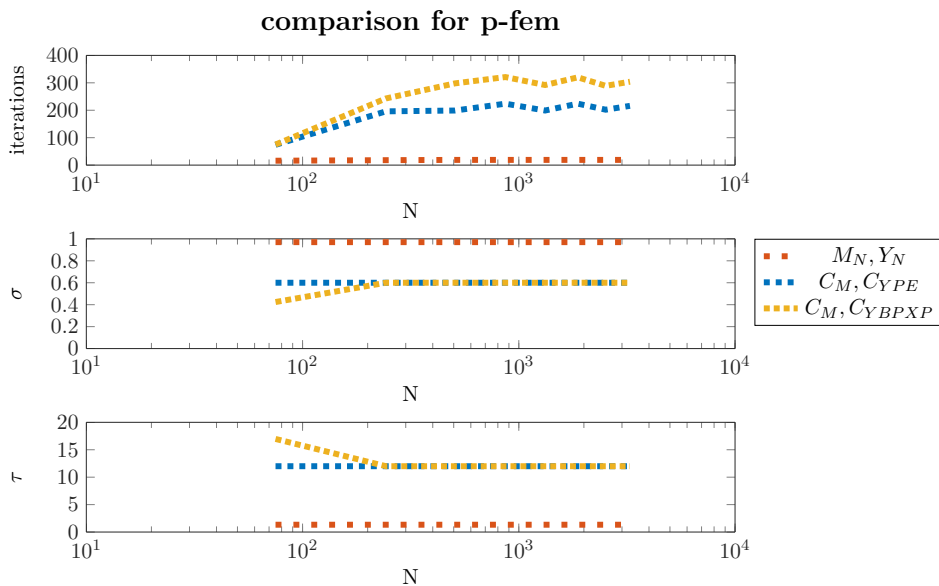


Figure 5.7: comparison of different preconditioners for uniform p -refinement

(5.11) and (5.12), it is quite hard to choose $\tilde{\varepsilon}$ sufficiently in order to avoid non positive definiteness for the scalar product. Since the algorithm by Herzog and Sachs [83] to adjust σ and τ in case the positive definiteness fails, does not lead to good results in the examples presented in here, a fixed value for σ and τ is used a-priori sometimes.

It has to be stated that the numerical results for the Schöberl-Zulehner PCG confirm the theory, i.e. a condition number independent of the discretization parameters h , p and the regularization parameter α and only depending on the parameters σ and τ . However, the estimation of the parameters σ and τ is quite nasty and especially in the case of uniform p -fem other preconditioners should be tried in order to decrease the number of iterations.

Remark 5.41. In order to decrease the iteration numbers for example multigrid methods can be used. They lead to a faster convergence in general which has two advantages: Firstly, that c_{1M} and c_{1Y} can be estimated more easily, since a very coarse mesh can be used to get a very good estimate. Second, the iteration numbers are usually lower in case of using multigrid methods. Therewith an overall reduction of the iterations numbers is expected.

Remark 5.42. The results are also valid for hp -discretizations. However, an extension to hanging nodes is not straightforward, due to scalar product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$, where a projector has to come into play in order to get a conform solution. This problem can be avoided by using triangular elements.

5.5.1.2 MINRES

Next, results for the preconditioned MINRES are given. Here, as preconditioner $\mathcal{P}_{\text{minres,d}}^{-1}$ is used. For the preconditioners \hat{A}_N and \hat{S}_N different choices are tried. In figure 5.8 the results for uniform h - and uniform p -refinement and the regularization parameter $\alpha = 1$ are given. The results for the choices M_N and Y_N show the best possible results, in case of uniform h -fem (color: middle blue) that means about 16 iterations, in case of uniform p -fem

that leads to about 29 iterations. If \hat{M}_N is taken as preconditioner and \hat{Y}_N is substituted by C_{YPE} , the iteration numbers increase, but are bounded by 24. If \hat{Y}_N is chosen to be C_{YBPXP} the iterations numbers are much higher. The reason therefore is, that the convergence (to a constant condition number) in case of the BPX preconditioner is quite slow. Nevertheless, since a semilog plot is used and the iteration number decreases, the constant condition number can also be confirmed in that case. In case of p -refinement the iteration numbers seems to oscillate. This is due to the behaviour of the preconditioners for even and odd polynomial degrees.

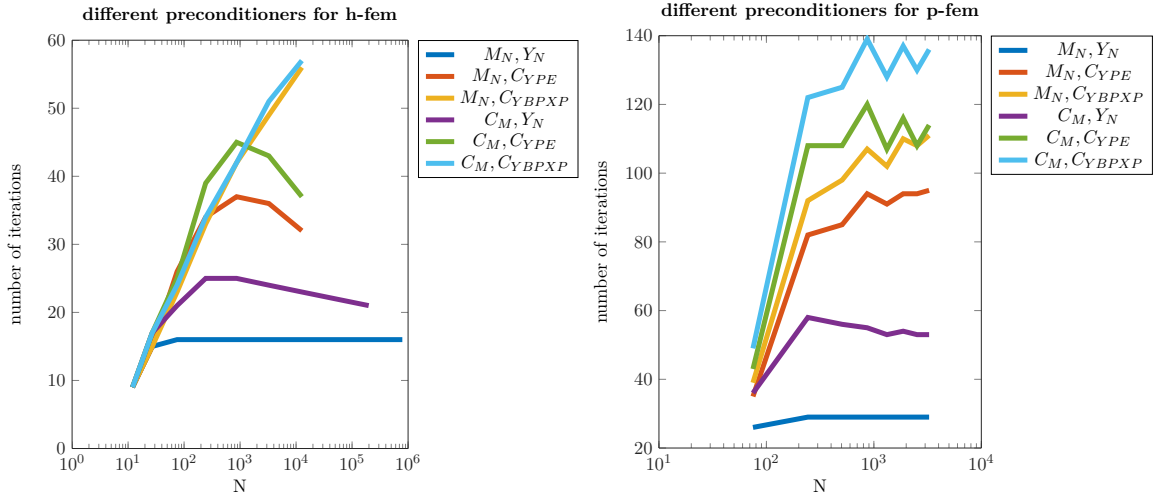


Figure 5.8: results for preconditioned MINRES for uniform h - and uniform p -refinement

In general – comparing both refinements – it can be said, that the preconditioner for \hat{Y}_N , i.e. C_{YBPXP} or C_{YPE} have to be improved in order to get better results when only preconditioners and no exact inversion is used, since the choice $\hat{M}_N = C_M$ and \hat{Y}_N lead to better results than $\hat{M}_N = M_N$ with either C_{YBPXP} or C_{YPE} . Furthermore it can be seen, that the choice C_{YPE} leads to better results than C_{YBPXP} . This behaviour is expected, since in C_{PE} the basis functions for $p = 1$ are inverted directly.

In figure 5.9 results for bc-refinement are presented. As in case of uniform h - or uniform p -refinement, the preconditioner C_{YPE} is better than the preconditioner C_{YBPXP} . However, due to the hanging nodes in case of applying C_{YBPXP} the constant condition number gets lost and a $\log(h)$ term comes into play (compare (5.39) for theory). This behaviour does not appear unexpectedly, since it is also observed in chapter 3 in figure 3.11, figure 3.12 and figure 3.16. As in case of uniform h - and uniform p -refinement, in order to improve the iterations numbers better preconditioners for \hat{Y}_N have to be used, since the preconditioner C_M is much better than the preconditioner C_{YPE} (about 60 iterations number versus 130 iteration numbers). The best possible iteration numbers which can be obtained are about 25 iteration numbers.

Remark 5.43. *In all applications of the MINRES, a big difference by the results which can be obtained in the best possible case ($\hat{M}_N = M_N$, $\hat{Y}_N = Y_N$) and the results yielded by using preconditioners can be observed. Therewith, it is supposed to use – as in the case of the Schöberl-Zulehner PCG – multigrid methods in order to lower the number of iterations.*

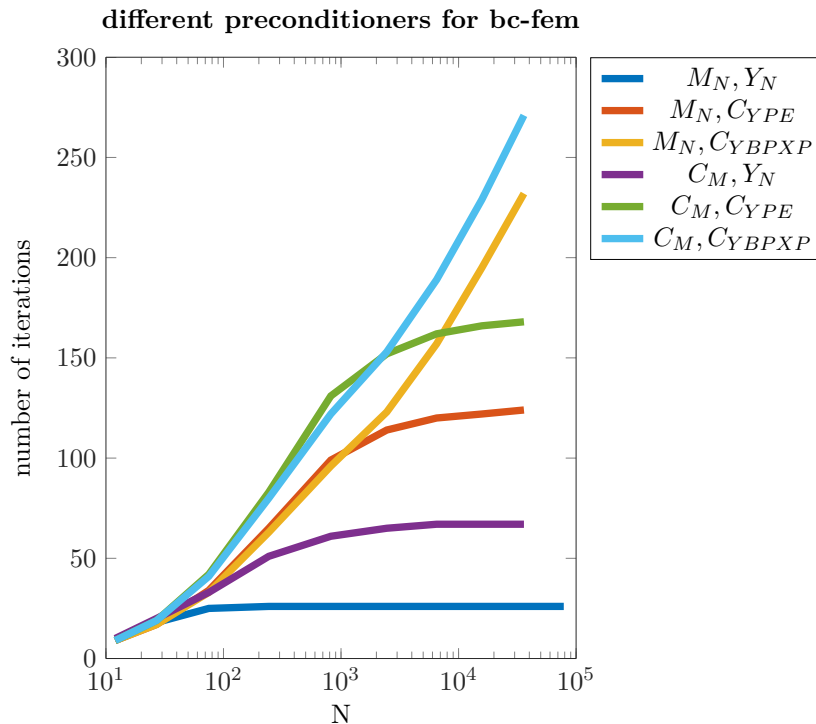


Figure 5.9: results for preconditioned MINRES for bc-refinement

5.5.1.3 GMRES

Results for the preconditioned GMRES are given next. In figure 5.10 different preconditioners are tried. Again, the choice M_N and Y_N show the best possible results.

The best possible iteration numbers in case of GMRES are about 12 steps in case of uniform h -refinement and about 16 iterations in case of uniform p -refinement. Moreover, figure 5.10 shows constant iteration numbers for both refinement. As in case of MINRES, the choice $\hat{M}_N = C_M$ and $\hat{Y}_N = Y_N$ leads to better results than the choice $\hat{M}_N = M_N$ and $\hat{Y}_N = C_{YPE}$. Moreover, the preconditioner C_{YPE} leads to much better results than the preconditioner C_{YBXP} . Furthermore, as in the MINRES an oscillating behavior in case of p -refinement is observed.

These results show discretization parameter independent behavior, even this is not confirmed by theory.

Remark 5.44. *The results for the GMRES were obtained with the left preconditioned GMRES. In case of using the right preconditioned GMRES the results change a bit. A further problem is the loss of the orthogonality in the basis vectors calculated in the GMRES algorithm due to machine precision. This seems to depend on the values of α as of the choice of left or right preconditioning. In order to avoid that problem, it is suggested to use the truncated GMRES method, even there, the convergence at least after N iterations gets lost.*

Remark 5.45. *An overall comparison between the three applied Krylov subspace methods with suitable preconditioners shows that the best possible results are obtained with the preconditioned CG method. The preconditioned MINRES leads to the highest iteration numbers in case of comparing the best possible results. Nevertheless, the difference is not that big*

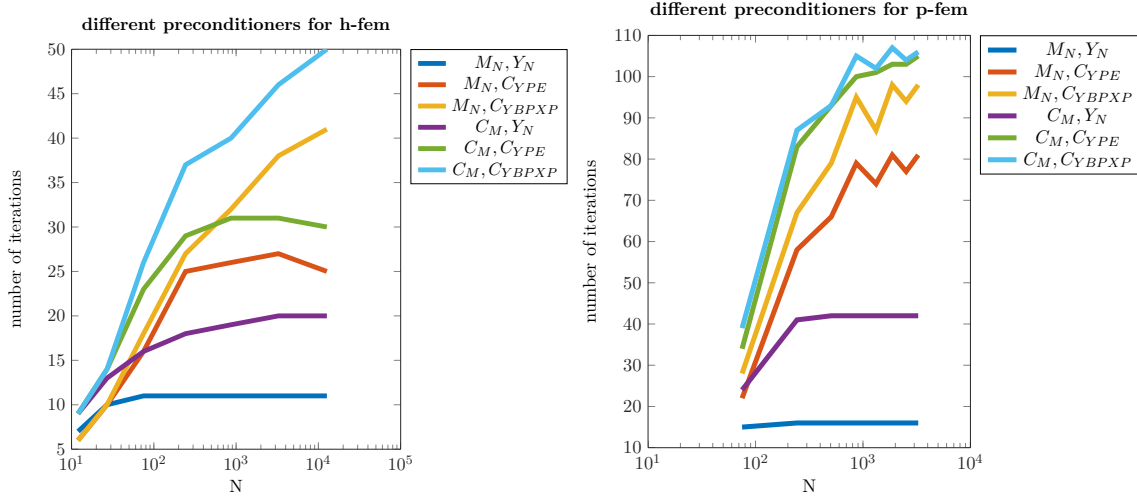


Figure 5.10: results for preconditioned GMRES for different refinements

and the preconditioned MINRES seems to be the easiest iterative method in practice. That is caused by the fact, that in the preconditioned CG some parameters have to be chosen correctly (which is quite nasty), and in the preconditioned GMRES there are no theoretical results and furthermore an application of the untruncated GMRES does not seem to be useful.

5.5.2 Hole

As second example a different domain Ω is considered and solved with the preconditioned MINRES. As in example 5.4.4.3, a square with a hole in the middle is given as domain. The differential equations stays the same as in the example before, i.e. the following equations are considered: the state equation is given by

$$\begin{aligned} -\Delta y(x) + y(x) &= u(x) + f(x) && \text{in } \Omega \\ \frac{\partial y}{\partial n}(x) &= 0 && \text{on } \Gamma \end{aligned}$$

and the adjoint equation by

$$\begin{aligned} -\Delta q(x) + q(x) &= y_d(x) - y(x) && \text{in } \Omega \\ \frac{\partial q}{\partial n}(x) &= 0 && \text{on } \Gamma. \end{aligned}$$

Since no active constraints are used, the projection formula is

$$u(x) = \frac{1}{\alpha} q(x) \text{ in } \Omega.$$

In this example the desired state is given by

$$y_d(x) = 10 \sin(\pi x_1) + 5 \cos(\pi x_2^2),$$

whereas the solution is unknown.

The reason why the preconditioned MINRES is chosen for this example is the quite nasty estimation of the parameters in the Schöberl-Zulehner PCG. The preconditioned MINRES is favourable to the GMRES, since in that case at least a discretization parameter independence can be proven, although the α -independence which holds for the Schöberl-Zulehner PCG gets lost. A further drawback which can occur in the GMRES, is the non-orthogonality of the calculated basis vectors in the algorithm due to machine precision. In order to avoid that problem, truncated GMRES can be used. Nevertheless, the property to reach the solution at least after N iterations gets lost.

Remark 5.46. *Especially in the case of quadrilateral elements with hanging nodes the preconditioned MINRES has the advantage that the projector in order to get a conform solution can be easily incorporated .*

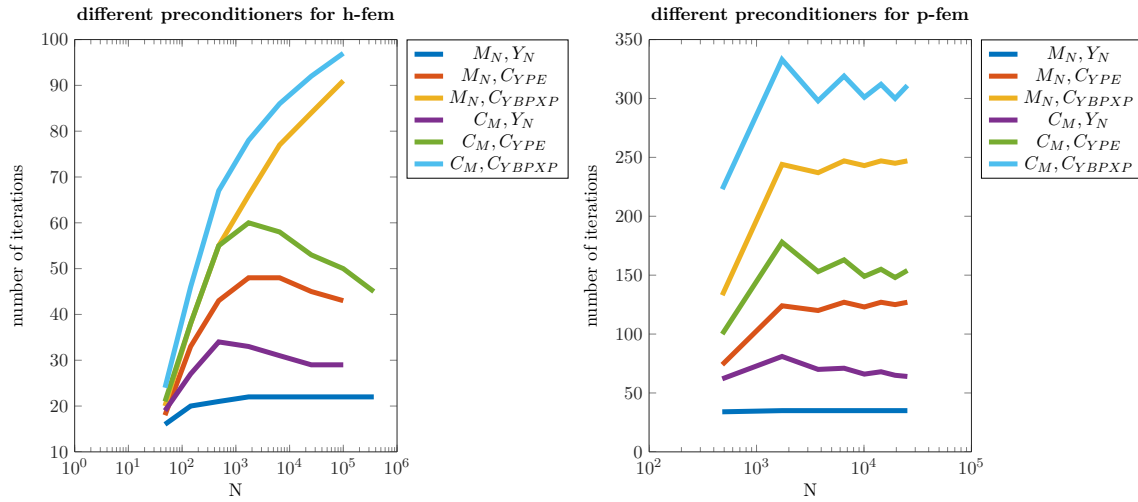


Figure 5.11: iteration numbers for the hole-example

In figure 5.11 the iteration numbers with respect to the number of unknowns for uniform h -, and uniform p -fem are given. The results show discretization parameter independence, which follows by theorem 5.32. As in the example before, different preconditioners – including the exact inverses – are used in order to compare them.

A comparison between uniform h - and uniform p -refinement shows that the preconditioners are discretization parameter independent. In both cases the choice $\hat{M}_N = C_M$ with $\hat{Y}_N = C_{YBPXP}$ leads to the worst results. Moreover, the results show that it is necessary to investigate in the preconditioner for \hat{Y}_N in order to get lower iteration numbers. The preconditioner C_M (with exact inversion of Y_N) is pretty good. Therefore it is recommended to use multigrid methods as least for \hat{Y}_N in order to decrease the iteration numbers.

Results for bc-refinement are presented in figure 5.12. There, the iteration numbers are h and p independent except in the case of using the BPX in the preconditioner. The reason therefore is the appearance of hanging nodes, which leads to a $\log(h)$ dependence (see also chapter 3 and figure 3.11).

The best possible iteration numbers yielded with the exact inverses as preconditioners are about 31, the results for the choice $\hat{M}_N = C_M$ and $\hat{Y}_N = C_{YPE}$ are 162 for the finest bc-mesh, whereas the choice $\hat{M}_N = C_M$ with $\hat{Y}_N = C_{YBPXP}$ yields 297 iteration numbers.

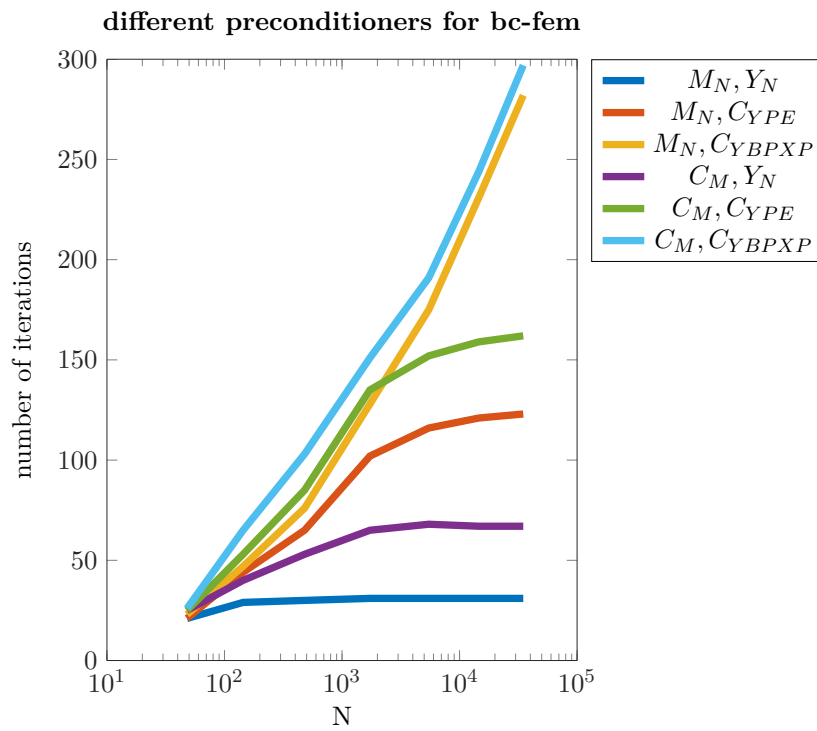
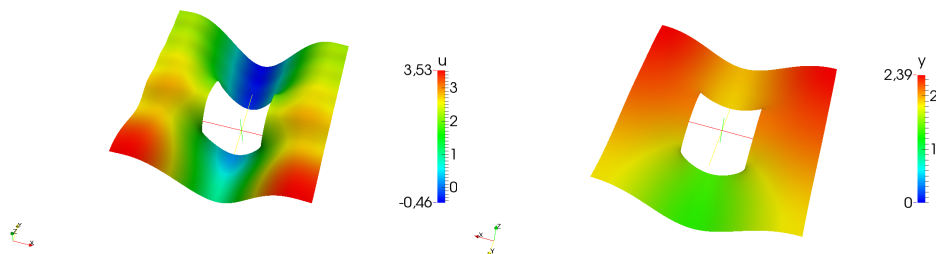


Figure 5.12: iteration numbers for the hole-example

Figure 5.13: control u and state y

In figure 5.13 and in figure 5.14 the control u , the state y , the polynomial degree distribution of the elements for bc-refinement and the adjoint q are plotted. These figures show that the adjoint and the control have the same values, which is due to the fact that the regularization parameter has the value 1 for the given results. The state is between 0 and 2.39, whereas the control and the adjoint are between -0.46 and 3.53 . The polynomial distribution of the elements show that all elements on the boundary – especially all L -corners in the interior – are h -refined. In case of using an edge-orientated refinement – that means h -refinement if an element has an edge on the boundary – the elements in the L -corners have to be added manually since they do not have an edge on the boundary.

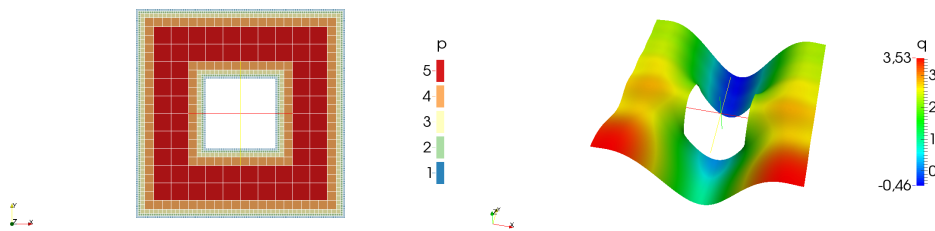


Figure 5.14: polynomial degree p and adjoint q

In order to improve the iteration numbers, different preconditioners – especially multigrid methods – are expected to lead to better results. Furthermore, it is recommended to use triangular elements or elements whose hp -refinement can easily be done without hanging nodes. In that case a projector to ensure getting a conform solution can be avoided, which makes the Schöberl-Zulehner PCG easier applicable for hp -discretizations. Furthermore, it is expected to cause less difficulties with the orthogonality of the basis vectors calculated in the GMRES.

Conclusion and Outlook

In this thesis hp -fem is applied to linear quadratic optimal control problems subject to elliptic differential equations. There, two kind of optimal control problems, optimal boundary control and distributed optimal control are considered in detail. For both problems variational discretization by Hinze [88] is used. This especially has the advantage that the error of the control u can be estimated by the discretization error of the adjoint q .

Optimal boundary control

For the boundary control problems mainly numerical examples are presented.

In the two-dimensional case, instead of the usual bc-refinement proposed in [98] and applied to such kind of problems already in [31], is compared with a Neumann boundary refinement. Moreover, it is suggested to choose a suitable starting mesh with Neumann boundary refinement instead of the usual uniform h -refinement if the vertex-concentrated refinement proposed in [163] is used. This is especially useful in case of oscillating solutions, since a suitable mesh has to be quite fine in order to find all jumps between active and inactive sets.

Furthermore, results in three dimensions are presented. There, first a cube with known analytical solution is discretized with bc-refinement and its solution is calculated with the semismooth Newton method. Moreover, the results are compared with uniform h -fem. The comparison shows that the L_2 -error decreases faster than in uniform h -fem with respect to the number of degrees of freedom. Moreover, the solution to an optimal boundary control problem for different geometries is calculated.

These results show that the proposed bc-refinement works quite well applied to the suggested kind of problems. However, in order to calculate results for more degrees of freedom, it is highly recommended to implement suitable preconditioners, since the condition number of the (inner) equation system becomes quite bad. Nevertheless, since it is not possible to get optimal complexity in case of applying the semismooth Newton method, a rewriting of the problem in a saddle point formulation in combination with suitable preconditioners, is suggested.

Furthermore, in order to reduce the number of degrees of freedom in three dimensions, an extension of the vertex-concentrated refinement applied in [163] for two dimensions would be helpful. In three dimensions it is expected, that an edge-based h -refinement should lead to good results.

Distributed optimal control

In most parts of this thesis a distributed optimal control problem is considered. There, two hp -refinement strategies for such kind of problems are proposed. Both refinement strategies are based on the projection formula, since on all elements where active and inactive parts of u appear, the regularity is lower. Therewith, on all these elements h -refinement is suggested.

The refinement of all other elements can either be performed by using a-priori information on the regularity of the domain, or by using error estimators, for example the error estimators by Melenk and Wohlmuth [114].

In order to solve the problem two different approaches, the semismooth Newton method and the rewriting as saddle point problem are considered.

Semismooth Newton

The semismooth Newton method is applied to two-dimensional problems with active constraints. There, it is important to calculate the integral over the mass matrix on the inactive part of the domain. The drawback herein is, that this leads to an integral over only parts of an element in case of using quadrilateral elements. However, even if an exact calculation of these integrals is theoretically possible, too many cases appear which makes the realization impractical. In order to avoid such difficulties it is suggested to implement triangular elements for hp -fem. On triangular elements the shape of the inactive set of elements with active and inactive parts is no part of a hyperbola as it can be in the quadrilateral case, but a triangle or a quadrangle. Therewith, it is easy to calculate these integrals exactly. Furthermore, by using [142], the exact evaluation of $M_{\mathcal{J}}$ on interface elements for triangular elements and polynomial degree $p = 2$ is possible.

Due to the numerical calculation of the matrix $M_{\mathcal{J}}$ an error occurs. That error can be estimated in further works. It is important to first calculate the error of the inner iteration and then consider the impact of that error to the outer iteration. In order not to increase the overall error, the error has to be below ch^2 .

A further possibility to avoid an error when integrating over the inactive set is to use the conventional discretization approach and discretize the control u with constant elements. For increasing the convergence rate, an extension of [129] with suitable a-posteriori techniques might be possible.

In order to apply the semismooth Newton method in each inner solve, K_N^{-1} has to be applied. In two dimensions this can be done fastly by using direct solvers. However, this is not possible in three dimensions any longer. Therewith, a different approach, the rewriting of the problem as saddle point formulation is considered.

Saddle point formulation

In case of the saddle point formulation a simplified problem, i.e. the box constraints $u_a = -\infty$ and $u_b = \infty$ is considered.

As in [139] the equation system can be written as saddle point problem. In order to solve the problem, different Krylov subspace methods are used.

Due to [139], where an uniform h -fem discretization is taken, by using the non-standard scalar product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ and a suitable preconditioner the equation system gets symmetric and positive definite with respect to that scalar product, which enables the application of the PCG. The advantage of that approach is, that the obtained iteration numbers are independent of the mesh size and the regularization parameter α . In this thesis the work in [139] is extended to hp -fem discretization, which keep the independence of the discretization parameters h and p as well as to the regularization parameter α in case of applying suitable preconditioners. Furthermore, estimates on the work for applying the Schöberl-Zulehner PCG for bc-fem for

certain preconditioners are given. Since optimal complexity (in case of no hanging nodes) is yielded, this is a promising approach.

When applying the Schöberl-Zulehner PCG two parameters, σ and τ have to be estimated in order to get low iteration numbers and keep the positive definiteness of the system matrix. In further work methods to find suitable values of these parameters with low effort can be considered. Moreover, multigrid methods can be used in order to decrease the iteration numbers, since the applied preconditioners still leads to very high iteration numbers.

As second method a preconditioned MINRES is applied. There, as preconditioner a block diagonal preconditioner is used. In case of using the exact inverses in the block diagonal preconditioner, the results of h , p and α independence can be kept. For using proper preconditioners in the blocks, the independence of the discretization parameters h and p is shown. Furthermore, estimates on the cost for applying these methods in case of bc-refinement are given.

Possible extensions to this ansatz are the implementation of the exact inverses as preconditioners and compare those results with the results from the block diagonal preconditioner. Moreover, other classes of preconditioners as multigrid methods should be considered in order to decrease the number of degrees of freedom. In order to avoid the $\log(h)$ dependence for bc-refinement with the BPX preconditioner triangular elements shall be used in further works.

As third and last Krylov subspace method a preconditioned GMRES is applied. Although the results are quite promising especially for the exact inverses as preconditioners, in case of usual preconditioners it is suggested to use truncated GMRES methods in further works. The reason therefore is that in some cases the orthogonality of the basis in the GMRES gets lost due to machine precision.

In general, in further works triangular elements shall be used for numerical examples. Then, suitable routines for handling the hanging nodes can be avoided, which would at least avoid an application of a kind of projector in the Schöberl-Zulehner PCG. That is especially tricky, since the evaluation of the scalar product with arbitrary entries shall be avoided in order to not have to evaluate \hat{S} .

Further possible extensions are taking account of active box constraints in case of applying the saddle point method. Therewith, it might be possible to use similar approaches as in [83], even there the conventional ansatz is used for discretization. However, in that case the h , p and α independence in the Schöberl-Zulehner PCG might get lost.

Moreover, in further work the preconditioners for saddle point problems, whose theoretical results also hold in three dimensions, could be applied in three dimensions. It is expected that this leads – in combination with suitable hp -refinement (analogue to the one presented in here) – to a significant reduction of degrees of freedom and in combination with suitable preconditioners therewith to a faster solution of equation systems.

Bibliography

- [1] R. A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.
- [2] M. Ainsworth and T. O. John. *A posteriori error estimation in finite element analysis*. Pure and applied mathematics. John Wiley, New York, Chichester, Weinheim, 2000. A Wiley-Interscience publication.
- [3] M. Ainsworth and B. Senior. Aspects of an adaptive *hp*-finite element method: adaptive strategy, conforming approximation and efficient solvers. *Comput. Methods Appl. Mech. Engrg.*, 150(1-4):65–87, 1997. Symposium on Advances in Computational Mechanics, Vol. 2 (Austin, TX, 1997).
- [4] H. W. Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Springer Lehrbuch. Springer, 2002.
- [5] T. Apel, J. Pfefferer, and A. Rösch. Graded meshes in optimal control for elliptic partial differential equations: an overview. In *Trends in PDE constrained optimization*, volume 165 of *Internat. Ser. Numer. Math.*, pages 285–302. Birkhäuser/Springer, Cham, 2014.
- [6] T. Apel, J. Pfefferer, and A. Rösch. Finite element error estimates on the boundary with application to optimal control. *Math. Comp.*, 84(291):33–70, 2015.
- [7] T. Apel, A. Rösch, and D. Sirch. L^∞ -error estimates on graded meshes with application to optimal control. *SIAM J. Control Optim.*, 48(3):1771–1796, 2009.
- [8] T. Apel and D. Sirch. A priori mesh grading for distributed optimal control problems. In *Constrained optimization and optimal control for partial differential equations*, volume 160 of *Internat. Ser. Numer. Math.*, pages 377–389. Birkhäuser/Springer Basel AG, Basel, 2012.
- [9] J. H. Argyris. *Energy theorems and structural analysis. A generalised discourse with applications on energy principles of structural analysis including the effects of temperature and non-linear stress-strain relations*. Co-author of Part II, S. Kelsey. Butterworths, London - Toronto - Sydney - Wellington - Durban, 1960.
- [10] V. Arnăutu and P. Neittaanmäki. Discretization estimates for an elliptic control problem. *Numer. Funct. Anal. Optim.*, 19(5-6):431–464, 1998.
- [11] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.*, 19(1):7–32, 1985.

-
- [12] D. N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Discontinuous Galerkin methods for elliptic problems. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 89–101. Springer, Berlin, 2000.
- [13] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02.
- [14] O. Axelsson and M. Neytcheva. Eigenvalue estimates for preconditioned saddle point matrices. *Numer. Linear Algebra Appl.*, 13(4):339–360, 2006.
- [15] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta. Efficient preconditioning for the p -version finite element method in two dimensions. *SIAM J. Numer. Anal.*, 28(3):624–661, 1991.
- [16] I. Babuška, H. C. Elman, and K. Markley. Parallel implementation of the hp -version of the finite element method on a shared-memory architecture. *SIAM J. Sci. Statist. Comput.*, 13(6):1433–1459, 1992.
- [17] I. Babuška and B. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. I. Boundary value problems for linear elliptic equation of second order. *SIAM J. Math. Anal.*, 19(1):172–203, 1988.
- [18] I. Babuška and B. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. II. The trace spaces and application to the boundary value problems with nonhomogeneous boundary conditions. *SIAM J. Math. Anal.*, 20(4):763–781, 1989.
- [19] I. Babuška and B. Guo. Approximation properties of the h - p version of the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 133(3-4):319–346, 1996.
- [20] I. Babuška, B. Guo, and E. P. Stephan. The h - p version of the boundary element method with geometric mesh on polygonal domains. *Comput. Methods Appl. Mech. Engrg.*, 80(1-3):319–325, 1990. Spectral and high order methods for partial differential equations (Como, 1989).
- [21] I. Babuška and M. Suri. The p and h - p versions of the finite element method, basic principles and properties. *SIAM Rev.*, 36(4):578–632, 1994.
- [22] R. E. Bank, T. F. Dupont, and H. Yserentant. The hierarchical basis multigrid method. *Numer. Math.*, 52(4):427–458, 1988.
- [23] R. E. Bank, A. H. Sherman, and A. Weiser. Refinement algorithms and data structures for regular local mesh refinement. In *Scientific computing (Montreal, Que., 1982)*, IMACS Trans. Sci. Comput., I, pages 3–17. IMACS, New Brunswick, NJ, 1983.
- [24] R. E. Bank, B. D. Welfert, and H. Yserentant. A class of iterative methods for solving saddle point problems. *Numer. Math.*, 56(7):645–666, 1990.
- [25] A. Bećirović, P. Paule, V. Pillwein, A. Riese, C. Schneider, and J. Schöberl. Hypergeometric summation algorithms for high-order finite elements. *Computing*, 78(3):235–249, 2006.

-
- [26] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: basic concept. *SIAM J. Control Optim.*, 39(1):113–132 (electronic), 2000.
- [27] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.
- [28] M. Benzi and V. Simoncini. On the eigenvalues of a class of saddle point matrices. *Numer. Math.*, 103(2):173–196, 2006.
- [29] S. Beuchler. Wavelet solvers for *hp*-FEM discretizations in 3D using hexahedral elements. *Comput. Methods Appl. Mech. Engrg.*, 198(13-14):1138–1148, 2009.
- [30] S. Beuchler, T. Eibner, and U. Langer. Primal and dual interface concentrated iterative substructuring methods. *SIAM J. Numer. Anal.*, 46(6):2818–2842, 2008.
- [31] S. Beuchler, K. Hofer, D. Wachsmuth, and J.-E. Wurst. Boundary concentrated finite elements for optimal control problems with distributed observation. *Comput. Optim. Appl.*, 62(1):31–65, 2015.
- [32] S. Beuchler, C. Pechstein, and D. Wachsmuth. Boundary concentrated finite elements for optimal boundary control problems of elliptic PDEs. *Comput. Optim. Appl.*, 51(2):883–908, 2012.
- [33] S. Beuchler, M. Pester, and A. Meyer. Spc-pm3adh v1.0 – programmer’s manual. *Preprintreihe des Chemnitzer SFB 393*, (01-08):1–79, 2005.
- [34] S. Beuchler, V. Pillwein, J. Schöberl, and S. Zaglmayr. Sparsity optimized high order finite element functions on simplices. In *Numerical and symbolic scientific computing*, Texts Monogr. Symbol. Comput., pages 21–44. SpringerWienNewYork, Vienna, 2012.
- [35] S. Beuchler and M. Purrucker. Schwarz type solvers for *hp*-FEM discretizations of mixed problems. *Comput. Methods Appl. Math.*, 12(4):369–390, 2012.
- [36] D. Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag, Berlin Heidelberg New York, 2003.
- [37] J. H. Bramble. *Multigrid methods*, volume 294 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, 1993.
- [38] J. H. Bramble and J. E. Pasciak. Corrigenda: “A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems”. *Math. Comp.*, 51(183):387–388, 1988.
- [39] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181):1–17, 1988.
- [40] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring. I. *Math. Comp.*, 47(175):103–134, 1986.

-
- [41] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring. II. *Math. Comp.*, 49(179):1–16, 1987.
- [42] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.
- [43] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.
- [44] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.
- [45] X.-C. Cai and O. B. Widlund. Multiplicative Schwarz algorithms for some nonsymmetric and indefinite problems. *SIAM J. Numer. Anal.*, 30(4):936–952, 1993.
- [46] G. F. Carey and J. T. Oden. *Finite elements. Vol. VI*. The Texas Finite Element Series, VI. Prentice Hall, Inc., Englewood Cliffs, NJ, 1986. Fluid mechanics.
- [47] E. Casas. Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems. *Adv. Comput. Math.*, 26(1-3):137–153, 2007.
- [48] E. Casas and M. Mateos. Error estimates for the numerical approximation of Neumann control problems. *Comput. Optim. Appl.*, 39(3):265–295, 2008.
- [49] E. Casas, M. Mateos, and F. Tröltzsch. Necessary and sufficient optimality conditions for optimization problems in function spaces and applications to control theory. In *Proceedings of 2003 MODE-SMAI Conference*, volume 13 of *ESAIM Proc.*, pages 18–30 (electronic). EDP Sci., Les Ulis, 2003.
- [50] Y. Chen, F. Huang, N. Yi, and W. Liu. A Legendre-Galerkin spectral method for optimal control problems governed by Stokes equations. *SIAM J. Numer. Anal.*, 49(4):1625–1648, 2011.
- [51] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.
- [52] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49:1–23, 1943.
- [53] K. Deckelnick, A. Günther, and M. Hinze. Finite element approximation of elliptic control problems with constraints on the gradient. *Numer. Math.*, 111(3):335–350, 2009.
- [54] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM J. Numer. Anal.*, 45(5):1937–1953 (electronic), 2007.
- [55] K. Deckelnick and M. Hinze. A-priori error bounds for finite element approximation of elliptic optimal control problems with gradient constraints. In *Trends in PDE constrained optimization*, volume 165 of *Internat. Ser. Numer. Math.*, pages 365–382. Birkhäuser/Springer, Cham, 2014.

- [56] M. Delfour, W. Hager, and F. Trochu. Discontinuous Galerkin methods for ordinary differential equations. *Math. Comp.*, 36(154):455–473, 1981.
- [57] L. Demkowicz. *Computing with hp-adaptive finite elements. Vol. 1.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2007. One and two dimensional elliptic and Maxwell problems, With 1 CD-ROM (UNIX).
- [58] L. Demkowicz, J. Kurtz, D. Pardo, M. Paszyński, W. Rachowicz, and A. Zdunek. *Computing with hp-adaptive finite elements. Vol. 2.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2008. Frontiers: three dimensional elliptic and Maxwell problems with applications.
- [59] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.
- [60] T. Eibner. *Randkonzentrierte und adaptive hp-fem*. PhD thesis, Technische Universität Chemnitz, 1 2006.
- [61] T. Eibner and J. M. Melenk. Fast algorithms for setting up the stiffness matrix in hp-fem: a comparison. *Preprintreihe des Chemnitzer SFB 393*, (05-08):1–43, 2005.
- [62] T. Eibner and J. M. Melenk. A local error analysis of the boundary-concentrated hp-FEM. *IMA J. Numer. Anal.*, 26(4):752–778, 2006.
- [63] T. Eibner and J. M. Melenk. Multilevel preconditioning for the boundary concentrated hp-FEM. *Comput. Methods Appl. Mech. Engrg.*, 196(37-40):3713–3725, 2007.
- [64] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, second edition, 2014.
- [65] V. Faber, J. Liesen, and P. Tichý. Properties of worst-case GMRES. *SIAM J. Matrix Anal. Appl.*, 34(4):1500–1519, 2013.
- [66] R. S. Falk. Nonconforming finite element methods for the equations of linear elasticity. *Math. Comp.*, 57(196):529–550, 1991.
- [67] R. Fletcher. *Practical methods of optimization. Vol. 2.* John Wiley & Sons, Ltd., Chichester, 1981. Constrained optimization, A Wiley-Interscience Publication.
- [68] B. Galerkin. Beams and plates. series in some questions of elastic equilibrium of beams and plates. *Vestnik Ingenerov*, 19:897–908, 1915. In Russian.
- [69] M. J. Gander and G. Wanner. From Euler, Ritz, and Galerkin to modern computing. *SIAM Rev.*, 54(4):627–666, 2012.
- [70] A. George and J. W. H. Liu. *Computer solution of large sparse positive definite systems*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981. Prentice-Hall Series in Computational Mathematics.

-
- [71] G. H. Golub, C. Greif, and J. M. Varah. An algebraic analysis of block diagonal preconditioner for saddle point systems. *SIAM J. Matrix Anal. Appl.*, 27(3):779–792 (electronic), 2005.
- [72] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.
- [73] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
- [74] M. Griebel. *Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen*. Teubner Skripten zur Numerik. [Teubner Scripts on Numerical Mathematics]. B. G. Teubner, Stuttgart, 1994.
- [75] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
- [76] C. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen*. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks]. B. G. Teubner, Stuttgart, second edition, 1994.
- [77] B. Guo and I. Babuška. Regularity of the solutions for elliptic problems on nonsmooth domains in \mathbf{R}^3 . I. Countably normed spaces on polyhedral domains. *Proc. Roy. Soc. Edinburgh Sect. A*, 127(1):77–126, 1997.
- [78] B. Guo and I. Babuška. Regularity of the solutions for elliptic problems on nonsmooth domains in \mathbf{R}^3 . II. Regularity in neighbourhoods of edges. *Proc. Roy. Soc. Edinburgh Sect. A*, 127(3):517–545, 1997.
- [79] G. Haase, U. Langer, and A. Meyer. The approximate Dirichlet domain decomposition method. I. An algebraic approach. *Computing*, 47(2):137–151, 1991.
- [80] G. Haase, U. Langer, and A. Meyer. The approximate Dirichlet domain decomposition method. II. Applications to 2nd-order elliptic BVPs. *Computing*, 47(2):153–167, 1991.
- [81] W. Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985.
- [82] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Vieweg + Teubner, [online-ausg.] edition, 2009. Martin Hanke-Bourgeois.
- [83] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.
- [84] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

-
- [85] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (2003), 2002.
- [86] M. Hintermüller and M. Ulbrich. A mesh-independence result for semismooth Newton methods. *Math. Program.*, 101(1, Ser. B):151–184, 2004.
- [87] E. Hinton and D. R. J. Owen. *An introduction to finite element computations*. Pineridge Press Ltd., Swansea, 1979.
- [88] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30(1):45–61, 2005.
- [89] M. Hinze and U. Matthes. A note on variational discretization of elliptic Neumann boundary control. *Control Cybernet.*, 38(3):577–591, 2009.
- [90] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
- [91] M. Hinze and F. Tröltzsch. Discrete concepts versus error analysis in PDE-constrained optimization. *GAMM-Mitt.*, 33(2):148–162, 2010.
- [92] M. Hinze and M. Vierling. A globalized semi-smooth Newton method for variational discretization of control constrained elliptic optimal control problems. In *Constrained optimization and optimal control for partial differential equations*, volume 160 of *Internat. Ser. Numer. Math.*, pages 171–182. Birkhäuser/Springer Basel AG, Basel, 2012.
- [93] M. Jung and U. Langer. *Methode der finiten Elemente für Ingenieure*. Springer, 2013.
- [94] M. Jung, U. Langer, A. Meyer, W. Queck, and M. Schneider. Multigrid preconditioners and their applications. In *Third Multigrid Seminar (Biesenthal, 1988)*, volume 89 of *Rep. MATH*, pages 11–52. Akad. Wiss. DDR, Berlin, 1989.
- [95] C. Kanzow. *Numerik linearer Gleichungssysteme: Direkte und iterative Verfahren*. Springer-Lehrbuch. Springer, 2004.
- [96] G. E. Karniadakis and S. J. Sherwin. *Spectral/hp element methods for computational fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, second edition, 2005.
- [97] B. N. Khoromskij and J. M. Melenk. An efficient direct solver for the boundary concentrated FEM in 2D. *Computing*, 69(2):91–117, 2002.
- [98] B. N. Khoromskij and J. M. Melenk. Boundary concentrated finite element methods. *SIAM J. Numer. Anal.*, 41(1):1–36, 2003.
- [99] N. Kikuchi. *Finite Element Methods in Mechanics*. Cambridge University Press, Cambridge Cambridgeshire ; New York, 1986.
- [100] V. Korneev, U. Langer, and L. S. Xanthis. On fast domain decomposition solving procedures for *hp*-discretizations of 3-D elliptic problems. *Comput. Methods Appl. Math.*, 3(4):536–559 (electronic), 2003.

-
- [101] V. G. Korneev and S. Jensen. Domain decomposition preconditioning in the hierarchical p -version of the finite element method. *Appl. Numer. Math.*, 29(4):479–518, 1999.
- [102] K. Kunisch and B. Vexler. Constrained Dirichlet boundary control in L^2 for a class of evolution equations. *SIAM J. Control Optim.*, 46(5):1726–1753, 2007.
- [103] A. Kunoth. Fast iterative solution of saddle point problems in optimal control based on wavelets. *Comput. Optim. Appl.*, 22(2):225–259, 2002.
- [104] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Research Nat. Bur. Standards*, 45:255–282, 1950.
- [105] G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, editors. *Trends in PDE constrained optimization*, volume 165 of *International Series of Numerical Mathematics*. Birkhäuser/Springer, Cham, 2014.
- [106] J. Liesen and P. Tichý. Convergence analysis of Krylov subspace methods. *GAMM Mitt. Ges. Angew. Math. Mech.*, 27(2):153–173 (2005), 2004.
- [107] J. Liesen and P. Tichý. The worst-case GMRES for normal matrices. *BIT*, 44(1):79–98, 2004.
- [108] J.-L. Lions. *Optimal control of systems governed by partial differential equations*. Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York-Berlin, 1971.
- [109] J.-F. Maitre and O. Pourquier. Condition number and diagonal preconditioning: comparison of the p -version and the spectral element methods. *Numer. Math.*, 74(1):69–84, 1996.
- [110] H. Maurer and H. D. Mittelmann. Optimization techniques for solving elliptic control problems with control and state constraints. II. Distributed control. *Comput. Optim. Appl.*, 18(2):141–160, 2001.
- [111] S. McCormick. Book Review: An introduction to multigrid methods. *Bull. Amer. Math. Soc. (N.S.)*, 28(2):373–375, 1993.
- [112] A. Meister. *Numerik linearer Gleichungssysteme*. Friedr. Vieweg & Sohn, Braunschweig, 1999. Eine Einführung in moderne Verfahren. [An introduction to modern procedures].
- [113] J. M. Melenk, K. Gerdes, and S. Christoph. Fully discrete hp-finite elements: Fast quadrature. *Research Report ETH Zürich*, (99-15):1–28, 1999.
- [114] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in hp-FEM. *Adv. Comput. Math.*, 15(1-4):311–331 (2002), 2001. A posteriori error estimation and adaptive computational methods.
- [115] G. Meurant and J. Duintjer Tebbens. The role eigenvalues play in forming GMRES residual norms with non-normal matrices. *Numer. Algorithms*, 68(1):143–165, 2015.
- [116] A. Meyer. Projected pcgm for handling hanging nodes in adaptive finite element procedures. *Preprintreihe des Chemnitzer SFB 393*, (99-25):1–14, 1999.

-
- [117] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. *SIAM J. Control Optim.*, 43(3):970–985 (electronic), 2004.
- [118] C. Meyer and A. Rösch. L^∞ -estimates for approximated optimal control problems. *SIAM J. Control Optim.*, 44(5):1636–1649 (electronic), 2005.
- [119] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972 (electronic), 2000.
- [120] J. T. Oden. Finite elements: an introduction. In *Handbook of numerical analysis, Vol. II*, Handb. Numer. Anal., II, pages 3–15. North-Holland, Amsterdam, 1991.
- [121] J. T. Oden and G. F. Carey. *Finite elements. Vol. V*. The Texas Finite Element Series, V. Prentice Hall, Inc., Englewood Cliffs, NJ, 1984. Special problems in solid mechanics.
- [122] S. A. Orszag. Spectral methods for problems in complex geometries. *J. Comput. Phys.*, 37(1):70–92, 1980.
- [123] D. R. J. Owen and E. Hinton. *Finite elements in plasticity: theory and practice*. Pineridge Press Ltd., Swansea, 1980.
- [124] C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.
- [125] L. F. Pavarino. Additive Schwarz methods for the p -version finite element method. *Numer. Math.*, 66(4):493–515, 1994.
- [126] V. Pillwein. Private communication.
- [127] A. Quarteroni and A. Valli. *Domain decomposition methods for partial differential equations*. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1999. Oxford Science Publications.
- [128] W. Ritz. Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. *J. Reine Angew. Math.*, 135:1–61, 1908.
- [129] A. Rösch and D. Wachsmuth. A-posteriori error estimates for optimal control problems with state and control constraints. *Numer. Math.*, 120(4):733–762, 2012.
- [130] M. Rozložník and V. Simoncini. Krylov subspace methods for saddle point problems with indefinite preconditioning. *SIAM J. Matrix Anal. Appl.*, 24(2):368–391, 2002.
- [131] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.
- [132] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.
- [133] S. A. Sauter and C. Schwab. *Boundary element methods*, volume 39 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2011. Translated and expanded from the 2004 German original.

-
- [134] W. Sautter. *Fehlerfortpflanzung und Rundungsfehler bei der verallgemeinerten Inversion von Matrizen*. PhD thesis, Technische Universität München, 7 1971.
- [135] M. Schanz and O. Steinbach, editors. *Boundary element analysis*, volume 29 of *Lecture Notes in Applied and Computational Mechanics*. Springer, Berlin, 2007. Mathematical aspects and applications.
- [136] J. Schöberl, J. M. Melenk, C. Pechstein, and S. Zanglmayr. Additive Schwarz preconditioning for p -version triangular and tetrahedral finite elements. *IMA J. Numer. Anal.*, 28(1):1–24, 2008.
- [137] J. Schöberl, J. M. Melenk, C. G. A. Pechstein, and S. C. Zanglmayr. Schwarz preconditioning for high order simplicial finite elements. In *Domain decomposition methods in science and engineering XVI*, volume 55 of *Lect. Notes Comput. Sci. Eng.*, pages 139–150. Springer, Berlin, 2007.
- [138] J. Schöberl, R. Simon, and W. Zulehner. A robust multigrid method for elliptic optimal control problems. *SIAM J. Numer. Anal.*, 49(4):1482–1503, 2011.
- [139] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773 (electronic), 2007.
- [140] C. Schwab. *p - and hp -finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.
- [141] H. A. Schwarz. *Gesammelte mathematische Abhandlungen. Band I, II*. Chelsea Publishing Co., Bronx, N.Y., 1972. Nachdruck in einem Band der Auflage von 1890.
- [142] D. Sevilla and D. Wachsmuth. Polynomial integration on regions defined by a triangle and a conic. In *ISSAC 2010—Proceedings of the 2010 International Symposium on Symbolic and Algebraic Computation*, pages 163–170. ACM, New York, 2010.
- [143] R. Simon and W. Zulehner. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numer. Math.*, 111(3):445–468, 2009.
- [144] V. Simoncini. On the convergence of restarted Krylov subspace methods. *SIAM J. Matrix Anal. Appl.*, 22(2):430–452, 2000.
- [145] V. Simoncini. Block triangular preconditioners for symmetric saddle-point problems. *Appl. Numer. Math.*, 49(1):63–80, 2004.
- [146] V. Simoncini. On a non-stagnation condition for GMRES and application to saddle point matrices. *Electron. Trans. Numer. Anal.*, 37:202–213, 2010.
- [147] B. F. Smith. An optimal domain decomposition preconditioner for the finite element solution of linear elasticity problems. *SIAM J. Sci. Statist. Comput.*, 13(1):364–378, 1992.

- [148] B. F. Smith. Domain decomposition methods for partial differential equations. In *Parallel numerical algorithms (Hampton, VA, 1994)*, volume 4 of *ICASE/LaRC Interdiscip. Ser. Sci. Eng.*, pages 225–243. Kluwer Acad. Publ., Dordrecht, 1997.
- [149] B. F. Smith, P. E. Bjørstad, and W. D. Gropp. *Domain decomposition*. Cambridge University Press, Cambridge, 1996. Parallel multilevel methods for elliptic partial differential equations.
- [150] P. Šolín, J. Červený, and I. Doležel. Arbitrary-level hanging nodes and automatic adaptivity in the *hp*-FEM. *Math. Comput. Simulation*, 77(1):117–132, 2008.
- [151] O. Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme. Finite Elemente und Randelemente*. Teubner, 2003.
- [152] B. Szabó and I. Babuška. *Finite element analysis*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1991.
- [153] L. T. Tenek and J. Argyris. *Finite element analysis for composite structures*, volume 59 of *Solid Mechanics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1998. With 1 IBM-PC floppy disc (3.5 inch; HD).
- [154] P. Tichý, J. Liesen, and V. Faber. On worst-case GMRES, ideal GMRES, and the polynomial numerical hull of a Jordan block. *Electron. Trans. Numer. Anal.*, 26:453–473, 2007.
- [155] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.
- [156] F. G. Tricomi. *Vorlesungen über Orthogonalreihen*. Zweite, Korrigierte Auflage. Die Grundlehren der mathematischen Wissenschaften, Band 76. Springer-Verlag, Berlin-New York, 1970.
- [157] H. Triebel. *Interpolation theory, function spaces, differential operators*, volume 18 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam-New York, 1978.
- [158] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*. Vieweg+Teubner Verlag, 2005.
- [159] M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–842 (2003), 2002.
- [160] M. Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, volume 11 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.
- [161] R. Unger. *Unterraum-CG-Techniken zur Bearbeitung von Kontaktproblemen*. PhD thesis, Technische Universität Chemnitz, 2 2007.

-
- [162] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Advances in numerical mathematics. Wiley-Teubner, 1996.
- [163] D. Wachsmuth and J.-E. Wurst. Exponential convergence of hp -finite element discretization of optimal boundary problems with elliptic partial differential equations. *Preprint 328, Institut für Mathematik, Universität Würzburg*, 2015.
- [164] D. Wachsmuth and J.-E. Wurst. An interior point method designed for solving linear quadratic optimal control problems with hp finite elements. *Optimization Methods and Software*, 30(6):1276–1302, 2015.
- [165] D. Wachsmuth and J.-E. Wurst. Optimal control of interface problems with hp -finite elements. *Numerical Functional Analysis and Optimization*, to appear 2016.
- [166] J.-E. Wurst. *Hp-Finite Elements for PDE-Constrained Optimization*. PhD thesis, Universität Würzburg, 4 2015.
- [167] H. Yserentant. Über die Konvergenz von Mehrgitterverfahren für nichtuniform verfeinerte Familien von Gittern. *Z. Angew. Math. Mech.*, 64(5):324–326, 1984.
- [168] S. Zaglmayr. *High Order Finite Element Methods for Electromagnetic Field Computation*. PhD thesis, Johannes Kepler Universität, Linz, 7 2006.
- [169] X. Zhang. Multilevel Schwarz methods. *Numer. Math.*, 63(4):521–539, 1992.
- [170] W. Zulehner. Analysis of iterative methods for saddle point problems: a unified approach. *Math. Comp.*, 71(238):479–505 (electronic), 2002.
- [171] W. Zulehner. *Numerische Mathematik: eine Einführung anhand von Differentialgleichungsproblemen. Band 1. Stationäre Probleme*. Mathematik Kompakt. [Compact Mathematics]. Birkhäuser Verlag, Basel, 2008.
- [172] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 32(2):536–560, 2011.