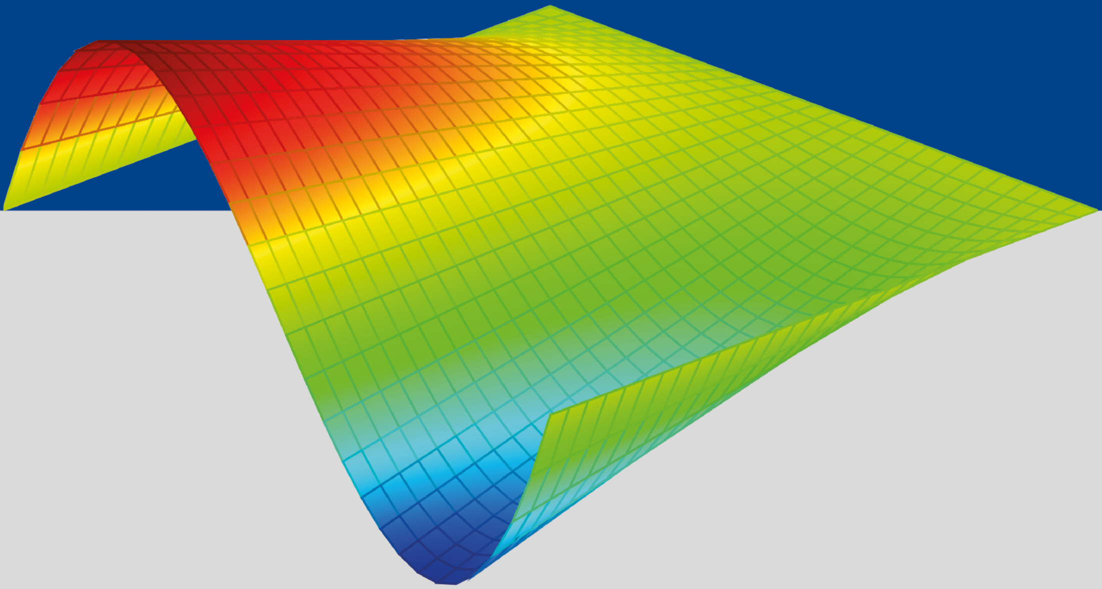


Jan-Eric Wurst

Hp-Finite Elements for PDE-Constrained Optimization

Würzburg University Press

Jan-Eric Wurst

Hp-Finite Elements for PDE-Constrained Optimization

Jan-Eric Wurst

Hp-Finite Elements for PDE-Constrained Optimization

Würzburg
University Press

Dissertation, Julius-Maximilians-Universität Würzburg
Fakultät für Mathematik und Informatik, 2015
Gutachter: Prof. Dr. Daniel Wachsmuth, Prof. Dr. Boris Vexler

Impressum

Julius-Maximilians-Universität Würzburg
Würzburg University Press
Universitätsbibliothek Würzburg
Am Hubland
D-97074 Würzburg
www.wup.uni-wuerzburg.de

Print on Demand

ISBN 978-3-95826-024-5 (print)
ISBN 978-3-95826-025-2 (online)
URN urn:nbn:de:bvb:20-opus-115027

Contents

Preface

Abstract

This thesis deals with the hp-finite element method (FEM) for linear quadratic optimal control problems. Here, a tracking type functional with control costs as regularization shall be minimized subject to an elliptic partial differential equation. In the presence of control constraints, the first order necessary conditions, which are typically used to find optimal solutions numerically, can be formulated as a semi-smooth projection formula. Consequently, optimal solutions may be non-smooth as well. The hp-discretization technique considers this fact and approximates rough functions on fine meshes while using higher order finite elements on domains where the solution is smooth.

The first main achievement of this thesis is the successful application of hp-FEM to two related problem classes: Neumann boundary and interface control problems. They are solved with an a-priori refinement strategy called boundary concentrated (bc) FEM and interface concentrated (ic) FEM, respectively. These strategies generate grids that are heavily refined towards the boundary or interface. We construct an elementwise interpolant that allows to prove algebraic decay of the approximation error for both techniques. Additionally, a detailed analysis of global and local regularity of solutions, which is critical for the speed of convergence, is included.

Since the bc- and ic-FEM retain small polynomial degrees for elements touching the boundary and interface, respectively, we are able to deduce novel error estimates in the L^2- and L^∞-norm. The latter allows an a-priori strategy for updating the regularization parameter in the objective functional to solve bang-bang problems.

Furthermore, we apply the traditional idea of the hp-FEM, i.e., grading the mesh geometrically towards vertices of the domain, for solving optimal control problems (vc-FEM). In doing so, we obtain exponential convergence with respect to the number of unknowns. This is proved with a regularity result in countably normed spaces for the variables of the coupled optimality system.

The second main achievement of this thesis is the development of a fully adaptive hp-interior point method that can solve problems with distributed or Neumann control. The underlying barrier problem yields a non-linear optimality system, which

poses a numerical challenge: the numerically stable evaluation of integrals over possibly singular functions in higher order elements. We successfully overcome this difficulty by monitoring the control variable at the integration points and enforcing feasibility in an additional smoothing step.

In this work, we prove convergence of an interior point method with smoothing step and derive a-posteriori error estimators. The adaptive mesh refinement is based on the expansion of the solution in a Legendre series. The decay of the coefficients serves as an indicator for smoothness that guides between h- and p-refinement.

Acknowledgements

I owe many thanks to my supervisor Prof. Dr. Daniel Wachsmuth for providing me the opportunity to work on the project 'higher-order finite element methods for optimal control problems'. He introduced me to this interesting topic, gave many valuable ideas and was constantly willing to discuss problems occurring both in theoretical and numerical mathematics. Through his patience and professional guidance he created a brilliant scientific atmosphere which finally lead to this thesis and the submission of four scientific papers.

Many thanks also go out to Prof. Dr. Sven Beuchler who answered many questions regarding the hp-finite element discretization. He provided access to the Fortran code that handles the storage and basic refinement of the FE grid. Furthermore, the code[1] that was developed at TU Chemnitz under SFB 393 'Parallele Numerische Simulation für Physik und Kontinuumsmechanik' was used for visualizing two dimensional FE solutions.

I also express my appreciation to Prof. Dr. Boris Vexler who agreed to co-referee this thesis.

Additionally, I gratefully acknowledge the support of my colleague Katharina Hofer, who substantially helped to develop the C/C++ code that provides the numerical results for this thesis. Many fruitful discussions on computational issues lead to a flexible and efficient implementation of hp-finite elements for optimal control problems.

Moreover, I thank the members of the institute of mathematics of the Julius-Maximilians-Universität Würzburg, ahead all colleagues of the chair of optimization and numerical mathematics lead by Prof. Dr. Kanzow. I appreciated the very pleasant working atmosphere and supportive team spirit.

I also enjoyed the support of my family and friends, which helped to persevere in times when one step towards a mathematical solution was followed by (what felt like) two steps backwards. The quality of this thesis was improved by fruitful advice from Dr. Axel Dreves, Dr. Monika Wolfmayr, and Juri Merger.

Last but not least, I thank the Austrian Science Fund FWF which financially supported me under grant P23484-N18 during my time in Würzburg.

Jan-Eric Wurst,
Würzburg, February 2015

[1] http://www-user.tu-chemnitz.de/~pester/graf2d/

List of Symbols and Abbreviations

General Notation

$\exists$	There exists
$\exists!$	There exists exactly one
$\forall$	For all
$\vee$	Logical ’or’
$\wedge$	Logical ’and’
$\mathbb{N}$	The set of natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$
$\mathbb{Z}$	The set of integers
$\mathbb{R}, \mathbb{R}^+, \mathbb{R}_0^+$	The real, positive real, and non-negative real numbers
$\mathbb{R}^n$	n-dimensional vector space of real numbers
$e_i \in \mathbb{R}^n$	The unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ with 1 located at the i-th position
$\overline{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{C}$	The field of complex numbers
$\Re(\lambda)$	The real part of a complex number λ
$\Im(\lambda)$	The imaginary part of a complex number λ
$\lvert \cdot \rvert_\infty$	The maximum norm in $\mathbb{R}^n$
$\mathrm{meas}_n(M)$	The n-dimensional Lebesgue measure of a measurable set M
$\#M$	The cardinality of a set M
α, β	A multi-index $(\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ (analogous for β)
$\lvert\alpha\rvert$	$\sum_{i=1}^n \alpha_i$
$\alpha!$	$\prod_{i=1}^n \alpha_i!$
$\mathrm{dist}(\cdot, \cdot)$	The distance of two points, sets, or a set and a point in $\mathbb{R}^n$

Partial Differential Equations

Ω	An open, bounded, and polygonal domain of $\mathbb{R}^2$
$\partial\Omega, \Gamma$	The boundary of Ω
$\overline{\Omega}$	The closure of Ω, i.e., $\Omega \cup \partial\Omega$
$\{\Omega_i\}_{i \in I}$	A $2d$-network of pairwise disjoint, polygonal subdomains, such that $\overline{\Omega} = \bigcup_{i \in I} \overline{\Omega}_i$

$\Gamma_{\mathcal{D}}$	The subset of Γ with Dirichlet boundary conditions
$\Gamma_{\mathcal{N}}$	The subset of Γ with Neumann boundary conditions
$\mathcal{I}$	The interface $\bigcup_{i\in I} \partial\Omega_i \setminus \partial\Omega$
$\mathcal{X}$	The set of vertices of Ω or $\{\Omega_i\}_{i\in I}$
ω_X	The opening angle in $(0, 2\pi)$ at a vertex $X \in \mathcal{X}$
$\mathcal{V} \supset \mathcal{X}$	A finite set of points from Γ
$\partial_{x_i}^k v$	k-th partial derivative of v in direction x_i
D^α	The differential operator $\partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n}$
∇	The differential operator $(\partial_{x_i})_{i=1,\dots,n}$ as column vector
$\nabla^p v$	$\lvert\nabla^p v\rvert^p = \sum\limits_{\alpha_1=1,\dots,\alpha_p=1}^{n} \lvert\partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n} v\rvert^p = \sum\limits_{\lvert\alpha\rvert=p} \frac{p!}{\alpha!} \lvert D^\alpha v\rvert^p$
$a(\cdot,\cdot)$	A bounded, coercive bilinear form
$A = -\nabla \cdot (D\nabla) + c$	An elliptic differential operator with matrix valued D and scalar valued c
B	An operator that maps data into the image space of A
I	The unit matrix
id	The identity mapping
$v \equiv c \in \mathbb{R}$	emphasizes that a function f takes constant values
Δv	The Laplacian $\nabla \cdot \nabla v$
$v\vert_{\Gamma_{\mathcal{D}}}$	The restriction/trace of a function v on $\Gamma_{\mathcal{D}}$
∂_n	The normal derivative $\vec{n} \cdot \nabla$ with the outward unit normal vector $n = \vec{n}$
∂_{n_D}	The co-normal derivative $D(x)n \cdot \nabla$
(ρ, θ)	Polar coordinates centered at the origin
$B_r(x)$	An open ball of radius $r > 0$ around $x \in \mathbb{R}^n$
$B_r(x)^+$	The half ball $B_r(x) \cap \{x_n = 0\}$
$\mathcal{B}$	A collection of balls

Functional Analysis (see Subsection 2.1.1)

$\Vert \cdot \Vert_V$	The norm in the vector space V
$\dim V$	The dimension of a finite dimensional vector space V
$\mathcal{L}(V, W)$	The set of linear and continuous operators from V to W
V^*	The dual space $\mathcal{L}(V, \mathbb{R})$
$(\cdot,\cdot)_H$	The inner product in a Hilbert space H
$\langle\cdot,\cdot\rangle_{V^*,V}$	The duality pairing of elements from V^*, V
$\to$	Strong convergence
$\rightharpoonup$	Weak convergence
$\hookrightarrow$	Continuous embedding

Function Spaces (see Subsection 2.1.3)

$C^k(\Omega)$	The vector space of k-times continuously differentiable functions with image in $\mathbb{R}$
$C^k(\overline{\Omega})$	Functions in $C^k(\Omega)$ where each derivative continuously extends to $\overline{\Omega}$
$C^{k,\beta}(\Omega)$	Functions in $C^k(\Omega)$ whose k-th derivative is Hölder-continuous of order $0 < \beta \leq 1$
$C^\infty(\Omega)$	$\bigcap_{k=1}^\infty C^k(\Omega)$
$C_c^\infty(\Omega)$	Functions from $f \in C^\infty(\Omega)$ with compact support in Ω
$L^p(\Omega)$	The Banach space of p-times Lebesgue-integrable functions
$L^\infty(\Omega)$	The Banach space of essentially bounded functions
$W^{k,p}(\Omega)$	Sobolev space of functions whose weak derivatives up to order k lie in $L^p(\Omega)$
$W^{s,p}(\Omega)$	Sobolev-Slobodeckij with $s = k + \sigma$ and $\sigma \in (0,1)$
$H^s(\Omega)$	The Hilbert space $W^{s,2}(\Omega)$
$H^1_{\Gamma_\mathcal{D}}(\Omega)$	The subspace of $H^1(\Omega)$ with vanishing trace on $\Gamma_\mathcal{D}$
$r_\Gamma, r_\mathcal{V}$	Weight functions measuring distances to $\Gamma, \mathcal{V}$, respectively
$H^{l,l}_\beta(\Omega)$	Space of functions whose highest derivative belongs to a weighted $L^2(\Omega)$-space
$B^l_\beta(\Omega, C, \gamma)$	Countably normed space with specific constants $C, \gamma > 0$

Optimal Control

(**P**)	The model problem
(**P**$_h$)	A discretized version of (**P**)
J	The objective functional
(N)	The constraining Neumann problem
(T)	The constraining transmission problem
y_d	The desired state
u, y, q	The control, state, and adjoint variable
u_a, u_b	The box constraints with $u_a \leq u_b$
$U \subset \overline{\Omega}$	The domain where the control acts
U_{ad}	The admissible set $U_{ad} := \{u \in L^2(U) \mid u_a \leq u \leq u_b$ almost everywhere in $U\}$
$\nu \geq 0$	The regularization parameter
$\mu > 0$	The homotopy parameter for interior point methods
$(u, y, q)^*$	Optimal variables solving the optimization problem (**P**)
$(u, y, q)^*_h$	Discrete, optimal variables solving problem (**P**$_h$)
$(u, y, q)_\mu$	The central path (see Definition 6.1.8)
$P_{U_{ad}}$	The projection operator onto the feasible set U_{ad}
$\mathfrak{A}$	The active set $\{x \in U \mid u^*(x) = u_a(x) \ \vee \ u^*(x) = u_b(x)\}$

Finite Elements (see Chapter 4)

τ	An admissible triangulations of Ω
τ_h	A boundary concentrated mesh of boundary size $h < 1$
τ_ς^m	A geometric mesh boundary with $m + 1 \geq 1$ layers and grading factor $\varsigma \in (0, 1)$
K	A finite element $K \in \tau$
$\hat{K}$	The reference element
F_K	The Diffeomorphism that maps $\hat{K} \to K$
p_K, h_K	The polynomial degree and diameter, respectively, associated to an element K
$\mathbf{p}$	The polynomial degree vector $\mathbf{p} := (P_K)_{K\in\tau}$
α	The slope of a linear $\mathbf{p}$ (see Definitions 4.3.2, 4.4.3)
$S^{\mathbf{p}}(\tau)$	The *hp*-finite element ansatz/trial space (see Definition 4.1.8)
N	The number of unknowns/degrees of freedom $N := \dim S^{\mathbf{p}}(\tau)$
Φ	The set of basis functions spanning $S^{\mathbf{p}}(\tau)$
$\mathcal{A}$	The matrix that is obtained after discretizing the differential operator A
$\mathcal{K}, \mathcal{M}$	The stiffness and mass matrix

Abbreviations

PDE(s)	Partial differential equation(s)
FE(M)	Finite element (method)
bc	Boundary concentrated
ic	Interface concentrated
vc	Vertex concentrated
dof	Degree of freedom (plural: ddof)
a.e.	Almost everywhere
a.a.	Almost all
i.m.p.	Injective modulo polynomials
h-refine	Split elements $K \in \tau$ into more elements of diameter smaller than h_K
p-refine	Increase the polynomial degree p_K of elements $K \in \tau$

CHAPTER 1

Introduction

Optimal control theory is a versatile mathematical discipline with applications in many fields. It has gained interest over the last decades mainly because increasing computational power allowed to tackle large and complex real life problems numerically. For offering reliable results, a thorough theoretical analysis of solution algorithms, their convergence properties, and approximation quality is inevitable. The investigation of optimal control problems started in the 1970's (see [101]) and is still the center of attention for many books such as [81, 98, 99, 144].

This thesis is concerned with numerical solution techniques for linear quadratic optimal control problems with elliptic partial differential equations (PDEs). There are two major issues that need to be handled:

- the discretization of the PDE, building on results from regularity and approximation theory,
- the optimization process, building on functional analysis and (numerical) optimization.

Depending on the order of these concepts, the terminology *first-discretize-then-optimize* or *first-optimize-then-discretize* is common.

We are going to follow the *first-optimize-then-discretize* approach and stay in infinite–dimensional function spaces for the construction of optimization algorithms. The semi-smooth Newton-method and the interior point method will be of major interest. Only after the construction of solution algorithms, the problem is discretized with the finite element method (FEM).

1.1 The Finite Element Method

The term Finite Element Method (FEM) appeared first in 1960, but the ideas behind it are even older and can be traced back to the engineering sciences. Being a powerful numerical tool for a variety of engineering disciplines, the FEM quickly found its way into applied mathematics where stability, discretization errors, and convergence rates are thoroughly investigated (see [85]). It has been a very active field of research ever since.

The standard way to gain accuracy of the approximate solution is to refine the triangulation of the computational domain, which introduces more degrees of freedom. A vast amount of publications deals with this so called h-version. We merely mention [34, 35, 43].

An alternative way of enlarging the number of unknowns is to increase the polynomial degree of finite elements. This, so called, p-version is less common and its convergence speed strongly depends on the regularity of the solution to the PDE. See [17, 26, 93, 135].

Let us explain this briefly using an example from [48]. Consider the smooth function e^x and the non-differentiable function

$$j(x) = \begin{cases} -1 & \text{if } x < -1/2, \\ 1 & \text{if } x > 1/2, \\ 2x & \text{if } x \in [-1/2, 1/2]. \end{cases}$$

We project the functions onto two different discrete approximation spaces:

- polynomials on $I := [-1, 1]$ of degree p collected in $\mathbb{R}[x]_p$,
- piecewise linear function on an equidistant partition of I.

Figure 1.1 shows the maximal error between $e^x, j(x)$ and the projections $e_p(x), j_p(x)$ onto $\mathbb{R}[x]_p$ with respect to increasing values for p.

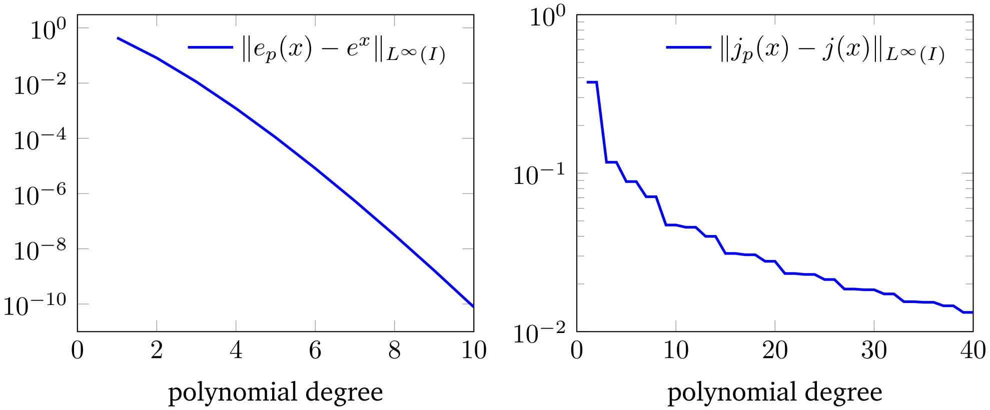


Figure 1.1. The approximation error of polynomials of increasing degree.

We see that the convergence speed for $e_p(x)$ and the smooth function e^x is much faster than for $j_p(x)$ and the non-differentiable function $j(x)$. In fact, the p-version converges exponentially, whereas the piecewise linear approximation of e^x shows an algebraic error decay (see Figure 1.2). If we project $j(x)$ onto the space of piecewise linear functions, the approximation j_h is exact ($j_h(x) = j(x)$) after two repeated bisections of $[-1, 1]$.

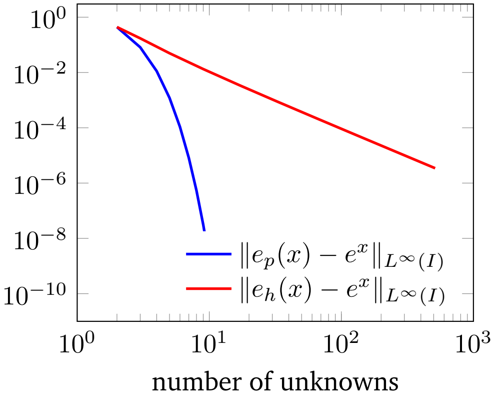


Figure 1.2. The approximation error for high-degree polynomials and piecewise linear functions.

This example heuristically shows that h-refinement should be favored for non-smooth functions, whereas p-refinement is efficient for smooth functions.

The core idea of the hp-finite element method (hp-FEM) is to combine both refinement techniques as follows: Approximate functions by

- polynomials of high degree on large elements in regions of high regularity,
- polynomials of low degree on small elements in regions of low regularity.

This strategy has been thoroughly investigated in the 1980's and 1990's in [9, 10, 11, 12, 13, 14] and [52, 119, 123]. The monographs [51, 53, 87, 135, 152] and [106] give an overall and self-contained access to the topic.

We now have a feeling for the two main ingredients that are necessary for establishing error estimates: the regularity of functions and the 'character' of the approximation space. Only if both concepts 'match each other', it is possible to obtain optimal convergence rates. In this thesis, we will encounter countably normed spaces with weight functions that are specifically designed for the approximation space.

1.2 Application to Control Constrained Optimization Problems

The main goal of this thesis is the application of hp-FEM to the optimal control problem (**P**), which is rigorously introduced in Section 2.2.

$$(\mathbf{P}) \qquad \begin{cases} \text{minimize } J(u,y) := \frac{1}{2}\|\, y - y_d \,\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\, u \,\|^2_{L^2(U)} \\ \qquad \text{subject to} \\ \qquad\qquad Ay = Bu \quad \text{on } \Omega, \\ \qquad\qquad u \in U_{ad}. \end{cases}$$

This formulation covers the case of distributed, Neumann, interface, and other control problems all at once. We follow the classical notation of [144] and denote the control by u and the state by y. The adjoint state, which is introduced for a compact formulation of the first order necessary conditions (see Theorem 2.3.4), is denoted by q. In presence of control constraints and $\nu > 0$, the optimal control can be expressed as (see Theorem 2.3.5)

$$u^* = P_{U_{ad}}\left(-\frac{1}{\nu}B^* q^*\right), \qquad (\star)$$

where $P_{U_{ad}}$ is the projection onto the feasible set U_{ad}.

This fact poses a challenge for numerical solution algorithms using the hp-FEM: the non-smooth projection formula restricts the global regularity of optimal controls in general. It may occur that the optimal control u^* has discontinuous derivatives at the interface of active and inactive sets. What weighs even more, the location of these 'kinks' is not known a-priori. Further bounds on the regularity of the state and adjoint variable are set by the domain Ω, which is assumed to be polygonal. Usually, the solution of a PDE displays singular behavior at the vertices of the boundary $\partial\Omega$.

In order to profit from the hp-idea, we carefully investigate the smoothness of the optimal variables and then examine how well functions are approximated by the underlying finite-dimensional FE space.

There are very few works that use the hp-FEM for solving optimal control problems with box constraints on the control (see [27, 28, 157, 158]). The reasons for this are twofold. First, the general implementation of higher order methods is more involved due to the need of

- high-order quadrature formulas for discretizing the weak formulation of the PDE constraint,
- handling the degrees of freedom through connectivity arrays and orientation information,
- sophisticated methods that guide between h- and p-refinement.

Second, the projection formula ($\star$) is challenging if it is applied to functions of higher polynomial degree. On a linear finite element, approximations q_h^* of the adjoint q^* are typically represented by the values at the nodes of the discretized domain. A projection is necessary if and only if at least one of these values is out of bounds. The values that represent q_h^* on higher order finite elements, however, may all be admissible but still $q_h^* \notin U_{ad}$. This is due to 'oscillatory' behavior of non-linear polynomials. Identifying the points that need to be projected or the corresponding sets is challenging but inevitable for the implementation of practical solvers.

1.3 Solution Techniques and their Main Ideas

The contribution of this thesis is the development of solution algorithms for (**P**) and an elliptic state equation using the hp-FEM. The latter has hardly been used for optimal control problems. Rather, it is common to employ solution techniques like (projected) gradient methods, (semi-smooth) Newton methods, SQP or interior point methods with the h-FEM.

1.3.1 The Semi-Smooth Newton Method and A-Priori Discretizations

The first approach to find optimal solutions is the solution of the optimality condition ($\star$) with a semi-smooth Newton method. An a-priori error analysis regarding the accuracy of piecewise constant functions with respect to the mesh size h was started by the early works [64, 67]. Since then, different ways of discretization have been thoroughly investigated for problems with Neumann boundary control (see, e.g., [4, 38, 39, 80, 104, 125]) and distributed controls (see, e.g., [7, 40, 110, 126, 127]). Some of the results even hold for semi-linear state equations and non-quadratic cost functionals. A more detailed survey on the exact order of accuracy is given at the beginning of Chapter 5.

In the same spirit of these publications, but equipped with the hp-FEM, we are going to investigate the accuracy of approximation and derive a-priori error estimates for Neumann boundary control problems. We carefully examine the regularity of the optimal variables of (**P**) because this critically influences the speed of convergence, which we have illustrated by the above example.

We investigate different FE methods on hp-meshes that are heavily h-refined towards regions of non-smoothness:

- the boundary concentrated (bc) FEM (see [57, 88]),
- the interface concentrated (ic) FEM,
- the vertex concentrated (vc) FEM (in the spirit of [9, 12]).

The bc-FEM is designed to solve problems with smooth data everywhere except for the boundary of the domain. While larger elements of higher polynomial degree approximate the solution in the interior of the domain, lower order elements of smaller size are employed with decreasing distance to the boundary. The literature describes such meshes as geometrically refined towards the boundary. The method can be viewed as a generalization of the boundary element method (BEM) (see [76, 129] and the references therein), but it is applicable to a wider range of elliptic problems.

The ic-FEM is a further generalization insofar as it is designed for $2d$-networks with piecewise analytic data. It consists of applying the bc-idea on each subdomain, which is the reason why most results carry over. In this thesis, we investigate the application to transmission problems, which are characterized by a low global regularity of the solutions.

The vc-FEM carries the idea of geometric mesh refinement to an extreme and uses lower order elements only near the vertices of the domain and near points where the data is irregular. This procedure is well known from PDEs, where a correct discretization yields exponential error decay (see [9, 12, 106, 135]) similar to the 1-dimensional example in Figure 1.2. We are going to extend this result for the application to optimal control problems with control constraints (see Section 3.3 and 4.4), which is one of the key findings of this work. In order to judge the quality of the hp-algorithms, we relate the error bounds to the number of unknowns because the concept of a mesh size h does not need to be available for higher order methods.

1.3.2 The Interior Point Method and A-Posteriori Discretizations

The second approach to find optimal solutions of (**P**) is via an adaptive interior point algorithm. Adaptive mesh refinement relies on a-posteriori estimates which aim at reducing the error in a certain norm (see, e.g., [2, 16, 62, 149] and the references therein) or reducing a quantity of interest (see the survey [20]). This idea has been transferred to control problems with box constraints (see the survey [124] and [19, 66, 78, 83, 90, 91, 100, 102, 103, 151]) as well as integral constraints ([41, 70]). In the latter case, the projection formula ($\star$) only consists of scaling the adjoint, which makes the application of hp-FEM simpler (see [42]).

The advantage of using interior point methods is the fact that the first order necessary conditions become non-linear but smooth. They possess a smoothing property that can be used for guaranteeing feasibility during the iterations of the algorithm. Theoretical and algorithmic properties have been thoroughly investigated not only for the control constrained case (e.g., [122, 130, 134, 148, 160] but also for state constraints (e.g., [131, 132, 133]). Here, we refer the to reader the overview at the beginning of Chapter 6.

The known results are exploited to construct a fully adaptive interior point method. In order to benefit from higher order methods, the smoothness of the control variable is estimated a-posteriori. To the best of our knowledge, this is the first adaptive method that solves (**P**) with the hp-FEM.

1.4 Outline of the Thesis

Chapter 2 begins with functional analytic preliminaries and several remarks on the physical domain Ω. Afterwards, we introduce a collection of function spaces which allow the study of solutions to elliptic PDEs. Then we rigorously introduce the model problem and establish existence, uniqueness, and first order optimality conditions.

Chapter 3 is dedicated to the study of regularity of optimal solutions in different function spaces. We start by classical Sobolev-Slobodeckij spaces and then treat countably normed spaces that contain weight functions to control the blow-up towards different regions of Ω, such as the boundary, interface, or vertices.

Chapter 4 presents the hp-FEM. We provide general concepts, implementational details, and approximation results. For this purpose, an interpolation operator is constructed, which maps smooth functions into a special polynomial space. Depending on the type of discretization, we show algebraic or exponential convergence with respect to mesh size or number of unknowns, respectively.

In Chapter 5, we formulate the discrete numerical algorithm for solving the model problem. Afterwards, several convergence results are established. We perform numerical experiments for the bc-,vc-, and ic-FEM in order to support our theoretical findings. Additionally, we compare the algorithms to each other and discuss their advantages. The chapter also contains an excursion to problems with bang-bang character.

Chapter 6 offers a very different algorithmic approach: the interior point method as an adaptive path-following method. We collect results on the underlying barrier problem and prove convergence of an interior point method with smoothing step. After that, we give a detailed description of the implementation, which adaptively steers hp-mesh refinement and the update of the barrier parameter. We also derive a-posteriori error estimates for this purpose. The chapter is closed by numerical tests regarding hp-adaptivity and path-following properties.

Chapter 7 sums up the results of this thesis: the hp-FEM as an efficient, accurate and flexible discretization method for optimal control problems. For future research, we list possible extensions and further applications of higher order methods in the context of control theory.

Parts of the thesis are published, see [27, 156, 157, 158].

CHAPTER 2

Optimal Control of Partial Differential Equations

Optimal control theory is a branch of mathematical optimization that is concerned with influencing a physical system in an optimal way. The system is usually described by differential equations, which may be coupled and subjected to additional constraints. While the theoretical investigation of real life applications can be very challenging, increasing computational power and storage of modern workspaces allow to solve more and more problems numerically. Applications comprise problems from aerodynamics, fluid mechanics, biology, engineering, mechanics, and many more disciplines.

In this thesis, we are interested in controlling an elliptic partial differential equation (PDE), where the state variable y can be influenced by a control u. An optimal control shall minimize a tracking type cost functional J, which favors a desired state y_d. We will introduce a model problem which allows us to examine basic features both from the theoretical and numerical point of view, while being general enough for applications.

In Section 2.1 we collect functional analytic preliminaries before we introduce the domain Ω and various functions spaces. Sobolev, Sobolev-Slobodeckij, and countably normed spaces are introduced because they are necessary for approximation theory. Section 2.2 is concerned with the rigorous formulation of the model problem for smooth Neumann and transmission problems. Finally, existence, uniqueness, and optimality conditions of optimal solutions are formulated.

2.1 Preliminaries

We start by some theory from functional analysis and collect several properties of the domain Ω afterwards. As a variety of function spaces is needed for regularity investigations, we provide their definitions and main characteristics as well as density results and embedding theorems.

More details on functional analysis can be found in the textbook [163]. The theory of Sobolev spaces is covered in [1] or textbooks on PDEs such as [63, 72, 114]. We also mention [95, 96] regarding weighted function spaces.

2.1.1 Functional Analysis

We restrict ourselves to real Banach spaces V, W with norms $\|\cdot\|_V, \|\cdot\|_W$, respectively. An operator $A : V \to W$ is called *bounded* if there exists a constant $c \geq 0$ such that $\| Av \|_W \leq c\| v \|_V$ for all $v \in V$. The set of all linear and bounded operators is denoted by $\mathcal{L}(V, W)$ and is a Banach space under the norm

$$\| A \|_{\mathcal{L}(V,W)} := \sup_{\| v \|_V = 1} \| Av \|_W .$$

It is well known that for linear operators the concepts of boundedness and continuity are equivalent.

The *dual space* of V is defined by $V^* := \mathcal{L}(V, \mathbb{R})$ and V is called reflexive if and only if $(V^*)^* = V$. We will make use of the duality pairing

$$\langle v', v \rangle_{V^*,V}, \quad v' \in V^*, v \in V.$$

We say a sequence $\{v_k\}_{k\in\mathbb{N}} \subset V$ *converges weakly* to $v \in V$ $(v_k \rightharpoonup v)$ if and only if

$$\lim_{k\to\infty} \langle v', v_k - v \rangle_{V^*,V} = 0 \quad \forall v' \in V^*.$$

If the norm $\|\cdot\|_H$ of a Banach space H is induced by an inner product, we speak of a *Hilbert space* and denote its inner product by $(\cdot, \cdot)_H$.

The following two results are classic and proved in, e.g., [163, Theorem III.6, Theorem III.7].

Theorem 2.1.1 (Riesz Representation). *Let H be a real Hilbert space. For all $l' \in H^*$, there exists an $l \in H$ such that $(l, v)_H = \langle l', v \rangle_{H^*,H}$ and $\| l' \|_{H^*} = \| l \|_H$.*

Theorem 2.1.2 (Lax-Milgram). *Let H be a Hilbert space and $a : H \times H \to \mathbb{R}$ a bilinear mapping which is coercive, i.e.,*

$$|a(x, x)| \geq c_0 \| x \|^2 \quad \forall x \in H,$$

and bounded, i.e.,

$$|a(x, y)| \leq c_1 \| x \| \| y \| \quad \forall x, y \in H.$$

Then there exists a uniquely determined bounded linear operator $S : H \to H$ with a bounded linear inverse S^{-1} such that

$$a(x, Sy) = (x, y)_H \quad \forall x, y \in H.$$

Additionally, $\| S \|_{\mathcal{L}(H,H)} \leq c_0^{-1}$ and $\| S^{-1} \|_{\mathcal{L}(H,H)} \leq c_1$.

For $w' \in W^*$ and an operator $A : V \to W$, the adjoint operator $A^* : W^* \to V^*$ is defined by

$$(A^* w')v := w'(Av).$$

With the duality pairing, we can write $\langle A^* w', v\rangle_{V^*,V} = \langle w', Av\rangle_{W^*,W}$. It is well known, that the mapping of an operator to its adjoint is linear and isometric. Since $(A^*)^{-1} = (A^{-1})^*$, we use the abbreviated notation A^{-*} for the inverse of A^*. For the case of $V = W$ being a Hilbert space, we call A *self-adjoint* if and only if $(A^*)^* = A$.

An operator $A : V \to W$ is called *Gâteaux-differentiable* at v in direction $h \in V$, if and only if there exists a map $A' \in \mathcal{L}(V, W)$ such that

$$\lim_{t \searrow 0} \frac{A(v + th) - A(v)}{t} = A'h + o(\| h \|_V).$$

We say that V is *continuously embedded* in W ($V \hookrightarrow W$) if and only if there is a continuous, injective and linear mapping $i : V \to W$. Since continuity of linear mappings is equivalent to boundedness, it follows that $\| i(v) \|_W \leq c(i) \| v \|_V$, where $c(i)$ is a positive constant depending on the map i. This definition aims at continuous embeddings of function spaces.

2.1.2 The Domain

Generally, a domain Ω is an open and connected subset of $\mathbb{R}^n$, $n \in \mathbb{N}$. We will only investigate the case of bounded domains in $\mathbb{R}^2$ whose boundary $\partial\Omega =: \Gamma$ is a polygon. The boundary is split into a Dirichlet and Neumann part, denoted by $\Gamma_{\mathcal{D}}$ and $\Gamma_{\mathcal{N}}$, respectively. It may be partitioned into l straight line segments Γ_i, $i = 1, \dots, l$ that are collected in $\mathcal{E}$. Each line segment either belongs to the Dirichlet boundary ($\Gamma_i \in \mathcal{E}_{\mathcal{D}}$) or Neumann boundary ($\Gamma_i \in \mathcal{E}_{\mathcal{N}}$). We denote the end-points of these segments by $X_1, \dots, X_l$, and set $\mathcal{X} := \{X_1, \dots, X_l\}$. Furthermore, it holds $\overline{\Gamma}_i \cap \overline{\Gamma}_{i+1} = X_i$ with the convention $\Gamma_{l+1} := \Gamma_1$. A point X_i is called *vertex*.

We say the boundary of a domain is $C^{0,1}$ or Lipschitz if for each $x \in \partial\Omega$ there is a neighborhood V of x and new orthogonal coordinates (y_1, y_2) such that

1. $V = \{(y_1, y_2) \in \mathbb{R}^2 \ : \ |y_i| < a_i \in \mathbb{R}^+, \ i = 1, 2\}$,
2. there exists a Lipschitz continuous function $\Psi : (-a_1, a_1) \to (-a_2, a_2)$ such that
   - $\Omega \cap V = \{y \in V \ : \ y_2 < \Psi(y_1)\}$,
   - $\Gamma \cap V = \{y \in V \ : \ y_2 = \Psi(y_1)\}$.

Obviously, polygonal domains have Lipschitz-boundary, which in turn implies the segment- or cone-property (see [1]).

We speak of a $2d$-*network* $\{\Omega_i\}_{i\in I}$ with $I = \{1,\dots,n_I\}$ if Ω can be partitioned into $n_I \in \mathbb{N}$ pairwise disjoint, polygonal subdomains Ω_i such that $\overline{\Omega} = \cup_{i\in I}\overline{\Omega}_i$. We further stipulate a compatibility condition among the subsets, i.e., exactly one of the following holds for $i, j \in I,\ i \neq j$:

- $\overline{\Omega}_i \cap \overline{\Omega}_j = \emptyset$,
- $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a common vertex,
- $\overline{\Omega}_i \cap \overline{\Omega}_j$ is a common side, denoted by $\overline{\gamma}_{i,j}$. These parts of the boundary are referred to as *interface* and are collected in $\mathcal{I}$.

In the context of $2d$-networks, the set of vertices $\mathcal{X}$ is enlarged by the set of interior vertices. The restriction of a function $u : \Omega \to \overline{\mathbb{R}}$ to one subdomain Ω_i is denoted by $u_i : \Omega_i \to \overline{\mathbb{R}}$.

2.1.3 Function Spaces

For $1 \leq p < \infty$, we denote the space of functions whose p-th power is integrable by $L^p(\Omega)$. The underlying measure is the Lebesgue-measure, which will sometimes be denoted by meas_i for the dimension $i = 1, 2$. Essentially bounded functions are collected in $L^\infty(\Omega)$.

A property of $v \in L^p(\Omega)$ holds almost everywhere (a.e.) in Ω, if it is violated only on a set of Lebesgue-measure zero. In fact, functions are only defined up to a set of measure zero, which is why L^p-spaces contain equivalence classes of functions.

Endowed with the norm

$$\begin{aligned} \| v \|_{L^p(\Omega)} &:= \left(\int_\Omega |v(x)|^p \,\mathrm{d}x \right)^{1/p} && \text{for } 1 \leq p < \infty, \\ \| v \|_{L^p(\Omega)} &:= \operatorname{ess\,sup}_\Omega |v(x)| && \text{for } p = \infty, \end{aligned}$$

$L^p(\Omega)$ becomes a Banach space. In the sequel, we sometimes omit Ω in the notation of Lebesgue or other function spaces if the context forbids confusion or the domain is of no relevance. For the special case $p = 2$ and the inner product

$$(v, w)_{L^2(\Omega)} := \int_\Omega v(x)w(x) \,\mathrm{d}x,$$

$L^2(\Omega)$ is a Hilbert space. Owing to the boundedness of the domain Ω, we have the embedding $L^p(\Omega) \hookrightarrow L^q(\Omega)$ with $p > q$. This is proved with Hölder's inequality

$$\int_\Omega vw \,\mathrm{d}x \leq \left(\int_\Omega |v|^p \,\mathrm{d}x \right)^{1/p} \left(\int_\Omega |w|^q \,\mathrm{d}x \right)^{1/q},$$

where $1/p + 1/q = 1$ (and the obvious extension for $p = \infty$). The case $p = q = 2$ reproduces the Cauchy-Schwarz inequality.

It is well known that for $1 \leq p < \infty$ the dual space of $L^p(\Omega)$ can be identified with $L^q(\Omega)$, where q is the dual exponent $q = p/(p-1)$. The well-definedness of the dual pairing

$$\langle v, w \rangle_{L^p(\Omega), L^q(\Omega)} = \int_\Omega vw \, \mathrm{d}x$$

follows, again, from Hölder's inequality.

Sobolev spaces have turned out to be suitable for developing a notion of weak solutions of PDEs. For k being a positive integer, we denote by $W^{k,p}(\Omega)$ the space of functions which are k-times weakly differentiable and whose derivatives lie in $L^p(\Omega)$. For a multi-index $\alpha \in \mathbb{N}_0^n$ we denote (weak) derivatives by

$$D^\alpha v = \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n} v.$$

The Sobolev space is complete under the norm

$$\| v \|_{W^{k,p}(\Omega)}^p := \sum_{0 \leq |\alpha| \leq k} \| D^\alpha v \|_{L^p(\Omega)}^p.$$

We agree on the notation $W^{0,p}(\Omega) = L^p(\Omega)$. The highest derivative is measured by the semi-norm

$$|v|_{W^{k,p}(\Omega)}^p := \sum_{|\alpha| = k} \| D^\alpha v \|_{L^p(\Omega)}^p.$$

The Sobolev-Slobodeckij space $W^{s,p}(\Omega)$ with a real value $s = k + \sigma > 0,\ k \in \mathbb{N},\ \sigma \in (0,1)$ extends the above definitions. It comprises functions from $W^{k,p}(\Omega)$ with the finite norm

$$\| v \|_{W^{s,p}(\Omega)}^p := \| v \|_{W^{k,p}(\Omega)}^p + \int_\Omega \int_\Omega \sum_{|\alpha| = k} \frac{|D^\alpha v(x) - D^\alpha v(y)|^p}{|x - y|^{2+\sigma p}} \, \mathrm{d}x \, \mathrm{d}y. \tag{2.1}$$

The expression $|x - y| = \operatorname{dist}(x, y)$ denotes the Euclidean distance between two points. An alternative way to define spaces with 'fractional' derivatives, is to proceed by real interpolation. For more details see [1, Chapter 7], [35, Chapter 14], and [22]. We now use $s \in \mathbb{R}_0^+$ as exponent and agree on the notation $H^s(\Omega) := W^{s,2}(\Omega)$ because we are dealing with a Hilbert space. The well known *trace theorem* states the existence of a linear and bounded trace operator

$$T_{\Gamma_\mathcal{D}} : W^{s,p}(\Omega) \to W^{s-1/p,p}(\Gamma_\mathcal{D}), \quad v \mapsto T_{\Gamma_\mathcal{D}} v = v|_{\Gamma_\mathcal{D}}$$

if $s \leq 1$ and $s - 1/p > 0$ is not integer, see [1, 72]. This is a generalization of the classical restriction of continuous functions to the boundary. The trace space $W^{s-1/p}(\Gamma_\mathcal{D})$ can be characterized as in (2.1) with the exponent $1 + \sigma p$ in the denominator and $\Gamma_\mathcal{D}$ as domain of integration, see [117, Subsection 1.2.3].

The trace operator allows us to define functions which are zero at (parts of) the boundary, i.e.,

$$W_{\Gamma_\mathcal{D}}^{s,p}(\Omega) := \{ v \in W^{s,p}(\Omega) \ : \ T_{\Gamma_\mathcal{D}} v = 0 \}.$$

The classical result that $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ is dense in $W^{k,p}(\Omega)$ holds for general open sets Ω and $1 \leq p < \infty$ (see [1, Theorem 3.17]). If Ω satisfies the segment-property (which is weaker than $\Gamma \in C^{0,1}$), the set of restrictions to Ω of functions $C_c^\infty(\mathbb{R}^n)$ is dense as well (see [1, Theorem 3.22]). This result extends to $W^{s,p}(\Omega)$ with $s > 0$ if the domain Ω allows for a continuous extension operator from $W^{s,p}(\Omega)$ to $W^{s,p}(\mathbb{R}^n)$ (see [115, Theorem 2.4]).

We also mention the standard Sobolev embeddings.

Theorem 2.1.3. *Let $\Omega \subset \mathbb{R}^n$ be bounded with Lipschitz boundary. Then the following continuous embeddings hold true for $k, k' \in \mathbb{N}$, $1 \leq p, p' < \infty$ and $\beta \in (0,1)$:*

$$\begin{aligned} W^{k,p}(\Omega) &\hookrightarrow W^{k',p'}(\Omega), & k \geq k', k - \frac{n}{p} &\geq k' - \frac{n}{p'}, \\ W^{k,p}(\Omega) &\hookrightarrow C^{m,\beta}(\overline{\Omega}), & k > 0,\ k - \frac{n}{p} &\geq m + \beta. \end{aligned}$$

The result is proved in, e.g., [1, Theorem 4.12]. For the special case $s \in [0,1]$, we can combine Theorem 5.4 and Theorem 6.5 of [115] to obtain

$$W^{s,p} \hookrightarrow L^q(\Omega) \quad \begin{cases} q \in [1, \frac{np}{n-sp}] & \text{if } sp < n, \\ q \in [1, \infty) & \text{if } sp \geq n. \end{cases} \tag{2.2}$$

Proposition 2.1.4. *Let $\sigma \in (0,1)$ and $n = 2$, then*

$$H^{1+\sigma}(\Omega) \hookrightarrow C^{0,\sigma}(\overline{\Omega}). \tag{2.3}$$

Proof. First, we see that Theorem 2.1.3 yields $H^{1+\sigma}(\Omega) \hookrightarrow H^1(\Omega) \hookrightarrow L^q(\Omega)$ for $1 \leq q < \infty$. For $v \in H^{1+\sigma}(\Omega)$, we find $\nabla v \in H^\sigma(\Omega) \hookrightarrow L^{q'}(\Omega)$ with $q' \in [1, \frac{2}{1-\sigma}]$ by (2.2). Consequently,

$$H^{1+\sigma}(\Omega) \hookrightarrow W^{1,\min\{q,q'\}}(\Omega) = W^{1,q'}(\Omega).$$

Applying Theorem 2.1.3 to $W^{1,2/(1-\sigma)}(\Omega)$ proves the result. □

It is commonly known that irregular domains promote singularities in the solution of elliptic PDEs (see Section 3.1). For instance, a blow-up in the first derivative can occur in reentrant corners. Weighted Sobolev spaces allow a more rigorous description of such phenomena and are a generalization of the previous spaces because weak derivatives no longer have to be integrable.

Definition 2.1.5. *A function $r : \Omega \to \overline{\mathbb{R}}$ is called **weight function** if and only if r is measurable and attains only positive values almost everywhere.*

Obviously,

$$r_M(x) := \inf_{X \in M} \min\{1, \operatorname{dist}(x, X)\}, \quad M \subset \overline{\Omega} \tag{2.4}$$

satisfies the properties of Definition 2.1.5.

Let us mention two important examples.

- The boundary weight function

  $$r_\Gamma(x) := \inf_{X\in\Gamma} \min\{1, \operatorname{dist}(x, X)\}, \quad x \in \overline{\Omega} \tag{2.5}$$

  for the boundary $\Gamma = \partial\Omega$.

- The vertex weight function

  $$r_\mathcal{V}(x) := \prod_{X\in\mathcal{V}} \min\{1, \operatorname{dist}(x, X)\}, \quad x \in \overline{\Omega} \tag{2.6}$$

  for a finite set of points (usually vertices) $\mathcal{X} \subset \mathcal{V} = \{X_1, \dots, X_m\} \subset \Gamma$ with $m \geq l$. Note that this function is defined slightly different from (2.4) because it contains a product over the distances to the vertices. Since Ω is assumed to be polygonal, we can find a constant $c > 0$ such that

  $$c^{-1} r_\mathcal{V}(x) \leq \inf_{X\in\mathcal{V}} \min\{1, \operatorname{dist}(x, X)\} \leq c r_\mathcal{V}(x).$$

  The definition in (2.6) has the advantage that a specific weight can be assigned to each vertex. Let $\beta \in \mathbb{R}^m$ be a multi-index satisfying $\beta = (\beta_1, \dots, \beta_m) \in (0, 1)$, where the inclusion is understood component-wise. We set for $x \in \overline{\Omega}$, $p \in \mathbb{Z}$

  $$r_\mathcal{V}(x)^{p+\beta} := \prod_{X\in\mathcal{V}} \min\{1, \operatorname{dist}(x, X)\}^{p+\beta_i}.$$

The weight function r_Γ will play an important role in Sections 3.2 and 4.3, whereas $r_\mathcal{V}$ appears in Sections 3.3 and 4.4. Weight functions give rise to the space $L^p(\Omega, r)$, which is defined as the set of measurable functions with finite norm

$$\| v \|_{L^p(\Omega,r)} = \left(\int_\Omega r(x)|v(x)|^p \,\mathrm{d}x \right)^{1/p}.$$

Additionally, $W^{k,p}(\Omega, r)$ comprises all functions whose weak derivatives up to order $k \in \mathbb{N}$ lie in $L^p(\Omega, r)$. If there exist constants $0 < c_1 \leq r(x) \leq c_2$, we obviously have

$$W^{k,p}(\Omega, r) = W^{k,p}(\Omega).$$

For a bounded weight function r, such as $r_\mathcal{V}, r_\Gamma$, embeddings for function spaces with decreasing powers of r, e.g.,

$$W^{k,p}(\Omega, r^{\varepsilon_1}) \hookrightarrow W^{k,p}(\Omega, r^{\varepsilon_2}) \quad \text{with } 0 \leq \varepsilon_1 \leq \varepsilon_2 < \infty.$$

are straightforward. Interestingly enough, the space $W^{k,p}(\Omega, r)$ is not necessarily a Banach space and density results may fail to hold (see [121]). We mention the Muckenhoupt class A_p of weight functions, satisfying

$$\sup_{B\subset\mathbb{R}^n} \left(\frac{1}{|B|} \int_B r \,\mathrm{d}x \right) \left(\frac{1}{|B|} \int_B r^{-\frac{1}{p-1}} \,\mathrm{d}x \right)^{p-1} < \infty, \quad p > 1.$$

This condition guarantees the completeness as well as the density of smooth functions (see [69]) and is satisfied if $r^{-1/(p-1)} \in L^1(\Omega)$. It is directly shown in [96, Proposition 7.6] that $C^\infty(\overline{\Omega})$ is dense in $W^{k,p}(\Omega, r^\varepsilon)$ for $\varepsilon \geq 0$ and $r \in \{r_\mathcal{V}, r_\Gamma\}$.

A further generalization of the above function spaces is obtained if each derivative D^α is weighted with $r^{\varepsilon(\alpha)}$. Let

$$V^{k,p}_\beta(\Omega) := \{v \in L^p(\Omega) \ : \ r^{\beta+|\alpha|-k} D^\alpha v \in L^p(\Omega), \quad |\alpha| \leq k\}.$$

This space has already been used in the seminal work [92]. Various properties and embeddings for $V^{k,p}_\beta$ can be found in [117, Chapter 1]. For sufficiently large β, i.e., if $\beta > kp - 1$, [95, Proposition 9.6] states that the spaces $V^{k,p}_{\beta/p}(\Omega)$ and $W^{k,p}(\Omega, r^\beta)$ are the same and their norms equivalent. Note that the weight function in $V^{k,p}_\beta$ generally can blow up for small values of $|\alpha|$, which forces functions to vanish at points where r is zero (see [117, Theorem 1.23]).

Furthermore, we denote by $H^{l,l}_\beta(\Omega)$ with $\beta \geq 0$ the completion of $C^\infty(\bar{\Omega})$ with respect to the norm

$$\| v \|^2_{H^{l,l}_\beta(\Omega)} = \| v \|^2_{H^{l-1}(\Omega)} + \| r^{\beta+p-l} \nabla^p v \|^2_{L^2(\Omega)}, \qquad l \geq 1, \tag{2.7a}$$

$$\| v \|^2_{H^{0,0}_\beta(\Omega)} = \| v \|^2_{H^0_\beta(\Omega)} = \| r^{\beta+p} \nabla^p v \|^2_{L^2(\Omega)}, \qquad l = 0, \tag{2.7b}$$

where

$$\| r^{\beta+p} \nabla^p v \|^2_{L^2(\Omega)} := \sum_{\alpha_1=1,\ldots,\alpha_p=1}^{n=2} \| r^{\beta+p} D^\alpha v \|^2_{L^2(\Omega)}.$$

The following result can be found in [15, Equation (2.2)].

Proposition 2.1.6. *For $l \geq 2$ and $\Omega \subset \mathbb{R}^2$, we have the continuous embedding*

$$H^{l,l}_\beta(\Omega) \hookrightarrow C^{l-2}(\overline{\Omega}). \tag{2.8}$$

For controlling the derivatives near the boundary of the domain, we introduce countably normed spaces for constants $C, \gamma > 0$ and $\beta \in [0, 1)$.

$$B^l_\beta(\Omega, C, \gamma) := \{v \in H^{l,l}_\beta(\Omega) \ : \ \| v \|_{H^{l,l}_\beta(\Omega)} \leq C, \ \| r^{\beta+p} \nabla^{p+l} v \|_{L^2(\Omega)} \leq C\gamma^p p!, \ p \in \mathbb{N}_0\}.$$
(2.9)

Functions belonging to countably normed spaces are analytic away from the zeros of the weight function (see [13, Lemma 2.4]).

Theorem 2.1.7. *Let $u \in B^l_\beta(\Omega, C, \gamma)$ for some $l \in \mathbb{N}_0$ and $\beta \in (0, 1)$. Then u is analytic on $\overline{\Omega} \setminus \{x \in \overline{\Omega} \ : \ r(x) = 0\}$.*

Remark 2.1.8. *Countably normed spaces can measure the blow up of solutions to elliptic PDEs, which typically occurs near a vertex $X \in \mathcal{X}$. The singularity for the Laplacian is known to have the form $s_{\lambda_X} = \rho_X^{\lambda_X} \sin(\lambda_X \theta)$ with $\lambda_X \in (0, 1)$ (see (3.10)) in polar coordinates (ρ_X, θ) with origin in X. By changing variables in the definition of B^2_β (see [12, Theorem 1.1]), it can be proved that $s_{\lambda_X} \in B^2_\beta$ with $\beta \in (1 - \lambda_X, 1)$ but $s_{\lambda_X} \notin B^2_{1-\lambda_X}$ (confer [12, Remark 1]).*

The set of traces of $B^2_\beta(\Omega, C, \gamma)$ is defined by

$$B^{3/2}_\beta(\Gamma, C, \gamma) := \{v \in C^0(\Gamma) \ : \ \exists \tilde{v} \in B^2_\beta(\Omega, C, \gamma) \text{ with } \tilde{v}|_\Gamma = v\}$$

in view of (2.8). An intrinsic characterization of the trace space is given in [13].

We stress that countably normed spaces depend on the weight function. As the weight function differs in parts of our exposition, we will specify which r is used in the corresponding chapter.

2.2 The Model Problem

Let us denote by $U \subset \overline{\Omega} \subset \mathbb{R}^2$ the domain where the control acts. We investigate the linear quadratic optimal control problem

$$\textbf{(P)} \qquad \begin{cases} \text{minimize } J(u,y) := \dfrac{1}{2}\| \, y - y_d \, \|^2_{L^2(\Omega)} + \dfrac{\nu}{2}\| \, u \, \|^2_{L^2(U)} \\ \qquad \text{subject to} \\ \qquad\qquad Ay = Bu \quad \text{on } \Omega, \\ \qquad\qquad u \in U_{ad}. \end{cases}$$

The constraint $Ay = Bu$ is called *state equation* and is an elliptic PDE, where A represents an elliptic differential operator in its *weak formulation*. The operator B models the way the control variable u influences the physical system modeled by the equation.

The functional J is said to be of *tracking type* because the minimization process intends to drive the state variable y to a desired state y_d. In order to model the costs for controlling the system, a regularization term $\frac{\nu}{2}\| \, u \, \|^2_{L^2(U)}$ with $\nu \geq 0$ is included.

The optimization is subject to additional control constraints which are described by the admissible set

$$U_{ad} := \{u \in L^2(U) \mid u_a \leq u \leq u_b \quad \text{a.e. in } U\}.$$

Naturally, we demand $u_a \leq u_b$.

The state equation will take different forms in the course of this thesis. For now, it will be convenient to distinguish between two types of problems:

- smooth Neumann problems,
- transmission problems.

This separation makes sense because of the differences regarding global regularity, numerical treatment, and the way the control enters ($U = \Gamma_\mathcal{N}$ or $U = \mathcal{I}$). Here, we generally stipulate $u_a, u_b \in H^{1/2}(U)$ because we want the optimal control to inherit as much smoothness as possible through the projection formula. Additionally, we always assume $y_d \in L^2(\Omega)$.

Only in Chapter 6, we will encounter distributed control problems where $U = \Omega$. In this context, the bounds need to be essentially bounded because we apply path-following techniques.

2.2.1 Smooth Neumann Problems

The following problem is distinguished by the smoothness properties of the differential operator. This will allow special regularity results in countably normed spaces (see Section 3.2). Let $Ay = Bu$ be the Neumann boundary value problem

$$\text{(N)} \qquad \begin{cases} -\nabla \cdot (D(x)\nabla y) + c(x)y = f & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma_{\mathcal{D}}, \\ \partial_{n_D} y = u & \text{on } \Gamma_{\mathcal{N}} \end{cases}$$

with a matrix valued D and scalar c. Here, ∂_{n_D} is the co-normal derivative $n(x)\cdot D(x)\nabla$ at a point $x \in \Gamma_{\mathcal{N}}$. For the Laplacian ($D \equiv I$), ∂_{n_D} becomes the normal derivative ∂_n with the outward unit normal vector $n(x)$. The properties of the involved functions are summed up in the following assumption.

Assumption 2.2.1. *Let $f \in L^2(\Omega)$, $u \in L^2(\Gamma_{\mathcal{N}})$ and let $c(x) \geq 0$ on $\overline{\Omega}$ ($c(x) \geq c_0 > 0$ if $\Gamma_{\mathcal{D}} = \emptyset$). We assume that D is symmetric and positive definite on $\overline{\Omega}$, i.e., there exists $\lambda_{\min} > 0$ such that for all $x \in \overline{\Omega}$ it holds $\xi^T D(x)\xi \geq \lambda_{\min}|\xi|^2$ for all $x, \xi \in \mathbb{R}^2$. Moreover, the functions D, c are assumed to be analytic on $\overline{\Omega}$, i.e., there are constants $C_D, \gamma_D, C_c, \gamma_c > 0$ such that*

$$\| \nabla^p D \|_{L^\infty(\Omega)} \leq C_D \gamma_D^p p!, \tag{2.10}$$

$$\| \nabla^p c \|_{L^\infty(\Omega)} \leq C_c \gamma_c^p p! \qquad \forall p \in \mathbb{N}_0. \tag{2.11}$$

Solutions $y \in H^1_{\Gamma_{\mathcal{D}}}$ are understood in the weak sense. They satisfy the equation

$$\int_\Omega D(x)\nabla y \cdot \nabla v + c(x)yv \,\mathrm{d}x = (f, v)_{L^2(\Omega)} + (u, v)_{L^2(\Gamma_{\mathcal{N}})} \quad \forall v \in H^1_{\Gamma_{\mathcal{D}}}(\Omega). \tag{2.12}$$

Theorem 2.2.2. *Under Assumption 2.2.1, the weak formulation (2.12) of the Neumann problem (N) possesses a unique solution that satisfies*

$$\| y \|_{H^1_{\Gamma_{\mathcal{D}}}(\Omega)} \leq C(\| f \|_{L^2(\Omega)} + \| u \|_{L^2(\Gamma_{\mathcal{N}})})$$

for a positive constant C depending on D, c.

Proof. The bilinear form of the weak formulation (2.12) is bounded and coercive due to Assumption 2.2.1. Furthermore, the right hand side of (2.12) lies in the dual space of $H^1_{\Gamma_{\mathcal{D}}}(\Omega)$. An application of Theorem 2.1.1 and 2.1.2 yields the result. □

It is clear that unique existence of a solution can be established under weaker assumptions. We need the analyticity of c and D for obtaining analytic regularity results and approximation results in finite dimensional spaces. Problem (**P**) subject to (N) is referred to as *Neumann control problem*.

2.2.2 Transmission Problems

Transmission problems are an extension of the previous smooth Neumann problems insofar as they are posed on a $2d$-network, which can model multi-composite materials in real life applications. The differential operator, on the other hand, is less general and consists of the Laplacian weighted by a constant coefficient κ.

The strong formulation of the transmission problem on a $2d$-network $\{\Omega_i\}_{i\in I}$ reads

$$\text{(T)} \qquad \begin{cases} -\kappa_i \Delta y_i = f_i & \text{in} \quad \Omega_i, \\ y_i - y_j = 0 & \text{on} \quad \gamma_{i,j} \in \mathcal{I}, \\ \kappa_i \partial_{n_i} y_i + \kappa_j \partial_{n_j} y_j = u & \text{on} \quad \gamma_{i,j} \in \mathcal{I}, \\ y_i = 0 & \text{on} \quad \Gamma_i \in \mathcal{E}_{\mathcal{D}}, \\ \kappa_i \partial_{n_i} y_i = h_i & \text{on} \quad \Gamma_i \in \mathcal{E}_{\mathcal{N}}. \end{cases}$$

The discontinuity of the coefficient κ across subdomain boundaries causes a jump in the normal derivative (see the third subequation of (T)), which is often written as

$$[\kappa \partial_n y(x)] := \lim_{\varepsilon \searrow 0} (\nabla y(x + \varepsilon n) - \nabla y(x - \varepsilon n)) \cdot n, \quad x \in \mathcal{I}.$$

Note that the expression is independent of the sign of the normal vector n. We also speak of (T) as an interface problem.

Again, we sum up the properties of the data in a general assumption. We define

$$u \in H^s(\mathcal{I}) \quad :\Leftrightarrow \quad u|_{\gamma_{i,j}} \in H^s(\gamma_{i,j}) \quad \forall \gamma_{i,j} \in \mathcal{I}$$

for a compact notation of the regularity of interface functions on $\mathcal{I}$.

Assumption 2.2.3. *Assume that $\{\Omega_i\}_{i\in I}$ is a $2d$-network and that $\Gamma_{\mathcal{D}} \neq \emptyset$. Furthermore, let*

$$h \in H^{1/2}(\Gamma_{\mathcal{N}}),\ u \in H^{1/2}(\mathcal{I}),\ \kappa_i > 0,\ f_i \in L^2(\Omega_i) \quad \textit{for } i \in I.$$

Using the characteristic function χ for sets, we define $\kappa(x) := \sum_{i\in I} \chi_{\Omega_i} \kappa_i$ and write down the weak formulation of (T). Find $y \in H^1_{\Gamma_{\mathcal{D}}}(\Omega)$ such that

$$\int_\Omega \kappa \nabla y \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x + \sum_{\gamma_{i,j}\in\mathcal{I}} (u, v)_{L^2(\gamma_{i,j})} + \sum_{\Gamma_i \in \mathcal{E}_{\mathcal{N}}} (h_i, v)_{L^2(\Gamma_i)} \quad \forall v \in H^1_{\Gamma_{\mathcal{D}}}(\Omega). \tag{2.13}$$

Theorem 2.2.4. *Under Assumption (2.2.3), the transmission problem (T) possesses a unique solution y, which satisfies*

$$\| y \|_{H^1_{\Gamma_{\mathcal{D}}}(\Omega)} \leq C(\| f \|_{L^2(\Omega)} + \| h \|_{H^{1/2}(\Gamma_{\mathcal{N}})} + \| u \|_{H^{1/2}(\mathcal{I})})$$

for a positive constant C depending on $\min_{i\in I} \kappa_i$.

Proof. The bilinear form of the weak formulation (2.13) is bounded and coercive due to Assumption 2.2.3. Furthermore, the right hand side of (2.13) lies in the dual space of $H^1_{\Gamma_{\mathcal{D}}}(\Omega)$. Again, the application of Theorem 2.1.1 and 2.1.2 yields the result. □

Since we want to promote solutions in the space $H^{1+\sigma}(\Omega)$ (see Section 3.1) we stipulated more regularity in the data u, h than necessary for an existence and uniqueness result. We refer to (**P**) subject to (T) as *interface control problem*.

Transmission problems have been thoroughly studied in literature. The homogeneous case $u \equiv 0$ is treated in [55, 75, 120, 161] while inhomogeneous data is discussed for general elliptic operators in [117, 118]. Note that the normal jump generally limits the global regularity of a solution to only $H^{3/2-\varepsilon}(\Omega)$ for small $\varepsilon > 0$.

2.3 Solvability and Optimality Conditions

The following assumption is met by our model problem (**P**) because of Theorem 2.2.2 and 2.2.4 for smooth Neumann and transmission problems, respectively.

Assumption 2.3.1. *We assume that $y_d \in L^2(\Omega)$ and $A \in \mathcal{L}(Y, Z)$, $B \in \mathcal{L}(L^2(U), Z)$ with Banach spaces Y, Z. Moreover,*

- *U_{ad} is nonempty, convex, and closed in $L^2(U)$. In the case $\nu = 0$, it is assumed to be bounded.*
- *the operator A has a bounded inverse $A^{-1} \in \mathcal{L}(Z, Y)$.*

This assumption allows to formulate a general existence result for optimal solutions.

Definition 2.3.2. *A pair $(u^*, y^*) \in U_{ad} \times Y$ is called* ***(optimal) solution*** *of problem* (**P**) *if and only if $Ay^* = Bu^*$ and*

$$J(u^*, y^*) \leq J(u, y) \quad \forall (u, y) \in U_{ad} \times Y \text{ with } Ay = Bu.$$

Theorem 2.3.3. *Under Assumption 2.3.1, problem* (**P**) *possesses an optimal solution. If $\nu > 0$ or $A^{-1}B$ is injective, the solution is unique.*

The proof is standard (e.g., [81, 144]) and only shown for completeness.

Proof. As A is boundedly invertible, we obtain an equivalent problem formulation if we eliminate the state from the optimization by replacing y with $A^{-1}Bu$. The reduced problem reads

$$\text{minimize } \hat{J}(u) := J(u, A^{-1}Bu) \quad \text{subject to} \quad u \in U_{ad}$$

and depends only on u. As $\hat{J} \geq 0$ and $U_{ad} \neq \emptyset$, we can take a minimizing sequence $\{u_k\}_{k\in\mathbb{N}} \subset U_{ad}$ with $J(u_k) \to \hat{J}^* := \inf_{u \in U_{ad}} \hat{J}(u)$. The sequence is bounded because either U_{ad} is bounded or $\hat{J}(u_k) \geq \nu/2\| u_k \|_{L^2(U)}$. Further, the admissible set is weakly sequentially compact due to the Theorem of Mazur [163, Theorem V.1.2] and Eberlein-Shmulyan [163, Chapter A.V.4], which is why we can pass from u_k to a subsequence (also denoted by u_k) that converges weakly to $u^* \in U_{ad}$. As the image of weakly convergent sequences under bounded, linear mappings (here $A^{-1}B$) remain weakly convergent, we find

$$y_k := A^{-1}Bu_k \rightharpoonup A^{-1}Bu^* =: y^*.$$

Thus, the pair (u^*, y^*) satisfies $Ay^* = Bu^*$ and is admissible. Since $\hat{J}$ is continuous and convex, it is also weakly lower semi-continuous ([61, Corollary I.2.2]). Consequently, we have

$$J^* = \liminf_{k\to\infty} \hat{J}(u_k) \geq \hat{J}(u^*) \geq J^*.$$

If $\nu > 0$ or $A^{-1}B$ is injective, the objective functional is strictly convex, which implies uniqueness of the solution pair (u^*, y^*). □

The solution can be characterized by optimality conditions.

Theorem 2.3.4. *Let Assumption 2.3.1 hold. An optimal solution (u^*, y^*) to (**P**) satisfies the* ***first order necessary conditions***

$$Ay^* = Bu^*, \tag{2.14a}$$
$$(B^*q^* + \nu u^*, u - u^*)_{L^2(U)} \geq 0 \quad \forall u \in U_{ad}, \tag{2.14b}$$
$$A^*q^* = y^* - y_d. \tag{2.14c}$$

Proof. Similarly as in the previous proof, we eliminate the state in the objective J and obtain the reduced problem. The functional $\hat{J} : U_{ad} \to \mathbb{R}$ is Gâteaux-differentiable on its whole domain and in every direction. From the optimality $\hat{J}(u^*) \leq \hat{J}(u)\ \forall u \in U_{ad}$ we derive

$$\hat{J}(u^*)'(u-u^*) = (A^{-1}Bu - y_d, A^{-1}B(u-u^*))_{L^2(U)} + \nu(u, u-u^*)_{L^2(U)} \geq 0.$$

Introducing the *adjoint variable* q^* as a solution to (2.14c) yields the first order necessary conditions in form of the *variational inequality*

$$(B^*q^* + \nu u^*, u - u^*)_{L^2(U)} \geq 0 \quad \forall u \in U_{ad}. \tag{2.15}$$

□

In the case of convex optimization, the conditions of (2.14) are also sufficient. There is a convenient reformulation of (2.14b) in form of a projection formula. Again, we refer to [81, 144].

Theorem 2.3.5. *Let $P_{U_{ad}} : L^2(U) \to U_{ad}$ be the pointwise projection*

$$\max\{u_a(x), \min\{u(x), u_b(x)\}\}, \quad x \in U \tag{2.16}$$

with the obvious modification if less than two bounds are present. If $\nu > 0$, then (2.14b) is equivalent to the ***projection formula***

$$u^* = P_{U_{ad}}\left(-\frac{1}{\nu}B^*q^*\right). \tag{2.17}$$

Definition 2.3.6. *The set*

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

*is called **active set**. The complement $\mathfrak{A}^c = U \setminus \mathfrak{A}$ is called inactive set. Accordingly, the points of $\mathfrak{A}$ and $\mathfrak{A}^c$ are called active and inactive, respectively. If $\mathfrak{A}^c$ is a set of Lebesgue-measure zero, u^* is called **bang-bang** control.*

Note that the active set is only defined up to a set of Lebesgue-measure zero unless the optimal control happens to be continuous. We close this section by elucidating how the projection $P_{U_{ad}}$ promotes regularity of the optimal control (see also [5, 97]).

Theorem 2.3.7. *Assume that $v \in W^{\sigma,p}(U)$ along with $u_a, u_b \in W^{\sigma,p}(U)$ and $\sigma \in [0,1]$. Then it holds $u := P_{U_{ad}}(v) \in W^{\sigma,p}(U)$ for $U = \Omega, \Gamma_{\mathcal{N}}, \mathcal{I}$.*

Proof. The case $\sigma = 0$ is trivial. For $\sigma = 1$, it is well known that $\max\{0, v\}$ and also $\min\{0, v\}$ are in $W^{1,p}(U)$. The result follows by rewriting (2.16) as

$$\max\{u_a, \min\{v, u_b\}\} = v + \min\{0, u_b - v\} + \max\{0, u_a - v - \min\{0, u_b - v\}\}.$$

Let $\sigma \in (0,1)$ and $\dim(U) := 2$ for $\mathrm{meas}_2(U) = 2$ and $\dim(U) := 1$, otherwise. We follow [5, Equation (4.10)] and directly compute

$$\begin{aligned}
\| u \|^p_{W^{\sigma,p}(U)} &= \| u \|^p_{L^p(U)} + \int_U \int_U \frac{|u(x) - u(y)|^p}{|x-y|^{\dim(U)+\sigma p}} \, \mathrm{d}x \, \mathrm{d}y \\
&= \| P_{U_{ad}} v \|^p_{L^p(U)} + \int_U \int_U \frac{|P_{U_{ad}} v(x) - P_{U_{ad}} v(y)|^p}{|x-y|^{\dim(U)+\sigma p}} \, \mathrm{d}x \, \mathrm{d}y \\
&\leq \| v \|^p_{L^p(U)} + \int_U \int_U \frac{|v(x) - v(y)|^p}{|x-y|^{\dim(U)+\sigma p}} \, \mathrm{d}x \, \mathrm{d}y = \| v \|^p_{W^{\sigma,p}(U)}.
\end{aligned}$$

The last step is a consequence of the Lipschitz continuity of the projection and can also be seen by distinguishing the nine possible cases of how $v(x)$ and $v(y)$ lie in or around $[u_a(x), u_b(x)]$ and $[u_a(y), u_b(y)]$, respectively. □

The above theorem shows that smoother controls occur if the PDE fosters regularity in the adjoint variable, provided that the control constraints are smooth enough. This matter will be investigated in the next chapter.

Remark 2.3.8. *For the Neumann or interface control problem, the optimal adjoint q^* lies in $H^1(\Omega)$ and $u_a, u_b \in H^{1/2}(U)$ with $U = \Gamma_{\mathcal{N}}$ or $U = \mathcal{I}$, respectively. Therefore, Theorem 2.3.7 implies $u^* \in H^{1/2}(U)$.*

CHAPTER 3

Regularity Results

In this chapter we are going to deduce regularity results for the state and adjoint equation of the optimal control problem (**P**) because the smoothness of solutions is crucial for the derivation of approximation results.

In Section 3.1, we show how the shape of the domain and an abstract eigenvalue problem determine over higher regularity in classical Sobolev-Slobodeckij spaces. Typically, the solution of an elliptic equation can be decomposed into a regular part $y_0 \in H^2(\Omega)$ and singularities limiting the global smoothness to $H^{1+\sigma}(\Omega)$ with $\sigma \in [0,1]$. We mention the expansion for Poisson's equation (Remark 3.1.2), and the transmission problem (Theorem 3.1.7, Theorem 3.1.12). The results relate to the seminal work [92], are collected from [49, 117, 118], and can also be found in [158].

Afterwards, Section 3.2 builds on the classical regularity and investigates the affiliation of solutions to $Ay = Bu$ to countably normed spaces $B^2_{1-\sigma}$ with the weight function r_Γ. This weight is designed to control the singularities near the boundary and allows to bound all orders of derivatives. With the help of [88], we establish regularity of the state and adjoint variable (see Theorem 3.2.1 and Corollary 3.2.2, respectively) for the smooth Neumann problems (N). We carry over the results to transmission problems (T) on $2d$-networks (Corollary 3.2.3).

The last Section 3.3 contains the lengthy and technical proof of regularity in vertex weighted spaces (Theorem 3.3.24). The weight function $r_\mathcal{V}$ only measures the distance to a finite set of points where the solution possesses singular components. It damps the blow up of derivatives and allows to prove the affiliation of the optimal variables to the countably normed space B^2_β of [9, 12, 13] with a multi-index β. A key ingredient is the coupled optimality system of the Neumann control problem, which restricts the result to the optimal variables (u^*, y^*, q^*). The result is submitted for publication ([156]).

3.1 Classical Sobolev-Slobodeckij Spaces

In this section, we aim at the maximal regularity that can be obtained for solutions to elliptic PDEs on polygonal domains or $2d$-networks. This may require additional regularity of the data.

Convention: The control problem (or state equation $Ay = Bu$) is referred to as $\mathbf{H}^{1+\sigma}$**-regular** with $\sigma \in (0,1]$ if and only if A^{-1} and A^{-*} map into $H^{1+\sigma}(\Omega)$ for 'sufficiently smooth data'.

For instance, the Neumann boundary problem (N) can only be $H^{1+\sigma}$-regular with $\sigma \in (1/2, 1]$ if $u \in H^{\sigma-1/2}(\Gamma_{\mathcal{N}})$. An optimal Neumann control u^* satisfies this assumption because of Remark 2.3.8 and the standing assumption $u_a, u_b \in H^{1/2}(\Gamma_{\mathcal{N}})$. The minimal regulariy $u \in L^2(\Gamma_{\mathcal{N}})$ is enough to facilitate $\sigma \in (0, 1/2]$ (confer [86]).

It turns out that the properties of Ω and the differential operator A decide over that value of σ. We will derive expansion results that split the solution into a regular part and singular contributions which decide over the global regularity.

Since the procedure is complicated and tedious, we start by sketching the basic ideas for a simpler case, mainly following the exposition in [96].

3.1.1 Expansions in Regular and Singular Components

Poisson's Equation

We look for weak solutions to the Dirichlet problem

$$-\Delta y = f \quad \text{in } \Omega, \quad y = 0 \quad \text{on } \Gamma. \tag{3.1}$$

It is well known that the Laplacian Δ in Cartesian coordinates transforms to

$$\frac{1}{\rho}\partial_\rho\left(\rho\partial_\rho\right) + \frac{1}{r^2}\partial_\theta^2 \tag{3.2}$$

in polar coordinates with $(x_1, x_2) = (\rho\cos\theta, \rho\sin\theta)$ and $\rho \geq 0,\ \theta \in [0, 2\pi)$. Analogous formulas are available for the n-dimensional case which allows for treating higher dimensional problems.

If we formally set $\rho\partial_\rho =: \lambda$, Poisson's equation becomes

$$(\lambda^2 + \partial_\theta^2)y = \rho^2 f.$$

Neglecting boundary and interface conditions, this equation is uniquely solvable if and only if

$$\lambda^2 v + \partial_\theta^2 v = 0 \tag{3.3}$$

has only trivial solutions $v \equiv 0$. This illustrates the important role of the non-linear eigenvalue problem (3.3).

Let us give a short but more rigorous outline of the derivation of the (Sturm-Liouville-) eigenvalue problem and the resulting expansion of the solution of (3.1). As regularity is a local concept, we only need to worry about the smoothness of y at the vertices of Ω. For interior balls

$$B_d(x_0) \quad \text{with} \quad x_0 \in \Omega, \quad d < \operatorname{dist}(x_0, \Gamma),$$

we know that $y \in H^2(B_d(x_0))$. The same is true for smooth parts of the boundary, which can be locally flattened and become half balls $B^+(x_0) := \{x \in \mathbb{R}^2 \mid x_2 = 0\}$. After the change of coordinates $y \in H^2(B_d^+(x_0))$ for d small enough and $x_0 \in \Gamma, x_0 \notin \mathcal{X}$.

For treating the irregularities in a vertex $X \in \mathcal{X}$ of the domain, we localize the problem using a smooth cut-off function $\eta_X = \eta_X(\operatorname{dist}(x, X)) \in C^\infty(\mathbb{R}^2)$. We stipulate $0 \leq \eta \leq 1$ such that $\eta_X \equiv 1$ for all x near X and η_X decreases rapidly to 0 so that all other vertices are not 'visible'. Locally, a solution of (3.1) satisfies

$$-\Delta(\eta_X y) = g := -(\Delta\eta_X)y + \eta_X f + 2\nabla\eta_X \cdot \nabla y, \tag{3.4}$$

where we can change the coordinates such that the domain becomes a cone with origin in X and opening angle $\omega_X \in (0, 2\pi)$, i.e.,

$$C_X := \{(\rho, \theta) \in \mathbb{R}^2 \mid \rho \in \mathbb{R}^+, \theta \in (0, \omega_X)\}.$$

Applying this change to the differential operator yields the equivalent problem for the new variable $y := \eta_X y$ (we dispense with writing y_X everywhere)

$$\begin{aligned} -\left(\partial_\rho^2 y + \frac{1}{\rho}\partial_\rho y + \frac{1}{\rho^2}\partial_\theta^2 y\right) &= g \quad \text{in } C_X, \\ y(\rho, 0) = y(\rho, \omega_X) &= 0. \end{aligned} \tag{3.5}$$

The operation which manages to send $\rho\partial_\rho$ to $\lambda \in \mathbb{C}$ is the Mellin-transform. It reads for fixed $\theta \in (0, \omega_X)$

$$M[y(.,\theta)](\lambda) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \rho^{-\lambda-1} y(\rho, \theta)\, \mathrm{d}\rho =: Y(\lambda, \theta).$$

It is closely connected to the Fourier-Transform $\mathcal{F}[y(.,\theta)](z \in \mathbb{C})$ by Euler's change of variables, which substitutes $\rho =: e^\tau$, $\tau \in (-\infty, \infty)$. If we suppress the dependence on θ, which is assumed to be arbitrary but fixed in $[0, 2\pi)$, the use of the definitions of both transforms yields the relation

$$M[y](\lambda) = F[y](z), \quad z = -i\lambda.$$

Several known properties from the Fourier-Transform, therefore, carry over to the Mellin-Transform. We only state a few important ones (see [96]).

Theorem 3.1.1. *Let $\Re$ and $\Im$ denote the real and imaginary part of a complex number. For $\lambda \in \mathbb{C}$ and $h := -\Re(\lambda)$ fixed, the Mellin-Transform M*

- *is an isomorphism*

$$M : \left\{ f \mid \int_0^\infty |f(\rho)|^2 \rho^{2h-1} \, d\rho < \infty \right\} \to L^2(-h + i\mathbb{R}),$$

- *possesses an inverse mapping M_h^{-1} given by*

$$M_h^{-1}[Y](\rho) := \frac{1}{i\sqrt{2\pi}} \int_{-h-i\infty}^{-h+i\infty} \rho^\lambda Y(\lambda) \, d\lambda,$$

- *satisfies*

$$M[(\rho\partial_\rho)^k y](\lambda) = \lambda^k M[y](\lambda), \quad k \in \mathbb{N}.$$

We can apply M to (3.5) and solve the simpler problem for $Y(\lambda, \theta)$, i.e.,

$$\begin{aligned} \lambda^2 Y + \partial_\theta^2 Y &= M[-\rho^2 g], \quad \text{in } (0, \omega_X), \\ Y(\lambda, 0) = Y(\lambda, \omega_X) &= 0. \end{aligned} \tag{3.6}$$

In order to solve (3.6) with the help of Theorem 3.1.1, the integrability of $-\rho^2 g$ as defined in (3.5) has to be investigated. Since the cut-off function η_X is smooth and $y \in H^1(\Omega)$, the behavior of f near the vertex X is crucial.

Let $f \in V_\beta^{0,2}(\Omega)$ with $\beta \geq 0$ and set $\Re(\lambda) = 1 - \beta = -h$, then

$$\int_{C_X} \rho^{2\beta} g(x)^2 \, \mathrm{d}x = \int_0^{\omega_X} \int_0^\infty \rho^{2\beta} g(\rho, \theta)^2 \rho \, \mathrm{d}\rho \, \mathrm{d}\theta = \int_0^{\omega_X} \int_0^\infty \rho^{2h-1} |\rho^2 g(\rho, \theta)|^2 \, \mathrm{d}\rho \, \mathrm{d}\theta. \tag{3.7}$$

is finite. This proves that the Mellin-Transform of $-\rho^2 g$ exists for $\Re(\lambda) \leq 1$. The case $\Re(\lambda) = 1$ corresponds to $f \in L^2(\Omega)$.

The inverse mapping of the Mellin-Transform works on lines parallel to the imaginary axis. We find a solution y of (3.4) if we apply the inverse mapping of Theorem 3.1.1 to solutions Y of (3.6) on a line $\{\Re(\lambda) = -h\}$ where

$$\begin{aligned} \lambda^2 V + \partial_\theta^2 V &= 0, \\ V(\lambda, 0) = V(\lambda, \omega_X) &= 0 \end{aligned} \tag{3.8}$$

has only trivial solutions (implying that Y is unique). The behavior of the resulting function for $r \to \infty$ depends on the value of $\Re(\lambda) = 1 - \beta = -h$ that was chosen for the inverse transform. We have the following regularity ([92, 96]):

$$M_h^{-1}[Y(\lambda)] \in V_\beta^{2,2}(C_X). \tag{3.9}$$

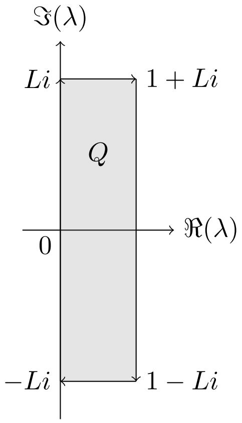


Figure 3.1. The domain of integration for evaluating inverse Mellin-Transforms.

An expansion of the solution to Poisson's equation is obtained as follows. The solutions of (3.8) are given by

$$V = C\sin(\lambda\theta) + D\cos(\lambda\theta), \quad C, D \in \mathbb{C}.$$

The constants need to be adjusted to fit the boundary conditions, which implies that only the case $\lambda_k = k\pi/\omega$ allows non-trivial solutions for $k \in \mathbb{Z}\setminus\{0\}$, i.e., $V_k = \sin(\frac{k\pi}{\omega_X}\theta)$. Obviously, the eigenvalues are real and distributed symmetrically around zero.

For $\omega_X > \pi$ the lines $\{\Re(\lambda) = 1\}$ and $\{\Re(\lambda) = 0\}$ are free of eigenvalues. Due to [92], the original problem (3.5) has a solution in $V_1^{2,2}(\Omega)$ which corresponds to the inverse Mellin-Transform with $\Re(\lambda) = h = 0$ because of (3.9). Consequently, we need to evaluate

$$\frac{1}{i\sqrt{2\pi}}\int_{1-i\infty}^{1+i\infty} Y(\lambda,\theta)\rho^{\lambda}\,\mathrm{d}\lambda.$$

This is done with the help of the residue theorem and the box domain Q depicted in Figure 3.1.

$$\begin{aligned}\sqrt{2\pi}\lim_{L\to\infty}\int_Q Y\rho^{\lambda}\,\mathrm{d}\lambda = \frac{1}{i\sqrt{2\pi}}\int_{-i\infty}^{+i\infty} Y(\lambda,\theta)\rho^{\lambda}\,\mathrm{d}\lambda - \frac{1}{i\sqrt{2\pi}}\int_{1-i\infty}^{1+i\infty} Y(\lambda,\theta)\rho^{\lambda}\,\mathrm{d}\lambda\\ = \sqrt{2\pi}\sum_{\lambda\in Q}\mathrm{Res}(Y(\lambda,\theta)\rho^{\lambda})\end{aligned}$$

because the integrals for the horizontal parts of Q vanish in the limit ([96]).

The only pole in the domain of integration is located at $\lambda_1 = \pi/\omega_X$, where the residue reads

$$c_X\rho^{\pi/\omega_X}\sin(\theta\pi/\omega_X), \quad c_X \in \mathbb{R}. \tag{3.10}$$

On account of the regularity property (3.9), we obtain the expansion

$$y(\rho,\theta) = w_X(\rho,\theta) + c_X\rho^{\pi/\omega_X}\sin(\theta\pi/\omega_X) \tag{3.11}$$

with $w_X \in H^2(C_X)$.

This procedure can be done for all $X \in \mathcal{X}$. Remember that we implicitly agreed on $y := \eta_X y$ previously, so it follows for the true solution y of (3.1) that

$$y = \sum_{X\in\mathcal{X}}\eta_X^2 y + \left(1 - \sum_{X\in\mathcal{X}}\eta_X^2\right)y,$$

where $\eta_X y$ has the form (3.11).

Thus,

$$y = \sum_{X \in \mathcal{X}} \eta_X c_X \, \rho_j^{\pi/\omega_X} \sin(\theta\pi/\omega_X) + y_0 \tag{3.12}$$

with $\rho_X := \operatorname{dist}(\cdot, X)$ and the regular part

$$y_0 = \sum_{X \in \mathcal{X}} \eta_X w_X + (1 - \sum_{X \in \mathcal{X}} \eta_X^2) y$$

being clearly in $H^2(\Omega)$.

Remark 3.1.2. *From the expansion* (3.12)*, we see that the regularity of a solution y to* (3.1) *is limited to $H^{1+\sigma}(\Omega)$, where*

$$\sigma = \min_{X \in \mathcal{X}} \{\pi/\omega_X\} - \varepsilon, \quad \varepsilon > 0$$

on domains with reentrant corners. On convex domains, there is no pole in Q yielding a smooth solution $y \in H^2(\Omega)$.

Transmission Problems

The following investigations are made under the standing Assumption 2.2.3. Proceeding as before, we obtain the following non-linear eigenvalue problem for $2d$-networks. Suppose J subdomains Ω_j meet at a vertex $X \in \mathcal{X}$, see Figure 3.2 for the notations. Then the non-linear eigenvalue problem is given by

$$\begin{aligned} \lambda^2 V + \partial_\theta^2 V &= 0 && \text{in } (\theta_j, \theta_{j+1}),\ j = 1, \dots, J, && \text{(3.13a)} \\ V(\lambda, \theta_j + 0) &= V(\lambda, \theta_j - 0) && j = 2, \dots, J, && \text{(3.13b)} \\ \kappa_j \partial_\theta V(\lambda, \theta_j + 0) &= \kappa_{j-1} \partial_\theta V(\lambda, \theta_j - 0) && j = 2, \dots, J, && \text{(3.13c)} \end{aligned}$$

compare also (3.6).

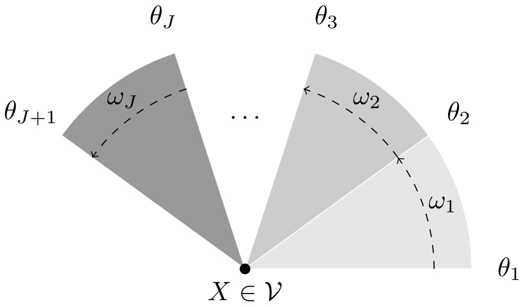


Figure 3.2. A vertex X in the domain Ω (after a change of variables) where J different materials meet.

If no exterior boundary is involved, i.e., $X \cap \partial\Omega = \emptyset$, we set $\theta_{J+1} = \theta_1$ and let both sums in (3.13b),(3.13c) run from $1, \dots, J$ with the convention $\kappa_0 = \kappa_J$. Otherwise, we have additional boundary conditions

$$V(\lambda, \theta_1 = 0) = 0 \quad \text{if } \partial\Omega_1 \cap \Gamma_{\mathcal{D}} \neq \emptyset \quad \vee \quad \partial_\theta V(\lambda, \theta_1 = 0) = 0 \quad \text{if } \partial\Omega_1 \cap \Gamma_{\mathcal{N}} \neq \emptyset, \tag{3.13d}$$

$$V(\lambda, \theta_{J+1}) = 0 \quad \text{if } \partial\Omega_J \cap \Gamma_{\mathcal{D}} \neq \emptyset \quad \vee \quad \partial_\theta V(\lambda, \theta_{J+1}) = 0 \quad \text{if } \partial\Omega_J \cap \Gamma_{\mathcal{N}} \neq \emptyset. \tag{3.13e}$$

A candidate for an eigensolution is (as for the Laplacian) the function

$$V_j = C_j \sin(\lambda(\theta - \theta_j)) + D_j \cos(\lambda(\theta - \theta_j)), \quad \theta \in (\theta_j, \theta_{j+1}). \tag{3.14}$$

The boundary and transmission conditions at the interface give rise to a system of equations for the unknowns C_j, D_j.

It is proved by induction ([117, Example 2.29]) that the Dirichlet-Dirichlet problem (3.13) with $V(\lambda, 0) = V(\lambda, \theta_{J+1}) = 0$ in (3.13d), (3.13e) is solved by

$$C_{j+1} = \frac{D_1}{\Pi_{\nu=2}^{i} \kappa_\nu} d_j^{\mathcal{D}}(\lambda), \quad D_{j+1} = \frac{D_1}{\Pi_{\nu=2}^{j+1} \kappa_\nu} d_j^{\mathcal{M}}(\lambda)$$

with the recursion formula

$$d_1^{\mathcal{D}}(\lambda) = \sin(\lambda\omega_1), \tag{3.15a}$$

$$d_1^{\mathcal{M}}(\lambda) = \kappa_1 \cos(\lambda\omega_1), \tag{3.15b}$$

$$d_j^{\mathcal{D}}(\lambda) = \kappa_j \cos(\lambda\omega_j) d_{j-1}^{\mathcal{D}}(\lambda) + \sin(\lambda\omega_j) d_{j-1}^{\mathcal{M}}(\lambda), \tag{3.15c}$$

$$d_j^{\mathcal{M}}(\lambda) = -\kappa_j^2 \sin(\lambda\omega_j) d_{j-1}^{\mathcal{D}}(\lambda) + \kappa_j \cos(\lambda\omega_j) d_{j-1}^{\mathcal{M}}(\lambda). \tag{3.15d}$$

The Dirichlet condition $V(\lambda, \theta_{J+1}) = 0$ in (3.13e) is equivalent to $D_{J+1} = 0$. Hence, there are non-trivial solutions if and only if $d_J^{\mathcal{D}}(\lambda)$ is zero. Let us note that the condition $d_J^{\mathcal{M}}(\lambda) = 0$ determines the eigenvalues for the transmission problem with mixed boundary conditions, i.e., $V(\lambda, 0) = 0$ in (3.13d) and $\partial_\theta V(\lambda, \theta_{J+1}) = 0$ in (3.13e).

In an analogous way, the discriminant $d_J^{\mathcal{N}}(\lambda)$ for the Neumann-Neumann problem with the boundary condition $\partial_\theta V(\lambda, 0) = 0$ in (3.13d) and $\partial_\theta V(\lambda, \theta_{J+1}) = 0$ in (3.13e) can be derived. It involves the discriminant $d_J^{\mathcal{M}'}(\lambda)$ for the mixed transmission problem with $\partial_\theta V(\lambda, 0) = 0$ in (3.13d) and $V(\lambda, \theta_{J+1}) = 0$ in (3.13e). We find

$$C_{j+1} = \frac{C_1}{\Pi_{\nu=2}^{j} \kappa_\nu} d_j^{\mathcal{M}'}(\lambda), \quad D_{j+1} = -\frac{C_1}{\Pi_{\nu=2}^{j+1} \kappa_\nu} d_j^{\mathcal{N}}(\lambda)$$

with the recursion formula

$$d_1^{\mathcal{N}}(\lambda) = \kappa_1 \sin(\lambda\omega_1), \tag{3.15e}$$

$$d_1^{\mathcal{M}'}(\lambda) = \cos(\lambda\omega_1), \tag{3.15f}$$

$$d_j^{\mathcal{N}}(\lambda) = \kappa_j^2 \sin(\lambda\omega_j) d_{j-1}^{\mathcal{M}'}(\lambda) + \kappa_j \cos(\lambda\omega_j) d_{j-1}^{\mathcal{N}}(\lambda), \tag{3.15g}$$

$$d_j^{\mathcal{M}'}(\lambda) = \kappa_j \cos(\lambda\omega_j) d_{j-1}^{\mathcal{M}'}(\lambda) - \sin(\lambda\omega_j) d_{j-1}^{\mathcal{N}}(\lambda). \tag{3.15h}$$

Remark 3.1.3. *The eigenvalues of the transmission problem are real numbers because the problem can be written as a self-adjoint operator (see [116, Theorem 2.2]). Additionally, the set of eigenvalues is countable without a cluster point (see [68] with the result of [118, Theorem 3.4]). Furthermore, the roots of $d_J^{\mathcal{D}}(\lambda)$, $d_J^{\mathcal{N}}(\lambda)$, $d_J^{\mathcal{M}}(\lambda)$, $d_J^{\mathcal{M}'}(\lambda)$ are symmetric around zero.*

Note that, at interior vertices, the function $V \equiv \mathrm{const}$ solves the eigenvalue problem for $\lambda = 0$ because there are no boundary conditions present.

Let us introduce some new notation and definitions that allow us to rigorously formulate the main result of this section, i.e., an asymptotic expansion for the transmission problem.

Definition 3.1.4. *Let $X \in \mathcal{X}$ and $L_X(\lambda)$ denote the differential operator corresponding to (3.13a) and $B_X(\lambda)$ denote the operator collecting (3.13b)-(3.13e) depending on the problem posed at vertex X. The eigenvalue problem is abbreviated by*

$$\mathcal{A}_X(\lambda) := (L_X(\lambda), B_X(\lambda)).$$

As the eigenvalue problem for $\mathcal{A}_X(\lambda)$ of Definition 3.1.4 is non-linear, we provide a generalized Definition of eigenvalues, eigensolutions, and generalized Jordan chains. The following definitions are made under the assumption of one arbitrary but fixed vertex $X \in \mathcal{X}$.

Definition 3.1.5. *Let $C_X^{\cap s} := C_X \cap \{\rho = 1\}$ be the intersection of the cone C_X with the one-dimensional sphere. A number $\lambda_0 \in \mathbb{C}$ is an **eigenvalue** of the operator $\mathcal{A}_X(\lambda)$ if there is a non-trivial function $s_{\lambda_0,0} \in H^2(C_X^{\cap s})$ (called **eigensolution**) with*

$$\mathcal{A}_X(\lambda_0)s_{\lambda_0,0} = 0.$$

If λ_0 is an eigenvalue of $\mathcal{A}(\lambda)$, there are $\dim \mathrm{Ker}(\mathcal{A}_X(\lambda)) =: I_{\lambda_0}$ linearly independent eigensolutions $s_{\lambda_0,i,0}$ with $i = 1, \dots, I_{\lambda_0}$. Besides them, there may exist $N_{\lambda_0,i}$ associated eigenfunctions $s_{\lambda_0,i,j}$.

Definition 3.1.6. *The system $\{s_{\lambda_0,i,j}\}$ with $i = 1, \dots I_{\lambda_0}$, $j = 0, \dots, N_{\lambda_0,i}$ consists of eigensolutions and **associated eigensolutions** (called a system of Jordan chains) if*

$$\sum_{\nu=0}^{k} \frac{1}{\nu!} \partial_\lambda^\nu \mathcal{A}_X(\lambda_0) s_{\lambda_0,i,k-\nu} = 0$$

*for $k = 0, \dots, N_{\lambda_0,i}$, where $N_{\lambda_0,i}$ (decreasing with respect to i) denotes the **size of the Jordan chain**. An eigenvalue λ_0 is called **simple** if no associated eigensolutions exist, i.e., $N_{\lambda_0} = 0$.*

Theorem 3.1.7. *Let $\lambda_{X,j}$, $j = 1, \dots, N_X$ denote all eigenvalues of $\mathcal{A}_X(\lambda)$ in $(0, 1]$ and assume that $\lambda_{X,j} \neq 1$. Then the solution of (T) with $u = 0$ on a 2d-network admits the expansion*

$$y = y_0 + \sum_{X \in \mathcal{V}} \sum_{j=1}^{N_X} c_{X,j}\, \eta_X\, \rho_X^{\lambda_{X,j}}\, s_{X,j}(\theta)$$

with $\rho_X := \operatorname{dist}(\cdot, X)$. Here, $u_{0,i} \in H^2(\Omega_i)$ $i \in I$, $c_{X,j} \in \mathbb{R}$, $s_{X,j} \in H^1(]0, \theta_{J(X)}[)$ and η_X is a smooth cut-off function.

This result comprises Remark 3.1.2 and is rigorously proved in [117, Theorem 2.27]. The eigenvalues of (3.13) are simple, which is why a similar result holds in weighted Sobolev spaces (see [117, Theorem 3.6]).

An expansion of the solution y into a regular part y_0 and singular contributions can be derived for more general $2m$-coercive problems ($m \geq 1$) and transmission problems. The proofs are more involved and use (semi-)Fredholm properties of the general operators. At the core, however, a non-linear eigenvalue problem similar to (3.8) and its solvability is discussed on the infinite cone (see [117, 118] and the references therein).

Allowing non-homogeneous jumps in the normal derivative, i.e., $u \neq 0$ in (T), seriously complicates the analysis. In order to formulate an expansion result for the inhomogeneous problem, it is necessary to introduce the concept of injectivity modulo polynomials, which is described in [118] or [49].

For a cone C_X, we define the restriction $C_{X,i}$ to be the part of C_X on which $\kappa = \kappa(\theta) = \kappa_i$. Analogous to the notation on $2d$-networks $\{\Omega_i\}_{i\in I}$, we use $\mathcal{E}_\mathcal{N}, \mathcal{E}_\mathcal{D}$, and $\mathcal{I}$ for the Neumann boundary, Dirichlet boundary, and interface, respectively. Furthermore, $y_{X,i} := y|_{C_{X,i}}$ for functions y defined on C_X.

Definition 3.1.8. *Let D be an open subset of $\mathbb{R}^n$ and let $l \in \mathbb{Z}$ and $X \in \mathcal{X}$. We define the* ***homogeneous polynomial spaces of degree l*** *as*

$$
\begin{aligned}
P_l^H(D) &:= \{q \mid q \text{ is a homogeneous polynomial of degree } l \text{ defined on } D\}, \quad l \geq 0,\\
P_l^H(D) &:= \{0\}, \quad l < 0,\\
P_l^H(C_X) &:= \{q : C_X \to \mathbb{R} \mid q_i \in P_l^H(C_{X,i})\}.
\end{aligned}
$$

For the data of (T), we define the polynomial space

$$
\begin{aligned}
\Upsilon_l^H(C_X) := P_{l-2}^H(C_X) \times \prod_{\gamma_{i,i'} \in \mathcal{I}} P_l^H(\gamma_{i,i'}) \times \prod_{\gamma_{i,i'} \in \mathcal{I}} P_{l-1}^H(\gamma_{i,i'})\\
\times \prod_{\Gamma_i \in \mathcal{E}_\mathcal{D}} P_l^H(\Gamma_i) \times \prod_{\Gamma_{ii'} \in \mathcal{E}_\mathcal{N}} P_{l-1}^H(\Gamma_i).
\end{aligned}
$$

Definition 3.1.9. *We say $\mathcal{A}_X(\lambda)$ is* ***injective modulo polynomials of order l*** *(short i.m.p) for $l \in \mathbb{N}$ on C_X if and only if any solution w_X that solves*

$$
\left\{
\begin{aligned}
-\kappa_i \Delta w_{X,i} &= f_i && \text{in} && C_X,\\
w_{X,i} - w_{X,i'} &= d_{i,i'} && \text{on} && \gamma_{i,i'} \in \mathcal{I},\\
\kappa_i \partial_{\theta_i} w_{X,i} - \kappa_{i'} \partial_{\theta_{i'}} w_{X,i'} &= u_{i,i'} && \text{on} && \gamma_{i,i'} \in \mathcal{I},\\
w_{X,i} &= g_i && \text{on} && \Gamma_i \in \mathcal{E}_\mathcal{D},\\
w_{X,i} &= h_i && \text{on} && \Gamma_i \in \mathcal{E}_\mathcal{N}
\end{aligned}
\right. \tag{3.16}
$$

with a right hand side in $\Upsilon_l^H(C_X)$ belongs to the space $P_l^H(C_X)$.

Remark 3.1.10. *If the operator is i.m.p., then every solution w_X of (3.16) with polynomial data is itself polynomial.*

In order to answer the question on injectivity modulo polynomials in practice, a characterization of w_X is necessary.

Proposition 3.1.11. *Let w_X be a solution of (3.16) with polynomial data from $\Upsilon_l^H(C_X)$ with $l > 0$. Then the restriction $w_{X,i}$ to one subdomain $C_{X,i}$ looks like (we suppress the index i for better readability)*

$$w_{X,l}(\rho_X, \theta) = P_{X,l}(\rho_X, \theta) + \sum_{\lambda_{X,j}=l} c_{X,j}\ \rho_X^l(\ln(\rho_X)s_{X,j}(\theta) + \theta\partial_\theta s_{X,j}(\theta)), \tag{3.17}$$

where $\rho_X = \operatorname{dist}(\cdot, X)$ and $P_{X,l}$ is a homogeneous polynomial of degree l (in Cartesian coordinates).

For the proof, we refer the reader to [117, Theorem 3.10], where the result extends to $l \geq 0$. Note that general elliptic equations lead to higher powers of $\ln(\rho_X)$ owing to the presence of more associated eigenfunctions (see [118, Lemma 7.1]).

A discussion of injectivity modulo polynomials for elliptic boundary value problems can be found in [49]. Under the assumption of injective modulo polynomials we have the following result.

Theorem 3.1.12. *Assume that $\mathcal{A}_X(\lambda)$ is i.m.p. of order 1 for all $X \in \mathcal{X}$ and the line $\{\Re(\lambda) = 1\}$ contains no eigenvalue of $\mathcal{A}(\lambda)$ except possibly at $\lambda = 1$. Then there exists a solution y to (T) that satisfies the expansion*

$$y = y_0 + \sum_{X\in\mathcal{V}} \sum_j c_{X,j}\ \eta_X\ \rho_X^{\lambda_{X,j}}\ s_{X,j}(\theta)$$

with $\rho_X = \operatorname{dist}(\cdot, X)$. Here, $y_0 \in L^2(\Omega)$ and $y_{0,i} \in H^2(\Omega_i)$, $c_{X,j} \in \mathbb{R}$, $s_{X,j} \in H^2(]0, \sum_i \omega_{X,i}[)$ and η_X is a smooth cut-off function.

Proof. The proof is the same as [118, Theorem. 7.4] where we set $k = 0$, $m = 1$, $p = 2$ to cover the situation considered here. We only point out some important steps. The fact $u \in H^{1/2}(\mathcal{I})$ allows to construct a lift function $v \in V_0^{2,2}(\Omega)$ that exactly fulfills the transmission conditions [118, Lemma 4.3]. A unique solution and its expansion with a regular part in the weighted Sobolev space $V_0^{0,p}(\Omega)$ follows from [118, Corollary 4.4]. The result for $p = 2$ is then obtained by an interpolation argument which exploits the property of injectivity modulo polynomials of order 1. □

The eigenvalues of $\mathcal{A}_X(\lambda)$, which is defined for transmission problems (see Definition 3.1.4), are real (Remark 3.1.3). Hence, the only prerequisite we have to check for is the injectivity modulo polynomials. Note that the result also holds for more general problems and corresponding operators $\mathcal{A}_X(\lambda)$ (see [118]).

3.1.2 Lower Bounds for eigenvalues

An expansion like the one provided in Theorem 3.1.7 or Theorem 3.1.12 allows to establish higher regularity (locally and globally) by bounding the eigenvalues of $\mathcal{A}(\lambda)$ from below. The singular functions in the expansion of Theorem 3.1.7 or 3.1.12 satisfy

$$\eta_X \rho_X^\lambda s_X \in H^{1+\lambda-\varepsilon}(\Omega) \quad \text{and} \quad \eta_X \rho_X^\lambda s_X \notin H^{1+\lambda}(\Omega) \tag{3.18}$$

for small $\varepsilon > 0$ (see [71, Theorem 1.2.18] or [117, Theorem 1.26]). Hence, the lowest exponent of ρ_X decides on the global regularity of a solution to the transmission problem.

Corollary 3.1.13. *Assume that $\mathcal{A}_X(\lambda)$ is i.m.p. of order 1 for all $X \in \mathcal{V}$ and that there is a $\sigma \in (0,1]$ such that $\lambda_{X,j} > \sigma$. Then the solution y of* (T) *satisfies for small $\varepsilon > 0$*

$$y_i \in H^{1+\sigma}(\Omega_i), \quad y \in H^{1+\min\{1/2-\varepsilon,\sigma\}}(\Omega).$$

Proof. For the homogeneous case $u = 0$, the result follows from Theorem 3.1.7 and [71, Theorem 1.2.18]. In the case of $\lambda_{X,j} \neq 1$, we have to rely on Sobolev embeddings and sharper results in L^p-spaces (see [120, Corollary 2.1]).

In the inhomogeneous case, we resort to Theorem 3.1.12 and [71, Theorem 1.2.18]. The global regularity is a general consequence of the observation that $y \in H^{3/2-\varepsilon}(\Omega)$. □

The Theorem could be refined by allowing lower bounds σ_i for the vertices of each subdomain Ω_i of the $2d$-network.

Quasi-Monotone Distributions

Without additional assumptions, it is impossible to find a lower bound for the eigenvalue distribution of $\mathcal{A}_X(\lambda)$, which the following example shows. For a mixed boundary value problem with $\omega_1 = \omega_2 = \pi/2$, it holds

$$d_2^{\mathcal{M}}(\lambda) = -\kappa_2^2 \sin^2(\lambda\pi/2) + \kappa_1\kappa_2 \cos^2(\lambda\pi/2).$$

A vanishing discriminant $d_2^{\mathcal{M}}(\lambda) = 0$ leads to

$$\tan(\lambda\pi/2) = \pm\sqrt{\frac{\kappa_1}{\kappa_2}}.$$

Letting $\kappa_2 \to \infty$ sends $\lambda \to 0$.

Such phenomena can be avoided with the help of quasi-monotone distributions of diffusion coefficients (see [120]), first introduced in [56].

Definition 3.1.14. *Let $\kappa_i > 0$ with $i \in I$ be the distribution of diffusion coefficients for a $2d$-network. Assume that*

$$i \neq j, \ \operatorname{meas}_1(\overline{\Omega}_i \cap \overline{\Omega}_j) > 0 \quad \Rightarrow \quad \kappa_i \neq \kappa_j.$$

Let at $X \in \mathcal{V}$ meet J different domains with $\kappa_j,\ j = 1, \ldots, J$. We denote by $\kappa_{a_j}, \kappa_{b_j}$ the material constants of the domains which abut on material j with positive one dimensional measure (one of them being zero if there is only one neighbor). The distribution of κ_i is called ***quasi-monotone*** *if the following assumptions hold for all $X \in \mathcal{V}$:*

- *If X lies in $\overline{\Omega} \setminus \Gamma_{\mathcal{D}}$, then*

$$\exists!\ j \in \{1, \ldots, J\} : \quad \kappa_j > max(\kappa_{a_j}, \kappa_{b_j}).$$

- *If X lies on $\Gamma_{\mathcal{D}}$, then $\kappa_j > max(\kappa_{a_j}, \kappa_{b_j})$ implies that $\operatorname{meas}_1(\overline{\Omega}_j \cap \Gamma_{\mathcal{D}}) > 0$.*

Remark 3.1.15. *Three materials meeting at an interior point are automatically distributed in a quasi-monotone way, i.e., the first condition only poses restrictions for $k \geq 4$. The second condition states that, locally, the domain with maximal material constant has to touch the Dirichlet boundary.*

Using this definition, we can prove the following result ([120, Theorem 2.8 and 2.9]).

Theorem 3.1.16. *If the coefficients κ_i are quasi-monotone, then the eigenvalues $\lambda_{X,i}$ of $\mathcal{A}_X(\lambda)$ satisfy for each $X \in \mathcal{X}$*

$$\lambda_{X,i} > \frac{1}{4}.$$

Proposition 3.1.17. *Let Ω consist only of two subdomains Ω_1, Ω_2. For each $X \in \mathcal{X}$ and opening angle $\omega_X \in (0, 2\pi]$, we have in the case of $\mathcal{E}_{\mathcal{N}} = \emptyset$*

$$\lambda_{X,i} \geq \frac{1}{2}.$$

Proof. The result for $X \in \Gamma$ can be found in [46, Theorem 8.1]. For interior vertices, we refer to [89, Section 2.2]. □

Two numerical examples for the eigenvalue distribution are shown in Figure 3.3: there the dependence of eigenvalues of $\mathcal{A}_X(\lambda)$ is depicted for the case where two materials meet with angles ω_1, ω_2 at a boundary and an interior vertex, respectively. We refer the reader to [120] and [89] regarding a further discussion of lower bounds.

Corollary 3.1.18. *Suppose that the coefficients κ_i are distributed in a quasi-monotone way and that $\mathcal{A}_X(\lambda)$ is i.m.p of order 1 for all $X \in \mathcal{X}$. Under Assumption 2.2.3, the solution of* (T) *lies in $H^{5/4}(\Omega)$.*

Proof. Due to Theorem 3.1.16, we can apply Corollary 3.1.13 with the lower bound $\sigma = 1/4$. □

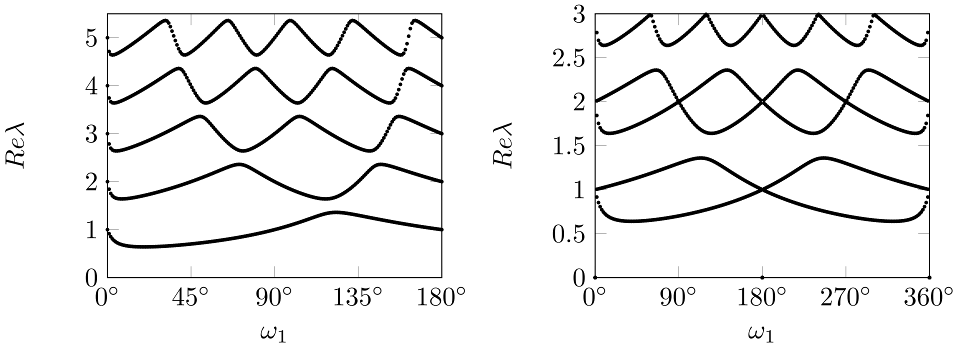
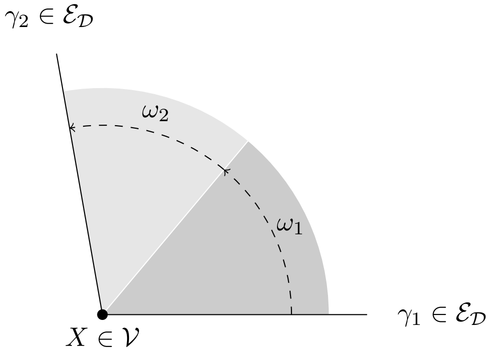


Figure 3.3: eigenvalue distribution for a boundary (left, $\omega_1 + \omega_2 = 180°$) and interior (right, $\omega_1 + \omega_2 = 360°$) vertex with $\kappa_1 = 0.25,\ \kappa_2 = 5$.

Examples

Let us discuss the concept of injectivity modulo polynomials of order 1 for the transmission problem with $\mathcal{E}_\mathcal{N} = \emptyset$ and some showcase vertices $X \in \mathcal{X}$. After that, we provide several examples for elliptic PDEs that are $H^{1+\sigma}$-regular, which is a property that is often needed in the following chapters.

Let C_X be the cone at an exterior vertex $X \in \mathcal{X}$ with Dirichlet-Dirichlet boundary conditions and two materials (see Figure 3.4).



$\gamma_2 \in \mathcal{E}_\mathcal{D}$
ω_2
ω_1
$\gamma_1 \in \mathcal{E}_\mathcal{D}$
$X \in \mathcal{V}$

Figure 3.4. The Dirichlet-Dirichlet problem at a conical point with two materials.

From (3.15c) we get

$$d_2^{\mathcal{D}}(\lambda) = \kappa_2 \cos(\lambda\omega_2)\sin(\lambda\omega_1) + \kappa_1 \sin(\lambda\omega_2)\cos(\lambda\omega_1).$$

For $\omega_1 = \omega_2$ it is obvious that $\lambda = 1$ is an eigenvalue if $\sin(\omega_1) = 0$, i.e., $\omega_1 = \pi/2$. For the decomposition in (3.17), we find with (3.14) that

$$s_{X,1} = \sin(\theta), \quad s_{X,2} = \cos(\theta - \pi/2) = \sin(\theta).$$

Observing that

$$P_X = \begin{cases} \frac{1}{\kappa_1}\rho_X \cos(\theta) & \text{if } \theta \in (0, \pi/2), \\ -\frac{1}{\kappa_2}\rho_X \cos(\theta) & \text{if } \theta \in (\pi/2, \pi), \end{cases}$$

is a polynomial that solves (3.16)v with polynomial data from $\Upsilon_1^H(C_X)$, we set

$$w_{X,i} = P_{X,i} + \frac{\rho_X}{\kappa_i}(\ln(\rho_X)\sin(\theta) + \theta\cos(\theta)).$$

A simple calculation shows that $\Delta w_X = 0$. Continuity of w_X at $\theta = \pi/2$ is also fulfilled as well as the jump in the normal derivative

$$\kappa_1 \partial_\theta w_{X,1} = -\rho_X - \pi/2 = -(\rho_X + \pi/2) = -\kappa_2 \partial_\theta w_{X,2}.$$

So w_X solves a problem with polynomial data but is itself non-polynomial. Consequently, the operator $A_X(\lambda)$ is not i.m.p. of order 1 at such a vertex $X \in \mathcal{X}$.

The situation is different if 1 is no eigenvalue of $\mathcal{A}_X(\lambda)$. Then the sum in (3.17) is empty and Proposition 3.1.11 yields a unique (polynomial) solution, which in turn guarantees i.m.p. of order 1. This can be also seen in the left diagram of Figure 3.3. There, eigenvalues λ were computed for

$$\kappa_1 = 0.25, \quad \kappa_2 = 5,$$
$$0 < \omega_1 < \pi, \quad \omega_1 + \omega_2 = \pi.$$

According to Figure 3.3, $\lambda = 1$ is only an eigenvalue if $\omega_1 = \pi/2$. As vertices with only one material do not pose a problem due to [49, Section 4], the operator $\mathcal{A}(\lambda)$ is i.m.p. of order 1 for the domain shown in Figure 3.5.

For two materials and an interior vertex, $\lambda = 1$ is never an eigenvalue. This can be seen as follows. Setting $J = 2$ in (3.13) and inserting the solution candidate from (3.14) yields a system of equation with the determinant (see also [117, Example 2.30])

$$d(\lambda) = (\kappa_1 - \kappa_2)^2 \sin^2(\lambda(\pi - \omega_1)) + (\kappa_1 + \kappa_2)^2 \sin^2(\lambda\pi).$$

Assume $d(1) = 0$, then it follows

$$\omega_1 = 0 \quad \vee \quad \omega_1 = \pi,$$

which can also be observed in the right diagram of Figure 3.3. However, these two cases do not allow for an interior vertex with two materials. Hence, the operator $\mathcal{A}(\lambda)$ is i.m.p on the model domain of Figure 3.6.

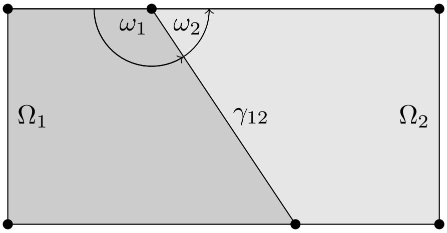


Figure 3.5. A suitably shaped domain to apply Theorem 3.1.12: For $\omega_1 \neq \pi/2$ the exterior vertices exclude $\lambda = 1$ as eigenvalue, which results in injectivity modulo polynomials.

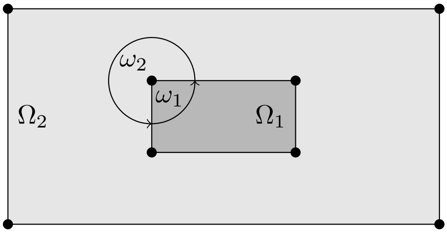


Figure 3.6. A suitably shaped domain to apply Theorem 3.1.12: For $\omega_1 \neq \pi$ the interior vertices exclude $\lambda = 1$ as eigenvalue, which results in injectivity modulo polynomials.

Let us comment on the consequences of the asymptotic expansions of solutions to elliptic PDEs on polygonal domains. The lower bound of an abstract eigenvalue problem decides over the exponents of the singular components of the solution and, hence, over the $H^{1+\sigma}$-regularity of the PDE. For instance, a convex domain implies H^2-regularity for Poisson's equation with homogeneous Dirichlet conditions (confer Remark 3.1.2 or [72]). This result continues to hold for general boundary conditions if each corner where $\Gamma_{\mathcal{D}}$ and $\Gamma_{\mathcal{N}}$ meet has an opening angle $\omega \leq \pi/2$ (see [72, Corollary 4.4.3.8]). The analysis of the corresponding eigenvalue problem for mixed and Neumann boundary conditions ([96, Section 2.1]) shows that the state equation is $H^{1+\sigma}$-regular with $\sigma = 1/4$ on general domains. This also holds true for $\sigma = 1/2$ and the pure Neumann problem as shown in [86].

The transmission problem is $H^{1+\sigma}$-regular with $\sigma = 1/4$ if the the coefficients are distributed quasi-monotone and the operator $\mathcal{A}(\lambda)$ is i.m.p. of order 1 (combine Theorem 3.1.13 and 3.1.16). Optimal global regularity, i.e., $\sigma = 1/2 - \varepsilon$ is obtained under the assumptions of Proposition 3.1.17. Note that the local regularity of a solution on a $2d$-network can be higher. Instead of appealing to theoretical results, we can numerically evaluate the eigenvalues for each vertex of a specific problem and obtain and estimate for σ, confer Figure 3.3.

3.2 Analytic Regularity far From the Boundary and Interface

Assuming a higher regularity of the PDE constraint, we now show that the state and adjoint variable y, q belong to the countably normed space $B^2_{1-\sigma}(\Omega)$, $\sigma \in (0, 1]$. Neumann control problems are treated with the weight function r_Γ, whereas interface problems use $r_{\mathcal{I}\cup\Gamma}$ in the sense of (2.4). In the view of Theorem 2.1.7, this establishes analytic regularity distant from the boundary and interface, respectively.

We recall a result from [88].

Theorem 3.2.1. *Let $f \in B^0_{1-\sigma}(C_f, \gamma_f)$ with $C_f, \gamma_f > 0$, $\sigma \in (0, 1]$ and the weight function r_Γ. Additionally, let $y \in H^{1+\sigma}(\Omega)$ solve the differential equation*

$$-\nabla \cdot (D(x)\nabla y) + c(x)y = f \qquad \text{in } \Omega,$$

which fulfills Assumption 2.2.1. Then there exist constants $C, \gamma > 0$ that depend only on Ω, C_D, C_f, γ_D, γ_f, and σ such that

$$\|r_\Gamma^{p+1-\sigma} \nabla^{p+2} y\|_{L_2(\Omega)} \leq C\gamma^p p! \left(C_f + \|y\|_{H^{1+\sigma}(\Omega)}\right) \quad \forall p \in \mathbb{N}_0,$$

which implies $y \in B^2_{1-\sigma}\left(C(C_f + \|y\|_{H^{1+\sigma}(\Omega)}),\ \gamma\right)$.

Proof. By closely inspecting the technical proof of [88, Theorem A.1], which builds on [106], one can see that the assumptions on f are sufficient for obtaining the theorem. □

We would like to stress that this regularity result holds without any assumptions on the boundary data because the weight function r_Γ damps possible singularities. Therefore, Theorem 3.2.1 is suitable for Neumann control problems whose solution u^* may have kinks stemming from the projection formula.

Corollary 3.2.2. *Let the assumptions of Theorem 3.2.1 hold. Let additionally $y_d \in B^0_{1-\sigma}(C_d, \gamma_d)$ for $C_d, \gamma_d > 0$ and the weight function r_Γ. If $q \in H^{1+\sigma}(\Omega)$ is a solution of the adjoint equation*

$$-\nabla \cdot (D(x)\nabla q) + c(x)q = y - y_d \qquad \text{in } \Omega,$$

then there exists constants $C, \gamma > 0$ such that $q \in B^2_{1-\sigma}(C, \gamma)$.

Proof. From Theorem 3.2.1 we get $y \in B^2_{1-\sigma}(C_y, \gamma_y)$ with $C_y, \gamma_y > 0$. As r_Γ is bounded by one, we easily obtain $y \in B^0_{1-\sigma}(C_y, \gamma_y)$. The regularity of y_d allows to apply Theorem 3.2.1 again, which concludes the proof. □

The countably normed spaces $B^2_{1-\sigma}$ use the weight function r_Γ to control the blow-up towards the whole boundary and are of importance for convergence of the boundary concentrated finite element method (see Section 4.3). The above results can be extended to piecewise analytic data or subdomain observation by using appropriate weight functions for the countably normed spaces (see [27]). Similarly, we use $r_{\mathcal{I}\cup\Gamma}$ instead of r_Γ and obtain analogous regularity features for the transmission problem (T).

Corollary 3.2.3. *Let $\{\Omega_i\}_{i\in I}$ be a 2d-network and $y, q \in H^{1+\sigma_i}(\Omega_i)$ be solutions of*

$$-\nabla \cdot (\kappa_i \nabla y) = f_i, \quad -\nabla \cdot (\kappa_i \nabla q) = y_i - y_{d,i} \quad \text{in } \Omega_i,$$

with $\kappa_i > 0, \sigma_i \in (0, 1]$ and $f_i, y_{d,i} \in B^0_{\sigma_i}(\Omega_i, C, \gamma)$ with the weight function $r = r_{\partial\Omega_i}$. Then there exist constants $\overline{C}, \overline{\gamma} > 0$ that depend only on the data such that $y, q \in B^2_{1-\sigma}(\Omega, \overline{C}, \overline{\gamma})$ for the weight function $r_{\mathcal{I}\cup\Gamma}$ and $\sigma := \min_{i\in I}\{\sigma_i\}$.

Proof. Applying Theorem 3.2.1 on each subdomain Ω_i, we obtain $y, q \in B^2_{1-\sigma_i}(\Omega_i, C_i, \gamma_i)$ for $r_{\partial\Omega_i}$ and positive constants C_i, γ_i. Setting $\overline{C} := \max_{i\in I}\{C_i\}$ and $\overline{\gamma} := \max_{i\in I}\{\gamma_i\}$ yields the result. □

3.3 Analytic Regularity far From Singular Points

The main result of this section is Theorem 3.3.24, where global analytic regularity of the optimal variables (u^*, y^*, q^*) is established. It can be viewed as a generalization of the previous section because we now take the regularity of the boundary data into account. This manifests in the use of a different weight function, which measures the distance to a finite set of 'singular' points.

3.3.1 Preliminary Assumptions and Remarks

We only consider the Neumann control problem and refer to Subsection 3.3.5 for remarks on the transmission problem. The smoothness of the optimal variables is limited by

- the vertices $\mathcal{X}$ of Ω, which may cause a blow-up in higher derivatives (confer Section 3.1),

- the projection formula (2.17), which may introduce kinks in the optimal control.

The projection representation (2.17) implies that the optimal control inherits regularity from the trace of the adjoint state. This allows to conclude higher regularity of the solution of (**P**).

However, the regularity of the optimal control is also limited by the non-smooth structure of the projection. In general, the optimal control will be at most Lipschitz continuous due to the appearance of kinks at points $x \in \Gamma_{\mathcal{N}}$, where $u_a(x) = -\frac{1}{\nu}q^*(x)$ or $u_b(x) = -\frac{1}{\nu}q^*(x)$. In the subsequent analysis, we will assume that the set of kinks in the control is finite.

Assumption 3.3.1. *We assume that there exists a finite set $\mathcal{S}$ of switching points of u^*. That is, there exists a finite set $\mathcal{S}$ such that u^* fulfills one of the equations*

$$\begin{aligned} u^* &= u_a, \\ u^* &= -\frac{1}{\nu}q^*, \\ u^* &= u_b \end{aligned}$$

on every connected component of $\Gamma_{\mathcal{N}} \setminus \mathcal{S}$.

This assumption will be of essential importance in Subsection 3.3.4. If the optimal control u^* is continuous, the assumption can be replaced by: the boundary of the active set $\mathfrak{A} := \{x \in \Gamma_{\mathcal{N}} \ : \ u^*(x) = u_a$ or $u^*(x) = u_b(x)\}$ is finite. This is not fulfilled in general, we refer to [104, Remark 4.1] for an example of a smooth control with infinitely many switching points. Moreover, the assumption implies that the switching points of u^* are known a-priori. In Section 4.2 we describe how we cope with this difficulty in the numerical computations.

Remark 3.3.2. *Assumption 3.3.1 is slightly stronger than other regularity assumptions used in the literature. In [104], the following assumption was used to prove a-priori finite element error estimates: the set of elements $K \subset \Gamma_{\mathcal{N}}$ such that u^* is not in $H^s(K)$ for some $s \in (\frac{3}{2}, \frac{5}{2})$ has measure proportional to the mesh-size h. If the set of switching points is finite and the elements K are of size h, then clearly this assumption is fulfilled.*

Assumption 3.3.1 ensures that the amount of points where the control changes from inactive to active behavior (or vice versa) is finite and can be included into the weight function.

Definition 3.3.3. *Define the set $\mathcal{V} := \mathcal{X} \cup \mathcal{S} = \{X_1, \ldots, X_m\}$. Let $\beta \in \mathbb{R}^m$ be a multi-index satisfying $(\beta_1, \ldots, \beta_m) \in (0,1)$ (understood component-wise). For $x \in \overline{\Omega}$, $p \in \mathbb{Z}$, we set*

$$r(x) := r_{\mathcal{V}}(x)^{p+\beta} = \prod_{i=1}^{m} \min\{1, \operatorname{dist}(x, X_i)\}^{p+\beta_i}, \quad p \in \mathbb{Z}. \tag{3.19}$$

Accordingly, we partition the boundary into straight line segments $\Gamma = \bigcup_{i=1}^m \overline{\Gamma}_i$ such that each intersection $\overline{\Gamma}_i \cap \overline{\Gamma}_j \neq \emptyset, i \neq j$ lies in $\mathcal{V}$.

The derivation of regularity in boundary weighted spaces relied on the assumption $f, y_d \in B^0_{1-\sigma}(\Omega, C, \gamma_f)$ with the weight function r_Γ. Similar assumptions are necessary if r only measures the distance to a finite set of points.

Assumption 3.3.4. *Let us assume that $f \in L^2(\Omega) \cap B^0_\beta(\Omega, C_f, \gamma_f)$ and $y_d \in L^2(\Omega) \cap B^0_\beta(\Omega, C_d, \gamma_d)$ for a given multi-index $\beta \in (0,1)$ and the weight function $r_{\mathcal{X}}$, i.e., for $p \geq 0$*

$$\begin{aligned} \| r_{\mathcal{X}}^{p+\beta} \nabla^p f \|_{L^2(\Omega)} &\leq C_f \gamma_f^p p!, \\ \| r_{\mathcal{X}}^{p+\beta} \nabla^p y_d \|_{L^2(\Omega)} &\leq C_d \gamma_d^p p!. \end{aligned} \tag{3.20}$$

In addition, we assume that $u_a, u_b \in B^2_\beta(\Omega, C_g, \gamma_g)$.

The assumptions on the control bounds imply that $u_a, u_b \in B^{3/2}_\beta(\Gamma_{\mathcal{N}}, C, \gamma)$ and strengthen the standing assumption of $u_a, u_b \in H^{1/2}(\Gamma_{\mathcal{N}})$. In particular, the embedding $H^{2,2}_\beta(\Omega) \hookrightarrow C^0(\overline{\Omega})$ implies the continuity of the optimal control.

Remark 3.3.5. *In the following investigations, it will be necessary to work with the weight function $r_{\mathcal{V}}$ from above, which satisfies $r_{\mathcal{V}} \leq r_{\mathcal{X}}$ on $\overline{\Omega}$. Hence, Assumption 3.3.4 implies that (3.20) remains valid if we exchange $r_{\mathcal{X}}$ for $r_{\mathcal{V}}$.*

Remark 3.3.6. *The results can also be extended to non-homogeneous Dirichlet boundary conditions by a lifting trace argument provided that the boundary values are regular enough.*

The global regularity result (Theorem 3.3.24) is obtained with the help of the following strategy:

1. Cover the domain with balls with a finite overlap property.
2. Apply local estimates for interior balls or half balls with boundary conditions.
3. Add up the estimates and obtain global bounds on the derivatives.

We will carry out this strategy in the following sections.

3.3.2 Covering Results

We will denote by $B_r(x)$ the open ball of radius $r > 0$ around $x \in \mathbb{R}^2$.

Definition 3.3.7. *Let $\mathcal{B}$ be a collection of open balls. We say $\mathcal{B}$ is a dichotomic with respect to Ω if one of the following two conditions is satisfied for each $B \in \mathcal{B}$.*

1. *B is contained in Ω, i.e., $B \cap \Omega = \Omega$,*
2. *$B \cap \Omega$ is a half-ball with the same center and radius as B.*

We will now show the existence of a dichotomic covering of the polygonal domain Ω that helps to resolve singularities of the solution near the vertices $\mathcal{X}$ of Ω and $\partial\mathfrak{A}$. Moreover, the covering resolves the active and inactive sets: if for a ball B of the covering $B \cap \Gamma_{\mathcal{N}} \neq \emptyset$, then it holds either $u^*|_{\Gamma_{\mathcal{N}} \cap B} = u_a$ or $u^*|_{\Gamma_{\mathcal{N}} \cap B} = -\nu^{-1} q^*$ or $u^*|_{\Gamma_{\mathcal{N}} \cap B} = u_b$. Thus, locally on B the optimal control problem has no inequality constraints.

The distance that helps to localize the area around vertices is defined as

$$\delta := \min \left\{ 1, \min_{\substack{X_i, X_j \in \mathcal{V}, \\ X_i \neq X_j}} \operatorname{dist}(X_i, X_j), \ \inf\{r > 0 \mid \ B_r(X) \cap \Omega \text{ is a sector } \forall X \in \mathcal{V}\} \right\}. \tag{3.21}$$

The last component of the set in (3.21) is included because straight parts Γ_i of the boundary may have arbitrarily small distance to points $X \in \mathcal{V}$ with $X \cap \overline{\Gamma}_i = \emptyset$. Controlling the mutual distance of points is, therefore, not enough, as Figure 3.7 illustrates.

Observe that $\delta > 0$ is well defined, because each of the finitely many points in $\mathcal{V}$ has an opening angle in $(0, 2\pi)$.

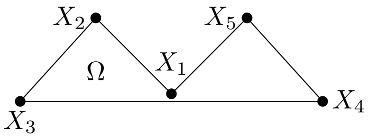


Figure 3.7: A domain where small neighborhoods around a vertex X_1 may not be sectors in Ω.

Before we construct the desired covering of Ω, we need the following technical Lemma, which collects some necessary trigonometric relations.

Lemma 3.3.8. *Let $\underline{\omega} \in (0, \pi)$ and $\varepsilon \in (0, 1)$ be given. Then there exists positive numbers $c \in (0, 1/2)$ and α satisfying the following conditions:*

$$0 < \alpha < \arctan(c) < \underline{\omega}/4, \tag{3.22a}$$

$$\arcsin((1+\varepsilon)c) < 2\alpha < \arcsin(2c), \tag{3.22b}$$

$$\tan(\arcsin(\sin(2\alpha) - c)) < c. \tag{3.22c}$$

Proof. Let us first choose $c \in (0, 1/2)$ such that $\arctan(c) < \underline{\omega}/4$, $\quad \frac{2}{1+c^2} > 1 + \varepsilon$. The latter inequality implies $(1 + \epsilon)c < \frac{2c}{1+c^2} = \sin(2\arctan(c))$, and by monotonicity it follows $\arcsin((1+\epsilon)c) < 2\arctan(c)$. Let us now choose α such that $\arcsin((1+\epsilon)c) < 2\alpha < 2\arctan(c)$ holds. Thus, it follows

$$\sin(2\alpha) - c < \sin(2\arctan(c)) - c = \frac{2c}{c^2+1} - c = \frac{c(1-c^2)}{c^2+1} < \min\left(c,\ \frac{2c}{c^2+1}\right).$$

With the identity $\tan(\arcsin x) = \frac{x}{\sqrt{1-x^2}}$ it follows

$$\tan(\arcsin(\sin(2\alpha) - c)) < \frac{c(1-c^2)}{c^2+1} \frac{1}{\sqrt{1 - (\frac{2c}{c^2+1})^2}} = c.$$

Hence, with c and α as chosen above, the conditions (3.22a)–(3.22c) are satisfied. □

Lemma 3.3.9. *Let δ be given by (3.21). For each $\varepsilon \in (0, 1)$ there exist $c \in (0, 1/2)$ depending on the shape of Ω and a countable set $\mathcal{B}$ of open balls $B_i = B_{r_i}(x_i)$, $i \in \mathbb{N}$, such that the following conditions hold.*

C1. The balls $B_i \in \mathcal{B}$ satisfy

$$B_i = \begin{cases} B_{c\delta/4}(x_i) & \text{if } \operatorname{dist}(x_i, \mathcal{V}) \geq \delta/4, \\ B_{c \operatorname{dist}(x_i, X_j)}(x_i) & \text{if } \operatorname{dist}(x_i, X_j) < \delta/4. \end{cases}$$

Furthermore, $\mathcal{B}$ is dichotomic with respect to Ω.

C2. $\mathcal{B}$ covers Ω, i.e., $\Omega \subset \cup_{i\in\mathbb{N}} B_{r_i}(x_i)$.

C3. $\mathcal{B}$ has finite overlap, which means that there exists $N \in \mathbb{N}$ such that

$$\#\{i \in \mathbb{N} \mid x \in B_{r_i}(x_i)\} \leq N \quad \forall x \in \Omega.$$

C4. The family of stretched balls

$$\hat{\mathcal{B}} := \{\hat{B}_i \mid \hat{B}_i = B_{(1+\varepsilon)r_i}(x_i),\ B_i \in \mathcal{B}\}$$

is dichotomic with respect to Ω and covers Ω with finite overlap, thus also satisfies C2 and C3.

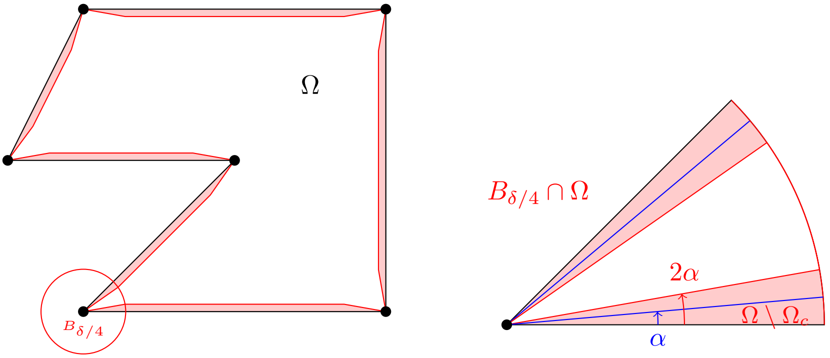


Figure 3.8: Schematic visualization of Ω_c for the construction of a dichotomic covering.

Proof. Let $\varepsilon \in (0,1)$ be given. Let us denote by $\underline{\omega}$ the minimal opening angle $\underline{\omega} := \min_{X_i \in \mathcal{V}}\{\omega_i\}$. Let $c \in (0, 1/2)$ and $\alpha \in (0, \arctan(c))$ be given by Lemma 3.3.8.

In the proof we will use local polar coordinates near vertices $X_j \in \mathcal{V}$. Let $x \in \Omega$ with $\operatorname{dist}(x, X_j) < \delta$. Then we will denote by $(\operatorname{dist}(x, X_j), \phi(x))$ the polar coordinates of x centered at $X_j \in \mathcal{V}$. We will choose $\phi(x)$ as the smaller one of the (positive) angles between the line from X_j to x and the two adjacent edges of Ω, leading to $\phi(x) \in (0, \omega_j/2)$.

We will first construct a covering of Ω by balls centered on Γ and in points with a certain distance to the boundary. To this end, let us define the set of centers by

$$\Omega_c := \left\{x \in \Omega : \operatorname{dist}(x, X_j) < \delta/4,\ \phi(x) \in [2\alpha, \omega_j/2]\right\} \\ \cup \left\{x \in \Omega : \operatorname{dist}(x, \mathcal{V}) \geq \delta/4, \operatorname{dist}(x, \Gamma) \geq \sin(2\alpha)\delta/4\right\}.$$

Finally, we define the cover

$$\mathcal{B}_u := \bigcup_{x_i \in (\Gamma \setminus \mathcal{V}) \cup \Omega_c} B(x_i), \quad B(x_i) := \begin{cases} B_{c\delta/4}(x_i) & \text{if } \operatorname{dist}(x_i, \mathcal{V}) \geq \delta/4, \\ B_{c \operatorname{dist}(x_i, X_j)}(x_i) & \text{if } \operatorname{dist}(x_i, X_j) < \delta/4. \end{cases} \tag{3.23}$$

Apart from being an uncountable set, the balls in $\mathcal{B}_u$ satisfy C1 by construction. The dichotomy follows from the dichotomy of the scaled balls, which will be shown below. In order to prove C2, we need to show that the points from $\Omega \setminus \Omega_c$ are covered. An example of this area is depicted as the shaded set of Figure 3.8.

First, let $x \in \Omega \setminus \Omega_c$ with $\operatorname{dist}(x, X_j) < \delta/4$ be given. Suppose that its azimuth angle satisfies $\phi(x) \in (0, \alpha]$. Let $\bar{x} \in \Gamma$ be such that $\operatorname{dist}(x, \bar{x}) = \operatorname{dist}(x, \Gamma)$. We find with the help of $\tan(\alpha) < c$ by (3.22a)

$$\operatorname{dist}(x, \bar{x}) = \tan(\phi(x)) \operatorname{dist}(\bar{x}, X_j) \leq \tan(\alpha) \operatorname{dist}(\bar{x}, X_j) < c \operatorname{dist}(\bar{x}, X_j).$$

Hence, it holds $x \in B_{c\,\mathrm{dist}(\bar{x},X_j)}(\bar{x})$. Analogously, we can show that points with $\mathrm{dist}(x, X_j) < \delta/4$ and $\phi(x) \in (\alpha, 2\alpha)$ are covered by $\mathcal{B}_u$.

Second, let x with $\mathrm{dist}(x, \mathcal{V}) \geq \delta/4$ be given. Again, let $\bar{x} \in \Gamma$ be such that $\mathrm{dist}(x, \Gamma) = \mathrm{dist}(x, \bar{x})$. Let $\bar{x}_c$ be on the ray from $\bar{x}$ through x such that $\bar{x}_c \in \partial\Omega_c$. Now, if $\mathrm{dist}(x, \bar{x}_c) < c\delta/4$ then x is covered by the ball in $\mathcal{B}_u$ with center $\bar{x}_c$. If $\mathrm{dist}(x, \bar{x}_c) \geq c\delta/4$ and $\mathrm{dist}(\bar{x}, \mathcal{V}) \geq \delta/4$, then by (3.22b)

$$\mathrm{dist}(x, \bar{x}) = \mathrm{dist}(\bar{x}, \bar{x}_c) - \mathrm{dist}(x, \bar{x}_c) \leq \sin(2\alpha)\delta/4 - c\delta/4 < c\delta/4,$$

and x is covered by the ball in $\mathcal{B}_u$ with center $\bar{x}$.

It remains to study the case $\mathrm{dist}(x, \bar{x}_c) \geq c\delta/4$ and $\mathrm{dist}(\bar{x}, X_j) < \delta/4$ for some j. This implies $\mathrm{dist}(x, \bar{x}) \leq (\sin(2\alpha) - c)\delta/4$, and $\sin(\phi(x)) \leq \sin(2\alpha) - c$. Using (3.22c), we find

$$\frac{\mathrm{dist}(x, \bar{x})}{\mathrm{dist}(\bar{x}, X_j)} = \tan(\phi(x)) \leq \tan(\arcsin(\sin(2\alpha) - c)) < c.$$

This implies $\mathrm{dist}(x, \bar{x}) < c\,\mathrm{dist}(\bar{x}, X_j)$, and x is covered by the ball around $\bar{x}$ with radius $c\,\mathrm{dist}(\bar{x}, X_j)$. Hence, it follows that $\mathcal{B}_u$ indeed covers Ω.

Now, let us argue that balls with stretched radius fulfill the dichotomy C4. Let $x \in \Omega_c$ with $\mathrm{dist}(x, \mathcal{V}) \geq \delta/4$. Since $\mathrm{dist}(x, \Gamma) \geq \sin(2\alpha)\delta/4 > (1+\epsilon)c\delta/4$ by (3.22b), the ball $B_{(1+\epsilon)c\delta/4}(x)$ is contained in Ω. Now take $x \in \Gamma$ with $\mathrm{dist}(x, X_j) < \delta/4$ and $\phi(x) = 0$. The ball $\hat{B} := B_{(1+\epsilon)c\,\mathrm{dist}(x,X_j)}(x)$ intersects the sector

$$\{x : \ \mathrm{dist}(x, X_j) \leq \delta/2, \ \phi(x) \in (0, \arcsin((1+\epsilon)c))\}.$$

on a half-ball with the same radius and center. Since (3.22a) and (3.22b) imply

$$\arcsin((1+\epsilon)c) < 2\alpha < \omega_j,$$

the intersection of $\hat{B}$ with Ω has the same properties. Analogously, one argues that for $\phi(x) = \omega_j$ the intersection of the stretched ball $\hat{B}$ with Ω is a half-ball. Moreover, for $\phi(x) \in [2\alpha, \omega_j/2]$, the ball $\hat{B}$ is contained in Ω. Hence, the stretched balls form a dichotomic covering, thus the dichotomies of C1 and C4 are proven.

With the help of the Besicovitch covering theorem [164, Theorem 1.3.5], which works for open balls as well, we can pass to a countable subset $\mathcal{B}$ of $\mathcal{B}_u$ which covers Ω (C2) and has finite overlap (C3). The finite overlap (C4) of $\hat{\mathcal{B}}$ can be proven as in [109, Lemma A.1] by setting $M = \mathcal{V}$. □

We will use the covering provided by Lemma 3.3.9 above to transfer between local and global regularity of functions in weighted spaces.

Lemma 3.3.10. *Let $\mathcal{B} = \{B_i \mid i \in \mathbb{N}\}$ with $B_i := B_{r_i}(x_i)$ be a covering that satisfies C1, C2, and C3 of Lemma 3.3.9, for $c \in (0, \frac{1}{2})$, δ given by (3.21). Let a multi-index $\beta \in (0,1)$ and $l \in \mathbb{N}_0$ be given.*

Define

$$\beta_i' := \begin{cases} \beta_j & \text{if } \mathrm{dist}(x_i, X_j) < \frac{\delta}{4} \text{ for some } j \in \{1, \dots, m\}, \\ 1 & \text{otherwise.} \end{cases}$$

Let $f \in B^l_\beta(\Omega, C_f, \gamma_f)$ be given. Then there are positive constants γ and $C(i)$, for $i \in \mathbb{N}$, depending only on C_f, γ_f, and $\tilde{C}$ independent of f such that

$$\begin{aligned} \| \nabla^{p+l} f \|_{L^2(\Omega \cap B_i)} &\leq \tilde{C} \, \frac{C(i)}{r_i^{\beta_i'}} \left(\frac{\gamma}{r_i} \right)^p p! \quad \forall p \in \mathbb{N}_0, \; i \in \mathbb{N}, \\ C(i) &\leq \sqrt{\frac{4}{3}} C_f, \\ \sum_{i=1}^{\infty} C(i)^2 &\leq \frac{4}{3} N C_f^2 < \infty. \end{aligned} \tag{3.24}$$

Conversely, let f in $H^{l,l}_\beta(\Omega)$ be given. Suppose that there are positive constants $\tilde{c}$, $\tilde{\gamma}$, $c(i)$, for $i \in \mathbb{N}$, such that

$$\begin{aligned} \| \nabla^{p+l} f \|_{L^2(\Omega \cap B_i)} &\leq \frac{c(i)}{r_i^{\beta_i'}} \left(\frac{\tilde{\gamma}}{r_i} \right)^p p! \quad \forall p \in \mathbb{N}_0, \; i \in \mathbb{N}, \\ \sum_{i=1}^{\infty} c(i)^2 &\leq \tilde{c}^2 < \infty. \end{aligned} \tag{3.25}$$

Then there exist positive constants C_f, which depends on $\tilde{c}$, and γ_f, which depends on $\tilde{\gamma}$, such that $f \in B^l_\beta(\Omega, C_f, \gamma_f)$.

The proof follows the lines of the proof of a similar result [106, Lemma 4.2.17] concerning regularity on sectors.

Proof. Suppose $f \in B^l_\beta(\Omega, C_f, \gamma_f)$. Let us define for $i \in \mathbb{N}$

$$C(i)^2 := \sum_{p=0}^{\infty} \frac{1}{(p!)^2 (2\gamma_f)^{2p}} \| r^{\beta+p} \nabla^{p+l} f \|^2_{L^2(\Omega \cap B_i)}.$$

By (2.9), it holds $\| r^{\beta+p} \nabla^{p+l} f \|_{L^2(\Omega \cap B_i)} \leq C_f \gamma_f^p p!$. Hence, the series in the definition of $C(i)$ is convergent, and we can estimate

$$C(i)^2 \leq C_f^2 \sum_{p=0}^{\infty} \frac{1}{4^p} = \frac{4}{3} C_f^2.$$

The finite overlap property of the covering $\mathcal{B}$ yields

$$\sum_{i=1}^{\infty} \| r^{p+\beta} \nabla^{p+l} f \|^2_{L^2(\Omega \cap B_i)} \leq N \| r^{p+\beta} \nabla^{p+l} f \|^2_{L^2(\Omega)}.$$

Consequently, the series $\sum_{i=1}^{\infty} C(i)^2$ is convergent, and we obtain as above

$$\sum_{i=1}^{\infty} C(i)^2 \leq \frac{4}{3} N C_f^2.$$

The definition of $C(i)$ also implies

$$\| r^{p+\beta} \nabla^{p+l} f \|_{L^2(\Omega \cap B_i)} \leq C(i)(2\gamma_f)^p p!. \tag{3.26}$$

Now we relate the weight function $r(x)$ to the radius r_i to prove (3.24). Let us take $B_i \in \mathcal{B}$ with center x_i and radius r_i.

Assume first that there is $X_j \in \mathcal{V}$ such that $\operatorname{dist}(x_i, X_j) < \delta/4$. By property C1 of $\mathcal{B}$, we obtain $r_i = c \operatorname{dist}(x_i, X_j)$.

Let $x \in B_i$. Then we have

$$|\min(1, \operatorname{dist}(X_j, x_i)) - \min(1, \operatorname{dist}(X_j, x))| \leq r_i = c \operatorname{dist}(x_i, X_j) = c \min(1, \operatorname{dist}(x_i, X_j)),$$

which implies

$$\min(1, \operatorname{dist}(X_j, x)) \geq (1-c) \min(1, \operatorname{dist}(X_j, x_i)) = \frac{1-c}{c} r_i \geq r_i,$$

where we used $\frac{1-c}{c} \geq 1$ for $c \in (0, 1/2)$. Now let $k \neq j$. Then we obtain

$$\delta \leq \operatorname{dist}(X_k, X_j) \leq \operatorname{dist}(X_k, x) + \operatorname{dist}(x, x_i) + \operatorname{dist}(x_i, X_j) \leq \operatorname{dist}(X_k, x) + \frac{\delta}{2},$$

and consequently it holds $\operatorname{dist}(X_k, x) \geq \delta/2$, which implies by $\delta < 1$ that

$$\min(1, \operatorname{dist}(X_k, x)) \geq \frac{\delta}{2}.$$

Define $|\beta| := \sum_{k=1}^{m} \beta_k$. By construction of β', we have $\beta'_i = \beta_j$. Using the lower bounds from above, we can estimate

$$r(x)^{p+\beta} = \prod_{k=1}^{m} \min(1, \operatorname{dist}(X_k, x))^{p+\beta_i} \geq \left(\frac{\delta}{2}\right)^{mp+|\beta|} r_i^{p+\beta'_i} \geq C_1^{-1}\, \gamma_1^{-p}\, r_i^{p+\beta'_i}, \tag{3.27}$$

where we set

$$C_1^{-1} := \left(\frac{\delta}{2}\right)^{|\beta|}, \quad \gamma_1^{-1} := \left(\frac{\delta}{2}\right)^{m}.$$

Secondly, assume that $\operatorname{dist}(x_i, X_j) \geq \delta/4$, for all $x_j \in \mathcal{V}$. Property C1 of the covering yields $r_i = c\frac{\delta}{4}$. Then we obtain as above for $j = 1, \ldots, m$

$$|\min(1, \operatorname{dist}(X_j, x_i)) - \min(1, \operatorname{dist}(X_j, x))| \leq r_i = c\frac{\delta}{4} \leq c \min(1, \operatorname{dist}(X_j, x_i)),$$

which yields

$$\min(1, \operatorname{dist}(X_j, x)) \geq (1-c) \min(1, \operatorname{dist}(X_j, x_i)) \geq (1-c)\frac{\delta}{4} = \frac{1-c}{c} r_i \geq r_i.$$

Using the definition of r and the inequality $(1-c)\frac{\delta}{4} < 1$, we find

$$r(x)^{p+\beta} \geq \left((1-c)\frac{\delta}{4}\right)^{mp+|\beta|} r_i^{p+\beta'_i} = C_2^{-1}\, \gamma_2^{-p}\, r_i^{p+\beta'_i} \tag{3.28}$$

with

$$C_2^{-1} := \left((1-c)\frac{\delta}{4}\right)^{|\beta|}, \quad \gamma_2^{-1} := \left((1-c)\frac{\delta}{4}\right)^{m}.$$

Now, inequalities (3.27) and (3.28) constitute lower bounds of $r(x)$ in terms of r_i, where $x \in B_i$. Define $\tilde{C} := \max(C_1, C_2)$, $\tilde{\gamma} := \max(\gamma_1, \gamma_2)$. Then we have for $x \in B_i$,

$$r(x)^{p+\beta} \geq \tilde{C}^{-1}\, \tilde{\gamma}^{-p}\, r_i^{p+\beta_i'}. \tag{3.29}$$

Combining (3.26) and (3.29), we find

$$C(i)(2\gamma_f)^p p! \geq \| \, r^{p+\beta} \nabla^{p+l} f \, \|_{L^2(\Omega\cap B_i)} \geq \| \, \nabla^{p+l} f \, \|_{L^2(\Omega\cap B_i)} \tilde{C}^{-1}\, \tilde{\gamma}^{-p}\, r_i^{p+\beta_i'},$$

which proves with the choice $\gamma := 2\gamma_f\tilde{\gamma}$

$$\| \, \nabla^{p+l} f \, \|_{L^2(\Omega\cap B_i)} \leq \tilde{C}\frac{C(i)}{r_i^{\beta_i'}}\left(\frac{2\gamma_f\tilde{\gamma}}{r_i}\right)^p p!.$$

Now let us assume that $f \in H_\beta^{l,l}(\Omega)$ satisfies (3.25). To prove the claim, we first derive upper bounds of $r(x)$ in terms of r_i for $x \in B_i$. For $x \in B_i$, we find

$$\min(1, \operatorname{dist}(X_j, x)) \leq r_i + \min(1, \operatorname{dist}(X_j, x_i)) \quad \forall j = 1, \dots, m.$$

If on one hand $\operatorname{dist}(X_j, x_i) < \frac{\delta}{4}$ for some j, then $r_i = c\operatorname{dist}(X_j, x_i)$, and it holds for $x \in B_i$

$$\min(1, \operatorname{dist}(X_j, x)) \leq (1 + c^{-1})r_i.$$

If $k \neq j$, we exploit that the contribution of $\operatorname{dist}(X_k, x)$ as a factor in $r(x)$ is bounded by one and, therefore,

$$r(x)^{p+\beta} \leq (1 + c^{-1})^{p+1} r_i^{p+\beta_i'}.$$

If on the other hand $\operatorname{dist}(X_j, x_i) \geq \frac{\delta}{4}$ for all j then it holds $r_i = c\frac{\delta}{4}$. Hence with $\beta_i' = 1$, we estimate

$$r(x)^{p+\beta} \leq \left(\frac{4}{c\delta}\right)^{p+1} r_i^{p+\beta_i'}.$$

Let us define

$$C_3 := \left(1 + c^{-1} + \frac{4}{c\delta}\right), \quad \gamma_3 := \left(1 + c^{-1} + \frac{4}{c\delta}\right).$$

Then for all $x \in B_i$, i arbitrary, it holds

$$r(x)^{p+\beta} \leq C_3\, \gamma_3^p\, r_i^{p+\beta_i'}. \tag{3.30}$$

Finally, we obtain

$$\begin{aligned}\| r^{p+\beta}\nabla^{p+l} f \|^2_{L^2(\Omega)} &\leq \sum_{i=1}^{\infty} \| r^{p+\beta}\nabla^{p+l} f \|^2_{L^2(\Omega\cap B_i)} \\ &\leq (C_3\gamma_3^p)^2 \sum_{i=1}^{\infty} r_i^{2(p+\beta_i')} \| \nabla^{p+l} f \|^2_{L^2(\Omega\cap B_i)} \\ &\leq (C_3\gamma_3^p)^2 \sum_{i=1}^{\infty} r_i^{2(p+\beta_i')} \left(\frac{c(i)}{r_i^{\beta_i'}} \left(\frac{\tilde{\gamma}}{r_i}\right)^p p! \right)^2 \\ &\leq (C_3\tilde{c}\,(\gamma_3\tilde{\gamma})^p\, p!)^2,\end{aligned}$$

which is the claim if we set $C_f := C_3\tilde{c}$, $\gamma_f := \gamma_3\tilde{\gamma}$. □

Note that changing from local to global estimates (and vice versa) enlarges the constants $\gamma, \tilde{\gamma}$. Thus, both directions of the result are not exact reverses of each other.

Corollary 3.3.11. *Let $\varepsilon \in (0,1)$ be given. Let $\mathcal{B}$ be the covering given by Lemma 3.3.9. Let $\hat{\mathcal{B}}$ denote the family of stretched balls $\hat{B}_i$, $i \in \mathbb{N}$. Let a multi-index $\beta \in (0,1)$ be given. Then there is a constant $C > 0$ such that*

$$r_i^{\beta_i'-1} \leq C\, r(x)^{\beta-1}$$

for all $x \in \hat{B}_i \cap \Omega$ and for all $\hat{B}_i \in \hat{\mathcal{B}}$.

Proof. This can be proven analogously to the inequality (3.30) in the proof of the previous lemma. □

3.3.3 Local Regularity

Due to the previous results, it suffices to prove local regularity results on balls and half-balls. These regularity results hold for domains in $\mathbb{R}^n$, although we will only need regularity in $\mathbb{R}^2$ for the optimal control problem. In this section, we follow the exposition of [106], which builds on techniques of [112] and [12].

Let us set $B_R := B_R(0) \subset \mathbb{R}^n$ for $R > 0$. Furthermore, we will work with half-balls $B_R^+ := B_R(0) \cap \{x : x_n > 0\}$. We set $\Gamma_R := \{x \in B_R \mid x_n = 0\}$.

Define

$$[p] := \max\{1, p\} \quad \text{for} \quad p \in \mathbb{Z}.$$

Let now p, q be integers. Following [106, 112], we will use the following notation to capture local regularity:

$$\begin{aligned}
N_{R,p}(v) &:= \frac{1}{[p]!} \sup_{R/2 \le r < R} (R-r)^{p+2} \| \nabla^{p+2} v \|_{L^2(B_r)}, \quad p \ge -2, \\
N^+_{R,p}(v) &:= \frac{1}{[p]!} \sup_{R/2 \le r < R} (R-r)^{p+2} \| \nabla^{p+2} v \|_{L^2(B^+_r)}, \quad p \ge -2, \\
N'_{R,p}(v) &:= \begin{cases} \frac{1}{p!} \sup\limits_{R/2 \le r < R} (R-r)^{p+2} \| \nabla^2 \nabla^p_x v \|_{L^2(B^+_r)}, & p \ge 0, \\ \sup\limits_{R/2 \le r < R} (R-r)^{p+2} \| \nabla^{2+p} v \|_{L^2(B^+_r)}, & p = -2, -1, \end{cases} \\
N'_{R,p,q}(v) &:= \frac{1}{[p+q]!} \sup_{R/2 \le r < R} (R-r)^{p+q+2} \| \partial^{q+2}_y \nabla^p_x v \|_{L^2(B^+_r)}, \quad p \ge 0, \ q \ge -2.
\end{aligned}$$

Here, ∇_x means the differentiation in tangential directions $x_1, \dots, x_{n-1}$. The normal derivative ∂_{x_n} is denoted by ∂_y. Hence, $N'_{R,p}(v)$ is used to control regularity of tangential derivatives, whereas $N'_{R,p,q}(v)$ controls normal derivatives. Estimates of $N_{R,p}(v)$ and $N^+_{R,p}(v)$ will be used later in order to prove the global regularity. Controlling terms as $N_{R,p}(v)$ is intimately connected with the analyticity of functions. We mention [113] and also [112, Section 5.7].

First, let us state a result that allows to estimate $N^+_{R,p}(v)$ against $N'_{R,p,q}(v)$.

Lemma 3.3.12. *Let $0 < R < R' \le 1$ be given. Let $v \in H^1(B^+_{R'})$ such that $N'_{R,p,q}(v)$ is finite for all $p \ge 0$, $q \ge -2$. Assume that there exists positive constants $C_v > 0$, γ_1, γ_2 such that*

$$N'_{R,p,q}(v) \le C_v \, \gamma_1^p \gamma_2^{q+2},$$

for all $p \ge 0$, $q \ge -2$ with $p + q \ne -2$. Then it holds with $\gamma = \sqrt{2} \max(\gamma_1, \gamma_2)$

$$N^+_{R,p}(v) \le C_v \, \gamma^{p+2},$$

for all $p \ge -1$.

Proof. Let $p \ge -1$. Then per definition it holds

$$\| \nabla^{p+2} v \|^2_{L^2(B^+_r)} = \sum_{q=-2}^{p} \| \partial^{q+2}_y \nabla^{p-q} v \|^2_{L^2(B^+_r)}.$$

By definition of $N^+_{R,p}$ and $N'_{R,p,q}$ we obtain

$$\begin{aligned}
N^+_{R,p}(v)^2 &= \frac{1}{[p]!^2} \sup_{R/2 \le r < R} (R-r)^{2(p+2)} \| \nabla^{p+2} v \|^2_{L^2(B^+_r)} \\
&\le \frac{1}{[p]!^2} \sum_{q=-2}^{p} \sup_{R/2 \le r < R} (R-r)^{2(p+2)} \| \partial^{q+2}_y \nabla^{p-q} v \|^2_{L^2(B^+_r)} \le \sum_{q=-2}^{p} (N'_{R,p-q,q})^2.
\end{aligned}$$

Let $\gamma := \sqrt{2}\max(\gamma_1, \gamma_2)$. Then if $p \geq 0$,

$$\sum_{q=-2}^{p} (N'_{R,p-q,q})^2 \leq C_v^2 \sum_{q=-2}^{p} \gamma_1^{2(p-q)} \gamma_2^{2(q+2)} \leq C_v^2 \gamma^{2(p+2)} \sum_{q=-2}^{p} \left(\frac{\gamma_1}{\gamma}\right)^{2(p-q)} \left(\frac{\gamma_2}{\gamma}\right)^{2(q+2)}$$
$$\leq C_v^2 \gamma^{2(p+2)} (p+3) 2^{-(p+2)}.$$

The function $x \mapsto (x+3)2^{-(x+2)}$ is monotonically decreasing for $x \geq 0$, it follows

$$\sum_{q=-2}^{p} (N'_{R,p-q,q})^2 \leq \frac{3}{4} C_v^2 \gamma^{2(p+2)}.$$

In the case $p = -1$, we have

$$N^+_{R,-1}(v)^2 \leq (N'_{R,0,-1})^2 + (N'_{R,1,-2})^2 \leq C_v^2 (\gamma_2^2 + \gamma_1^2) \leq C_v^2 \gamma^2.$$

□

Regularity Results for Optimal Control Problem on Half-Balls

Now, we establish regularity results for optimal control problems on half-balls. Here the control u^* acts on boundary $\Gamma_{R'}$, and it is coupled to the adjoint state with the condition $\nu u^* + q^* = 0$ on $\Gamma_{R'}$. Thus, these results cover the situation, where the control constraints are inactive. Consequently, this local optimal control problem has no control constraints.

Theorem 3.3.13 (Regularity for local optimality systems)**.** *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B^+_{R'}$. Let (u^*, y^*, q^*) solve the following system*

$$\begin{aligned} -\nabla \cdot (D\nabla y^*) + cy^* &= f \quad \textit{in } B^+_{R'}, & -\nabla \cdot (D\nabla q^*) + cq^* &= y^* - y_d \quad \textit{on } \Gamma_{R'}, && (3.31a)\\ \partial_{n_D} y^* &= u^* \quad \textit{in } B^+_{R'}, & \partial_{n_D} q^* &= 0 \quad \textit{on } \Gamma_{R'}, && (3.31b)\\ \nu u^* + q^* &= 0 \quad \textit{on } \Gamma_{R'}. &&&& (3.31c)\end{aligned}$$

Assume that there are positive constants C_d, C_f, γ_d, γ_f such that f, y_d satisfy

$$\begin{aligned} \| \nabla^p f \|_{L^2(B^+_{R'})} &\leq C_f \left(\frac{\gamma_f}{R'}\right)^p p!, \\ \| \nabla^p y_d \|_{L^2(B^+_{R'})} &\leq C_d \left(\frac{\gamma_d}{R'}\right)^p p! \quad \forall p \in \mathbb{N}_0. \end{aligned} \tag{3.32}$$

Then there exist a constant $\gamma > 0$ depending only on the constants in Assumption 2.2.1 and (3.32) and on ν such that y^*, q^* satisfy for $p \geq -1$,

$$\begin{aligned} N^+_{R,p}(y^*) &\leq C\, \gamma^{p+2}, \\ N^+_{R,p}(q^*) &\leq C\, \gamma^{p+2} \end{aligned} \tag{3.33}$$

with

$$\begin{aligned} C := R\| \nabla y^* \|_{L^2(B_R^+)} + R^2(C_f + C_c\| y^* \|_{L^2(B_R^+)}) + R^2\| y^* \|_{L^2(B_R^+)} \\ + R\| \nabla q^* \|_{L^2(B_R^+)} + R^2(C_d + C_c\| q^* \|_{L^2(B_R^+)}) + R\| q^* \|_{L^2(B_R^+)}. \end{aligned} \tag{3.34}$$

Proof. The proof will be given at the end of this section. □

The remainder of this subsection is dedicated to the proof of this theorem. Here, the following steps are important: first we need to prove that weak derivatives of y^* and q^* of arbitrary order exist. Then regularity of tangential derivatives is proven, which is followed by the proof of regularity of normal derivatives.

Let us first cite a result from [106]. In order to state this result, let us define for $p \in \mathbb{N}_0$

$$H_{R,p}(v) := \frac{1}{[p-1]!} \sup_{R/2 \le r < R} (R-r)^{p+1} \left[\| \nabla_x^p v \|_{L^2(B_r^+)} + \frac{R-r}{[p]} \| \nabla_x^p \nabla v \|_{L^2(B_r^+)} \right],$$

$$M'_{R,p}(v) := \frac{1}{p!} \sup_{R/2 \le r < R} (R-r)^{p+2} \| \nabla_x^p v \|_{L^2(B_r^+)}.$$

Lemma 3.3.14. *Let $R \in (0,1]$. Let the coefficient function D of the differential operator A fulfill the conditions of Assumption 2.2.1 on $\Omega = B_R^+$. Let $p \in \mathbb{N}_0$ be such that $f \in H^p(B_R^+)$ and $G \in H^{p+1}(B_R^+)$. Then there exists a constant $C_B > 0$ depending solely on the properties of D and the space dimension, but not on p, F, and G such that any solution $y \in H^1(B_R^+)$ of the Neumann problem*

$$-\nabla \cdot (D\nabla y) = f \quad \text{in } B_R^+, \quad \partial_{n_D} y = G \quad \text{on } \Gamma_R$$

satisfies

$$N'_{R,p}(y) \le C_B \left[M'_{R,p}(f) + H_{R,p}(G) + S_{R,p}(y) + N'_{R,p-1}(y) + N'_{R,p-2}(y) \right] \tag{3.35}$$

with

$$\begin{aligned} S_{R,p}(y) = \sum_{q=1}^{p+1} \binom{p+1}{q} \left[\left(\frac{R}{2}\right)^q \| \nabla^q D \|_{L^\infty(B_R^+)} + \left(\frac{R}{2}\right)^{q-1} q \| \nabla^{q-1} D \|_{L^\infty(B_R^+)} \right] \\ \cdot \frac{[p-q]!}{p!} N'_{R,p-q}(y). \end{aligned} \tag{3.36}$$

For $p = 0$, we have the sharper bound

$$N'_{R,0}(y) \le C_B \left[R^2 \| f \|_{L^2(B_R^+)} + R\| G \|_{L^2(B_R^+)} + R^2 \| \nabla G \|_{L^2(B_R^+)} + R\| \nabla y \|_{L^2(B_R^+)} \right].$$
(3.37)

Proof. The proof follows the lines of the proof of [106, Lemma 5.5.23]. The bound (3.37) is from [106, Lemma 5.5.26]. □

We formulate two further lemmas regarding the relation between $H_{R,p}, M'_{R,p}, N'_{R,p}$.

Lemma 3.3.15. *Let $R \in (0,1]$. Let v be such that $M'_{R,p}(v)$ and $N'_{R,p}(v)$ are well defined for all $p \in \mathbb{N}_0$. Then it holds*

$$M'_{R,p}(v) \leq \frac{1}{[p][p-1]} N'_{R,p-2}(v) \quad \forall p \in \mathbb{N}_0.$$

Proof. We use the fact that $(R-r)^j \leq 1$ for $j \geq 0$ and $r \in [R/2, R]$. Together with the definition of $M'_{R,p}(v)$ we estimate for $p \geq 2$

$$\begin{aligned} M'_{R,p}(v) &= \frac{1}{p!} \sup_{R/2 \leq r < R} (R-r)^{p+2} \| \nabla_x^p v \|_{L^2(B_r^+)} \\ &\leq \frac{1}{p!} \sup_{R/2 \leq r < R} (R-r)^{p} \| \nabla^2 \nabla_x^{p-2} v \|_{L^2(B_r^+)} = \frac{1}{p(p-1)} N'_{R,p-2}(v). \end{aligned}$$

For $p = 0, 1$, the same estimate without the term $\frac{1}{p(p-1)}$ is valid. □

Lemma 3.3.16. *Let $R \in (0,1]$. Let v be such that $H_{R,p}(v)$ and $N'_{R,p}(v)$ are well defined for all $p \in \mathbb{N}_0$. Then it holds*

$$H_{R,p}(v) \leq \frac{1}{[p-1]} N'_{R,p-2}(v) + N'_{R,p-1}(v) \quad \forall p \in \mathbb{N}_0.$$

Proof. The estimate can be established analogously to the proof of Lemma 3.3.15. □

Finally, we need higher regularity of the optimal variables on half balls. A part of the proof is based on the following supporting lemma

Lemma 3.3.17. *Let $p \in \mathbb{N}_0$. Then it holds*

$$\sum_{j=1}^{p+1} \frac{p+1}{[p-j+1]} \frac{1}{2^j} \leq 3. \tag{3.38}$$

Proof. First, we see

$$\frac{p+1}{p-j+1} = 1 + \frac{j}{p-j+1} \leq 1 + j.$$

The addend for $j = p+1$ in (3.38) is equal to $(p+1)2^{-(p+1)}$, which is smaller than $(p+2)2^{-(p+1)}$. Consequently, it holds

$$\sum_{j=1}^{p+1} \frac{p+1}{[p-j+1]} \frac{1}{2^j} \leq \sum_{j=1}^{p+1} (1+j) \frac{1}{2^j}.$$

Using the power series

$$(1-x)^{-2} = \sum_{j=0}^{\infty}(1+j)x^j, \ |x|<1,$$

we can estimate

$$\sum_{j=1}^{p+1} \frac{p+1}{[p-j+1]} \frac{1}{2^j} \leq \sum_{j=1}^{\infty}(1+j)\frac{1}{2^j} = \left(1-\frac{1}{2}\right)^{-2} - 1 = 3.$$

□

Lemma 3.3.18. *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B_{R'}^+$. Let $(u^*, y^*, q^*) \in L^2(\Gamma_{R'}) \times H^1(B_{R'}^+) \times H^1(B_{R'}^+)$ be a solution of the local control problem (3.31) on the half ball $B_{R'}^+$. Then it holds*

$$u^* \in H^{p-1/2}(\Gamma_R), \quad y^*,\ q^* \in H^p(B_R^+) \quad \forall p \geq 2.$$

The proof basically exploits the optimality system (3.31) for a bootstrapping argument

$$u^* \in H^{1/2} \quad \Rightarrow \quad y^* \in H^2 \quad \Rightarrow \quad q^* \in H^4 \quad \Rightarrow \quad u^* \in H^{3.5} \quad \Rightarrow \quad \dots \tag{3.39}$$

on half balls with decreasing radii.

Proof. Let $0 < R < R'$. First, we note that for $f \in H^m(B_{R'}^+), g \in H^{m-1/2}(\Gamma_{R'})$ a solution v to

$$\begin{aligned} Av &= f \quad \text{in } B_R^+, \\ \partial_{n_D} v &= g \quad \text{on } \Gamma_R \end{aligned}$$

lies in $H^{m+2}(B_R^+)$. This can be shown by an induction on m (see the proof [63, Theorem 6.5]) triggered by the method of difference quotients developed by Nirenberg (see the proof of [63, Theorem 6.4] or [72, Theorem 2.2.2.5]).

Define $r(p) := R + \frac{R'-R}{p}$. As $q^* \in H^1(B_{R'}^+)$, the optimality system yields $u^* \in H^{1/2}(\Gamma_{r(1)})$. Hence, $y^* \in H^2(B_{r(2)}^+)$. As the adjoint equation has smooth boundary data, it follows $q^* \in H^4(B_{r(4)}^+)$. Applying the trace operator yields $u^* \in H^{3.5}(\Gamma_{r(4)})$. The assertion then follows by induction and the fact that $r(p) \geq R$. □

We remark that the result solely proves the regularity. The proof does not offer a way to control norms of derivatives, which will be done in the next Lemma 3.3.19 below.

Lemma 3.3.19. *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B_{R'}^+$. Let (u^*, y^*, q^*) be a solution of the local control problem (3.31) with f and y_d satisfying (3.32).*

Define

$$C_t := R\| \nabla y^* \|_{L^2(B_R^+)} + R^2(C_f + C_c\| y^* \|_{L^2(B_R^+)}) + R\| q^* \|_{L^2(B_R^+)} + R^2\| \nabla q^* \|_{L^2(B_R^+)} \\ + R\| \nabla q^* \|_{L^2(B_R^+)} + R^2(C_d + C_c\| q^* \|_{L^2(B_R^+)}) + R^2\| y^* \|_{L^2(B_R^+)}. \tag{3.40}$$

There there is a constant $\gamma_t > 0$ depending only on the constants in Assumption 2.2.1 and (3.32) and on ν such that for all $p \geq -1$

$$N'_{R,p}(y^*) \leq C_t\gamma_t^{p+2}, \tag{3.41a}$$

$$N'_{R,p}(q^*) \leq C_t\gamma_t^{p+2}. \tag{3.41b}$$

Proof. Let us choose a γ_t larger than $\max\{1, \gamma_f/2, \gamma_d/2, \gamma_c, \gamma_D\}$ such that

$$C_B\left(2\gamma_t^{-2} + (2C_c + 1)\,\gamma_t^{-2} + 6C_D(\gamma_D + 1)\gamma_t^{-1} + (1 + \nu^{-1})(\gamma_t^{-2} + \gamma_t^{-1})\right) \leq 1 \tag{3.42}$$

with the constant C_B from Lemma 3.3.14. This constant only depends on the data of the problem.

Let us prove (3.41) for $p = -1$. From the definition of $N'_{R,p}$ we obtain

$$N'_{R,-1}(y^*) = \sup_{R/2\leq r<R} (R - r)\| \nabla y^* \|_{L^2(B_r^+)} \leq \frac{R}{2}\| \nabla y^* \|_{L^2(B_R^+)} \leq C_t \leq C_t\gamma_t.$$

Similarly, we can prove $N'_{R,-1}(q^*) \leq C_t\gamma_t$.

Let now $p = 0$. Using the sharp bound (3.37) of Lemma 3.3.14 we find

$$N'_{R,0}(y^*) \leq C_B\Big(R^2\| f - cy^* \|_{L^2(B_R^+)} + R\nu^{-1}\| q^* \|_{L^2(B_R^+)} \\ + R^2\nu^{-1}\| \nabla q^* \|_{L^2(B_R^+)} + R\| \nabla y^* \|_{L^2(B_R^+)}\Big). \tag{3.43}$$

With the help of Assumption 2.2.1 and (3.32), we estimate

$$\| f - cy^* \|_{L^2(B_R^+)} \leq C_f + C_c\| y^* \|_{L^2(B_R^+)}.$$

Inserting these estimates in (3.43) yields by the definition (3.40) of C_t and $\gamma_t \geq 1$

$$N'_{R,0}(y^*) \leq C_BC_t(1 + \nu^{-1}) = C_t\frac{C_B(1 + \nu^{-1})}{\gamma_t}\gamma_t \leq C_t\gamma_t \leq C_t\gamma_t^2.$$

Using again (3.37), we obtain analogously

$$N'_{R,0}(q^*) \leq C_B\left(R^2\| y^* \|_{L^2(B_R^+)} + R^2C_d + R^2C_c\| q^* \|_{L^2(B_R^+)} + R\| \nabla q^* \|_{L^2(B_R^+)}\right) \\ = C_BC_t \leq C_t\frac{C_B}{\gamma_t}\gamma_t \leq C_t\gamma_t^2.$$

We finish by induction. Let $p \geq 1$ be given. Suppose (3.41) holds for all p' with $-1 \leq p' < p$. Because of the regularity results of Lemma 3.3.18 we can apply Lemma 3.3.14. In combination with the estimates of Lemmas 3.3.15 and 3.3.16, we can derive the bound

$$\begin{aligned} N'_{R,p}(y^*) &\leq C_B\Big(M'_{R,p}(f - cy^*) + S_{R,p}(y^*) + H_{R,p}\Big(\frac{-q^*}{\nu}\Big) + N'_{R,p-1}(y) + N'_{R,p-2}(y)\Big) \\ &\leq C_B\ \Big(M'_{R,p}(f - cy^*) + S_{R,p}(y^*) + \frac{1}{\nu[p-1]}N'_{R,p-2}(q^*) + \frac{1}{\nu}N'_{R,p-1}(q^*) \\ &\qquad\qquad + N'_{R,p-1}(y) + N'_{R,p-2}(y)\Big) \\ &\leq C_B\ \Big(M'_{R,p}(f) + M'_{R,p}(cy^*) + S_{R,p}(y^*) + (1 + \nu^{-1})C_t(\gamma_t^p + \gamma_t^{p+1})\Big)\,. \end{aligned}$$

(3.44)

It remains to estimate $M'_{R,p}(f)$, $M'_{R,p}(cy^*)$ and $S_{R,p}(y^*)$. By assumption (3.32), we have

$$\begin{aligned} M'_{R,p}(f) &\leq \frac{1}{p!}\sup_{R/2\leq r<R}(R-r)^{p+2}\|\,\nabla^p f\,\|_{L^2(B_r^+)} \leq \left(\frac{R}{2}\right)^{p+2} C_f\left(\frac{\gamma_f}{R}\right)^p \\ &\leq R^2 C_f\gamma_t^p\left(\frac{\gamma_f}{2\gamma_t}\right)^p \leq C_t\gamma_t^p. \end{aligned} \tag{3.45}$$

By [106, Lemma 5.5.13], we have the following upper bound of $M'_{R,p}(cy^*)$

$$\begin{aligned} M'_{R,p}(cy^*) \leq C_c\sum_{q=0}^{p-1}\left(\gamma_c\frac{R}{2}\right)^q\left(\frac{R}{2}\right)^2\frac{[p-q-2]}{(p-q)!}N'_{R,p-q-2}(y^*) \\ + \left(\gamma_c\frac{R}{2}\right)^p C_c\left(\frac{R}{2}\right)^2 N'_{R,-2}(y^*). \end{aligned}$$

Let us note that $N'_{R,-2}(y^*)$ does not satisfy the induction hypothesis, rather it holds $N'_{R,-2}(y^*) \leq \|\,y^*\,\|_{L^2(B_R^+)}$. We continue the estimation procedure using the inequality $C_c\left(\frac{R}{2}\right)^2 N'_{R,-2}(y^*) \leq C_t$.

$$\begin{aligned} M'_{R,p}(cy^*) &\leq C_c\sum_{q=0}^{p-1}\left(\gamma_c\frac{R}{2}\right)^q\left(\frac{R}{2}\right)^2 C_t\gamma_t^{p-q} + C_t\left(\gamma_c\frac{R}{2}\right)^p \\ &\leq C_c C_t\gamma_t^p\sum_{q=0}^{p-1}\frac{1}{2^q} + C_t\gamma_t^p \leq (2C_c+1)\,C_t\gamma_t^p. \end{aligned} \tag{3.46}$$

The next step is estimating $S_{R,p}(y^*)$, for its definition we refer to (3.36). Here, we obtain by Assumption 2.2.1

$$\frac{R}{2}\|\,\nabla^q D\,\|_{L^\infty(B_R^+)} + q\|\,\nabla^{q-1}D\,\|_{L^\infty(B_R^+)} \leq \frac{R}{2}C_D\gamma_D^q q! + C_D\gamma_D^{q-1}q! \leq C_D(\gamma_D+1)\gamma_D^{q-1}q!.$$

Hence, we get the following bound

$$
\begin{aligned}
S_{R,p}(y^*) &\leq \sum_{q=1}^{p+1} \binom{p+1}{q} \left(\frac{R}{2}\right)^{q-1} C_D(\gamma_D+1)\gamma_D^{q-1} q! \frac{[p-q]!}{p!} \underbrace{N'_{R,p-q}(y^*)}_{\leq C_t \gamma^{p-q+2}} \\
&= C_D C_t 2(\gamma_D+1)\gamma_t^{p+1} \sum_{q=1}^{p+1} \frac{(p+1)[p-q]!}{(p-q+1)!} \frac{1}{2^q} \left(\frac{R\gamma_D}{\gamma_t}\right)^{q-1} \\
&\leq C_D C_t 2(\gamma_D+1)\gamma_t^{p+1} \sum_{q=1}^{p+1} \frac{p+1}{[p-q+1]} \frac{1}{2^q}.
\end{aligned}
$$

Applying Lemma 3.3.17 yields

$$
S_{R,p}(y^*) \leq 6C_D(\gamma_D+1)C_t\gamma_t^{p+1}. \tag{3.47}
$$

Inserting the estimates (3.45), (3.46), (3.47) of $M'_{R,p}(f)$, $M'_{R,p}(cy^*)$, and $S_{R,p}(y^*)$ in (3.44) results in

$$
\begin{aligned}
&N'_{R,p}(y^*) \\
&\quad\leq C_B\Big(C_t\gamma_t^p + (2C_c+1)\,C_t\gamma_t^p + 6C_D(\gamma_D+1)C_t\gamma_t^{p+1} + (1+\nu^{-1})C_t(\gamma_t^p+\gamma_t^{p+1})\Big) \\
&\quad= C_t\gamma_t^{p+2}C_B\left(\gamma_t^{-2} + (2C_c+1)\,\gamma_t^{-2} + 6C_D(\gamma_D+1)\gamma_t^{-1} + (1+\nu^{-1})(\gamma_t^{-2}+\gamma_t^{-1})\right).
\end{aligned}
$$

By (3.42), we find

$$
N'_{R,p}(y^*) \leq C_t\gamma_t^{p+2},
$$

which finishes the prove of the estimate of $N'_{R,p}(y^*)$.

Let us briefly show the relevant arguments for estimating $N'_{R,p}(q^*)$. First, by Lemma 3.3.14, we obtain for $p \geq 1$

$$
\begin{aligned}
N'_{R,p}(q^*) \leq C_B\Big(&M'_{R,p}(y^*) - M'_{R,p}(y_d) - M'_{R,p}(cq^*) \\
&+ S_{R,p}(q^*) + N'_{R,p-1}(q^*) + N'_{R,p-2}(q^*)\Big).
\end{aligned}
$$

By Lemma 3.3.15, it holds $M'_{R,p}(y^*) \leq N'_{R,p-2}(y^*)$. Analogously to (3.45), (3.46), and (3.47) we can prove

$$
\begin{aligned}
M'_{R,p}(y_d) &\leq C_t\gamma_t^p, \\
M'_{R,p}(cq^*) &\leq (2C_c+1)\,C_t\gamma_t^p, \\
S_{R,p}(q^*) &\leq 6C_D(\gamma_D+1)C_t\gamma_t^{p+1}.
\end{aligned}
$$

Since γ_t satisfies (3.42), we finish the proof of (3.41) by estimating

$$
\begin{aligned}
&N'_{R,p}(q^*) \\
&\quad\leq C_B\Big(C_t\gamma_t^p + C_t\gamma_t^p + (2C_c+1)\,C_t\gamma_t^p + 6C_D(\gamma_D+1)C_t\gamma_t^{p+1} + C_t\gamma_t^{p+1} + C_t\gamma_t^p\Big) \\
&\quad\leq C_t\gamma_t^{p+2}C_B\left(2\gamma_t^{-2} + (2C_c+1)\,\gamma_t^{-2} + 6C_D(\gamma_D+1)\gamma_t^{-1} + \gamma_t^{-1} + \gamma_t^{-2}\right) \leq C_t\gamma_t^{p+2}.
\end{aligned}
$$

□

Lemma 3.3.20. *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B^+_{R'}$. Let (u^*, y^*, q^*) be a solution of the local control problem (3.31) with f and y_d satisfying (3.32). Then there exist constants $\gamma_1, \gamma_2 > 0$ depending only on the constants in Assumption 2.2.1 and (3.32) and on ν such that y^*, q^* satisfy for $p \in \mathbb{N}_0$, $q \geq -2$, and $p + q \neq -2$*

$$N'_{R,p,q}(y^*) \leq C\ \gamma_1^p\ \gamma_2^{q+2}, \tag{3.48a}$$

$$N'_{R,p,q}(q^*) \leq C\ \gamma_1^p\ \gamma_2^{q+2}, \tag{3.48b}$$

with C given by (3.34).

Proof. The proof essentially relies on the proof of [106, Proposition 5.5.2] with Lemma 3.3.19 as an induction start. Please note that the technique of the proof of [106, Proposition 5.5.2] is independent of the boundary conditions of the local problem. Due to the coupling between y^* and q^* in the right-hand side of the adjoint equation the induction proof for the estimates of $N'_{R,p,q}(y^*)$ and $N'_{R,p,q}(q^*)$ has to be performed simultaneously.

The only modification of the proof concerns the estimate of the right-hand side y^* of the adjoint equation. There, one needs to estimate

$$\frac{1}{[p+q]!} \sup_{R/2 \leq r < R} (R-r)^{p+q+2} \|\partial_y^q \nabla_x^p (\tilde{d} y^*)\|_{L^2(B_r^+)} \tag{3.49}$$

for $p \geq 0$ and $q \geq 0$, where $\tilde{d} := D_{nn}^{-1}$, and D_{nn} is the (n,n)-entry of the coefficient matrix D. According to [106, Lemma 5.5.19], there are γ'_D and C'_D depending only on the constants in Assumption 2.2.1, such that $\|\nabla^p \tilde{d}\|_{L^\infty(B_R^+)} \leq C'_D \gamma_D'^p p!$ for all $p \geq 0$. By [106, Lemma 5.5.18], one can now estimate the term (3.49) analogously to the estimates of mixed derivatives of $\tilde{c}y^* := (\tilde{dc}) \cdot y^*$ and $\tilde{c}q^* := (\tilde{dc}) \cdot q^*$. The rest of the proof of [106, Proposition 5.5.2] can be transferred line-by-line to our situation, and the induction can be concluded. □

Proof of Theorem 3.3.13. The claim of the theorem follows from Lemma 3.3.20 and Lemma 3.3.12. □

Regularity Results for Coupled State-Adjoint Systems

In this part, we briefly formulate regularity on balls and half-balls for the coupled system of state and adjoint equation. First, we consider the situation that the control constraint is active on the normal boundary of the half-ball.

Theorem 3.3.21 (Coupled state-adjoint system - Neumann case)**.** *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B^+_{R'}$. Let (y^*, q^*) solve (3.31a) with the boundary conditions*

$$\partial_{n_D} y^* = u_a, \quad \partial_{n_D} q^* = 0 \quad \text{on } \Gamma_{R'}. \tag{3.50}$$

Assume that there are positive constants C_d, C_f, C_g, γ_d, γ_f, γ_g such that f, y_d, u_a satisfy (3.32) *as well as*

$$\| u_a \|_{L^2(B^+_{R'})} \leq C_g, \quad \| \nabla^{p+1} u_a \|_{L^2(B^+_{R'})} \leq C_g \left(\frac{\gamma_g}{R'}\right)^p p! \quad \forall p \in \mathbb{N}_0. \tag{3.51}$$

Then there exist a constant $\gamma > 0$ depending only on the constants in Assumption 2.2.1*,* (3.32)*, and* (3.51) *such that y^*, q^* satisfy* (3.33) *with*

$$\begin{aligned} C := {} & R\| \nabla y^* \|_{L^2(B^+_R)} + R^2(C_f + C_c\| y^* \|_{L^2(B^+_R)}) + R^2\| y^* \|_{L^2(B^+_R)} \\ & + R\| \nabla q^* \|_{L^2(B^+_R)} + R^2(C_d + C_c\| q^* \|_{L^2(B^+_R)}) + R\| u_a \|_{L^2(B^+_{R'})} + R^2\, C_g. \end{aligned} \tag{3.52}$$

Proof. We briefly sketch the modifications of the proof compared to the proof of Theorem 3.3.13 in the previous section. The regularity result of Lemma 3.3.18 remains valid due to the regularity of the Neumann datum u_a. Some estimates in Lemma 3.3.19 have to be modified to take the boundary data into account: First, from the assumptions it follows $\| \nabla u_a \|_{L^2(B^+_{R'})} \leq C_g$, which implies with Lemma 3.3.14, that $N'_{R,0}(y^*) \leq C_B(R^2(C_f + C_c\| y^* \|_{L^2(B^+_R)}) + R\| u_a \|_{L^2(B^+_{R'})} + R^2\, C_g + R\| \nabla y^* \|_{L^2(B^+_R)})$ holds, compare (3.37) and (3.43). Second, one can derive the bound $H_{R,p}(u_a) \leq R^2 C_g(\gamma_g^{p-1} + \gamma_g^p)$ for all $p \geq 1$. Using $\gamma_t \geq \gamma_g/2$, this term can be compensated in the induction argument of Lemma 3.3.19. Thus, estimate (3.40) holds with the modified constant

$$\begin{aligned} C_t := {} & R\| \nabla y^* \|_{L^2(B^+_R)} + R^2(C_f + C_c\| y^* \|_{L^2(B^+_R)}) + R\| u_a \|_{L^2(B^+_{R'})} + R^2\, C_g \\ & + R\| \nabla q^* \|_{L^2(B^+_R)} + R^2(C_d + C_c\| q^* \|_{L^2(B^+_R)}) + R^2\| y^* \|_{L^2(B^+_R)}. \end{aligned}$$

The rest of the proof is completely analogous to the proof of Theorem 3.3.13. □

The second regularity result concerns half-balls with Dirichlet boundary.

Theorem 3.3.22 (Coupled state-adjoint system - Dirichlet case). *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption* 2.2.1 *on $\Omega = B^+_{R'}$. Let (y^*, q^*) solve* (3.31a) *with the boundary conditions*

$$y^* = 0, \quad q^* = 0 \quad \textit{on } \Gamma_{R'}. \tag{3.53}$$

Assume that there are positive constants C_d, C_f, γ_d, γ_f such that f, y_d satisfy (3.32)*. Then there exist a constant $\gamma > 0$ depending only on the constants in Assumption* 2.2.1 *and* (3.32) *such that y^*, q^* satisfy* (3.33) *with*

$$\begin{aligned} C := {} & R\| \nabla y^* \|_{L^2(B^+_R)} + R^2(C_f + C_c\| y^* \|_{L^2(B^+_R)}) + R^2\| y^* \|_{L^2(B^+_R)} \\ & + R\| \nabla q^* \|_{L^2(B^+_R)} + R^2(C_d + C_c\| q^* \|_{L^2(B^+_R)}). \end{aligned} \tag{3.54}$$

Proof. The proof is completely analogous to the proofs of Theorems 3.3.13 and 3.3.21. One only has to exchange Lemma 3.3.14 by the analogous result [106, Lemma 5.5.15] for Dirichlet problems. The analogue of Lemma 3.3.18 only involves the state and adjoint variable and can be proved similarly. □

Finally, let us state the result for the case of an interior ball.

Theorem 3.3.23 (Coupled state-adjoint system - interior case)**.** *Let $0 < R < R' \leq 1$ be given. Let the differential operator A fulfill Assumption 2.2.1 on $\Omega = B_{R'}$. Let (y^*, q^*) solve (3.31a) on $B_{R'}$. Assume that there are positive constants C_d, C_f, γ_d, γ_f such that f, y_d satisfy (3.32) with $L^2(B^+_{R'})$-norms replaced with $L^2(B_{R'})$-norms. Then there exist a constant $\gamma > 0$ depending only on the constants in Assumption 2.2.1 and (3.32) such that y^*, q^* satisfy*

$$N_{R,p}(y^*) \leq C\,\gamma^{p+2}, \quad N_{R,p}(q^*) \leq C\,\gamma^{p+2}, \tag{3.55}$$

with C given by (3.54).

Proof. Here, the only change consists of replacing Lemma 3.3.14 by [106, Lemma 5.5.12] and Lemma 3.3.18 by [63, Theorem 6.3] in order to account for interior regularity. □

Note that the bound (3.55) holds for $N_{R,p}$ replaced by $N_{R',p}$ because [63, Theorem 6.3] yields regularity on balls of the same radius, in contrast to Lemma 3.3.18.

3.3.4 Global Regularity

Equipped with the Besicovitch covering of Lemma 3.3.9, the local estimates on balls and half balls of the previous section, and the correlation of local and global estimates (Lemma 3.3.10), it is now possible to prove the main result.

Theorem 3.3.24. *Let Ω be a polygonal domain. Let Assumption 2.2.1 be satisfied. Suppose that (u^*, y^*, q^*) is the solution to* (**P**) *and that u^* has finitely many switching points as in Assumption 3.3.1.*

Then there exist multi-indices $\beta, \tilde{\beta} \in (0,1)$ such that for data satisfying Assumption 3.3.4 with the weight function r of (2.1.5) *there are constants $C_*, \gamma_*, C_u, \gamma_u > 0$ such that*

$$u^* \in B^{3/2}_{\tilde{\beta}}(\Gamma_{\mathcal{N}}, C_u, \gamma_u), \quad y^*, q^* \in B^2_\beta(\Omega, C_*, \gamma_*)$$

holds. The constants $C_*, \gamma_*, C_u, \gamma_u$ depend only on the data A, f, y_d, u_a, u_b and the domain Ω.

The proof will be given at the end of the section. We first start with proving the regularity of y^* and q^* in weighted H^2-spaces.

Theorem 3.3.25. *Let Ω be a polygonal domain. Let Assumption 2.2.1 be satisfied. Suppose that (u^*, y^*, q^*) is the solution to* (**P**) *and that u^* has finitely many switching points as in Assumption 3.3.1. There exists a multi-index $\beta \in (0,1)$ such that for data satisfying Assumption 3.3.4 we have $y^*, q^* \in H^{2,2}_\beta(\Omega)$.*

Proof. Recall that we denote the edges of Ω by Γ_j, $j = 1, \dots, m$. Let us renumber the vertices in $\mathcal{V}$ if necessary to obtain $\overline{\Gamma}_j \cap \overline{\Gamma}_{j+1} = X_j$ for all $j = 1, \dots, m$ with $\Gamma_{m+1} := \Gamma_1$.

We split the domain into different (overlapping) parts. The first part is a neighborhood of the corners, consisting of sectors, defined by $\Omega_0 := \cup_{X_j \in \mathcal{V}} S_j$ with $S_j := \Omega \cap B_{\delta/2}(X_j)$. Here, we need the assumption that the number of switching points is finite. The second part is a covering of the boundary away from the corners. Take $\varepsilon \in (0,1)$. Let $\mathcal{B}$ be given by Lemma 3.3.9. Define the index set I by $I := \{i \in \mathbb{N} : B_i \cap \Omega \text{ is a half-ball}, \operatorname{dist}(x_i, \mathcal{V}) > \delta/4\}$, and set $\Omega_1 := \cup_{i \in I} B_i$. Then $\Omega_0 \cup \Omega_1$ covers a neighborhood of the boundary. Moreover, we can choose $\Omega_2 \subset \Omega$ such that Ω_2 has positive distance to the boundary and such that it holds $\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2$.

Let us first prove the regularity of the adjoint state q^*.

1. Depending on the type of boundary condition on $\partial S_j \cap \Gamma$, i.e., homogeneous Dirichlet, Neumann, or mixed boundary conditions, we apply [106, Proposition 5.3.2, 5.4.4, or 5.4.7] and obtain the existence of $\beta_j \in (0,1)$ such that

$$\| \operatorname{dist}(x, X_j)^{\beta_j} \nabla^2 q^* \|_{L^2(S_j)} \le C(\| y - y_d \|_{L^2(\Omega)} + C_c \| q^* \|_{L^2(\Omega)}) \quad \forall X_j \in \mathcal{V}.$$

   Here it is important to note that the value of β_j only depends on the opening angle ω_j at the corner X_j and on the coefficient $D(X_j)$.

2. The solution q^* is H^2-regular on the half-balls $B_i \cap \Omega$, $i \in I$, which follows from [106, Proposition 5.5.7 and 5.5.9]. Here, we obtain

$$\| \nabla^2 q^* \|_{L^2(\Omega \cap B_i)} \le C(\| y^* - y_d \|_{L^2(\Omega \cap \hat{B}_i)} + C_c \| q^* \|_{L^2(\Omega \cap \hat{B}_i)}).$$

   Since the balls B_i all have the same radius, the estimate is uniform in $i \in I$. Due to the finite-overlapping property, this yields $\| \nabla^2 q^* \|_{L^2(\Omega_1)} \le C(\| y^* - y_d \|_{L^2(\Omega)} + C_c \| q^* \|_{L^2(\Omega)})$.

3. The closed set $\overline{\Omega}_2$ has positive distance to the boundary, and by standard interior regularity, e.g., [63, Theorem 6.2], it holds

$$\| \nabla^2 q^* \|_{L^2(\Omega_2)} \le C(\| y - y_d \|_{L^2(\Omega)} + C_c \| q^* \|_{L^2(\Omega)}).$$

Since $\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2$, we conclude by the previous bounds $\| q^* \|_{H^{2,2}_\beta(\Omega)} < \infty$.

The line of reasoning to prove regularity of y^* is very similar, and we only point out the necessary modifications. Assume that $\Gamma_j \subset \Gamma_{\mathcal{N}}$. By the construction of $\mathcal{V}$, the optimal control satisfies on Γ_j one of the conditions $u^*|_{\Gamma_j} = u_a|_{\Gamma_j}$, $u^*|_{\Gamma_j} = -\nu^{-1} q^*|_{\Gamma_j}$, or $u^*|_{\Gamma_j} = u_b|_{\Gamma_j}$. In either case, $u^*|_{\Gamma_j} \in H^{1/2}(\Gamma_j)$. Furthermore, due to the regularity of q^* and the assumptions on u_a, u_b, the control $u^*|_{\Gamma_j}$ is the restriction of a function $g_j \in H^{2,2}_\beta(\Omega)$ to Γ_j, where g_j is one of q^*, u_a, u_b.

Now assume that $\partial S_j \cap \Gamma \subset \Gamma_{\mathcal{N}}$. Then by [106, Proposition 5.4.4] it follows

$$\begin{aligned} &\| \operatorname{dist}(x, X_j)^{\beta_j} \nabla^2 y^* \|_{L^2(S_j)} \le C(\| f \|_{L^2(\Omega)} + C_c \| y^* \|_{L^2(\Omega)} \\ &+ \| g_j \|_{L^2(S_j)} + \| \operatorname{dist}(x, X_j)^{\beta_j} g_j \|_{L^2(S_j)} + \| g_{j+1} \|_{L^2(S_j)} + \| \operatorname{dist}(x, X_j)^{\beta_j} g_{j+1} \|_{L^2(S_j)}). \end{aligned}$$

Analogous results holds if there are mixed boundary condition on S_j. For balls B_i with $B_i \cap \Gamma \subset \Gamma_{\mathcal{N}}$, we have by H^2-regularity

$$\| \nabla^2 y^* \|_{L^2(\Omega\cap B_i)} \leq C(\| f \|_{L^2(\Omega\cap\hat{B}_i)} + C_c \| y^* \|_{L^2(\Omega\cap\hat{B}_i)} + \| u^* \|_{H^{1/2}(\Gamma\cap\hat{B}_i)}).$$

The proof can now be completed as above, and this proves $y^* \in H^{2,2}_\beta(\Omega)$. □

Remark 3.3.26. *Let us briefly discuss the operator $A = -\Delta + id$. If X_j is a vertex of the domain, then the value of β_j satisfies $\beta_i \in [0,1) \cap (1 - \pi/\omega_i, 1)$ as stated in [106, Proposition 5.2.1]. If X_j is a switching point of the optimal control but not a vertex of the domain, then due to the H^2-regularity on half-balls, we can choose $\beta_j > 0$ arbitrary.*

Remark 3.3.27. *Let us remark that $u_a, u_b \in B^1_\beta(\Omega, C_g, \gamma_g)$ is sufficient to prove $y^*, q^* \in H^{2,2}_\beta(\Omega)$ in Theorem 3.3.25 and $y^*, q^* \in B^2_\beta(\Omega, C_*, \gamma_*)$ in Theorem 3.3.24.*

Corollary 3.3.28. *Let the assumptions of Theorem 3.3.25 be satisfied. Then it holds $u^* \in C(\Gamma_{\mathcal{N}})$.*

Proof. We have $u_a, u_b, q^* \in H^{2,2}_\beta(\Omega)$ by Assumption 3.3.4 and Theorem 3.3.25. Then by the continuity of the embedding $H^{2,2}_\beta(\Omega) \hookrightarrow C(\overline{\Omega})$, see [15], and by the projection formula (2.17), we conclude $u^* \in C(\Gamma_{\mathcal{N}})$. □

Lemma 3.3.29. *Let (ρ, ϕ) be polar coordinates centered at the origin. Define $S_\varepsilon(\omega) := \{x \in \mathbb{R}^2 \mid 0 < \rho(x) < \varepsilon,\ 0 < \phi(x) < \omega\}$ for $\omega \in (0, 2\pi)$ and $\varepsilon \in (0,1)$. Then there is a constant $C = C(\varepsilon) > 0$ independent of ω such it holds*

$$\| \rho^{\beta-1} y \|_{L^2(S_\varepsilon(\omega))} \leq C(\| \rho^\beta \nabla y \|_{L^2(S_\varepsilon(\omega))} + \| y \|_{L^2(S_\varepsilon(\omega))})$$

for all y such that the right-hand side is bounded.

Proof. We use Hardy's inequality in one dimension (see [106, Lemma A.1.6]) to compute

$$\begin{aligned}
\| \rho^{\beta-1} y \|^2_{L^2(S_\varepsilon(\omega))} &= \int_0^\omega \int_0^\varepsilon \rho^{2\beta-1} y^2 \,\mathrm{d}\rho \,\mathrm{d}\phi \\
&\leq C \int_0^\omega \left(\int_0^\varepsilon \rho^{1+2\beta} (\partial_\rho y)^2 \,\mathrm{d}\rho + \int_{\varepsilon/2}^\varepsilon y^2 \,\mathrm{d}\rho \right) \mathrm{d}\phi \\
&\leq C(\| \rho^\beta \nabla y \|^2_{L^2(S_\varepsilon(\omega))} + \| y \|^2_{L^2(S_\varepsilon(\omega))}).
\end{aligned}$$

□

Lemma 3.3.30. *Let β be a multi-index with $\beta \in (0,1)$. Then there is a constant $C > 0$ such that*

$$\|r^{\beta-1}v\|_{L^2(\Omega)} \leq C\|v\|_{H^{1,1}_\beta(\Omega)}$$

for all $v \in H^{1,1}_\beta(\Omega)$.

Proof. Let δ be given by (3.21). Then $S_i := \Omega \cap B_{\delta/4}(X_i)$ is a sector for all $X_i \in \mathcal{V}$ for all $i = 1,\dots,m$. Let us define $\rho_i(x) := \operatorname{dist}(x, X_i)$ for $x \in S_i$. Then by the previous Lemma 3.3.29, we obtain

$$\| \rho_i^{\beta_i-1}v \|_{L^2(S_i)} \leq C(\|v\|_{L^2(S_i)} + \|\rho_i^{\beta_i}\nabla v\|_{L^2(S_i)}).$$

For $x \in S_i$, we have $\min(1, \operatorname{dist}(x, X_i)) = \rho_i(x)$ and $\delta/4 \leq \min(1, \operatorname{dist}(x, X_j)) \leq 1$ for $i \neq j$. Thus it follows for $x \in S_i$.

$$r^{\beta-1}(x) \leq \rho_i(x)^{\beta_i-1}\,(\delta/4)^{|\beta|-m}\,, \quad \rho_i(x)^{\beta_i} \leq r^{\beta}(x).$$

This proves $\| r^{\beta-1}v \|_{L^2(S_i)} \leq C(\|v\|_{L^2(S_i)} + \|r^{\beta}\nabla v\|_{L^2(S_i)})$. On $S_0 := \Omega \setminus \cup S_i$ it holds $r(x) \geq \delta/4$. Hence $\| r^{\beta-1}v \|_{L^2(S_0)} \leq (\delta/4)^{|\beta|-m}\| v \|_{L^2(S_0)}$. Combining the estimates on S_i for $i = 0,\dots,m$ proves the claim. □

Now, we turn to the proof of Theorem 3.3.24.

Proof of Theorem 3.3.24. The goal is to establish estimates of type (3.25) for y^*, q^*. Then, the result follows from Lemma 3.3.10.

Step 1: Covering.

Let the vertex set $\mathcal{V}$ with elements X_j be as in Definition 2.1.5. Let $\varepsilon \in (0,1)$ be given. Then by Lemma 3.3.9 we obtain the countable covering $\mathcal{B} = \{B_i,\ i \in \mathbb{N}\}$. Let us denote by r_i the radius of the ball B_i. Let $\hat{B}_i$ denote the ball with same center as B_i and with stretched radius $\hat{r}_i := (1+\varepsilon)r_i$. Set $\hat{\mathcal{B}} := \{\hat{B}_i \mid i \in \mathbb{N}\}$.

Let the multi-index β be the one of Theorem 3.3.25 and set

$$\beta_i' := \begin{cases} \beta_j & \text{if } \operatorname{dist}(x_i, X_j) < \frac{\delta}{4} \text{ for some } j \in \{1,\dots,m\},\\ 1 & \text{otherwise,} \end{cases}$$

with δ as in (3.21).

Step 2: Local estimates of the data.

Due to the regularity assumption 3.3.4 and Lemma 3.3.10, there exist positive constants γ_f, $C_f(i)$ for $i \in \mathbb{N}$ with $\sum_{i=1}^{\infty} C_f(i)^2 < \infty$ such that

$$\| \nabla^p f \|_{L^2(\Omega\cap\hat{B}_i)} \leq \frac{C_f(i)}{\hat{r}_i^{\beta_i'}}\left(\frac{\tilde{\gamma}_f}{\hat{r}_i}\right)^p p! \quad \forall p \in \mathbb{N}_0. \tag{3.56}$$

Analogously, there are constants γ_d, $C_d(i)$ for $i \in \mathbb{N}$ with $\sum_{i=1}^{\infty} C_d(i)^2 < \infty$, γ_g, $C_g(i)$ for $i \in \mathbb{N}$ with $\sum_{i=1}^{\infty} C_g(i)^2 < \infty$, such that it holds

$$\| \nabla^p y_d \|_{L^2(\Omega\cap\hat{B}_i)} \leq \frac{C_d(i)}{\hat{r}_i^{\beta_i'}}\left(\frac{\tilde{\gamma}_d}{\hat{r}_i}\right)^p p! \quad \forall p \in \mathbb{N}_0 \tag{3.57}$$

and

$$\| \nabla^{p+1} u_a \|_{L^2(\Omega\cap\hat{B}_i)},\ \| \nabla^{p+1} u_b \|_{L^2(\Omega\cap\hat{B}_i)} \leq \frac{C_g(i)}{\hat{r}_i^{\beta_i'}} \left(\frac{\tilde{\gamma}_g}{\hat{r}_i}\right)^p p! \quad \forall p \in \mathbb{N}_0. \tag{3.58}$$

Step 3: Local estimates of the solution.
Let $B_i \in \mathcal{B}$. If on one hand B_i is such that $B_i \subset \Omega$ then (y^*, q^*) satisfy (3.31a) on B_i. Due the construction of the covering, on the other hand (y^*, q^*) with u^* satisfy (3.31a) on B_i with one of the following sets of conditions on $B_i \cap \Gamma$: (3.31b)–(3.31c), (3.31b) and (3.50), or (3.53). Thus, we can use the results of Subsection 3.3.3 to estimate the regularity of solutions on $B_i \cap \Omega$. We will use these estimates with $R := \hat{r}_i$ and $r := r_i$. Applying one of the Theorems 3.3.13, 3.3.21, 3.3.22, or 3.3.23 we obtain for $p \in \mathbb{N}_0$

$$(\hat{r}_i - r_i)^{p+2} \| \nabla^{p+2} y^* \|_{L^2(\Omega\cap B_i)} \leq C(i)\gamma^{p+2} p!$$

with

$$\begin{aligned} C(i) := {} & \hat{r}_i \| \nabla y^* \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i^{2-\beta_i'} C_f(i) + \hat{r}_i^2 C_c \| y^* \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i^2 \| y^* \|_{L^2(\Omega\cap\hat{B}_i)} \\ & + \hat{r}_i \| \nabla q^* \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i^{2-\beta_i'} C_d(i) + \hat{r}_i^2 C_c \| q^* \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i \| q^* \|_{L^2(\Omega\cap\hat{B}_i)} \\ & + \hat{r}_i^{2-\beta_i'} C_g(i) + \hat{r}_i \| u_a \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i \| u_b \|_{L^2(\Omega\cap\hat{B}_i)}. \end{aligned}$$

Observe that we can choose γ independent of the index i. The constant $C(i)$ is a combination of (3.34), (3.52), and (3.54). Here, we used the inequalities (3.56)–(3.58) to estimate the contributions of the data f, y_d, u_a, u_b. Hence, we obtain

$$\| \nabla^{p+2} y^* \|_{L^2(B_i\cap\Omega)} \leq C(i) \left(\frac{\gamma}{\varepsilon r_i}\right)^{p+2} p! = \gamma^2\, C(i) r_i^{\beta_i'-2} \frac{1}{r_i^{\beta_i'}} \left(\frac{\gamma}{\varepsilon r_i}\right)^p p!$$

Step 4: Global estimates of the solution.
In order to invoke Lemma 3.3.10, it remains to prove $\sum_{i=1}^\infty r_i^{2(\beta_i'-2)} C(i)^2 < \infty$. Due to the properties of $C_f(i)$, $C_d(i)$, $C_g(i)$, it holds

$$\sum_{i=1}^\infty r_i^{2(\beta_i'-2)} \left(\hat{r}_i^{2-\beta_i'} C_f(i) + \hat{r}_i^{2-\beta_i'} C_d(i) + \hat{r}_i^{2-\beta_i'} C_g(i)\right)^2 < \infty.$$

By the finite overlap property $\hat{\mathcal{B}}$, we find using $r_i < 1$

$$\sum_{i=1}^\infty r_i^{2(\beta_i'-2)} \hat{r}_i^4 \| y^* \|^2_{L^2(\Omega\cap\hat{B}_i)} \leq (1+\varepsilon)^4 \sum_{i=1}^\infty r_i^{2\beta_i'} \| y^* \|^2_{L^2(\Omega\cap\hat{B}_i)} \leq (1+\varepsilon)^4 N \| y^* \|^2_{L^2(\Omega)}.$$

Analogously, we obtain

$$\sum_{i=1}^\infty r_i^{2(\beta_i'-2)} \left(\hat{r}_i^2 C_c \| y^* \|_{L^2(\Omega\cap\hat{B}_i)} + \hat{r}_i^2 C_c \| q^* \|_{L^2(\Omega\cap\hat{B}_i)}\right)^2 < \infty.$$

Let us turn to estimate the contribution of $\hat{r}_i\| \nabla y^* \|_{L^2(\Omega\cap\hat{B}_i)}$. Using Corollary 3.3.11 and $\hat{r}_i = (1+\varepsilon)r_i$ we find

$$\begin{aligned}\sum_{i=1}^{\infty} r_i^{2(\beta_i'-2)}\hat{r}_i^2\| \nabla y^* \|^2_{L^2(\Omega\cap\hat{B}_i)} &= (1+\varepsilon)^2\sum_{i=1}^{\infty} r_i^{2(\beta_i'-1)}\| \nabla y^* \|^2_{L^2(\Omega\cap\hat{B}_i)}\\ &\leq (1+\varepsilon)^2 C^{-2}\sum_{i=1}^{\infty}\| r^{\beta-1}\nabla y^* \|^2_{L^2(\Omega\cap\hat{B}_i)}\\ &\leq C'\| r^{\beta-1}\nabla y^* \|^2_{L^2(\Omega)},\end{aligned}$$

where in the last step we relied on the finite overlap property of $\hat{\mathcal{B}}$. Using Lemma 3.3.30, yields

$$\begin{aligned}\sum_{i=1}^{\infty} r_i^{2(\beta_i'-2)}\hat{r}_i^2\| \nabla y^* \|^2_{L^2(\Omega\cap\hat{B}_i)} &\leq C'\| r^{\beta-1}\nabla y^* \|^2_{L^2(\Omega)}\\ &\leq C\| \nabla y^* \|_{H^{1,1}_\beta(\Omega)} \leq C\| y^* \|_{H^{2,2}_\beta(\Omega)} < \infty.\end{aligned}$$

Similarly, we can prove

$$\begin{aligned}&\sum_{i=1}^{\infty} r_i^{2(\beta_i'-2)}\hat{r}_i^2(\| \nabla q^* \|^2_{L^2(\Omega\cap\hat{B}_i)} + \| q^* \|^2_{L^2(\Omega\cap\hat{B}_i)} + \| u_a \|^2_{L^2(\Omega\cap\hat{B}_i)} + \| u_b \|^2_{L^2(\Omega\cap\hat{B}_i)})\\ &\qquad\leq C(\| q^* \|_{H^{2,2}_\beta(\Omega)} + \| q^* \|_{H^{1,1}_\beta(\Omega)} + \| u_a \|_{H^{1,1}_\beta(\Omega)} + \| u_b \|_{H^{1,1}_\beta(\Omega)}) < \infty.\end{aligned}$$

Thus, the convergence $\sum_{i=1}^{\infty} r_i^{2(\beta_i'-2)}C(i)^2 < \infty$ is proven. By Lemma 3.3.10, we find y^*, $q^* \in B^2_\beta(\Omega, C_*, \gamma_*)$ for some positive constants C_*, γ_*.

Step 5: Regularity of u^*.

To prove the regularity $u^* \in B^{3/2}_{\tilde{\beta}}(\Gamma_{\mathcal{N}}, C_u, \gamma_u)$ we need to construct an extension $\tilde{u}^* \in B^2_{\tilde{\beta}}(\Omega, C_u, \gamma_u)$. Let us partition Γ into pieces Γ_i, $i = 1, \ldots, m$ as in the proof of Theorem 3.3.25. Further, let us take Γ_i with $\Gamma_i \subset \Gamma_{\mathcal{N}}$. Then the optimal control $u^*|_{\Gamma_i}$ is the trace of a function in $B^2_\beta(\Omega)$. Using the trace theorem [13, Theorem 4.1] shows that $u^*|_{\Gamma_i} \in B^{k_i}_{\tilde{\beta}_i}(\Gamma_i)$ with $\beta_i \in \mathbb{R}^2$ a multi-index satisfying $\beta_i \in (0,1)$, and $k_i \in \{1,2\}$, where the value of k_i depends on β_i, β_{i+1}.

Let us note that the optimal control u^* is continuous on $\Gamma_{\mathcal{N}}$ because (see [15])

$$q^* \in B^2_\beta(\Omega, C_*, \gamma_*) \subset H^{2,2}_\beta(\Omega) \hookrightarrow C^0(\overline{\Omega}).$$

Then, we can apply the extension theorem [13, Theorem 4.3] to obtain the regularity $u^* \in B^{3/2}_{\tilde{\beta}}(\Gamma_{\mathcal{N}}, C_u, \gamma_u)$, where $\tilde{\beta} \in (0,1)$ is a multi-index satisfying $\tilde{\beta}_i \in (\beta_i, 1)$. □

3.3.5 Comments on the Derivation of Analyticity

The boundary weighted spaces in Section 3.2 damp any singularity towards the boundary and, therefore, allow to prove regularity results independent from the smoothness of the boundary data. A 'sequential' argument of the following type is possible:

$$u \in H^{1/2}(\Gamma_{\mathcal{N}}) \quad \Rightarrow \quad y \in B^2_{1-\sigma}(\Omega) \quad \Rightarrow \quad q \in B^2_{1-\sigma}(\Omega).$$

Unfortunately, this procedure is not applicable in Section 3.3 because the boundary data has to be considered. The optimality system serves as remedy and is exploited for obtaining higher regularity.

The amount of lemmas and tedious observations of various constants around the proof of Theorem 3.3.24 raise the question whether a 'simpler' or at least shorter line of reasoning is available. The trick of our proof is to localize the domain by a dichotomic covering and to stay on the local level to use the projection formula for a bootstrapping argument. This way, the local regularity results can be added up to obtain a global result.

It is possible to take the viewpoint of [106], where the author only investigates the regularity on sectors. The results are then transferred to polygonal domains by a partition of unity argument. It is unclear whether such a two-step procedure would lead to a simpler proof. Our approach has the advantage of requiring only one localization step.

The restriction to sectors is also done in [12]. Here, a sector is transformed into an infinite strip with the Mellin-transformation (see Subsection 3.1.1). An induction argument then yields the affiliation to countably normed spaces. This line of reasoning seems impossible for the coupled state-adjoint system because no a-priori regularity of the optimal variables is given. Instead, the projection formula has to be exploited for deriving smoothness in the manner of (3.39). In the induction step, however, the involved constants increase because one has to switch between the global and local view (Lemma 3.3.10). Hence, no global bound for the derivatives of the optimal variables seems to be available.

We showed analytic regularity for the solution of (**P**) subject to (N) with the help of local regularity estimates and a dichotomic covering. A shift theorem in the weighted space $H^{2,2}_{\beta}(\Omega)$ enabled us to add up the local results and obtain a global estimate. We believe that an analogous version of Theorem 3.3.24 holds for interface control problems. The covering of Lemma 3.3.9 can be extended to $2d$-networks. Moreover, a local regularity result holds ([106, Proposition 5.5.4]), as well as a shift theorem in the space $H^{2,2}_{\beta}$ ([106, Proposition A.2.1]). Note that the latter needs to deal with the case of jumping coefficients κ_i in the differential operator A (confer [106, Remark 5.2.3]). We presume that a bootstrapping technique with the projection formula allows to prove an analogous version of Lemma 3.3.19 that bounds tangential derivatives. This would amend the available results by a local regularity result at the interface (analogous to Theorem 3.3.13). For brevity, we omit a rigorous proof of analytic regularity in the context of transmission problems, confer [75].

CHAPTER 4

The *hp*-Finite Element Method

In the 1980's and 1990's, a solid theoretical and algorithmic foundation for the hp–version of the FEM as a reliable and efficient discretization technique was laid by [9, 10, 11, 12, 13, 14] and [52, 119, 123]. The articles [59, 107, 108, 139] are younger. We also mention the monographs [51, 53, 87, 135, 152] and [106], which give an overall and self-contained access to the topic.

The hp-FEM is a solution technique for PDEs that achieves higher accuracy by both refining the triangulation of Ω and increasing the polynomial degree of finite elements. This discretization method tries to approximation functions by

- polynomials of high degree on large elements in regions of high regularity,
- polynomials of low degree on small elements in regions of low regularity.

Obviously, the hp-FEM combines the h-version of FEM, which gains accuracy by decreasing the sizes of elements of fixed (low) polynomial degree, and the p-version of FEM, which gains accuracy by increasing the polynomial degree on a fixed discretization of the domain. See, e.g, [34, 35, 43] and [17, 26, 93, 135], respectively.

The number of unknowns N, i.e., the dimension of the approximation space, is the typical reference variable for comparing the efficiency of the different methods. Standard error estimates for higher order methods are of type

$$\| y - y_h \|_{H^1(\Omega)} \leq CN^{-t}, \quad t > 0 \tag{4.1}$$

for the solution y and its FE approximation y_h. As N appears polynomially in (4.1), we speak of *algebraic convergence*.

The symbol y_h originates from the uniform h-FEM, where the accuracy is measured with respect to the mesh size h. The polynomial relation $\sqrt{N} \sim h^{-1}$, or equivalently $h \sim N^{-1/2}$, yields bounds $\mathcal{O}(h^s), s > 0$ for the discretization error of $2d$ problems. Measuring the error in different norms, e.g., the *energy norm* as in (4.1) or Lebesgue norms, leads to different orders of N or h.

As indicated above, it is inevitable to thoroughly study the smoothness of the true solutions to the PDE, as done in the previous chapter. Now, we examine the (finite-dimensional) FE space as regards its approximation quality for functions of the given regularity. The chapter is organized as follows.

We start by introducing basic notations, definitions and ideas of the (hp-)FEM in Section 4.1. After that, we provide details on the implementation of the code that produced the numerical results for this thesis. The reader who does not care about programming issues is encouraged to skip Section 4.2.

In Section 4.3, we prove approximation results for the boundary concentrated (bc) FEM of [88] (see also [57]). This type of discretization heavily refines the mesh near the boundary of the domain while keeping the element size and polynomial degree large in the interior. The main result of this section is Theorem 4.3.13, which proves (4.1) with $t = \sigma$ for $H^{1+\sigma}$-regular problems. This result is possible because of the novel interpolant of Subsection 4.3.1 and because the number of unknowns of the bc-FEM is related to the boundary mesh size h via $h \sim N^{-1}$ (see Theorem 4.3.3). We formulate a similar result (Corollary 4.3.15) for interface concentrated (ic) finite elements, which is an application of the bc-FEM on the subdomains of a $2d$-network. Additionally, we present new estimates in the L^2- and L^∞-norm at the boundary.

Section 4.4 closes the chapter with the famous result of [9] that states that the hp-approximation error can be bounded in terms of $e^{-b\sqrt[3]{N}}$ for a constant $b > 0$ (Theorem 4.4.4). This *exponential convergence* can be achieved by meshing the domain with so called geometric mesh patches that are heavily refined towards vertices, where solutions tend to be singular (confer Section 3.1). The results will later be used in Chapter 5 for error estimates of what we call the vertex concentrated (vc) FEM.

Some of the presented results can also be found in [27, 156].

4.1 General Concepts

Finite element methods typically use the Galerkin method on a weak formulation of a PDE in an infinite-dimensional space V, which we generally write as

$$a(y, v) = l_u(v) \quad \forall v \in V. \tag{4.2}$$

Here, $a(\cdot, \cdot)$ is a coercive bilinear form that maps $V \times V \to \mathbb{R}$ and l_u is a member of V^* (see (2.12) or (2.13) as concrete examples).

The solution $y \in V$ is approximated in a finite-dimensional subspace

$$V_h := \operatorname{span}\{\phi_1, \ldots, \phi_N\} \subset V.$$

Since the approximating space is contained in the infinite-dimensional one, this method is referred to as conformal discretization. We do not allow discontinuous Galerkin methods (see the overview article [44] by Cockburn) which are also amenable to the hp-idea (see, e.g., [18, 32, 36]).

Functions in $v^h \in V_h$ can be uniquely represented as a vector $\tilde{v}^h \in \mathbb{R}^N$, where

$$v^h(x) = \sum_{j=1}^{N} \tilde{v}_j^h \phi_j(x) \tag{4.3}$$

and $\Phi := \{\phi_j\}$ is a basis of V_h.

Definition 4.1.1. *A coefficient $\tilde{v}_j^h$ in (4.3) is called **(global) degree of freedom** (short: dof, plural: ddof). We also refer to N as the **number of unknowns**.*

There is a one-to-one correspondence between $v^h \in V_h$ and $\tilde{v}^h \in \mathbb{R}^N$, which is why we implicitly change between the two and generally write v^h. Solving the discrete equations means finding a y^h that satisfies

$$a(y^h, v^h) = l_u(v^h) \quad \forall v^h \in V_h. \tag{4.4}$$

Note that we use the same space for testing the equation and representing the solution. The unique solvability of both (4.2) and (4.4) is guaranteed by Theorem 2.1.2.

In practice, we plug (4.3) into (4.4) and obtain a linear system of equations $\mathcal{A}y^h = \bar{l}_u$. Let the indices i, j run from $1, \ldots, N$ and let (4.4) be the discrete weak formulation of $-\Delta y + y = u$ in Ω. It holds

$$\mathcal{A}y^h = \mathcal{K}y^h + \mathcal{M}y^h = \bar{l}_u$$

with the stiffness and mass matrix

$$\mathcal{K}_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \,\mathrm{d}x, \quad \mathcal{M}_{ij} = \int_\Omega \phi_i\phi_j \,\mathrm{d}x,$$

and the load vector

$$\bar{l}_u := \int_\Omega u\phi_i \,\mathrm{d}x.$$

The following result is standard.

Theorem 4.1.2. *Let y and y^h be solutions to (4.2) and (4.4), respectively. Then **Galerkin-orthogonality** holds, i.e.,*

$$a(y - y^h, v^h) = 0 \quad \forall v^h \in V_h.$$

*Furthermore, we have **Cea's lemma***

$$\| y - y^h \|_V \leq C \inf_{v^h \in V_h} \| y - v^h \|_V.$$

The finite-dimensional approximation space V_h is constructed from a triangulation of Ω.

Definition 4.1.3. *A **k-irregular triangulation** τ of Ω is a collection of open, convex, and nonempty elements K such that*

- $\overline{\Omega} = \bigcup_{K \in \tau} \overline{K}$,
- $\overline{K}_i \cap \overline{K}_j$ *is either empty, a vertex or an edge,*
- *Each edge contains at most k irregular nodes (see Definition 4.1.4).*

*The word **mesh** will be used interchangeably with triangulation.*

Definition 4.1.4. *An element vertex is called a **regular** node if and only if it is a vertex to each neighboring element. **Irregular** or **hanging** nodes are those vertices that are not regular.*

The elements K usually are quadrilaterals or triangles formed by the edges $e_K \in E_K$.

Definition 4.1.5. *Let τ be a triangulation of Ω and h_K be the diameter of an element $K \in \tau$. We say τ is γ-shape regular if each K is the image of a reference element $\hat{K} \subset \mathbb{R}^2$ under a diffeomorphism F_K and there is a constant $\gamma > 0$ such that*

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

Here, F_K' denotes the Jacobian.

Definition 4.1.6. *Let τ be a 1-irregular mesh which is γ-shape regular. If all mappings F_K are affine linear, we say that τ is **admissible**.*

The word admissible is sometimes used in the context of regular/irregular meshes but merely signifies the class of triangulation we want to use in this thesis. Admissible meshes are compatible with Ω in so far, as we assumed in Section 2.1.2 that Ω and $2d$-networks are polygonal. Therefore, admissible triangulations can capture the boundary and interface, which are not curved.

The mapping F_K^{-1} allows a *pull-back* of $K \in \tau$ to $\hat{K}$ which is often used both for theoretical investigations (e.g., scaling arguments) and for numerical computations.

An example for a conform finite element space $V_h \subset V$ is

$$V_h^p := \{ v \in H^1(\Omega) \ : \ v|_K \text{ is a polynomial of degree } p \}.$$

If p is kept constant (h-FEM), accuracy is gained by refining the triangulation. An alternative would be to work with a fixed triangulation and increase only the polynomial degree (p-FEM). We are going to investigate a combination of both methods: hp-FEM. The FE approximation space is enlarged by refining the triangulation (h-refinement) on the one hand and by increasing the polynomial degree of elements (p-refinement) on the other hand.

Definition 4.1.7. *Let τ be an admissible triangulation. The collection of polynomial degrees $p_K \in \mathbb{N}$ on elements $K \in \tau$ is called **polynomial degree vector** $\mathbf{p} := (p_K)_{K \in \tau}$. The edges of K are collected in the set E_K and each edge $e_K \in E_K$ has an associated polynomial degree*

$$p_{e_K} := \min\{ p_{K'} \ : \ e_K \cap \overline{K}' \neq \emptyset, \ K' \in \tau \}. \tag{4.5}$$

The minimum condition has to be seen in the context of designing conform $V_h \subset H^1(\Omega)$ that contain continuous functions across edges (which may have hanging nodes).

Definition 4.1.8. *Let τ be an admissible mesh and $\mathbf{p}$ its polynomial degree vector. Define the space of polynomials*

$$\mathbb{R}[x]_p := \operatorname{span}\{x^i\}_{i=0,\dots,p},$$
$$\mathbb{R}[x,y]_p := \operatorname{span}\{x^i y^j\}_{0\le i,j\le p}$$

(with $0 \le i+j \le p$ for meshes consisting of triangles). Then the FE approximation space $S^{\mathbf{p}}(\tau)$ is defined as

$$S^{\mathbf{p}}(\tau) := \Big\{ v \in H^1(\Omega) \ : \ v|_K \circ F_K \in \mathbb{R}[x,y]_{p_K},\ v|_{e_K} \circ F_K \in \mathbb{R}[x]_{p_{e_K}} \\ \textit{for all } K \in \tau \textit{ and all } e_K \in E_K \Big\}.$$

In view of Cea's lemma, the properties of $S^{\mathbf{p}}(\tau)$ determine the numerical accuracy that is achieved by the hp-FEM (see section 4.3, 4.4).

4.2 Implementational Remarks

This section is intended to briefly describe the main issues that appear during the implementation of an hp-FEM code. While higher order methods are distinguished by good approximation properties, their implementation is more challenging.

There are several methods for designing the approximation space $S^{\mathbf{p}}(\tau)$, efficiently computing and assembling element matrices as well as solving the arising system of equations. The coding techniques should be sophisticated because a trivial approach often leads to high complexity, especially if the space dimension of the PDE is larger than one.

The basis function in the physical domain Ω need to be constructed in a way that promotes sparsity while retaining small condition numbers. The construction is achieved by shape functions on the reference domain, which are then mapped into the physical domain by F_K. Since the mappings are affine linear (see Definition 4.1.6), we obtain a sub-parametric mapping as soon as the polynomial degree of an element is greater than one.

4.2.1 Designing Basis Functions

Two strategies for constructing a basis Φ of the hp-FEM space are possible: the *modal* and *nodal* approach (see [87]). A nodal basis is constructed from shape functions that are Lagrange-polynomials on a set of discretization points. Evaluated at such a discretization point, the polynomial vanishes at all but one point, where it takes the value 1. The advantage of such a basis is the simple connection of the discrete function $v^h(x)$ and its representation as a linear combination of basis vectors. The drawback, however, is the fact that every time we increase the polynomial degree of an element K in the discretization τ, many basis functions need to be recomputed.

That is why we use a modal (hierarchical) basis. When p-refining the mesh, we keep the old basis and enrich the approximation space by additional functions. This way, the previous basis spans a subspace of the new approximation space.

The choice of the basis significantly determines the numerical properties of hp-methods and many different types have been discussed in literature (see [8, 37, 93] and the references therein). It is typical for hp-FEM to categorize the basis functions in the following way.

$$S^{\mathbf{p}}(\tau) = \operatorname{span}\Phi = \operatorname{span}(\Phi_V \cup \Phi_E \cup \Phi_I).$$

- Φ_V comprises all *hat functions* which are 1 at exactly one regular node of the mesh, and zero at all other regular nodes.

- Φ_E comprises so called *edge bubble functions*, which are non-zero at exactly one edge e with regular beginning and end node. They vanish at all other edges. Its support is formed by the union of all elements containing e.

- Φ_I comprises so called *element bubble functions*, whose support is contained in exactly one element.

The global basis functions shall have small support in the physical domain Ω so that the stiffness and mass matrix become sparse. An exemplary stiffness matrix is depicted in Figure 4.1, where we can see a 3×3 block structure resulting from the above categories.

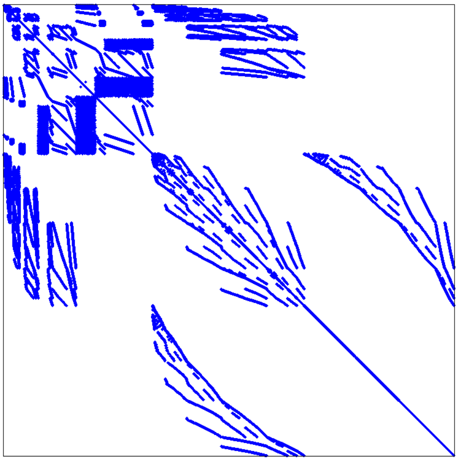


Figure 4.1. The sparsity pattern of a 12000×12000 stiffness matrix for a typical hp-discretization.

A basis function $\phi_i \in \Phi$ is constructed from contributions of local shape functions defined on the reference element $\hat{K}$ (see Figure 4.2).

It is desired to obtain sparse element matrices with small condition numbers. Suppose that we use Legendre polynomials, i.e., orthogonal polynomials with respect to the L^2 inner product. Then, for example, the mass matrix on the reference element is the identity matrix and very easy to invert. In order to exploit a similar orthogonality effect for the stiffness matrix, we use integrated Legendre polynomials as a hierarchical basis (see [26, 93]).

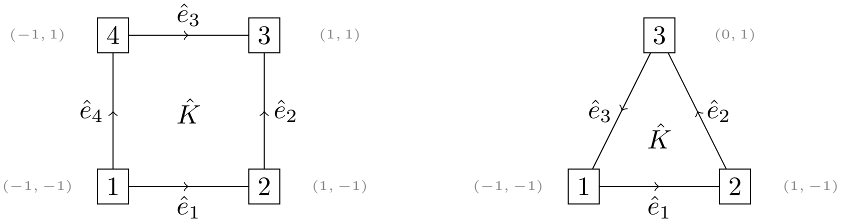


Figure 4.2. The quadrilateral and triangular reference element.

In order to emphasize that the following functions are constructed on reference domains, we use the notation $\hat{x} \in (-1, 1)$. The (integrated) Legendre polynomials $L_i(\hat{x})$ (respectively $\hat{L}_i(\hat{x})$) are defined as

$$L_i(\hat{x}) := \frac{1}{2^i i!} \frac{d^i}{d\hat{x}^i} (\hat{x}^2 - 1)^i, \quad i \in \mathbb{N}_0, \tag{4.6}$$

$$\hat{L}_i(\hat{x}) := (-1)^i \gamma_i \int_{-1}^{\hat{x}} L_i(s) \, \mathrm{d}s, \quad i \geq 2, \tag{4.7}$$

with the scaling factor

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

Furthermore, we set

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

Proposition 4.2.1. *[26, Lemma 2.1]. The Legendre polynomials L_i and integrated Legendre polynomials $\hat{L}_i$ satisfy the following properties:*

1. $\frac{d}{d\hat{x}} \hat{L}_i(\hat{x}) = (-1)^i \gamma_i L_i(\hat{x})$.
2. *The $L_i(\hat{x})$ are orthogonal with respect to the $L^2((-1, 1))$ inner product, i.e.,*

$$\int_{-1}^{1} L_i(\hat{x}) L_j(\hat{x}) \, d\hat{x} = \delta_{ij} \frac{2}{2i + 1}.$$

3. *The polynomials obey the recurrence formulas*

$$(i + 1)L_{i+1}(\hat{x}) + iL_{i-1}(\hat{x}) = (2i + 1)\hat{x}L_i(\hat{x}), \quad i \geq 1, \tag{4.8}$$

$$\hat{L}_i(\hat{x}) = (-1)^i \sqrt{\frac{(2i+1)(2i-3)}{4(2i-1)}}(L_i(\hat{x}) - L_{i-2}(\hat{x})), \quad i \geq 2. \tag{4.9}$$

We mention that the Legendre polynomials are a special case of Jacobi polynomials $P_i^{(\alpha,\beta)}(\hat{x})$ which are orthogonal on $(-1,1)$ with respect to the weight $(1-\hat{x})^\alpha(1+\hat{x})^\beta$, $\alpha, \beta > -1$. Jacobi polynomials are well investigated (see [143]) and can be used for the evaluation of higher derivatives of $\hat{L}$ (see [87, Appendix C]) or sparsity optimized shape functions on triangles/simplices (see [29, 30, 31]).

For the reference square $\hat{K} = [-1,1]^2 \subset \mathbb{R}^2$, we construct shape functions by tensorizing integrated Legendre polynomials (see Figure 4.3). Let

$$\hat{L}_{ij} = \hat{L}_{ij}(\hat{x},\hat{y}) := \hat{L}_i(\hat{x})\hat{L}_j(\hat{y}), \quad \deg(\hat{L}_{ij}) := \max\{\deg(\hat{L}_i), \deg(\hat{L}_j)\}.$$

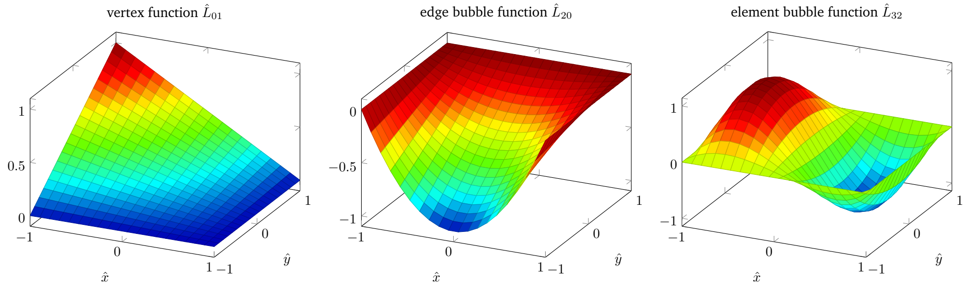


Figure 4.3. Shape basis functions in the hp-FEM on the reference square.

We adopt the same notation for the Legendre polynomials. The local shape functions up to degree $p > 0$, i.e., $\hat{\Phi} := \{\hat{L}_{ij}\}_{0 \leq i,j \leq p}$ benefit from the special properties of $\hat{L}_i$ in one dimension and produce sparse element matrices (see [26]).

The local numbering of shape functions $\hat{L}_{ij}(\hat{x},\hat{y})$ on the reference element $\hat{K}$ automatically introduces the concept of *local ddof* $\hat{v}_K^h$, which are a selection of the global ddof $v^h \in \mathbb{R}^N$ from a function in V_h in a special order. This connection of local and global ddof is described by a distribution mapping Ω_K (see [3] and [87, Section 4.2]) which satisfies the equation

$$\hat{v}_K^h = \Omega_K v^h$$

with a sparse matrix Ω_K that requires a memory saving implementation.

Let us mention an important fact about working with a hierarchical basis. Unlike nodal based functions, the sign of edge bubble functions (see Figure 4.3) may depend on the orientation of an edge. As most of the mathematical operations are performed on the reference element, both the edges of $\hat{K}$ and the physical element K possess a direction. If an orientation changed during the pull-back, the sign of the affected local ddof needs to be inverted in order to retain conformity. This phenomenon does not occur when using a nodal basis.

A Transformation to Legendre Polynomials

Now we show how the coefficients of a FE function $v^h \in S^{\mathbf{p}}(\tau)$ transform if the basis is changed from integrated Legendre polynomials to Legendre polynomials. This is necessary for estimating the smoothness of v which helps to decide between h- and p-refinement, see [59] and Subsection 6.3.2.

We start with the basis transformation in one dimension. First,

$$\hat{L}_0(\hat{x}) = -\frac{1}{2}(\hat{x}-1) = \frac{1}{2} - \frac{\hat{x}}{2} = \frac{1}{2}L_0(\hat{x}) - \frac{1}{2}L_1(\hat{x}), \tag{4.10}$$

$$\hat{L}_1(\hat{x}) = \frac{1}{2}(\hat{x}+1) = \frac{1}{2} + \frac{\hat{x}}{2} = \frac{1}{2}L_0(\hat{x}) + \frac{1}{2}L_1(\hat{x}). \tag{4.11}$$

Second, we have from (4.9) that for $i \geq 2$

$$\hat{L}_i(\hat{x}) = t_i(L_i(\hat{x}) - L_{i-2}(\hat{x})), \quad \text{where} \quad t_i := (-1)^i \sqrt{\frac{(2i+1)(2i-3)}{4(2i-1)}}.$$

Using these results, we can already set up the matrix for the basis transformation, i.e.,

$$T_p := \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & \cdots & & & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & \cdots & & & 0 \\ -t_2 & 0 & t_2 & 0 & \cdots & & 0 \\ 0 & -t_3 & 0 & t_3 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \ddots & \ddots & \vdots \\ 0 & \cdots & & 0 & -t_p & 0 & t_p \end{pmatrix}.$$

Let $p \geq 2$ and $i, j \in \{0, \ldots, p\}$. Let λ_i be the coordinates of a univariate polynomial $P \in \mathbb{R}[\hat{x}]_p$ in the ordered basis $\{\hat{L}_i\}$. We compute

$$P = \sum_i \hat{\lambda}_i \hat{L}_i = \sum_i \hat{\lambda}_i \sum_j (T_p)_{ij} L_j = \sum_i L_i \sum_j (T_p^\top)_{ij} \hat{\lambda}_j =: \sum_i L_i \lambda_i.$$

Hence, the matrix $T_p^\top$ describes the transformation of $\hat{\lambda}_i$ to the coordinates λ_i that represent P in the ordered basis $\{L_i\}$.

Definition 4.2.2. *Let $A \in \mathbb{C}^{k \times l}$ and $B \in \mathbb{C}^{m \times n}$ with $k, l, m, n \in \mathbb{N}$. The matrix*

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1l}B \\ \vdots & \ddots & \vdots \\ a_{k1}B & \cdots & a_{kl}B \end{pmatrix}$$

is called ***Kronecker product****.*

It is easy to verify that

$$(A \otimes B)^\top = A^\top \otimes B^\top. \tag{4.12}$$

As the $2d$ shape functions were constructed by tensorizing integrated Legendre polynomials, the Kronecker product provides a compact notation of the $2d$-transformation matrix if the both type of shape functions are arranged in the following lexicographic order.

$$\big(L_0(\hat{x}) \cdot L_0(\hat{y}), \ldots, L_0(\hat{x}) \cdot L_p(\hat{y}), L_1(\hat{x}) \cdot L_0(\hat{y}), \ldots, L_1(\hat{x}) \cdot L_p(\hat{y}), \\ \ldots, L_p(\hat{x}) \cdot L_{p-1}(\hat{y}), L_p(\hat{x}) \cdot L_p(\hat{y})\big)^\top.$$

Let $K \in \tau$ be an element with degree p_K. We collect and order the local ddof of $v \in S^{\mathbf{p}}(\tau)$ in $\hat{v}_K$. Recall that this corresponds to the pull-back which may involve the sign change of some coefficients as a consequence of changing orientations. Some additional zeros may have to be inserted into $\hat{v}_K$ if the finite element has different polynomial degrees on its edges. The coefficients $\tilde{v}_K^h$ in the ordered basis $\{L_i\}$, corresponding to the coefficients $\hat{v}_K$ in the ordered basis $\{\hat{L}_i\}$, can be computed with the following formula:

$$\tilde{v}_K^h = (T_p \otimes T_p)^\top \hat{v}_K^h = (T_p^\top \otimes T_p^\top)\hat{v}_K^h.$$

4.2.2 The Assembly Process

Element Matrices and Sum Factorization

We have seen that the three categories Φ_V, Φ_E, Φ_I of the global basis functions are transferred to the elemental level such that $\hat{\Phi} = \hat{\Phi}_V \cup \hat{\Phi}_E \cup \hat{\Phi}_I$ (see Figure 4.3). If the basis functions follow this order, an element matrix A_K, $K \in \tau$ is structured as follows.

$$A_K = \begin{pmatrix} A_{VV} & A_{VE} & A_{VI} \\ A_{VE}^\top & A_{EE} & A_{EI} \\ A_{VI}^\top & A_{EI}^\top & A_{II} \end{pmatrix}. \tag{4.13}$$

The condition number, number on non-zero entries, and preconditioners of such element matrices has been investigated in many publications. See, e.g., [8, 26, 31, 60, 93] and the references therein.

Computing an element matrix A_K for the degree p and a n-dimensional domain Ω means calculating $\mathcal{O}(p^{2n})$ entries. This is usually done by Gaussian quadrature with $\mathcal{O}(p^n)$ integration points. Hence, the overall complexity is of order $\mathcal{O}(p^{3n})$. This is too slow since the polynomial degree is increased significantly in hp-FEM.

In our code we made use of sum factorization (see, e.g., [57, 87]) which reduces the costs for computing an element matrix to $\mathcal{O}(p^{2n+1})$. Its basic idea is to precompute integrals so that some parts of the code can be moved out of nested for-loops. This intuitive code optimization is appealing because it is easy to implement and allows to treat PDEs with non-constant coefficients.

Suppose that we want to set up a c-weighted mass matrix

$$\mathcal{M}_K = \int_K c(x,y)\phi_r(x,y)\phi_s(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ = \int_{\hat{K}} (c \circ F_K)(\hat{x},\hat{y}) \cdot (\hat{L}_{ij}\hat{L}_{kl})(\hat{x},\hat{y}) \cdot |\det(F'_K(\hat{x},\hat{y}))|\,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y}$$

for an element of degree p. Let G be a Gaussian quadrature rule on $[-1,1]$ consisting of pairs of integration points $\hat{x}$ and their weights $w_{\hat{x}}$. For integrating the product of two basis functions, G needs to be at least exact up to degree $2p$. We abbreviate $c \circ F_K = \hat{c}$ and approximate the element matrix by

$$\mathcal{M}_K \approx \sum_{(\hat{x},w_{\hat{x}})\in G}\ \sum_{(\hat{y},w_{\hat{y}})\in G} \hat{c}(\hat{x},\hat{y})\hat{L}_i(\hat{x})\hat{L}_j(\hat{y})\hat{L}_k(\hat{x})\hat{L}_l(\hat{y})|\det(F'_K(\hat{x},\hat{y}))|w_{\hat{x}}w_{\hat{y}} \\ = \sum_{(\hat{x},w_{\hat{x}})\in G} \hat{L}_i(\hat{x})\hat{L}_k(\hat{x})w_{\hat{x}} \underbrace{\sum_{(\hat{y},w_{\hat{y}})\in G} \hat{L}_j(\hat{y})\hat{L}_l(\hat{y})\hat{c}(\hat{x},\hat{y})|\det(F'_K(\hat{x},\hat{y}))|w_{\hat{y}}}_{=:H_{j,l,x}}\,.$$

Precomputing $H_{j,l,x}$ and using its values for numerical quadrature saves costs of order $\mathcal{O}(p)$ (one 'for-loop' in the implementation). Sum factorization can be extended to triangular or tetrahedral elements as well.

Other methods for speeding up the assembly of the full system or matrix vector multiplication for iterative solvers can be found in [57].

Connectivity Arrays and Constrained Degrees of Freedom

In order to compute the full matrix $\mathcal{A}$ corresponding to (4.4), the element matrices need to be assembled in a global matrix. This is done by connectivity matrices $\Lambda_K = \Omega_K^\top$ which describe the connection of local to global ddof and the process of adding up the contributions of the single element matrices. The system $\mathcal{A}$ can be computed as

$$\mathcal{A} = \sum_{K\in\tau} \Lambda_K A_K \Lambda_K^\top. \tag{4.14}$$

Some diligence is required when setting up Λ_K for irregular meshes because the global basis functions need to be continuous across edges which contain a hanging node in its interior. Local ddof that correspond to an element with an irregular vertex are *constrained* by the values of global ddof. The following considerations are made according to Figure 4.4 and shall explain how constrained ddof are (uniquely) determined by the value of those ddof stemming from elements with only regular vertices. The procedure is similar to the basis transformation from Section 4.2.1.

Suppose that $e \subset K \in \tau$ is an edge that contains an irregular node in its interior. The hanging node is a vertex of two elements K_l, K_r and beginning/end node of $e_l := e \cap \overline{K}_l, e_r := e \cap \overline{K}_r$. As we pull-back e to $\hat{e} = [-1,1]$ orientated from left to right, we demand that the orientation of e_l, e_r point into the same direction as the orientation of e.

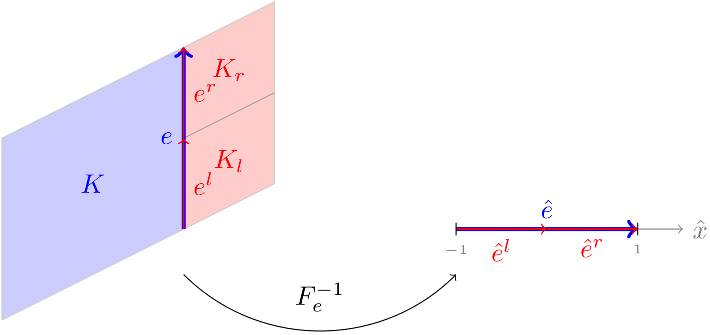


Figure 4.4. The pull-back for a constraining edge e onto the reference interval $[-1, 1]$.

The basis functions on $\hat{e}$ are the integrated Legendre functions $\hat{L}_i(\hat{x})$. They can be expressed by the basis functions

$$\hat{L}_i^l(\hat{x}) := \begin{cases} \hat{L}_i(2\hat{x}+1) & \text{if } 0 \geq \hat{x} \geq -1, \\ 0 & \text{if } 0 < \hat{x}, \end{cases} \tag{4.15}$$

$$\hat{L}_i^r(\hat{x}) := \begin{cases} \hat{L}_i(2\hat{x}-1) & \text{if } 0 \leq \hat{x} \leq 1, \\ 0 & \text{if } 0 > \hat{x}, \end{cases} \tag{4.16}$$

whose support is $\hat{e}^l$ and $\hat{e}^r$, respectively. For the vertex based functions $\hat{L}_{0,1}(\hat{x})$, we obviously have

$$\hat{L}_0(\hat{x}) = \hat{L}_0^l(\hat{x}) + \frac{1}{2}(\hat{L}_1^l(\hat{x}) + \hat{L}_0^r(\hat{x})), \tag{4.17}$$

$$\hat{L}_1(\hat{x}) = \hat{L}_1^r(\hat{x}) + \frac{1}{2}(\hat{L}_1^l(\hat{x}) + \hat{L}_0^r(\hat{x})). \tag{4.18}$$

For abbreviation, we introduce $\hat{L}^m(\hat{x}) := \hat{L}_1^l(\hat{x}) + \hat{L}_0^r(\hat{x})$ which corresponds to a hat function for the hanging node.

It remains to find a representation

$$\hat{L}_i(\hat{x}) = a_i\hat{L}^m(\hat{x}) + \sum_{j\geq 2} a_{i,j}^l \hat{L}_i^l(\hat{x}) + \sum_{j\geq 2} a_{i,j}^r \hat{L}_i^r(\hat{x}), \quad i \geq 2.$$

For representing $\hat{L}_i$, we only need Legendre polynomials up to the same degree. That is why the matrices $a_{i,j}^l, a_{i,j}^r$ are lower triangular.

If we list $\hat{L}_0^l, \hat{L}_1^r, \hat{L}^m$ first, followed by the left and right edge bubble functions $\hat{L}_{i\geq 2}^l$, $\hat{L}_{i\geq 2}^r$, the transformation matrix for $p \geq 2$ reads

$$P_p := \begin{pmatrix} 1 & 0 & \frac{1}{2} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & \frac{1}{2} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & a_2 & a_{2,2}^l & & & a_{2,2}^r & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \vdots & \ddots & \\ 0 & 0 & a_p & a_{p,2}^l & \cdots & a_{p,p}^l & a_{p,2}^r & \cdots & a_{p,p}^r \end{pmatrix}.$$

The matrix P_p describes how the basis functions on $\hat{e}^l, \hat{e}^r$ add up to the basis functions on $\hat{e}$. It can be computed by equating the coefficients. With the notation

$$(r)_n = r \cdot (r+1) \cdot \ldots \cdot (r+n-1), \quad r \in \mathbb{R}, n \in \mathbb{N}$$

and the convention $a_i = \gamma_i \hat{a}_{i,1}^r,\ a_{i,j}^r = \frac{\gamma_i}{\gamma_j} \hat{a}_{i,j}^r$ we find

$$\begin{aligned} \hat{a}_{2i,1}^r &= (-1)^{i+1} \frac{(-1/2)_i}{i!}, && i \geq 1, \\ \hat{a}_{2i+1,1}^r &= 0, && i \geq 1, \\ \hat{a}_{i,i}^r &= 2^{-i}, && i \geq 2, \\ \hat{a}_{2i,2}^r &= \frac{3}{2} \hat{a}_{2,i}, && i \geq 2, \\ \hat{a}_{2i+1,2}^r &= (-1)^i \frac{(-3/2)_{i+1}}{(i+1)!}, && i \geq 1, \\ \hat{a}_{i,j}^r &= 0, && i > j \geq 2. \end{aligned}$$

The remaining values can be computed with the recurrence formula

$$\hat{a}_{i,j}^r = -\hat{a}_{i-2,j-1}^r + \frac{1}{2} \hat{a}_{i-1,j-1}^r - \hat{a}_{i-1,j}^r + \frac{1}{2} \hat{a}_{i-1,j+1}^r, \quad i \geq 4,\ j \geq 3.$$

The entries of a^l read

$$\begin{aligned} a_{2i,1}^l &= a_{2i,1}^r, && j \geq i \geq 2, \\ a_{2i,j}^l &= (-1)^{j+1} a_{2i,j}^r, && i \geq 1,\ j \geq 2,\ i < j, \\ a_{2i+1,j}^l &= (-1)^j a_{2i+1,j}^r, && i \geq 1,\ j \geq 2,\ i < j. \end{aligned}$$

Given the matrix P_p, it is easy to determine the values of the constrained ddof: Let $v \in S^{\mathbf{p}}(\tau)$ and $v_{\hat{e}}$ be the coefficients corresponding to edge bubble functions of an edge $\hat{e}$ with irregular node in its interior. Denote by $v_{l,m,r}$ the ddof belonging to the left, middle (constrained) and right node that form $\hat{e}$. Denote by $v_{\hat{e}^l},\ v_{\hat{e}^r}$ the constrained ddof corresponding to the left and right sub edge of $\hat{e}$ (see Figure 4.4). Then the constrained ddof are determined by

$$(v_l, v_r, v_m, v_{\hat{e}^l}, v_{\hat{e}^r})^\top = P_p^\top (v_l, v_r, v_{\hat{e}})^\top.$$

With this knowledge, the connectivity matrices Λ_K can be set up as described in [3] .

Note that the assembly according to (4.14) does not consider Dirichlet boundary conditions. These have to be enforced in a second step. Suppose i is the index of a Dirichlet node. Then we overwrite the i-th row of $\mathcal{A}$ with the unit vector $e_i \in \mathbb{R}^N$. The desired Dirichlet value is written to the corresponding entry of the load vector l. Non-linear boundary data as well as higher order boundary elements are more complicated because the data has to be mapped into the right polynomial space and the coefficients for bubble functions need to be computed.

Inverting the full $N \times N$ system $\mathcal{A}$ means solving a discretized PDE. Direct solvers are desirable in the context of optimal control problems because optimization methods, such as projected gradient, semi-smooth Newton, interior point or SQP methods, require a lot of solves of the state and adjoint equation. Thus, it is efficient to store the decomposition of $\mathcal{A}$ and reuse it during the optimization process. Since the matrix $\mathcal{A}$ is generally very sparse (see Figure 4.1), a direct solver must exploit the sparsity pattern for being efficient. We used UMFPACK[2] as a sparse LU solver. Other possibilities are PARDISO[3] or SUPERLU[4].

4.3 Algebraic Convergence of the Boundary Concentrated Finite Element Method

In the following, we will provide approximation results for a special hp-strategy: the boundary concentrated finite element method (bc-FEM). A novel interpolant is constructed in Subsection 4.3.1 (see also [27]). It helps establish the main result of this chapter (Theorem 4.3.13), which states that the error decay is algebraic with respect to the boundary mesh size. Such an energy-norm estimate also holds for the interface concentrated (ic-) FEM (see Corollary 4.3.15), which is an application of bc-FEM on the subdomains of a $2d$-network.

Additionally, we are going to derive a-priori error estimates measured by Lebesgue-norms in Subsection 4.3.2. Although the theory is formulated for quadrilaterals it extends to triangular elements.

Definition 4.3.1. *Let τ_h be an admissible triangulation and $h := \min_{\overline{K}\cap\Gamma\neq\emptyset}\{h_K\}$<1 be a measure for the mesh size at the boundary. We speak of τ_h as a* ***boundary concentrated*** *(bc) mesh, if and only if there exist constants $c_1, c_2 > 0$ such that for all $K \in \tau_h$:*

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 $c_1 \inf_{x\in K} \operatorname{dist}(x, \Gamma) \leq h_K \leq c_2 \sup_{x\in K} \operatorname{dist}(x, \Gamma)$.*

Note that condition 1 implies that the discretization of the boundary, originating from the triangulation of Ω, is quasi-uniform. This justifies to speak of *the boundary mesh size h*.

[2] http://www.cise.ufl.edu/research/sparse/umfpack/

[3] http://www.pardiso-project.org/

[4] http://crd-legacy.lbl.gov/~xiaoye/SuperLU/

Definition 4.3.2. *Let τ_h be a bc-mesh on Ω with mesh size h. The polynomial degree vector $\mathbf{p} = (p_K)_{K \in \tau_h}$ is said to be **linear** with **slope** $\alpha > 0$ if there exist constants c_1, $c_2 > 0$ such that*

$$1 + \alpha c_1 \log \frac{h_K}{h} \leq p_K \leq 1 + \alpha c_2 \log \frac{h_K}{h} \,. \tag{4.19}$$

We speak of *boundary concentrated (bc)* FEM if we solve an elliptic PDE with $V_h = S^{\mathbf{p}}(\tau)$ in (4.4), where $S^{\mathbf{p}}(\tau)$ is the approximation space arising from discretizations according to Definition 4.3.1 and 4.3.2

In the context of $2d$-networks $\{\Omega_i\}_{i \in I}$ and interface problems, we can apply the bc-FEM on each subdomain Ω_i, which leads to the *interface-concentrated (ic)* FEM. This corresponds to replacing the boundary Γ by $\Gamma \cup \mathcal{I}$ in Definition 4.3.1.

Theorem 4.3.3. *Let τ_h be a bc-mesh with polynomial degree of slope $\alpha > 0$. Then there exists a $C > 0$ independent of h such that*

$$\sum_{K \in \tau_h} 1 \leq C h^{-1},$$

$$\max_{K \in \tau_h} p_K \leq C |\ln h|,$$

$$\dim(S^{\mathbf{p}}(\tau)) \sim \sum_{K \in \tau} p_K^2 \leq C h^{-1}$$

The result is proved in [88, Proposition 2.7] and extends to the ic-FEM. It points out that the number of unknowns increases linearly if the boundary mesh size is decreased. For uniform refinement in h-FEM, we have the relation $h \sim N^{-1/2}$ on two dimensional domains. Thus, the bc- and ic-FEM solve a two dimensional problem 'for the costs of a one dimensional problem', which is the reason for its fast convergence (see Chapter 5).

To some degree, the bc-FEM can be regarded as a generalization of the boundary element method (BEM), see [76, 129] and the references therein. There, the PDE is reformulated to an equation posed only on the boundary, which yields the same reduction in the dimension. However, BEM is restricted to equations where a fundamental solution is known, which generally fails to be the case for non-constant differential operators. The bc-FEM, on the other hand, can handle analytic coefficients in the elliptic equation.

4.3.1 Energy Norm Estimates

In the sequel, we will construct a bc-FEM interpolation operator. Since we allow hanging nodes, the result generalizes [88]. Regarding the interpolation error, we obtain approximation results comparable to those obtained in [88] for regular meshes.

Estimates of local element size and polynomial degrees

In the interpolation estimates below, it will be important to have comparable element size and element polynomial degree for neighboring elements. For meshes without hanging nodes, we have the following result from [107, Lemma 2.3], its extension to meshes with hanging nodes as used here is straightforward.

Lemma 4.3.4. *Let τ be a γ-shape-regular mesh. Then there exists a constant $C(\gamma)$ such that for two neighboring elements K, K' with $\overline{K} \cap \overline{K}' \neq \emptyset$ there holds*

$$C(\gamma)^{-1} h_K \leq h_{K'} \leq C(\gamma) h_K. \tag{4.20}$$

Theorem 4.3.5. *Let τ_h be a bc-mesh on Ω with a linear polynomial degree vector $\mathbf{p}$ of slope α. Then there is a constant $C(\alpha)$ depending on γ such that for two neighboring elements K, K' with $\overline{K} \cap \overline{K}' \neq \emptyset$ it holds*

$$C(\alpha)^{-1} p_K \leq p_{K'} \leq C(\alpha) p_K.$$

Moreover, $C(\alpha) \in \mathcal{O}(\alpha)$.

Proof. The constants c_1, c_2 defining the linear degree vector naturally satisfy $c_2 > c_1$, cf. Definition 4.3.2. Using the properties of the linear degree vector and Lemma 4.3.4 we can estimate

$$\begin{aligned} p_{K'} &\leq 1 + \alpha c_2 \log(h_{K'}/h) \\ &\leq 1 + \alpha c_2 \log(C(\gamma) h_K/h) \\ &\leq 1 + \alpha c_2 \log(h_K/h) + \alpha c_2 \log(C(\gamma)) \\ &\leq c_2 c_1^{-1} (1 + \alpha c_1 \log(h_K/h) + \alpha c_2 \log(C(\gamma))) \\ &\leq c_2 c_1^{-1} (p_K + p_K \alpha c_2 \log(C(\gamma))) \\ &\leq c_2 c_1^{-1} (1 + \alpha c_2 \log(C(\gamma))) p_K. \end{aligned}$$

The same computation yields a bound of p_K from above. This proves the claim with $C(\alpha) := \frac{c_2}{c_1}(1 + \alpha c_2 \log(C(\gamma)))$. □

Extension and projection operators

The reference element we have in mind is the square $(-1,1)^2$, but we will keep the notation relatively neutral to make the results applicable to triangles as well. The index i is taken from $\{1,2,3(,4)\}$. We denote the reference element by $\hat{K}$ and recall the space $\mathbb{R}[x,y]_p := \operatorname{span}\{x^i y^j \;:\; 0 \leq i,j \leq p\}$. Triangles would require the space $\operatorname{span}\{x^i y^j \;:\; 0 \leq i+j \leq p\}$.

As our mesh will have hanging nodes, we assume that each edge e_i of the reference element has an associated polynomial degree $p_i := p_{e_i}$ (see also (4.5)) with $p_i \leq p_K$. Here, p_K is the polynomial degree of an arbitrary element K that is pulled back to $\hat{K}$. The constructed approximant will lie in

$$P_{\mathbf{p}(K)}(\hat{K}) := \{f \in \mathbb{R}[x,y]_p \;:\; \deg(f|_{e_i}) \leq p_i\} \quad \text{with} \quad \mathbf{p}(K) := (p_K, p_1, \ldots, p_4). \tag{4.21}$$

We first need an extension operator acting from $\partial\hat{K}$ to $\hat{K}$ (see [106, Lemma 3.2.3]).

Lemma 4.3.6. *Let $f \in C(\partial\hat{K})$ be a polynomial of degree p_i on the i-th edge of the reference element. There exists a linear extension mapping $E : C(\partial\hat{K}) \to P_{\mathbf{p}(K)}(\hat{K})$ with the following properties:*

$$(Ef)|_{e_i} = f, \tag{4.22}$$

$$\| Ef \|_{L^\infty(\hat{K})} + p^{-2}\| \nabla Ef \|_{L^\infty(\hat{K})} \leq c\| f \|_{L^\infty(\partial\hat{K})}. \tag{4.23}$$

Proof. We prove this only in the case of $\hat{K}$ being the reference square. The extension to triangular $\hat{K}$ is straightforward, see, e.g., [106, Lemma 3.2.3].

By subtracting a bilinear function from f we can assume that it vanishes on the vertices of the reference element. For each $f_i := f|_{e_i}$ we construct an extension $E_i(f_i) \in P_{\mathbf{p}(K)}(\hat{K})$ which is zero at all other edges e_j, $j \neq i$.

Let us demonstrate the construction of $E_i(f_i)$ for e_1, $e_1 := \{(x, y) \in \mathbb{R}^2 \ : \ x \in [-1, 1], y = -1\}$. Here we define $E_1(f_1) := \frac{1-y}{2} f(x)$. Analogously we define the extension from the edges e_i, $i > 1$. This way we get an extension $F := E(f) := \sum_i E_i(f_i)$.

With the inverse estimate $\| \nabla F \|_{L^\infty(\hat{K})} \leq cp^2\| F \|_{L^\infty(\hat{K})}$ ([135, Theorem 4.76]) with $p \geq p_i$ we only need to show $\| F \|_{L^\infty(\hat{K})} \leq c\| f \|_{L^\infty(\partial\hat{K})}$. This is a trivial estimate: $\|E_1(f_1)\|_{L^\infty(\hat{K})} \leq \|f_1\|_{L^\infty(e_1)}$, as $\frac{1-y}{2} \leq 1$ on $\hat{K}$.

In the case that f does not vanish in the vertices let us denote by F_0 the bilinear interpolation of f that is exact in the vertices. Then we set $Ef := F_0 + \sum_i E_i(f_i - F_0)$. It is now easy to argue that the extension fulfills the claim. □

Element-wise Interpolation on Boundary Concentrated Meshes

For the remaining part of the subsection, we adopt the notation from the theory/numerics of PDEs and use the variable u for a general function and not the control variable. The aim of this section is to construct an interpolant on the reference element. It is desired to interpolate a function u living on the physical domain Ω by pulling it back to the reference element for each element of the finite element discretization τ.

The constructed interpolator will be needed for elements in the interior of Ω. There, we need to distinguish between elements possessing a hanging node or not.

At first, we will construct the interpolator for elements without hanging nodes. The following theorem is similar to [88, Lemma 2.9]. We give a proof here in order to track the dependence of the constants on the parameter α of the linear degree vector.

In the sequel, we will denote by $GL(q, f)$ the one-dimensional Gauss-Lobatto interpolant of degree $q \geq 1$ for the function f on $I = [-1, 1]$. The Gauss-Lobatto interpolation points $x_0, \ldots, x_q$ are defined as the roots of $(1 - \hat{x}^2)L'_{q-1}(\hat{x})$ (see (4.6)). It is well known that these roots are real, distinct, and lie in $[-1, 1]$ (see [143, Theorem 3.3.1]). Hence, the definition is well-posed and we have the representation

$$GL(q, f) = \sum_{i=0}^{q} f(x_i) \prod_{\substack{j=0\\ j\neq i}}^{q} \frac{x - x_j}{x_i - x_j}.$$

Theorem 4.3.7. *Let $\hat{K}$ be the reference element. Let u be a function on Ω whose pull back $\hat{u} = u \circ F_K$ is analytic on $\overline{\hat{K}}$ and satisfies*

$$\| \nabla^{q+2} \hat{u} \|_{L^\infty(\hat{K})} \leq C_u \gamma_u^q q!, \quad q = 0, 1, 2, \dots.$$

Then there exists an interpolant $I(u) \in P_{\mathbf{p}(K)}(\hat{K})$ such that

1. *$I(\hat{u})|_{e_i} = GL(p_i, \hat{u}|_{e_i})$,*
2. *$\| I(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq C_\alpha C_u e^{-bp_m}$,*

where $b > 0$ depends on γ_u, and $C_\alpha > 0$ depends on γ_u and α with $C_\alpha = \mathcal{O}(\alpha^6)$ for $\alpha \to \infty$.

Here, p_m denotes the minimal polynomial degree is defined by $p_m := \min_i\{p_i\}$ and naturally $p_m \leq p_i \leq p$ with p being the degree of the image of I, i.e., $P_{\mathbf{p}(K)}(\hat{K})$.

Proof. We restrict $\hat{u} \in C(\hat{K})$ to the boundary $\partial\hat{K}$ and define the piecewise Gauss-Lobatto interpolation operator

$$i : C(\partial\hat{K}) \to \{f \in C(\partial\hat{K}) \ : \ f|_{e_i} \text{ is polynom with degree } p_i\},$$
$$i(\hat{u})(x) = GL(p_i, \hat{u}|_{e_i})(x) \quad \forall x \in \partial\hat{K}.$$

Let us define the finite-dimensional subspace

$$V := \{u \in P_{\mathbf{p}(K)}(\hat{K}) \ : \ \hat{u}|_{\partial\hat{K}} = 0\}.$$

Since V is finite-dimensional, there is a linear and bounded projection operator $\Pi : P_{\mathbf{p}(K)}(\hat{K}) \to V$ with $\|\Pi\|_{\mathcal{L}(C(\overline{\hat{K}}),C(\overline{\hat{K}}))} \leq \sqrt{\dim V}$, confer [106, Theorem A.4.1]. As $V \subset P_{\mathbf{p}(K)} \subset \mathbb{R}[x,y]_p(\hat{K})$, we have $\dim(V) \leq (p+1)^2$, which shows $\|\Pi\|_{\mathcal{L}(C(\overline{\hat{K}}),C(\overline{\hat{K}}))} \leq p+1$.

The interpolation operator I is now defined by

$$I(\hat{u}) := E(i(\hat{u})) + \Pi(\hat{u} - E(i(\hat{u})))$$

with the extension operator E from Lemma 4.3.6. By construction, the first property is fulfilled. If $\hat{u} \in P_{\mathbf{p}(K)}(\hat{K})$ it follows that $i(\hat{u}) = \hat{u}|_{\partial\hat{K}}$ and therefore $\hat{u} - E(i(\hat{u})) \in V$. Thus, I interpolates functions of $P_{\mathbf{p}(K)}(\hat{K})$ exactly.

Let $\hat{u} \in C(\overline{\hat{K}})$ be given. Let us first estimate the norm of I by

$$\begin{aligned}\| I(\hat{u}) \|_{L^\infty(\hat{K})} &\leq c\| i(\hat{u}) \|_{L^\infty(\partial\hat{K})} + (p+1)\| \hat{u} - E(i(\hat{u})) \|_{L^\infty(\hat{K})} \\ &\leq c(1+\ln p)\| \hat{u} \|_{L^\infty(\partial\hat{K})} + (p+1)\| \hat{u} \|_{L^\infty(\hat{K})} \\ &\quad + c(1+\ln p)(p+1)\| \hat{u} \|_{L^\infty(\partial\hat{K})},\end{aligned}$$

where we used [106, Lemma 3.2.1] to bound the Gauss-Lobatto-interpolation operator i. Exploiting $\| \hat{u} \|_{L^\infty(\partial\hat{K})} \leq \| \hat{u} \|_{L^\infty(\hat{K})}$ for $\hat{u} \in C(\hat{K})$ yields the estimate

$$\| I(\hat{u}) \|_{L^\infty(\hat{K})} \leq C_I p(1+\ln p)\| \hat{u} \|_{L^\infty(\hat{K})}.$$

Regarding approximation properties, it now follows with arbitrary $v \in P_{\mathbf{p}(K)}(\hat{K})$ and using $v = Iv$ that

$$\begin{aligned} \| \hat{u} - I(\hat{u}) \|_{L^\infty(\hat{K})} &= \| (\hat{u} - v) - I(\hat{u} - v) \|_{L^\infty(\hat{K})} \\ &\leq (1 + C_I p(1 + \ln p)) \| \hat{u} - v \|_{L^\infty(\hat{K})}. \end{aligned}$$

In order to achieve an approximation property in $W^{1,\infty}(\hat{K})$, we need to estimate the first derivatives of $\hat{u} - I(\hat{u})$:

$$\begin{aligned} &\| \nabla(\hat{u} - I(\hat{u})) \|_{L^\infty(\hat{K})} \\ &\quad = \| \nabla((\hat{u} - v) - I(\hat{u} - v)) \|_{L^\infty(\hat{K})} \\ &\quad \leq \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} + \| \nabla(I(\hat{u}) - v) \|_{L^\infty(\hat{K})} \\ &\quad \leq \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} + Cp^2 \| (I(\hat{u}) - v) \|_{L^\infty(\hat{K})} \\ &\quad \leq \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} + Cp^2 (\| (I(\hat{u}) - \hat{u}) \|_{L^\infty(\hat{K})} + \| \hat{u} - v \|_{L^\infty(\hat{K})}) \\ &\quad \leq \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} + Cp^2 (2 + C_I p(1 + \ln p)) \| \hat{u} - v \|_{L^\infty(\hat{K})}. \end{aligned}$$

In the last two estimates, we can pass to the infimum because v was arbitrary, which shows

$$\begin{aligned} \| \hat{u} - I(\hat{u}) \|_{L^\infty(\hat{K})} &\leq \hat{C}_1 p(1 + \ln p) \inf_{v \in P_{\mathbf{p}(K)}(\hat{K})} \| \hat{u} - v \|_{L^\infty(\hat{K})}, \\ \| \nabla(\hat{u} - I(\hat{u})) \|_{L^\infty(\hat{K})} &\leq \inf_{v \in P_{\mathbf{p}(K)}(\hat{K})} \{ \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} \\ &\qquad + \hat{C}_2 p^3 (1 + \ln p) \| \hat{u} - v \|_{L^\infty(\hat{K})} \}. \end{aligned}$$

Relying on best approximation results in the space $P_{\mathbf{p}(K)}(\hat{K})$, we have [106, Theorem 3.2.19]

$$\begin{aligned} \inf_{v \in P_p(\hat{K})} \| \hat{u} - v \|_{L^\infty(\hat{K})} &\leq C C_u e^{-b' p_m}, \\ \inf_{v \in P_p(\hat{K})} \| \nabla(\hat{u} - v) \|_{L^\infty(\hat{K})} &\leq C C_u e^{-b' p_m} \end{aligned}$$

with constants C, b' depending both on γ_u. Collecting the estimates above, we obtain

$$\begin{aligned} \| I(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq \hat{C}_1 p(1 + \ln p) C C_u e^{-b' p_m} \\ + C C_u e^{-b' p_m} + \hat{C}_2 p^3 (1 + \ln p) \hat{C}_1 p (1 + \ln p) C C_u e^{-b' p_m}. \end{aligned}$$

We have from Theorem 4.3.5 that $C(\alpha)^{-1} p_{K'} \leq p_K \leq C(\alpha) p_{K'}$ for two neighboring elements K, K'. Hence, we can bound $p \leq C(\alpha) p_m$ because the minimal polynomial degree is determined by at least one neighbor. This way we get

$$\| I(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq \hat{C}_3 C(\alpha)^6 p_m^6 C C_u e^{-b' p_m}. \tag{4.24}$$

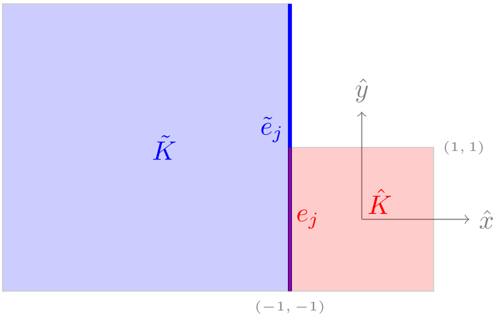


Figure 4.5. A reference element with hanging node and its neighbor (possibly distorted).

Absorbing p_m^6 by decreasing the constant b' yields

$$\| I(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq C_\alpha C_u e^{-bp_m}$$

with C_α depending on α, γ_u and b on γ_u, and $C_\alpha = \mathcal{O}(\alpha^6)$ for $\alpha \to \infty$. □

Remark 4.3.8. *We cannot avoid the constant $p(1+\ln p)$ in the estimates of the interpolation errors $\| \hat{u} - I(\hat{u}) \|_{L^\infty(\hat{K})}$ and $\| \nabla(\hat{u} - I(\hat{u})) \|_{L^\infty(\hat{K})}$ because we allow different polynomial degrees in the interior and on the edges of elements.*

In the second step, we will construct an interpolation operator that can deal with hanging nodes. To begin with, we cite an one-dimensional interpolation result of [106, Lemma 3.2.6].

Lemma 4.3.9. *Let u be analytic on the interval $I = (-1,1)$ and satisfy for some C_u, γ_u*

$$\| \nabla^{q+2} u \|_{L^\infty(I)} \leq C_u \gamma_u^q q!, \quad q = 0,1,2,\dots.$$

There are constants $C, b > 0$ depending on γ_u such that $GL(q,u)$ satisfies for $p = 1,2,\dots$

$$\| u - GL(p,u) \|_{W^{1,\infty}(I)} \leq C C_u e^{-bp}.$$

Proof. In [106, Lemma 3.2.6], the estimate $\| u - GL(p,u) \|_{W^{1,\infty}(I)} \leq \kappa C_u \left(\frac{1}{1+\sigma}\right)^{p+1}$ is proved with $\kappa, \sigma > 0$ depending on γ_u. With $C = \kappa(1+\sigma)^{-1}$ and $b = \ln(1+\sigma)$ we obtain the desired estimate. □

Let us describe now the construction of an interpolator on elements with hanging nodes. Depending on the position of the hanging nodes, we prolong the local edge e_j to the full coarse edge $\tilde{e}_j$, with j from $\{1,2,3(,4)\}$. An exemplary situation is depicted in Figure 4.5.

Theorem 4.3.10 (hanging nodes). *Let $\hat{K}$ be the reference element. Let u be a function on Ω whose pull back $\hat{u}$ is analytic on $\overline{\hat{K}}$ and satisfies*

$$\| \nabla^{q+2} \hat{u} \|_{L^\infty(\hat{K})} \leq C_u \gamma_u^q q!, \quad q = 0, 1, 2, \ldots . \tag{4.25}$$

Let the indices i represent the free edges, whereas j denotes constrained edges due to the existence of hanging nodes. If additionally it holds

$$\| \nabla^{q+2} \hat{u} \|_{L^\infty(\tilde{e}_j)} \leq C_u \gamma_u^q q!, \quad q = 0, 1, 2, \ldots \tag{4.26}$$

with $C_u, \gamma_u > 0$, then there exists an interpolant $\tilde{I}(\hat{u}) \in P_{\mathbf{p}(K)}(\hat{K})$ such that

1. *$\tilde{I}(\hat{u})|_{e_i} = GL(p_i, \hat{u}|_{e_i})$,*
2. *$\tilde{I}(\hat{u})|_{e_j} = GL(p_j, \hat{u}|_{\tilde{e}_j})|_{e_j}$,*
3. *$\| \tilde{I}(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq \tilde{C}(\alpha) C_u e^{-bp_m}$,*

where b depends on γ_u, γ_2. The constant $\tilde{C}(\alpha)$ is at most $\mathcal{O}(\alpha^6)$ for $\alpha \to \infty$.

Let us comment on the impact of Theorem 4.3.10. Because of property 1. and 2., it is possible to construct a complete interpolant in an element by element fashion. Together with Theorem 4.3.7 it is guaranteed that the resulting function is continuous across each edge and, therefore, lies in the conforming finite element space $S^{\mathbf{p}}(\tau_h)$. This is possible as the definition of the finite element space enforces that the polynomial degree on a constrained edge coincides with the polynomial degree on the corresponding coarse edge.

Proof. We define the piecewise Gauss-Lobatto interpolation operator as

$$\begin{aligned} \tilde{i} : C(\partial\hat{K} \cup \bigcup_j \tilde{e}_j) \to \{ f \in C(\partial\hat{K}) : \; f_i|_{e_i} \text{ is polynomial of degree } p_i \}, \\ \tilde{i}(\hat{u})(x) = GL(p_i, \hat{u}|_{e_i})(x), \quad x \in e_i, \\ \tilde{i}(\hat{u})(x) = GL(p_j, \hat{u}|_{\tilde{e}_j})|_{e_j}, \quad x \in e_j. \end{aligned}$$

The function $\hat{u} = u \circ F_K$ can also be evaluated at points outside of $\hat{K}$ since the mapping F_K is analytic. Thus, the Gauss-Lobatto interpolation on $\tilde{e}$ is well defined.

With the operators defined in the proof of Theorem 4.3.7 we define the interpolation operator as

$$\tilde{I} = E(\tilde{i}(\hat{u})) - \Pi(\hat{u} - E(i(\hat{u}))).$$

We compute

$$\| \tilde{I}(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} \leq \| I(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} + \| I(\hat{u}) - \tilde{I}(\hat{u}) \|_{W^{1,\infty}(\hat{K})},$$

where I is given by Theorem 4.3.7. The first addend is bounded by $C_\alpha C_u e^{-bp_m}$ (Theorem 4.3.7). So we only need to estimate the second one. Using Lemma 4.3.6 we find

$$\begin{aligned}
&\| I(\hat{u}) - \tilde{I}(\hat{u}) \|_{W^{1,\infty}(\hat{K})} \\
&\quad = \| E(i(u)) - E(\tilde{i}(u)) \|_{W^{1,\infty}(\hat{K})} \leq cp^2 \| i(\hat{u}) - \tilde{i}(\hat{u}) \|_{L^\infty(\partial \hat{K})} \\
&\quad = cp^2 \| \sum_j GL(p_j, \hat{u}|_{e_j}) - GL(p_j, \hat{u}|_{\tilde{e}_j})|_{e_j} \|_{L^\infty(e_j)} \\
&\quad \leq cp^2 \sum_j \left(\| GL(p_j, \hat{u}|_{e_j}) - \hat{u} \|_{L^\infty(e_j)} + \| GL(p_j, \hat{u}|_{\tilde{e}_j})|_{e_j} - \hat{u} \|_{L^\infty(e_j)} \right). \qquad (4.27)
\end{aligned}$$

The first addends are bounded due to (4.25) and Lemma 4.3.9.

$$\sum_j \| GL(p_j, \hat{u}|_{e_j}) - \hat{u} \|_{L^\infty(e_j)} \leq \sum_j CC_u e^{-b_1 p_j} \leq 4CC_u e^{-b_1 p_m}. \qquad (4.28)$$

If we use an affine mapping from $\tilde{e}_j$ to $[-1, 1]$, the prerequisite (4.26) transforms into

$$\| \nabla^{q+2} \hat{u} \|_{L^\infty(-1,1)} \leq C_u (2\gamma_u)^q q!, \quad q = 0, 1, 2, \dots .$$

Using again Lemma 4.3.9 we find

$$\begin{aligned}
\sum_j \| GL(p_j, \hat{u}|_{\tilde{e}_j})|_{e_j} - \hat{u} \|_{L^\infty(e_j)} &\leq \sum_j \| GL(p_j, \hat{u}|_{\tilde{e}_j}) - \hat{u} \|_{L^\infty(\tilde{e}_j)} \\
&\leq \sum_j CC_u e^{-b_2 p_j} \leq 4CC_u e^{-b_2 p_m}.
\end{aligned}$$

Due to the inequality $cp^2 \leq cC(\alpha)^2$, the final estimate reads

$$\begin{aligned}
\| \tilde{I}(\hat{u}) - \hat{u} \|_{W^{1,\infty}(\hat{K})} &\leq C_\alpha C_u e^{-bp_m} + cC(\alpha)^2 (CC_u e^{-b_1 p_m} + CC_u e^{-b_2 p_m}) \\
&\leq \tilde{C}_\alpha C_u e^{-\tilde{b} p_m}
\end{aligned}$$

with $\tilde{C}_\alpha$ depending on α, γ_u and $\tilde{b}$ on γ_u. As $C_\alpha \in \mathcal{O}(\alpha^6)$, it follows $\tilde{C}_\alpha \in \mathcal{O}(\alpha^6)$. □

Remark 4.3.11. *Note that the interpolation operator projects $\hat{u} - E(i(\hat{u}))$ instead of $\hat{u} - E(\tilde{i}(\hat{u}))$ onto the subspace V of polynomials vanishing at the boundary of the element. This simplifies the interpolation error estimates (see (4.27)).*

The Final Error Estimate

First we establish an easy lemma to conveniently check the prerequisites of Theorem 4.3.7 and 4.3.10.

Lemma 4.3.12. *Let u be a function on Ω that satisfies*

$$\| \nabla^q u \|_{L^2(\Omega)} \leq C_u \gamma_u^q q!, \quad q = 0, 1, 2, \dots . \tag{4.29}$$

Then u is analytic on $\overline{\Omega}$ and scaling constants $C_s, c_s > 0$ exist such that

$$\| \nabla^q u \|_{C(\overline{\Omega})} \leq C_s C_u (c_s \gamma_u)^q q!, \quad q = 0, 1, 2, \dots .$$

Proof. For an arbitrary but fixed q, we have $\nabla^q u \in H^2(\Omega)$. A Sobolev embedding implies

$$\| \nabla^q u \|_{C(\overline{\Omega})} \leq C \| \nabla^q u \|_{H^2(\Omega)}.$$

Estimating each derivative of u appearing in the $H^2(\Omega)$-norm separately with (4.29) yields

$$\| \nabla^q u \|_{C(\overline{\Omega})} \leq C(1 + \gamma_u + \gamma_u^2) C_u \gamma_u^q (q+2)!.$$

Choosing $C_s := 2C(1 + \gamma_u + \gamma_u^2)$ and $c_s = 6$, which implies $c_s^q \geq (q+2)(q+1)$ for $q \geq 1$, proves the estimate, which in turn gives analyticity of u on $\overline{\Omega}$. □

The proof of the following theorem is inspired by [88, Proposition 2.10].

Theorem 4.3.13. *Let τ_h be a bc-mesh on Ω with a linear degree vector $\mathbf{p}$ of slope α. Let $u \in B^2_{1-\sigma}(\Omega, C_u, \gamma_u)$ for some $\sigma \in (0, 1]$ and $r = r_\Gamma$. Then it holds for sufficiently large α*

$$\inf \left\{ \| u - v \|_{H^1(\Omega)} \ : \ v \in S^{\mathbf{p}}(\tau_h) \right\} \leq C \, C_u \, h^\sigma .$$

Here, C depends on Ω, γ_u, α and the shape regularity constant γ but not on C_u. The choice of α also depends on all these constants but not on C_u.

We want to construct the interpolant element by element. On elements abutting the boundary we will use the linear interpolant because the linear degree vector does not allow larger polynomial degrees on elements of size h.

For elements not abutting the boundary we want to take advantage of the increased polynomial degree to achieve good approximation quality. The previous error estimates of the interpolants, however, depend on the minimal polynomial degree p_m which is determined by at least one neighbor element. To guarantee that the neighbor's polynomial degree (and thus p_m) can be increased sufficiently, we introduce a second layer of elements near the boundary.

Proof. Overall we distinguish the following cases:

1. Elements K collected in τ_b abutting the boundary, i.e., $\overline{K} \cap \partial\Omega \neq \emptyset$.

2. Elements in the 'second' layer near the boundary, i.e., $K \in \tau_h$ such that $\overline{K} \cap \partial\Omega = \emptyset$ and $\exists K' \in \tau_h$ with $\overline{K} \cap \overline{K}' \neq \emptyset$, $\overline{K}' \cap \partial\Omega \neq \emptyset$. These elements are collected in τ_s.

3. Elements without hanging nodes which do not belong to $\tau_b \cup \tau_s$. They are collected in τ_f (*f*ree elements).

4. Elements that do not fall into the previous categories, i.e., elements with hanging (*c*onstrained) nodes which do not belong to $\tau_b \cup \tau_s$. They form the set τ_c.

Let $u \in B^2_{1-\sigma}(C_u, \gamma_u)$. In the following, we simply write r for r_Γ. For an element K, we define the constant C_K by

$$C_K^2 = \sum_{q=0}^{\infty} \frac{1}{(2\gamma_u)^{2q}(q!)^2} \| r^{q+1-\sigma}\nabla^{q+2}u \|^2_{L^2(K)}.$$

It holds

$$\| r^{q+1-\sigma}\nabla^{q+2}u \|_{L^2(K)} \leq C_K(2\gamma_u)^q q!, \tag{4.30}$$

$$\sum_{K\in\tau_h} C_K^2 \leq \frac{4}{3}C_u^2. \tag{4.31}$$

Additionally, we define

$$\tilde{C}_K^2 := C_K^2 + \sum_{K':\overline{K}\cap\overline{K}'\neq\emptyset} C_{K'}^2,$$

which implies $\sum_{K\in\tau_h} \tilde{C}_K^2 \leq (c+1)\frac{4}{3}C_u^2$, where c is a general upper bound for the amount of neighbors that an element $K \in \tau_h$ can have.

We construct an interpolant $u_h \in S^{\mathbf{p}}(\tau_h)$ of u for each element K falling into one of the four categories above. In the sequel, the index q will always be from $\mathbb{N}_0$.

1. $K \in \tau_b$. Let I_{lin} denote the linear or bilinear interpolation. We set $u_h|_K := I_{lin}u|_K$. We use [88, Appendix B.4] and the property 1. of Definition 4.3.1 to obtain

$$\| u - u_h \|_{H^1(K)} \leq \| u - I_{lin}(u) \|_{H^1(K)} \leq Ch_K^\sigma \| r^{1-\sigma}\nabla^2 u \|_{L^2(K)} \leq Ch^\sigma C_K.$$

3. $K \in \tau_f$. The pullback $\hat{u}$ of u on $\hat{K}$ satisfies

$$\begin{aligned} \| \nabla^{q+2}\hat{u} \|_{L^2(\hat{K})} &\leq Ch_K^{q+1} \| \nabla^{q+2}u \|_{L^2(K)} \\ &\leq Ch_K^{q+1} \| r^{q+1-\sigma}\nabla^{q+2}u \|_{L^2(K)} \frac{1}{\inf_{x\in K} r(x)^{q+1-\sigma}}. \end{aligned} \tag{4.32}$$

Since $r(x)$ for $x \in K$ is bounded from below by the diameter of the largest inscribed circle of a neighboring element, γ-shape-regularity yields

$$\inf_{x\in K} r(x) \geq \tilde{c}(\gamma)h_K$$

for a $\tilde{c}(\gamma) > 0$. Consequently,

$$\| \nabla^{q+2}\hat{u} \|_{L^2(\hat{K})} \leq CC_K h_K^\sigma (2\tilde{c}\gamma_u)^q q!,$$

where C is possibly rescaled by $\tilde{c}(\gamma)$.

We set $u_h|_K := I(\hat{u}) \circ F_K^{-1}$, where I is given by Theorem 4.3.7. Owing to Lemma 4.3.12, we can apply Theorem 4.3.7 and get

$$\| u - u_h \|_{H^1(K)} \leq C_\alpha C C_K h_K^\sigma e^{-bp_{m,K}}$$

with b, C_α given by Theorem 4.3.7 depending on γ_u but not on C_u and K. Using

$$p_m = p_{K'} \geq c\alpha \ln(h_{K'}/h)$$

for a neighbor K' of element K, we arrive at

$$\| u - u_h \|_{H^1(K)} \leq C_\alpha C C_K h_{K'}^{\sigma - \alpha b} h^{\alpha b}.$$

Using $h_{K'} \geq ch$ yields

$$h_{K'}^{\sigma - \alpha b} h^{\alpha b} \leq h^{\min\{\sigma, \alpha b\}}.$$

4. $K \in \tau_c$. We set $\hat{K} := F_K^{-1}(K)$ and denote the edges of $\hat{K}$ that possess a hanging node by e_j, $j \in \{1, \dots, 4\}$. The coarse edge that contains e_j is denoted by $\tilde{e}_j$ in reference coordinates. Let K_j denote the neighboring element of K that contains the same hanging node, i.e., $\overline{K}_j \cap F_K(\tilde{e}_j) \neq \emptyset$, and set $\hat{K}_j := F_K^{-1}(K_j)$. For an illustration see Figure 4.6.

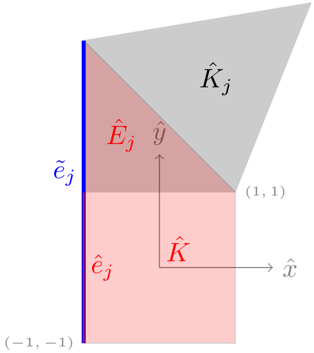


Figure 4.6. A reference element $\hat{K}$ enlarged to $\hat{E}_j$ to handle a hanging node.

In order to apply Theorem 4.3.10, we have to estimate L^∞-norms of the pullback on the extended edge $\tilde{e}_j$. With the properties of the elements in τ_c we deduce

$$\| \nabla^{q+2} \hat{u} \|_{L^\infty(\tilde{e}_j)} \leq \| \nabla^{q+2} \hat{u} \|_{C(\tilde{e}_j)} \leq C \| \nabla^{q+2} \hat{u} \|_{C(\hat{E}_j)}$$

with $\hat{E}_j = \operatorname{int} \operatorname{conv}(\hat{K} \cup \hat{e}_j) \subset \hat{K} \cup \hat{K}_j$. Let us emphasize that the constant C depends on $\tilde{E}_j$ but not on $\tilde{K}_j$. Hence, C is independent of K_j, and thus it is independent of the mesh.

Since $h_{K_j'}$ and h_K are comparable (Lemma 4.3.4), we deduce analogously to (4.32)

$$\| \nabla^{q+2}\hat{u} \|_{L^2(\hat{K}_j)} \le C C_{K_j} h_K^{\sigma} (2\gamma_u)^q q! \tag{4.33}$$

with a possibly larger constant C independent of K, K_j. The two estimates (4.32) and (4.33) yield

$$\| \nabla^{q+2} u \|_{L^2(\hat{E}_j)} \le C(C_{K_j} + C_K) h_K^{\sigma} (2\gamma_u)^q q!$$

and Lemma 4.3.12 shows that the prerequisites for Theorem 4.3.10 are fulfilled. So we set $u_h|_K := \tilde{I}(\hat{u}) \circ F_K^{-1}$ with $\tilde{I}$ given by Theorem 4.3.10. The result of this theorem yields

$$\begin{aligned} \| u - u_h \|_{H^1(K)} &\le \tilde{C}_\alpha C \left(C_K + \sum_{K' : \overline{K} \cap \overline{K}' \neq \emptyset} C_{K'} \right) h_K^{\sigma} e^{-b p_{m,K}} \\ &= \tilde{C}_\alpha C \tilde{C}_K h_K^{\sigma} e^{-b p_{m,K}}. \end{aligned}$$

Arguing as in the case $K \in \tau_f$, we find

$$\| u - u_h \|_{H^1(K)} \le \tilde{C}_\alpha C \tilde{C}_K h^{\min\{\sigma, \alpha b\}}.$$

2. $K \in \tau_s$. Here, we set $u_h|_K := I(\hat{u}) \circ F_K^{-1}$ if K has no hanging nodes or $u_h|_K := \tilde{I}(\hat{u}) \circ F_K^{-1}$ otherwise. Analogously as in the cases $K \in \tau_f$, $K \in \tau_c$, we obtain

$$\| u - u_h \|_{H^1(K)} \le \tilde{C}_\alpha C \tilde{C}_K h_K^{\sigma} e^{-b p_{m,K}}.$$

However, we cannot apply $p_m \ge \alpha \ln(h_K/h)$ because $p_m = 1$, and thus p_m is fixed and cannot be increased. In bc-meshes, the element size h_K is proportional to the size of a neighboring element. In the second layer, there is a neighbor abutting the boundary, so we find $C(\gamma)^{-1} h \le h_K \le C(\gamma) c_2 h$. Thus, we obtain for a possibly adapted C

$$\| u - u_h \|_{H^1(K)} \le \tilde{C}_\alpha C \tilde{C}_K h^{\sigma}.$$

Overall we now estimate

$$\begin{aligned} \sum_{K \in \tau_h} \| u - u_h \|_{H^1(K)}^2 \le C^2 \Big(& \sum_{K \in \tau_b} C_K^2 h^{2\sigma} + \tilde{C}_\alpha^2 \sum_{K \in \tau_s} \tilde{C}_K^2 h^{2\sigma} \\ & + C_\alpha^2 \sum_{K \in \tau_f} C_K^2 h^{2\min\{\sigma, \alpha b\}} + \tilde{C}_\alpha^2 \sum_{K \in \tau_c} \tilde{C}_K^2 h^{2\min\{\sigma, \alpha b\}} \Big). \end{aligned}$$

As b is independent of α, we can choose α large enough to obtain

$$\sum_{K \in \tau} \| u - u_h \|_{H^1(K)}^2 \le C^2 \, C_u^2 \, h^{2\sigma}.$$

By construction u_h is a continuous function on $\overline{\Omega}$. Thus, it holds $u_h \in H^1(\Omega)$ and

$$\| u - u_h \|_{H^1(\Omega)} \leq C\, C_u\, h^\sigma .$$

□

Remark 4.3.14. *The proof only works for affine linear or bilinear mappings F_K. The reason is that prolonged edges of the reference element have to be straight lines under F_K, so that in global coordinates they coincide with the coarse edges. Together with the property that hanging nodes are in the middle of a coarse edge, the described procedure and usage of interpolation operators works.*

Just as in Section 3.2 we can extend this result to $2d$-networks. Since ic-FEM is the local application of bc-FEM on each subdomain, we deduce the following corollary.

Corollary 4.3.15. *Let τ_h be a ic-mesh on the $2d$-network $\{\Omega_i\}_{i\in I}$ with a linear degree vector $\mathbf{p}$ of slope α. Let $u \in B^2_{1-\sigma}(\Omega, C_u, \gamma_u)$ for some $\sigma \in (0,1]$ and $r = r_{\mathcal{I}\cup\Gamma}$. Then it holds for sufficiently large α*

$$\inf \Big\{ \| u - v \|_{H^1(\Omega)} \ :\ v \in S^{\mathbf{p}}(\tau_h) \Big\} \leq C\, C_u\, h^\sigma .$$

Here, C depends on Ω, γ_u, α and the shape regularity constant γ but not on C_u. The choice of α also depends on all these constants but not on C_u.

The corollary can be shown by replacing r_Γ with $r_{\mathcal{I}\cup\Gamma}$ in the previous proof.

4.3.2 Lebesgue Norm Estimates at the Boundary

It is well known from h-FEM that the error in the L^2-norm decays twice as fast than in the energy-norm. This can be shown by the well known Nitsche trick ([34]). Furthermore, uniform estimates in the L^∞-norm are available. Inspired by these results, we are going to establish similar error estimates for the bc-FEM.

The Nitsche Trick

Theorem 4.3.16. *Let τ_h be a bc-mesh on Ω with mesh size h and $\mathbf{p}$ be a linear degree vector of slope α. Let $y \in H^{1+\sigma}(\Omega)$ with $\sigma \in (0,1]$ be the weak solution of the Neumann problem (N), i.e.,*

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

which satisfies Assumption 2.2.1. Let $f \in B^0_{1-\sigma}(C_f, \gamma_f)$ for $C_f, \gamma_f > 0$ and $r = r_\Gamma$. Denote by $y_h \in S^{\mathbf{p}}(\tau_h)$ the FE solution of the corresponding discretized problem. If α is sufficiently large and the stability estimate

$$\| y \|_{H^{3/2}(\Omega)} \leq C(\| f \|_{L^2(\Omega)} + \| u \|_{L^2(\Gamma_{\mathcal{N}})})$$

holds, then there is a constant $C > 0$ independent of h and y such that

$$\|y - y_h\|_{L^2(\Gamma_{\mathcal{N}})} \leq Ch^{\sigma+\frac{1}{2}} \left(C_f + \|y\|_{H^{1+\sigma}(\Omega)}\right).$$

Proof. We proof this result by the standard Nitsche trick. Let z denote the solution of the dual problem

$$\begin{aligned} -\nabla \cdot (D(x)\nabla z) + c(x)z &= 0 && \text{in } \Omega, \\ z &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ \partial_{n_D} z &= y - y_h && \text{on } \Gamma_{\mathcal{N}} \end{aligned}$$

with $z_h \in S^{\mathbf{p}}(\tau_h)$ being its bc-FEM approximation.

Then it holds by Galerkin-orthogonality

$$\|y - y_h\|^2_{L^2(\Gamma_{\mathcal{N}})} = a(z, y - y_h) = a(z - z_h,\, y - y_h) = a(z - z_h,\, y - I_h y),$$

where I_h is the bc-FEM interpolation operator from Theorem 4.3.13. According to Theorem 4.3.13 the solution y satisfies

$$\|r^{p+1-\sigma}\nabla^{p+2} y\|_{L^2(\Omega)} \leq C_y \gamma_y^p p! \left(C_f + \|y\|_{H^{1+\sigma}(\Omega)}\right) \quad \forall p \in \mathbb{N}_0$$

with $C_y, \gamma_y > 0$ independent of u. By Theorem 4.3.13 we obtain the interpolation error estimate

$$\|y - I_h y\|_{H^1(\Omega)} \leq C\, C_y \left(C_f + \|y\|_{H^{1+\sigma}(\Omega)}\right) h^\sigma$$

holds for sufficiently large α. The solution z of the dual problem satisfies

$$\|r^{p+1-\sigma}\nabla^{p+2} z\|_{L^2(\Omega)} \leq C_z \gamma_z^p p! \|z\|_{H^{1+\sigma}(\Omega)} \quad \forall p \in \mathbb{N}_0$$

with $\sigma = \frac{1}{2}$ and $C_z, \gamma_z > 0$ independent of $y - y_h$, cf. Theorem 4.3.13.

With the same arguments as above and applying Cea's lemma as well as Theorem 4.3.13 we conclude

$$\|z - z_h\|_{H^1(\Omega)} \leq C\|z - I_h z\|_{H^1(\Omega)} \leq C\, C_z \|z\|_{H^{3/2}(\Omega)} h^{1/2}$$

with $C_z > 0$ independent of $y - y_h$ and sufficiently large α, where the choice of α is independent of $y - y_h$ and thus independent of the discretization. Using the stability assumption in the $H^{3/2}$-norm implies

$$\|z - z_h\|_{H^1(\Omega)} \leq C\, C_z \|z\|_{H^{3/2}(\Omega)} h^{\frac{1}{2}} \leq C\, C_z\, h^{\frac{1}{2}} \|y - y_h\|_{L^2(\Gamma_{\mathcal{N}})}.$$

Hence, we obtain the estimate

$$\begin{aligned} \|y - y_h\|^2_{L^2(\Gamma_{\mathcal{N}})} &\leq C\|z - z_h\|_{H^1(\Omega)} \|y - I_h y\|_{H^1(\Omega)} \\ &\leq Ch^{\sigma+\frac{1}{2}} \left(C_f + \|y\|_{H^{1+\sigma}(\Omega)}\right) \|y - y_h\|_{L^2(\Gamma_{\mathcal{N}})}, \end{aligned}$$

which ends the proof. □

Unfortunately, this technique cannot be applied to obtain an enhanced estimate in $L^2(\Omega)$. The dual problem would contain $y - y_h$ as a source term, which no longer lies in $B^2_{1-\sigma}$ and prevents Theorem 4.3.13 from being applicable.

Using the Bramble-Hilbert Lemma

Lemma 4.3.17 (Bramble-Hilbert). *Let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz boundary. Futher, let $L \in \mathcal{L}(H^k(\Omega), V)$ be a bounded, linear map from $H^k(\Omega)$ to a normed space V with $k \geq 2$. If the polynomials up to degree $k-1$ are a subset of $\ker L := \{v \in H^k(\Omega) : L(v) = 0\}$, there is a $C > 0$ depending on Ω and L such that*

$$\| Lv \| \leq C|v|_{H^k(\Omega)} \quad \forall v \in H^k(\Omega).$$

See [34, Lemma II.6.3] for the proof. We use the lemma with the domain $\hat{K}$ and set $k = 2$, $L := \operatorname{id} - I_{lin}$ with the linear interpolator operator I_{lin} and obtain

$$\| v - I_{lin}(v) \|_{H^2(\hat{K})} \leq C|v|_{H^2(\hat{K})}. \tag{4.34}$$

For an L^∞-estimate on the boundary, we need two further estimates. First, for $1 \leq p < \infty$ and domains with Lipschitz boundary, there is a constant $C > 0$ such that (see [35, Theorem 1.6.6])

$$\| v \|_{L^p(\Gamma)} \leq C\| v \|_{L^p(\Omega)}^{1-1/p} \| v \|_{W^{1,p}(\Omega)}^{1/p} \quad \forall v \in W^{1,p}(\Omega). \tag{4.35}$$

Second, let τ_h be a quasi-uniform triangulation of a Lipschitz domain $\Omega \subset \mathbb{R}^2$, then the linear interpolation operator I_{lin} satisfies

$$\left(\sum_{K \in \tau} \| v - I_{lin}v \|_{H^l(K)}^2 \right)^{1/2} \leq Ch^{2-l}|v|_{H^2(\Omega)} \quad \forall v \in H^2(\Omega),\ l = 0, 1, 2, \tag{4.36}$$

for a $C > 0$, independent of h. This result is well known from approximation theory and proved in, e.g., [35, Theorem 4.4.20], [34, Theorem II.6.7].

Theorem 4.3.18. *Let the Neumann problem (N) be H^2-regular and satisfy Assumption 2.2.1. Denote by y the solution to (N) and by $y_h \in S^{\mathbf{p}}(\tau_h)$ its FE approximation on a bc-mesh with linear polynomial degree vector of sufficiently large slope. Then*

$$\| y - y_h \|_{L^\infty(\Gamma)} \leq Ch.$$

The proof is an adaptation of [34].

Proof. First observe that the bc-FEM interpolator in the proof of Theorem 4.3.13 coincides with the linear interpolator I_{lin} on all elements K abutting the boundary Γ. The continuous embedding $H^2(K) \hookrightarrow C(K)$ (Theorem 2.1.3), and (4.34) yield

$$\| y - I_{lin}y \|_{L^\infty(\hat{K})} \leq C\| y - I_{lin}y \|_{H^2(K)} \leq c|y|_{H^2(\hat{K})} \quad \forall K \in \tau_h.$$

We set

$$\Omega_h := \bigcup \{K \in \tau_h : \overline{K} \cap \Gamma \neq \emptyset\}.$$

As τ_h is quasi-uniform of size h at the boundary (see Definition 4.3.1), a scaling argument yields

$$\begin{aligned}\| y - I_{lin}y \|_{L^\infty(K\cap\Gamma)} &\leq \| y - I_{lin}y \|_{L^\infty(K)} = \| \hat{y} - I_{lin}\hat{y} \|_{L^\infty(\hat{K})} \\ &\leq c|\hat{y}|_{H^2(\hat{K})} \leq ch|y|_{H^2(K)} \\ &\leq ch|v|_{H^2(\Omega_h)} \quad \forall K \subset \Omega_h.\end{aligned}$$

Taking the maximum over the boundary elements leads to

$$\| y - I_{lin}y \|_{L^\infty(\Gamma)} \leq ch|v|_{H^2(\Omega_h)}. \tag{4.37}$$

We use the scaling argument again and exploit the equivalence of norms in finite-dimensional spaces (with a constant c_f), i.e.,

$$\begin{aligned}\| v_h \|_{L^\infty(K\cap\Gamma)} = \| \hat{v}_h \|_{L^\infty((-1,1))} &\leq c_f \| \hat{v}_h \|_{L^2((-1,1))} \\ &\leq c_f (2h)^{-1/2} \| v_h \|_{L^2(K\cap\Gamma)} \quad \forall v_h \in S^{\mathbf{p}}(\tau_h).\end{aligned}$$

Hence, the following inverse estimate holds (see also [35, Theorem 4.5.11]):

$$\| v_h \|_{L^\infty(\Gamma)} \leq c_f (2h)^{-1/2} \| v \|_{L^2(\Gamma)} \quad \forall v_h \in S^{\mathbf{p}}(\tau_h). \tag{4.38}$$

Putting (4.37) and (4.38) together leads to

$$\begin{aligned}\| y - y_h \|_{L^\infty(\Gamma)} &\leq \| y - I_{lin}y \|_{L^\infty(\Gamma)} + \| y_h - I_{lin}y \|_{L^\infty(\Gamma)} \\ &\leq ch|y|_{H^2(\Omega_h)} + c_f (2h)^{-1/2} \| y_h - I_{lin}y \|_{L^2(\Gamma)} \\ &\leq ch|y|_{H^2(\Omega_h)} + c_f (2h)^{-1/2} (\| y_h - y \|_{L^2(\Gamma)} + \| y - I_{lin}y \|_{L^2(\Gamma)}).\end{aligned} \tag{4.39}$$

Using Theorem 4.3.16, (4.35), and (4.36) yields

$$\begin{aligned}&\| y_h - y \|_{L^2(\Gamma)} + \| y - I_{lin}y \|_{L^2(\Gamma)} \\ &\qquad \leq Ch^{3/2}(C_f + \| y \|_{H^2(\Omega)}) + C\| y - I_{lin}y \|_{L^2(\Omega)}^{1/2} \| y - I_{lin}y \|_{H^1(\Omega)}^{1/2} \\ &\qquad \leq Ch^{3/2}(C_f + \| y \|_{H^2(\Omega)}) + Ch^{3/2}|y|_{H^2(\Omega)},\end{aligned} \tag{4.40}$$

where C is a generic constant independent of h. Inserting (4.40) into (4.39) concludes the proof. □

Remark 4.3.19. *Note that (4.36), which is used in the previous proof, requires a quasi-uniform discretization of Ω. Since $\Omega_h|_\Gamma$ is quasi-uniform, we can (virtually) extend Ω_h to a quasi-uniform triangulation of whole Ω and mesh size h*

4.4 Exponential Convergence of the Vertex Concentrated Finite Element Method

The results of this section are formulated for quadrilaterals but extend also to triangular elements (see [136]). We follow the exposition of [135] which gives a self-contained outline of the hp-FEM and contains the famous result on exponential convergence with respect to the number of unknowns (originally due to [9]).

Definition 4.4.1. *An* ***irregular geometric mesh patch*** $\hat{\tau}_\varsigma^m$ *(with* $m+1$ ***layers*** *and* ***grading factor*** $\varsigma \in (0,1)$*) is an (admissible) triangulation of* $(0,1)^2$ *which is defined recursively. If* $m = 0$*,* $\hat{\tau}_\varsigma^0 := (0,1)^2$*. For a given* $\hat{\tau}_\varsigma^m$*, generate* $\hat{\tau}_\varsigma^{m+1}$ *by subdividing* $K \in \hat{\tau}_\varsigma^m$ *containing the point* $(0,0)$ *into four smaller rectangles. The refinement is achieved by dividing the sides of element* K *in a* $\varsigma/(1-\varsigma)$ *ratio.*

An exemplary irregular geometric mesh patch is depicted in Figure 4.7.

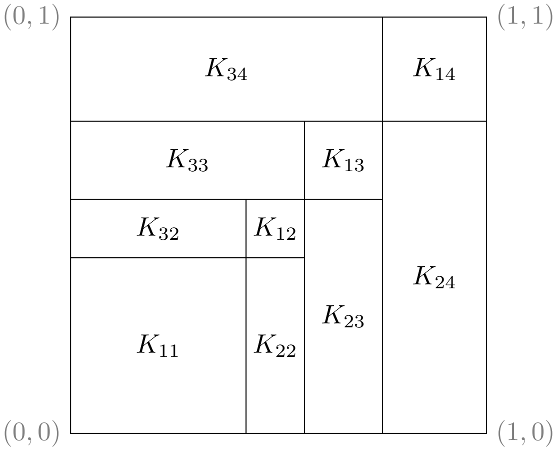


Figure 4.7. An irregular geometric mesh patch $\hat{\tau}_\varsigma^m$ with $m = 3$, $\varsigma = 0.75$.

Regular geometric mesh patches can be obtained by refining those elements which possess a hanging node as a mid-side node into triangles.

Definition 4.4.2. *An* ***(irregular) geometric mesh*** τ_ς^m *of* Ω *with* $m+1$ *layers is an admissible triangulation that is obtained by linearly mapping a combination of geometric mesh patches to a finite set of points* $\mathcal{V} \subset \overline{\Omega}$*. The possibly remaining part of* Ω *is meshed with finitely many quadrilaterals.*

It remains to specify the polynomial degree vector $\mathbf{p}$.

Definition 4.4.3. *A polynomial degree vector* $\mathbf{p}$ *to the triangulation* $\hat{\tau}_\varsigma^m = \{K_{ij} \,:\, i = 1,2,3,\ 1 \leq j \leq m+1\}$ *(see Figure 4.7 for the numbering) is called* ***linear*** *with* ***slope*** $\alpha > 0$ *if and only if* $\mathbf{p} = (p_{K_{ij}})_{K_{ij} \in \hat{\tau}_\varsigma^m}$ *satisfies*

$$p_{K_{11}} = 1, \quad p_{K_{ij}} = \max\{2, \lfloor \alpha j \rfloor\},\ j > 1$$

with the numbering K_{ij} *according to Figure 4.7. The number of layers is hereby proportional to the maximal polynomial degree, i.e.,* $m \sim |\mathbf{p}|_\infty$.

Irregular geometric meshes are defined by heavily h-refining a mesh towards $\mathcal{V}$ which will always contain the vertices $\mathcal{X}$ of Ω in the following. Therefore, we speak of *vertex concentrated (vc)* finite elements, in resemblance of the boundary concentrated (*bc*) FEM.

The approximation error for this version of the hp-FEM decays exponentially with respect to the number of unknowns.

Theorem 4.4.4. *Let* $\Omega \subset \mathbb{R}^2$ *be a polygonal domain and* $u \in B^2_\beta(\Omega, C_u, \gamma_u)$ *with* $r_\mathcal{V}$ *and a multi-index* $\beta \in (0,1)$. *Let* τ_ς^m *be a geometric mesh whose polynomial degree vector* $\mathbf{p}$ *is linear with sufficiently large slope* α *(assumed identical in each geometric mesh patch). Then there exist constants* $C, b > 0$ *such that*

$$\inf\left\{\| u - v \|_{H^1(\Omega)} \ :\ v \in S^{\mathbf{p}}(\tau)\right\} \leq Ce^{-b\sqrt[3]{N}},$$

where $N := \dim(S^{\mathbf{p}}(\tau))$. *The constants* C, b *are independent of* N.

The proof can be found in [9] or [135, Theorem 4.63]. It is possible to choose ς to maximize the constant $b = b(\varsigma)$, in order to achieve optimal convergence in an asymptotic sense, confer [9, Remark 1]. All of our numerical experiments were conducted with $\varsigma = 0.5$.

CHAPTER 5

Numerical Investigations

In this chapter, we investigate the numerical solution of the model problem (**P**) with the hp-FEM. As mentioned before, higher order methods are efficient if they approximate functions with large elements of high polynomial degree in regions of high regularity, whereas small elements with low polynomial degree are used in regions of low regularity. The results of Chapter 3 and 4 answered the question regarding regularity and approximation quality of finite-dimensional FE spaces, respectively. Hence, we have laid the theoretical foundation for numerical investigations of control problems.

The chapter is organized as follows. We introduce a discretized model problem ($\mathbf{P}_h$) in Section 5.1 and show how the semi-smooth Newton method can be used to find an approximate solution.

In Section 5.2, we carry over the convergence rates of the bc- and vc-FEM to optimal control problems (Theorem 5.2.1 and 5.2.3, respectively). Additionally, we compare our higher order methods with the traditional h-FEM to judge the performance with respect to the number of unknowns. Since the vc-FEM is distinguished by exponential convergence, the other techniques are not expected to be able to compete because they only allow algebraic error estimates.

We point out Subsection 5.2.3, where the bc-FEM is applied to a problem with bang-bang character. This behavior may occur if $\nu = 0$ in the model problem. Building on L^∞-results, we derive an a-priori update strategy for the regularization parameter ν, which is sent to zero in an outer iteration of the semi-smooth Newton method.

Section 5.3 contains the proof of the convergence of the ic-FEM applied to interface control problems (Theorem 5.3.1). We also use the vc-FEM for solving this problem class.

Let us give an overview on existing results regarding the discretization of Neumann control problems and our contributions. Generally, the estimates from h-FEM are of the type

$$\|u^* - u_h^*\|_{L_2(\Gamma)} \leq Ch^s. \tag{5.1}$$

This estimate is proved in [39] with $s = 1$ for a piecewise constant approximation u_h^* of u^* in the case of a convex domain. A piecewise linear discretization yields (5.1) with $s = 3/2 - \epsilon$, as proved in [38]. The variational discretization with low order elements allows $s = 3/2$ and an L^∞-estimate with $s = 2$ including the additional factor $|\log h|$ for smooth domains (see [80]). We also mention [104], which contains approximation results with $s \in [1, 2]$ depending on the angles of a (possibly non-convex) domain Ω. In the convex case, the rate of $s = 2 - 1/p$ can be shown if the optimal state y is in $W^{2,p}(\Omega)$. In the non-convex case and $D(x) \equiv I$, convergence rates with $1 < s < 1/2 + \pi/\omega$ are obtained, where ω is the largest inner angle of the domain. The collaboration [4] shows how the order $s = 3/2$ can be obtained for non-convex domains by using sufficiently graded meshes. All these results were obtained for finite elements with fixed polynomial degree for the discretization of the elliptic equation.

The bc-FEM (also investigated in [27, 28]) is an efficient technique because the number of unknowns N behaves like h^{-1} (Theorem 4.3.3), where h denotes the mesh size on the boundary (Definition 4.3.1). A uniform mesh of the classical h-FEM has $\mathcal{O}(h^{-d})$ unknowns for $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$. Loosely speaking, we can solve a two dimensional problem for the costs of a one dimensional one if we use the bc-FEM. Therefore, the bc-FEM is superior to many common h-discretizations.

Only recently, the authors of [5] show for suitably graded meshes that

$$\|u^* - u_h^*\|_{L_2(\Gamma_\mathcal{N})} \leq Ch^2 |\ln h|^{3/2}$$

holds even for non-convex domains. This puts graded h-FEM into a competitive position, additionally because less regularity on the source term (only $L_2(\Omega)$) is needed. Their result is, however, restricted to the case of the Laplacian, where the exact structure of the singularities at the vertices of Ω is known (see Section 3.1), while our results remain valid for general elliptic operators. We refer the reader to the discussion of Table 5.1 - 5.3 for a further comparison of the h- and bc-FEM.

Finally, we mention the publications [27, 156, 158], where similar numerical experiments are conducted.

5.1 Variational Discretization

The optimal control problem (**P**) has theoretically been investigated in Chapter 2. Regularity results and approximation estimates have been established in Chapter 3 and 4, respectively. We now have the machinery to derive convergence results for a numerical approach. Let us define the following discrete optimal control problem:

$$(\mathbf{P}_h) \qquad \begin{cases} \text{minimize } J(u_h, y_h) := \dfrac{1}{2}\| y_h - y_d \|^2_{L^2(\Omega)} + \dfrac{\nu}{2}\| u_h \|^2_{L^2(U)} \\ \qquad\qquad \text{subject to} \\ \qquad\qquad a(y_h, v_h) = l_{u_h}(v_h) \quad \forall v_h \in V_h \\ \qquad\qquad\qquad u_h \in U_{ad}. \end{cases}$$

In our experiments, we will choose $V_h = S^{\mathbf{p}}(\tau)$ with different hp-spaces from the previous chapter. We emphasize that u_h still stems from U_{ad} and is not discretized. With the same standard arguments as for problem (**P**), we can derive existence, uniqueness, and first order necessary conditions for an optimal solution.

Theorem 5.1.1. *The discrete optimal control problem* ($\mathbf{P}_h$) *with* $\nu > 0$ *possesses a unique solution* $(u_h^*, y_h^*) \in U_{ad} \times V_h$*. There is an adjoint* $q_h^* \in V_h$ *such that the following optimality system holds:*

$$\begin{aligned} a(y_h^*, v_h) &= l_{u_h^*}(v_h) && \forall v_h \in V_h, && (5.2a)\\ (B^* q_h^* + \nu u_h^*, u - u_h^*)_{L^2(U)} &\geq 0 && \forall u \ \in U_{ad}, && (5.2b)\\ a^*(q_h^*, v_h) &= (y_h^* - y_d, v_h)_{L^2(\Omega)} && \forall v_h \in V_h. && (5.2c) \end{aligned}$$

Here, $a^*(\cdot,\cdot)$ *denotes the bilinear form corresponding to* A^**.*

Note that the optimality system (5.2) for ($\mathbf{P}_h$) is a discrete version of the first order necessary conditions (2.14) for the original problem ($\mathbf{P}$). Since the control has not been discretized in the above formulation, Theorem 2.3.5 allows to rewrite the variational inequality (5.2b) as

$$u_h^* = P_{U_{ad}} \left(-\frac{1}{\nu} B^* q_h^* \right) \qquad (5.3)$$

if $\nu > 0$.

This representation is the typical *variational discretization* due to [79]. The control $u_h^* \in U_{ad}$ is given implicitly by the discrete adjoint variable q_h^* and is not necessarily a 'member' of the chosen hp-space.

Definition 5.1.2. *Define* $S := A^{-1}B$ *and* $S^* := A^{-*}$ *as the solution operators to the state and adjoint equation, respectively. For a general FE space* V_h*, we denote by* S_h *and* S_h^* *the approximate solution operators, arising from the discrete version of the weak formulation of the state and adjoint equation, respectively (confer Section 4.1).*

We can solve (5.2a) and (5.2c) and eliminate q_h^* in (5.3). With Definition 5.1.2, this leads to

$$u_h^* = P_{U_{ad}} \left(-\frac{1}{\nu} B^* S_h^* (S_h u_h^* - y_d) \right). \qquad (5.4)$$

It is well known that (5.4) is semi-smooth and can be solved with a semi-smooth Newton-method ([146, 81]), which is equivalent ([84]) to a primal dual active set strategy (see [24]). Implementational details for this approach can be found in [28]. Note that some adaptations are necessary because we study problems with distributed observation ($y_d \in L^2(\Omega)$) instead of boundary observation ($y_d \in L^2(\Gamma_{\mathcal{N}})$).

Remark 5.1.3. *For Neumann control problems* ($U = \Gamma_{\mathcal{N}}$)*, the projection operator* $P_{U_{ad}}$ *acts on one dimensional elements and the result can be computed easily for elements of low order. In the case of distributed controls* ($U = \Omega$)*, functions that live on* $2d$*-elements have to be projected. The numerical implementation is straightforward for triangular elements of degree one but degree two or more seriously complicates the analysis (see [137]). This also holds true for quadrilateral elements of polynomial degree one or higher. Therefore, discretization techniques seek to retain linear elements at points where the control switches from active to inactive and vice versa.*

Generally, a globalization technique is required to guarantee convergence of the optimization algorithm because Newton's method only converges locally. Most of the following examples have large regions of convergence and the algorithm terminates successfully for a wide range of initial guesses. In the remaining cases, the (projected) gradient method (see, e.g., [81, 144]) produces iterates that are close enough to the solution and suitable for launching the semi-smooth Newton method.

The beauty of the solution strategy is the fact that the error in the state and control variable is mainly determined by accuracy of the FE discretization.

Theorem 5.1.4. *Denote by (u^*, y^*, q^*) and (u_h^*, y_h^*, q_h^*) the solutions to the optimal control problem* (**P**) *and* ($\mathbf{P}_h$)*, respectively. Let $y^h := S_h u^*$ and $q^h := S_h^*(y^* - y_d)$. Then the following estimate holds:*

$$\nu \| u^* - u_h^* \|_{L^2(U)}^2 + \| y^* - y_h^* \|_{L^2(\Omega)}^2 \leq \frac{1}{\nu} \| q^* - q^h \|_{L^2(U)}^2 + \| y^* - y^h \|_{L^2(\Omega)}^2. \tag{5.5}$$

In particular, there is a $C > 0$ independent of u^*, y^*, q^* such that

$$\| y^* - y_h^* \|_{H^1(\Omega)} \leq C \| u^* - u_h^* \|_{L^2(U)} + \| y^* - y^h \|_{H^1(\Omega)}, \tag{5.6}$$

$$\| q^* - q_h^* \|_{H^1(\Omega)} \leq C \| y^* - y_h^* \|_{L^2(\Omega)} + \| q^* - q^h \|_{H^1(\Omega)}. \tag{5.7}$$

Proof. The proof of inequality (5.5) is given in [80, Theorem 1]. Observe that

$$\begin{aligned} \| y^* - y_h^* \|_{H^1(\Omega)} &= \| S u^* - S_h u_h^* \|_{H^1(\Omega)} \\ &\leq \| (S - S_h) u^* \|_{H^1(\Omega)} + \| S_h (u^* - u_h^*) \|_{H^1(\Omega)}. \end{aligned} \tag{5.8}$$

The first addend is nothing but $\| y^* - y^h \|_{H^1(\Omega)}$. With the definition of y^h, it follows that

$$\begin{aligned} \| y^h - y_h^* \|_{H^1(\Omega)}^2 &\leq a(y^h - y_h^*, y^h - y_h^*) = a(S_h(u^* - u_h^*), y^h - y_h^*) \\ &= (u^* - u_h^*, y^h - y_h^*)_{L^2(U)} \leq C \| u^* - u_h^* \|_{L^2(U)} \| y^h - y_h^* \|_{H^1(\Omega)}. \end{aligned} \tag{5.9}$$

In the last step, we used the Cauchy-Schwarz inequality together with the trace theorem if $U = \Gamma_{\mathcal{N}}$ and the stronger energy norm if $U = \Omega$. Dividing (5.9) by $\| y^h - y_h^* \|_{H^1(\Omega)}$ and inserting the result into (5.8) proves (5.6). The remaining estimate (5.7) can be established similarly. □

We would like to emphasize that the estimate (5.5) is not robust with respect to $\nu \searrow 0$, which can be overcome by using L^∞- estimates (see [153, 154]). Similar estimates for more general optimal control problems can be found in [91, Theorem 2.2]. Theorem 5.1.4 bounds the error towards the optimal variables in terms of the general approximation error resulting from a FEM discretization of the state and adjoint equation.

Remark 5.1.5. *Note that the right hand side of* (5.5) *only uses L^2-norms whereas most of our approximation results in Chapter 4 use the energy norm. This gap can be closed with the trace theorem (or Theorem 4.3.16) and $\| \cdot \|_{L^2(\Omega)} \leq \| \cdot \|_{H^1(\Omega)}$. The latter estimate is the reason why the results in the following sections do not yield optimal rates when measuring the convergence speed in L^2. Proving enhanced results is expected to be connected with duality arguments in the spirit of Aubin-Nitsche. However, it is hard to transfer the ideas from the h-FEM because higher order techniques require high regularity of the data. We mention [27], where Theorem 4.3.16 was successfully used to prove faster rates for control problems with boundary observation, i.e., $y_d \in L^2(\Gamma_{\mathcal{N}})$. To the best of our knowledge, no optimal $L^2(\Omega)$ estimates are available for the bc-FEM.*

The best currently available L_2-estimate for $H^{1+\sigma}$-regular problems is proven in [58]. It is shown that for every compact $\Omega' \subset\subset \Omega$ there exists $\sigma' \in [0, \sigma]$ such that for all elements $K \subset\subset \Omega'$ the error estimate $\|y - y_h\|_{L_2(K)} \leq Ch^{\sigma+\sigma'}$ holds. However, σ' depends on Ω', and it is unclear under which conditions $\sigma = \sigma'$ can be proven.

5.2 Neumann Control Problems

We present the control problem for a series of numerical tests. The objective reads

$$\text{minimize } J(u,y) := \frac{1}{2}\| \, y - y_d \, \|^2_{L^2(\Omega)} + \frac{\nu}{2}\| \, u \, \|^2_{L^2(\Gamma_{\mathcal{N}})} + (e_q, y)_{L^2(\Gamma_{\mathcal{N}})} \tag{5.10a}$$

and is a modification of the target functional of (**P**). The state equation reads

$$\begin{aligned} -\nabla \cdot (D(x)\nabla y) + c(x)y &= f && \text{in } \Omega, \\ y &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ \partial_{n_D} y &= u + e_y && \text{on } \Gamma_{\mathcal{N}}. \end{aligned} \tag{5.10b}$$

The adjoint equation (2.14c) now becomes

$$\begin{aligned} -\nabla \cdot (D(x)\nabla q) + c(x)q &= y - y_d && \text{in } \Omega, \\ q &= 0 && \text{on } \Gamma_{\mathcal{D}}, \\ \partial_{n_D} q &= e_q && \text{on } \Gamma_{\mathcal{N}}. \end{aligned}$$

The optimization is subject to

$$u \in U_{ad} := \{ u \in L^2(\Gamma_{\mathcal{N}}) \mid u_a \leq u \leq u_b \text{ a.e. in } \Gamma_{\mathcal{N}} \}, \tag{5.10c}$$

where $u_a, u_b \in \mathbb{R}$ with $u_a \leq u_b$. If $e_q \equiv e_y \equiv 0$, we recover the *Neumann control problem*, i.e., (**P**) subject to (N). The analysis of the previous chapters is not affected by our modifications because

- the boundary conditions do not influence the regularity or approximation in the context of the bc-FEM,
- we ensure that e_q, e_y is smooth enough in the context of the vc-FEM.

Hence, we formulate the following convergence results for (**P**) subject to (N) and not for (5.10), which is only referenced during computational investigations. Generally, the inhomogeneities e_q, e_y can be used for the construction of test problems with a known optimal solution.

We assume throughout this subsection that (N) satisfies Assumption 2.2.1. If we speak of (u_h^*, y_h^*, q_h^*) as the numerical approximation of (u^*, y^*, q^*), we mean the optimal variables of $(\mathbf{P}_h)$.

5.2.1 Boundary Concentrated FEM

Theorem 5.2.1. *Let (u^*, y^*, q^*) be the optimal variables of the Neumann control problem and (u_h^*, y_h^*, q_h^*) be their numerical approximations. Let τ_h be a bc-mesh and $\mathbf{p}$ a linear degree vector of sufficiently large slope α. Let (**P**) subject to (N) be $H^{1+\sigma}$-regular and $f, y_d \in B^0_{1-\sigma}(\Omega, C_f, \gamma_f)$ with $C_f, \gamma_f > 0$ and $r = r_\Gamma$. Then there exists a constant $C > 0$, independent of h, such that*

$$\| u^* - u_h^* \|_{L^2(\Gamma_\mathcal{N})} + \| y^* - y_h^* \|_{H^1(\Omega)} + \| q^* - q_h^* \|_{H^1(\Omega)} \leq C h^\sigma. \tag{5.11}$$

Proof. Because of (5.5), Cea's lemma, and $(y^*, q^*) \in B^2_{1-\sigma}(\Omega) \times B^2_{1-\sigma}(\Omega)$ (Theorem 3.2.1, Corollary 3.2.2), we obtain $\| u - u_h^* \|_{L^2(\Gamma_\mathcal{N})} \in \mathcal{O}(h^\sigma)$ as a consequence of the best approximation properties of the space $S^\mathbf{p}(\tau_h)$ on bc-meshes (Theorem 4.3.13). We use Theorem 4.3.13 again in (5.6) and (5.7) to obtain the desired estimates for the state and adjoint, respectively. □

Mesh Refinement Strategy

In our computations, the bc-meshes are obtained by suitably refining a given coarse mesh, which consists of quadrilateral elements with polynomial degree equal to one. In each refinement step we either refine a finite element (h-refinement) or increase its polynomial degree (p-refinement). Since each element is refined one way or the other, this method can be regarded as a 'uniform' refinement strategy.

The type of refinement depends on the location of an element $K \in \tau_h$ according to the definition of bc-meshes: If $\overline{K} \cap \Gamma \neq \emptyset$, the element is h-refined. Otherwise, we perform p-refinement. Note that this procedure only takes a-priori information, i.e., the location of the boundary, into account. An exemplary mesh is shown in Figure 5.1.

Example 1: Non-Convex Domain

We first study the numerically observed convergence of the bc-FEM for problems on general domains before turning to the special case of convex domains. Let us take a numerical example from [104]. The elliptic operator is of reaction-diffusion type ($D \equiv I, c \equiv 1$) on the L-shape domain

$$\Omega = (-1,1)^2 \setminus ([0,1] \times [-1,0]), \quad \Gamma_\mathcal{N} = \partial\Omega. \tag{5.12a}$$

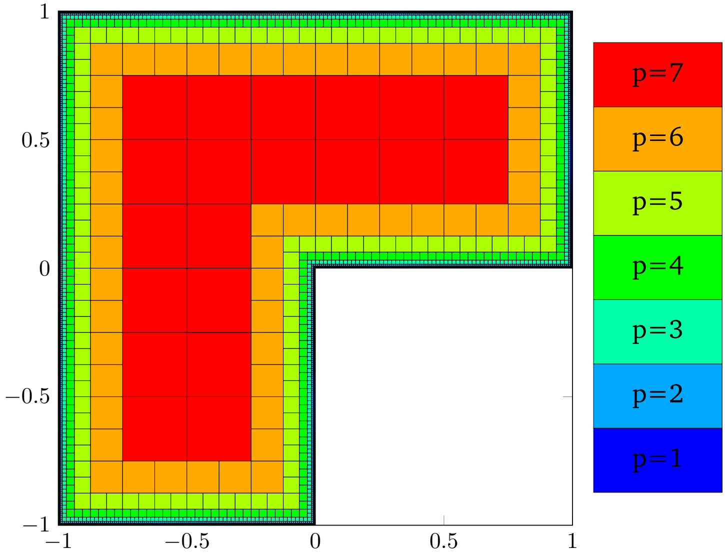


Figure 5.1. A bc-mesh of boundary size $h \approx 0.008$ for an L-shape domain.

Furthermore, $\nu = 1$, $u_a = -0.78$, $u_b = 0.55$. The rest of the data is given in polar coordinates:

$$y_d = -r^\lambda \cos(\lambda\theta), \tag{5.12b}$$

$$e_q = -\partial_n y_d, \tag{5.12c}$$

$$f \equiv 0 \tag{5.12d}$$

with $\lambda = 2/3$. It is shown in [104] that the unique solution is given by

$$y^* \equiv 0, \quad u^* = P_{[u_a,u_b]}\left(y_d\right), \quad q^* = -y_d. \tag{5.13}$$

The adjoint q^* admits the typical singularity appearing in problems posed on domains with reentrant corners. We have $q^* \in H^{1+\lambda-\varepsilon}(\Omega)$ for $\varepsilon > 0$ because of [72, Theorem 1.2.18].

The computations begin on an initial mesh consisting of 12 quadrilaterals of polynomial degree equal to one. We employ a warm start strategy and use the solution on a coarser grid as initial guess for computations on a finer discretization. In the spirit of the variational discretization, we prolong q_h^* on the fine mesh.

First, we examine the convergence with respect to L^2-norms. Theorem 5.2.1 predicts an error decay of order $\mathcal{O}(h^{2/3-\varepsilon})$. However, Figure 5.2 shows that the state and adjoint variable rather converge with order $\mathcal{O}(h^{4/3-\varepsilon})$. This faster rate ($4/3 = 2\lambda$) has already been observed in [28], but the proof remains open.

The control converges with algebraic order $\lambda + 1/2 - \varepsilon = 7/6 - \varepsilon$ in h, which is the best approximation rate obtained by bc-FEM at the boundary (see Theorem 4.3.16). In order to improve the rate in Theorem 5.2.1, an enhanced $L^2(\Omega)$-estimate for the state variable would be necessary. See also Remark 5.1.5 regarding the theoretical gap for Lebesgue norms.

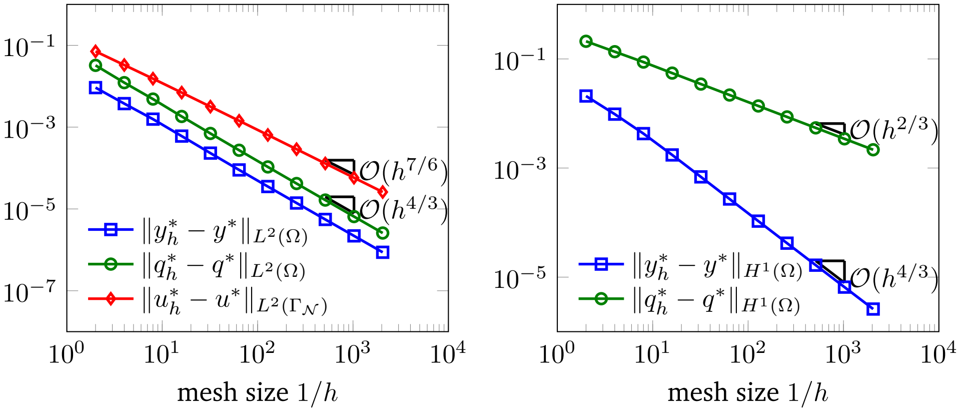


Figure 5.2. The convergence history of the bc-FEM applied to (5.10) with data (5.12).

The $H^1(\Omega)$-error for the adjoint variable decays with order $\mathcal{O}(h^{\lambda-\varepsilon})$, which is exactly the approximation rate of Theorem 5.2.1. A reason for the fast convergence order $\mathcal{O}(h^{2\lambda})$ in the state variable (see Figure 5.2) could be the fact that $y^* \equiv 0$. In this example, the error in the optimal variables is of the same order as the general approximation error of the finite element method.

We summarize known approximation results for the optimal control from the h-FEM in Table 5.1 in order to compare the convergence speed with respect to the number of unknowns N.

Non-Convex Ω	h-FEM	proved in	type of disc-	note
$\Rightarrow \sigma < 1$	$L^2(\Gamma_{\mathcal{N}})$		retization	
[104]	$h^{\sigma+1/2}$	Thm. 6.2,	post-process	$A = -\Delta + \mathrm{id}$
[104]	$h^{\sigma+1/2}$	Chap. 7,	variational	$A = -\Delta + \mathrm{id}$
[5]	$h^2 \lvert \ln h \rvert^{3/2}$	Thm. 7.7,	graded mesh	$A = -\Delta + \mathrm{id}$

Table 5.1: The convergence rates of the h-FEM applied to Neumann control problems on non-convex domains.

Theorem 4.3.3 shows $h \sim N^{-1}$ for the bc-FEM and, naturally, $h \sim N^{-1/2}$ for the uniform h-FEM. For this test problem, we observed with $\sigma \approx \lambda$

$$\| u^* - u_h^* \|_{L^2(\Gamma_\mathcal{N})} \in \mathcal{O}(h^{\sigma+1/2}) = \mathcal{O}(N^{-\sigma-1/2}),$$

which is better than the rates of [104]. The optimal rate of $h^2 |\ln h|^{3/2} \sim N^{-1}(\ln N)^{3/2}$ from [5] cannot be obtained by the bc-FEM. The latter, however, is designed for general elliptic equations, while the mesh grading factor in [5] is given in terms of the singular exponents of an expansion of the solution, which may be unknown.

Example 2: Convex Domain

The Laplacian on a convex domain is an H^2-regular example for which we were able to derive an L^∞-result of the bc-FEM (Theorem 4.3.18). We now extend this result to the approximation of the optimal control. The proof of the following theorem is similar to [80, Theorem 2].

Theorem 5.2.2. *Under the assumptions of Theorem 5.2.1 we have the error bound*

$$\| u^* - u_h^* \|_{L^\infty(\Gamma_\mathcal{N})} \leq Ch |\ln h|^{1/2}.$$

Proof. Define $q^h := S_h^*(y^* - y_d)$ and $y^h := S_h u^*$. Due to the variational discretization, we can exploit the projection formula and obtain the estimate

$$\begin{aligned} \| u^* - u_h^* \|_{L^\infty(\Gamma_\mathcal{N})} &\leq \| P_{U_{ad}}(-\nu^{-1} q^*) - P_{U_{ad}}(-\nu^{-1} q_h^*) \|_{L^\infty(\Gamma_\mathcal{N})} \\ &\leq \frac{1}{\nu} \| q^* - q_h^* \|_{L^\infty(\Gamma_\mathcal{N})} \leq \frac{1}{\nu} \| q^* - q^h \|_{L^\infty(\Gamma_\mathcal{N})} + \| q^h - q_h^* \|_{L^\infty(\Gamma_\mathcal{N})}. \end{aligned}$$

The first addend is of order $\mathcal{O}(h)$ according to Theorem 4.3.18. The second addend is estimated with the help of [162, Lemma 4.4] and the trace theorem. It holds

$$\| q^h - q_h^* \|_{L^\infty(\Gamma)} \leq C |\ln h|^{1/2} \| q^h - q_h^* \|_{H^{1/2}(\Gamma)} \leq C |\ln h|^{1/2} \| q^h - q_h^* \|_{H^1(\Omega)}.$$

We proceed analogously to (5.9) and obtain

$$\| q^h - q_h^* \|_{H^1(\Omega)}^2 \leq C \| y^* - y_h^* \|_{L^2(\Omega)} \| q^h - q_h^* \|_{H^1(\Omega)}.$$

The last two estimates and Theorem 5.2.1 yield $\| q^h - q_h^* \|_{L^\infty(\Gamma)} \leq C |\ln h|^{1/2} h$. Hence, it holds

$$\| u^* - u_h^* \|_{L^\infty(\Gamma_\mathcal{N})} \leq C(h + h |\ln h|^{1/2}),$$

which concludes the proof. □

The inverse estimate of [162, Lemma 4.4] is established with generalized discrete harmonic extensions on a quasi-uniform triangulation $\tilde{\tau}$ of Ω. As explained in Remark 4.3.19, we can apply the result in the context of bc-meshes.

We use the same numerical solution strategy as before for a problem of [27] on the convex domain $\Omega = (0,1)^2$. The optimal variables read

$$\begin{aligned} q^* &= -x_1 x_2^2 e^{x_1+x_2}, \\ u^* &= P_{[u_a,u_b]}\left(x_1 x_2^2 e^{x_1+x_2}|_{\Gamma_{\mathcal{N}}}\right), \\ y^* &= (4x_1x_2 + 2x_1x_2^2)e^{x_1+x_2}. \end{aligned}$$

We choose

$$f := -(4x_1x_2^2 + 16x_1x_2 + 4x_2^2 + 8x_2 + 12x_1)e^{x_1+x_2}, \tag{5.14a}$$

$$y_d := (2x_2^2 + 2x_1)e^{x_1+x_2} \tag{5.14b}$$

with the inhomogeneities

$$e_q := \begin{cases} -2x_2^2 e^{1+x_2} & \text{if } x_1 = 1, \\ -3x_1 e^{1+x_1} & \text{if } x_2 = 1, \end{cases} \tag{5.14c}$$

and

$$e_y := -u^* + \begin{cases} (8x_2 + 4x_2^2)e^{1+x_2} & \text{if } x_1 = 1, \\ 14x_1 e^{x_1+1} & \text{if } x_2 = 1. \end{cases} \tag{5.14d}$$

The rest of the data is given by

$$\nu = 1, \quad u_a \equiv 1, \quad u_b \equiv 6, \quad D \equiv I, \quad c \equiv 0, \tag{5.14e}$$

$$\Gamma_{\mathcal{N}} = \{x_1 = 1\} \cup \{x_2 = 1\}, \quad \Gamma_{\mathcal{D}} = \Gamma \setminus \Gamma_{\mathcal{N}}. \tag{5.14f}$$

Our computations start on a uniform mesh consisting of 16 elements of degree one. The L^2-errors of y_h^*, u_h^* decay with order $\mathcal{O}(h^2)$ (see [27, Figure 2]), while Theorem 5.2.1 only predicts $\mathcal{O}(h)$. This shows that the theory is not optimal in L^2, as already noted in Remark 5.1.5.

Here, we investigate the pointwise errors of the state and adjoint variable in view of Theorem 5.2.2.

Convex Ω $\Rightarrow \sigma = 1$	h-FEM $L^\infty(\Gamma_{\mathcal{N}})$	proved in	type of discretization	notes
[39]	h^1	Thm. 4.8	full	u_h piecewise constant
[38]	h^1	Thm. 6.7	full	u_h piecewise linear
[80]	$h^{2-2/p}\lvert\ln h\rvert$	Ex 2	variational	smooth $\Gamma, q^* \in W^{2,p}(\Omega)$

Table 5.2: The L^∞ convergence rates in of the h-FEM applied to Neumann control problems on convex domains.

Since the projection $P_{[u_a,u_b]}$ is non-expansive, the $L^\infty(\Gamma_\mathcal{N})$-error of q_h^* gives a bound for the error in the control variable. We observe that the bc-FEM converges significantly faster than the h-FEM with respect to the number of unknowns (Figure 5.3).

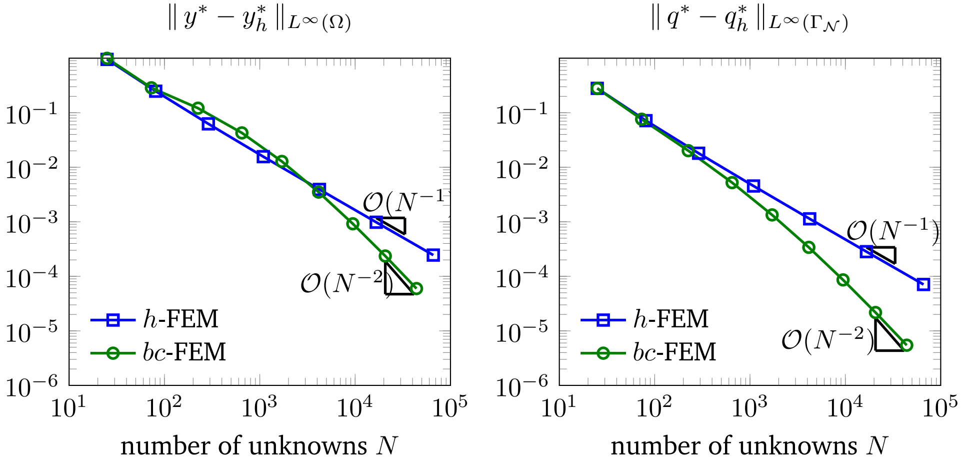


Figure 5.3. The convergence histories in L^∞ of the h, bc-FEM applied to (5.10) with data (5.14).

Let us amend the previous overview on convergence results by enhanced L^2 estimates for convex domains (Table 5.3).

Convex Ω $\Rightarrow \sigma = 1$	h-FEM $L^2(\Gamma_\mathcal{N})$	proved in	type of discretization	notes
[39]	h^1	Thm. 4.9	full	u_h piecewise constant
[38]	$h^{3/2}$	Thm. 6.6	full	u_h piecewise linear
[38]	$h^{3/2-\varepsilon}$	Thm. 5.4	variational	semi-linear problem
[80]	$h^{3/2}$	Ex 1	variational	smooth $\Gamma, q^* \in W^{2,p}(\Omega)$
[104]	$h^{2-1/p}$	Thm. 5.2	post-process	$\omega \leq \pi/2, q^* \in W^{2,p}(\Omega)$
[104]	$h^{3/2}$	Thm. 5.2	post-process	$\omega < \pi, q^* \in W^{2,p}(\Omega)$
[5]	$h^2 \lvert \ln h \rvert^{3/2}$	Thm. 7.7	graded mesh	$A = -\Delta + \mathrm{id}$

Table 5.3: The L^2 convergence rates of the h-FEM applied to Neumann control problems on convex domains.

As regards the $L^\infty(\Gamma_\mathcal{N})$ error, the result of Theorem 5.2.2 and $h \sim N^{-1}$ make the higher order method superior to the low order methods of [39, 38] and about equal to [80]. Owing to Theorem 5.2.1, the bc-FEM converges faster in $L^2(\Gamma_\mathcal{N})$ than the uniform h-FEM in [39, 38, 80], for which we have $h^{3/2} \sim N^{-3/4}$. The same arguments show that the proven error decay of the bc-FEM is basically of the same quality as in [104, 5].

We observe that the convergence of the bc-FEM in $L^\infty(\Gamma_\mathcal{N})$ is faster than predicted by Theorem 5.2.2.

5.2.2 Vertex Concentrated FEM

Now we turn to the vc-FEM and its approximation quality for Neumann control problems. We carry over the approximation results of the hp-FEM to the discrete problem $(\mathbf{P}_h)$.

Theorem 5.2.3. *Let (u^*, y^*, q^*) be the optimal variables of the Neumann control problem and (u_h^*, y_h^*, q_h^*) be their numerical approximations. Let the assumptions of Theorem 3.3.24 be valid. Let τ_ς^m be a geometric mesh with a linear polynomial degree vector $\mathbf{p}$ of sufficiently large slope α. Then there exist constants C, b, independent of the number of unknowns $N = dim(S^\mathbf{p}(\tau))$, such that*

$$\| u_h^* - u^* \|_{L^2(\Gamma_\mathcal{N})} + \| y_h^* - y^* \|_{H^1(\Omega)} + \| q_h^* - q^* \|_{H^1(\Omega)} \leq C e^{-b\sqrt[3]{N}}.$$

Proof. The proof is analogous to the proof of Theorem 5.2.1. First, we replace Theorem 3.2.1 and Corollary 3.2.2 by Theorem 3.3.24. Second, we replace Theorem 4.3.13 by Theorem 4.4.4. □

Mesh Refinement Strategy

The mesh refinement strategy for the vc-FEM mainly follows the ideas of the bc-FEM. Our computations start on a coarse grid consisting of quadrilaterals with polynomial degree one. In the consecutive refinement steps, we obtain finer discretizations by either refining an element (h-refinement) or increasing its polynomial degree (p-refinement).

In the analysis above we assumed that the points of singularity of solutions are known. Singularities arising from the differential equation are known to be confined to the vertices of the domain. Singularities from switching points are features of optimal control problems with inequality constraints. In general the location of switching points is unknown.

Let us describe how we cope with this difficulty. Once the discrete problem on a given mesh is solved, we can compute the switching points of the discrete optimal control u_h^*. Then we h-refine the elements containing these switching points. In addition, we h-refine their neighbors that are closest to the switching point. If the discretization error is small enough we expect that the switching points of u^* are contained in these h-refined elements. In this way, switching points are treated like points where the boundary conditions changes, and we stay consistent with the usual geometric mesh refinement, see [135, Chapter 4]. As we expect that the number of switching points stays bounded, the number of geometric mesh patches is finite. Let us emphasize that it is still open under what assumptions these meshes satisfy the requirements of Theorem 5.2.3.

In addition, elements containing the vertices of the domain or vertices, where the type of boundary condition changes, are h-refined. Elements that are not h-refined will be p-refined. This lead to the characteristic discretization with geometric mesh patches as shown in Figure 5.4.

On finer discretizations, we use the same warm start strategy as for the bc-FEM in order to ensure that the first iterate of Newton's method is close enough to the true solution.

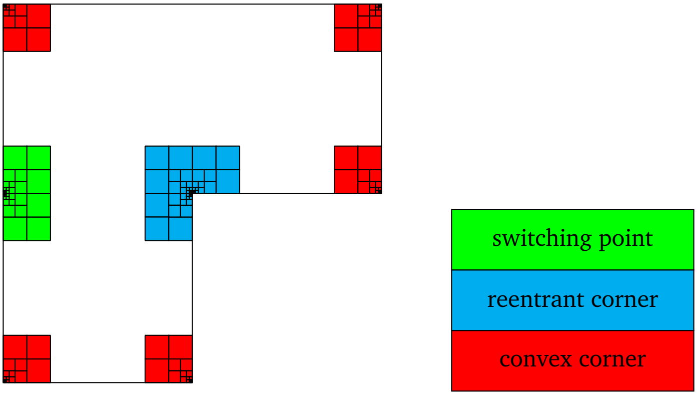


Figure 5.4. Schematic refinement of corners and switching points of the control.

Example 1 Revisited

We apply the vc-FEM with the mixed a-priori a-posteriori refinement strategy from above to problem (5.10) with data (5.12). The solution is given in (5.13) and we already know that $q^* \in B_\beta^2(C, \gamma, \Omega)$ with $\beta \in (1 - \lambda, 1)$ and the weight function $r_\mathcal{V}$, but $q^* \notin B_{1-\lambda}^2$ (Remark 2.1.8). For the parameters determining the geometric mesh, we choose $\sigma = 0.5$ and $\alpha = 1$. Again, the initial mesh consists of 12 quadriaterals of degree one.

Theorem 5.2.3 predicts an exponential error decay of the control and state variable in the L^2- and H^1-norm, respectively. The numerical results, which are depicted in Figure 5.5, reflect this behavior.

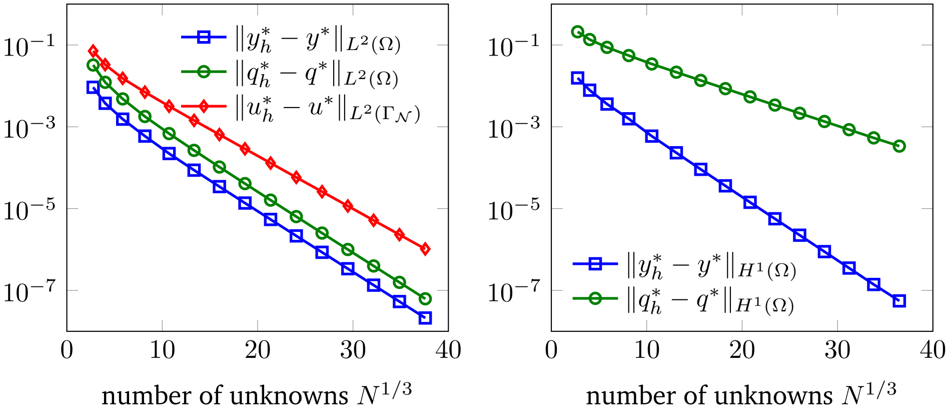


Figure 5.5. The error decay of the vc-FEM applied to (5.10) with data (5.12).

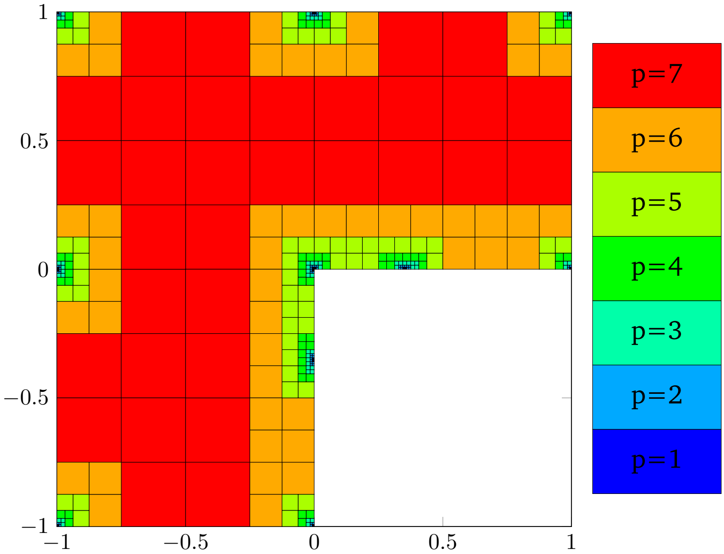


Figure 5.6. The geometric mesh of level 8 and the vc-FEM applied to Problem (5.10) with data (5.12).

We point out that the approximation result in Theorem 5.2.3 was established for a geometric mesh that exactly captures the switching points. However, the mesh patches from the mixed a-priori a-posteriori refinement strategy do not exactly reproduce the geometric meshes as defined in Section 6.3. A resulting mesh is depicted in Figure 5.6. Sometimes, kinks in the optimal control u_h^* , i.e., points where the active and inactive set meet, move into neighboring elements. The algorithm, therefore, has to make sure that the polynomial degree is kept low in a sufficiently large neighborhood (in practice, two/three elements sufficed) because the projection formula has to be applied. This leads to slightly different mesh patches in the course of successive refinements.

In order to provide a detailed error analysis, we list the discretization error for states and adjoints with respect to the energy norm in Table 5.4 for the different refinement levels. Note that there is no conception of a mesh size for geometric meshes.

For investigating the exponential decay numerically, we introduce the *experimental exponent of convergence* (eec).

$$eec(y, L^2(\Omega)) := \frac{\ln \| y^* - y^*_{h_1} \|_{L^2(\Omega)} - \ln \| y^* - y^*_{h_2} \|_{L^2(\Omega)}}{\sqrt[3]{N_2} - \sqrt[3]{N_1}}, \tag{5.15}$$

where $N_1 < N_2$ are the degrees of freedom for two consecutive refinements. The EEC is the numerical approximation for the constant b in Theorem 5.2.3. The convergence history of Table 5.4 indicates that the EEC is bounded from below, which supports the exponential convergence result.

$level$	N	$\|y_h^* - y^*\|_{H^1(\Omega)}$	$eec(y, H^1(\Omega))$	$\|q_h^* - q^*\|_{H^1(\Omega)}$	$eec(q, H^1(\Omega))$
1	21	$1.56 \cdot 10^{-2}$	-	$2.11 \cdot 10^{-1}$	-
2	65	$7.86 \cdot 10^{-3}$	$5.46 \cdot 10^{-1}$	$1.36 \cdot 10^{-1}$	$3.48 \cdot 10^{-1}$
3	199	$3.60 \cdot 10^{-3}$	$4.29 \cdot 10^{-1}$	$8.75 \cdot 10^{-2}$	$2.42 \cdot 10^{-1}$
4	535	$1.57 \cdot 10^{-3}$	$3.64 \cdot 10^{-1}$	$5.52 \cdot 10^{-2}$	$2.02 \cdot 10^{-1}$
5	1,167	$5.92 \cdot 10^{-4}$	$4.05 \cdot 10^{-1}$	$3.44 \cdot 10^{-2}$	$1.97 \cdot 10^{-1}$
6	2,251	$2.32 \cdot 10^{-4}$	$3.64 \cdot 10^{-1}$	$2.16 \cdot 10^{-2}$	$1.80 \cdot 10^{-1}$
7	3,845	$9.12 \cdot 10^{-5}$	$3.64 \cdot 10^{-1}$	$1.36 \cdot 10^{-2}$	$1.81 \cdot 10^{-1}$
8	6,075	$3.61 \cdot 10^{-5}$	$3.60 \cdot 10^{-1}$	$8.57 \cdot 10^{-3}$	$1.79 \cdot 10^{-1}$
9	9,049	$1.43 \cdot 10^{-5}$	$3.58 \cdot 10^{-1}$	$5.40 \cdot 10^{-3}$	$1.78 \cdot 10^{-1}$
10	12,875	$5.65 \cdot 10^{-6}$	$3.56 \cdot 10^{-1}$	$3.40 \cdot 10^{-3}$	$1.78 \cdot 10^{-1}$
11	17,649	$2.24 \cdot 10^{-6}$	$3.56 \cdot 10^{-1}$	$2.14 \cdot 10^{-3}$	$1.78 \cdot 10^{-1}$
12	23,465	$8.88 \cdot 10^{-7}$	$3.57 \cdot 10^{-1}$	$1.35 \cdot 10^{-3}$	$1.78 \cdot 10^{-1}$
13	30,419	$3.52 \cdot 10^{-7}$	$3.57 \cdot 10^{-1}$	$8.50 \cdot 10^{-4}$	$1.79 \cdot 10^{-1}$
14	38,607	$1.4 \cdot 10^{-7}$	$3.58 \cdot 10^{-1}$	$5.36 \cdot 10^{-4}$	$1.79 \cdot 10^{-1}$
15	48,565	$5.55 \cdot 10^{-8}$	$3.44 \cdot 10^{-1}$	$3.37 \cdot 10^{-4}$	$1.72 \cdot 10^{-1}$

Table 5.4: The convergence history for the vc-FEM applied to (5.10) with data (5.12).

Let us comment on the multi-indices $\beta, \tilde{\beta}$ in Theorem 3.3.24 for the current example. After relabeling the vertices $X_i \in \mathcal{V}$, we can assume that X_1 is the origin and X_2 the right switching point on the horizontal line $\{x_2 = 0\}$ (confer Figure 5.6). The reentrant corner X_1 of Ω yields the index $\beta_1 \in (1 - \pi/\omega_{X_1}, 1) = (1/3, 1)$ while the remaining vertices or switching points allow for arbitrary $\beta_j \in (0,1), j \neq 1$, see Remark 3.3.26.

Now we investigate the value of $\tilde{\beta}$ for the regularity of the optimal control. The extension of u^* in the proof of Theorem 3.3.24 is constructed with [13, Theorem 4.3], which relies on the continuity of u^* (Corollary 3.3.28). The idea is to subtract a polynomial that coincides with u^* at all vertices, which restricts the extension problem to the case $u^*(X_j) = 0$ for all $X_j \in \mathcal{V}$. Then, each $u^*|_{\Gamma_j}$ is extended separately with the help of [13, Theorem 4.2] and the summation over j yields a global extension $u^* \in B^2_{\tilde{\beta}}(\Omega)$.

We exemplary show the generation of the values $\tilde{\beta}$ on the segment $[X_1, X_2] = \overline{\Gamma}_2$, where the optimal control is inactive as can be seen from (5.13). Let $i = 1, 2$ and $-\nu^{-1} q^* \in B^2_{\beta}(\Omega)$ with $\beta_i \in (1/2, 1)$. Then the trace theorem [13, Theorem 4.1] yields $u^*|_{\Gamma_2} \in B^1_{\hat{\beta}}(\Gamma_2)$ with $\hat{\beta}_i \in (\beta_i - 1/2, 1/2)$. An extension from Γ_2 into the domain, constructed with [13, Theorem 4.2], lies in $B^2_{\tilde{\beta}}(\Omega)$ with $\tilde{\beta}_i \in (\hat{\beta}_i + 1/2, 1)$. We see that $\tilde{\beta}_i$ is strictly larger than β_i because the values need to be chosen from the open interval (confer [13, Remark 4.2]). If on the other hand $\beta_i \in (0, 1/2)$, then $u^*|_{\Gamma_2} \in B^2_{\hat{\beta}}(\Gamma_2)$ with $\hat{\beta}_i \in (\beta_i + 1/2, 1)$ and the extension to the domain belongs to $B^2_{\tilde{\beta}}(\Omega)$ with $\tilde{\beta}_i \in (\hat{\beta}_i - 1/2, 1/2)$. Again, the value of $\tilde{\beta}$ has increased compared to β.

Observe that the application of Theorem 4.1 and 4.2 of [13] is only possible if the values of β_j, β_{j+1} stem from the same interval $(0, 1/2)$ or $(1/2, 1)$ on both ends of Γ_j. That is why a value $\beta_j \in (0, 1/2)$, e.g., $\beta_1 \in (1/3, 1/2)$ at the reentrant corner, may need to be increased as soon as one $\beta_{j'}, j \neq j'$ lies in $(1/2, 1)$ on a connected component of the Neumann boundary.

We close this subsection with a survey on the error decay with respect to the number of ddof for different mesh refinement strategies. Our hp-strategies can be regarded as uniform refinement techniques because each element is refined. A comparison with uniform h-FEM, therefore, seems to be adequate.

There are plenty of strategies which aim at enhancing the speed of convergence for the h-FEM such as mesh grading (see [4, 5]), extended finite element methods (see the overview article [21]), and adaptive mesh refinement (h-FEM ad.) based on a-posteriori error estimators (see [90, 91]). We implemented the last approach with the residual based error estimator of [106] (see also Subsection 6.4.3). Only those elements whose error is greater than 50% of the maximal error among the elements are marked for refinement. The different FEM discretizations are compared in Figure 5.7 regarding the approximation error with respect to the number of unknowns.

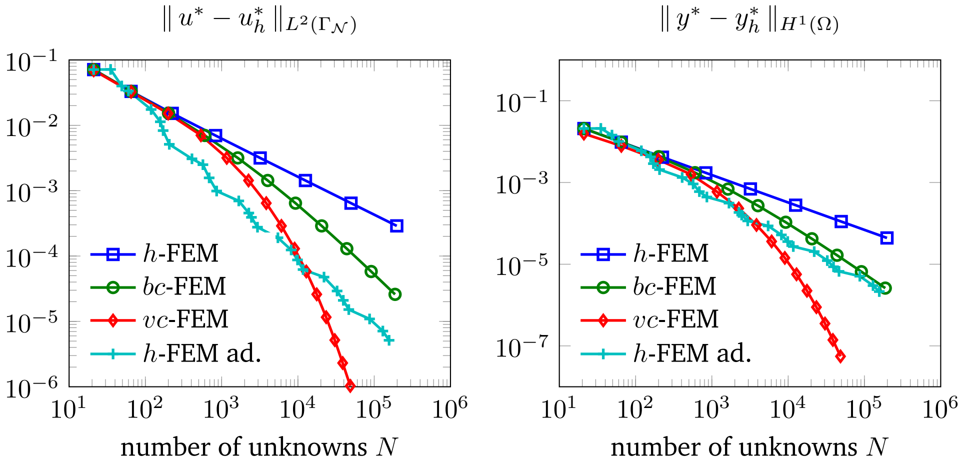


Figure 5.7. The error decay for different FEM strategies applied to Problem (5.10) with data (5.12).

As we would expect, the vc-FEM is superior regarding the approximation quality with respect to the number of unknowns and eventually beats all other strategies. Both the vc- and bc-FEM converge faster than the uniform h-FEM, which compensates for the higher implementational efforts. Admittedly, the adaptive h-FEM converges faster than bc-FEM, which on the other hand does not depend on an a-posteriori error estimator.

Example 2 Revisited

The numerical results for the vc-FEM and the optimization problem with data (5.14) posed on the unit square are similar to the previous example. The plots show exponential convergence and the EEC is bounded from below. We also note that the vc-FEM can be applied to problems with small regularization parameter ν as done in the preprint of [156].

Again, we compare the different FE methods for the data (5.14). We observe the same qualitative behavior as for Example 1, except for the fact that the adaptive h-FEM is no longer superior to the bc-FEM. This is a consequence of the convex domain, which yields an H^2-regular problem. In this case, the uniform h-FEM achieves optimal order of convergence. All in all, the higher order methods are superior as regards the convergence speed with respect to the number of unknowns (see Figure 5.8).

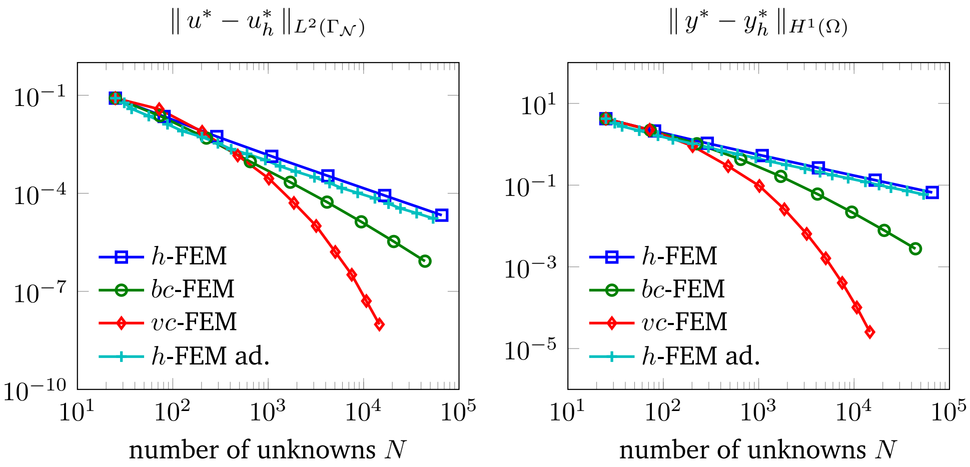


Figure 5.8. The error decay for different FEM strategies applied to Problem (5.10) with data (5.14).

5.2.3 Bang-Bang Problems

Recall from Definition 2.3.6 that an optimal control u^* is called bang-bang if it is active almost everywhere on its domain U. The previous numerical examples heavily exploit the projection formula, which is an equivalent reformulation of the variational inequality (2.14b) of the first order necessary conditions (2.14). This result fails to hold for $\nu = 0$, which is why we resort to the following equivalent representation (see [144]):

$$u^*(x) = \begin{cases} u_a(x) & \text{if } B^*q^*(x) + \nu u^*(x) > 0, \\ \in [u_a(x), u_b(x)] & \text{if } B^*q^*(x) + \nu u^*(x) = 0, \\ u_b(x) & \text{if } B^*q^*(x) + \nu u^*(x) < 0. \end{cases} \tag{5.16}$$

As before, we treat the case $U = \Gamma_\mathcal{N}$.

A Neumann control is bang-bang for $\nu = 0$ if the set $\{x \in \Gamma_\mathcal{N} \mid q^*(x) = 0\}$ has Lebesgue-measure zero. We control the measure of the level sets of $|q(x)|$ with the following inequality:

$$\operatorname{meas}_1\{x \in \Gamma_\mathcal{N} \mid |q^*(x)| \leq \varepsilon\} \leq c\varepsilon. \tag{5.17}$$

For $\varepsilon \searrow 0$, we see that q^* indeed produces bang-bang controls. Such a structural assumption involving q^* already appears in [50, 65, 147, 155, 159].

In the remaining part of this section, we add the index ν to the control problem to stress the dependence on the regularization parameter. The optimal variables are denoted by $(u, y, q)_\nu$, where the index ν instead of the asterisk * signifies optimality.

For the un-regularized problem with $\nu = 0$, the existence of a solution was established under the additional assumption that U_{ad} is bounded (see Assumption 2.3.1). An alternative is to prove existence under the assumption that y_d is in the range of $A^{-1}B$. Uniqueness only follows if $A^{-1}B$ is injective. Given a slight perturbation $y_{d,\delta} \approx y_d$ in the data, $y_{d,\delta}$ may leave the range of the solution operator and the optimal solution ceases to exist. Hence, the un-regularized problem is ill-posed in the sense of Hadamard because (unique) existence of solutions or continuous dependence on the data may be violated.

It is a common technique to solve un-regularized problem ($\mathbf{P}^0$) by a sequence of regularized problems ($\mathbf{P}^\nu$) with $\nu \searrow 0$ and to split the error into

- the discretization error $\| u_\nu - u_{\nu,h} \|_{L^2(U)}$ (see [154, Theorem 2.3]),
- the regularization $\| u_0 - u_\nu \|_{L^2(U)}$ (see [154, Lemma 2.2]).

To control the latter, parameter choice rules have been investigated in the context of inverse problems (see the overview in [153]). An important strategy for regularized inverse problems is the discrepancy principle (see [111]), which is applied to continuous optimal control problems in [155]. We also mention [50, 159, 154].

Changing the parameter ν in the course of computations is challenging with the hp-FEM because the method constructs meshes for *one* particular solution. Adjusting ν alters the problem and, consequently, the solution may no longer be smooth in regions where higher order elements aimed at exploiting regularity. Vice versa, singular parts of the solution may have moved away such that strongly refined elements lie in wrong regions.

This is the reason why we do not apply the vc-FEM to bang-bang problems but rather the bc-FEM. A changing solution does not affect the approximation properties of the FE space because the method keeps low order elements on the whole Neumann boundary.

We use the results of [154] to establish an a-priori parameter choice rule.

Theorem 5.2.4. *Let the model problem* ($\mathbf{P}$)=($\mathbf{P}^\nu$) *and* $u_a, u_b \in L^\infty(\Gamma_\mathcal{N}) \cap H^{1/2}(\Gamma_\mathcal{N})$ *be given subject to an* H^2*-regular Neumann boundary value problem* (N). *Let* (5.17) *hold. Denote by* $(u, y, q)_0$ *the solution to* ($\mathbf{P}^0$) *and by* $(u, y, q)_{\nu(h)>0,h}$ *a sequence of solutions to* ($\mathbf{P}_h{}^{\nu(h)}$)*, obtained by the update strategy* $\nu(h) \sim h$*. Under the same assumptions as in Theorem* 5.2.1*, there exists a* $C > 0$ *such that*

$$\begin{aligned} \| u_0 - u_{\nu(h),h} \|_{L^1(\Gamma_\mathcal{N})} &\leq Ch, \\ \| y_0 - y_{\nu(h),h} \|_{L^2(\Omega)} &\leq Ch, \\ \| q_0 - q_{\nu(h),h} \|_{L^\infty(\Gamma_\mathcal{N})} &\leq Ch. \end{aligned}$$

The idea behind the strategy $\nu(h) \sim h$ is to balance the error stemming from the regularization with the error of discretization. The latter is measured in the L^2- and L^∞-norm, which is why the assumptions on the state equation are taken from Section 4.3.2, where we derived error estimates for Lebesgue-norms.

Proof. From Theorem 3.2.1 we have $y \in B^2_0(\Omega)$ for a solution y to the state equation. Hence, Theorem 4.3.13 and 4.3.16 yield

$$\| (S - S_h)u \|_{L^2(\Omega)} + \| (S^* - S_h^*)(y - y_d) \|_{L^2(\Gamma_\mathcal{N})} \leq c(h + h^{3/2}) =: \delta_2(h).$$

Applying Theorem 4.3.18 yields

$$\| (S - S_h)u \|_{L^\infty(\Gamma_\mathcal{N})} \leq ch =: \delta_\infty(h).$$

The a-priori estimates δ_2, δ_∞ can be inserted ([154, Remark 2.3]) into [154, Theorem 3.1], which proves the result. □

We apply the a-priori regularization strategy $\nu(h) \sim h$ to the optimal control problem (5.10) with $\nu = 0$ and $e_q \equiv 0$. We choose the Laplacian ($D \equiv I$, $c \equiv 0$) on the following domain:

$$\begin{aligned} \Omega &= (0,1)^2, && \text{(5.18a)} \\ \Gamma_\mathcal{D} &= (\{x_1 = 0\} \cap \partial\Omega) \cup (\{x_2 = 0\} \cap \partial\Omega), && \text{(5.18b)} \\ \Gamma_\mathcal{N} &= (\{x_1 = 1\} \cap \partial\Omega) \cup (\{x_2 = 1\} \cap \partial\Omega). && \text{(5.18c)} \end{aligned}$$

The rest of the data is inspired by [144]. We construct the data such that the optimal control has bang-bang character in a checkerboard pattern. We set

$$\begin{aligned} y_d &= \sin(\pi x_1)\sin(\pi x_2) + \sin(2.5\pi x_1)\sin(2.5\pi x_2), && \text{(5.18d)} \\ f &= 2\pi^2 \sin(\pi x_1)\sin(\pi x_2), && \text{(5.18e)} \\ e_y &= \begin{cases} \pi(\sin(\pi x_2)) - \operatorname{sign}(\sin(2.5\pi x_2)) & \text{if } x_1 = 1, \\ \pi(\sin(\pi x_1)) - \operatorname{sign}(\sin(2.5\pi x_1)) & \text{if } x_2 = 1 \end{cases} && \text{(5.18f)} \end{aligned}$$

with the sign-function

$$\operatorname{sign} : \mathbb{R} \to \{-1, 0, 1\}, \quad \operatorname{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

For $u_a \equiv -1$, $u_b \equiv 1$, straightforward forward calculations show that the unique solution is given by

$$\begin{aligned} y_0 &= \sin(\pi x_1)\sin(\pi x_2), \\ u_0 &= \begin{cases} \operatorname{sign}(\sin(2.5\pi x_2)) & \text{if } x_1 = 1, \\ \operatorname{sign}(\sin(2.5\pi x_1)) & \text{if } x_2 = 1, \end{cases} \\ q_0 &= -\frac{2}{25\pi^2}\sin(2.5\pi x_1)\sin(2.5\pi x_2). \end{aligned}$$

Since the domain is convex, we deduce H^2-regularity of the state equation (see [72]). Applying the arguments of the proof of [50, Lemma 3.2] to the setting of boundary controls shows that (5.17) holds true. Thus, all prerequisites of Theorem 5.2.4 are met by the test example.

The numerical results depicted in Figure 5.9 are in very good concordance with the proved error decay of order $\mathcal{O}(h)$.

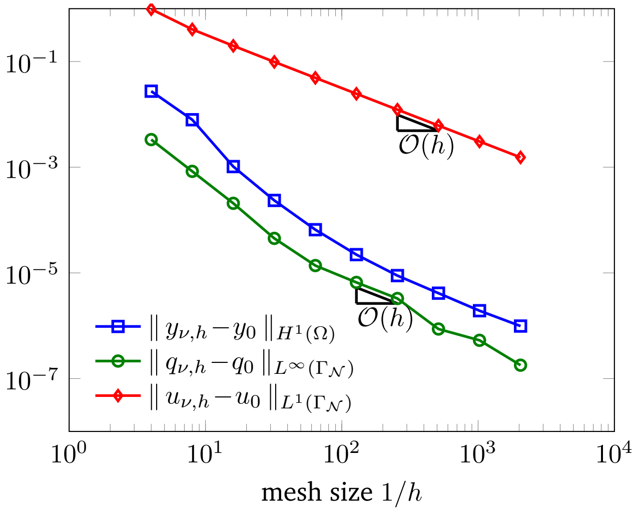


Figure 5.9. The problem (5.10) with data (5.18) is solved with the bc-FEM and the a-priori regularization $\nu = \nu(h) \sim h$ of Theorem 5.2.4.

5.3 Interface Control Problems

Very much related to the previous examples, solved with the bc-FEM, are *interface control problems* posed on $2d$-networks. We rigorously introduced the differential constraint, also called transmission problem, in Subsection 2.2.2 and guaranteed well-posedness through Assumption 2.2.3.

For the readers convenience, we recall the interface control problem, i.e., (**P**) subject to (T). It reads

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

subject to $u \in U_{ad} := \{u \in L^2(\mathcal{I}) \mid u_a \leq u \leq u_b \text{ a.e. in } \mathcal{I}\}$ with $u_a, u_b \in H^{1/2}(\mathcal{I})$ and

$$\text{(T)} \qquad \left\{ \begin{aligned} -\kappa_i \Delta y_i &= f_i && \text{in } \Omega_i, \\ y_i - y_j &= 0 && \text{on } \gamma_{i,j} \in \mathcal{I}, \\ \kappa_i \partial_{n_i} y_i + \kappa_j \partial_{n_j} y_j &= u && \text{on } \gamma_{i,j} \in \mathcal{I}, \\ y_i &= 0 && \text{on } \Gamma_i \in \mathcal{E}_{\mathcal{D}}, \\ \kappa_i \partial_{n_i} y_i &= h_i && \text{on } \Gamma_i \in \mathcal{E}_{\mathcal{N}}. \end{aligned} \right.$$

For the remainder of this chapter, we assume that transmission problem satisfies Assumption 2.2.3. Therein, we find $u \in H^{1/2}(\mathcal{I})$ which is to be understood for the optimal control that automatically possesses this regularity (Remark 2.3.8).

Theorem 5.3.1. *Let (u^*, y^*, q^*) be the optimal variables of the interface control problem and (u^*_h, y^*_h, q^*_h) be their numerical approximations. Let τ_h be an ic-mesh on a $2d$-network $\{\Omega_i\}_{i\in I}$ with a linear degree vector $\mathbf{p}$ of sufficiently large slope α. Suppose that the state equation allows solutions that belong to $H^{1+\sigma_i}(\Omega_i)$ with $\sigma_i \in (0,1]$ for all $i \in I$. If $f, y_d \in B^0_{1-\sigma_i}(\Omega_i, C_f, \gamma_f)$ with $C_f, \gamma_f > 0$ and the weight function $r_{\partial\Omega_i}$, there exists a constant $C > 0$, independent of h, such that*

$$\| u^* - u^*_h \|_{L^2(\mathcal{I})} + \| y^* - y^*_h \|_{H^1(\Omega)} + \| q^* - q^*_h \|_{H^1(\Omega)} \leq Ch^\sigma \tag{5.19}$$

with $\sigma := \min_{i\in I}\{\sigma_i\}$.

By stipulating local $H^{1+\sigma_i}(\Omega_i)$-regularity, the theorem does not depend on Section 3.1, where we derived an expansion for the solution of transmission problems.

Proof. We use Theorem 5.1.4 and then apply Cea's lemma. The optimal variables y^*, q^* lie in $B^2_{1-\sigma}(\Omega, \overline{C}, \overline{\gamma})$ for h-independent constants $\overline{C}, \overline{\gamma} > 0$ and $r_{\mathcal{I}\cup\Gamma}$ thanks to Corollary 3.2.3. The result now follows from the best approximation properties of the FE space $S^{\mathbf{p}}(\tau_h)$ on ic-meshes (Corollary 4.3.15) and the estimates (5.5)-(5.7). □

The domain for the numerical example is depicted in Figure 5.10. We have $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 = [0,2]^2$ with the interface $\mathcal{I} = \partial\Omega_1$ and $\Gamma_{\mathcal{D}} = \partial\Omega$. The data is chosen on the two subdomains Ω_1, Ω_2 as

$$f(x) = \begin{cases} f_1 \equiv 10, \\ f_2 \equiv 10, \end{cases} \qquad y_d(x) = \begin{cases} y_{d,1} \equiv 16, \\ y_{d,2} \equiv 10. \end{cases} \tag{5.20a}$$

with $\kappa_1 = 5$ and $\kappa_2 = 0.25$. Slightly different from [158], we choose $u_b \equiv 5.5$ and $\nu = 0.01$.

A strict lower bound σ for the eigenvalues $\lambda_{X,j}$ of $\mathcal{A}_X(\lambda)$ with $X \in \mathcal{X}$ is obtained by applying Theorem 3.1.16 ($\sigma = 1/4$) or the refined estimate in Proposition 3.1.17 ($\lambda_{X,j} \geq 1/2$). By numerically evaluating the eigenvalues at the interior vertices of Ω, we find

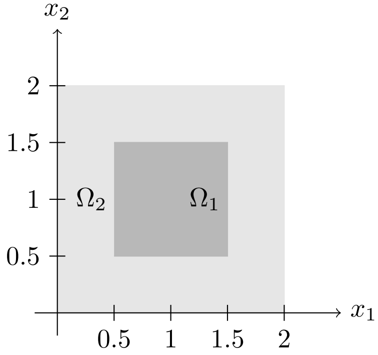


Figure 5.10. The $2d$-network for the interface control problem.

$$\lambda_1 = 0.70114949, \quad \lambda_2 = 1.2988505, \quad \lambda_3 = 2.7011495, \quad \ldots.$$

Note that all four vertices possess the same eigenvalues. Hence, we have for all $i \in I$ that $y^*_i \in H^{1+\lambda_1-\varepsilon}(\Omega_i)$ for small $\varepsilon > 0$ (Corollary 3.1.13).

Because of Theorem 5.3.1, we expect the error for the state and control variable to decay approximately like $\mathcal{O}(h^{0.7})$. We compute the error in the optimal variables by taking the solution on the finest discretization as reference (see Figure 5.11 and 5.12).

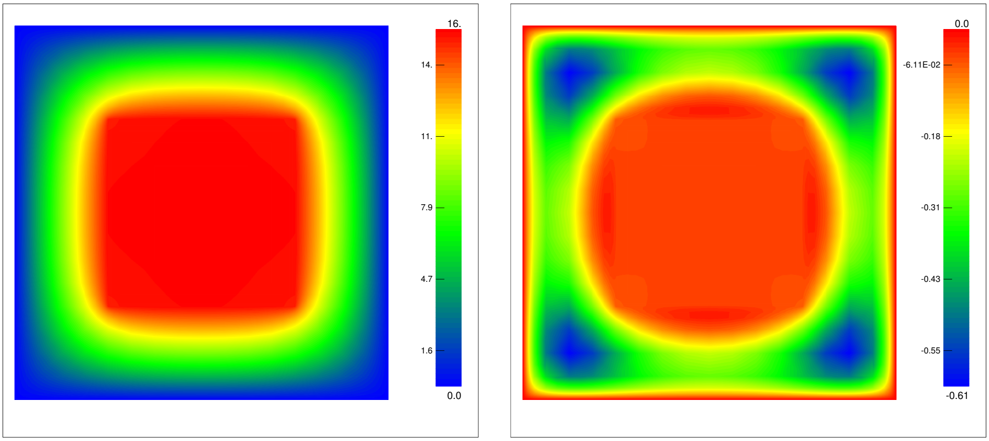


Figure 5.11: The approximate optimal state (left) and adjoint (right) of the interface control problem with data (5.20) on the finest discretization of the *ic*-FEM.

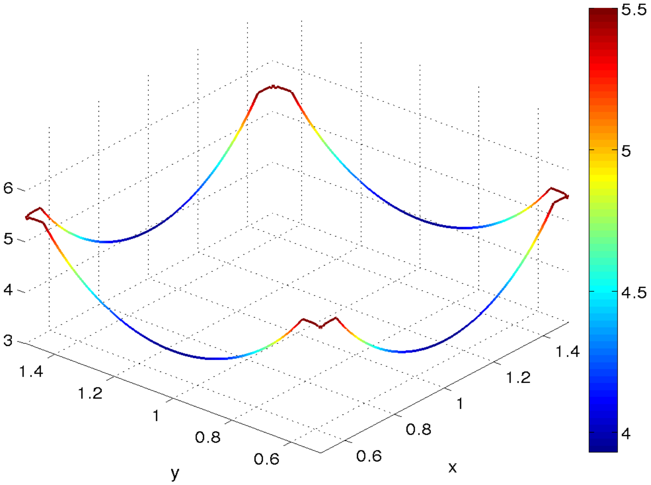


Figure 5.12. The approximate optimal control of the interface control with data (5.20) on the finest discretization of the *ic*-FEM.

For investigating the convergence behavior, we introduce the experimental order of convergence (eoc).

$$eoc(y, H^1(\Omega)) := \frac{\ln \| y^* - y^*_{h_1} \|_{H^1(\Omega)} - \ln \| y^* - y^*_{h_2} \|_{H^1(\Omega)}}{\ln(h_1) - \ln(h_2)}. \tag{5.21}$$

Here, $h_2 < h_1$ denote the mesh sizes of two consecutive discretizations.

We start the computation on a coarse mesh that consists of 64 quadrilaterals and has boundary mesh size 0.25. The convergence history for the state and adjoint variable is shown in Table 5.5 and 5.6, respectively.

The eoc for the state y in the $H^1(\Omega)$-norm is in very good compliance with the theoretical results and the estimate of $\sigma = \lambda_1 - \varepsilon \approx 0.7$. The convergence for the adjoint variable q and the $H^1(\Omega)$ norm is significantly faster, which could be explained by the fact that the optimal adjoint is close to zero in Ω_1 (see Figure 5.11). The singularities at the interior vertices, therefore, do not have much impact on the approximation quality.

h	N	$\|y_h^* - y^*\|_{L^2(\Omega)}$	$eoc(y, L^2(\Omega))$	$\|y_h^* - y^*\|_{H^1(\Omega)}$	$eoc(y, H^1(\Omega))$
0.25	81	$3.04 \cdot 10^{-1}$	-	5.69	-
0.125	289	$1.04 \cdot 10^{-1}$	1.55	3.51	$6.95 \cdot 10^{-1}$
0.0625	1,073	$3.82 \cdot 10^{-2}$	1.44	2.13	$7.21 \cdot 10^{-1}$
0.0312	3,425	$1.39 \cdot 10^{-2}$	1.45	1.27	$7.46 \cdot 10^{-1}$
0.0156	9,209	$5.08 \cdot 10^{-3}$	1.46	$7.71 \cdot 10^{-1}$	$7.20 \cdot 10^{-1}$
0.00781	22,177	$1.85 \cdot 10^{-3}$	1.46	$4.69 \cdot 10^{-1}$	$7.16 \cdot 10^{-1}$
0.00391	49,857	$6.63 \cdot 10^{-4}$	1.48	$2.83 \cdot 10^{-1}$	$7.30 \cdot 10^{-1}$
0.00195	107,329	$2.25 \cdot 10^{-4}$	1.56	$1.65 \cdot 10^{-1}$	$7.76 \cdot 10^{-1}$
0.000977	224,777	$6.17 \cdot 10^{-5}$	1.87	$8.64 \cdot 10^{-2}$	$9.37 \cdot 10^{-1}$
0.000488	462,593	-	-	-	-

Table 5.5: The convergence history for the state variable and the ic-FEM applied to the interface control problem with data (5.20).

As regards the error at boundary parts, we proved in Theorem 4.3.16 that the L^2 approximation quality is of order $\mathcal{O}(h^{\sigma+1/2})$ at best. This rate can be observed in Table 5.6 and is also valid for $\| u^* - u_h^* \|_{L^2(\partial\Omega_1)}$ because the adjoint and control variable are related through the (non-expansive) projection operator $P_{U_{ad}}$.

h	N	$\|q_h^* - q^*\|_{L^2(\mathcal{I})}$	$eoc(q, L^2(\mathcal{I}))$	$\|q_h^* - q^*\|_{H^1(\Omega)}$	$eoc(q, H^1(\Omega))$
0.25	81	$3.54 \cdot 10^{-3}$	-	2.45	-
0.125	289	$1.38 \cdot 10^{-3}$	1.36	1.29	$9.18 \cdot 10^{-1}$
0.0625	1,073	$5.55 \cdot 10^{-4}$	1.31	$5.99 \cdot 10^{-1}$	1.11
0.0312	3,425	$2.39 \cdot 10^{-4}$	1.21	$2.42 \cdot 10^{-1}$	1.31
0.0156	9,209	$1.08 \cdot 10^{-4}$	1.15	$9.29 \cdot 10^{-2}$	1.38
0.00781	22,177	$4.79 \cdot 10^{-5}$	1.17	$3.60 \cdot 10^{-2}$	1.37
0.00391	49,857	$2.08 \cdot 10^{-5}$	1.21	$1.48 \cdot 10^{-2}$	1.28
0.00195	107,329	$8.44 \cdot 10^{-6}$	1.30	$6.70 \cdot 10^{-3}$	1.14
0.000977	224,777	$2.79 \cdot 10^{-6}$	1.60	$3.08 \cdot 10^{-3}$	1.12
0.000488	462,593	-	-	-	-

Table 5.6: The convergence history for the adjoint variable and the ic-FEM applied to the interface control problem with data (5.20) .

Similarly, the convergence order $\mathcal{O}(h^{2\sigma})$ of the state variable measured in $L^2(\Omega)$ is not covered by Theorem 5.3.1. The ic-FEM basically consists of using the bc-FEM, for which faster convergence in Lebesgue norms has already been observed, on each subdomain Ω_i of the $2d$-network (see Figure 5.13).

Remark 5.3.2. *A best approximation error estimate $\| q - q_h \|_{L^2(\mathcal{I})} \leq Ch^{\sigma+1/2}$ can be established with the Aubin-Nitsche trick (analogue to Theorem 4.3.16). Here, the idea is to apply results from the bc-FEM on the subdomains of the $2d$-network (confer the proof of Theorem 5.3.1). For proving error estimates in the optimal variables, the same theoretical gap as mentioned in Remark 5.1.5 appears.*

We did not derive a regularity result in countably normed spaces for the optimal variables of the interface control problem. However, the remarks in Subsection 3.3.5 suggest that analytic regularity can be proved. The extension of the approximation result on geometric meshes (Theorem 4.4.4) to $2d$-networks should to be straightforward. Then, an analogous approximation result for the vc-FEM and interface control problems could be derived.

Due to the above expectations, we apply the vc-FEM to the test problem with data (5.20) and estimate the speed of convergence. We start the computations from the same discretization as the ic-FEM. As expected, the algorithm designs geometric mesh patches at the vertices and the kinks of the optimal control (see Figure 5.13).

Table 5.7 shows the eec, which is bounded from below.

level	N	$\|y_h^* - y^*\|_{L^2(\Omega)}$	$eec(y, L^2(\Omega))$	$\|q_h^* - q^*\|_{L^2(\mathcal{I})}$	$eec(q, L^2(\mathcal{I}))$
1	81	$3.04 \cdot 10^{-1}$	-	$3.54 \cdot 10^{-3}$	-
2	249	$1.07 \cdot 10^{-1}$	$5.30 \cdot 10^{-1}$	$1.93 \cdot 10^{-3}$	$3.10 \cdot 10^{-1}$
3	765	$8.81 \cdot 10^{-2}$	$6.87 \cdot 10^{-2}$	$8.32 \cdot 10^{-4}$	$2.94 \cdot 10^{-1}$
4	1,853	$3.18 \cdot 10^{-2}$	$3.25 \cdot 10^{-1}$	$3.86 \cdot 10^{-4}$	$2.45 \cdot 10^{-1}$
5	3,685	$6.66 \cdot 10^{-3}$	$4.94 \cdot 10^{-1}$	$1.41 \cdot 10^{-4}$	$3.18 \cdot 10^{-1}$
6	6,489	$2.38 \cdot 10^{-3}$	$3.20 \cdot 10^{-1}$	$6.59 \cdot 10^{-5}$	$2.38 \cdot 10^{-1}$
7	10,481	$1.14 \cdot 10^{-3}$	$2.29 \cdot 10^{-1}$	$3.63 \cdot 10^{-5}$	$1.85 \cdot 10^{-1}$
8	15,877	$6.25 \cdot 10^{-4}$	$1.84 \cdot 10^{-1}$	$2.2 \cdot 10^{-5}$	$1.54 \cdot 10^{-1}$
9	22,893	$3.75 \cdot 10^{-4}$	$1.57 \cdot 10^{-1}$	$1.42 \cdot 10^{-5}$	$1.34 \cdot 10^{-1}$
10	35,349	$1.12 \cdot 10^{-4}$	$2.74 \cdot 10^{-1}$	$4.71 \cdot 10^{-6}$	$2.49 \cdot 10^{-1}$
11	48,333	$5.72 \cdot 10^{-5}$	$1.85 \cdot 10^{-1}$	$2.36 \cdot 10^{-6}$	$1.92 \cdot 10^{-1}$
12	64,177	-	-	-	-

Table 5.7: The convergence history for the vc-FEM applied to the interface control problem with data (5.20).

All in all, the various numerical tests that were conducted within this work clearly show that the different hp-discretizations can successfully be integrated within a semismooth Newton method to solve the inequality constrained Neumann or interface control problem. We recover fast convergence rates with respect to the number of unknowns that are predicted by theory.

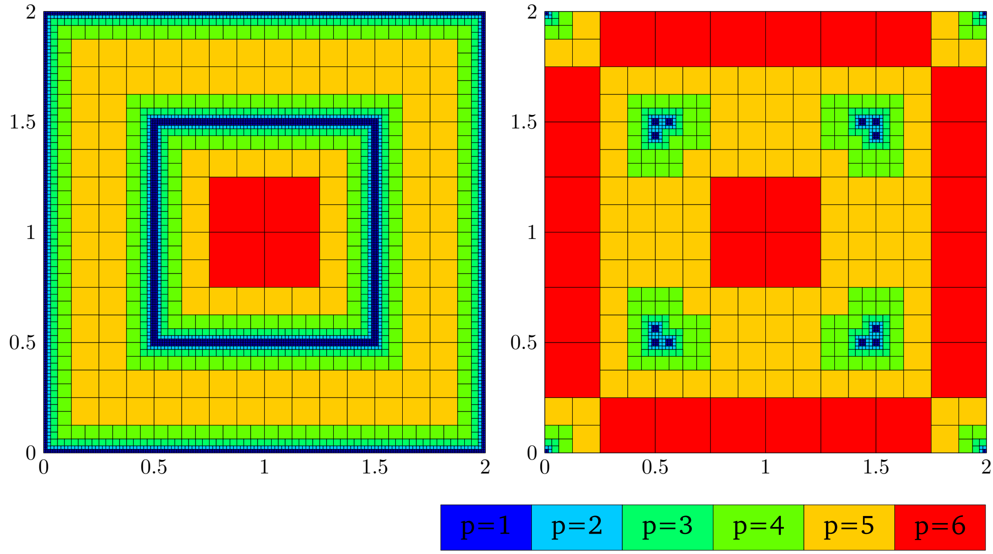


Figure 5.13: Left: An interface concentrated mesh of size $h \approx 0.008$ for the $2d$-network of Figure 5.10. Right: The corresponding geometric mesh of level 6 from the vc-FEM.

CHAPTER 6

A Path-Following Approach

We recall the model problem

$$\text{(P)} \qquad \begin{cases} \text{minimize } J(u,y) := \dfrac{1}{2}\| y - y_d \|^2_{L^2(\Omega)} + \dfrac{\nu}{2}\| u \|^2_{L^2(U)} \\ \qquad \text{subject to} \\ \qquad\qquad Ay = Bu, \\ \qquad\qquad u \in U_{ad}, \end{cases}$$

where we now assume $U_{ad} = \{u \in L^2(U) \mid u_a \leq u \leq u_b \text{ a.e. on } U\}$ and $u_a, u_b \in L^\infty(U)$ with $|u_a - u_b| \geq \vartheta > 0$. This allows theoretical investigations for $u \in L^q(U)$ and $2 < q \leq \infty$ (see [148]).

Instead of solving the problem with the semi-smooth Newton method, we dedicate this chapter to the *interior point method*. This method solves (**P**) as a sequence of perturbed problems, where the control constraints are dropped and barrier terms in the cost functional enforce feasibility. The solution of a perturbed problem is a point on the so called *central path* and the interior point method produces strictly feasible iterates that stay close to the central path in order to remain within the convergence region of Newton's method. This explains the term *path-following method*, which will be used interchangeably with interior point method. We stress that both distributed and Neumann control problems can be solved with this technique. The contents of this chapter are published in [157].

Let us comment on existing literature. A control reduced algorithm has successfully been applied in [130, 134, 160]). Here, the control is eliminated from the optimality system, which is similar to the concept of variational discretization from [79]. Additionally, super-linear convergence in problem specific norms is proved. This approach has also turned out to be suitable for treating state constraints, see [160, 133, 131].

Interior point methods working with primal and dual variables and projection or smoothing steps are explored in [148]. We also mention [122] for a problem with mixed control and state constraints. In [147], an affine scaling method is augmented by a smoothing step to prove super-linear convergence, which is related to semi-smooth Newton methods in function space [145]. All authors use h-FEM with a fixed polynomial degree (usually linear elements) as discretization.

We solve the model problem with an interior point method in function space. The implementation uses the hp-FEM with adaptive mesh refinement during the individual steps of the discrete algorithm. We heavily exploit the fact that the necessary optimality conditions of the barrier problems can be written as a smooth system of equations. Thus, the optimization method does not introduce artificial non-smoothness that could influence the hp-refinement. This fact was the main motivation to study an interior point method with this type of discretization. To the best of our knowledge, there are no available references about the application of *adaptive* hp-FEM to control constrained problems with pointwise inequality constraints.

We mention [41, 70], where the authors pursue an adaptive strategy for distributed control problems with the integral control constraint $\int_\Omega u \, \mathrm{d}x \geq 0$. Here, the regularity of the optimal control is not restricted by the constraint, different from pointwise constraints u_a, u_b (see also Theorem 2.3.7).

In the following section, we rigorously define the barrier problem and provide main results such as existence of the central path and first order necessary optimality conditions. The formulation of the interior point method and its convergence theorem is located in Section 6.2. We use the general framework of [148], which can treat solutions that touch the control bounds at isolated points.

In Section 6.3 we explain the implementation of a discrete version of the path-following method. Adaptive updates for the barrier parameter and mesh refinement will ensure that the iterates stay within the area of convergence for Newton's method. After that, a-posteriori error estimators for the central path as well as the Newton system are derived. Both rely on residual based error estimators of the underlying PDE.

Finally, in Section 6.5, we solve a test problem with known solution, where the convergence radius for Newton's method is large around the central path. Here, the hp-character of our method will be the center of investigation. Then, a problem with very small regularization parameter ν is solved. Due to adaptivity, the path-following algorithm successfully manages to stay within the region of attraction of Newton's method, which is very sensitive to reductions in the homotopy parameter μ.

6.1 The Barrier Problem and Central Path

We follow [130] to introduce barrier functionals and their most important properties. Proofs and further references can be found in [130, Chapter 2]. Recall that $U \ni x$ is the domain where the control $u = u(x)$ acts.

Definition 6.1.1. *Let $B \subset U \times \mathbb{R}$ be measurable, such that $B(x) := \{u \in \mathbb{R} \mid (x, u) \in B\}$ is closed, convex with non-empty interior for all $x \in U$. A function $l(x, u) : U \times \mathbb{R} \to \overline{\mathbb{R}}$ is called barrier function if it fulfills*

- *$l(., u)$ is measurable for any $u \in \mathbb{R}$.*
- *$l(x, .)$ is convex, continuous and differentiable on $\operatorname{int}(B(x))$.*

- $u \in \partial B(x) \Leftrightarrow l(x,u) = \infty$ *and* $\operatorname{dist}(\partial B(x), u) \geq d > 0 \Leftrightarrow l(x,u) < L$ *for* $L \in \mathbb{R}$ *depending only on* d.
- $l(x,u)$ *can be minorized by* $a(x) - c|u|$ *with* $a \in L^1(U)$.

The minorizing criterion guarantees that $\int_U l(x,u(x)) > -\infty$ for $u \in L^1(U)$.

A barrier function generates a barrier functional with the following definition.

$$b_\mu : L^1(U) \to \overline{\mathbb{R}}, \quad u \mapsto \mu \int_\Omega l(x,u(x))\,\mathrm{d}x.$$

Theorem 6.1.2. *For* $1 \leq p \leq \infty$ *the barrier functional* $b_\mu : L^p(U) \to \overline{\mathbb{R}}$ *is convex and lower semi-continuous. Furthermore,* $b_\mu(u) < \infty \Rightarrow u \in \operatorname{int}(B)$ *a.e.*

The result is shown in, e.g., [130, Corollary 2.1.6]. As the barrier functional suffers from reduced regularity, it is necessary to apply subdifferential calculus.

Definition 6.1.3. *Let* $f : V \to \overline{\mathbb{R}}$ *be a convex function. We define the set valued mapping*

$$\begin{aligned} \partial : V &\rightrightarrows V^* \\ v &\mapsto \{\partial f(v)\} \end{aligned}$$

where the image of v *contains all linear and bounded functionals* $m \in V^*$*, such that*

$$f(v + \delta v) + \langle m, \delta v\rangle_{V^*,V} \ \leq f(v)$$

In other words, the subdifferential $\partial f(v)$ contains the slope of all affine minorants that are exact at v.

Definition 6.1.4. *We say* $u \in U_{ad}$ *is strictly feasible if and only if there is an* $\varepsilon > 0$ *such that*

$$|u(x) - u_a(x)| + |u_b(x) - u(x)| > \varepsilon \quad \text{a.e. in } U.$$

Theorem 6.1.5. *Consider* $b : L^p(U) \to \overline{\mathbb{R}}$ *with* $1 \leq p \leq \infty$*. If* $p < \infty$ *or* u *is strictly feasible, then*

$$\partial b_\mu(u) \neq \emptyset \quad \Rightarrow \quad \partial b_\mu(u) = \{b'_\mu(u)\}$$

with the first variation

$$b'_\mu(u)\delta u = \mu \int_U \partial_u l(x,u(x))\delta u(x)\,dx.$$

In particular, we have $b'_\mu(u) \in L^p(U)^*$.

The result is proved in [130, Lemma 2.1.10, 2.1.11] and will be applied for the first order necessary conditions of the barrier problem.

In this work, we will work with $B(x) = [u_a(x), u_b(x)] \subset \mathbb{R}$ a.e. on U and use the logarithmic barrier function

$$l(x,u(x)) = -\ln(u(x) - u_a(x)) - \ln(u_b(x) - u(x)), \quad x \in U.$$

We enforce $u \in U_{ad}$ by adding a barrier functional to J. This leads to the problem ($\mathbf{P}_\mu$).

$$(\mathbf{P}_\mu) \qquad \begin{cases} \text{minimize } J_\mu(u,y) := J(u,y) + b_\mu(u) \\ \qquad\quad \text{subject to} \\ \qquad\qquad\quad Ay = Bu, \\ \qquad\qquad\quad u \in U_{ad}. \end{cases}$$

First, we have to assure that our homotopy approach makes sense insofar as the subproblem ($\mathbf{P}_\mu$) admits a unique solution.

Theorem 6.1.6. *The problem* ($\mathbf{P}_\mu$) *admits a unique solution* (u_μ, y_μ) *for all* $\mu > 0$. *Its value of the objective functional is finite.*

The assumptions on u_a, u_b imply that there is a $\check{u}$ with $u_a < \check{u} < u_b$ a.e. Then, the proof is standard (see, e.g., [130, Lemma 2.2.2]) and basically builds on the minima of convex functions (see [61, Proposition II.1.2]). For a more general setting, see [148].

In order to characterize the solution, we derive optimality conditions.

Theorem 6.1.7. *The first order necessary optimality system for* ($\mathbf{P}_\mu$) *reads*

$$Ay_\mu = Bu_\mu, \tag{6.1a}$$

$$\nu u_\mu + B^* q_\mu + b'_\mu(u_\mu) = 0, \tag{6.1b}$$

$$A^* q_\mu = y_\mu - y_d. \tag{6.1c}$$

These conditions are also sufficient for a minimizer (u_μ, y_μ).

We remark that (6.1) is very similar to the optimality system (2.14), the only difference being the non-linear equation (6.1b), which replaces the variational inequality (2.14b) due to barrier terms in J_μ.

Proof. We eliminate the state from the objective function by inverting the state equation. Take a minimizer u_μ from Theorem 6.1.6 and observe

$$J_\mu(u) \geq J_\mu(u_\mu) = J_\mu(u_\mu) + \langle 0, u - u_\mu \rangle \quad \forall u \in U_{ad}.$$

So we conclude $0 \in \partial J_\mu(u^*) \neq \emptyset$ and compute the derivative with the sum and chain rule ([61, Proposition I.5.6, I.5.7]). The fact that $b'_\mu(u_\mu) \in L^2(U)$ follows from Theorem 6.1.5. The adjoint variable q_μ is part of the optimality system as a Lagrange multiplier for the state equation. Since the optimization problem is convex, the necessary conditions are also sufficient. □

Definition 6.1.8. *For notational convenience, we set*

$$Y := Q := H^1(\Omega).$$

The solution (u_μ, y_μ) *together with the corresponding adjoint variable* q_μ *are referred to as the central path*

$$(u, y, q)_\mu \in L^2(U) \times Y \times Q.$$

Because of the bounds on the control and the boundedness of the solution operators for the state and adjoint equation, we get for all $\mu \in (0, \mu_0]$

$$\| (u, y, q)_\mu \|_{L^2(U) \times Y \times Q} \leq C_{\mu_0}.$$

Remark 6.1.9. *As the operators A, A^* are boundedly invertible, we can write down an equivalent formulation of the optimality system: $(u, y, q)_\mu \in L^2(U) \times Y \times Q$ solves (6.1) if and only if*

$$F_\mu(u) = \nu u_\mu + B^* A^{-*}(A^{-1} B u_\mu - y_d) + b'_\mu(u_\mu) = 0. \tag{6.2}$$

This equation holds a.e. in U.

Theorem 6.1.10. *The central path is Hölder continuous with index $1/2$, i.e.,*

$$\| (u, y, q)_\mu - (u, y, q)_\eta \|_{U \times Y \times Q} \leq L_c \sqrt{|\mu - \eta|} \tag{6.3}$$

*for all $\mu, \eta \in (0, \mu_0]$. Moreover, $(u^*, y^*, q^*) = \lim_{\mu \searrow 0} (u, y, q)_\mu$ exists and is the global solution of (**P**).*

For the proof we refer the reader to [148, Lemma 9]. Recall that a general solvability result for $\mu = 0$ is also established in Theorem 2.3.3. Passing to the limit ($\eta \searrow 0$) in (6.3) yields an error bound of order $\mathcal{O}(\sqrt{\mu})$ for the central path and the true solution. This bound can be improved for some $p < 2$, see [33].

Remark 6.1.11. *If the controls on the central path have positive distance to the bounds u_a, u_b in the L^∞ sense, the central path is differentiable (see, e.g., [130, 148]) and admits the bound*

$$\| \partial_\mu (u, y, q)_\mu \|_{L^\infty(U) \times Y \times Q} \leq C \mu^{-1/2}.$$

6.2 The Interior Point Method

Interior point methods can be regarded as methods that systematically solve a perturbed optimality system of the original minimization problem. This perturbed system is itself the exact optimality system for a barrier problem, which is characterized by an additional term in the cost functional that penalizes the violation of constraints. This strategy allows to deal with control and/or state constraints where the resulting equations have the advantage of being smooth.

6.2.1 An Abstract Algorithm in Function Space

Let the first order necessary conditions of the central path be given by $F_\mu(x) = 0$. A general interior point algorithm in function space tries find a solution with Newton's method and drives $\mu \searrow 0$ at the same time (see [148, Algorithm PDPFS]). A schematic outline of the method is given in Algorithm 1. We describe its building blocks in the subsequent sections.

Algorithm 1 Interior Point Method in Function Space

```
1: Choose initial point $x_0$ and barrier parameter $\mu_0 \in (0, \mu_{-\infty}]$.
2: $k := 0$.
3: Solve the Newton system
       $\partial F_{\mu_k}(x_k)\delta x = -F_{\mu_k}(x_k).$
4: $\tilde{x}_{k+1} := x_k + \delta x$.
5: Apply a smoothing step $x_{k+1} := Z(\tilde{x}_{k+1})$.
6: Choose $\sigma_k \in [\sigma_{min}, 1)$ and set $\mu_{k+1} := \sigma_k \mu_k$.
7: $k := k + 1$ and go to 3.
```

We remark that the smoothing step is not always necessary for convergence of the algorithm. Note that more details and investigations are essential for implementing this path-following algorithm in practice.

6.2.2 Well-Posedness

We take $F_\mu(u)$ from (6.2) and investigate the Newton system

$$\partial F_{\mu_k}(u_k)\delta u_k = -F_{\mu_k}(u_k). \tag{6.4}$$

Let $\mu_k > 0$ and u_k denote an iterate of the interior point method. Moreover, let $1 \leq p \leq \infty$ and

$$D^p := \{u \in U_{ad} \mid (u - u_a)^{-1}, (u_b - u)^{-1} \in L^p(U)\}. \tag{6.5}$$

A formal derivation at $u_k \in D_p$ in the direction h for $2 \leq p \leq \infty$ yields

$$\partial F_\mu(u_k)h = \nu h + B^* A^{-*} A^{-1} B h + \frac{\mu h}{(u_k - u_a)^2} + \frac{\mu h}{(u_b - u_k)^2}. \tag{6.6}$$

In order to guarantee that the Newton system can be solved at the new iterate $u_k + \delta u_k$, the smoothing operator is designed accordingly.

Lemma 6.2.1. *The μ-dependent function*

$$\beta : (u_a, u_b) \to \mathbb{R}, \quad x \mapsto \nu x - \frac{\mu}{x - u_a} + \frac{\mu}{u_b - x}. \tag{6.7}$$

is invertible and the Nemytzkij operator

$$\beta^{-1} : L^p(U) \to \{v \in L^p(U) \,|u_a < v < u_b\},\ (\beta^{-1}v)(x) := \beta^{-1}(u(x)),$$

is Lipschitz continuous for $p \in [2, \infty)$.

The lemma is proved in [148]. Note that we use the same symbol for the scalar function β^{-1} and its associated Nemytzkij operator β^{-1} (also named superposition operator as in [6]). For $2 \leq p < \infty$, we set

$$\begin{aligned} W^p &:= L^p(U) \times Y \times Q \times L^p(U) \times L^p(U) \text{ with} \\ w &:= (w_u, w_y, w_q, w_a, w_b) \in W^p \end{aligned}$$

to address each component. Furthermore, we use the notation $y(u) := A^{-1}Bu$ for the solution of the state equation and $q(u) = q(y(u))$ for the solution to the adjoint equation, respectively.

As already in the previous chapters, we need an additional regularity assumption for the constraint $Ay = Bu$ at this point.

Assumption 6.2.2. *Assume that the state equation is $H^{1+\sigma}$-regular. For a fixed p, we assume that $2 \leq p < q \leq \infty$ with the continuous embedding*

$$H^{1+\sigma}(\Omega) \hookrightarrow L^q(U).$$

Examples are given at the end of Subsection 3.1.2. See also Equation (2.3). We use the symbol σ to stay consistent with our previous notation and issue the warning not the mistake it for the step size selection parameter $\sigma_k \in [\sigma_{min}, 1)$ in the interior point method.

Definition 6.2.3. *The smoothing operators $z_u : L^p(U) \to L^q(U)$ and $Z : W^p \to W^q$ are defined as*

$$Z(w) := \begin{pmatrix} z_u(w_u) \\ z_y(w_u) \\ z_q(w_u) \\ z_a(w_u) \\ z_b(w_u) \end{pmatrix} = \begin{pmatrix} \beta^{-1}(q(w_u)) \\ y(z_u(w_u)) \\ q(w_u) \\ \mu/(z_u(w_u) - u_a) \\ \mu/(u_b - z_u(w_u)) \end{pmatrix} \tag{6.8}$$

The definition of z_u and z_a, z_b is taken from (70) and (72),(73) of [148]. For the convergence of the interior point method, it is essential that the smoothing step satisfies a Lipschitz property.

Lemma 6.2.4. *Let $(\mu, u) \in \mathbb{R}^+ \times L^p(U)$ and $s := (u_\mu, y_\mu, q_\mu, z_a(u_\mu), z_b(u_\mu))$ be a solution to ($\mathbf{P}_\mu$). Under Assumption 6.2.2, the smoothing operators z_u, Z are well defined and there exist constants $L_u, L > 0$ such that*

$$\| z_u(w_u) - s_u \|_{L^q(\Omega)} \leq L_u \| w_u - s_u \|_{L^p(\Omega)}, \tag{6.9}$$

$$\| Z(w) - s \|_{W^q} \leq L \| w - s \|_{W^p}, \quad \forall w \in W^p. \tag{6.10}$$

Moreover,

$$z_u(w_u) \in D^q.$$

Proof. Obviously, $s \in W^q$ and $s = Z(s)$ due to the first order necessary conditions. Lemma 6.2.1 and the invertibility of A, A^* yield

$$\| z_u(w_u) - z_u(s_u) \|_{L^q(U)} \leq \frac{1}{\nu} \| q(w_u) - q(s_u) \|_{L^p(U)} \leq \frac{C}{\nu} \| w_u - s_u \|_{L^p(U)}.$$

This proves the first claim. Similarly, we find

$$\| z_y(w_u) - z_y(s_u) \|_Y \leq C \| u - u_\mu \|_{L^p(U)}, \quad \| z_q(w_u) - z_q(s_u) \|_Q \leq C \| w_u - s_u \|_{L^p(U)}.$$

The remaining components of Z can be treated as in the proof of [148, Theorem 2] and we obtain with constants L_a, L_b

$$\| z_a(w_u) - z_a(s_u) \|_{L^q(U)} \leq L_a \| w_u - s_u \|_{L^p(\Omega)} \tag{6.11}$$
$$\| z_b(w_u) - z_b(s_u) \|_{L^q(U)} \leq L_b \| w_u - s_u \|_{L^p(\Omega)}. \tag{6.12}$$

Thus, we can bound $\| Z(w) - Z(s) \|_{W^q}$ by $\| w_u - s_u \|_{L^p(U)} \leq \| w - s \|_{W^p}$. It holds

$$\| Z(w) - s \|_{W^q} \leq \max\{L_u, C, L_a, L_b\} \| w - s \|_{W^p}.$$

It remains to show the last claim, i.e., $z_u(w_u) \in D^q$.

From the definition of z_u we get $\beta(z_u(w_u)) = q(w_u)$. Writing out the terms of β yields

$$\nu z_u(w_u) + z_a(z_u(w_u)) + z_b(z_u(w_u)) = q(w_u).$$

Since $z_u(w_u) \in U_{ad}$, it follows that $\nu z_u(w_u)$ is essentially bounded. With $q(w_u) \in H^{1+\sigma}(\Omega)$, we conclude

$$z_a(z_u(w_u)) + z_b(z_u(w_u)) \in L^q(U).$$

The possible singularities of the two addends do not interfere because $|u_a - u_b| \geq \vartheta > 0$. Consequently, each of them is q-times integrable and $z_u(w_u)$ belongs to D^q. □

Finally, we establish the invertibility of the Newton system (see also [148, Section 5]).

Theorem 6.2.5. *Let $\mu_k > 0$ and Assumption 6.2.2 be fulfilled. For any $u_k \in D^q$ and $r \in L^q(U)^*$, there is a $\delta u_k : U \to \overline{\mathbb{R}}$ such that*

$$\partial F_\mu(z_u(u_k))\delta u_k = r \quad \text{a.e. on } U \tag{6.13}$$

If $r \in L^q(U)$ we also have $\| \delta u_k \|_{L^t(U)} \leq C \| r \|_{L^p(U)}$ for all $t \in [p, q]$, where the constant C can be chosen uniformly on bounded subsets $(\mu, u) \in \mathbb{R} \times D^q$.

Proof. The result follows from the proof of [148, Lemma 4] if we show that $Z(u_k)$ lies in

$$\begin{aligned} N_{-\infty,q}(\mu) := \{(u, y, q, w_a, w_b) \in W_q \mid u \in U_{ad},\ (u - u_a)w_a \geq \gamma\mu,\ (u_b - u)w_b \geq \gamma\mu, \\ w_a|_{u>(u_a+u_b)/2} \leq \Theta_\mu,\ w_b|_{u<(u_a+u_b)/2} \leq \Theta_\mu, \\ \min\{w_a, w_b\}|_{u=(u_a+u_b)/2} \leq \Theta_\mu\} \end{aligned}$$

with $\Theta_\mu = 2\max\{\mu_{-\infty}, \mu/\gamma\}/\vartheta$.

The value $\gamma \in (0,1)$ is arbitrary and for $\gamma \searrow 0$ the set $N_{-\infty,q}(\mu)$ becomes $U_{ad} \times Y \times Q \times \{z_a \in L^q \mid z_a \geq 0\} \times \{z_b \in L^q \mid z_b \geq 0\}$. In order to make $\mu_{-\infty}$ an upper bound for the homotopy parameters μ_k (see Algorithm 1), we simply pick a large but finite value.

Remember that $u_b - u_a \geq \vartheta > 0$ almost everywhere. From the definition of the smoothing operator Z it is clear that $z_u(u_k) \in U_{ad}$ and

$$(z_u(u_k) - u_a)z_a(u) = \mu \geq \gamma\mu, \quad (u_b - z_u(u_k))z_b(u) = \mu \geq \gamma\mu.$$

Define $U_k^+ := \left\{x \in U \mid z_u(u_k) > \frac{u_a+u_b}{2}\right\}$. Then,

$$z_a(u_k)|_{U_k^+} = \frac{\mu}{(z_u(u_k) - u_a)}\bigg|_{U_k^+} \leq \frac{2\mu}{u_b - u_a}\bigg|_{U_k^+} \leq \frac{2\mu}{\vartheta} \leq \Theta_\mu.$$

The same estimate yields $z_b(u)|_{U_k^-} \leq \Theta_\mu$ with $U_k^- := \left\{x \in U \mid z_u(u_k) < \frac{u_a+u_b}{2}\right\}$.

For $z_u(u_k) = (u_a + u_b)/2$ we directly get

$$z_a(u_k) = z_b(u_k) = \frac{2\mu}{u_b - u_a} \leq \Theta_\mu,$$

which implies $Z(u_k) \in N_{-\infty,q}(\mu)$. □

6.2.3 Convergence

In this section we show that the Algorithm 1 converges. The Lipschitz continuity of Z, which was proved in Lemma 6.2.4, is vital because it closes a $p-q$ gap (see [148]) in the convergence analysis of the interior point method.

Theorem 6.2.6. *Let $\mu_0, \rho_0 > 0$ be fixed. Assume that $u^* = \lim_{\mu \searrow 0} u_\mu$ satisfies strict complementarity (as in [148, Definition 2]). Then there exists a sequence $\sigma_{min,k} \leq \sigma_{max} < 1$ with $\sigma_k \in [\sigma_{min,k}, \sigma_{max}]$ such that the iterates u_k generated by Algorithm 1 with the smoothing operator z_u are well defined and satisfy*

$$\| u_k - u_\mu \|_{L^q(U)} \leq C\sqrt{\mu_{k-1}}, \tag{6.14}$$
$$\| u_k - u^* \|_{L^q(U)} \leq (C + L_c)\sqrt{\mu_{k-1}}. \tag{6.15}$$

where C is some constant and L_c the Hölder constant of the central path (see Theorem 6.1.10) on $(0, \mu_0]$.

Proof. The proof of convergence is achieved by showing equivalence of Algorithm 1 with the path-following method of [148]. The latter is designed for iterates in W^q. By choosing μ_0 large enough we can launch the algorithms with

$$u_0 := \frac{u_a + u_b}{2} \in D^q,$$

respectively

$$w_0 := (u_0, y(u_0), q(u_0), \mu(u_a - u_0)^{-1}, \mu(u_b - u_0)^{-1})^\top \in W^q$$

sufficiently close to the central path.

It suffices to show that the iterates u_k of Algorithm 1 (smoothed under z_u) have a corresponding sequence w_k with $u_k = w_{k,u}$ that is generated by the interior point method of [148] (smoothed under Z). Convergence then follows from [148, Theorem 3] because the smoothing operator Z satisfies a Lipschitz property (due to Lemma 6.2.4).

The iterates in W_q are given by $w_{k+1} = Z(w_k + \delta w_k)$ with the Newton update

$$\begin{pmatrix} \nu & 0 & B^* & -I & I \\ 0 & I & A^* & 0 & 0 \\ B & A & 0 & 0 & 0 \\ w_{k,a} & 0 & 0 & w_{k,u} - u_a & 0 \\ -w_{k,b} & 0 & 0 & 0 & u_b - w_{k,u} \end{pmatrix} \delta w_k = - \begin{pmatrix} w_{k,y} - y_d + A^* w_{k,q} \\ \nu w_{k,u} + w_{k,q} - w_{k,a} + w_{k,b} \\ A w_{k,y} - B w_{k,u} \\ w_{k,a}(w_{k,u} - u_a) - \mu \\ w_{k,b}(u_b - w_{k,u}) - \mu \end{pmatrix}. \tag{6.16}$$

Now we show that $\delta w_{k,u} = \delta u_k$ for $k \geq 0$. By construction, we have $u_k = (w_k)_u$ and

$$A w_{k,y} - B w_{k,u} = 0, \quad w_{k,a}(w_{k,u} - u_a) = \mu, \quad w_{k,b}(u_b - w_{k,u}) = \mu w_{k,b} \tag{6.17}$$

for $k = 0$.

Assume by induction that these equalities hold for $k \geq 0$. From (6.17) we deduce that the last three components of the right hand side of (6.16) are zero. Eliminating $\delta w_{k,q}, \delta w_{k,a}, \delta w_{k,b}$ from the first row of (6.16) yields (6.4) and proves that $\delta w_{k,u}$ equals δu_k.

As the smoothing operator only depends on $(w_k)_u = u_k$, we find

$$u_{k+1} = z_u(u_k + \delta u_k) = z_u(w_k + \delta w_k) = Z(w_k + \delta w_k)_u = w_{k+1,u}.$$

The construction of Z guarantees that (6.17) remains valid. Thus, the induction is complete and our algorithm is equivalent to the one of [148], where all claims are established. □

6.3 Discretization

First, we explain the discretization of the optimality system by finite elements. We develop our ideas starting with the discretization of (6.1) and end up with a fully discrete version for solving (6.2). This will show why some diligence is necessary for the treatment of the non-linear optimality system when using higher order elements. After that, we present an implementable path-following algorithm, which adaptively controls the homotopy parameter, the area of convergence for Newton's method, and the discretization errors.

6.3.1 The Optimality System

In the following, we suppress the dependence on μ in (6.1). The state and adjoint equation can be discretized with the hp-FEM as described in Section 4.1. Let τ be an admissible triangulation of Ω and $S^{\mathbf{p}}(\tau) = \operatorname{span}\{\phi_1, \dots, \phi_M\}$ the approximation space. For the case of distributed controls, we obtain the discrete version of (6.1a) and (6.1c).

$$\mathcal{K}y^h - \mathcal{M}u^h = 0, \tag{6.18}$$

$$\mathcal{K}^* q^h - \mathcal{M}y^h + \bar{y}_d = 0, \tag{6.19}$$

where $\mathcal{M}$ is the mass matrix $\mathcal{M}_{ij} = \int_U \phi_i \phi_j \, \mathrm{d}x$ and $\mathcal{K}, \mathcal{K}^*$ shall represent the matrices corresponding to the differential operator A, A^*. If $A = \Delta = A^*$, this is the stiffness matrix with $\mathcal{K}_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, \mathrm{d}x$. The load vector is denoted by $\bar{y}_d := \int_\Omega y_d \phi_i \, \mathrm{d}x$.

The main question is how to discretize the control in (6.1b). If we used a finite element function $u^h \in S^{\mathbf{p}}(\tau)$, it would be hard to check whether a new Newton iterate $u^h_{k+1} = u^h_k + \delta u^h_k$ is feasible and produces finite integrals

$$\int_U \mu(u^h_{k+1} - u_a)^{-1} \, \mathrm{d}x, \qquad \int_U \mu(u_b - u^h_{k+1})^{-1} \, \mathrm{d}x.$$

An implementation has to ensure that the values at the integration points x_j lie in $(u_a(x_j), u_b(x_j))$. But this is challenging for polynomials of degree greater than one.

This issue is solved by representing the control not as a member of $S^{\mathbf{p}}(\tau)$, but as a vector consisting of values at the integration points, which are also used for the evaluation of the barrier terms.

We approximate

$$\int_U u\phi_i \, \mathrm{d}x \approx \sum_{j=1}^{M} u(x_j)\phi_i(x_j)\omega_j \tag{6.20}$$

where x_j are integration points in U with weights ω_j stemming from a Gaussian quadrature rule (see Section 6.3.2). We set

$$R^\top = \phi_i(x_j), \quad D = \operatorname{diag}(\omega_j), \quad i = 1, \dots, N, \quad j = 1, \dots, M. \tag{6.21}$$

and use u^h_j, $j = 1, \dots, M$ as a discrete control variable. Finding the values of $v^h \in S^{\mathbf{p}}(\tau)$ at x_j is achieved by

$$v^h_j := v^h(x_j) = (Rv^h)_j.$$

The discrete version of the first order necessary optimality condition only enforces (6.2) at each integration point, i.e.,

$$\nu u^h_j + (Rq^h)_j - \frac{\mu}{u^h_j - u_a} + \frac{\mu}{u_b - u^h_j} = 0 \quad \forall j = 1, \dots, M.$$

We take the values of u^h as a column vector (and understand non-linear terms as vectors as well) and rewrite the last equation as

$$\nu u^h + Rq^h - \frac{\mu}{u^h - u_a} + \frac{\mu}{u_b - u^h} = 0.$$

If we correct (6.18) according to the conventions in (6.20),(6.21), we have to solve the discrete optimality system

$$F_\mu^h(u^h, y^h, q^h) := \begin{pmatrix} \mathcal{K}y^h - R^\top D u^h \\ \mathcal{K}^* p^h - \mathcal{M} y^h + \bar{y}_d \\ \nu u^h + R q^h - \frac{\mu}{u^h - u_a} + \frac{\mu}{u_b - u^h} \end{pmatrix} = 0$$

It can easily be verified that this is the exact optimality system of the following discrete problem:

$$(\mathbf{P}_\mu^h) \qquad \begin{cases} \text{minimize } J_\mu^h(y^h, u^h) := \dfrac{1}{2} \| y^h - y_d \|^2_{L^2(\Omega)} + \dfrac{\nu}{2} \displaystyle\sum_{j=1}^{M} \omega_i (u_j^h)^2 \\ \qquad\qquad - \mu \displaystyle\sum_{j=1}^{M} \omega_i (\ln(u_j^h - u_a) + \ln(u_b - u_j^h)) \\ \qquad \text{subject to} \\ \qquad\qquad \mathcal{K} y^h = R^\top D u^h, \end{cases}$$

This non-linear system of equations is to be solved with Newton's method. After rearranging the rows of F_μ^h, the linearization reads

$$\partial F_\mu^h(u^h, y^h, p^h) = \begin{pmatrix} \nu I + \mu \operatorname{diag}((u_i^h - u_a)^{-2} + (u_b - u_i^h)^{-2}) & 0 & R \\ 0 & \mathcal{M} & -\mathcal{K} \\ R^\top D & -\mathcal{K} & 0 \end{pmatrix}$$

If we multiply the first row by D, the discretized optimality system in the variables (u^h, y^h, q^h) is symmetric. It can be solved either with direct methods or an iterative solver such as MINRES.

As in the continuous case, where we eliminated the state and adjoint variable, we can invert $\mathcal{K}, \mathcal{K}^*$ and get a discrete equation only in the variable u_k^h. We find

$$\partial F_\mu^h(u_k^h)\delta u_k^h := (\nu I + \mu \operatorname{diag}((u_k^h - u_a)^{-2} + (u_b - u_k^h)^{-2}) + R\mathcal{K}^{*^{-1}} \mathcal{M} \mathcal{K}^{-1} R^\top D)\delta u_k^h =$$
$$- \nu u + \mu (u_k^h - u_a)^{-1} - \mu (u_b - u_k^h)^{-1} - R\mathcal{K}^{*^{-1}} (\mathcal{M}\mathcal{K}^{-1} R^\top D u_k^h - \bar{y}_d) =: -F_\mu^h(u_k^h).$$

which is a discretization of the continuous Newton system (6.6). Multiplying this equation with $D = D^\top$ from the left and reordering yields

$$\Big(\nu D + \mu D \operatorname{diag}((u_k^h - u_a)^{-2} + (u_b - u_k^h)^{-2}) + DR\mathcal{K}^{*^{-1}} \mathcal{M}\mathcal{K}^{-1} R^\top D\Big) u_{k+1}^h =$$
$$\mu D (u_k^h - u_a)^{-1} - \mu D (u_b - u_k^h)^{-1} + \mu D \operatorname{diag}\left((u_k^h - u_a)^{-2} + (u_b - u_k^h)^{-2}\right) u_k^h + DR\mathcal{K}^{*^{-1}} \bar{y}_d.$$

For the addends of the left hand side of the system, we now have

- νD is positive definite (> 0) if we use an integration scheme with positive weights only,
- $\operatorname{diag}(u_k^h - u_a)^{-2} > 0$ and $\operatorname{diag}(u_b - u_k^h)^{-2} > 0$ if $u_k^h(x_j) \in (u_a, u_b)$,
- $DR\mathcal{K}^{*^{-1}}\mathcal{M}\mathcal{K}^{-1}R^\top D > 0$ because of $\mathcal{M}, \mathcal{K}, \mathcal{K}^* > 0$.

The symmetry of the left hand side is obvious. Thus, the system can be inverted by a (P)CG-solver (see, e.g., [34]).

The pointwise smoothing operator Z calls for numerical solutions of the cubic equation (6.7), which can be implemented in a numerically stable way (see [130, Section 8.5]). It guarantees that the values u_j^h are feasible and makes the numerical algorithm well-defined.

The adaptive interior point method will have to control the discretization error and perform hp mesh refinements. There are different strategies to guide between a finer triangulation and higher order elements. We choose the estimate of the smoothness of u_k based on the expansion in a Legendre series ([59]). Several modifications are possible because the variables $y_k(u_k)$ and/or $q_k(u_k)$ can also be examined and included in the decision process.

6.3.2 Estimating Smoothness and *hp*-Adaptivity

Let us now comment on the estimation of the smoothness of u_k^h, which is complicated by the fact that u_k^h is represented by the values at the integration points and, therefore, does not fit into the framework of [59]. Assume that $U = \Omega$ (the case of boundary control is analogous) and let $K \in \tau$ be an element with polynomial degree $p_K \geq 1$. As we want to assemble element mass and stiffness matrix on K without errors, we work with a Gaussian quadrature that which is at least exact for polynomials of order $2p_K$. Hence, we tensorize a one dimensional integration scheme with $p_K + 1$ points and obtain $(p_K + 1)^2$ points. A vector of values at these integration points uniquely determines a polynomial of order p_K. We therefore obtain a one to one mapping Ψ_K

$$\Psi_K : \mathbb{R}^{(p_K+1)^2} \to Q(p_K) := \Big\{ v : K \to \mathbb{R} \ : \ v = \sum_{i,j} a_{ij} x^i y^j, \quad 0 \leq i, j \leq p_K, \\ a_{ij} \in \mathbb{R}, \ (x, y) \in K \Big\}. \tag{6.22}$$

Let u_K^h be the vector that consists of all u_j^h with $x_j \in K$ ordered by the value of j. The transformation of u^h to a finite element function is realized by

$$\begin{aligned} \Psi : \mathbb{R}^N &\to S^{\mathbf{p}}(\Omega, \tau) \\ u^h &\mapsto \arg \min_{u \in S^{\mathbf{p}}(\Omega,\tau)} \frac{1}{2} \| u - \sum_{K \in \tau} \chi_K \Psi_K(u_K^h) \|_{L^2(\Omega)}^2. \end{aligned} \tag{6.23}$$

The solution of this L^2-projection can be computed by inverting the mass matrix for the load vector with components given by

$$\int_\Omega \Phi_i \cdot \sum_{K\in\tau} \chi_K \Psi_K(u^h_K)\,\mathrm{d}x, \quad \Phi_i \in S^{\mathbf{p}}(\Omega,\tau).$$

Let us briefly describe how the smoothness of a discrete function $v^h \in S^{\mathbf{p}}(\Omega,\tau)$ is assessed. We recall the Legendre polynomials $L_i(\hat{x})$ of degree $i \geq 0$ on $[-1,1]$, which were defined as

$$L_i(\hat{x}) := \frac{1}{2^i i!}\frac{d^i}{d\hat{x}^i}(\hat{x}^2-1)^i.$$

in the implementational remarks on the hp-FEM in Chapter 4. It is well known that they are orthogonal with respect to the $L^2((-1,1))$ inner product. This property can be retained in two space dimensions by tensorization, i.e., setting $L_{ij}(\hat{x},\hat{y}) = L_i(\hat{x})L_j(\hat{y})$. For each $K \in \tau$ and $v^h \in S^{\mathbf{p}}(\Omega,\tau)$, we can write $v^h|_K$ as a linear combination of Legendre shape functions. Let F_K be the function that maps the reference element $\hat{K} = (-1,1)^2$ to the physical element $K \in \tau$. The orthogonality property of $L_{ij}(\hat{x},\hat{y})$ yields

$$(v^h \circ F_K)(\hat{x},\hat{y}) = \sum_{0\leq i,j\leq p_K} v^K_{ij} L_{ij}(\hat{x},\hat{y}),$$

where $v^K_{ij} = \int_{\hat{K}} (v^h \circ F_K)(\hat{x},\hat{y}) L_{ij}(\hat{x},\hat{y})\,\mathrm{d}x\,\mathrm{d}y$. In our implementation, we can also use the basis transformation from Subsection 4.2.1.

The decay of the Legendre coefficients v^K_{ij} is exponential if the function is analytic, which means there are constants $C_K, b_K > 0$ such that $|v^K_{ij}| \leq Ce^{-b(i+j)}$. Given v^K_{ij} we compute C_K and b_K by a least-square fit. Elements whose estimated error η^2_K is larger than a fraction of the mean value, i.e., $\sigma \sum_{K\in\tau} \eta^2_K/\#\tau$ are marked for refinement. The value of b_K is then used to decide between h-refinement ($b_K \geq \delta$) and p-refinement ($b_K < \delta$), see [59, Algorithm 5]. We chose $\sigma = 0.75$ and $\delta = 1$.

In order to obtain sensible estimates of b_K, sufficiently many Legendre coefficients need to be computed, which is why the the initial mesh must consist of elements with polynomial degree no less than two or three. In practice, the smoothness of all optimal variables, the state, control, and adjoint variables can be considered. We now have a way of numerically estimating the smoothness of FE functions which allows the implementation of Algorithm 1.

6.3.3 A Fully Adaptive Interior Point Method

The implementation of Algorithm 1 is done as described in [130]. We borrow ideas from [133, 131] as regards the adaptive update of μ, which builds on general results about Newton and homotopy methods in [54]. The ideas are formulated for iterates u_k of Newton's method, which intends to solve the optimality conditions of the central path u_μ, i.e., the non-linear equation $F_\mu(u_\mu) = 0$.

A numerical realization of the homotopy method can only compute an inexact Newton step in function space, i.e.,

$$\delta u_k = -\partial F_\mu(u_k)^{-1} F_\mu(u_k) + e_k.$$

In order to ensure that the inexact method converges, the error e_k and the contraction

$$\theta_k(\mu) := \frac{\| \, \partial F_\mu(u_k)^{-1}[\partial F_\mu(u_k)(u_k - u_\mu) - (F_\mu(u_k) - F_\mu(u_\mu))] \, \|}{\| \, u_k - u_\mu \, \|} \tag{6.24}$$

are controlled. The simple calculation

$$\begin{aligned}\| \, u_{k+1} - u_\mu \, \| &= \| \, u_k + \delta u_k - u_\mu \, \| = \| \, u_k - \partial F_\mu^{-1} F_\mu(u_k) + e_k - u_\mu \, \| \\ &\le \| \, \partial F_\mu^{-1}(u_k)[\partial F_\mu(u_k)(u_k - u_\mu) - (F_\mu(u_k) - F_\mu(u_\mu))] \, \| + \| \, e_k \, \| \\ &\le \theta_k(\mu) \| \, u_k - u_\mu \, \| + \| \, e_k \, \|. \end{aligned} \tag{6.25}$$

shows that linear convergence

$$\| \, u_{k+1} - u_\mu \, \| \le \gamma \| \, u_k - u_\mu \, \| \tag{6.26}$$

with a factor $\gamma \in (0,1)$ is obtained if

$$\theta_k(\mu) + \frac{\| \, e_k \, \|}{\| \, u_k - u_\mu \, \|} \le \gamma. \tag{6.27}$$

The estimates (6.26) and (6.27) describe the interaction of discretization error and non-linearity of the barrier problem. For mildly non-linear problems we have $\theta_k \ll 1$ and the relative discretization error is the main contribution in (6.27). Highly non-linear problems, on the other hand, may allow the algorithm to perform several steps without refining the mesh, because θ_k dominates in (6.27). Both types of problems are solved in Section 6.5.

Algorithmically, we work with a desired contraction θ_d and two further parameters:

- θ_t with $0 < \theta_d < \theta_t < 1$ to decide when the contraction parameter is too large and whether the Newton corrector is successful,
- θ_c with $\theta_t < \theta_c < 1$ as the critical contraction that models the removal of the iterates from the region of convergence. Too large values terminate the Newton corrector with a failure.

For obtaining a numerical estimate of θ_k, we assume $e_k = 0$, and insert u_{k+1} in (6.24) as the best possible guess for the unknown u_μ. This leads to

$$\begin{aligned}\theta_k(\mu) \approx [\theta_k(\mu)] &:= \frac{\| \, \partial F_\mu(u_k)^{-1}[\partial F_\mu(u_k)(u_k - u_{k+1}) - (F_\mu(u_k) - F_\mu(u_{k+1}))] \, \|}{\| \, u_k - u_{k+1} \, \|} \\ &= \frac{\| \, u_k - u_{k+1} - \partial F_\mu(u_k)^{-1}(F_\mu(u_k) - F_\mu(u_{k+1})) \, \|}{\| \, u_k - u_{k+1} \, \|} = \frac{\| \, u_k - \bar{u}_{k+1} \, \|}{\| \, u_k - u_{k+1} \, \|} \end{aligned} \tag{6.28}$$

with the simplified Newton iterate

$$\bar{u}_{k+1} := u_{k+1} + \mathit{\Delta} u_k := u_{k+1} - \partial F_\mu(u_k)^{-1} F_\mu(u_{k+1}). \tag{6.29}$$

The interior point method, as described in Algorithm 1, uses a smoothing step $u_{k+1} := z_u(u_k + \delta u_k)$. Inserting this into (6.28) and (6.29) yields

$$[\theta_k(\mu)] = \frac{\| \, u_k - (z_u(u_k + \delta u_k) - \partial F_\mu(u_k)^{-1} F_\mu(z_u(u_k + \delta u_k))) \, \|}{\| \, z_u(u_k + \delta u_k) - u_k \, \|}. \tag{6.30}$$

The discretization error e_k is estimated with a robust a-posteriori error estimator (see Section 6.4). If the distance of u_{k+1} to the central path u_{μ_k} is below a certain accuracy tol_d and the contraction rate is acceptable ($[\theta_k] < \theta_t$), the Newton corrector is considered successful and a new homotopy parameter is computed. Otherwise, more Newton steps might be necessary. The simplified Newton step can be added to the current iterate in order to reduce the distance to the central path. If the estimated contraction is beyond the critical value θ_c, a more conservative value for μ_k is computed and the Newton corrector is relaunched.

An adaptive choice of σ_k as an update of the barrier parameter μ_k shall ensure that the iterates do not leave the area of convergence of Newton's method. The two main ingredients are slope information η about the central path and estimates of the contraction θ that relates Newton updates and simplified Newton updates.

If the central path u_μ is differentiable, we approximate the slope η_k at the current iterate u_k by

$$\partial_\mu u_\mu \approx [\eta_k(\mu)] := -\partial F_\mu^{-1}(u_k) \partial_\mu F_\mu(u_k). \tag{6.31}$$

Here and in the following, we stay consistent with our notation but remark that $\partial_\mu u_\mu$ is the derivative of the central path with respect to μ evaluated at the point μ, which is more commonly written as $\partial_\mu u(\mu)$. If u_k is close to the central path, we expect this inexact quantity to be a good estimate.

We use the approximation of the slope for the termination criterion of the path-following algorithm. The fundamental theorem of calculus yields

$$u_{\mu_k} - u^* = \lim_{\underline{\mu} \searrow 0} \int_{\underline{\mu}}^{\mu_k} \partial_\mu u_\mu \, \mathrm{d}\mu.$$

From the result of Remark 6.1.11, i.e., $\| \, \partial_\mu u_\mu \, \| \in \mathcal{O}(\mu^{-1/2})$ we construct the approximation

$$\partial_\mu u_\mu \approx \sqrt{\frac{\mu_k}{\mu}} \partial_\mu u_{\mu_k}. \tag{6.32}$$

We approximate the distance by

$$\| \, u_{\mu_k} - u^* \, \| \approx \lim_{\underline{\mu} \searrow 0} \int_{\underline{\mu}}^{\mu_k} \left\| \sqrt{\frac{\mu_k}{\mu}} \partial_\mu u_{\mu_k} \right\| \mathrm{d}\mu = 2\mu_k \| \, [\eta_k(\mu_k)] \, \|_{L^2(U)} =: [\| \, u_{\mu_k} - u^* \, \|]. \tag{6.33}$$

As soon as a global tolerance tol is reached, the algorithm stops.

Assuming a linear model for the contraction we have

$$\theta_k(\mu) \leq \omega_k(\mu) \| \, u_k - u_\mu \, \|. \tag{6.34}$$

The role of $\omega_k(\mu)$ is related to an upper bound of an affine covariant Lipschitz condition as used in [54]. For $\mu = \mu_k$, the best available guess for the central path is u_{k+1} and leads us the the estimate

$$\omega_k(\mu_k) \approx [\omega_k(\mu_k)] := \frac{[\theta_k(\mu_k)]}{\| u_k - u_{k+1} \|}. \tag{6.35}$$

The numerical estimate of the contraction also provides an estimate for the error in the central path via (6.25)

$$\begin{aligned} \| u_{k+1} - u_{\mu_k} \| &\approx [\theta_k(\mu_k)] \| u_k - u_{\mu_k} \| + \| e_k \| \\ &\leq [\theta_k(\mu_k)](\| u_k - u_{k+1} \| + \| u_{k+1} - u_{\mu_k} \|) + \| e_k \|. \end{aligned} \tag{6.36}$$

Hence,

$$\| u_{k+1} - u_{\mu_k} \| \approx [\| u_{k+1} - u_{\mu_k} \|] := \frac{[\theta_k(\mu_k)]}{1 - [\theta_k(\mu_k)]} \| u_{k+1} - u_k \| + \| e_k \|. \tag{6.37}$$

In Section 6.4.2 we develop an a-posteriori error estimator for the error in the newton step $[\| e_k \|] \approx \| e_k \|$. As as an alternative to (6.37) one could also use the a-posteriori error estimator of section 6.4.1 to estimate $\| u_{k+1} - u_{\mu_k} \|$.

An adaptive step size selection aims at achieving

$$\omega_k(\mu) \| u_k - u_\mu \| \approx \theta_d \in [0.1, 0.75]. \tag{6.38}$$

Assuming the model $\omega_k(\mu) \in \mathcal{O}(\mu^{-1/2})$ leads us to

$$[\omega_k(\mu)] := [\omega_k(\mu_k)] \sqrt{\frac{\mu_k}{\mu}}.$$

Proceeding as in [132], we compute

$$\| u_{k+1} - u_{\mu_{k+1}} \|_{L^2(U)} \leq \| u_{k+1} - u_{\mu_k} \|_{L^2(U)} + \| u_{\mu_k} - u_{\mu_{k+1}} \|_{L^2(U)} \tag{6.39}$$

and approximate the first term on the right hand side as in (6.37). For the second one, we use (6.32) and plug in (6.31) to obtain

$$\begin{aligned} \| u_{\mu_k} - u_{\mu_{k+1}} \|_{L^2(U)} \leq \int_{\mu_{k+1}}^{\mu_k} \| \partial_\mu u_\mu \|_{L^2(U)} \, \mathrm{d}\mu &\approx \| [\eta_k](\mu_k) \|_{L^2(U)} \int_{\mu_{k+1}}^{\mu_k} \sqrt{\frac{\mu_k}{\mu}} \, \mathrm{d}\mu \\ &= \| [\eta_k](\mu_k) \|_{L^2(U)} 2\sqrt{\mu_k}(\sqrt{\mu_k} - \sqrt{\mu_{k+1}}). \end{aligned} \tag{6.40}$$

Since the algorithm sets $\mu_{k+1} = \sigma\mu_k$, we are lead to the following step size rule.

$$[\omega_k(\mu_k)]\sigma^{-1/2}(\| u_k - u_{\mu_k} \| + \| [\eta_k(\mu_k)] \|_{L^2(U)} 2\mu_k(1 - \sqrt{\sigma})) = \theta_d. \tag{6.41}$$

If the Newton corrector was not successful, we use same equation for computing a more conservative μ_k. We simply replace $[\eta_k]$ and $\| u_k - u_{\mu_k} \|$ by the estimates of the previous (successful) iterate (see also [131]). In detail, the conservative step size selection reads

$$[\omega_k(\mu_k)]\sigma^{-1/2}(\| u_{k-1} - u_{\mu_k} \| + \| [\eta_{k-1}(\mu_{k-1})] \|_{L^2(U)} 2\mu_k(1 - \sqrt{\sigma})) = \theta_d. \tag{6.42}$$

Now we have everything at hand to implement a version of Algorithm 1.

Algorithm 2 Interior Point Method in Finite Dimensional Space

```
Choose parameters $\Lambda_d, \sigma_{min}, \sigma_{max}, \theta_c, \theta_d, \theta_t, tol_d, tol$.
Choose $(u_0, \mu_0)$
$k := 0$
$\varepsilon_k := tol + 1$
do
    $(\tilde{u}, success)$ :=NEWTON CORRECTOR$(u_k, \mu_k)$
    if success then                                      ▷ Implementing Algorithm 1 line 6
        compute $[\eta_k]$                                ▷ see (6.31)
        compute $\sigma_k \in [\sigma_{min}, \sigma_{max}]$           ▷ see (6.41)
        $\mu_{k+1} := \sigma_k \mu_k$
        $u_{k+1} := \tilde{u}$
        $k := k + 1$
        compute $[\| u_k - u^* \|] =: \varepsilon_k$              ▷ see (6.33)
    else
        if $k > 0$ then
            compute conservative $\sigma_k \in [\sigma_{min}, \sigma_{max}]$   ▷ see (6.42)
            $\mu_k := \sigma_k \mu_{k-1}$
            restore mesh
        else
            terminate: 'bad initial guess $(\mu_0, u_0)$'
        end if
    end if
while $\varepsilon_k > tol$
```

Let us close this section with some further remarks on the implementation. The value of σ_{min} is motivated by the best error reduction we can expect from uniform h-refinements. For elliptic equations on convex domains, the error decays like h^2. As the central path is Hölder continuous with index $1/2$, we set $\sigma_{min} = 1/16$ to facilitate an error reduction of $1/4$. If the mesh is refined r times during one Newton corrector step, we set $\sigma_{min} = 1/16^r$.

By demanding the relative error reduction Λ_d in the Newton corrector, the implemented algorithm converges at least linearly.

Since the algorithm may require to prolong an iterate u_k to a finer discretization. This is difficult for a representation on integration points, which is why we keep the adjoint q_k that is used for the smoothing the control with z_u (see Definition 6.2.3). As an FE function, q_k can be displayed exactly on a finer grid.

In order to find a sensible value for μ_0, we compute the Newton update $\delta\breve{u}$ for $\breve{u} := (u_a + u_b)/2$. Starting from $\mu_0 = 1$, we enlarge the initial homotopy parameter by $1/\sigma_{max}$ if $\| z_u(\delta\breve{u}) \|_{L^2(\Omega)} > 1$. For values smaller than 0.2, we decrease μ_0 by $\sigma_{\min}$. Otherwise, we start the algorithm with $z_u(\breve{u} + \delta\breve{u})$ and the computed μ_0. This way, we get slope information at the very first iterate from (6.31) because $\partial_\mu F(u_0) \neq 0$. To avoid a too aggressive σ_0, one may demand $\sigma_0 \geq \underline{\sigma} > 0$.

Algorithm 3 Newton corrector in Finite Dimensional Space with Mesh Refinement

```
procedure NEWTON CORRECTOR($u_k, \mu_k$)        ▷ Implementing Algorithm 1 line 3,5
    do                                              ▷ Newton Step
        do                                          ▷ Adaptive Refinement
            refine marked elements
            solve $\partial F_{\mu_k}(u_k)\delta u_k = -F_{\mu_k}(u_k)$
            compute $[\| e_k \|]$                   ▷ see Theorem 6.4.2
            mark elements
        while $[\| e_k \|]/\| \delta u_k \| < tol_d$
        compute $[\theta_k(\mu_k)]$                  ▷ see (6.28)
        $\tilde{u} := z_u(u_k + \delta u_k)$        ▷ see (6.8)
        compute $[\| \tilde{u} - u_{\mu_k} \|]$     ▷ see (6.37)
        success := $([\| \tilde{u} - u_{\mu_k} \|] < \Lambda_d \| \tilde{u} - u_k \| \wedge [\theta_k(\mu_k)] < \theta_t)$ ?
        failure := $([\theta_k(\mu_k)] > \theta_c)$ ?
        $u_k := z_u(u_k + \delta u_k)$              ▷ see (6.8)
    while not(success $\vee$ failure)
    return $(\tilde{u}, success)$
end procedure
```

6.4 A-Posteriori Error Estimators

The following error estimators exploit the structure of the optimality system. For treating more problems one can proceed as in [90, 91]. A different approach was taken in [128] to obtain a-posteriori error estimates for problems with additional state constraints.

6.4.1 Error to the Central Path

Let an approximate solution (y^h, u^h, q^h) of $(\mathbf{P}_\mu^h)$ be given. We will derive an upper bound of $\| u_\mu - u^h \|_{L^2(U)}$, which will be amenable for numerical realizations, see Subsection 6.4.3 below.

Theorem 6.4.1. *Let (y_μ, u_μ) be the solution to $(\mathbf{P}_\mu)$, $\mu > 0$. Let a discrete point (y^h, u^h, q^h) be given that satisfies $b'_\mu(u_h) \in L^2(U)$. Then there is a constant $c > 0$ independent of μ, h, and (y^h, u^h, q^h), such that*

$$\| u_\mu - u^h \|^2_{L^2(U)} \leq c \left(\|r_y\|^2_{Y^*} + \|r_q\|^2_{Y^*} + \| r_u \|^2_{L^2(U)} \right)$$

with

$$\begin{aligned} r_u &:= \nu u^h + B^* q^h + b'_\mu(u^h), \\ r_y &:= A y^h - B u^h, \\ r_q &:= A^* q^h - (y^h - y_d). \end{aligned}$$

Proof. Let q_μ be the adjoint state such that (6.1) is satisfied. Subtracting $\nu u^h + B^* q^h + b'_\mu(u^h)$ from both sides of (6.1b), multiplying with $u^h - u_\mu$, and integrating on U yields

$$\nu \| u^h - u_\mu \|^2_{L^2(U)} + \int_U (b'_\mu(u^h) - b'_\mu(u_\mu))(u^h - u_\mu)\,\mathrm{d}x + \int_U B^*(q^h - q_\mu)(u^h - u_\mu)\,\mathrm{d}x \\ = \int_U r_u(u_\mu - u^h)\,\mathrm{d}x.$$

Due to monotonicity of the subdifferential, the second term is non-negative. Using equations (6.1a) and (6.1c) we obtain

$$\begin{aligned} \int_U B^*(q^h - q_\mu)(u^h - u_\mu)\,\mathrm{d}x &= \langle A(y^h - y_\mu), q^h - q_\mu \rangle_{Y^*,Y} - \langle r_y, q^h - q_\mu \rangle_{Y^*,Y} \\ &= \langle A^*(q^h - q_\mu), y^h - y_\mu \rangle_{Y^*,Y} - \langle r_y, q^h - q_\mu \rangle_{Y^*,Y} \\ &= \| y^h - y_\mu \|^2_{L^2(\Omega)} + \langle r_q, y^h - y_\mu \rangle_{Y^*,Y} - \langle r_y, q^h - q_\mu \rangle_{Y^*,Y}. \end{aligned}$$

Combining with the previous estimate, we find

$$\nu \| u^h - u_\mu \|^2_{L^2(U)} + \| y^h - y_\mu \|^2_{L^2(\Omega)} \le \int_U r_u(u_\mu - u^h)\,\mathrm{d}x - \langle r_q, y^h - y_\mu \rangle_{Y^*,Y} + \langle r_y, q^h - q_\mu \rangle_{Y^*,Y}.$$

It remains to estimate $y^h - y_\mu$ and $q^h - q_\mu$. On account of the invertibility of A, it follows

$$\| y^h - y_\mu \|_Y \le c \| A y^h - A y_\mu \|_{Y^*} \le c(\| r_y \|_{Y^*} + \| u^h - u_\mu \|_{L^2(U)}).$$

Similarly, we deduce

$$\| q^h - q_\mu \|_Y \le c(\| r_q \|_{Y^*} + \| y^h - y_\mu \|_{L^2(\Omega)}).$$

Collecting all these estimates, we arrive at

$$\begin{aligned} &\nu \| u^h - u_\mu \|^2_{L^2(U)} + \| y^h - y_\mu \|^2_{L^2(\Omega)} \\ &\le \| r_u \|_{L^2(U)} \| u_\mu - u^h \|_{L^2(U)} + \| r_q \|_{Y^*} \| y^h - y_\mu \|_Y + \| r_y \|_{Y^*} \| q^h - q_\mu \|_Y \\ &\le \left(\| r_u \|_{L^2(U)} + c \| r_q \|_{Y^*} \right) \| u_\mu - u^h \|_{L^2(U)} + c \| r_y \|_{Y^*} \left(2 \| r_q \|_{Y^*} + \| y^h - y_\mu \|_{L^2(\Omega)} \right). \end{aligned}$$

The claim follows now with Young's inequality. □

6.4.2 Error in the Newton System

In addition to the results in the previous section, we will now derive an error estimator for the discretization error of the Newton step. Recall that the first-order necessary conditions (6.2) (or equivalently (6.1)) of ($\mathbf{P}_\mu$) are solved with Newton's method.

Let an iterate (y_k, u_k, q_k) be given. Then the Newton step $(\delta u, \delta y, \delta q)$ is computed as the solution of the system

$$A\delta y = B\delta u - (Ay_k - Bu_k), \tag{6.43a}$$

$$A^*\delta q = \delta y - (A^* q_k - (y_k - y_d)) \tag{6.43b}$$

$$\nu\delta u + B^*\delta q + b''_\mu(u_k)\delta u = -(\nu u_k + B^* q_k + b'_\mu(u_k)) \tag{6.43c}$$

This Newton system (6.4) is itself the optimality system of the following quadratic subproblem under linearized constraints:

($\mathbf{P}^q_\mu$)

$$\begin{cases} \text{minimize } J^q_\mu(\delta y, \delta u) := (\nu u_k + B^* q_k + b'_\mu(u_k), \delta u)_{L^2(U)} - (A^* q_k - y_k + y_d, \delta y)_{L^2(\Omega)} \\ \qquad\qquad + \dfrac{1}{2}(\| \delta y \|^2_{L^2(\Omega)} + \nu \| \delta u \|^2_{L^2(U)} + b''_\mu(\delta u, \delta u)) \\ \quad \text{subject to} \\ \quad A\delta y - B\delta u = -(Ay_k - Bu_k), \\ \qquad \delta u \in U_{ad}. \end{cases}$$

Since b''_μ is non-negative, the necessary conditions (6.43) are also sufficient. Solvability of the Newton system (Theorem 6.2.5) automatically proves the existence of a minimizer $(\delta y, \delta u)$ of ($\mathbf{P}^q_\mu$).

Let a discrete approximation $(\delta y^h, \delta u^h, \delta p^h)$ of $(\delta u, \delta y, \delta q)$ be given. We will now derive an a-posteriori error estimator for $\|\delta u^h - \delta u\|_{L^2(U)}$. In the notation of Section 6.3, we want to compute an estimate $[\|e_k\|]$ with $\| \delta u - \delta u^h \|_{L^2(U)} \le c[\|e_k\|]$.

Theorem 6.4.2. *Let (y_k, u_k, q_k) be given such that $b'_\mu(u_k) \in L^2(U)$. Let $(\delta u, \delta y, \delta q)$ be the solution of (6.43), and let $(\delta y^h, \delta u^h, \delta p^h)$ be a discrete approximation. Then there is a constant $c > 0$ independent of μ, h, and (y^h, u^h, q^h), $(\delta u, \delta y, \delta q)$, such that*

$$\| \delta u - \delta u^h \|^2_{L^2(U)} \le c \left(\|r_{\delta y}\|^2_{Y^*} + \|r_{\delta q}\|^2_{Y^*} + \| r_{\delta u} \|^2_{L^2(U)} \right)$$

with

$$\begin{aligned} r_{\delta u} &:= \nu(u_k + \delta u^h) + B^*(q_k + \delta q^h) + b'_\mu(u_k) + b''_\mu(u_k)\delta u^h, \\ r_{\delta y} &:= A(y_k + \delta y^h) - B(u_k + \delta u^h), \\ r_{\delta q} &:= A^*(q_k + \delta q^h) - (y_k + \delta y^h - y_d). \end{aligned}$$

Proof. The claim can be proved with similar arguments as Theorem 6.4.1. It exploits the non-negativity of b''_μ. □

Remark 6.4.3. *The solution algorithm in the previous section strongly relies on the smoothing operator Z. The latter guarantees that the current iterate u_k of the interior point method lies in D_2 and that the Newton system is invertible. Thus, $b'_\mu(u_k) \in L^2(U)$ and the Theorems 6.4.1, 6.4.2 are applicable for Algorithm 2.*

6.4.3 Residual Based *hp* Error Estimates

Let us explain the estimates of Y^*-norms of residuals in state and adjoint equations as they appear in Theorems 6.4.1 and 6.4.2. We exemplarily show the derivation for the residual $r_y = Ay^h - Bu^h \in Y^*$ of the state equation with operator A chosen to be $Ay = -\Delta y + y$ and $B = id$ for distributed controls.

Let now y^h be the solution of the discrete equation (6.18) to the control u^h. Then it holds

$$\langle r_y, v^h \rangle_{Y^*,Y} = \langle Ay^h - Bu^h, v^h \rangle_{Y^*,Y} = 0 \quad \forall v^h \in S^{\mathbf{p}}(\Omega, \tau).$$

Let $v \in H^1(\Omega)$, $v^h \in S^{\mathbf{p}}(\Omega, \tau)$ be given. Then we obtain using integration by parts

$$\begin{aligned}\langle r_y, v \rangle_{Y^*,Y} &= \langle r_y, v - v^h \rangle_{Y^*,Y} \\ &= \int_\Omega \nabla y^h \nabla (v - v^h) + y^h (v - v^h) - u^h (v - v^h) \, \mathrm{d}x \\ &= \sum_{K \in \tau} \int_K (-\Delta y^h + y^h - u^h)(v - v^h) \, \mathrm{d}x + \int_{\partial K \cap \Omega} \partial_n y^h (v - v^h) \, \mathrm{d}s \\ &= \sum_{K \in \tau} \left(\int_K (-\Delta y^h + y^h - u^h)(v - v^h) \, \mathrm{d}x + \frac{1}{2} \sum_{e \in \partial K \cap \Omega} \int_e \left[\partial_n y^h \right] (v - v^h) \, \mathrm{d}s \right),\end{aligned} \tag{6.44}$$

where $e \subset \partial K \cap \Omega$ is an abbreviation for the iteration over the set of all edges of an element K that are not part of the boundary Γ. Moreover, $[\phi]$ denotes the jump of the quantity ϕ across an edge e.

We will now choose $v^h := I_h v$, where I_h is an Clément type interpolation operator taken from [108]. Let us briefly introduce some notation to describe the approximation properties of I_h. For a vertex V of the triangulation τ, let us define the patches

$$\begin{aligned}\omega_V^0 &:= \{V\}, \\ \omega_V^j &:= \cup \{K \in \tau \mid K \cap \omega_V^{j-1} \neq \emptyset\}, \quad j \geq 1,\end{aligned}$$

and set

$$\begin{aligned}h_V &:= \min\{h_K \mid V \in K, \ K \in \tau\}, \\ p_V &:= \max\{p_K + 1 \mid V \in K, K \in \tau\}.\end{aligned}$$

Then the interpolation operator I_h of [108] satisfies

$$\| I_h(v) - v \|_{L^2(\omega_V^1)} + \frac{h_V}{p_V} \| \nabla I_h(v) \|_{L^2(\omega_V^1)} + \sqrt{\frac{h_V}{p_V}} \| I_h(v) - v \|_{L^2(e)} \leq C \frac{h_V}{p_V} \| \nabla v \|_{L^2(\omega_V^8)}, \tag{6.45}$$

where $e \subset K$ is an edge with one of its endpoints being V. For an element $K \in \tau$, let V_K denote a vertex of K. Then it holds $h_{V_K} \leq h_K$ and $p_{V_K} \geq p_K$. Using the interpolation operator I_h in (6.44) and employing (6.45) we estimate

$$\langle r_y, v\rangle_{Y^*,Y} \leq C \sum_{K\in\tau} \left(\frac{h_K}{p_K} \| -\Delta y^h + y^h - u^h \|_{L^2(K)} \right.$$
$$\left. + \frac{1}{2} \sum_{e\in\partial K\cap\Omega} \left(\frac{h_K}{p_K}\right)^{1/2} \| \left[\partial_n y^h\right] \|_{L^2(e)} \right) \| \nabla v \|_{L^2(\omega^8_{V_e})} \leq C \left(\sum_{K\in\tau} \eta_K^2 \right)^{1/2} \| v \|_{H^1(\Omega)}$$

where

$$\eta_K^2 := \frac{h_K^2}{p_K^2} \| -\Delta y^h + y^h - u^h \|^2_{L^2(K)} + \frac{1}{2} \sum_{e\in\partial K\cap\Omega} \frac{h_K}{p_K} \| \left[\partial_n y^h\right] \|^2_{L^2(e)} . \tag{6.46}$$

As $v \in H^1(\Omega)$ was arbitrary, this implies

$$\|r_y\|_{Y^*} = \|Ay^h - Bu^h\|_{Y^*} \leq C \left(\sum_{K\in\tau} \eta_K^2 \right)^{1/2} .$$

Due to the construction, this error estimator is reliable, thus providing an upper bound on the error. For results regarding local efficiency of the estimator, see [108].

The residual in the adjoint equation, $r_q = A^*q - (y^h - y_d)$, can be estimated using similar arguments. For general y_d one has to take integration errors into account, leading to estimates involving data oscillation term.

Remark 6.4.4. *The above error estimators would profit from an enhanced L^2 a-posteriori error estimators. If the problem is H^2-regular, i.e., $A, A^{-*} \in \mathcal{L}(L^2(\Omega), H^2(\Omega))$, then Theorems 6.4.1 and 6.4.2 are valid if the Y^*-norms of the residuals are replaced by $(Y \cap H^2(\Omega))^*$-norms. These then could be estimated by $L^2(\Omega)$-error estimators because $\|r_y\|_{(Y\cap H^2(\Omega))^*} \leq c\|A^{-1}r_y\|_{L^2(\Omega)}$. Unfortunately, no L^2-error estimators are available so far. In h-FEM, they are constructed with the Aubin-Nitsche trick and estimates of the type*

$$\| v - I(v) \|_{L^2(K)} \leq Ch_K^2 |v|_{H^2(K)}, \quad \| v - I(v) \|_{L^2(\partial K)} \leq Ch_K^{3/2} |v|_{H^2(K)},$$

see [2, Section 3.3] for details. Only suboptimal hp-equivalents are available (confer [135, Remark 4.70]) and the estimates

$$\| v - I(v) \|_{L^2(K)} \leq C \left(\frac{h_K}{p_K}\right)^2 |v|_{H^2(K)}, \quad \| v - I(v) \|_{L^2(\partial K)} \leq C \left(\frac{h_K}{p_K}\right)^{3/2} |v|_{H^2(K)}$$

are expected to be true but remain to be proved.

6.5 Numerical Examples

In the following, we present numerical results of the discretized interior point method. First, we investigate the ability of the algorithm to identify regions where the optimal control is non-smooth. If u^* is continuous, the active set $\mathfrak{A}$ is uniquely defined and we expected fast convergence if the algorithm strongly refines the mesh towards $\partial\mathfrak{A}$, i.e., the interface where u^* changes from active to inactive and vice versa.

Second, we solve a problem with a very small regularization parameter. Here, the region of convergence for Newton's method is small and sensible to perturbations in μ. This fact allows us to examine the performance of the adaptive step size selection of the discrete algorithm.

6.5.1 Testing *hp*-Adaptivity

Example 1: Neumann Control Problem Revisited

We solve the problem (5.10) with data (5.12). We recall that for small $\varepsilon > 0$ the problem is $H^{5/3-\varepsilon}$-regular because it is posed on an L-shaped domain with a reentrant corner of angle $\frac{3}{2}\pi$.

The parameters of the path-following method are chosen as follows:

$$\theta_d = 0.1,\ \theta_t = 0.5,\ \theta_c = 0.8, \quad \sigma_{max} = 0.9,\ \sigma_{min} = \frac{1}{4}.$$

with the tolerances

$$tol_d = 0.7, \quad tol = 10^{-3}, \quad \Lambda_d = 0.7.$$

We start the path-following method with the homotopy parameter $\mu_0 = 0.05$ on an initial mesh consisting of 12 elements of degree two.

In order to investigate the ability of Algorithm 2 to adapt to the regularity properties of the optimal variables, we refine the mesh uniformly: either h- or p-refinement takes place on each element.

For elements abutting the Neumann boundary, we expand the control u in a Legendre series and estimate the decay of the Legendre coefficients. This allows to judge the smoothness and guide between h- or p-refinement (see [59] and Subsection 6.3.2). The same technique is applied to the remaining elements using the variables y, q. We favor h refinement and only increase the polynomial degree if both variables seem to be smooth.

The problem turns out to have a large region of convergence for Newton's method, which in addition is relatively robust with respect to changes in the homotopy parameter μ. That is why the method rapidly decreases μ_k and finds the solution in only 4 iterates. This fast convergence implies that many refinements are necessary during the Newton corrector iteration (Algorithm 3). Due to the observed quadratic convergence of the norm of the updates $\| \delta u_k \|$, the mesh is refined several times until the relative error $[\| e_k \|]/\| \delta u_k \| < tol_d$ is small, see steps 3–8 in Algorithm 3. This enforced refinement ensures that the discretization error does not prevail, and that linear convergence is achieved in function space.

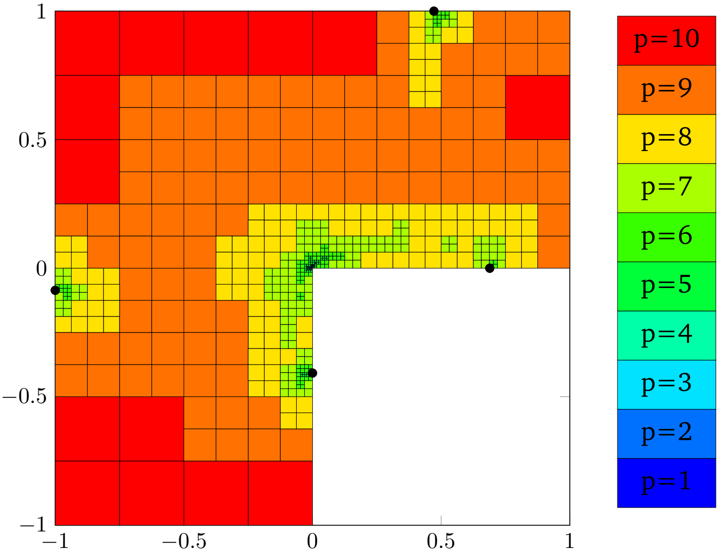


Figure 6.1. The adaptive mesh from Algorithm 2 after solving (5.10) with data (5.12). The kinks in the optimal control are marked with black points.

The final mesh at $\mu_4 \approx 2.29 \cdot 10^{-6}$ is depicted in Figure 6.1 and comprises $27,951$ ddof and 816 integration points at the Neumann boundary. We see that the algorithm identifies the re-entrant corner as an area where the variables are non-smooth. Furthermore, the elements are heavily refined towards the kinks of the optimal control (Figure 6.1).

The final number of integrations points for the boundary control is 816 and the errors read

$$\begin{aligned} \| u^* - u_{4,h} \|_{L^2(\Gamma_{\mathcal{N}})} &\approx 2.29 \cdot 10^{-3}, \\ \| y^* - y_{4,h} \|_{H^1(\Omega)} &\approx 8.72 \cdot 10^{-5}, \\ \| q^* - q_{4,h} \|_{H^1(\Omega)} &\approx 1.76 \cdot 10^{-3}. \end{aligned}$$

Regarding accuracy with respect to the number of unknowns, the interior point algorithm performs weaker than the discretization of the vc-FEM (see Table 5.4). This is only seems natural because a sequence of perturbed problems is solved with adaptive discretization. Regarding convergence with respect to $\mu \searrow 0$, we observe the rate $\mathcal{O}(\sqrt{\mu})$ of Theorem 6.1.10 (see Figure 6.2).

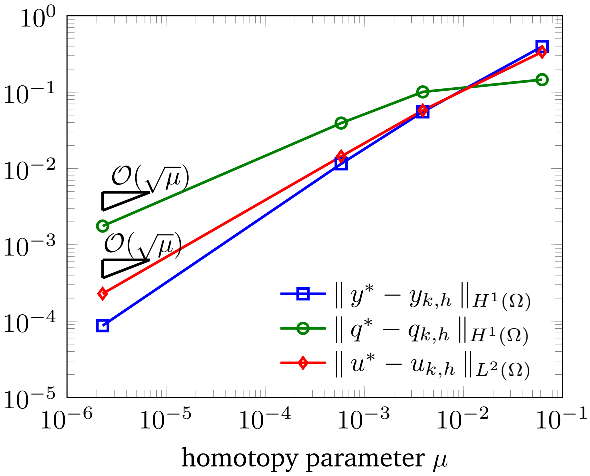


Figure 6.2. The convergence history of Algorithm 2 and problem (5.10) with data (5.12).

Example 2: Distributed Control Problem

Now we test Algorithm 2 for a distributed control problem on a convex domain. As a faster convergence in the L^2-norm has been observed in our previous numerical experiments of the bc-FEM, we modified η_K in (6.46) to contain the weights $(h_K/p_K)^4, (h_K/p_K)^3$ instead of $(h_K/p_K)^2, (h_K/p_K)$. This mimics an enhanced a-posteriori estimator as mentioned in Remark 6.4.4.

The problem under consideration is taken from [144] and reads

$$\text{minimize } J(u,y) = \frac{1}{2}\| y - y_d \|^2_{L^2(\Omega)} + \frac{\nu}{2}\| u \|^2_{L^2(\Omega)} + (e_q, y)_{L^2(\Gamma)} \tag{6.47a}$$

subject to

$$\begin{aligned} -\kappa\Delta y + y &= u + e_y && \text{in } \Omega, && (6.47b)\\ \partial_n y &= 0 && \text{on } \Gamma_{\mathcal{N}}, && (6.47c)\\ y &= 0 && \text{on } \Gamma_{\mathcal{D}}, && (6.47d) \end{aligned}$$

and $u_a \leq u \leq u_b$ a.e. in Ω for real numbers $u_a < u_b$. The problem is posed on

$$\Omega = (0,1)^2, \quad \Gamma_{\mathcal{N}} = \Gamma, \quad \Gamma_{\mathcal{D}} = \emptyset \tag{6.48a}$$

with the data $\nu = \kappa = 1,\ u_a = 0,\ u_b = 1$, and

$$y_d = -142/3 + 12\,\mathrm{dist}(x, x_0)^2, \quad e_q \equiv -12. \tag{6.48b}$$

The optimal solution is radially symmetric with origin in $x_0 = (0.5, 0.5)$ and reads

$$\begin{aligned} u^* &= P_{U_{ad}}(-\nu^{-1}q^*), \\ y^* &\equiv 1, \\ q^* &= -12\,\mathrm{dist}(x, x_0)^2 + 1/3, \end{aligned}$$

if the inhomogeneity e_y is set to $e_y = 1 - u^*$. The previous analysis is not affected by the appearance of e_y, e_q.

The parameters of the algorithm are

$$\theta_d = 0.1,\ \theta_t = 0.5,\ \theta_c = 0.8, \quad \sigma_{max} = 0.9,\ \sigma_{min} = \frac{1}{16}$$

with the tolerances

$$tol_d = 0.5, \quad tol = 0.005, \quad \Lambda_d = 0.6.$$

We start path-following method with the homotopy parameter $\mu_0 = 1$ on a mesh consisting of four elements of degree 2.

We apply the same solution strategy as before and either p or h refine an element if the algorithm decides to reduce the discretization error. The hp refinement method judges the smoothness of the control iterate u_k by expanding it in a Legendre series and estimating the decay of the Legendre coefficients (see [59] and Subsection 6.3.2). If the coefficients decay fast enough, the element will be p-refined, otherwise it will be h-refined.

The distributed control problem displays similar properties to the Neumann control problem on the L-shape domain. It is characterized by the fact that Newton's method has a large region of convergence that is robust with respect to changes in μ. Again, the mesh is refined several times to keep the relative discretization error small. This way linear convergence in function space is guaranteed because the discretization error does not prevail (confer steps 3–8 in Algorithm 3).

The adaptive path-following algorithm performs very well and nicely captures the interface γ where the optimal control is non-smooth, namely

$$\gamma = \{\mathrm{dist}(x - x_0) = \frac{1}{6}\} \cup \{\mathrm{dist}(x - x_0) = \frac{1}{3}\}.$$

In Figure 6.3 we depict the final mesh for $\mu_4 \approx 2.87 \cdot 10^{-5}$ consisting of $149,613$ ddof and $263,216$ integration points. In addition the interface γ is drawn as white circles. We observe a strong h-refinement near the interface, whereas large parts of the active and inactive sets are p-refined.

The convergence behavior is very similar to the previous example and the final errors read

$$\begin{aligned} \| u^* - u_{4,h} \|_{L^2(\Gamma_\mathcal{N})} &\approx 4.56 \cdot 10^{-4}, \\ \| y^* - y_{4,h} \|_{H^1(\Omega)} &\approx 2.50 \cdot 10^{-5}, \\ \| q^* - q_{4,h} \|_{H^1(\Omega)} &\approx 1.17 \cdot 10^{-5}. \end{aligned}$$

The full convergence history is depicted in Figure 6.4.

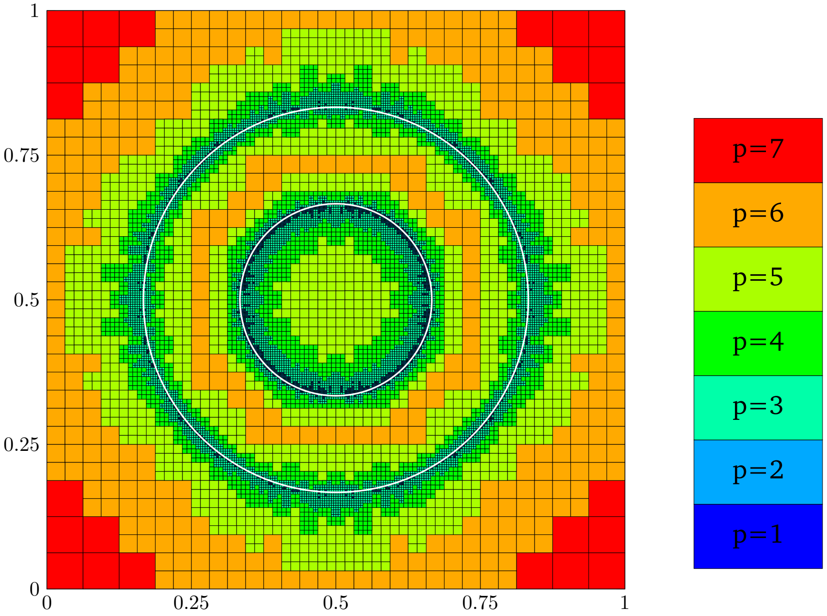


Figure 6.3. The adaptive mesh from Algorithm 2 after solving problem (6.47) with data (6.48). The interface γ is marked by white circles.

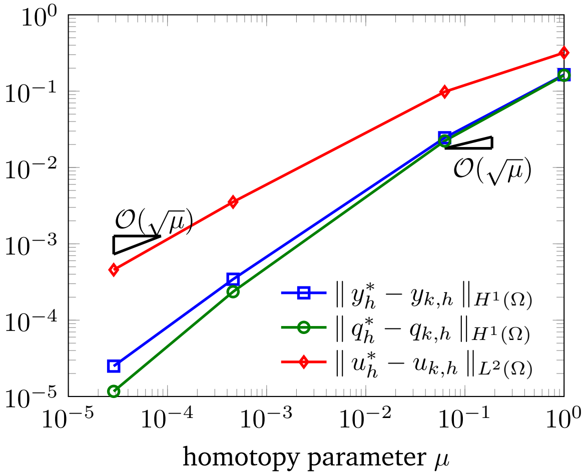


Figure 6.4. The convergence history of Algorithm 2 and problem (6.47) with data (6.48).

6.5.2 Testing Adaptive Path-Following

This test case is taken from [130] and turns out to be numerically challenging because of the very small regularization parameter ν. Unlike the previous examples, the radius of convergence around the central path is small and very sensitive with respect to changes in μ. Problem (6.47) is posed on

$$\Omega = (0,1)^2, \quad \Gamma_{\mathcal{N}} = \emptyset, \quad \Gamma_{\mathcal{D}} = \Gamma \tag{6.49a}$$

with data

$$\nu = 10^{-6},\ \kappa = \frac{1}{10},\ u_a = 0,\ u_b = 1. \tag{6.49b}$$

The desired state is rough with patch-wise behavior

$$y_d = 0.01 \cdot \chi_{(-1,0.2)\times(-1,0.6)} - 0.01 \cdot \chi_{(-1,0.2)\times(-0.6,1)} + 0.02 \cdot \chi_{(0.2,1)\times(-0.6,1)}. \tag{6.49c}$$

No inhomogeneities are used ($e_y \equiv e_q \equiv 0$). We choose the contraction parameters a little stricter.

$$\theta_d = 0.1,\ \theta_t = 0.3,\ \theta_c = 0.5, \quad \sigma_{max} = 0.9,\ \sigma_{min} = \frac{1}{16}.$$

Similar as before,

$$tol_d = 0.5, \quad tol = 10^{-2}, \quad \Lambda_d = 0.6.$$

A too aggressive σ_0 is avoided by $\sigma_0 \geq \underline{\sigma} = 1/4$. The path-following method is launched with $\mu_0 = 16^{-3}$ on a mesh consisting of 64 elements of degree three.

Since the error estimators of Section 6.4 suffer from small ν, we rescaled them to prohibit extensive mesh refinement at the beginning of the algorithm. If a mesh refinement is necessary, we adaptively refine the mesh according to Subsection 6.3.2.

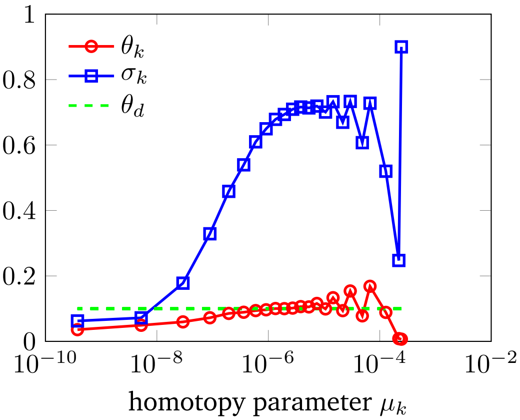


Figure 6.5. The step sizes and contraction factors of Algorithm 2 and problem (**P**) with data (6.49).

The algorithm manages to stay inside the convergence radius of Newton's method by choosing relatively large values for σ. This behavior occurs because of the high nonlinearity of the problem and the small convergence area for Newton's method resulting from a very small ν. The desired contraction θ_d is achieved nicely (Figure 6.5).

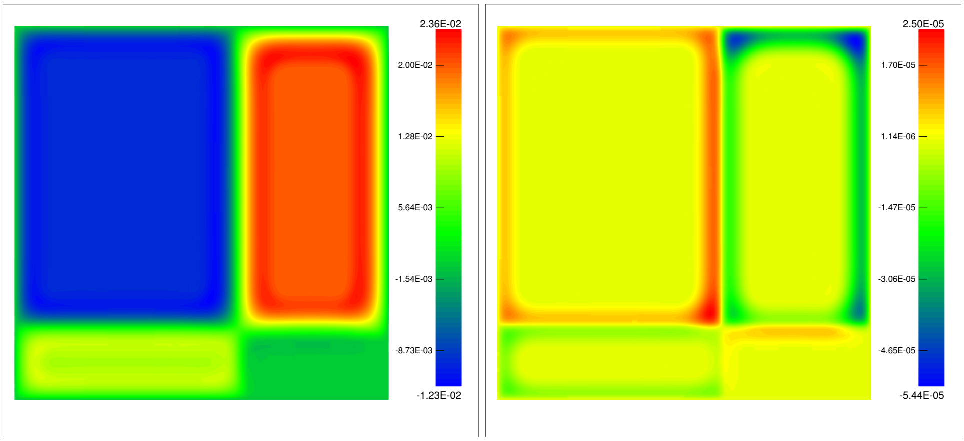


Figure 6.6: The state (left) and adjoint (right) variable on the final discretization of Algorithm 2 and problem (**P**) with data (6.49).

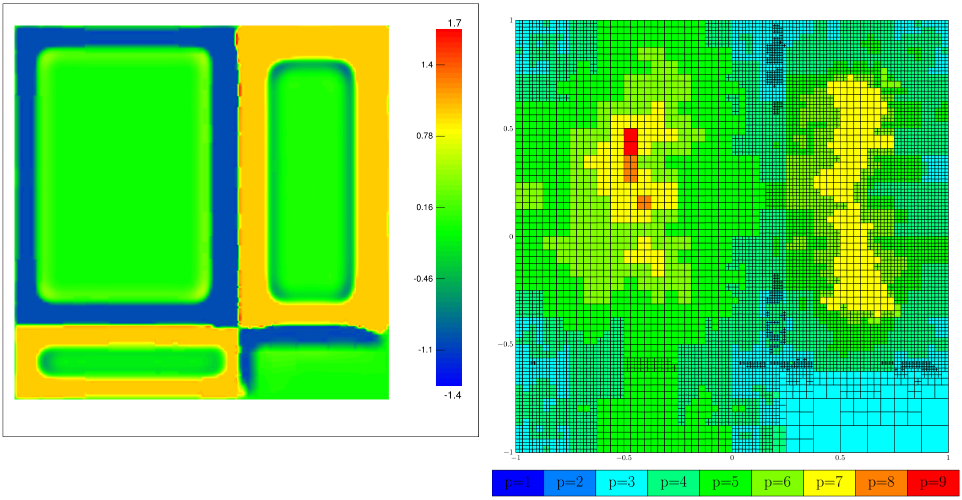


Figure 6.7: The control (left) variable on the final discretization (right) of Algorithm 2 and problem (**P**) with data (6.49).

The approximate optimal state and adjoint variable are shown in Figure 6.6. Additionally, Figure 6.7 displays the discrete optimal control and the final mesh ($208,550$ ddof and $312,552$ integration points). We see that the optimal control displays jumps and exhibits a patchwise behavior similar to y_d. Since the smoothness of u guided between h- and p-refinement, we roughly recover the structure of the optimal control in the final mesh.

On the whole, the numerical test examples clearly show that the adaptive hp-refinement can be successfully integrated within a Newton path-following method to solve the inequality constrained control problem.

CHAPTER 7

Conclusion and Outlook

We successfully used the hp-finite element method (hp-FEM) as a discretization technique for linear quadratic optimal control problems. The constraints consisted of an elliptic PDE as well as pointwise bounds on the control.

7.1 Semi-Smooth Newton Methods

The semi-smooth Newton method is capable of efficiently solving problems with control constraints by tackling the non-smooth optimality system. A general interpolation operator has been presented for the boundary concentrated finite element method (bc-FEM), which lead to energy-norm error estimates of algebraic order. The results were extended to the interface concentrated finite element method (ic-FEM) for interface control problems, which suffers from poor regularity. An expansion of solutions into a regular part and singular components can be derived with the help of the eigenvalues of the Mellin-transformed differential operator and the concept of injectivity modulo polynomials.

We also presented new error estimates in the L^2- and L^∞-norm at the boundary of the domain. The result formed the basis of an a-priori update rule for the regularization parameter in the context of problems with bang-bang character.

Additionally, we established a novel regularity result in countably normed spaces with a weight function that includes the vertices of the domain and the points where the optimal control exhibits kinks. Combined with known approximation results from the hp-FEM, a mixed a-priori a-posteriori refinement strategy is capable of achieving exponential convergence with respect to the number of unknowns (vc-FEM). We also compared the different hp-strategies to each other as well as to standard h-FEM methods. The numerical examples underline the fact that higher order methods are very efficient with respect to the number of unknowns.

Many of the presented results are expected to hold also in three space dimensions. Unfortunately, their proofs are expected to be even more technical, which is why the numerical analysis for this case is far less complete than for the $2d$ case. We mention [45, 47, 73, 74, 88].

The idea of ic-FEM can be extended to problems with distributed control where the active and inactive set are determined a-posteriori. A mesh would look similar to the one in Figure 6.3. A-priori information can be used to resolve corner singularities. This procedure would be analogue to vc-FEM and can be transferred to boundary control problems of $3d$ problems.

7.2 Path-Following Methods

Additionally to the mainly a-priori motivated refinement strategies, we were able to present a fully adaptive interior point method that allows to solve problems with control constraints. By monitoring functions at integration points, we guaranteed strict feasibility and obtained a sequence of solutions that converges to the optimal solution. To the best of our knowledge, this is the first algorithm for optimal control problems that works with hp-FEM and pure a-posteriori information. We tested adaptivity as well as path-following performance for different examples and obtained good results.

From the point if view of applications, it is interesting to solve problems on higher dimensional domains ($\Omega \subset \mathbb{R}^d, d > 2$). Since the interior point method and the hp-FEM also cover the three-dimensional setting, it is assumed that a working algorithm can be implemented in higher space dimensions, too.

In addition, more general cost functionals can be considered. The convergence of the interior point method carries over to general functionals provided that sufficient second-order optimality conditions are satisfied [148]. If the cost functional is analytic in y and u, one can expect that the convergence rate of the hp-method is unaffected.

7.3 Non-Linear State Equations

From the applicational point of view, it is interesting to extend the hp-idea to non-linear control problems. This is challenging because hp meshes are heavily adapted to a specific form of the active/inactive set which can change considerably along the solution iterates. Coarsening techniques and the use of an underlying base grid, as used for multi-grid methods, come to our mind. Alternatively, critical regions can be discretized with low order elements, similar to the bc- and ic-FEM. Confer [158], where a heating problem with non-local radiation operator is solved.

In some cases, a direct application is possible as the following example shows.

$$
\text{(S)} \qquad \left\{ \begin{array}{l} \text{minimize } J(u,y) := \dfrac{1}{2}\| \, y - y_d \, \|^2_{L^2(\Omega)} + \dfrac{\nu}{2}\| \, u \, \|^2_{L^2(\Gamma)} \\ \qquad\quad \text{subject to} \\ \quad -\Delta y + y^3 + y = f \quad \text{on } \Omega, \\ \qquad\qquad\quad\; \partial_n y = u \quad \text{on } \Gamma, \\ \qquad\qquad\qquad\; u \in U_{ad}. \end{array} \right.
$$

This semi-linear control problem is well posed and the optimality conditions can be reformulated as a semi-smooth projection formula (see [144, 146]).

We expect that the techniques of Section 3.3 can be used to establish a similar regularity result in countably normed spaces. Together with the approximation quality of geometric meshes, we assume that a numerical solutions converges with order $\mathcal{O}(e^{-b\sqrt[3]{N}})$ for $b > 0$. At least the numerical results suggest that exponential convergence is obtained (see Figure 7.1). The data is chosen as

$$\nu = 0.5, \quad u_a = -0.3, \quad y_d = 4x_1^2x_2^2, \quad f \equiv 2.$$

The solution on the finest discretization is taken as a reference for computing the errors.

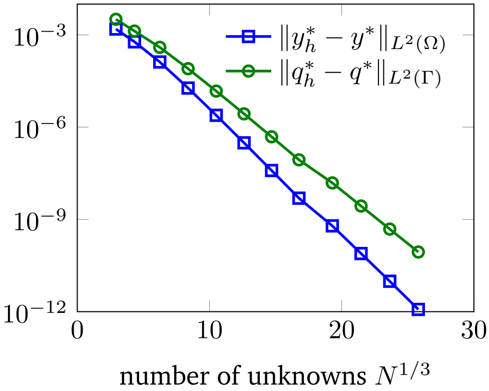


Figure 7.1. The error decay of the vc-FEM applied to the semi-linear problem (S).

Note that the solution method that we used in Chapter 5 is designed for positive definite matrices (see [28, Equation (4.7)]). In the context of non-linear equations, this is not necessarily the case any more. For simplicity, we chose ν large enough to avoid this issue.

7.4 State Constraints

Another aspect that often appears in the context of real-life applications, is the presence of state constraints. These model the fact that the controlled system shall stay within a feasible set of states. Instead of $u \in U_{ad}$, we demand

$$y \in Y_{ad} := \{y \in C(\Omega) \mid y_a \leq y \leq y_b\}, \tag{7.1}$$

where $y_a, y_b \in C(\Omega)$ with $y_a < y_b$ pointwise. It is well known that the challenge of such model problems is the fact that the Lagrange multipliers corresponding to (7.1) have low regularity. They can be identified with regular Borel measures, which are hard to approximate. Solution techniques range from regularization techniques (see, e.g, [23, 25, 77, 110]) to interior point methods (see, e.g., [82, 130, 131, 133]).

We can transfer the ideas of Chapter 6 and obtain a homotopy method for state constraints. The main issue is to guarantee feasibility of the state iterates. This can be ensured by a damped Newton step as explained in [131, Equation (40)]. Unfortunately, no smoothing operator like the one in (6.8) arises naturally from the optimality conditions. Consequently, the prolongation of solutions to finer meshes may become infeasible at the integration points if a naive implementation is chosen.

A-posteriori error estimators can be developed similar as in Section 6.4 by exploiting the optimality system. Depending on the computations, it is necessary to switch between an FE representation and a pointwise representation. Sophisticated techniques are necessary in order to retain feasibility and enforce Dirichlet boundary conditions.

We implemented a simplified interior point method, that uses a fixed step size reduction for $\mu_k = \sigma\mu_{k-1}$. Moreover, we dispensed with grid adaptivity and refined the mesh in each iteration. The obtained algorithm was applied to a test example from [82].

The computational domain is the unit square $\Omega = (0,1)^2$ with $\Gamma_\mathcal{N} = \partial\Omega$ and the elliptic constraint reads $-\Delta y + y = u$. Besides,

$$\nu = 10^{-3}, \quad y_d = 2x_1x_2, \quad y_b = 0.55.$$

Starting with $\mu_0 = 0.1, \sigma = 0.25$ on a uniform mesh with four quadrilateral elements of degree two, we obtain promising results after 7 iterations. The optimal variables (see Figure 7.2) look like those in [82] and the interface is captured relatively nicely (Figure 7.3).

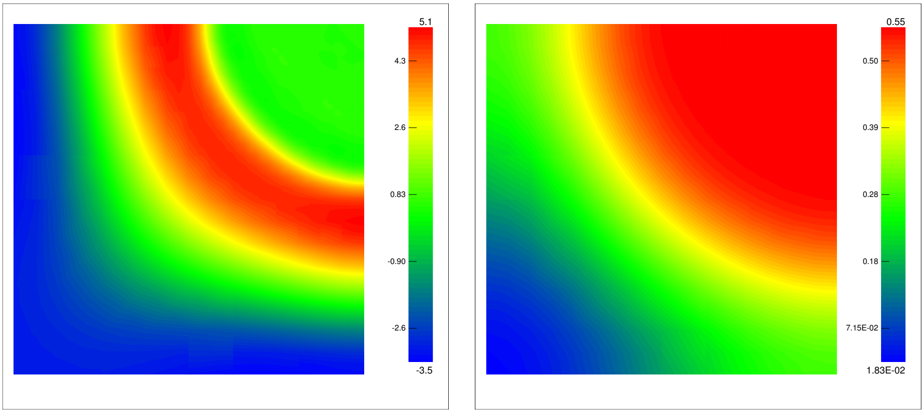


Figure 7.2: The optimal control (left) and state (right) variable for the state constrained problem.

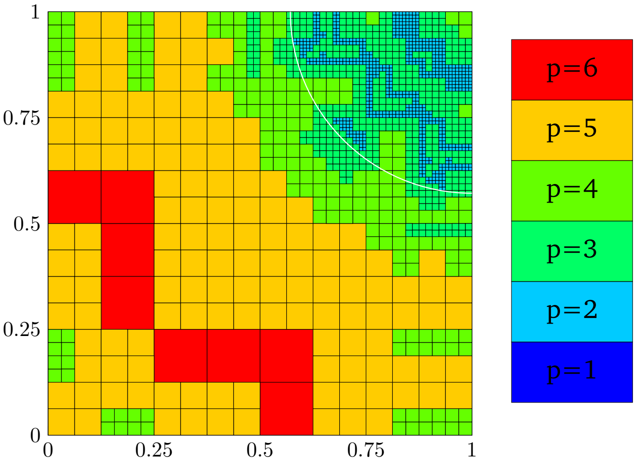


Figure 7.3. The final mesh of the interior point method applied to the state constrained problem. The approximate interface is marked with a white arc.

Hence, we are optimistic that the idea of using higher order elements for the numerical solution of state constrained problems is principally possible. However, numerical test showed that the implementation of a stable and fully adaptive path-following algorithm is difficult. A success could also lead to solvers that can handle *both* state and control constraints.

7.5 Miscellaneous

Time dependent problems have been successfully solved by higher order methods (see [94, 141, 142] and the references therein). It would be interesting to employ the known results in the context of non-stationary optimal control problems and investigate the approximation quality.

Additionally, the applicability of higher order methods to Dirichlet control problems can be studied. The theoretical and numerical analysis is challenging because the Dirichlet boundary condition is hard to integrate into the variational setting. Furthermore, the projection formula contains the normal derivative of the adjoint variable, which makes optimal convergence rates difficult to achieve. Often, very weak formulations of the state equation are used (see, e.g., [97, 105, 150] and the references therein).

In this work, all experiments were conducted on one spatial discretization for all variables. Working with different meshes (confer [138, 140]) for the state and adjoint variable could lead to higher numerical accuracy. Multi-grid methods could be an interesting approach for iterative solution algorithms.

Bibliography

1. R.A. Adams, *Sobolev spaces*, Pure and Applied Mathematics, vol. 65, Academic Press, New York London, 1975.

2. M. Ainsworth and J.T. Oden, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg. **142** (1997), no. 1-2, 1–88.

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** (1997), no. 1-4, 65–87.

4. T. Apel, J. Pfefferer, and A. Rösch, *Finite element error estimates for Neumann boundary control problems on graded meshes*, Comput. Optim. Appl. **52** (2012), no. 1, 3–28.

5. T. Apel, J. Pfefferer, and A. Rösch, *Finite element error estimates on the boundary with application to optimal control*, Math. Comp. **84** (2015), 33–70.

6. J. Appell and P.P. Zabrejko, *Nonlinear superposition operators*, Cambridge Tracts in Mathematics, vol. 95, Cambridge University Press, Cambridge, 1990.

7. N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl. **23** (2002), no. 2, 201–229.

8. I. Babuška, M. Griebel, and J. Pitkäranta, *The problem of selecting the shape functions for a p-type finite element*, Internat. J. Numer. Methods Engrg. **28** (1989), no. 8, 1891–1908.

9. I. Babuška and B.Q. Guo, *The h-p version of the finite element method. Part 1: The basic approximation results*, Comput. Mech. **1** (1986), no. 1, 21–41.

10. I. Babuška and B.Q. Guo, *The h-p version of the finite element method. Part 2: General results and applications*, Comput. Mech. **1** (1986), no. 3, 203–220.

11. I. Babuška and B.Q. Guo, *The h-p version of the finite element method for domains with curved boundaries*, SIAM J. Numer. Anal. **25** (1988), no. 4, 837–861.

12. I. Babuška and B.Q. 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** (1988), no. 1, 172–203.

13. I. Babuška and B.Q. 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** (1989), no. 4, 763–781.

14. I. Babuška and B.Q. Guo, *Approximation properties of the h-p version of the finite element method*, Comput. Methods Appl. Mech. Engrg. **133** (1996), no. 3-4, 319–346.

15. I. Babuška, R.B. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math. **33** (1979), no. 4, 447–471.

16. I. Babuška and W.C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal. **15** (1978), no. 4, 736–754.

17. I. Babuška, B.A. Szabo, and I.N. Katz, *The p-version of the finite element method*, SIAM J. Numer. Anal. **18** (1981), no. 3, 515–545.

18. C.E. Baumann and J.T. Oden, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg. **175** (1999), no. 3-4, 311–341.

19. R. Becker, M. Braack, D. Meidner, R. Rannacher, and B. Vexler, *Adaptive finite element methods for PDE-constrained optimal control problems*, Reactive Flows, Diffusion and Transport, Springer, Berlin, 2007, pp. 177–205.

20. R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer. **10** (2001), 1–102.

21. T. Belytschko, R. Gracie, and G. Ventura, *A review of extended/generalized finite element methods for material modeling*, Modelling Simul. Mater. Sci. Eng. **17** (2009), no. 4, 043001 (24pp).

22. J. Bergh and J. Löfström, *Interpolation spaces - An introduction*, vol. 223, Springer, Berlin New York, 1976, Die Grundlehren der mathematischen Wissenschaften.

23. M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of a Moreau-Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim. **11** (2000), no. 2, 495–521.

24. M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim. **37** (1999), no. 4, 1176–1194.

25. M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl. **22** (2002), no. 2, 193–224.

26. S. Beuchler, *Multi-level methods for degenerated problems with applications to p-versions of the fem*, Dissertation, TU Chemnitz, 2003.

27. S. Beuchler, K. Hofer, D. Wachsmuth, and J.-E. Wurst, *Boundary concentrated finite elements for optimal control problems with distributed observation*, Comput. Optim. Appl. (2015), 31–65.

28. S. Beuchler, C. Pechstein, and D. Wachsmuth, *Boundary concentrated finite elements for optimal boundary control problems of elliptic PDEs*, Comput. Optim. Appl. **51** (2012), no. 2, 883–908.

29. S. Beuchler, V. Pillwein, J. Schöberl, and S. Zaglmayr, *Sparsity optimized high order finite element functions on simplices*, Numerical and Symbolic Scientific Computing, Texts Monogr. Symbol. Comput., Springer, Wien, 2012, pp. 21–44.

30. S. Beuchler, V. Pillwein, and S. Zaglmayr, *Fast summation techniques for sparse shape functions in tetrahedral hp-FEM*, Domain Decomposition Methods in Science and Engineering XX, Lect. Notes Comput. Sci. Eng., vol. 91, Springer, Berlin Heidelberg, 2013, pp. 511–518.

31. S. Beuchler and J. Schöberl, *New shape functions for triangular p-FEM using integrated Jacobi polynomials*, Numer. Math. **103** (2006), no. 3, 339–366.

32. K.S. Bey and J.T. Oden, *hp-version discontinuous Galerkin methods for hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg. **133** (1996), no. 3-4, 259–286.

33. J.F. Bonnans and F.J. Silva, *Asymptotic expansion for the solutions of control constrained semilinear elliptic problems with interior penalties*, SIAM J. Control Optim. **49** (2011), no. 6, 2494–2517.

34. D. Braess, *Finite elements - Theory, fast solvers, and applications in elasticity theory*, 3rd ed., Cambridge University Press, Cambridge, 2007.

35. S.C. Brenner and L.R. Scott, *The mathematical theory of finite element methods*, 3rd ed., Texts in Applied Mathematics, vol. 15, Springer, New York, 2008.

36. A. Cangiani, E.H. Georgoulis, and P. Houston, *hp-Version discontinuous Galerkin methods on polygonal and polyhedral meshes*, Math. Models Methods Appl. Sci. **24** (2014), no. 10, 2009–2041.

37. G.F. Carey and E. Barragy, *Basis function selection and preconditioning high degree finite element and spectral methods*, BIT **29** (1989), no. 4, 794–804.

38. E. Casas and M. Mateos, *Error estimates for the numerical approximation of Neumann control problems*, Comput. Optim. Appl. **39** (2008), no. 3, 265–295.

39. E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl. **31** (2005), no. 2, 193–219.

40. Y. Chang and D. Yang, *Superconvergence for optimal control problem governed by nonlinear elliptic equations*, Numer. Funct. Anal. Optim. **35** (2014), no. 5, 509–538.

41. Y. Chen and Y. Lin, *A posteriori error estimates for hp finite element solutions of convex optimal control problems*, J. Comput. Appl. Math. **235** (2011), no. 12, 3435–3454.

42. Y. Chen, N. Yi, and W. Liu, *A Legendre-Galerkin spectral method for optimal control problems governed by elliptic equations*, SIAM J. Numer. Anal. **46** (2008), no. 5, 2254–2275.

43. P.G. Ciarlet, *The finite element method for elliptic problems*, vol. 4, North-Holland, Amsterdam, 1978, Studies in Mathematics and its Applications.

44. B. Cockburn, G.E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, Discontinuous Galerkin methods, Lect. Notes Comput. Sci. Eng., vol. 11, Springer, Berlin, 2000, pp. 3–50.

45. M. Costabel and M. Dauge, *General edge asymptotics of solutions of second-order elliptic boundary value problems. I, II*, Proc. Roy. Soc. Edinburgh Sect. A **123** (1993), no. 1, 109–155, 157–184.

46. M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, Math. Model. Numer. Anal. **33** (1999), no. 3, 627–649.

47. M. Costabel, M. Dauge, and S. Nicaise, *Analytic regularity for linear elliptic systems in polygons and polyhedra*, Math. Models Methods Appl. Sci. **22** (2012), no. 8, 1250015, 63pp.

48. C.L. Darby, W.W. Hager, and A.V. Rao, *An hp-adaptive pseudospectral method for solving optimal control problems*, Optim. Control Appl. Meth. **32** (2011), no. 4, 476–502.

49. M. Dauge, *Elliptic boundary value problems on corner domains*, Lecture Notes in Mathematics, vol. 1341, Springer, Berlin Heidelberg, 1988, Smoothness and asymptotics of solutions.

50. K. Deckelnick and M. Hinze, *A note on the approximation of elliptic control problems with bang-bang controls*, Comput. Optim. Appl. **51** (2012), no. 2, 931–939.

51. L. Demkowicz, *Computing with hp-adaptive finite elements - One and two dimensional elliptic and Maxwell problems*, Applied Mathematics and Nonlinear Science Series, vol. 1, Chapman & Hall/CRC, Boca Raton, FL, 2007.

52. L. Demkowicz, J. T. Oden, W. Rachowicz, and O. Hardy, *Toward a universal h-p adaptive finite element strategy. I. Constrained approximation and data structure*, Comput. Methods Appl. Mech. Engrg. **77** (1989), no. 1-2, 79–112.

53. L. Demkowicz et al., *Computing with hp-adaptive finite elements - Frontiers: Three dimensional elliptic and Maxwell problems with applications*, Applied Mathematics and Nonlinear Science Series, vol. 2, Chapman & Hall/CRC, Boca Raton, FL, 2008.

54. P. Deuflhard, *Newton methods for nonlinear problems - Affine invariance and adaptive algorithms*, Springer Series in Computational Mathematics, vol. 35, Springer, Heidelberg, 2011.

55. M. Dobrowolski, *Numerical approximation of elliptic interface and corner problems*, Habilitation, Friedrich-Wilhelms-Universität Bonn, 1981.

56. M. Dryja, M.V. Sarkis, and O.B. Widlund, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math. **72** (1996), no. 3, 313–348.

57. T. Eibner, *Randkonzentrierte und adaptive hp-fem*, Dissertation, TU Chemnitz, 2006.

58. T. Eibner and J.M. Melenk, *A local error analysis of the boundary-concentrated hp-FEM*, IMA J. Numer. Anal. **26** (2006), no. 4, 752–778.

59. T. Eibner and J.M. Melenk, *An adaptive strategy for hp-FEM based on testing for analyticity*, Comput. Mech. **39** (2007), no. 5, 575–595.

60. T. Eibner and J.M. Melenk, *Multilevel preconditioning for the boundary concentrated hp-FEM*, Comput. Methods Appl. Mech. Engrg. **196** (2007), no. 37-40, 3713–3725.

61. I. Ekeland and R. Témam, *Convex analysis and variational problems*, Classics in Applied Mathematics, vol. 28, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

62. K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta numerica, Cambridge University Press, Cambridge, 1995, pp. 105–158.

63. L.C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 2010.

64. R.S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl. **44** (1973), 28–47.

65. U. Felgenhauer, *On stability of bang-bang type controls*, SIAM J. Control Optim. **41** (2003), no. 6, 1843–1867.

66. A. Gaevskaya, R.H.W. Hoppe, Y. Iliash, and M. Kieweg, *Convergence analysis of an adaptive finite element method for distributed control problems with control constraints*, Control of coupled partial differential equations, Internat. Ser. Numer. Math., vol. 155, Birkhäuser, Basel, 2007, pp. 47–68.

67. T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér. **13** (1979), no. 4, 313–328.

68. I. Gohberg, S. Goldberg, and M.A. Kasshoek, *Classes of linear operators*, vol. 1, Birkhäuser, Basel, 1990.

69. V. Gol'dshtein and A. Ukhlov, *Weighted Sobolev spaces and embedding theorems*, Trans. Amer. Math. Soc. **361** (2009), no. 7, 3829–3850.

70. W. Gong, W. Liu, and N. Yan, *A posteriori error estimates of hp-FEM for optimal control problems*, Int. J. Numer. Anal. Model. **8** (2011), no. 1, 48–69.

71. P. Grisvard, *Singularities in boundary value problems*, Research Notes in Applied Mathematics, vol. 22, Masson, Paris, 1992.

72. P. Grisvard, *Elliptic problems in nonsmooth domains*, Classics in Applied Mathematics, vol. 69, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

73. B.Q. Guo, *The h-p version of the finite element method for solving boundary value problems in polyhedral domains*, Boundary value problems and integral equations in nonsmooth domains, Lecture Notes in Pure and Applied Mathematics, vol. 167, Dekker, New York, 1995, pp. 101–120.

74. B.Q. Guo and I. Babuška, *Regularity of the solutions for elliptic problems on nonsmooth domains in $\mathbf{R}^3$. I,II.*, Proc. Roy. Soc. Edinburgh Sect. A **127** (1997), no. 3, 517–545.

75. B.Q. Guo and H.S. Oh, *The h-p version of the finite element method for problems with interfaces*, Internat. J. Numer. Methods Engrg. **37** (1994), no. 10, 1741–1762.

76. W. Hackbusch, *Integral equations - Theory and numerical treatment*, International Series of Numerical Mathematics, vol. 120, Birkhäuser, Basel, 1995.

77. M. Hintermüller and M. Hinze, *Moreau-Yosida regularization in state constrained elliptic control problems: Error estimates and parameter adjustment*, SIAM J. Numer. Anal. **47** (2009), no. 3, 1666–1683.

78. M. Hintermüller, R.H.W. Hoppe, Y. Iliash, and M. Kieweg, *An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints*, ESAIM Control Optim. Calc. Var. **14** (2008), no. 3, 540–560.

79. M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl. **30** (2005), no. 1, 45–61.

80. M. Hinze and U. Matthes, *A note on variational discretization of elliptic Neumann boundary control*, Control Cybernet. **38** (2009), no. 3, 577–591.

81. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, Mathematical Modelling: Theory and Applications, vol. 23, Springer, New York, 2009.

82. M. Hinze and A. Schiela, *Discretization of interior point methods for state constrained elliptic optimal control problems: Optimal error estimates and parameter adjustment*, Comput. Optim. Appl. **48** (2011), no. 3, 581–600.

83. R.H.W. Hoppe, Y. Iliash, C. Iyyunni, and N. H. Sweilam, *A posteriori error estimates for adaptive finite element discretizations of boundary control problems*, J. Numer. Math. **14** (2006), no. 1, 57–82.

84. K. Ito and K. Kunisch, *Lagrange multiplier approach to variational problems and applications*, Advances in Design and Control, vol. 15, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

85. V. Jagota, A. Sethi, and K. Kumar, *Finite element method: An overview*, Walailak J. Sci. & Tech. **10** (2013), no. 1, 1–8.

86. D.S. Jerison and C.E. Kenig, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc. (N.S.) **4** (1981), no. 2, 203–207.

87. G.E. Karniadakis and S.J. Sherwin, *Spectral/hp element methods for computational fluid dynamics*, 2nd ed., Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2005.

88. B.N. Khoromskij and J.M. Melenk, *Boundary concentrated finite element methods*, SIAM J. Numer. Anal. **41** (2003), 1–36.

89. D. Knees, *Regularitätsaussagen für zweidimensionale elastische Felder in Kompositen*, Diplomarbeit, Universität Stuttgart, 2001.

90. K. Kohls, A. Rösch, and K.G. Siebert, *A posteriori error estimators for control constrained optimal control problems*, Constrained Optimization and Optimal Control for Partial Differential Equations, Internat. Ser. Numer. Math., vol. 160, Birkhäuser, Basel, 2012, pp. 431–443.

91. K. Kohls, A. Rösch, and K.G. Siebert, *A posteriori error analysis of optimal control problems with control constraints*, SIAM J. Control Optim. **52** (2014), no. 3, 1832–1861.

92. V.A. Kondratjev, *Boundary value problems for elliptic equations in domains with conical or angular points*, Trudy Moskovkogo Mat. Obschetsva **16** (1967), 209–292.

93. V.G. Korneev and S. Jensen, *Preconditioning of the p-version of the finite element method*, Comput. Methods Appl. Mech. Engrg. **150** (1997), no. 1-4, 215–238.

94. L. Korous and P. Šolín, *An adaptive hp-dg method with dynamically-changing meshes for non-stationary compressible euler equations*, Computing **95** (2013), no. 1, 425–444.

95. A. Kufner, *Weighted Sobolev spaces*, Teubner-Texte zur Mathematik, vol. 31, BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1980.

96. A. Kufner and A.-M. Sändig, *Some applications of weighted Sobolev spaces*, Teubner-Texte zur Mathematik, vol. 100, BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1987.

97. K. Kunisch and B. Vexler, *Constrained Dirichlet boundary control in L^2 for a class of evolution equations*, SIAM J. Control Optim. **46** (2007), no. 5, 1726–1753.

98. G. Leugering et al. (ed.), *Constrained optimization and optimal control of partial differential equations*, International Series of Numerical Mathematics, vol. 160, Birkhäuser, Basel, 2012.

99. G. Leugering et al. (ed.), *Trends in PDE constrained optimization*, International Series of Numerical Mathematics, vol. 165, Birkhäuser, Basel, 2014.

100. R. Li, W. Liu, H. Ma, and T. Tang, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim. **41** (2002), no. 5, 1321–1349.

101. J.L. Lions, *Optimal control of systems governed by partial differential equations*, Die Grundlehren der mathematischen Wissenschaften, vol. 170, Springer, Berlin, 1971.

102. W. Liu and N. Yan, *A posteriori error estimates for convex boundary control problems*, SIAM J. Numer. Anal. **39** (2001), no. 1, 73–99.

103. W. Liu and N. Yan, *A posteriori error estimates for distributed convex optimal control problems*, Adv. Comput. Math. **15** (2001), no. 1-4, 285–309 (2002).

104. M. Mateos and A. Rösch, *On saturation effects in the Neumann boundary control of elliptic optimal control problems*, Comput. Optim. Appl. **49** (2011), no. 2, 359–378.

105. S. May, R. Rannacher, and B. Vexler, *Error analysis for a finite element approximation of elliptic Dirichlet boundary control problems*, SIAM J. Control Optim. **51** (2013), no. 3, 2585–2611.

106. J.M. Melenk, *Hp-finite element methods for singular perturbations*, Lecture Notes in Mathematics, vol. 1796, Springer, Berlin Heidelberg, 2002.

107. J.M. Melenk, *Hp-interpolation of nonsmooth functions and an application to hp-a posteriori error estimation*, SIAM J. Numer. Anal. **43** (2005), no. 1, 127–155.

108. J.M. Melenk and B. Wohlmuth, *On residual-based a posteriori error estimation in hp-FEM*, Adv. Comput. Math. **15** (2001), no. 1-4, 311–331.

109. J.M. Melenk and B. Wohlmuth, *Quasi-optimal approximation of surface based Lagrange multipliers in finite element methods*, SIAM J. Numer. Anal. **50** (2012), no. 4, 2064–2087.

110. C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim. **43** (2004), no. 3, 970–985.

111. V.A. Morozov, *Regularization methods for ill-posed problems*, CRC Press, Boca Raton, FL, 1993.

112. C.B. Morrey, Jr., *Multiple integrals in the calculus of variations*, Die Grundlehren der mathematischen Wissenschaften, vol. 130, Springer, Boston, 1966.

113. C.B. Morrey, Jr. and L. Nirenberg, *On the analyticity of the solutions of linear elliptic systems of partial differential equations*, Comm. Pure Appl. Math. **10** (1957), 271–290.

114. J. Nečas, *Direct Methods in the Theory of Elliptic Equations*, Springer Monographs in Mathematics, Springer, Berlin Heidelberg, 2012.

115. E.D. Nezza, G. Palatucci, and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bull. Sci. Math. **136** (2012), no. 5, 521–573.

116. S. Nicaise, *Le Laplacien sur les réseaux deux-dimensionnels polygonaux topologiques*, J. Math. Pures Appl. (9) **67** (1988), no. 2, 93–113.

117. S. Nicaise, *Polygonal interface problems*, Methoden und Verfahren der Mathematischen Physik, vol. 39, Peter Lang, Frankfurt am Main, 1993.

118. S. Nicaise and A.-M. Sändig, *General interface problems - I, II*, Math. Methods Appl. Sci. **17** (1994), no. 6, 395–429, 431–450.

119. J.T. Oden, L. Demkowicz, W. Rachowicz, and T.A. Westermann, *Toward a universal h-p adaptive finite element strategy. II. A posteriori error estimation*, Comput. Methods Appl. Mech. Engrg. **77** (1989), no. 1-2, 113–180.

120. M. Petzold, *Regularity and error estimators for elliptic problems with discontinuous coefficients*, Dissertation, Freie Universität Berlin, 2001.

121. V.C. Piat and F.S. Cassano, *Some remarks about the density of smooth functions in weighted sobolev spaces*, J. Conv. Anal. **1** (1994), no. 2, 135–142.

122. U. Prüfert, F. Tröltzsch, and M. Weiser, *The convergence of an interior point method for an elliptic control problem with mixed control-state constraints*, Comput. Optim. Appl. **39** (2008), no. 2, 183–218.

123. W. Rachowicz, J.T. Oden, and L. Demkowicz, *Toward a universal h-p adaptive finite element strategy. III. Design of h-p meshes*, Comput. Methods Appl. Mech. Engrg. **77** (1989), no. 1-2, 181–212.

124. R. Rannacher, B. Vexler, and W. Wollner, *A posteriori error estimation in PDE-constrained optimization with pointwise inequality constraints*, Constrained Optimization and Optimal Control for Partial Differential Equations, Internat. Ser. Numer. Math., vol. 160, Birkhäuser, Basel, 2012, pp. 349–373.

125. A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw. **21** (2006), no. 1, 121–134.

126. A. Rösch and R. Simon, *Linear and discontinuous approximations for optimal control problems*, Numer. Funct. Anal. Optim. **26** (2005), no. 3, 427–448.

127. A. Rösch and R. Simon, *Superconvergence properties for optimal control problems discretized by piecewise linear and discontinuous functions*, Numer. Funct. Anal. Optim. **28** (2007), no. 3-4, 425–443.

128. A. Rösch and D. Wachsmuth, *A-posteriori error estimates for optimal control problems with state and control constraints*, Numer. Math. **120** (2012), no. 4, 733–762.

129. S.A. Sauter and C. Schwab, *Boundary element methods*, Springer Series in Computational Mathematics, vol. 39, Springer, Berlin, 2011.

130. A. Schiela, *The control reduced interior point method - a function space oriented algorithmic approach*, Dissertation, Freie Universität Berlin, 2006.

131. A. Schiela, *An interior point method in function space for the efficient solution of state constrained optimal control problems*, Math. Program. **138** (2013), no. 1-2, Ser. A, 83–114.

132. A. Schiela and A. Günther, *Interior point methods in function space for state constraints - inexact Newton and adaptivity*, ZIB-Report, Konrad Zuse-Zentrum für Informationstechnik, Berlin, 01 2009.

133. A. Schiela and A. Günther, *An interior point algorithm with inexact step computation in function space for state constrained optimal control*, Numer. Math. **119** (2011), no. 2, 373–407.

134. A. Schiela and M. Weiser, *Superlinear convergence of the control reduced interior point method for PDE constrained optimization*, Comput. Optim. Appl. **39** (2008), no. 3, 369–393.

135. C. Schwab, *p- and hp-finite element methods - Theory and applications in solid and fluid mechanics*, Numerical Mathematics and Scientific Computation, Clarendon Press, Oxford, 1998.

136. C. Schwab, *Exponential convergence of simplicial hp-FEM for H^1 functions with isotropic singularities*, SAM Research Report 15, ETH Zürich, 2014.

137. D. Sevilla and D. Wachsmuth, *Polynomial integration on regions defined by a triangle and a conic*, ISSAC 2010 - Proceedings of the 2010 International Symposium on Symbolic and Algebraic Computation, ACM, New York, 2010, pp. 163–170.

138. P. Šolín, J. Cerveny, L. Dubcova, and D. Andrs, *Monolithic discretization of linear thermoelasticity problems via adaptive multimesh hp-FEM*, J. Comput. Appl. Math. **234** (2010), no. 7, 2350–2357.

139. P. Šolín and L. Demkowicz, *Goal-oriented hp-adaptivity for elliptic problems*, Comput. Methods Appl. Mech. Engrg. **193** (2004), no. 6-8, 449–468.

140. P. Šolín, L. Dubcova, and J. Kruis, *Adaptive hp-FEM with dynamical meshes for transient heat and moisture transfer problems*, J. Comput. Appl. Math. **233** (2010), no. 12, 3103–3112.

141. P. Šolín and L. Korous, *Adaptive higher-order finite element methods for transient PDE problems based on embedded higher-order implicit Runge-Kutta methods*, J. Comput. Phys. **231** (2012), no. 4, 1635 – 1649.

142. P. Šolín and L. Korous, *Space-time adaptive hp-FEM for problems with traveling sharp fronts*, Computing **95** (2013), no. 1, suppl., S709–S722.

143. G. Szegö, *Orthogonal polynomials*, American Mathematical Society, New York, 1959.

144. F. Tröltzsch, *Optimal control of partial differential equations - Theory, methods and applications*, Graduate Studies in Mathematics, vol. 112, American Mathematical Society, Providence, RI, 2010.

145. M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim. **13** (2003), no. 3, 805–842.

146. M. Ulbrich, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, MOS-SIAM Series on Optimization, vol. 11, Mathematical Optimization Society (MOS)/Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

147. M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim. **38** (2000), no. 6, 1938–1984.

148. M. Ulbrich and S. Ulbrich, *Primal-dual interior-point methods for PDE-constrained optimization*, Math. Program. **117** (2009), no. 1-2, Ser. B, 435–485.

149. R. Verfürth, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Advances in Numerical Mathematics, Wiley/Teubner, Chichester/Stuttgart, 1996.

150. B. Vexler, *Finite element approximation of elliptic Dirichlet optimal control problems*, Numer. Funct. Anal. Optim. **28** (2007), no. 7-8, 957–973.

151. B. Vexler and W. Wollner, *Adaptive finite elements for elliptic optimization problems with control constraints*, SIAM J. Control Optim. **47** (2008), no. 1, 509–534.

152. P. Šolín, K. Segeth, and I. Doležel, *Higher-order finite element methods*, Studies in Advanced Mathematics, Chapman & Hall/CRC, Boca Raton, FL, 2004.

153. D. Wachsmuth, *Adaptive regularization and discretization of bang-bang optimal control problems*, Electron. Trans. Numer. Anal. **40** (2013), 249–267.

154. D. Wachsmuth, *Robust error estimates for regularization and discretization of bang-bang control problems*, Comp. Optim. Appl. (2014), 19pp.

155. D. Wachsmuth and G. Wachsmuth, *Regularization error estimates and discrepancy principle for optimal control problems with inequality constraints*, Control Cybernet. **40** (2011), no. 4, 1125–1158.

156. D. Wachsmuth and J.-E. Wurst, *Exponential convergence of hp-finite element discretization of optimal boundary control problems with elliptic partial differential equations*, Preprint 328, Julius-Maximilians-Universität Würzburg, 2015, submitted to SIAM J. Control Optim.

157. D. Wachsmuth and J.-E. Wurst, *An interior point method designed for solving linear quadratic optimal control problems with hp finite elements*, Optim. Methods Softw. **30** (2015), no. 6, 1276–1302.

158. D. Wachsmuth and J.-E. Wurst, *Optimal control of interface problems with hp-finite elements*, Numer. Func. Anal. Optim. (2015), to appear.

159. G. Wachsmuth and D. Wachsmuth, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM Control Optim. Calc. Var. **17** (2011), no. 3, 858–886.

160. M. Weiser, T. Gänzler, and A. Schiela, *A control reduced primal interior point method for a class of control constrained optimal control problems*, Comput. Optim. Appl. **41** (2008), no. 1, 127–145.

161. R. Winkler, *Schwache Randwertprobleme von Systemen elliptischen Charakters auf konischen Gebieten*, Dissertation, Julius-Maximilians-Universität Würzburg, 2008.

162. J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev. **40** (1998), no. 4, 857–914.

163. K. Yosida, *Functional analysis*, Classics in Mathematics, Springer, Berlin, 1980.

164. W.P. Ziemer, *Weakly differentiable functions - Sobolev spaces and functions of bounded variation*, Graduate Texts in Mathematics, vol. 120, Springer, New York, 1989.

Optimal control theory is a versatile mathematical discipline with applications in many fields. It has gained interest over the last decades mainly because increasing computational power allowed to tackle large and complex real life problems numerically. For offering reliable results, a thorough theoretical analysis of solution algorithms, their convergence properties, and approximation quality is inevitable.

We follow this need and investigate linear quadratic optimal control problems with elliptic partial differential equations. The discretization with *hp*-finite elements is embedded in both Newton-type and interior point methods. Different efficient strategies are presented and accompanied by new results on regularity, approximation, and convergence theory.

ISBN 978-3-95826-024-5

9 783958 260245

Würzburg University Press