Structure exploiting Galerkin schemes for optimal control of PDEs with constraints on the involved variables

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Universität Hamburg

vorgelegt
im Fachbereich Mathematik

von
Andreas Günther

aus Bautzen

Hamburg
2010

Als Dissertation angenommen vom Fachbereich
Mathematik der Universität Hamburg

Auf Grund der Gutachten von Prof. Dr. Michael Hinze
und Prof. Dr. Arnd Rösch

Hamburg, den 30.11.2010

Prof. Dr. Vicente Cortés
Leiter des Fachbereiches Mathematik

Abstract.
This thesis is about structure exploiting Galerkin schemes for optimal control problems governed by elliptic partial differential equations under constraints onto the control, the state and its derivative.

Those tailored Galerkin concepts enter on the a priori part by the permanent application of the variational discretization concept proposed by Hinze in [Hin05]. This minimal invasive finite element discretization technique allows an elegant and funded a priori error analysis. We prove several a priori error estimates for the above mentioned optimal control problems. For control constrained Dirichlet boundary control we even improve these results by superconvergence effects caused by additional assumptions onto the underlying mesh of computation. All estimates are verified by numerous numerical examples and experimental order of convergence measurements.

Moreover on the a posteriori part the concept of variational discretization avoids the appearance of additional control error terms in error representations. We exploit the structure of the underlying optimal control problems by designing goal-oriented error estimators for control- and state-constrained problems. This builds up an extension of the DWR-method proposed by Becker and Rannacher in [BR96] for unconstrained optimization with PDEs. By only usage of the numerical solution we derive computable error estimators in order to efficiently resolve the optimal objective value. In a few numerical experiments we find appropriate adaptive meshes, which by model reduction help to substantially save degrees of freedom and hence CPU-time. We further study the efficiency indices of the derived estimators.

Contents

Danksagung

Zuallererst möchte ich meinem Betreuer Prof. Dr. Michael Hinze meinen tiefsten Dank für die vielseitige Unterstützung, Forderung und Förderung aussprechen. Ohne die besonderen Möglichkeiten durch meine DFG-Projekt-Stelle in Hamburg wäre diese Arbeit so nicht entstanden. Ich bedanke mich für die gemeinsame konstruktive Zeit des Forschens, für viele wertvolle Hinweise und für das entgegengebrachte Vertrauen, Resultate auch auf internationalen Konferenzen präsentieren zu dürfen.

Mein Dank gilt ebenso Prof. Dr. Klaus Deckelnick. Als wertvoller Unterstützer und Begleiter meiner Forschungstätigkeit hat er in vielen Teilen zum Gelingen dieser Arbeit beigetragen.

Ferner möchte ich Dr. Anton Schiela und Dr. Moulay Hicham Tber für die angenehme Zusammenarbeit und die Einladungen nach Berlin und Graz danken, wodurch sich für mich neue Herangehensweisen ergeben haben.

Ich danke Martin Kunkel für seine Ausdauer bei der Klärung meiner Linux- und Matlab-Fragen.

Besondere Erwähnung gebührt an dieser Stelle auch meinen Eltern für ihre umfassende moralische Unterstützung und das ständige Zutrauen in meine Arbeit. Gleiches gilt für meine Freundin Anja, der ich außerdem herzlich für den festen Zusammenhalt danken möchte.

Nomenclature

We use similar notations as in [EG04].

Basic notation

$\mathrm{card}(E)$	Cardinal number of the set E
$u_{\|E}$	Restriction of the function u to the set E
$\mathbb{1}_E$	Characteristic function on the set E
$\mathrm{span}\{\vec{v}_1, \ldots, \vec{v}_n\}$	Vector space spanned by the vectors $\vec{v}_1, \ldots, \vec{v}_n$
δ_{ij}	KRONECKER symbol: $\delta_{ij} = 1$ if $i = j$ and 0 otherwise

Vectors and matrices

$[u_1, \ldots, u_n]^T$	Cartesian components of the vector $\mathbf{u} \in \mathbb{R}^n$
$\mathbf{u} \cdot \mathbf{v}$	Euclidean scalar product in $\mathbb{R}^n$: $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \sum_{i=1}^n u_i v_i$
$\|\mathbf{u}\|$	Euclidean norm in $\mathbb{R}^n$: $\|\mathbf{u}\| = (\mathbf{u}^T\mathbf{u})^{\frac{1}{2}}$
$\mathbb{R}^{m\times n}$	Vector space of $m \times n$ matrices with $\mathbb{R}$-valued entries
$\mathbf{M}, \mathbf{A}$	Matrices
$\mathbf{I}$	Identity matrix
$\mathbf{0}$	Zero matrix
a_{ij}	Entry of $\mathbf{A}$ in the ith row and the jth column
$\mathbf{A}^T$	Transpose of the matrix $\mathbf{A}$
$\mathbf{diag}(\mathbf{u})$	Diagonal matrix with diagonal $\mathbf{u}$: $\mathbf{diag}(\mathbf{u}) = [\delta_{ij}u_i]_{i,j=1}^n$
$\mathbf{Au}$	Matrix-vector product : For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{Au} = [\sum_{j=1}^n a_{ij}u_j]_{i=1}^m$
$\mathbf{u} \otimes \mathbf{v}$	Tensor product for $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^m$: $\mathbf{u} \otimes \mathbf{v} = \mathbf{uv}^T$

Differential operators

$(x_1, \ldots, x_d)$	Cartesian coordinates in $\mathbb{R}^d$
$\partial_{x_i} y = y_{x_i}$	Distributional derivative of y with respect to x_i
$\partial_{x_i x_j} y = y_{x_i x_j}$	Distributional derivative of y with respect to x_i and x_j
$\partial^{\boldsymbol{\alpha}} y$	$\partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d} y$ where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^T \in \mathbb{N}_0^d$ is a multi-index
$\|\boldsymbol{\alpha}\|$	Length of $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^T \in \mathbb{N}_0^d$: $\|\boldsymbol{\alpha}\| = \alpha_1 + \cdots + \alpha_d$
∇y	Gradient: $\nabla y = [\partial_{x_i} y]_{i=1}^d \in \mathbb{R}^d$ if y is $\mathbb{R}$-valued
$\nabla \cdot \vec{v} = \mathrm{div}\vec{v}$	Divergence: $\nabla \cdot \vec{v} = \sum_{i=1}^d \partial_{x_i} v_i$ if $\vec{v}$ is $\mathbb{R}^d$-valued
Δy	LAPLACE operator: $\Delta y = \sum_{i=1}^d \partial_{x_i x_i} y$ if y is $\mathbb{R}$-valued
$D^2 y$	Hessian operator: $D^2 y = [\partial_{x_i x_j} y]_{i,j=1}^d$ if y is $\mathbb{R}$-valued

Function spaces

$\mathcal{L}(E;F)$	Vector space of the bounded linear operators from E to F
X'	Topological dual of the topological space X
$\mathcal{A}^*$	Dual operator of $\mathcal{A}$: if $\mathcal{A} \in \mathcal{L}(E;F), \mathcal{A}^* \in \mathcal{L}(F';E')$
$\|u\|_X$	Norm of u in the normed space X
$\mathbb{P}^k$	Vector space of polynomials in the variables $x_1, \ldots, x_d$ of global degree at most k
$\mathcal{D}(\Omega)$	Infinitely differentiable functions compactly supported in Ω
$C^0(\Omega), C^k(\Omega)$	Space of continuous functions on $\Omega \subset \mathbb{R}^d$, and space of k times continuously differentiable functions on Ω
$C^{k,\alpha}(\Omega)$ (resp., $C^{k,\alpha}(\bar{\Omega})$)	Space of functions whose derivatives up to order k are locally (resp., globally) α-HÖLDER continuous
δ_x	DIRAC measure at x
$L^p(\Omega)$	Functions whose p-th order is LEBESGUE integrable on Ω
p'	Conjugate of p, $\frac{1}{p} + \frac{1}{p'} = 1$
$W^{s,p}(\Omega)$	Functions whose derivatives up to order s are in $L^p(\Omega)$
$W_0^{s,p}(\Omega)$	Closure of $\mathcal{D}(\Omega)$ in $W^{s,p}(\Omega)$
$W^{-s,p'}(\Omega)$	Dual of $W_0^{s,p}(\Omega)$
$\|u\|_{L^p(\Omega)}$	Norm in $L^p(\Omega)$: $\|u\|_{L^p(\Omega)} = (\int_\Omega \|u\|^p)^{\frac{1}{p}}$
$\|u\|_{W^{s,p}(\Omega)}$	Seminorm in $W^{s,p}(\Omega)$: $\|u\|_{W^{s,p}(\Omega)} = \sum_{\alpha=s} \|\partial^\alpha u\|_{L^p(\Omega)}$
$\|u\|_{W^{s,p}(\Omega)}$	Norm in $W^{s,p}(\Omega)$: $\|u\|_{W^{s,p}(\Omega)} = \sum_{l \le s} \|u\|_{W^{l,p}(\Omega)}$
$H^s(\Omega), H_0^s(\Omega)$	$W^{s,2}(\Omega)$, $W_0^{s,2}(\Omega)$
$\|u\|_{H^s(\Omega)}$, $\|u\|_{H^s(\Omega)}$	$\|u\|_{W^{s,2}(\Omega)}$, $\|u\|_{W^{s,2}(\Omega)}$
(u,v)	Scalar product on $L^2(\Omega)$: $\int_\Omega uv$
$H(\mathrm{div};\Omega)$	$\{\vec{v} \in [L^2(\Omega)]^d : \nabla \cdot \vec{v} \in L^2(\Omega)\}$

Mesh-related symbols

$h_T = \mathrm{diam}(T)$	Diameter of $T \subset \mathbb{R}^d$
m	Number of geometrical nodes
nt	Number of cells (or elements) in the mesh

Finite element spaces

$P^k_{c,h}(\mathcal{T}_h)$	Vector space of functions that are piecewise in $\mathbb{P}^k$ and are continuous
$P^k_{td,h}(\mathcal{T}_h)$	Vector space of (totally discontinuous) functions that are piecewise in $\mathbb{P}^k$
$RT_{0,h}(\mathcal{T}_h)$	Space of lowest order Raviart–Thomas functions

Active and inactive (index) subsets

$\bullet$	All: either $\bullet = \{1, \ldots, m\}$ or $\bullet = \Omega$
$\oplus, \circledcirc, \oplus, \circledcirc$	Control constraints: inactive, lower active, upper active, active
$\ominus, \circledcirc, \ominus, \circledcirc$	State constraints: inactive, lower active, upper active, active

Introduction

This manuscript is about tailored Galerkin discretization strategies for optimization problems governed by *partial differential equations* (PDEs) under additional constraints onto the involved quantities. Generally speaking we deal with the problem

$$\begin{aligned} & f(\mathbf{x}) \to \min \\ \text{such that } & c(\mathbf{x}) = 0, \\ & g(\mathbf{x}) \leq 0, \end{aligned} \tag{0.1}$$

where f is the objective to be minimized under the condition, that a PDE modeled by $c(\mathbf{x}) = 0$ and additionally further nonlinear constraints $g(\mathbf{x}) \leq 0$ have to be satisfied. In order to efficiently solve these kind of problems one has to essentially exploit the structure of the involved equations. This is done on the one hand by a funded *a priori* analysis of the underlying optimization problem but on the other hand also by *a posteriori* error estimation techniques to the point of implementation issues.

In the last decades the subject of PDE-constrained optimization with all its surrounding topics became a key technology. This was not only by the increasing *high performance computing* (HPC) resources but also due to the tremendously raised importance of applications in engineering sciences such as mechanical and medical engineering, aerospace industry or materials science.

To meet the challenges nowadays of optimizing more and more complex models the techniques such as *multigrid* (MG), *automatic differentiation* (AD), *parallel computing*, *preconditioning* and *adaptivity* from the numerical point of view with appropriate discretization strategies making essentially use of the structure of the underlying system have to be united.

Down to this state of the present day is a long path in history, which comes into its own by the occurrence of famous mathematician's and physicist's names in a lot of widely spread notions. Within the next pages these traces become visible in the field of PDEs, optimization and of their interplay.

0.1. Partial differential equations

The notion of a *partial differential equation* is defined in [Jos07] as follows:

Definition 0.1.1. A partial differential equation is an equation involving derivatives of an unknown function $y : \Omega \to \mathbb{R}$, where Ω is an open subset of $\mathbb{R}^d$, $d \geq 2$.

PDEs can be classified due to their *order* of the highest-occurring derivative. For instance a PDE of second-order can be represented by

$$c(x, y, \partial_{x_i} y, \partial_{x_i x_j} y) = 0,$$

or to be consistent to problem (0.1) shortly by $c(\mathbf{x}) = 0$. Notice that this equation is usually formulated in an infinite dimensional function space. Now depending on

the properties of the function c we distinguish between linear and nonlinear such as semilinear or quasilinear PDEs. Moreover for developing a unified theory the literature distinguishes between *elliptic*, *parabolic* and *hyperbolic* PDEs. For each of them the POISSON-equation, the heat-equation and wave-equation are representatives. But even more PDEs are named after famous mathematicians and physicists such as CAUCHY, EULER, HELMHOLTZ, LAPLACE, MAXWELL, NAVIER, RIEMANN, SCHRÖDINGER and STOKES.

In order to guarantee the existence of a unique solution of the underlying PDE in the *domain* Ω one further has to state conditions at the *boundary* of the domain $\Gamma := \partial\Omega$. We distinguish between DIRICHLET-boundary conditions, where one fixes the unknown function y like a spanned membrane, NEUMANN-boundary conditions, where first derivatives of the solution are prescribed and ROBIN-boundary conditions, a mixture of both. Additionally for time dependent PDEs like the heat-equation a further initial condition has to hold.

The books of [Fri69], [LM72], [Jos07] and [Eva98] give a deep insight into the theory and a priori analysis of PDEs. Especially the books of [GT01] and [Hac10] concern elliptic PDEs, where also this manuscript focusses on. For basic assertions they all combine the well known arguments like the maximum principle together with established notions like BANACH and HILBERT spaces to derive extremely useful results as the existence of *weak* solutions by the help of the famous RIESZ representation theorem and the Lax-MILGRAM lemma. The regularity of such weak solutions is even improved due to the accomplishments in the field of functional analysis as the notion of SOBOLEV spaces and embedding theorems as they can be found in [Alt06] for example.

The huge variety of different PDEs is of course accompanied by a wide range of numerical solution concepts and computer software. All concepts *discretize* the underlying domain and PDE in a certain way in order to obtain a finite dimensional problem. The accuracy of the approximate solution is controlled by a discretization parameter tending to zero leading to increasing system sizes to be solved. This fact is common for the most important numerical solution concepts namely the *Finite Difference Method* (FDM), the *Finite Volume Method* (FVM) and the *Finite Element Method* (FEM). We are going to make us of the latter one.

The FEM came up at the late 1950s. Here the discretization of the domain is realized by a *partitioning* of it into so called cells in forms of triangles or quadrilaterals in 2D or tetrahedra in 3D for instance. Out from this partitioning having the mesh size h one defines a finite number of Ansatz-functions ϕ_h^p with local support. Usually these functions were constructed by the help of polynomials of maximal degree p or by splines. Now the approximate solution of the PDE as a linear combination of these Ansatz-functions is found by testing the variational formulation of the PDE with all ϕ_h^p giving us the so called GALERKIN scheme. Depending on which discretization parameter tends to a limit, we distinguish the FEMs between h-, p- and h-p-methods. Classical literature concerning the topic of the FEM are the books of [SF73] and [Cia80]. More recently the excellent monographs [Bra07], [BS08], [GR05] and [EG04] include the latest research in this field.

Let us now come to a basic pillar of this manuscript concerning the efficient solution of PDEs. As already announced, the accuracy of the approximate solution of such a GALERKIN system is controlled by a discretization parameter which directly scales the system size in terms of *degrees of freedom* (DOFs). This is just the half truth. Since there are two choices for the overall aim namely:

- minimize the CPU-time to obtain the solution within given accuracy, or
- maximize the accuracy of the solution within given CPU-time,

so called *model reduction* techniques can be applied. A certainly well established technique is *Proper Orthogonal Decomposition* (POD). It is widely used in the field of *Computational Fluid Dynamics* (CFD) for instance. Its analysis extracts out from *snapshots* a small set of eigenfunctions which describe the dominant behavior of a dynamical system. The approximate solution is then represented by just a few DOFs entering the linear combination of those eigenfunctions and is obtained out from a small but usually dense system. If one often has to simulate the dynamical system or for reasons of optimization the computation of those eigenfunctions pays off and hence this is one step towards the above mentioned overall aim.

We follow another *model reduction* technique, namely the concept of *adaptivity*. At least for the finite element h-method, which we are going to apply throughout this manuscript, roughly speaking a DOF can be localized in the computational domain and represents local information onto the solution at this certain area. Depending on what "accuracy of the solution" means, it may be reasonable to accumulate DOFs in a certain area of interest while the loss of DOFs at other parts of the domain has negligible impact onto the accuracy. Hence for a given CPU-time and therefore indirectly given amount of DOFs one may specifically place those in the domain in order to maximize the accuracy of the solution. The adaptive iteration consists of four parts:

$$\text{Solve} \rightarrow \text{Estimate} \rightarrow \text{Mark} \rightarrow \text{Re-mesh}.$$

A very good overview about this cascade is given in the books of [Ver96] and [BR03]. Especially the second part concerning the design of an estimator to ones demands is introduced in the book of [AO00]. Due to a rich variety of different settings, concerning the type of PDE under consideration, the possible presence of additional inequality constraints, and the quantities of interest, there are many techniques to be explored. We distinguish between residual type and goal oriented error estimators. An investigation of convergence of the *Adaptive Finite Element Method* (AFEM) has just started about 15 years ago. Basic results for its development are presented in the work of [BR78] or [BW85]. A first rigorous analysis of convergence of the AFEM is arranged in [Dör96]. A generalized proof taking data oscillations into account can be found in [MNS00]. A few years later [BDD04] showed convergence with optimal rates even though intermediate mesh coarsening was still necessary. Nowadays the frontiers are pushed further towards more complicated problems of consideration. For instance a convergent AFEM for optimal design problems is lately presented in [BC08]. Recently even a new approach in the convergence analysis of AFEM for control constrained optimal control problems is proposed in [KRS10]. However the convergence of the goal-oriented AFEM is totally open yet.

Even though a lot of questions are still unexplained, the AFEM has already proved to be a very successful concept. As a key technology in order to achieve the above mentioned overall aim a huge spectrum of available PDE solvers have been developed. Those classify into open source projects and commercial code but of course predominantly after their application, programming language and provided interfaces. As a representative solver for the FVM we arbitrarily mention FiPy[1]. But since we concentrate onto the FEM let us list for instance ALBERTA[2],

[1]FiPy: A Finite Volume PDE Solver Using Python, http://www.ctcms.nist.gov/fipy/
[2]ALBERTA: An adaptive hierarchical finite element toolbox, http://www.alberta-fem.de

Gascoigne3D[3], OpenFEM[4], Getfem++[5], deal.II[6], FEAST[7], and Kaskade 7[8]. They further distinguish between their implemented different macro-elements in different space dimensions concerning their shape and order and if parallelization techniques are used. On the side of commercial software let us mention in the first instance the Matlab Partial Differential Equation Toolbox[9] and COMSOL Multiphysics[10]. The former one is mainly used throughout this manuscript.

0.2. Optimization

In order to introductorily approach our topic of consideration let us investigate problem (0.1) under the aspect of optimization. Problem (0.1) captures already the three main ingredients in *mathematical programming*:

- the *unknown variable* $\mathbf{x} \in X$ to be optimized,
- the *objective function* $f : X \to \mathbb{R}$, and
- (in-)equality *constraints* $c : X \to Z_1$ ($g : X \to Z_2$).

Depending on the properties of X one primarily distinguishes between *finite* and *infinite* dimensional optimization problems. In this work both situations have their place. Although in PDE-constrained optimization the arising problems are originally formulated in some infinite dimensional function space X, say $L^2(\Omega)$, the approximate problems after applying the above mentioned discretization techniques could possibly be stated in the finite dimensional space $X = \mathbb{R}^n$. Both cases fall into the class of *continuous* optimization problems. Moreover due to the permanent presence of some underlying PDE to be satisfied our topic further falls into the field of *constrained optimization*. Both constraint functions c and g mark the so called *feasible region*. Apart from *linear programming* (LP), where all appearing quantities f, c, and g are linear functions we focus on *nonlinear programming* (NLP). Nevertheless we are in the field of *convex optimization* whenever f and g are convex and c is linear.

Generally in NLP a *global* optimal solution can hardly be found. However under certain smoothness assumptions of the involved functions *local* solutions can be characterized. With the help of the Lagrange multiplier method a local solution satisfies the *first order optimality conditions* or Karush-Kuhn-Tucker (KKT) conditions. It is clear that an efficient solution algorithm should essentially use these equations for fast convergence.

Let us briefly give an overview about the related literature. A comprehensive work is the book of [GK99] about unconstrained optimization. This topic is still of interest since constrained optimization problems can be relaxed to unconstrained ones as is explained in the continuation [GK02]. Both issues are also well addressed in the book of [NW06]. Moreover for finite convex optimization let us mention [BV04]. For infinite dimensional problems we highlight the impressive monograph [Fat99], which also already approaches the topic of PDE-constrained optimization.

[3]Gascoigne3D: High Performance Adaptive Finite Element Toolkit, http://www.numerik.uni-kiel.de/~mabr/gascoigne/
[4]OpenFEM: An Open-Source Finite Element Toolbox, http://www-rocq.inria.fr/OpenFEM/
[5]Getfem++, http://home.gna.org/getfem/
[6]deal.II: A Finite Element Differential Equations Analysis Library, http://www.dealii.org
[7]FEAST: Finite Element Analysis & Solutions Tools, http://www.feast.uni-dortmund.de
[8]Kaskade 7, http://www.zib.de/Numerik/numsoft/kaskade7/
[9]Matlab Partial Differential Equation Toolbox, http://www.mathworks.de/products/pde/
[10]COMSOL Multiphysics, http://www.comsol.de

The first optimization technique, which is known as steepest descent, goes back to GAUSS. Impelled by its enormous utility and challenged by complicated applications mathematical programs nowadays can handle even extremely nonlinear problems with a huge number of unknowns in the vector $\mathbf{x}$. For solving those *large scaled* optimization problems, inexact NEWTON methods, *Sequential Quadratic Programming* (SQP) methods or *Interior-Point* (IP) methods have been developed and implemented. For an excellent overview about available software we refer to the website "Decision Tree for Optimization Software"[11] but also to the book of [NW06]. Far from being complete we allude SQPlab[12], Ipopt[13] and very recently the sparse NLP solver WORHP[14]. Since it is used in some stages we also note the Matlab Optimization Toolbox[15].

0.3. PDE-constrained optimization

This manuscript joins the above introduced both topics. In PDE-constrained optimization one takes advantage on the one side of the long lasting expertise and deep knowledge from the field of PDEs and combines this together with well known techniques and matured skills from the other side of optimization. This was firstly carried out in the comprehensive monograph [Lio71] and continued by the already mentioned monograph [Fat99]. The book of [BGHvBW03] approaches the topic from an applicational and algorithmical point of view especially under the aspect of large scaled optimization. Recently the books of [Trö05], [NST06], [IK08] and [HPUU09] report the state of the art in PDE-constrained optimization.

Before we give a concrete example, it is useful to acquaint us with some specialties in optimal control theory. In this subject the variable $\mathbf{x} = (u, y) \in X$ is a partition of the *control variable* $u \in U$ and the *state variable* $y \in Y$. Now the PDE modeled by $c(\mathbf{x}) = 0$ implicitly describes how a certain control affects the state variable. As we will see, a control can influence a dynamical system in various manners. One can think of U to be an infinite dimensional function space on a *control domain*. Then one speaks of *distributed* or *boundary control*. In case of finite dimensional control the space U is isomorph to $\mathbb{R}^m$. Then a control prescribes the impact of a finite number of actuators like a certain shape interacting with the prescribed dynamical system coming up from a finite parametrization respectively. The precise description of the control space U will not be constituted at this stage. In order to complete this introductive examination we apply ourselves to a crucial topic of this manuscript. In PDE-constrained optimization it is for various reasons almost inevitable to include additional constraints such as $g(\mathbf{x}) \leq 0 \in Z_2$, which are exemplarily discussed later on. These constraints generally characterize the space of *admissible* controls $U_{ad} \subset U$ and the space of *admissible* states $Y_{ad} \subset Y$. Control and state constraints unfold their meaning depending on the BANACH space Z_2 and its equipped norm.

As already addressed let us investigate the announced keywords for a concrete class of problems at hand. The problem of optimal aerodynamic shape design is well suited because the arising subproblems are simultaneously easily conceivable and sufficiently complex. Roughly speaken in optimal aerodynamic shape design the task could be:

[11] Decision Tree for Optimization Software, http://plato.asu.edu/guide.html

[12] SQPlab: A Matlab solver of nonlinear optimization and optimal control problems, http://www-rocq.inria.fr/~gilbert/modulopt/optimization-routines/sqplab/sqplab.html

[13] Ipopt: Interior Point OPTimizer, https://projects.coin-or.org/Ipopt/

[14] WORHP: Large-scale sparse nonlinear optimization, http://www.worhp.de

[15] Matlab Optimization Toolbox, http://www.mathworks.de/products/optimization/

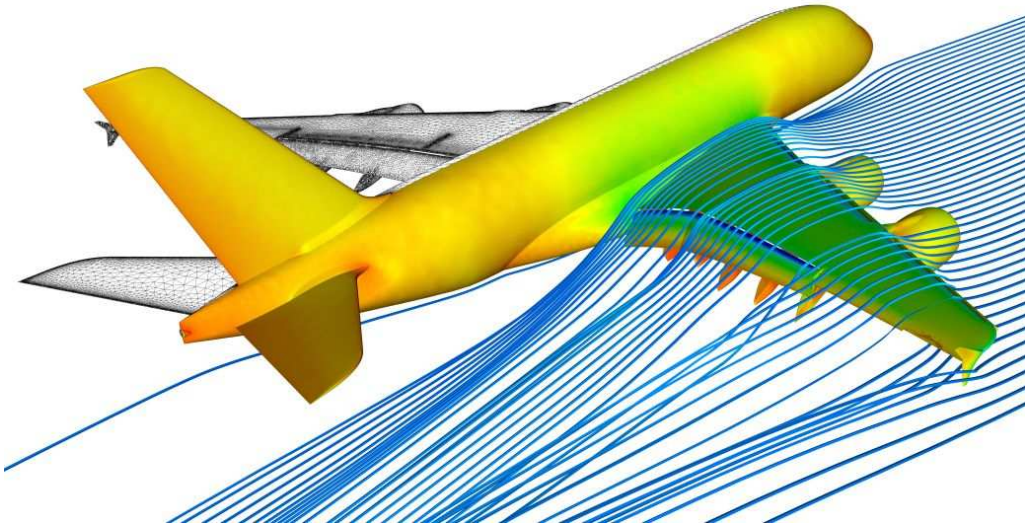


FIGURE 0.1. Result of a numerical flow simulation on an Airbus A380. The pressure distribution during flight can be seen on the fuselage and the flow distribution on the right wing. The left wing shows the used computational mesh. Copyright: DLR.

Optimize the shape of an aircraft such that it still flies.

"Optimize" means with respect to a multiobjective functional f reflecting goals like the minimization of the aircraft's drag, fuel consumption, costs of production and maintenance, inner material tensions to elongate the aircraft's lifetime and the maximization of the aircraft's lift and loading capacity.

The control u specifies the shape of the airplane from head over the wings to the end. Since the set of possible shapes is limited due to difficulties in manufacturing and empirical values, engineers usually dictate bounds on the shape and hence on the control in terms of specifying the set U_{ad}.

Now for a given shape u, traveling speed and traveling altitude of the aircraft via the compressible NAVIER-STOKES equations a corresponding state $y = (\vec{v}, p)$ of the air emerges, which consists of a velocity field $\vec{v}$ and a pressure distribution p. Let us easily write this physical interrelationship as $c(u, (\vec{v}, p)) = 0$ and keep in mind that this models a three dimensional, time dependent and extremely nonlinear partial differential equation due to the possible occurrence of turbulent flows. In order to refine the model even more equations can be added as for instance the heat equation on surfaces to include a temperature distribution θ. But also a force distribution $\vec{F}$ onto the wings via fluid-structure interaction can be considered. All this possibly leads to a state variable $y = (\vec{v}, p, \theta, \vec{F})$. Figure 0.1 shows the pressure distribution p on the fuselage as well as some streamlines from the flow $\vec{v}$ for an Airbus A380[16].

Besides the restrictions on the control additional state constraints naturally enter due to temperature bounds onto the used materials. Throughout all physical temperature distributions one should assure the temperature θ to be below the melting point especially in the critical area at the forefront of a hypersound aircraft. But also the appearing forces $\vec{F}$ may not have a too big magnitude.

[16]Airbus A380, DLR:
http://www.dlr.de/rd/en/Portaldata/1/Resources/portal_news/newsarchiv2007/A380_sim.jpg

We even take a step forward and impose constraints on the gradient of the state. This is indeed the case when not only the temperature θ itself is bounded but also the modulus of its gradient $|\nabla\theta|$ in order to avoid cracks in the material due to tension.

Optimal aerodynamic shape design is doubtlessly a suited realistic application to study the questions arising in PDE-constrained optimization. But there are similar challenging applications widely spread all over the fields of medicine, economics and industry. Besides the optimal hyperthermia treatment planning for cancer patients we mention the optimization of processes in chemical engineering like the cooling of glass or the growing of crystals. In the latter case one aims to optimally heat up the walls of a melting furnace in order to reach a planar phase transition between melted and solidified mass. Mathematically speaken this considers an optimal boundary control problem of a two-phase STEFAN problem where the moving of the phase transition is driven forward by the BOUSSINESQ approximation of the NAVIER-STOKES equations.

PDE-constrained optimization is a young field and has reached its importance not least by the progresses made in HPC. Within the last few years this discipline experienced to be in full bloom. On German's side the DFG-Priority Program 1253[17] has certainly made a contribution to this circumstance, to which the author also participates. Its major goal is to develop algorithms for optimal control problems with PDE constraints that satisfy the relation

$$\frac{\text{effort of optimization}}{\text{effort of simulation}} = \text{constant} \tag{0.2}$$

with a constant of moderate size. It goes without saying that for both numerator and denominator the best available methods should be used.

Recalling the example of optimal aerodynamic shape design under the aspect of degrees of freedom, it becomes clear that in the field of PDE-constrained optimization the discretized systems rapidly grow from 10^3 till 10^{10} number of unknowns. At the moment it is obviously impossible to approach all arising questions in order to obtain the goal (0.2) for a certain class of problem at once. Therefore and in order to show proofs we reduce the set of problems by simplifying the model equations and restricting assumptions. While picking up single questions in PDE-constrained optimization, this manuscript is going to make its contribution to achieve (0.2).

Therefore we exemplarily focus on stationary, linear elliptic PDEs in two or three space dimensions. The main nonlinearity enters through the presence of constraints involving the control, state and/or the gradient of the state. We develop structure exploiting GALERKIN schemes through a funded a priori and a posteriori error analysis. For deriving a priori estimates for the PDE-constrained problem of consideration we of course make use of the already well developed error analysis from PDEs. The furthermore we permanently apply the *variational discretization* concept proposed by Hinze in [Hin05]. From the numerical point of view we essentially make use of the involved KKT equations and apply regularization techniques. Since CPU-time and memory is bounded, one naturally asks for error control and optimal complexity. We combine techniques from linear algebra such as factorization and preconditioning and derive problem suited a posteriori error estimators in order do extend the technique of *Dual Weighted Residuals* (DWR) proposed by [BR96]

[17] DFG-Priority Program 1253, http://www.am.uni-erlangen.de/home/spp1253/

to the presence of additional constraints. The latter one falls into the context of *goal-oriented* adaptivity.

0.4. Outline

This manuscript is structured as follows: In the first chapter we state often required basic definitions and properties concerning elliptic partial differential equations and their discretizations in terms of domain partitions and finite element spaces.

Chapter 2 deals with elliptic optimal control problems under control constraints

$$g(\mathbf{x}) = \begin{pmatrix} u_a - u \\ u - u_b \end{pmatrix} \leq 0$$

and mainly consists of two parts. The first part imitates the results from the work [DGH09b] together with Deckelnick and Hinze where a finite element approximation of a DIRICHLET boundary control problem on two- and three-dimensional curved domains is considered. A priori error estimates are improved by additional assumptions on the underlying meshes. The second part of Chapter 2 addresses bounded, distributed control problems under the aspect of variational discretization. Besides the introduction of useful notations hints for its numerical implementation are given.

After citation of available literature about a priori error estimates for state constrained optimal control problems with

$$g(\mathbf{x}) = \begin{pmatrix} y_a - y \\ y - y_b \end{pmatrix} \leq 0$$

Chapter 3 focusses onto a posteriori error estimation and goal-oriented adaptivity for such problems. While the study of an unregularized, purely state constrained problem is leaned on the paper [GH08] together with Hinze, we further investigate a goal-oriented adaptive Moreau-Yosida algorithm for control- and state-constrained elliptic optimal control problems taken from the work [GT09] together with Tber.

Chapter 4 is devoted to constraints on the gradient of the state

$$g(\mathbf{x}) = |\nabla y| - \delta \leq 0.$$

Regularity theory requires to consider two different scenarios. On the one hand we consider a mixed formulation of a quadratic optimal control problem with additional control constraints descended from the work [DGH09c] together with Deckelnick and Hinze. Secondly a L^r-regularization of the control in the objective ($r > d$) allows to omit control bounds. The a priori error analysis of different discretization schemes originates from the article [GH09] together with Hinze.

Finally we conclude our findings in Chapter 5.

CHAPTER 1

Preliminaries

1.1. Elliptic partial differential equations

The idea of this section is to concretize possible PDE constraints of consideration, i.e. let us specify $c(\mathbf{x}) = 0$ in problem (0.1). Besides all different types of PDEs, the emphasis for this manuscript clearly lies on the second order elliptic kind.

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with LIPSCHITZ-boundary $\partial\Omega$. We define the differential operator

$$\mathcal{A}y := -\sum_{i,j=1}^{d} \partial_{x_j}\big(a_{ij} y_{x_i}\big) + \sum_{i=1}^{d} b_i y_{x_i} + cy, \tag{1.1}$$

along with its formal adjoint operator

$$\mathcal{A}^* y = -\sum_{i=1}^{d} \partial_{x_i}\Big(\sum_{j=1}^{d} a_{ij} y_{x_j} + b_i y\Big) + cy \tag{1.2}$$

where for the coefficients we assume $a_{ij}, b_i, \sum_{k=1}^{d} \partial_{x_k} b_k, c \in L^\infty(\Omega)$ for all $i, j = 1, \dots, d$.

Now for given functions $f_u : \Omega \to \mathbb{R}$, $g_u : \partial\Omega \to \mathbb{R}$ and boundary operator $\mathcal{B}$ we consider the problem of finding a function $y : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} \mathcal{A}y &= f_u \quad \text{in } \Omega, \\ \mathcal{B}y &= g_u \quad \text{in } \partial\Omega. \end{aligned} \tag{1.3}$$

DIRICHLET-boundary conditions can be modeled by $\mathcal{B} = id$, while NEUMANN-boundary conditions are obtained when choosing $\mathcal{B}y = \sum_{i,j=1}^{d} a_{ij} y_{x_i} \nu_j$. Here $\vec{\nu}$ denotes the unit outward normal to $\partial\Omega$. In order to ensure well-posedness of problem (1.3), we have to make several assumptions on the operator $\mathcal{A}$. To be more precise we subsequently assume that $\mathcal{A}$ is *elliptic*, which is specified in [EG04, Def. 3.1]:

Definition 1.1.1. The operator $\mathcal{A}$ from (1.1) is said to be *elliptic* if there exists $c_0 > 0$ such that

$$\forall\, \boldsymbol{\xi} \in \mathbb{R}^d, \qquad \sum_{i,j=1}^{d} a_{ij} \xi_i \xi_j \geq c_0 |\boldsymbol{\xi}|^2 \quad \text{a.e. in } \Omega.$$

Equation (1.3) is then called an *elliptic PDE*.

We associate with $\mathcal{A}$ the bilinear form

$$a(y, \phi) := \int_\Omega \Big(\sum_{i,j=1}^{d} a_{ij} y_{x_i} \phi_{x_j} + \sum_{i=1}^{d} b_i y_{x_i} \phi + cy\phi \Big), \qquad y, \phi \in H^1(\Omega) \tag{1.4}$$

and suppose that the form a is *coercive* on $V = H^1(\Omega)$ for NEUMANN-boundary conditions or $V = H_0^1(\Omega)$ for DIRICHLET-boundary conditions, i.e. there exists $c_1 > 0$

such that

$$a(\phi,\phi) \geq c_1 \|\phi\|_V^2 \qquad \text{for all } \phi \in V. \tag{1.5}$$

In [EG04, Thm. 3.8] sufficient conditions for a being coercive on V are proven. For the convenience of the reader we summarize this result in

Remark 1.1.2. Set $p = \operatorname{ess\,inf}_{x\in\Omega}(c - \frac{1}{2}\sum_{i=1}^d \partial_{x_i} b_i)$ and let c_Ω be the constant in the POINCARÉ inequality. For the DIRICHLET problem a is coercive on $V = H_0^1(\Omega)$ if $c + \min(0, \frac{p}{c_\Omega}) > 0$. For the NEUMANN problem a is coercive on $V = H^1(\Omega)$ if $p > 0$ and $\operatorname{ess\,inf}_{x\in\partial\Omega}(\sum_{i=1}^d b_i \nu_i) \geq 0$.

We introduce $f \in V'$ for the homogeneous DIRICHLET boundary problem by $f(\phi) = \int_\Omega f_u \phi$ and for NEUMANN boundary conditions to be $f(\phi) = \int_\Omega f_u\phi + \int_{\partial\Omega} g_u \phi$. Now we can rewrite the elliptic PDE (1.3) as

$$\text{Seek } y \in V \text{ such that } a(y,\phi) = f(\phi) \qquad \forall\, \phi \in V \tag{1.6}$$

to bring it into the framework of the Lax-MILGRAM lemma (see [EG04, Lem. 2.2])

Lemma 1.1.3. *Let V be a Hilbert space, let $a \in \mathcal{L}(V \times V; \mathbb{R})$, and let $f \in V'$. Assume that the bilinear form a is coercive with constant $c_1 > 0$. Then, problem* (1.6) *admits one and only one solution with a priori estimate*

$$\|y\|_V \leq \frac{1}{c_1}\|f\|_{V'} \qquad \forall\, f \in V'. \tag{1.7}$$

We denote the solution $y \in V$ of problem (1.6) by $y =: \mathcal{G}(f_u, g_u)$. Now we overload the meaning of the solution operator $\mathcal{G}$. Let $b_i = 0$ for $i = 1, \dots, d$ in (1.1). For a given function $f_u \in L^2(\Omega)$, $g_u \in C^{0,1}(\partial\Omega)$ it is well–known that the elliptic boundary value problem

$$\begin{aligned} \mathcal{A}y &= f_u \quad \text{in } \Omega, \\ y &= g_u \quad \text{on } \partial\Omega. \end{aligned} \tag{1.8}$$

can be written in mixed formulation. To this purpose we introduce

$$H(\operatorname{div};\Omega) := \{\vec{w} \in L^2(\Omega)^d \,:\, \operatorname{div}\vec{w} \in L^2(\Omega)\}$$

and denote $\vec{v} := A\nabla y$, where $A(x) := \big(a_{ij}(x)\big)_{i,j=1}^d$. Then $(y, \vec{v})$ satisfies

$$\int_\Omega A^{-1}\vec{v}\cdot\vec{w} + \int_\Omega y \operatorname{div}\vec{w} - \int_{\partial\Omega} g_u\, \vec{w}\cdot\vec{\nu} = 0 \qquad \forall\, \vec{w} \in H(\operatorname{div};\Omega) \tag{1.9a}$$

$$\int_\Omega z \operatorname{div}\vec{v} - \int_\Omega c y\, z + \int_\Omega f_u\, z = 0 \qquad \forall\, z \in L^2(\Omega). \tag{1.9b}$$

In what follows it will be convenient to write $(y, \vec{v}) = \mathcal{G}(f_u, g_u)$ for the solution of (1.9). The different meaning of $\mathcal{G}$ can be figured out through the number of components of $\mathcal{G}(f_u, g_u)$.

1.2. Finite element discretization

In order to carry out error analysis and to numerically solve the above PDEs, we are going to apply the finite element method. The aim of this section is to introduce simplicial partitions of the computational domain, to define the often recurring finite element spaces in the following chapters and to explain approximate finite element solutions for the involved PDEs. Moreover we will find some useful notations and cite basic required properties related to finite elements.

The following basic definitions can be found for instance in [EG04, Sec. 1.2].

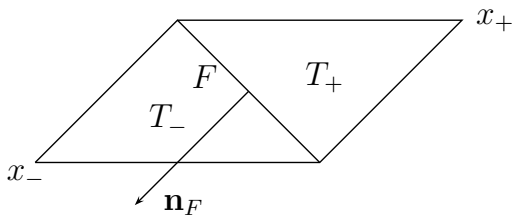


FIGURE 1.1. Setting in Definition 1.2.3.

Definition 1.2.1. Let $\{\mathbf{a}_0, \dots, \mathbf{a}_d\}$ be a family of points $\mathbb{R}^d$ ($d \geq 1$) such that the vectors $\mathbf{a}_1 - \mathbf{a}_0, \dots, \mathbf{a}_d - \mathbf{a}_0$ are linear independent in $\mathbb{R}^d$. The convex set

$$T = \text{conv hull}\{\mathbf{a}_0, \dots, \mathbf{a}_d\} \tag{1.10}$$

is a called a *d-simplex* in $\mathbb{R}^d$. Let $1 \leq i \leq d$. We say $\mathbf{a}_i$ is a *vertex*. The set

$$F_i := \text{conv hull}\{\mathbf{a}_j \,:\, 0 \leq j \leq d, j \neq i\} \subset \partial T \tag{1.11}$$

is called *face* of T opposite to $\mathbf{a}_i$. The vector $\mathbf{n}_i$ denotes the unit outward normal to F_i. Note that for $d = 2$ a face is also called an *edge*. The *unit d-simplex* in $\mathbb{R}^d$ is the set

$$\hat{T} := \{x \in \mathbb{R}^d \,:\, x_i \geq 0, 1 \leq i \leq d, \text{and } \sum_{i=1}^{d} x_i \leq 1\}. \tag{1.12}$$

An invertible, differentiable mapping $\vec{F}_T : \hat{T} \to T \subset \mathbb{R}^d$ is given by the affine linear parametrization

$$\vec{F}_T(\hat{x}) = \mathbf{A}_T \hat{x} + \mathbf{a}_0, \tag{1.13}$$

where $\mathbf{A}_T := [\mathbf{a}_1 - \mathbf{a}_0, \dots, \mathbf{a}_d - \mathbf{a}_0] \in \mathbb{R}^{d \times d}$. We further define the *diameter*

$$h_T := \text{diam}(T) = \max_{x_1, x_2 \in T} |x_1 - x_2| \tag{1.14}$$

and the *inball-radius*

$$\rho_T := \sup\{r \,:\, B_r \subset T \text{ is a } d\text{-ball of radius } r\} \tag{1.15}$$

of T. With $|T|$ we denote the LEBESGUE-d-measure of T.

Definition 1.2.2. Let $d \leq 3$ and $\mathcal{T}_h$ be a set of d-simplexes in $\mathbb{R}^d$. We say $\mathcal{T}_h$ is a *conforming simplicial mesh* or *conforming triangulation* of

$$\Omega := \text{int} \bigcup_{T \in \mathcal{T}_h} T \subset \mathbb{R}^d \tag{1.16}$$

if and only if for two different simplexes $T_1, T_2 \in \mathcal{T}_h$ the intersection $T_1 \cap T_2$ is either empty or a vertex or a complete face. We further introduce the *maximum* and *minimum mesh size* $h := \max_{T \in \mathcal{T}_h} h_T$ and $h_{\min} := \min_{T \in \mathcal{T}_h} h_T$ of $\mathcal{T}_h$. We refer to the vertices $x_1, \dots, x_m \in \bar{\Omega}$ the *set of nodes* $\mathcal{N}_h := \bigcup_{i=1}^{m} \{x_i\}$. The *number of elements* is denoted by $nt := \text{card}(\mathcal{T}_h)$. The *set of edges* $\mathcal{E}_h$ has cardinality $ne := \text{card}(\mathcal{E}_h)$, while the *set of faces* $\mathcal{F}_h$ has got $nf := \text{card}(\mathcal{F}_h)$ elements.

According to [BC05, Def. 4.2] we further introduce notation for elements that share a face $F \in \mathcal{F}_h$.

Definition 1.2.3. Let $F \in \mathcal{F}_h$ be an interior face of Ω. For the vertex $x_\pm$ opposite to F we define $T_\pm = \operatorname{conv}\operatorname{hull}\{F \cup \{x_\pm\}\}$ such that the face $F = \operatorname{conv}\operatorname{hull}\{\mathbf{a}_i : 0 \le i \le d, \mathbf{a}_i \neq x_\pm\}$ has the right hand rule orientation. Then $\mathbf{n}_F$ points outwards from T_+ to T_-. If $F \subset \partial\Omega$ is an exterior face, then $\mathbf{n}_F$ is the exterior normal and F defines T_+ (and T_- is undefined).

Definition 1.2.4. The family of triangulations $\{\mathcal{T}_h\}$, $0 < h \le 1$ is said to be *quasi-uniform* if there exists a constant $\kappa > 0$ (independent of h) such that each $T \in \mathcal{T}_h$ is contained in a ball of radius $\kappa^{-1}h$ and contains a ball of radius κh.

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded LIPSCHITZ-domain. We suppose that $\bar{\Omega}$ is the union of the elements of the triangulation $\mathcal{T}_h$. For the case that the domain boundary $\partial\Omega$ is sufficiently smooth we allow boundary elements to have one curved face. For further details we refer to the books of Ern and Guermond [EG04, Sec. 1.3], Ciarlet [Cia80] and Section 2.1.2. For ease of exposition we subsequently assume the domain to be polygonally $(d = 2)$ or polyhedrally $(d = 3)$ bounded.

We now give a precise description of the required finite element spaces.

Definition 1.2.5. The space of *piecewise constant finite elements* is

$$P^0_{td,h}(\mathcal{T}_h) := \{\phi_h \in L^2(\Omega) : \phi_{h|T} \in \mathbb{P}^0(T) \text{ for all } T \in \mathcal{T}_h\}. \tag{1.17}$$

The set $\{\mathbb{1}_T \in L^2(\Omega) : T \in \mathcal{T}_h\}$ is a nt-dimensional basis of $P^0_{td,h}(\mathcal{T}_h)$ consisting of totally discontinuous functions.

Definition 1.2.6. The space of *linear finite elements* is

$$P^1_{c,h}(\mathcal{T}_h) := \{\phi_h \in C^0(\bar{\Omega}) : \phi_{h|T} \in \mathbb{P}^1(T) \text{ for all } T \in \mathcal{T}_h\}. \tag{1.18}$$

Let $\phi_i \in P^1_{c,h}(\mathcal{T}_h)$ with $\phi_i(x_j) = \delta_{ij}$ for all $x_j \in \mathcal{N}_h$ and $i, j \in \{1, \dots, m\}$. Here, δ_{ij} represents the KRONECKER symbol. The set $\{\phi_i : i = 1, \dots, m\}$ is called the m-dimensional *standard nodal basis* or LAGRANGE *basis* of $P^1_{c,h}(\mathcal{T}_h)$. Furthermore, we introduce the LAGRANGE *interpolation operator* $I_h : C^0(\bar{\Omega}) \to P^1_{c,h}(\mathcal{T}_h)$ by

$$I_h v := \sum_{i=1}^{m} v(x_i)\phi_i \quad \text{for all } v \in C^0(\bar{\Omega}). \tag{1.19}$$

For $T \in \mathcal{T}_h$ and LAGRANGE basis function ϕ with $T \subset \operatorname{supp}(\phi)$ we introduce the local basis function $\varphi : T \to \mathbb{P}^1(T)$ as $\varphi := \phi_{|T}$.

Since beside others we are going to derive numerical solution algorithms associated to linear finite element solutions of PDEs, it will be useful to introduce vector- and matrix-notation for it.

Definition 1.2.7. For $\phi_i, \phi_j \in P^1_{c,h}(\mathcal{T}_h)$ we introduce the *mass-matrix* $\mathbf{M} \in \mathbb{R}^{m\times m}$ with entries

$$m_{ij} := \int_\Omega \phi_i \phi_j \,\mathrm{d}x \qquad i, j \in \{1, \dots, m\}.$$

Given a bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ we further define the *stiffness-matrix* $\mathbf{A} \in \mathbb{R}^{m\times m}$ with entries

$$a_{ij} := a(\phi_i, \phi_j) \qquad i, j \in \{1, \dots, m\}.$$

Remark 1.2.8. $\mathbf{M}$ is symmetric and positive definite. If a is a coercive bilinear form the matrix $\mathbf{A}$ is merely positive semidefinite.

Remark 1.2.9. With the help of the set of *neighboring indices*

$$\mathcal{N}_i := \{j \in \{1,\dots,m\} \,:\, \exists T \in \mathcal{T}_h \,:\, \{x_i, x_j\} \subset T\}$$

related to a vertex x_i one can characterize the sparsity structure of $\mathbf{M}$ by

$$m_{ij} > 0 \qquad \forall\, i \in \{1,\dots,m\}, j \in \mathcal{N}_i, \tag{1.20a}$$

$$m_{ij} = 0 \qquad \forall\, i \in \{1,\dots,m\}, j \notin \mathcal{N}_i. \tag{1.20b}$$

We are also going to make use of the following

Definition 1.2.10. We call the diagonal matrix $\tilde{\mathbf{M}} \in \mathbb{R}^{m\times m}$ with

$$\tilde{\mathbf{M}} := \operatorname*{diag}_{i=1,\dots,m} \left(\textstyle\sum_{j=1}^m |m_{ij}|\right) \tag{1.21}$$

the *lumped mass-matrix*. Its inverse is obviously given by

$$\tilde{\mathbf{M}}^{-1} = \operatorname*{diag}_{i=1,\dots,m} \left(\tfrac{1}{\sum_{j=1}^m |m_{ij}|}\right).$$

Definition 1.2.11. We represent a finite element function $v_h \in P^1_{c,h}(\mathcal{T}_h)$ by the corresponding vector $\mathbf{v} = [v_i]_{i=1}^m$ due to the equality

$$v_h(x) = \sum_{i=1}^m v_i \phi_i(x) \qquad \forall\, x \in \Omega.$$

Moreover for a given function $w \in L^2(\Omega)$ we introduce the vector $\hat{\mathbf{w}} \in \mathbb{R}^m$ by

$$\hat{\mathbf{w}} := \left[\textstyle\int_\Omega w\phi_i\right]_{i=1}^m \tag{1.22}$$

Clearly if $w_h \in P^1_{c,h}(\mathcal{T}_h)$ then $\hat{\mathbf{w}}_\mathbf{h} = \mathbf{M}\mathbf{w}$.

For later active set calculus we introduce the blockwise split of a matrix within

Definition 1.2.12. For a given matrix $\mathbf{B} \in \mathbb{R}^{m\times m}$ and a disjoint index decomposition $\mathcal{V}, \mathcal{A}, \mathcal{I} \subset \bullet := \{1,\dots,m\}$ we use the following notation for a *blockwise* decomposition of $\mathbf{B}$:

$$\mathbf{B} = \mathbf{B}_\bullet = \begin{bmatrix} \mathbf{B}_\mathcal{I} & \mathbf{B}_{\mathcal{I}\mathcal{A}} \\ \mathbf{B}_{\mathcal{A}\mathcal{I}} & \mathbf{B}_\mathcal{A} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_\mathcal{I} & \mathbf{B}_{\mathcal{I}\mathcal{V}} & \mathbf{B}_{\mathcal{I}\mathcal{A}} \\ \mathbf{B}_{\mathcal{V}\mathcal{I}} & \mathbf{B}_\mathcal{V} & \mathbf{B}_{\mathcal{V}\mathcal{A}} \\ \mathbf{B}_{\mathcal{A}\mathcal{I}} & \mathbf{B}_{\mathcal{A}\mathcal{V}} & \mathbf{B}_\mathcal{A} \end{bmatrix},$$

where $\mathcal{A} := \mathcal{V} \cup \mathcal{A}$. We neglect possible permutations in rows and columns. Furthermore we write $\mathbf{B}_\mathcal{I}$ instead of $\mathbf{B}_{\mathcal{I}\mathcal{I}}$ respectively. If $\mathbf{B}_\mathcal{I}$ is regular its inverse is denoted by $\mathbf{B}_\mathcal{I}^{-1}$ instead of $(\mathbf{B}_\mathcal{I})^{-1}$. Analogously we introduce this decomposition for $\mathcal{E}, \mathcal{F}, \Theta \subset \bullet$ and $\mathcal{B} := \mathcal{E} \cup \mathcal{F}$.

We now consider approximate linear finite element solutions to problem (1.6). Let $V_h = P^1_{c,h}(\mathcal{T}_h)$ for Neumann boundary conditions or $V_h = P^1_{c,h}(\mathcal{T}_h) \cap H^1_0(\Omega)$ for the homogeneous Dirichlet case. The approximate linear finite element solution $y_h = \mathcal{G}_h(f_u, g_u)$ to problem (1.6) in the finite dimensional subspace $V_h \subset V$ solves the problem:

$$\text{Seek } y_h \in V_h \text{ such that } a(y_h, \phi_h) = f(\phi_h) \qquad \forall\, \phi_h \in V_h. \tag{1.23}$$

Similar to the solution operator $\mathcal{G}$ we also overload the meaning of $\mathcal{G}_h$ due to the already announced mixed formulation (1.9). For our considerations it is sufficient to consider the mixed finite element method based on the lowest order Raviart–Thomas element from [RT77] (compare also [BC05]), which is part of

Definition 1.2.13. The space of *lowest order Raviart–Thomas elements* is

$$RT_0(\mathcal{T}_h) := \{\vec{w}_h \in H(\mathrm{div};\Omega) \,:\, \vec{w}_{h|T} \in RT_0(T) \text{ for all } T \in \mathcal{T}_h\}, \tag{1.24}$$

where

$$RT_0(T) := \{\vec{w} : T \to \mathbb{R}^d \,:\, \vec{w}(x) = \mathbf{a} + \beta x, \mathbf{a} \in \mathbb{R}^d, \beta \in \mathbb{R}\}.$$

Let $F \in F_h$ be a fixed face. According to Definition 1.2.3 there are either two elements T_+ and T_- in $\mathcal{T}_h$ with face $F = \partial T_+ \cap \partial T_-$ or there is exactly one element T_+ in $\mathcal{T}_h$ with $F \subset \partial T_+$. Then if $T_\pm = \text{conv hull}\{F \cup \{x_\pm\}\}$ for the vertex $x_\pm$ opposite to F of $T_\pm$ set

$$\vec{\psi}_F(x) := \begin{cases} \pm\frac{|F|}{d|T_\pm|}(x - x_\pm) & \text{for } x \in T_\pm, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.25}$$

The set $\{\vec{\psi}_F \,:\, F \in \mathcal{F}_h\}$ forms an nf-dimensional basis of $RT_0(\mathcal{T}_h)$.

A useful property of the basis functions $\vec{\psi}_F \in RT_0(\mathcal{T}_h)$ is stated from [BC05, Lem. 4.1] in the following

Lemma 1.2.14. *Let $F \in \mathcal{F}_h$. There holds*

$$\vec{\psi}_F \cdot \mathbf{n}_F = \begin{cases} 0 & \text{along } \bigcup_{F' \in \mathcal{F}_h, F' \neq F} F' \\ 1 & \text{along } F. \end{cases} \tag{1.26}$$

With the definition of the lowest order Raviart–Thomas elements at hand we can continue to define a discrete approximation operator $\mathcal{G}_h$ according to the mixed system (1.9). We denote the solution of

$$\int_\Omega A^{-1}\vec{v}_h \cdot \vec{w}_h + \int_\Omega y_h \,\mathrm{div}\vec{w}_h - \int_\Gamma g_u \,\vec{w}_h \cdot \vec{\nu} = 0 \qquad \forall\, \vec{w}_h \in RT_0(\mathcal{T}_h) \tag{1.27a}$$

$$\int_\Omega z_h \,\mathrm{div}\vec{v}_h - \int_\Omega c y_h \, z_h + \int_\Omega f_u \, z_h = 0 \qquad \forall\, z_h \in P^0_{td,h}(\mathcal{T}_h) \tag{1.27b}$$

by $(y_h, \vec{v}_h) =: \mathcal{G}_h(f_u, g_u) \in P^0_{td,h}(\mathcal{T}_h) \times RT_0(\mathcal{T}_h)$.

We now make a big jump towards a numerical tool to investigate proven orders of convergence for instance for $L^2(\Omega)$-errors of involved functions over the domain. Corresponding to different mesh parameters $h = h_1$ and $h = h_2$ we close this chapter with the following

Definition 1.2.15. For an error functional $E : (0,\infty) \to (0,\infty)$ and given parameters $h_1, h_2 \in (0,\infty)$ we define the *Experimental Order of Convergence* (EOC) by

$$\text{EOC} := \frac{\log E(h_1) - \log E(h_2)}{\log h_1 - \log h_2}.$$

CHAPTER 2

Control constraints

In this chapter let us investigate structure exploiting GALERKIN schemes for optimization problems governed by elliptic PDEs under additional control constraints.

Our considerations are twofold. At the first stage we study DIRICHLET boundary control problems, which are well suited for an introductory insight, since those kinds of problems can be reduced from the d-dimensional domain Ω to its $(d-1)$-dimensional manifold $\partial\Omega$. At the second stage we focus onto distributed control problems and give details to the numerical implementation of the variational discretization approach.

2.1. Optimal Dirichlet boundary control on curved domains

This section imitates the results from [DGH09b] which are also summarized in [DGH09a]. We consider the variational discretization of elliptic DIRICHLET optimal control problems with constraints on the control. The underlying state equation, which is considered on smooth two- and three-dimensional domains, is discretized by linear finite elements taking into account domain approximation. The control variable is not discretized. We obtain optimal error bounds for the optimal control in two and three space dimensions and derive a superconvergence result in 2d provided that the underlying mesh satisfies some additional condition. We confirm our analytical findings by numerical experiments.

This section is organized as follows. After giving an overview about the related literature in Subsection 2.1.0 we present the mathematical setting and formulate the optimal control problem in Subsection 2.1.1. In Subsection 2.1.2 we examine the finite element discretization of the state equation taking into account the approximation of the domain. In Subsection 2.1.3 we introduce the discrete control problem and obtain an optimal error estimate for the discrete controls. Subsection 2.1.4 deals with superconvergence properties of boundary controls induced by finite element partitions with certain regularity properties. In Subsection 2.1.5 we finally present numerical results which confirm our analytical findings.

2.1.0. Introduction. DIRICHLET boundary control plays an important role in many practical applications such as active boundary control of flows. If one is interested in control by blowing and suction on parts of the boundary only, boundary controls with low regularity should be admissible which even may develop jump discontinuities. Typical control spaces here would be $L^2(\Gamma)^d, L^\infty(\Gamma)^d$, where Γ denotes the part of the boundary of the domain where the control is applied and d is the spatial dimension. In model based optimization with boundary controls the flow often is modeled with the help of the NAVIER-STOKES equations whose classical variational formulation does not allow for DIRICHLET boundary data with jump discontinuities, see [FGH98, HK04], so that the concept of very weak solutions [LM72] has to be applied instead, see [Ber04] for a more detailed discussion. Moreover, pointwise

bounds on the control actions have to be considered in practice. The related projection in $L^2(\Gamma)$ then can be easily evaluated. For a survey of different formulations of DIRICHLET boundary control problems we refer to [KV07].

Here we consider as model problem DIRICHLET boundary control of an elliptic equation with L^2-boundary controls subject to pointwise bounds on the controls. The state equation is posed on a bounded, sufficiently smooth domain in $\mathbb{R}^d$, $d = 2, 3$. Our aim is to develop and analyze a finite element concept which is tailored to the numerical treatment of pointwise bounds, and at the same time is able to cope with the low regularity of the control and the state. To this purpose we propose an approximation of the state equation using piecewise linear, continuous finite elements taking into account domain approximation. The controls are not discretized explicitly, but implicitly (variationally) through the optimality conditions associated with the discrete optimal control problem. Our main result, see Theorem 2.1.4, is an $\mathcal{O}(h\sqrt{|\log h|})$ bound for the L^2–error of optimal control and state. In two space dimensions and under additional conditions on the underlying mesh we are able derive the improved error bound $\mathcal{O}(h^{\frac{3}{2}})$, which reflects a superconvergence effect.

There are only few contributions to DIRICHLET boundary control reported in the literature. [CR06] consider semilinear elliptic DIRICHLET boundary control problems with pointwise bounds on two-dimensional convex polygonal domains Ω. Denoting by u the optimal control they are able to prove the optimal result

$$\|u - u_h\|_{L^2(\partial\Omega)} \leq Ch^{1-1/p}.$$

Here, u_h denotes the optimal discrete boundary control which they find in the space of piecewise linear, continuous finite elements on $\partial\Omega$, and $p \geq 2$ depends on the smallest angle of the boundary polygon. This also had been numerically investigated by the author in [HPUU09, Sec. 3.2.7.4]. Therein as domain a duodecagon with maximum inner angle $\frac{5}{6}\pi$ is considered. Besides the classical approach using linear finite elements as in [CR06] also variational discretization combined with a mixed finite element approximation of the state equation based on the lowest order Raviart–Thomas elements is carried out.

For control functions of the form

$$B\mathbf{q} := \sum_{i=1}^{n} q_i f_i$$

with given $f_i \in H^{5/2}(\Gamma)$ and box-constrained $\mathbf{q} \in \mathbb{R}^n$, [Vex07] provides a finite element analysis for two-dimensional bounded polygonal domains and proves

$$|\mathbf{q} - \mathbf{q}_h| \leq Ch^2.$$

In [MRV08] May, Rannacher and Vexler consider DIRICHLET boundary control without control constraints on two-dimensional convex polygonal domains, where they present optimal error estimates for the state and the adjoint state. Important ingredients are duality techniques and an optimal error estimate in $H^{-1/2}(\Gamma)$ for the control.

Recently in [CS10] Casas and Sokolowski compare the solutions of control constrained optimal DIRICHLET control problems between a convex domain $\Omega \subset \mathbb{R}^2$ and its polygonally approximated domain Ω_h with maximal edge length h. Each of them are infinite dimensional problems without any further discretization. For the effect of small changes in the domain they prove

$$\|u - u_h \circ g_h^{-1}\|_{L^2(\partial\Omega)} \leq Ch,$$

where g_h is an appropriate bijective mapping from $\partial\Omega_h$ to $\partial\Omega$.

The paradox of observing a control error of order $\mathcal{O}(h^{\frac{3}{2}})$ for superconvergence meshes and the above estimate of order $\mathcal{O}(h)$ for the exclusive domain approximation is explained by Casas, Mateos and the author in the work [CGM10]. Therein it is proven that the order of $\mathcal{O}(h)$ is optimal by the construction of an analytic example without control constraints. This leads to the paradox, that the numerical solution is a better approximation of the optimal control than the exact one obtained just by changing the domain Ω to Ω_h.

Numerical analysis for NEUMANN- and ROBIN-type boundary control of general elliptic control problems is provided by Casas and Mateos in [CM08], where they investigate several discrete concepts for the controls including variational discretization. The latter concept is also applied by Hinze and Matthes to ROBIN- and NEUMANN-type boundary control in [HM09], where also L^∞-estimates for the error in the controls are provided.

2.1.1. Mathematical setting. Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with a C^3–boundary $\Gamma := \partial\Omega$ and consider the differential operator $\mathcal{A}$ from (1.1) with $a_{ij} = a_{ji}$ and $b_i = 0$ for $i, j = 1, \dots, d$ so that $\mathcal{A}$ is selfadjoint, i.e. $\mathcal{A} = \mathcal{A}^*$. In what follows we assume that the coefficients a_{ij} and c belong to $C^2(\bar{\Omega})$, $c \geq 0$ and that $\mathcal{A}$ is elliptic in the sense of Definition 1.1.1.

Given $f \in L^2(\Omega), u \in L^2(\Gamma)$ we consider the boundary value problem

$$\begin{aligned} \mathcal{A}y &= f \quad \text{in } \Omega, \\ y &= u \quad \text{on } \Gamma, \end{aligned} \tag{2.1}$$

which is obtained by setting $\mathcal{B} = id$, $g_u = u$ and $f_u = f$ in (1.3). This problem has a unique solution $y \in H^{\frac{1}{2}}(\Omega)$ which we denote by $y = \mathcal{G}(u)$ and which solves (2.1) in the sense that

$$\int_\Omega y\mathcal{A}\phi = \int_\Omega f\phi - \int_\Gamma u\partial_{\vec{\nu}_\mathcal{A}}\phi \qquad \forall \phi \in H^2(\Omega) \cap H_0^1(\Omega). \tag{2.2}$$

Here, $\partial_{\vec{\nu}_\mathcal{A}}\phi = \sum_{i,j=1}^d a_{ij}\phi_{x_j}\nu_i$ and $\vec{\nu}$ is the outer unit normal to Γ. Let us briefly sketch the existence of y in the case $f \equiv 0$: Denote by $T : L^2(\Omega) \to L^2(\Gamma)$ the linear operator which is defined by $T\psi := -\partial_{\vec{\nu}_\mathcal{A}}\phi$ where $\phi \in H^2(\Omega) \cap H_0^1(\Omega)$ is the unique solution of $\mathcal{A}\phi = \psi$ in Ω, $\phi = 0$ on Γ. Letting $y := T^*u$, where T^* is the adjoint of T, it is not difficult to verify that y satisfies (2.2). The fact that y belongs to $H^{\frac{1}{2}}(\Omega)$ follows from an estimate of the form

$$\left| \int_\Omega y\psi \right| \leq C\|u\|_{L^2(\Gamma)}\|\psi\|_{H^{-\frac{1}{2}}(\Omega)}$$

for $\psi \in L^2(\Omega)$, compare [Cas85] for a similar argumentation.
In order to define an approximation of (2.1) we recall from (1.4) the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ associated with the differential operator $\mathcal{A}$ as

$$a(y, \phi) = \sum_{i,j=1}^d \int_\Omega \big(a_{ij}y_{x_i}\phi_{x_j} + cy\phi\big).$$

By our assumptions and Remark 1.1.2 a is coercive on $H_0^1(\Omega)$ in terms of satisfying inequality (1.5).

Next, let $\alpha > 0$ and $y_0 \in W^{1,\bar{r}}(\Omega)$, $\bar{r} > d$ be given. We then consider the DIRICHLET boundary control problem

$$\begin{aligned} &\min_{u\in U_{ad}} J(u) = \frac{1}{2}\int_\Omega |y-y_0|^2 + \frac{\alpha}{2}\int_\Gamma |u|^2 \\ &\text{subject to } y = \mathcal{G}(u), \end{aligned} \tag{2.3}$$

where

$$U_{ad} = \{u \in L^2(\Gamma) \,:\, u_a \le u \le u_b \text{ a.e. on } \Gamma\}$$

and $u_a, u_b \in \mathbb{R}, u_a < u_b$. Existence of a unique solution $u \in U_{ad}$ of (2.3) follows by standard arguments. This solution is characterized by the variational inequality

$$\int_\Omega (y-y_0)(z-y) + \alpha\int_\Gamma u(v-u) \ge 0 \qquad \forall\, v \in U_{ad} \tag{2.4}$$

where $z = \mathcal{G}(v)$. Let us introduce the adjoint state $p \in H^2(\Omega)\cap H_0^1(\Omega)$ as the solution of the following boundary value problem:

$$\begin{aligned} \mathcal{A}p &= y - y_0 \quad &&\text{in } \Omega, \\ p &= 0 &&\text{on } \Gamma. \end{aligned} \tag{2.5}$$

It is not difficult to see that the optimal control u is given by

$$u = P_{[u_a,u_b]}\left(\frac{1}{\alpha}\partial_{\vec{\nu}_{\mathcal{A}}} p\right) \qquad \text{a.e. on } \Gamma \tag{2.6}$$

where $P_{[u_a,u_b]}$ denotes the pointwise projection onto the interval $[u_a, u_b]$.

Lemma 2.1.1. *Let $u \in U_{ad}$ be the solution of (2.3) with corresponding state y and adjoint state p. Then*

$$u \in C^{0,1}(\Gamma), \quad y \in H^{\frac{3}{2}}(\Omega), \quad p \in W^{3,r}(\Omega) \text{ for some } d < r \le \bar{r}.$$

PROOF. Since $p \in H^2(\Omega)$ we have $\partial_{\vec{\nu}_{\mathcal{A}}} p \in H^{\frac{1}{2}}(\Gamma)$ and hence $u \in H^{\frac{1}{2}}(\Gamma)$ in view of (2.6) (cf. [CR06, p. 1590]) which in turn yields $y \in H^1(\Omega)$. Elliptic regularity implies that $p \in H^3(\Omega)$ and then $\partial_{\vec{\nu}_{\mathcal{A}}} p \in H^{\frac{3}{2}}(\Gamma)$. Therefore $u \in H^1(\Gamma)$ and $y \in H^{\frac{3}{2}}(\Omega)$. Using an embedding theorem, the above regularity of $\partial_{\vec{\nu}_{\mathcal{A}}} p$ also implies that $u \in W^{1-\frac{1}{r},r}(\Gamma)$ for some $r > d$. Hence, $y \in W^{1,r}(\Omega)$ and since $y_0 \in W^{1,\bar{r}}(\Omega)$ we obtain $p \in W^{3,r}(\Omega)$ for some $d < r \le \bar{r}$ again by elliptic regularity. An embedding theorem now yields $p \in C^{1,1}(\bar{\Omega})$ and $\partial_{\vec{\nu}_{\mathcal{A}}} p \in C^{0,1}(\Gamma)$. Since $P_{[u_a,u_b]}$ is LIPSCHITZ we finally deduce that $u \in C^{0,1}(\Gamma)$. □

2.1.2. Finite element discretization. Let $\mathcal{T}_h$ be a triangulation of a polygonal domain Ω_h approximating Ω. We assume that all vertices on $\partial\Omega_h =: \Gamma_h$ also lie on Γ and that at most one face of a simplex $T \in \mathcal{T}_h$ belongs to Γ_h. Furthermore, we suppose that the triangulation is quasi-uniform in the sense of Definition 1.2.4. Recalling equation (1.13) for every $T \in \mathcal{T}_h$ there exists an invertible affine mapping

$$\vec{F}_T : \mathbb{R}^d \to \mathbb{R}^d, \quad \vec{F}_T(\hat{x}) = \mathbf{A}_T\hat{x} + \mathbf{b}_T,$$

which maps the standard d–simplex $\hat{T}$ onto T. Besides the triangulation $\mathcal{T}_h$ which will be used to define the discrete problem and to carry out the practical calculations we also introduce an exact triangulation $\tilde{\mathcal{T}}_h$ of Ω. The existence of such a triangulation for sufficiently small h is shown in [Ber89]. In essence, for every $T \in \mathcal{T}_h$ there is a mapping $\vec{\Phi}_T \in C^3(\hat{T};\mathbb{R}^d)$ such that $\tilde{\vec{F}}_T := \vec{F}_T + \vec{\Phi}_T$ maps $\hat{T}$ onto a curved d–simplex $\tilde{T} \subset \bar{\Omega}$ and $\bar{\Omega} = \bigcup_{\tilde{T}\in\tilde{\mathcal{T}}_h} \tilde{T}$. Furthermore, the mapping $\vec{G}_h$ which is

locally defined by $\vec{G}_{h|T} := \vec{\tilde{F}}_T \circ \vec{F}_T^{-1}$ is a homeomorphism between Ω_h and Ω. The construction in [Ber89] also implies that $\vec{\Phi}_T = 0$ if T has at most one vertex on Γ_h so that $\vec{G}_h \equiv id$ on all simplexes which are disjoint from Γ_h. Furthermore, we have the estimates

$$\begin{aligned} &\sup_{x\in T} \|(D\vec{G}_{h\,|\,T} - I)(x)\| \le Ch, \quad \|\vec{G}_h\|_{W^{3,\infty}(T)} \le C, \quad T \in \mathcal{T}_h \\ &\sup_{\hat{x}\in\hat{T}} \|D\vec{\tilde{F}}_T(\hat{x})\| \le C\|\mathbf{A}_T\|, \quad \sup_{x\in\tilde{T}} \|D\vec{\tilde{F}}_T^{\,-1}(x)\| \le C\|\mathbf{A}_T^{-1}\|, \quad T \in \mathcal{T}_h \\ &c_1|\det \mathbf{A}_T| \le |\det D\vec{\tilde{F}}_T(\hat{x})| \le c_2|\det \mathbf{A}_T|, \quad \hat{x} \in \hat{T} \end{aligned} \tag{2.7}$$

with constants which can be chosen independently of h. Here $\|\cdot\|$ denotes a norm in $\mathbb{R}^d$ (for instance the euclidean norm $|\cdot|$) and the resulting induced matrix norm.

For discretizing the state equation we choose the space of linear finite elements $Y_h := P^1_{c,h}(\mathcal{T}_h)$ defined in (1.18) as well as $Y_{h0} := Y_h \cap H^1_0(\Omega_h)$. Let γY_h be the restriction to Γ_h of functions in Y_h and denote by $P_h : L^2(\Gamma_h) \to \gamma Y_h$ the L^2–projection, i.e. for $v \in L^2(\Gamma_h)$ we have

$$\int_{\Gamma_h} v\chi_h = \int_{\Gamma_h} P_h v\ \chi_h \qquad \forall \chi_h \in \gamma Y_h. \tag{2.8}$$

Let us introduce an approximation to the solution operator $\mathcal{G}$ as follows. For a given function $u_h \in L^2(\Gamma_h)$ we denote by $y_h = \mathcal{G}_h(u_h) \in Y_h$ the unique solution of

$$\begin{aligned} a_h(y_h,\phi_h) = \textstyle\int_{\Omega_h} f_h\phi_h, \quad &\forall\, \phi_h \in Y_{h0}, \\ y_h = P_h(u_h) \qquad &\text{on } \Gamma_h, \end{aligned} \tag{2.9}$$

where

$$a_h(y_h,\phi_h) = \sum_{i,j=1}^d \int_{\Omega_h} \big(a_{h,ij}\, y_{h,x_i}\phi_{h,x_j} + c_{h,0}\, y_h\phi_h\big)$$

and $a_{h,ij} = a_{ij} \circ \vec{G}_h, c_{h,0} = c \circ \vec{G}_h$ and $f_h = f \circ \vec{G}_h$.

In order to deal with the problem that the solutions of (2.1) and (2.9) are defined on different domains we assign to each $\phi_h \in Y_h$ a function $\tilde{\phi}_h : \bar{\Omega} \to \mathbb{R}$ by $\tilde{\phi}_h := \phi_h \circ \vec{G}_h^{-1}$ and let

$$\tilde{Y}_h := \{\tilde{\phi}_h \,:\, \phi_h \in Y_h\} \quad \text{ as well as } \quad \gamma\tilde{Y}_h = \{\tilde{\phi}_{h|\Gamma} \,:\, \tilde{\phi}_h \in \tilde{Y}_h\}.$$

Using the transformation rule, the fact that $\nabla\tilde{y}_h = (\nabla y_h) \circ \vec{G}_h^{-1}\, D\vec{G}_h^{-1}$ and (2.7) we obtain

$$|a(\tilde{y}_h,\tilde{\phi}_h) - a_h(y_h,\phi_h)| \le Ch\|\tilde{y}_h\|_{H^1(A_h)}\|\tilde{\phi}_h\|_{H^1(A_h)} \qquad \forall\, y_h,\phi_h \in Y_h, \tag{2.10}$$

where $A_h = \{x \in \Omega \,:\, \operatorname{dist}(x,\Gamma) < \beta h\}$ and β is chosen so large that $\bigcup_{\tilde{T}\cap\Gamma\neq\varnothing} \tilde{T} \subset A_h$.

Next, by adapting the methods developed in [SZ90], [Ber89, Sect. 4-5], it is possible to show that there exists an interpolation operator $\tilde{\Pi}_h : L^1(\Omega) \to \tilde{Y}_h$ with $\tilde{\Pi}_{h|\tilde{Y}_h} = id_{\tilde{Y}_h}$ such that for $\phi \in W^{l,p}(\Omega)$ ($1 \le l \le 2$ if $p = 1$, $\frac{1}{p} < l \le 2$ otherwise)

$$\|\phi - \tilde{\Pi}_h\phi\|_{W^{m,p}(\Omega)} \le Ch^{l-m}\|\phi\|_{W^{l,p}(\Omega)}, \quad 0 \le m \le \min(1,l). \tag{2.11}$$

In addition it is possible to construct $\tilde{\Pi}_h$ in such a way that $\tilde{\Pi}_h\phi = 0$ on Γ provided that $\phi = 0$ on Γ. If $\phi \in C^0(\bar{\Omega})$ then we can also define the usual Lagrange interpolation operator $\tilde{I}_h : C^0(\bar{\Omega}) \to \tilde{Y}_h$ via $\tilde{I}_h\phi = [I_h(\phi \circ \vec{G}_h)] \circ \vec{G}_h^{-1}$ where I_h is the Lagrange interpolation operator corresponding to Y_h.

Abbreviating $\vec{g}_h := \vec{G}_{h|\Gamma_h}$ we define for $v \in L^2(\Gamma)$ the projection $\tilde{P}_h v := [P_h(v \circ g_h)] \circ g_h^{-1} \in \gamma \tilde{Y}_h$. In view of Lemma 3.1 in [IKP06] we have

$$\int_{\Gamma_h} v = \int_\Gamma v \circ \vec{g}_h^{-1}\, t_h \quad \text{where } t_h = \det D\vec{G}_h^{-1}\, |(D\vec{G}_h)^T \circ \vec{G}_h^{-1}\, \vec{\nu}|. \tag{2.12}$$

Applying (2.12) to (2.8) we see that $\tilde{P}_h$ is characterized by the relation

$$\int_\Gamma v\tilde{\chi}_h\, t_h = \int_\Gamma \tilde{P}_h v\ \tilde{\chi}_h\, t_h \qquad \forall \tilde{\chi}_h \in \gamma \tilde{Y}_h. \tag{2.13}$$

Furthermore one can show that

$$\|v - \tilde{P}_h v\|_{L^2(\Gamma)} \le C h^s \|v\|_{H^s(\Gamma)}, \quad v \in H^s(\Gamma),\ 0 \le s \le 2. \tag{2.14}$$

An important ingredient of our analysis will be an L^2–error estimate for the approximation given by (2.9), in particular for low regularity of the boundary values. A corresponding result in the case of $\mathcal{A} = -\Delta$ and a polygonal domain can be found in [Ber04].

Lemma 2.1.2. *Suppose that $f \in L^2(\Omega), u \in H^s(\Gamma)\,(0 \le s \le 1)$ and that $y \in H^{s+\frac{1}{2}}(\Omega)$, $y_h \in Y_h$ are the solutions of (2.1) and (2.9) with $u_h = u \circ \vec{g}_h$ respectively. Then there exists $h_0 > 0$ such that for $0 < h \le h_0$*

$$\|y - \tilde{y}_h\|_{L^2(\Omega)} \le C h^{s+\frac{1}{2}} \big(\|u\|_{H^s(\Gamma)} + \|f\|_{L^2(\Omega)}\big). \tag{2.15}$$

Proof. In view of the linearity of $\mathcal{A}$ it is sufficient to consider the problems where either $f \equiv 0$ or $u \equiv 0$.
Let us first assume that $f \equiv 0$ and take $s = 1$. We denote by $y^h \in H^{\frac{3}{2}}(\Omega)$ the solution of

$$\begin{aligned} a(y^h, \phi) &= 0 \qquad \forall \phi \in H_0^1(\Omega), \\ y^h &= \tilde{P}_h u \quad \text{on } \Gamma. \end{aligned} \tag{2.16}$$

Clearly,

$$\|y^h\|_{H^{s+\frac{1}{2}}(\Omega)} \le C \|\tilde{P}_h u\|_{H^s(\Gamma)}, \quad 0 \le s \le 1. \tag{2.17}$$

Let us choose $\tilde{\phi}_h = \tilde{\Pi}_h[y^h - \tilde{y}_h] = \tilde{\Pi}_h y^h - \tilde{y}_h$. Note that $\phi_h \in Y_{h0}$ since $y^h = \tilde{y}_h$ on Γ. The ellipticity of $\mathcal{A}$ and the fact that $c \ge 0$ imply together with (2.16) and (2.9)

$$\begin{aligned} c_0 \int_\Omega |\nabla(y^h - \tilde{y}_h)|^2 &\le a(y^h - \tilde{y}_h, y^h - \tilde{y}_h) \\ &= a(y^h - \tilde{y}_h, y^h - \tilde{\Pi}_h y^h) + a(y^h - \tilde{y}_h, \tilde{\Pi}_h y^h - \tilde{y}_h) \\ &= a(y^h - \tilde{y}_h, y^h - \tilde{\Pi}_h y^h) + [a_h(y_h, (\tilde{\Pi}_h y^h) \circ \vec{G}_h - y_h) - a(\tilde{y}_h, \tilde{\Pi}_h y^h - \tilde{y}_h)] \\ &\equiv I + II. \end{aligned} \tag{2.18}$$

Using Poincaré's inequality and (2.11) we infer

$$\begin{aligned} |I| &\le C\|y^h - \tilde{y}_h\|_{H^1(\Omega)} \|y^h - \tilde{\Pi}_h y^h\|_{H^1(\Omega)} \le C h^{\frac{1}{2}} \|\nabla(y^h - \tilde{y}_h)\|_{L^2(\Omega)} \|y^h\|_{H^{\frac{3}{2}}(\Omega)} \\ &\le \frac{c_0}{4} \|\nabla(y^h - \tilde{y}_h)\|^2_{L^2(\Omega)} + Ch \|y^h\|^2_{H^{\frac{3}{2}}(\Omega)}. \end{aligned}$$

In view of (2.10), (2.11), POINCARÉ's and YOUNG's inequality we have

$$\begin{aligned}
|II| &\le Ch\|\tilde{y}_h\|_{H^1(\Omega)}\|\tilde{\Pi}_h y^h - \tilde{y}_h\|_{H^1(\Omega)}\\
&\le Ch\big(\|y^h\|_{H^1(\Omega)} + \|y^h - \tilde{y}_h\|_{H^1(\Omega)}\big)\big(\|y^h - \tilde{\Pi}_h y^h\|_{H^1(\Omega)} + \|y^h - \tilde{y}_h\|_{H^1(\Omega)}\big)\\
&\le Ch\big(\|y^h\|_{H^{\frac{3}{2}}(\Omega)} + \|\nabla(y^h - \tilde{y}_h)\|_{L^2(\Omega)}\big)\big(Ch^{\frac{1}{2}}\|y^h\|_{H^{\frac{3}{2}}(\Omega)} + \|\nabla(y^h - \tilde{y}_h)\|_{L^2(\Omega)}\big)\\
&\le (\frac{c_0}{4} + Ch)\|\nabla(y^h - \tilde{y}_h)\|^2_{L^2(\Omega)} + Ch^{\frac{3}{2}}\|y^h\|^2_{H^{\frac{3}{2}}(\Omega)}.
\end{aligned}$$

Inserting the two estimates into (2.18) and choosing $h_0 > 0$ so small that $Ch_0 \le \frac{c_0}{4}$ we obtain for $0 < h \le h_0$ after another application of POINCARÉ's inequality

$$\|y^h - \tilde{y}_h\|_{H^1(\Omega)} \le C\sqrt{h}\|y^h\|_{H^{\frac{3}{2}}(\Omega)}. \tag{2.19}$$

In order to estimate the L^2–norm of $y - \tilde{y}_h$ we employ the usual duality argument, namely denote by $\psi \in H^2(\Omega)$ the solution of

$$\begin{aligned}
\mathcal{A}\psi &= y - \tilde{y}_h \quad &&\text{in } \Omega,\\
\psi &= 0 &&\text{on } \Gamma.
\end{aligned} \tag{2.20}$$

Then, (2.2) and integration by parts imply that

$$\int_\Omega |y - \tilde{y}_h|^2 = \int_\Omega (y - \tilde{y}_h)\mathcal{A}\psi = -a(\tilde{y}_h, \psi) - \int_\Gamma (u - \tilde{P}_h u)\partial_{\vec{\nu}_\mathcal{A}}\psi \equiv I + II.$$

Observing that $\psi, \tilde{I}_h\psi \in H^1_0(\Omega), I_h(\psi \circ \vec{G}_h) \in Y_{h0}$ we infer from (2.9) and (2.16)

$$\begin{aligned}
I &= a(y^h - \tilde{y}_h, \psi - \tilde{I}_h\psi) + [-a\big(\tilde{y}_h, \widetilde{I_h(\psi \circ \vec{G}_h)}\big) + a_h\big(y_h, I_h(\psi \circ \vec{G}_h)\big)]\\
&\le Ch^{\frac{3}{2}}\|y^h\|_{H^{\frac{3}{2}}(\Omega)}\|\psi\|_{H^2(\Omega)} + Ch\|\tilde{y}_h\|_{H^1(A_h)}\|\tilde{I}_h\psi\|_{H^1(A_h)}
\end{aligned}$$

by (2.19), (2.10) and an interpolation estimate. Next, using the continuous embeddings $H^{\frac{1}{2}}(\Omega) \hookrightarrow L^3(\Omega)$, $H^1(\Omega) \hookrightarrow L^6(\Omega)$ as well as (2.19) we obtain

$$\begin{aligned}
\|\tilde{y}_h\|_{H^1(A_h)} &\le \|y^h\|_{H^1(A_h)} + \|y^h - \tilde{y}_h\|_{H^1(A_h)}\\
&\le C|A_h|^{\frac{1}{6}}\|y^h\|_{W^{1,3}(A_h)} + C\sqrt{h}\|y^h\|_{H^{\frac{3}{2}}(\Omega)} \le Ch^{\frac{1}{6}}\|y^h\|_{H^{\frac{3}{2}}(\Omega)},\\
\|\tilde{I}_h\psi\|_{H^1(A_h)} &\le \|\psi\|_{H^1(A_h)} + \|\psi - \tilde{I}_h\psi\|_{H^1(A_h)}\\
&\le C|A_h|^{\frac{1}{3}}\|\psi\|_{W^{1,6}(A_h)} + Ch\|\psi\|_{H^2(\Omega)} \le Ch^{\frac{1}{3}}\|\psi\|_{H^2(\Omega)}.
\end{aligned}$$

Thus,

$$|I| \le Ch^{\frac{3}{2}}\|y^h\|_{H^{\frac{3}{2}}(\Omega)}\|\psi\|_{H^2(\Omega)} \le Ch^{\frac{3}{2}}\|\tilde{P}_h u\|_{H^1(\Gamma)}\|\psi\|_{H^2(\Omega)} \tag{2.21}$$

in view of (2.17). For II we obtain with the help of (2.13)

$$\begin{aligned}
II &= -\int_\Gamma (u - \tilde{P}_h u)\partial_{\vec{\nu}_\mathcal{A}}\psi\, t_h + \int_\Gamma (u - \tilde{P}_h u)\partial_{\vec{\nu}_\mathcal{A}}\psi(t_h - 1)\\
&= -\int_\Gamma (u - \tilde{P}_h u)(\partial_{\vec{\nu}_\mathcal{A}}\psi - \tilde{P}_h\partial_{\vec{\nu}_\mathcal{A}}\psi)t_h + \int_\Gamma (u - \tilde{P}_h u)\partial_{\vec{\nu}_\mathcal{A}}\psi(t_h - 1)
\end{aligned} \tag{2.22}$$

and hence using (2.14) and (2.7)

$$\begin{aligned}
|II| &\le Ch^{\frac{3}{2}}\|u\|_{H^1(\Gamma)}\|\partial_{\vec{\nu}_\mathcal{A}}\psi\|_{H^{\frac{1}{2}}(\Gamma)} + Ch^2\|u\|_{H^1(\Gamma)}\|\partial_{\vec{\nu}_\mathcal{A}}\psi\|_{L^2(\Gamma)}\\
&\le Ch^{\frac{3}{2}}\|u\|_{H^1(\Gamma)}\|\psi\|_{H^2(\Omega)}.
\end{aligned}$$

Combining this bound with (2.21), the stability of $\tilde{P}_h$ in $H^1(\Gamma)$ and a standard elliptic regularity result we deduce that

$$\|y - \tilde{y}_h\|_{L^2(\Omega)} \leq Ch^{\frac{3}{2}}\|u\|_{H^1(\Gamma)}. \tag{2.23}$$

Let us next look at the case $s = 0$ and define $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$ again via (2.20). As above we obtain

$$\int_\Omega |y - \tilde{y}_h|^2 \equiv I + II.$$

Using (2.21) together with an inverse inequality we have

$$|I| \leq Ch^{\frac{3}{2}}\|\tilde{P}_h u\|_{H^1(\Gamma)}\|\psi\|_{H^2(\Omega)} \leq Ch^{\frac{1}{2}}\|\tilde{P}_h u\|_{L^2(\Gamma)}\|\psi\|_{H^2(\Omega)}.$$

Returning to (2.22) we infer for the second term

$$\begin{aligned} |II| &\leq C\big(\|u\|_{L^2(\Gamma)} + \|\tilde{P}_h u\|_{L^2(\Gamma)}\big)\big(h^{\frac{1}{2}}\|\partial_{\vec{\nu}_\mathcal{A}}\psi\|_{H^{\frac{1}{2}}(\Gamma)} + h\|\partial_{\vec{\nu}_\mathcal{A}}\psi\|_{L^2(\Gamma)}\big) \\ &\leq Ch^{\frac{1}{2}}\|u\|_{L^2(\Gamma)}\|\psi\|_{H^2(\Omega)}. \end{aligned}$$

Combining the above two bounds we deduce that

$$\|y - \tilde{y}_h\|_{L^2(\Omega)} \leq Ch^{\frac{1}{2}}\|u\|_{L^2(\Gamma)}. \tag{2.24}$$

The case $0 < s < 1$ now follows by interpolation: To see this, denote by S the linear operator that maps u to $y - \tilde{y}_h$. The estimates (2.23) and (2.24) then imply that $\|S\|_{H^1(\Gamma)\to L^2(\Omega)} \leq Ch^{\frac{3}{2}}$ and $\|S\|_{L^2(\Gamma)\to L^2(\Omega)} \leq Ch^{\frac{1}{2}}$, so that (2.15) follows from Proposition 14.1.5 and Theorem 14.2.3 in [BS08].
If $u \equiv 0, f \in L^2(\Omega)$ we can proceed in a similar way as above, starting with a bound of the form $\|y - \tilde{y}_h\|_{H^1(\Omega)} \leq Ch\|f\|_{L^2(\Omega)}$ followed by a duality argument to give

$$\|y - \tilde{y}_h\|_{L^2(\Omega)} \leq Ch^{\frac{3}{2}}\|f\|_{L^2(\Omega)}.$$

Since our primary interest lies on the boundary values we leave the details to the reader. □

Our next aim is to bound the discrete solution corresponding to $f \equiv 0$ in terms of $\|u\|_{L^2(\Gamma)}$. In order to formulate the result we introduce the distance function $d_\Gamma(x) := \operatorname{dist}(x, \Gamma)$. It follows from [GT01, Sec. 14.6], that there exists $\delta > 0$ such that $d_\Gamma \in C^3(\Omega_\delta)$, where $\Omega_r := \{x \in \bar{\Omega} : d_\Gamma(x) < r\}$ for $r > 0$. Choose a function $\eta \in C^3(\bar{\Omega})$ such that $0 \leq \eta \leq 1$, $\eta(x) = 1, x \in \Omega_{\frac{\delta}{2}}$ and $\eta(x) = 0, x \in \bar{\Omega} \setminus \Omega_{\frac{2\delta}{3}}$. Then, $\rho(x) := \eta(x)d_\Gamma(x) + (1 - \eta(x))\frac{\delta}{2}$, $x \in \bar{\Omega}$ belongs to $C^3(\bar{\Omega})$ and satisfies

$$\rho(x) = d_\Gamma(x),\ x \in \Omega_{\frac{\delta}{2}}, \quad \rho(x) \geq \frac{\delta}{2},\ x \in \bar{\Omega} \setminus \Omega_{\frac{\delta}{2}}. \tag{2.25}$$

Furthermore, let

$$\omega(x) := \rho(x) + h, \quad x \in \bar{\Omega}.$$

Lemma 2.1.3. *Let $u \in L^2(\Gamma)$ and suppose that $z_h \in Y_h$ is the solution of*

$$\begin{aligned} a_h(z_h, \phi_h) &= 0 && \forall\, \phi_h \in Y_{h0}, \\ z_h &= P_h(u \circ \vec{g}_h) && \text{on } \Gamma_h. \end{aligned} \tag{2.26}$$

Then

$$\int_\Omega \big(|\tilde{z}_h|^2 + \omega|\nabla \tilde{z}_h|^2\big) \leq C\|u\|^2_{L^2(\Gamma)}.$$

PROOF. Let y^h be again the solution of (2.16). Since $(\tilde{P}_h u)\circ\vec{g}_h = P_h(u\circ\vec{g}_h)$ and $P_h^2 = P_h$, Lemma 2.1.2 for $s=0$ implies that

$$\|y^h - \tilde{z}_h\|_{L^2(\Omega)} \le C\sqrt{h}\|\tilde{P}_h u\|_{L^2(\Gamma)} \le C\sqrt{h}\|u\|_{L^2(\Gamma)}. \tag{2.27}$$

Combining this estimate with (2.17) we deduce

$$\|\tilde{z}_h\|_{L^2(\Omega)} \le \|y^h\|_{L^2(\Omega)} + C\sqrt{h}\|u\|_{L^2(\Gamma)} \le C\|u\|_{L^2(\Gamma)}. \tag{2.28}$$

On the other hand, an inverse estimate, (2.11), (2.17) and (2.27) yield

$$\begin{aligned}\|\nabla\tilde{z}_h\|_{L^2(\Omega)} &\le \|\nabla(\tilde{z}_h - \tilde{\Pi}_h y^h)\|_{L^2(\Omega)} + \|\nabla\tilde{\Pi}_h y^h\|_{L^2(\Omega)}\\ &\le Ch^{-1}\|\tilde{z}_h - \tilde{\Pi}_h y^h\|_{L^2(\Omega)} + C\|y^h\|_{H^1(\Omega)}\\ &\le Ch^{-1}\Big(\|\tilde{z}_h - y^h\|_{L^2(\Omega)} + \|y^h - \tilde{\Pi}_h y^h\|_{L^2(\Omega)}\Big) + C\|y^h\|_{H^1(\Omega)}\\ &\le Ch^{-1}\|\tilde{z}_h - y^h\|_{L^2(\Omega)} + C\|y^h\|_{H^1(\Omega)}\\ &\le Ch^{-\frac{1}{2}}\|u\|_{L^2(\Gamma)} + C\|\tilde{P}_h u\|_{H^{\frac{1}{2}}(\Gamma)} \le Ch^{-\frac{1}{2}}\|u\|_{L^2(\Gamma)}.\end{aligned} \tag{2.29}$$

It remains to bound $\displaystyle\int_\Omega \rho\,|\nabla\tilde{z}_h|^2$. The ellipticity of $\mathcal{A}$ and the fact that $c\ge 0$ imply

$$\begin{aligned}c_0\int_\Omega \rho\,|\nabla\tilde{z}_h|^2 &\le \sum_{i,j=1}^d \int_\Omega \rho a_{ij}\tilde{z}_{h,x_i}\tilde{z}_{h,x_j}\\ &\le a(\tilde{z}_h,\rho\tilde{z}_h) - \frac{1}{2}\sum_{i,j=1}^d\int_\Omega a_{ij}\rho_{x_i}(\tilde{z}_h^2)_{x_j} \equiv I + II.\end{aligned}$$

Since $\rho(x) = d_\Gamma(x) = 0, x\in\Gamma$, we have that $\phi_h := I_h\big((\rho\circ\vec{G}_h)z_h\big) \in Y_{h0}$. Hence, (2.26) and (2.10) yield

$$\begin{aligned}I &= a(\tilde{z}_h,\rho\tilde{z}_h - \tilde{I}_h(\rho\tilde{z}_h)) + [a(\tilde{z}_h,\tilde{I}_h(\rho\tilde{z}_h)) - a_h\big(z_h, I_h((\rho\circ\vec{G}_h)z_h)\big)]\\ &\le C\|\tilde{z}_h\|_{H^1(\Omega)}\|\rho\tilde{z}_h - \tilde{I}_h(\rho\tilde{z}_h)\|_{H^1(\Omega)} + Ch\|\tilde{z}_h\|_{H^1(A_h)}\|\tilde{I}_h(\rho\tilde{z}_h)\|_{H^1(A_h)}.\end{aligned} \tag{2.30}$$

For fixed $\tilde{T}\in\tilde{\mathcal{T}}_h$ we have observing (2.7) together with the fact that $z_h\in P_1(T)$

$$\begin{aligned}&\|\rho\tilde{z}_h - \tilde{I}_h(\rho\tilde{z}_h)\|_{H^1(\tilde{T})}\\ &\quad\le C\|(\rho\circ\vec{G}_h)z_h - I_h\big((\rho\circ\vec{G}_h)z_h\big)\|_{H^1(T)} \le Ch\|D^2[(\rho\circ\vec{G}_h)z_h]\|_{L^2(T)}\\ &\quad\le Ch\big(\|z_h D^2(\rho\circ\vec{G}_h)\|_{L^2(T)} + \|\nabla(\rho\circ\vec{G}_h)\otimes\nabla z_h\|_{L^2(T)}\\ &\qquad + \|\nabla z_h\otimes\nabla(\rho\circ\vec{G}_h)\|_{L^2(T)}\big)\\ &\quad\le Ch\|z_h\|_{H^1(T)} \le Ch\|\tilde{z}_h\|_{H^1(\tilde{T})},\end{aligned} \tag{2.31}$$

where $\otimes$ denotes the dyadic product of two vectors. In particular

$$\|\tilde{I}_h(\rho\tilde{z}_h)\|_{H^1(\tilde{T})} \le \|\rho\tilde{z}_h - \tilde{I}_h(\rho\tilde{z}_h)\|_{H^1(\tilde{T})} + \|\rho\tilde{z}_h\|_{H^1(\tilde{T})} \le C\|\tilde{z}_h\|_{H^1(\tilde{T})}. \tag{2.32}$$

Inserting (2.31) and (2.32) into (2.30) we deduce with the help of (2.28) and (2.29)

$$I \le Ch\|\tilde{z}_h\|^2_{H^1(\Omega)} \le C\|u\|^2_{L^2(\Gamma)}. \tag{2.33}$$

Finally, integration by parts and (2.28) imply

$$\begin{aligned} II &= \frac{1}{2}\sum_{i,j=1}^{d}\int_\Omega \big(a_{ij,x_j}\rho_{x_i} + a_{ij}\rho_{x_ix_j}\big)\tilde z_h^2 - \frac{1}{2}\sum_{i,j=1}^{d}\int_\Gamma \partial_{\vec\nu_{\mathcal A}}\rho\,\tilde z_h^2 \\ &\le C\big(\|\tilde z_h\|^2_{L^2(\Omega)} + \|\tilde z_h\|^2_{L^2(\Gamma)}\big) \le C\big(\|u\|^2_{L^2(\Gamma)} + \|\tilde P_h u\|^2_{L^2(\Gamma)}\big) \le C\|u\|^2_{L^2(\Gamma)}. \end{aligned}$$

Combining this estimate with (2.33) completes the proof. □

2.1.3. Error analysis for the control problem. We approximate (2.3) using the variational discretization from [Hin05]. This leads to the following control problem depending on h:

$$\begin{aligned} &\min_{u_h\in U_{h,ad}} J_h(u_h) = \frac{1}{2}\int_{\Omega_h}|y_h - y_{h,0}|^2 + \frac{\alpha}{2}\int_{\Gamma_h}|u_h|^2 \\ &\text{subject to } y_h = \mathcal G_h(u_h), \end{aligned} \tag{2.34}$$

where $U_{h,ad} = \{u_h \in L^2(\Gamma_h) : u_a \le u_h \le u_b \text{ a.e. on } \Gamma_h\}$ and $y_{h,0} = y_0 \circ \vec G_h$. It is not difficult to see that (2.34) has a unique solution $u_h \in U_{h,ad}$ and that this solution is characterized by the variational inequality

$$\int_{\Omega_h}(y_h - y_{h,0})(z_h - y_h) + \alpha\int_{\Gamma_h} u_h(v_h - u_h) \ge 0 \qquad \forall v_h \in U_{h,ad}. \tag{2.35}$$

Here $z_h = \mathcal G_h(v_h) \in Y_h$. It is easy to show that (compare (2.6))

$$u_h = P_{[u_a,u_b]}\left(\frac{1}{\alpha}\partial^h_{\vec\nu_{\mathcal A}} p_h\right),$$

where $p_h \in Y_{h0}$ and $\partial^h_{\vec\nu_{\mathcal A}} p_h \in \gamma Y_h$ are defined by

$$a_h(\phi_h, p_h) = \int_{\Omega_h}(y_h - y_{h,0})\phi_h \qquad \forall\,\phi_h \in Y_{h0}$$

and

$$\int_{\Gamma_h}(\partial^h_{\vec\nu_{\mathcal A}} p_h)w_h = a_h(w_h, p_h) - \int_{\Omega_h}(y_h - y_{h,0})w_h \qquad \forall\, w_h \in Y_h. \tag{2.36}$$

Theorem 2.1.4. *Let u and u_h be the solutions of* (2.3) *and* (2.34) *with corresponding states y and y_h respectively. Then*

$$\|u - \tilde u_h\|_{L^2(\Gamma)} + \|y - \tilde y_h\|_{L^2(\Omega)} \le Ch\sqrt{|\log h|}$$

for all $0 < h \le h_0$. Here, $\tilde u_h = u_h \circ \vec g_h^{-1}$.

Proof. Using $v = \tilde u_h \in U_{ad}$ in (2.4) and $v_h = u \circ \vec g_h \in U_{h,ad}$ in (2.35) we obtain

$$\int_\Omega (y - y_0)(y^h - y) + \alpha\int_\Gamma u(\tilde u_h - u) \ge 0 \tag{2.37a}$$

$$\int_{\Omega_h}(y_h - y_{h,0})(z_h - y_h) + \alpha\int_{\Gamma_h} u_h(u\circ\vec g_h - u_h) \ge 0 \tag{2.37b}$$

where $y^h = \mathcal G(\tilde u_h)$ and $z_h = \mathcal G_h(u\circ\vec g_h)$. Transforming (2.37b) to Ω and Γ respectively we obtain

$$\int_\Omega(\tilde y_h - y_0)(\tilde z_h - \tilde y_h)|\det D\vec G_h^{-1}| + \alpha\int_\Gamma \tilde u_h(u - \tilde u_h)t_h \ge 0$$

or equivalently with $\theta_h := \int_\Omega (\tilde{y}_h - y_0)(\tilde{z}_h - \tilde{y}_h)(|\det D\vec{G}_h^{-1}| - 1) + \alpha \int_\Gamma \tilde{u}_h(u - \tilde{u}_h)(t_h - 1)$

$$\int_\Omega (\tilde{y}_h - y_0)(\tilde{z}_h - \tilde{y}_h) + \alpha \int_\Gamma \tilde{u}_h(u - \tilde{u}_h) + \theta_h \geq 0 \tag{2.38}$$

where, using (2.7) together with the fact that $\|\tilde{y}_h\|_{L^2(\Omega)}, \|\tilde{u}_h\|_{L^2(\Gamma)} \leq C$,

$$\begin{aligned} |\theta_h| &\leq Ch\big(\|\tilde{z}_h - \tilde{y}_h\|_{L^2(\Omega)} + \|u - \tilde{u}_h\|_{L^2(\Gamma)}\big) \\ &\leq Ch\big(\|y - \tilde{y}_h\|_{L^2(\Omega)} + \|y - \tilde{z}_h\|_{L^2(\Omega)} + \|u - \tilde{u}_h\|_{L^2(\Gamma)}\big) \\ &\leq \varepsilon\big(\|y - \tilde{y}_h\|^2_{L^2(\Omega)} + \|u - \tilde{u}_h\|^2_{L^2(\Gamma)}\big) + C_\varepsilon h^2 + C\|y - \tilde{z}_h\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.39}$$

Combining (2.37a), (2.38) and (2.39) we deduce

$$\begin{aligned} \alpha\|u - \tilde{u}_h\|^2_{L^2(\Gamma)} &\leq \int_\Omega (y - y_0)(y^h - y) + \int_\Omega (\tilde{y}_h - y_0)(\tilde{z}_h - \tilde{y}_h) + \theta_h \\ &= -\int_\Omega (y - \tilde{y}_h)^2 + \int_\Omega (y - \tilde{y}_h)(y - \tilde{z}_h) - \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big) + \theta_h \\ &\leq -\frac{1}{2}\|y - \tilde{y}_h\|^2_{L^2(\Omega)} - \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big) \\ &\quad + \varepsilon\big(\|y - \tilde{y}_h\|^2_{L^2(\Omega)} + \|u - \tilde{u}_h\|^2_{L^2(\Gamma)}\big) + C_\varepsilon h^2 + C_\varepsilon\|y - \tilde{z}_h\|^2_{L^2(\Omega)} \end{aligned}$$

and hence after choosing $\varepsilon > 0$ small enough and recalling Lemma 2.1.2

$$\frac{\alpha}{2}\|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + \frac{1}{4}\|y - \tilde{y}_h\|^2_{L^2(\Omega)} \leq Ch^2 - \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big). \tag{2.40}$$

Using (2.5), (2.2), integration by parts, the definition of $\tilde{P}_h$ and the fact that $a_h(z_h - y_h, \phi_h) = 0$ for $\phi_h \in Y_{h0}$ we obtain

$$\begin{aligned} \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big) &= \int_\Omega (y - y^h)\mathcal{A}p - \int_\Omega (\tilde{z}_h - \tilde{y}_h)\mathcal{A}p \\ &= -\int_\Gamma (u - \tilde{u}_h)\partial_{\vec{\nu}_\mathcal{A}} p - a(p, \tilde{z}_h - \tilde{y}_h) + \int_\Gamma \tilde{P}_h(u - \tilde{u}_h)\partial_{\vec{\nu}_\mathcal{A}} p \\ &= -a(p - \tilde{I}_h p, \tilde{z}_h - \tilde{y}_h) - \int_\Gamma \big((u - \tilde{u}_h) - \tilde{P}_h(u - \tilde{u}_h)\big)\partial_{\vec{\nu}_\mathcal{A}} p \\ &\quad + [a_h(I_h(p \circ \vec{G}_h), z_h - y_h) - a(\tilde{I}_h p, \tilde{z}_h - \tilde{y}_h)] \\ &\equiv I + II + III. \end{aligned} \tag{2.41}$$

The first integral is estimated with the help of an interpolation inequality and Lemma 2.1.3:

$$\begin{aligned} |I| &\leq \big(\int_\Omega \omega^{-1}|\nabla(p - \tilde{I}_h p)|^2\big)^{\frac{1}{2}} \big(\int_\Omega \omega|\nabla(\tilde{z}_h - \tilde{y}_h)|^2\big)^{\frac{1}{2}} \\ &\leq Ch\|p\|_{W^{2,\infty}(\Omega)}\big(\int_\Omega \omega^{-1}\big)^{\frac{1}{2}} \|u - \tilde{u}_h\|_{L^2(\Gamma)} \\ &\leq Ch\|p\|_{W^{3,r}(\Omega)}\big(\int_\Omega \omega^{-1}\big)^{\frac{1}{2}} \|u - \tilde{u}_h\|_{L^2(\Gamma)}. \end{aligned}$$

In view of (2.25) and the coarea formula we have

$$\int_\Omega \omega^{-1} \leq \int_{\Omega_{\frac{\delta}{2}}} \frac{1}{d_\Gamma + h} + \int_{\Omega \setminus \Omega_{\frac{\delta}{2}}} \frac{2}{\delta} \leq C\int_0^{\frac{\delta}{2}} \int_{\{d_\Gamma = \tau\}} \frac{1}{\tau + h} dA d\tau + C \leq C|\log h|$$

where dA denotes the area element. Hence,

$$|I| \leq \varepsilon \|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + C_\varepsilon h^2 |\log h|. \tag{2.42}$$

Next, $II = II_1 + II_2$ where

$$II_1 = -\int_\Gamma \big((u - \tilde{u}_h) - \tilde{P}_h(u - \tilde{u}_h)\big)\partial_{\vec{\nu}_\mathcal{A}} p \, t_h$$
$$II_2 = \int_\Gamma \big((u - \tilde{u}_h) - \tilde{P}_h(u - \tilde{u}_h)\big)\partial_{\vec{\nu}_\mathcal{A}} p (t_h - 1).$$

We infer from (2.13) and (2.14) that

$$\begin{aligned} |II_1| &= \Big| -\int_\Gamma (u - \tilde{u}_h)\big(\partial_{\vec{\nu}_\mathcal{A}} p - \tilde{P}_h \partial_{\vec{\nu}_\mathcal{A}} p\big) t_h \Big| \\ &\leq C h^{\frac{3}{2}} \|\partial_{\vec{\nu}_\mathcal{A}} p\|_{H^{\frac{3}{2}}(\Gamma)} \|u - \tilde{u}_h\|_{L^2(\Gamma)} \leq \varepsilon \|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + C_\varepsilon h^3. \end{aligned}$$

On the other hand, (2.7) implies

$$|II_2| \leq Ch \|u - \tilde{u}_h\|_{L^2(\Gamma)} \|\partial_{\vec{\nu}_\mathcal{A}} p\|_{L^2(\Gamma)} \leq \varepsilon \|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + C_\varepsilon h^2$$

so that in conclusion

$$|II| \leq \varepsilon \|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + C_\varepsilon h^2. \tag{2.43}$$

Finally, recalling (2.10) we have

$$\begin{aligned} |III| &\leq Ch \|\tilde{I}_h p\|_{H^1(A_h)} \|\tilde{z}_h - \tilde{y}_h\|_{H^1(\Omega)} \leq Ch |A_h|^{\frac{1}{2}} \|p\|_{W^{1,\infty}(A_h)} \|\tilde{z}_h - \tilde{y}_h\|_{H^1(\Omega)} \\ &\leq Ch \|p\|_{W^{1,\infty}(\Omega)} \|u - \tilde{u}_h\|_{L^2(\Gamma)} \leq \varepsilon \|u - \tilde{u}\|^2_{L^2(\Gamma)} + C_\varepsilon h^2 \end{aligned} \tag{2.44}$$

in view of Lemma 2.1.3. Inserting (2.42), (2.43) and (2.44) into (2.40) and choosing ε small enough yields the result. □

2.1.4. Superconvergence. In the following subsection we demonstrate that it is possible to improve the order of convergence under additional conditions on the underlying mesh.

We assume from now on that $d = 2$. We are going to make use of the theory developed in [BX03], where the following definition can be found:

Definition 2.1.5. The triangulation $\mathcal{T}_h$ is called $\mathcal{O}(h^{2\sigma})$ *irregular* if the following holds:

a) The set of interior edges of $\mathcal{T}_h$ can be decomposed into two disjoint sets $\mathcal{E}_1$ and $\mathcal{E}_2$ with the following properties:
   - For each $e \in \mathcal{E}_1$ let $T, T' \in \mathcal{T}_h$ with $T \cap T' = e$. Then in the quadrilateral formed by $T \cup T'$ the lengths of any two opposite edges only differ by $\mathcal{O}(h^2)$.
   - $\sum_{e \in \mathcal{E}_2} (|T| + |T'|) = \mathcal{O}(h^{2\sigma})$.

b) The set of boundary vertices $\mathcal{P}$ can be decomposed into two disjoint sets $\mathcal{P}_1$ and $\mathcal{P}_2$ with the following properties:
   - For each vertex $x \in \mathcal{P}_1$ denote by $e \subset T, e' \subset T'$ the two boundary edges sharing x and let $\mathbf{t}$, $\mathbf{t}'$ be the unit tangents. Also denote by e, f, g and e', f', g' the edges obtained by making a clockwise traversal

of $\partial T, \partial T'$ respectively. Then

$$\begin{aligned}|\mathbf{t}-\mathbf{t}'| &= \mathcal{O}(h),\\ |e|-|e'| &= \mathcal{O}(h^2),\\ |f|-|f'| &= \mathcal{O}(h^2),\\ |g|-|g'| &= \mathcal{O}(h^2).\end{aligned}$$

- $\operatorname{card}(\mathcal{P}_2) \le C$ where C is independent of h.

The following result is essentially proved in [BX03, Lem. 2.5] for functions f belonging to $W^{3,\infty}(\Omega)$. Since we would like to use a corresponding estimate for the solution of the adjoint problem which only belongs to $W^{3,r}(\Omega)$ for some $r > 2$ we require a suitable modification allowing a boundary term of the discrete test function ϕ_h.

Lemma 2.1.6. *Suppose that the triangulation $\mathcal{T}_h$ is $\mathcal{O}(h^{2\sigma})$ irregular and let $f \in W^{3,r}(\Omega_h)$ for some $r > 2$. Then*

$$\begin{aligned}&\left|\int_{\Omega_h} \nabla(f - I_h f)\cdot\nabla\phi_h\right| \\ &\qquad \le C\|f\|_{W^{3,r}(\Omega_h)}\big(h^{1+\min(1,\sigma)}\|\phi_h\|_{H^1(\Omega_h)} + h^{\frac{3}{2}}\|\phi_h\|_{L^2(\Gamma_h)}\big) \quad \forall \phi_h \in Y_h.\end{aligned}$$

Proof. Lemma 2.3 in [BX03] gives

$$\begin{aligned}\int_{\Omega_h} \nabla(f - I_h f)\cdot\nabla\phi_h &= \sum_{T\in\mathcal{T}_h}\int_T \nabla(f - I_h f)\cdot\nabla\phi_h \\ &= \underbrace{\sum_{T\in\mathcal{T}_h}\sum_{e\subset\partial T}\int_e q_e\big(\alpha_e\frac{\partial^2 f}{\partial\mathbf{t}^2} + \beta_e\frac{\partial^2 f}{\partial\mathbf{t}\partial\mathbf{n}}\big)\frac{\partial\phi_h}{\partial t}}_{=:I_1} - \underbrace{\sum_{T\in\mathcal{T}_h}\int_T \sum_{|\lambda|=3,|\mu|=1}\gamma_{T,\lambda\mu}\,\partial^\lambda f\,\partial^\mu\phi_h}_{=:-I_2}.\end{aligned} \tag{2.45}$$

Here, q_e is the quadratic function vanishing at the endpoints of e and being equal to $\frac{1}{4}$ at the midpoint. Furthermore, $\mathbf{n}$ is the unit normal to e pointing away from T while $\mathbf{t}$ denotes the unit tangent with the tangents on ∂T being oriented counterclockwise. The numbers α_e, β_e and functions $\gamma_{T,\lambda\mu}$ depend on the geometry of T and their precise form can be found in [BX03]. For our purposes it is sufficient to note that the conditions in Definition 2.1.5 imply

$$|\alpha_e|,\ |\beta_e|,\ |\gamma_{T,\lambda\mu}| \le Ch^2, \quad e \in \mathcal{E}_1 \cup \mathcal{E}_2, \tag{2.46a}$$

$$|\alpha_e - \alpha_{e'}|,\ |\beta_e - \beta_{e'}| \le Ch^3, \quad T \cap T' = e \in \mathcal{E}_1, \tag{2.46b}$$

$$|\alpha_e - \alpha_{e'}|,\ |\beta_e - \beta_{e'}| \le Ch^3, \quad e, e' \subset \Gamma_h, e \cap e' = \{x\}, x \in \mathcal{P}_1. \tag{2.46c}$$

In view of (2.46a) we have

$$|I_2| \le Ch^2\|f\|_{H^3(\Omega_h)}\|\phi_h\|_{H^1(\Omega_h)}. \tag{2.47}$$

Next, we write as in [BX03]

$$I_1 = I_{11} + I_{12} + I_{13},$$

where

$$I_{1j} = \sum_{e\in\mathcal{E}_j}\int_e q_e\big\{(\alpha_e-\alpha_{e'})\frac{\partial^2 f}{\partial \mathbf{t}^2}+(\beta_e-\beta_{e'})\frac{\partial^2 f}{\partial \mathbf{t}\partial\mathbf{n}}\big\}\frac{\partial\phi_h}{\partial\mathbf{t}},\quad j=1,2,$$

$$I_{13} = \sum_{e\subset\Gamma_h}\int_e q_e\big\{\alpha_e\frac{\partial^2 f}{\partial \mathbf{t}^2}+\beta_e\frac{\partial^2 f}{\partial \mathbf{t}\partial\mathbf{n}}\big\}\frac{\partial\phi_h}{\partial\mathbf{t}}.$$

Arguing as in [BX03] we have

$$|I_{11}|+|I_{12}|\le C(h^2+h^{1+\sigma})\|f\|_{W^{2,\infty}(\Omega_h)}\|\phi_h\|_{H^1(\Omega_h)}. \tag{2.48}$$

In order to treat I_{13} we proceed in a slightly different manner compared to [BX03]. Let us set

$$B_e(f) := \alpha_e\frac{\partial^2 f}{\partial \mathbf{t}^2}+\beta_e\frac{\partial^2 f}{\partial \mathbf{t}\partial\mathbf{n}},\ e\subset\Gamma_h \quad \text{as well as } \bar B_e(f) := \frac{1}{|e|}\int_e B_e(f).$$

Then we can write

$$I_{13} = \sum_{e\subset\Gamma_h}\int_e q_e B_e(f)\frac{\partial\phi_h}{\partial\mathbf{t}} = \sum_{e\subset\Gamma_h}\int_e q_e\big(B_e(f)-\bar B_e(f)\big)\frac{\partial\phi_h}{\partial\mathbf{t}}+\sum_{e\subset\Gamma_h}\int_e q_e\bar B_e(f)\frac{\partial\phi_h}{\partial\mathbf{t}}.$$

A Poincaré type inequality along with a scaling argument yields for $g\in W^{1,\tilde q}(T)$

$$\|g-\frac{1}{|e|}\int_e g\|_{L^q(e)}\le Ch^{1+\frac1q-\frac{2}{\tilde q}}\|\nabla g\|_{L^{\tilde q}(T)},\quad e\subset\partial T, 1+\frac1q-\frac{2}{\tilde q}>0. \tag{2.49}$$

Applying this estimate with $q=\tilde q=2$ and using (2.46a) as well as an inverse inequality we deduce

$$\begin{aligned}&|\sum_{e\subset\Gamma_h}\int_e q_e\big(B_e(f)-\bar B_e(f)\big)\frac{\partial\phi_h}{\partial\mathbf{t}}|\\ &\quad\le C\sum_{e\subset\Gamma_h}\|B_e(f)-\bar B_e(f)\|_{L^2(e)}\|\nabla\phi_h\|_{L^2(e)}\le Ch^2\|f\|_{H^3(\Omega_h)}\|\nabla\phi_h\|_{L^2(\Omega_h)}.\end{aligned}$$

For the second term we write as in [BX03]

$$\begin{aligned}\sum_{e\subset\Gamma_h}\int_e q_e\bar B_e(f)\frac{\partial\phi_h}{\partial\mathbf{t}} &= \sum_{e\subset\Gamma_h}\bar B_e(f)\frac{\partial\phi_h}{\partial\mathbf{t}}\int_e q_e = \sum_{e\subset\Gamma_h}\bar B_e(f)\frac{\partial\phi_h}{\partial\mathbf{t}}\frac{|e|}{6}\\ &= \frac16\sum_{x\in\mathcal{P}_1}\big(\bar B_e(f)-\bar B_{e'}(f)\big)\phi_h(x)+\frac16\sum_{x\in\mathcal{P}_2}\big(\bar B_e(f)-\bar B_{e'}(f)\big)\phi_h(x),\end{aligned}$$

where e and e' are the edges sharing x. Using (2.46c) as well as $|\mathbf{t}-\mathbf{t}'|=\mathcal{O}(h)$ for $e\cap e'=\{x\}$ we have for $x\in\mathcal{P}_1$

$$\begin{aligned}&|\bar B_e(f)-\bar B_{e'}(f)|\\ &\quad\le|\bar B_e(f)-B_e(f)(x)|+|\bar B_{e'}(f)-B_{e'}(f)(x)|+|B_e(f)(x)-B_{e'}(f)(x)|\\ &\quad\le C\big(\|B_e(f)-\bar B_e(f)\|_{L^\infty(e)}+\|B_{e'}(f)-\bar B_{e'}(f)\|_{L^\infty(e')}\big)+Ch^3|D^2f(x)|\\ &\quad\le Ch^{3-\frac2r}\|f\|_{W^{3,r}(T\cup T')}\end{aligned}$$

by (2.49) with $q=\infty,\tilde q=r$. On the other hand we have for $x\in e\subset T$

$$|\phi_h(x)|\le\|\phi_h\|_{L^\infty(e)}\le Ch^{-\frac12}\|\phi_h\|_{L^2(e)}+Ch^{1-\frac{2}{r'}}\|\nabla\phi_h\|_{L^{r'}(T)}.$$

Thus,

$$
\begin{aligned}
&|\sum_{x\in\mathcal{P}_1}\big(\bar{B}_e(f)-\bar{B}_{e'}(f)\big)\phi_h(x)|\\
&\quad\le Ch^{3-\frac{2}{r}}\sum_{x\in\mathcal{P}_1}\|f\|_{W^{3,r}(T\cup T')}\big(h^{-\frac{1}{2}}\|\phi_h\|_{L^2(e)}+h^{1-\frac{2}{r'}}\|\nabla\phi_h\|_{L^{r'}(T)}\big)\\
&\quad\le Ch^{\frac{5}{2}-\frac{2}{r}}\Big(\sum_{T\in\mathcal{T}}\|f\|^r_{W^{3,r}(T)}\Big)^{\frac{1}{r}}\Big(\sum_{e\subset\Gamma_h}\|\phi_h\|^2_{L^2(e)}\Big)^{\frac{1}{2}}\Big(\sum_{x\in\mathcal{P}}1\Big)^{\frac{1}{2}-\frac{1}{r}}\\
&\qquad+Ch^{4-\frac{2}{r}-\frac{2}{r'}}\Big(\sum_{T\in\mathcal{T}_h}\|f\|^r_{W^{3,r}(T)}\Big)^{\frac{1}{r}}\Big(\sum_{T\in\mathcal{T}_h}\|\nabla\phi_h\|^{r'}_{L^{r'}(T)}\Big)^{\frac{1}{r'}}\\
&\quad\le Ch^{2-\frac{1}{r}}\|f\|_{W^{3,r}(\Omega_h)}\|\phi_h\|_{L^2(\Gamma_h)}+Ch^2\|f\|_{W^{3,r}(\Omega_h)}\|\nabla\phi_h\|_{L^2(\Omega_h)},
\end{aligned}
$$

since $\sum_{x\in\mathcal{P}}1\le Ch^{-1}$ and $r'<2$. Furthermore, recalling that $|\mathcal{P}_2|\le C$,

$$
\begin{aligned}
|\sum_{x\in\mathcal{P}_2}\big(\bar{B}_e(f)-\bar{B}_{e'}(f)\big)\phi_h(x)|&\le Ch^2\|D^2f\|_{L^\infty(\Gamma_h)}\|\phi_h\|_{L^\infty(\Gamma_h)}\\
&\le Ch^{\frac{3}{2}}\|f\|_{W^{3,r}(\Omega_h)}\|\phi_h\|_{L^2(\Gamma_h)}.
\end{aligned}
$$

In conclusion,

$$
|I_{13}|\le Ch^{\frac{3}{2}}\|f\|_{W^{3,r}(\Omega_h)}\|\phi_h\|_{L^2(\Gamma_h)}+Ch^2\|f\|_{W^{3,r}(\Omega_h)}\|\nabla\phi_h\|_{L^2(\Omega_h)}. \tag{2.50}
$$

Combining (2.50) with (2.48) and (2.47) we finish the proof of the lemma. □

Remark 2.1.7. Lemma 2.1.6 continues to hold if the triangulation $\mathcal{T}_h$ is *piecewise* $\mathcal{O}(h^{2\sigma})$ *irregular*, that is, if Ω_h can be written as the union of a bounded number of polygonal subdomains each of which is $\mathcal{O}(h^{2\sigma})$ irregular (cf. [BX03, Thm. 4.4]).

In order to simplify the subsequent analysis we assume from now on that $\Omega\subset\mathbb{R}^2$ is convex and that $\mathcal{A}=-\Delta$. As a consequence, $\Omega_h\subset\Omega$ and $y_h=\mathcal{G}_h(u_h)$ is defined by

$$
\begin{aligned}
\textstyle\int_{\Omega_h}\nabla y_h\cdot\nabla\phi_h=\int_{\Omega_h}f\phi_h,&\quad\forall\,\phi_h\in Y_{h0},\\
y_h=P_h(u_h)&\quad\text{on }\Gamma_h,
\end{aligned} \tag{2.51}
$$

where P_h is again given by (2.8). We extend a function $\phi_h\in Y_h$ to $\bar{\Omega}$ as follows: if Ω_e is the subset of $\Omega\setminus\Omega_h$ bounded by the boundary edge $e\subset T\cap\Gamma_h$ and the curved segment $\tilde{e}\subset\Gamma$, then $\tilde{\phi}_{h|\Omega_e}$ is given by the linear extension of ϕ_h from T. Furthermore, let $\vec{g}_h:\Gamma_h\to\Gamma$ be defined by

$$
\vec{g}_h(x):=x+\delta_h(x)\vec{\nu}_h(x),\quad x\in e\subset\Gamma_h,
$$

where $\vec{\nu}_h$ is the constant normal to Γ_h on e and $\delta_h(x)$ is chosen in such a way that $\vec{g}_h(x)\in\Gamma$. Note that in general the function $\vec{g}_h$ will be different from the one introduced in Subsection 2.1.2. Clearly, $\vec{g}_h$ is bijective for small h. Given $u\in H^s(\Gamma), 0\le s\le 1$, it follows from [BK94, Thm. 1] that

$$
\|y-\tilde{y}_h\|_{L^2(\Omega)}\le C\big(h^2\|f\|_{L^2(\Omega)}+h^{s+\frac{1}{2}}\|u\|_{H^s(\Gamma)}\big), \tag{2.52}
$$

where $y=\mathcal{G}(u)$ and $y_h=\mathcal{G}_h(u\circ\vec{g}_h)$. We are now in position to state the main result of this subsection.

Theorem 2.1.8. *Suppose that the triangulation $\mathcal{T}_h$ is piecewise $\mathcal{O}(h^2)$ irregular. Let u and u_h be the solutions of* (2.3) *and* (2.34) *(with $y_{h,0}=y_{0|\Omega_h}$). Then*

$$
\|u-\tilde{u}_h\|_{L^2(\Gamma)}+\|y-\tilde{y}_h\|_{L^2(\Omega)}\le Ch^{\frac{3}{2}}
$$

for all $0 < h \le h_0$. *Here,* $\tilde{u}_h = u_h \circ \vec{g}_h^{-1}$ *and* y, y_h *are the corresponding states respectively.*

PROOF. As in the proof of Theorem 2.1.4 let $y^h = \mathcal{G}(\tilde{u}_h), z_h = \mathcal{G}_h(u \circ \vec{g}_h)$. We again have

$$\int_\Omega (\tilde{y}_h - y_0)(\tilde{z}_h - \tilde{y}_h) + \alpha \int_\Gamma \tilde{u}_h(u - \tilde{u}_h) + \theta_h \ge 0 \tag{2.53}$$

where now

$$\theta_h = -\int_{\Omega\setminus\Omega_h} (\tilde{y}_h - y_0)(\tilde{z}_h - \tilde{y}_h) + \alpha \int_\Gamma \tilde{u}_h(u - \tilde{u}_h)(t_h - 1).$$

Since $|t_h - 1| \le Ch^2$ in our setting we obtain

$$|\theta_h| \le \big(\|y_0\|_{L^2(\Omega\setminus\Omega_h)} + \|\tilde{y}_h\|_{L^2(\Omega\setminus\Omega_h)}\big)\|\tilde{z}_h - \tilde{y}_h\|_{L^2(\Omega\setminus\Omega_h)} + Ch^2\|u - \tilde{u}_h\|_{L^2(\Gamma)}. \tag{2.54}$$

Using Lemma 2 in [BK94] we infer that

$$\|y_0\|_{L^2(\Omega\setminus\Omega_h)} \le C\big(h\|y_0\|_{L^2(\Gamma)} + h^2\|y_0\|_{H^1(\Omega)}\big) \le Ch.$$

On the other hand it follows from (2.10) in [BK94] that for $\phi_h \in Y_h$

$$\|\tilde{\phi}_h\|_{L^2(\Omega\setminus\Omega_h)} \le C\big(h\|\phi_h\|_{L^2(\Gamma_h)} + h^2\|\tilde{\phi}_h\|_{H^1(\Omega)}\big) \le C\big(h\|\phi_h\|_{L^2(\Gamma_h)} + h^2\|\phi_h\|_{H^1(\Omega_h)}\big). \tag{2.55}$$

Combining the bounds

$$\begin{aligned} \|y_h\|_{H^1(\Omega_h)} &\le C\big(h^{-\frac{1}{2}}\|u_h\|_{L^2(\Gamma_h)} + \|f\|_{L^2(\Omega_h)}\big) \le Ch^{-\frac{1}{2}}, \\ \|z_h - y_h\|_{H^1(\Omega_h)} &\le Ch^{-\frac{1}{2}}\|u \circ \vec{g}_h - u_h\|_{L^2(\Gamma_h)} \end{aligned}$$

with (2.55) we deduce from (2.54)

$$|\theta_h| \le Ch^2\big(\|u \circ \vec{g}_h - u_h\|_{L^2(\Gamma_h)} + \|u - \tilde{u}_h\|_{L^2(\Gamma)}\big) \le Ch^2\|u - \tilde{u}_h\|_{L^2(\Gamma)}. \tag{2.56}$$

Thus, we deduce from (2.37a), (2.53) and (2.56) similarly as in the proof of Theorem 2.1.4

$$\begin{aligned} \alpha\|u - \tilde{u}_h\|^2_{L^2(\Gamma)} \le -\frac{1}{2}\|y - \tilde{y}_h\|^2_{L^2(\Omega)} - \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big) \\ +\varepsilon\big(\|y - \tilde{y}_h\|^2_{L^2(\Omega)} + \|u - \tilde{u}_h\|^2_{L^2(\Gamma)}\big) + C_\varepsilon h^4 + C\|y - \tilde{z}_h\|^2_{L^2(\Omega)} \end{aligned}$$

and hence after choosing ε sufficiently small and applying (2.52) with $s = 1$

$$\frac{\alpha}{2}\|u - \tilde{u}_h\|^2_{L^2(\Gamma)} + \frac{1}{4}\|y - \tilde{y}_h\|^2_{L^2(\Omega)} \le Ch^3 - \int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big). \tag{2.57}$$

Using (2.5) for our case $\mathcal{A} = -\Delta$ as well as integration by parts we have

$$\begin{aligned} &\int_\Omega (y - y_0)\big((y - y^h) - (\tilde{z}_h - \tilde{y}_h)\big) \\ &\quad = -\int_\Omega (y - y^h)\Delta p + \int_{\Omega_h} (z_h - y_h)\Delta p + \int_{\Omega\setminus\Omega_h} (\tilde{z}_h - \tilde{y}_h)\Delta p \\ &\quad = -\int_\Gamma (u - \tilde{u}_h)\partial_{\vec{\nu}} p - \int_{\Omega_h} \nabla(z_h - y_h)\cdot\nabla p + \int_{\Gamma_h} P_h\big((u \circ \vec{g}_h) - u_h\big)\partial_{\vec{\nu}_h} p \\ &\qquad + \int_{\Omega\setminus\Omega_h} (\tilde{z}_h - \tilde{y}_h)\Delta p \equiv S_1 + S_2 + S_3 + S_4. \end{aligned} \tag{2.58}$$

Taking into account (2.51) and the fact that $I_h p \in Y_{h0}$ we infer with the help of Lemma 2.1.6

$$\begin{aligned}|S_2| &= |\int_{\Omega_h} \nabla(z_h - y_h)\cdot\nabla(p - I_h p)| \\ &\le C\|p\|_{W^{3,r}(\Omega_h)}\big(h^2\|z_h - y_h\|_{H^1(\Omega_h)} + h^{\frac{3}{2}}\|z_h - y_h\|_{L^2(\Gamma_h)}\big) \\ &\le Ch^{\frac{3}{2}}\|u\circ\vec{g}_h - u_h\|_{L^2(\Gamma_h)} \le Ch^{\frac{3}{2}}\|u - \tilde{u}_h\|_{L^2(\Gamma)}.\end{aligned}$$

Since $p \in H^3(\Omega)$ we deduce similarly as above that

$$|S_4| \le Ch^2\|u - \tilde{u}_h\|_{L^2(\Gamma)}.$$

Next, recalling the relation $[\tilde{P}_h v]\circ\vec{g}_h = P_h(v\circ\vec{g}_h)$ as well as (2.12) we have

$$\begin{aligned}S_3 &= \int_\Gamma \tilde{P}_h(u - \tilde{u}_h)[\nabla p\cdot\vec{\nu}_h]\circ\vec{g}_h^{-1}\, t_h \\ &= \int_\Gamma \tilde{P}_h(u - \tilde{u}_h)\partial_{\vec{\nu}} p\, t_h + \int_\Gamma \tilde{P}_h(u - \tilde{u}_h)\big([\nabla p\cdot\vec{\nu}_h]\circ g_h^{-1} - \nabla p\cdot\vec{\nu}\big)t_h.\end{aligned}$$

In order to deal with the second term we let $\tilde{x} = \vec{g}_h(x) \in \Gamma$. Since $p = 0$ on Γ we have that $\nabla p = \partial_{\vec{\nu}} p\,\vec{\nu}$ on Γ. Hence

$$\begin{aligned}[\nabla p\cdot\vec{\nu}_h](\vec{g}_h^{-1}(\tilde{x})) - (\nabla p\cdot\vec{\nu})(\tilde{x}) &= \nabla p(x)\cdot\vec{\nu}_h(x) - \nabla p(\vec{g}_h(x))\cdot\vec{\nu}(\vec{g}_h(x)) \\ &= \big(\nabla p(x) - \nabla p(\vec{g}_h(x))\big)\cdot\vec{\nu}_h(x) + \partial_{\vec{\nu}} p(\vec{g}_h(x))\,\vec{\nu}(\vec{g}_h(x))\cdot(\vec{\nu}_h(x) - \vec{\nu}(\vec{g}_h(x)) \\ &= \big(\nabla p(x) - \nabla p(\vec{g}_h(x))\big)\cdot\vec{\nu}_h(x) - \frac{1}{2}\partial_{\vec{\nu}} p(\vec{g}_h(x))|\vec{\nu}(\vec{g}_h(x)) - \vec{\nu}_h(x)|^2.\end{aligned}$$

As a consequence,

$$|[\nabla p\cdot\vec{\nu}_h]\circ\vec{g}_h^{-1} - \nabla p\cdot\vec{\nu}| \le Ch^2 \qquad \text{on } \Gamma$$

since $|\vec{g}_h(x) - x| \le Ch^2$, $|\vec{\nu}(\vec{g}_h(x)) - \vec{\nu}_h(x)| \le Ch$, which follows, roughly speaking, from the fact that a boundary edge can be seen as a linear approximation to the corresponding part of Γ. Finally, we may write

$$S_1 + S_3 = -\int_\Gamma\big((u - \tilde{u}_h) - \tilde{P}_h(u - \tilde{u}_h)\big)\partial_{\vec{\nu}} p t_h + r_h = -\int_\Gamma (u - \tilde{u}_h)\big(\partial_{\vec{\nu}} p - \tilde{P}_h\partial_{\vec{\nu}} p\big)\, t_h + r_h$$

where $|r_h| \le Ch^2\|u - \tilde{u}_h\|_{L^2(\Gamma)}$. Now, (2.14) implies that

$$|S_1 + S_3| \le Ch^{\frac{3}{2}}\|\partial_{\vec{\nu}} p\|_{H^{\frac{3}{2}}(\Gamma)}\|u - \tilde{u}_h\|_{L^2(\Gamma)} + |r_h| \le Ch^{\frac{3}{2}}\|u - \tilde{u}_h\|_{L^2(\Gamma)}.$$

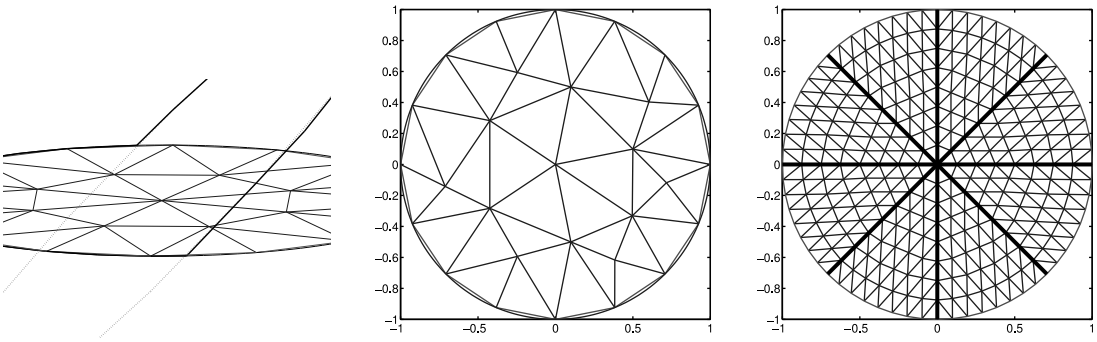
Returning to (2.58) we finally obtain

$$|\int_\Omega (y - y_0)\big((y - y^h) - (z_h - y_h)\big)| \le Ch^{\frac{3}{2}}\|u - \tilde{u}_h\|_{L^2(\Gamma)}$$

and the result follows after inserting this estimate into (2.57). □

2.1.5. Numerical experiments. For our numerical experiments we consider the variational discretization (2.34) of problem (2.3) with the unit circle $\Omega = B_1(0) \subset \mathbb{R}^2$ as domain and $\mathcal{A} = -\Delta$ as differential operator. We set $\alpha = 1$, $u_a = 0$ and $u_b = 1$. For the numerical solution of the optimal control problem (2.34) we apply the fixpoint iteration

- $v \in U_{h,ad}$ given
- $v^+ := P_{[u_a,u_b]}\left(\frac{1}{\alpha}\partial^h_{\vec{\nu}_\mathcal{A}} p_h(v)\right)$
- $v := v^+$.



(a) Variational discretization. (b) Arbitrary mesh ($i = 2$). (c) Superconvergence mesh ($j = 4$).

FIGURE 2.1. Variational discretization and considered triangulations.

Here, for given $v \in U_{h,ad}$ the function $\partial^h_{\vec{\nu}_\mathcal{A}} p_h(v)$ is defined by (2.36) with $y_h = \mathcal{G}_h(v)$. We note that the variational discrete solution u_h may admit active sets whose boundaries do not coincide with finite element nodes, compare Figure 2.1(a), where the boundary control u_h (bold) is depicted on a coarse mesh together with function $\frac{1}{\alpha}\partial^h_{\vec{\nu}_\mathcal{A}} p_h(u_h)$ (dotted).

Remark 2.1.9. The above fixpoint iteration converges for sufficiently large $\alpha > 0$, since $P_{[u_a,u_b]}$ is non-expanding and $v \mapsto \partial^h_{\vec{\nu}_\mathcal{A}} p_h(v)$ is the decomposition of discrete solution operators. The mapping $v \mapsto P_{[u_a,u_b]}\left(\frac{1}{\alpha}\partial^h_{\vec{\nu}_\mathcal{A}} p_h(v)\right)$ then is contractive for $\alpha > 0$ sufficiently large. In the opposite case one could use a more sophisticated boundary element method as described in [OPS10].

We consider two examples and investigate the error functionals

$$E^0_u(h) = \|u - \tilde{u}_h\|_{L^2(\Gamma)}, \quad E^0_y(h) = \|y - y_h\|_{L^2(\Omega_h)}, \quad E^1_y(h) = \|y - y_h\|_{H^1(\Omega_h)},$$
$$E^0_p(h) = \|p - p_h\|_{L^2(\Omega_h)}, \quad E^1_p(h) = \|p - p_h\|_{H^1(\Omega_h)},$$

both on a sequence of arbitrary meshes and on a sequence of congruently refined, piecewise $\mathcal{O}(h^2)$ irregular meshes. Figure 2.1(b) shows an arbitrary mesh while Figure 2.1(c) depicts a grid of the type which we use to numerically confirm our superconvergence result of Theorem 2.1.8.

Remark 2.1.10. The triangulation in Figure 2.1(c) is piecewise $\mathcal{O}(h^2)$ irregular, but only $\mathcal{O}(h)$ irregular. It is automatically constructed by congruent refinement from the initial grid formed by the 8 bold sector boarders together with the corresponding sector secants. Here we note that new boundary nodes are projected onto the unit circle. The resulting triangulation in each of the 8 sectors then is $\mathcal{O}(h^2)$ irregular.

Piecewise $\mathcal{O}(h^2)$ irregular meshes are often generated automatically by congruent refinement, say from an initial grid $\mathcal{T}_0$ containing finitely many triangles T combined with projecting boundary nodes onto smooth domain boundaries. Every sub-triangulation obtained in this way from some $T \in \mathcal{T}_0$ then is $\mathcal{O}(h^2)$ irregular. This in view of Theorem 2.1.8 explains why in practice one often observes better rates of convergence than expected from the general theory, compare the discussion in [BX03].

Tables 2.1 and 2.2 summarize the mesh-properties in terms of the number of triangles nt, the number of nodes m and the mesh parameter h.

i	nt	m	h
1	8	9	1.000000
2	40	29	0.596568
3	170	102	0.298819
4	684	371	0.149721
5	2680	1393	0.074921
6	10812	5511	0.037497
7	44568	22489	0.018749
8	179292	90051	0.009375
9	701964	351791	0.004687

TABLE 2.1. Mesh parameters for the sequence of arbitrary meshes.

j	nt	m	h
1	8	9	1.000000
2	32	25	0.571070
3	128	81	0.302195
4	512	289	0.155086
5	2048	1089	0.078516
6	8192	4225	0.039498
7	32768	16641	0.019809
8	131072	66049	0.009919
9	524288	263169	0.004963

TABLE 2.2. Mesh parameters for the sequence of piecewise $\mathcal{O}(h^2)$ irregular meshes.

Finally for an arbitrary function $g : B_1(0) \to \mathbb{R}$ we abbreviate $\hat{g}(r,\phi) := g(r\cos\phi, r\sin\phi)$, where $(r,\phi) \in (0,1]\times[0,2\pi)$. For constructing analytical examples it is helpful to recall

$$\Delta g = \hat{g}_{rr} + \frac{1}{r}\hat{g}_r + \frac{1}{r^2}\hat{g}_{\phi\phi}$$

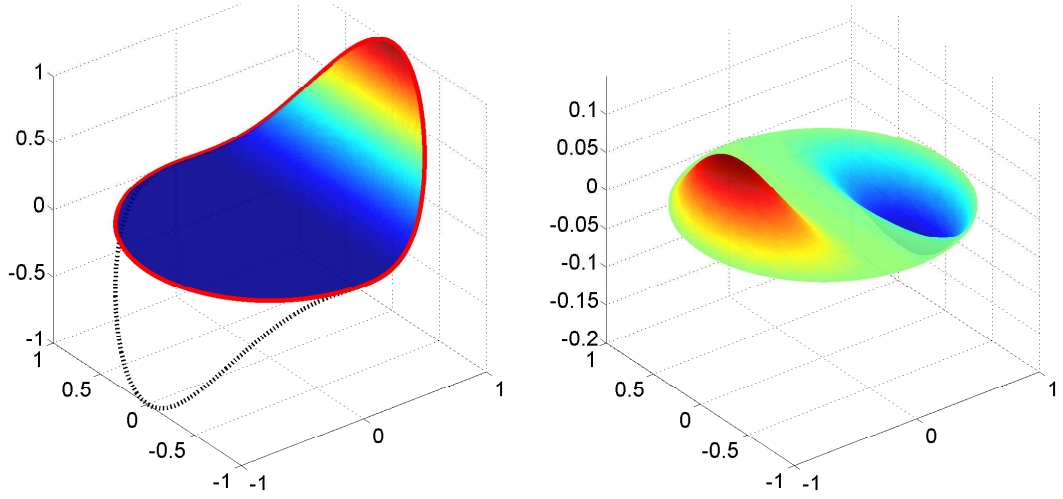
for $g \in C^2(\Omega)$.

Example 2.1.11. In our first example we consider problem (2.3) with continuous data f and smooth data y_0. For this purpose we set

$$\begin{aligned}\hat{y}(r,\phi) &= r^3\max(0,\cos^3\phi)\\ \widehat{y_0}(r,\phi) &= (7r^2\cos^2\phi + 6r^2 - 6r)\cos\phi + \hat{y}(r,\phi) \text{ and}\\ \hat{f}(r,\phi) &= -6r\max(0,\cos\phi).\end{aligned}$$

Then it is easy to check that $\hat{u}(1,\phi) = \hat{u}(\phi) = \max(0,\cos^3\phi)$ solves (2.3) and the associated adjoint variable is given by $\hat{p}(r,\phi) = r^3(r-1)\cos^3\phi$. In the present example we deal with classical solutions in the sense that $y, p \in C^2(\bar{\Omega})$ and $u \in C^2(\Gamma)$, see Figures 2.2(a) and 2.2(b). Table 2.3 summarizes the numerical results for the sequence of arbitrary meshes from Table 2.1. In addition to the EOCs for two consecutive meshes also the average and the EOC between coarsest and finest grid is computed in the rows $\varnothing$ and $\frac{1}{9}$. The EOC for E_u^0 behaves as predicted by Theorem 2.1.4, whereas the L^2-error of the state E_y^0 converges with a rate of 1.5 faster than predicted. In Table 2.4 we present the numerical results for our sequence of $\mathcal{O}(h^2)$ irregular meshes. One clearly observes the superconvergence effect for piecewise $\mathcal{O}(h^2)$ irregular grids predicted by Theorem 2.1.8. Again the rate of convergence for E_u^0 behaves as expected whereas the EOC for E_y^0 is nearly quadratic.

Example 2.1.12. Next, let us construct an analytical solution to problem (2.3) in the same way as in the previous example but in which the optimal control now only



(a) State y, control u and $\frac{1}{\alpha}\partial_{\bar{\nu}} p|_\Gamma$. (b) Adjoint state p.

FIGURE 2.2. Analytical solution of Example 2.1.11.

i	E_u^0	EOC	E_y^0	EOC	E_y^1	EOC	E_p^0	EOC	E_p^1	EOC
1	0.277414	-	0.149239	-	0.983167	-	0.073546	-	0.313464	-
2	0.071514	2.624	0.040577	2.521	0.441360	1.550	0.039436	1.207	0.287896	0.165
3	0.070380	0.023	0.023135	0.813	0.407958	0.114	0.012772	1.631	0.175988	0.712
4	0.018892	1.903	0.005006	2.215	0.158316	1.370	0.003133	2.034	0.085166	1.050
5	0.011166	0.760	0.001868	1.424	0.104513	0.600	0.000771	2.024	0.041827	1.027
6	0.006742	0.729	0.000762	1.295	0.081769	0.355	0.000197	1.970	0.021083	0.990
7	0.004180	0.690	0.000341	1.159	0.078123	0.066	0.000050	1.978	0.010630	0.988
8	0.002040	1.035	0.000124	1.456	0.050939	0.617	0.000012	2.013	0.005287	1.008
9	0.000994	1.037	0.000044	1.513	0.033625	0.599	0.000003	2.004	0.002635	1.005
$\frac{1}{9}$		1.050		1.518		0.629		1.879		0.891
$\varnothing$		1.100		1.550		0.659		1.858		0.868

TABLE 2.3. Errors and EOCs for arbitrary meshes of Example 2.1.11.

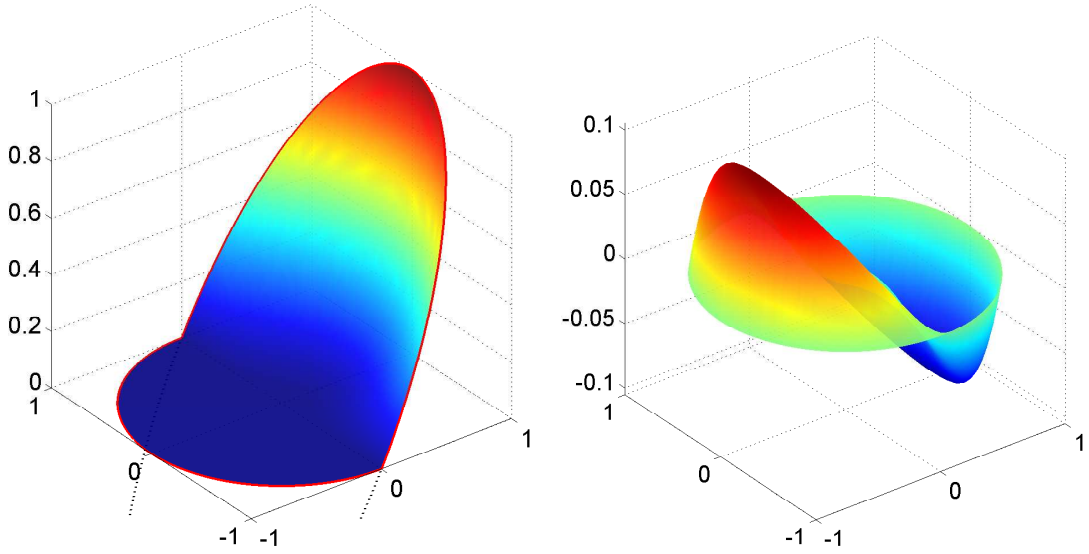
j	E_u^0	EOC	E_y^0	EOC	E_y^1	EOC	E_p^0	EOC	E_p^1	EOC
1	0.277414	-	0.149239	-	0.983167	-	0.073546	-	0.313464	-
2	0.170809	0.866	0.099800	0.718	0.904301	0.149	0.050445	0.673	0.330714	-0.096
3	0.096494	0.897	0.033170	1.731	0.587454	0.678	0.017067	1.703	0.207877	0.730
4	0.044380	1.164	0.010026	1.794	0.336040	0.837	0.004614	1.961	0.110293	0.950
5	0.018420	1.292	0.003010	1.768	0.184591	0.880	0.001177	2.007	0.056021	0.995
6	0.007163	1.375	0.000878	1.794	0.098095	0.920	0.000296	2.010	0.028126	1.003
7	0.002676	1.427	0.000248	1.833	0.050908	0.950	0.000074	2.007	0.014078	1.003
8	0.000976	1.458	0.000068	1.864	0.026013	0.971	0.000019	2.004	0.007041	1.002
9	0.000351	1.477	0.000019	1.884	0.013161	0.984	0.000005	2.002	0.003521	1.001

TABLE 2.4. Errors and EOCs for piecewise $\mathcal{O}(h^2)$ irregular meshes of Example 2.1.11.

belongs to $C^{0,1}(\Gamma)$. We choose

$$\hat{y}(r,\phi) = r^3 \max(0, \cos\phi),$$
$$\widehat{y_0}(r,\phi) = (15r^2 - 8r)\cos\phi + \hat{y}(r,\phi)$$

and set $f := -\Delta y$. Then $\hat{u}(1,\phi) = \hat{u}(\phi) = \max(0,\cos\phi)$ solves (2.3) and the associated adjoint variable is given by $\hat{p}(r,\phi) = r^3(r-1)\cos\phi$. Let us note that



(a) State y, control u and $\frac{1}{\alpha}\partial_{\vec{\nu}} p|_\Gamma$. (b) Adjoint state p.

FIGURE 2.3. Analytical solution of Example 2.1.12.

$f = -\Delta y$ has to be understood in the distributional sense, i.e.

$$\langle f, \zeta \rangle = -\int_{\mathfrak{D}} 8r(x_1, x_2)\cos(\phi(x_1, x_2))\zeta(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 - \int_{-1}^{1} x_2^2 \zeta(0, x_2)\,\mathrm{d}x_2 \qquad \forall\, \zeta \in C_0^\infty(\Omega),$$

where $\mathfrak{D} = \{(x_1, x_2) \in \bar{\Omega} : x_1 > 0\}$. In particular, $f \notin L^2(\Omega)$. Nevertheless, the state equation and the corresponding boundary control problem are still meaningful, compare [LM72, p. 188].

Figure 2.3(a) shows the optimal state y with the optimal boundary control u and Figure 2.3(b) presents the associated adjoint state p. The convergence behaviour of our error functionals is similar to that observed in the previous example. For arbitrary meshes E_u^0 converges linearly as is shown in Table 2.5. On our sequence of piecewise $\mathcal{O}(h^2)$ irregular meshes the convergence rate of this error functional improves to 1.5 as displayed in Table 2.6. Again in both cases the behaviour of E_y^0 is better than predicted and the convergence rate on our sequence of piecewise $\mathcal{O}(h^2)$ irregular meshes is higher than on the sequence of arbitrary meshes.

Example 2.1.11 with a mixed formulation. Let us briefly consider a mixed formulation for the underlying state equation (2.1) as carried out in (1.9a)-(1.9b) with $f_u := f$ and $g_u := u$. Discretization with the help of lowest order Raviart–Thomas elements gives a finite dimensional linear system (1.27a)-(1.27b), whose solution $(y_h, \vec{v}_h) \in P^0_{td,h}(\mathcal{T}_h) \times RT_{0,h}(\mathcal{T}_h)$ we denote by $\mathcal{G}_h(u)$. Similar to problem (2.34) we again apply variational discretization to problem (2.3), which yields the following control problem depending on h:

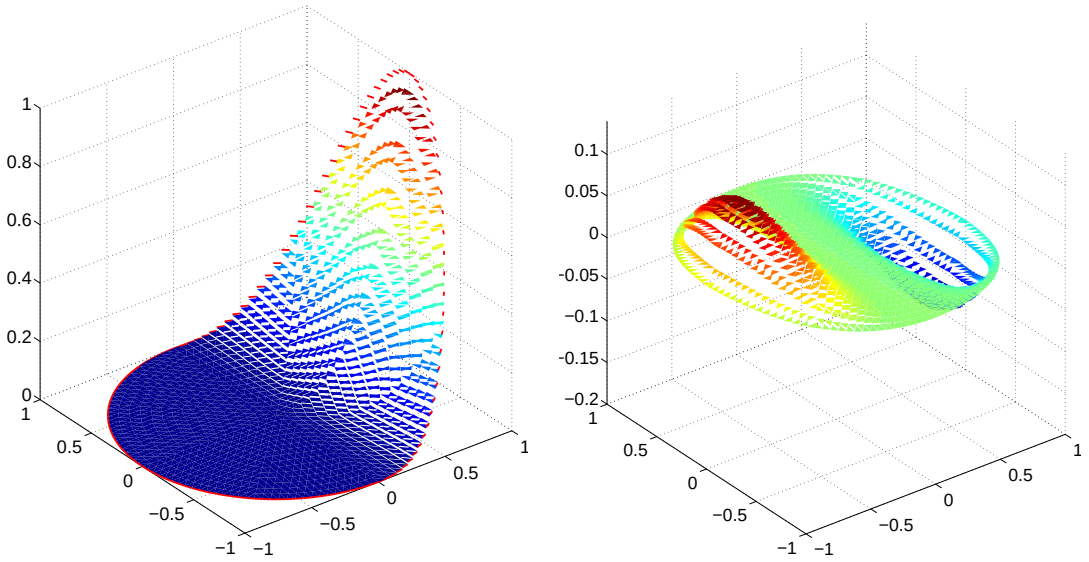
$$\min_{u_h \in U_{h,ad}} J_h(u_h) = \frac{1}{2}\int_{\Omega_h} |y_h - y_{h,0}|^2 + \frac{\alpha}{2}\int_{\Gamma_h} |u_h|^2 \quad \text{subject to } (y_h, \vec{v}_h) = \mathcal{G}_h(u_h). \tag{2.59}$$

i	E_u^0	EOC	E_y^0	EOC	E_y^1	EOC	E_p^0	EOC	E_p^1	EOC
1	0.297512	-	0.179552	-	0.904504	-	0.092648	-	0.376590	-
2	0.138275	1.483	0.057686	2.198	0.639030	0.673	0.049097	1.229	0.353687	0.121
3	0.098899	0.485	0.029068	0.991	0.557107	0.198	0.015616	1.657	0.214636	0.722
4	0.019660	2.338	0.005318	2.458	0.181005	1.627	0.003832	2.033	0.103822	1.051
5	0.016497	0.253	0.002845	0.904	0.167813	0.109	0.000959	2.000	0.051436	1.014
6	0.008651	0.932	0.001009	1.498	0.113754	0.562	0.000244	1.979	0.025931	0.989
7	0.005008	0.789	0.000422	1.256	0.097192	0.227	0.000061	1.987	0.013014	0.995
8	0.002531	0.985	0.000158	1.421	0.066351	0.551	0.000015	2.003	0.006483	1.005
9	0.001241	1.028	0.000057	1.469	0.045102	0.557	0.000004	1.993	0.003234	1.003
$\frac{1}{9}$		1.022		1.502		0.559		1.881		0.887
$\varnothing$		1.037		1.524		0.563		1.860		0.863

TABLE 2.5. Errors and EOCs for arbitrary meshes of Example 2.1.12.

j	E_u^0	EOC	E_y^0	EOC	E_y^1	EOC	E_p^0	EOC	E_p^1	EOC
1	0.325180	-	0.138937	-	0.856030	-	0.093237	-	0.377756	-
2	0.208354	0.795	0.105716	0.488	1.047727	-0.361	0.062975	0.700	0.405792	-0.128
3	0.121702	0.845	0.038899	1.571	0.720985	0.587	0.021157	1.714	0.256137	0.723
4	0.057121	1.134	0.012582	1.692	0.435196	0.757	0.005715	1.962	0.136028	0.949
5	0.023779	1.287	0.003810	1.755	0.245236	0.843	0.001459	2.006	0.069111	0.995
6	0.009233	1.377	0.001096	1.813	0.131543	0.907	0.000367	2.010	0.034699	1.003
7	0.003442	1.430	0.000305	1.853	0.068408	0.947	0.000092	2.007	0.017368	1.003
8	0.001254	1.460	0.000083	1.877	0.034941	0.971	0.000023	2.004	0.008686	1.002
9	0.000451	1.478	0.000022	1.895	0.017675	0.984	0.000006	2.002	0.004344	1.001

TABLE 2.6. Errors and EOCs for piecewise $\mathcal{O}(h^2)$ irregular meshes of Example 2.1.12.



(a) Optimal state y_h with boundary control u_h. (b) Adjoint state p_h.

FIGURE 2.4. Numerical solution of Example 2.1.11 for $j = 5$ with mixed formulation.

It naturally comes out, that the adjoint state is also of mixed structure $(p_h, \vec{\chi}_h) \in P^0_{td,h}(\mathcal{T}_h) \times RT_{0,h}(\mathcal{T}_h)$. Figure 2.4 depicts the numerical solution of Example 2.1.11 with this discretization scheme for the mesh $j = 5$.

j	E_u^0	EOC
1	0.465830	-
2	0.190478	1.596
3	0.081835	1.327
4	0.038559	1.128
5	0.018951	1.044
6	0.009433	1.015
7	0.004711	1.006
8	0.002355	1.003

TABLE 2.7. Errors and EOCs for piecewise $\mathcal{O}(h^2)$ irregular meshes of Example 2.1.11 with mixed formulation.

Since by (1.26) $\vec{\nu}_h \cdot \vec{w}_h$ is piecewise constant on Γ_h for $\vec{w}_h \in RT_{0,h}(\mathcal{T}_h)$ and the optimal control $u_h \in L^2(\Gamma_h)$ satisfies

$$u_h = P_{[u_a,u_b]}\left(\frac{1}{\alpha}\vec{\nu}_h \cdot \vec{\chi}_h\right) \qquad \text{on } \Gamma_h,$$

u_h is also piecewise constant, and hence is of simple structure. However, the best one would expect for the convergence of the $L^2(\Gamma)$-error for the control is of order $\mathcal{O}(h)$. This is independent from which computational meshes are used. Especially for piecewise $\mathcal{O}(h^2)$ irregular meshes we do not expect a superconvergence effect. This is the case as one can read out from Table 2.7.

2.2. Optimal distributed control on polygonal domains

In this section we focus onto distributed elliptic optimal control problems on polygonal domains Ω with control constraints. To be more accurate we consider in (1.3) functions f_u depending on $u : \Omega \to \mathbb{R}$ and $g_u = g : \partial\Omega \to \mathbb{R}$ independent from u. In Subsection 2.2.0 we introductorily mention applications of distributed optimal control and give an overview about related literature in this field. The specialty introduced by additional control constraints is the appearance of variational inequalities in first order optimality systems as we will see in Subsection 2.2.1. Therefore we do not focus onto a concrete class of state equations. In fact, the specific state equation is not determined. In Subsection 2.2.2 we aim to compare a classical discretization approach with variational discretization in terms of sensitivities and algorithmical realization. Finally in Subsection 2.2.3 we investigate and visualize our findings in terms of a simple numerical example.

2.2.0. Introduction. Optimal control problems governed by partial differential equations with control constraints practically emerge, when for instance a distributed heat source as control influences a dynamical system. This can be achieved for example by electro magnetic induction or by emission of micro waves. To have a certain application in mind we remember onto the already mentioned optimization in glass cooling processes or in crystal growing. Due to physical bounds concerning the heating power naturally control constraints come into play.

The numerical analysis of optimization problems governed by PDEs with control constraints goes back to the 70s when Falk [Fal73] and Geveci [Gev79] present convergence analysis for elliptic optimization problems with distributed controls and piecewise constant Ansatz functions for the controls. Malanowski in [Mal82]

investigates convex parabolic optimal control problems with piecewise constant as well as piecewise linear Ansatz functions for the controls.

Optimal error estimates for distributed control of semi-linear elliptic equations with piecewise constant controls are presented by Arada, Casas and Tröltzsch in [ACT02]. The authors prove linear convergence for the error of the controls in both the L^2- and the L^∞-norm. Only a few optimal results are known for continuous, piecewise linear approximations of the control. In [Rös06], Rösch proves convergence of order $h^{3/2}$ for the controls for a one-dimensional linear-quadratic elliptic model problem under special assumptions on the continuous solution. Similar results are obtained by Casas and Tröltzsch in [CT03], where also boundary control problems are investigated.

Meyer and Rösch prove a super-convergence result for piecewise constant discrete controls $\bar{u}_h$ in [MR04] and use this result to show quadratic convergence of the post-processed control $u = P_{[u_a,u_b]}(-\frac{1}{\alpha}p_h(y_h(\bar{u}_h)))$ for elliptic distributed control problems in two space dimensions under mild assumptions on the intersection of the active set of the optimal control and the finite element grid. Here, $p_h(y_h(\bar{u}_h))$ denotes the discrete adjoint associated to $y_h(\bar{u}_h)$. In particular Meyer and Rösch have to require that the $(d-1)$-dimensional Hausdorff measure of the discrete active set induced by the optimal control only intersects with a certain number of simplexes of the triangulation ([HPUU09, Sec. 3.2.6.2]). The same authors prove L^∞–estimates for elliptic control problems with piecewise linear, continuous controls in [MR06]. Apel and Rösch extend the results of [MR04] to non-convex domains with corner singularities and prove for appropriately graded meshes together with Winkler in [ARW06] quadratic convergence in L^2, and together with Sirch in [ARS09] $h^2|\log h|$ convergence in L^∞.

In [Hin05] Hinze presents a general abstract variational discretization concept together with a tailored algorithmic concept for linear-quadratic problems with control constraints. The concept allows to compute discrete controls without discretizing the control space. It is applicable to a large class of control constrained optimal control problems with PDEs, including parabolic equations [HPUU09, Chap. 3], and the (time-dependent) Stokes system. In particular it applies to the problems considered in [MR04, MR06, ARW06, ARS09], and leads to error estimates for the controls of at least the same quality as presented there.

In [Sch06] Schiela also applied this concept for PDE-constrained optimization with control constraints for an interior point function space algorithm. Besides its convergence analysis numerical experiments concerning the behavior of variational discretization for linear and quadratic finite elements are investigated.

Recently Hinze and Vierling combined variational discretization and semi-smooth Newton methods in [HV09] to a numerical algorithm to whom they address implementation issues, convergence analysis and globalization techniques.

Optimal control problems for parabolic equations in the presence of control constraints are considered by Meidner and Vexler in [MV08a, MV08b]. They use discontinuous Galerkin methods in time and finite elements in space to discretize the state equations, and among other things obtain optimal convergence results for variational discretization [Hin05] of the controls. This is also reported by Hinze in [HPUU09, Chap. 3].

For reasons of clarity we abandon to provide a bibliographic overview concerning adaptive approaches for control constrained optimization of PDEs at this stage.

Since we are going to focus onto goal-oriented adaptivity for control and state constrained problems in Section 3.2.3, we illuminate available literature accumulated in Section 3.2.1.

2.2.1. Mathematical setting. Let $\Omega \subset \mathbb{R}^d$ (d=2, 3) be a bounded domain. For a simplified discussion we additionally assume that Ω is polygonally bounded, since then there exists an exact partition $\mathcal{T}_h$ of Ω. The finite element analysis for smooth bounded domains is already carried out in Subsection 2.1.2.

We consider the distributed optimal control problem

$$\begin{aligned} &\min_{u\in U_{ad}} J(u) = \frac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - u_0\|^2_U \\ &\text{subject to } y = \mathcal{G}(u), \end{aligned} \tag{2.60}$$

with $U = L^2(\Omega)$. Here $\alpha > 0$ and the functions y_0, $u_0 \in L^2(\Omega)$ are given. The space of admissible controls is

$$U_{ad} = \{u \in U \,:\, u_a \le u \le u_b\}$$

for some fixed $u_a, u_b \in \mathbb{R}$ with $u_a < u_b$. U_{ad} is a convex, closed subset of the HILBERT-space U. Since we are only going to focus onto the influence of control constraints, we do not specify the state equation and the corresponding solution operator $\mathcal{G}$ here. In particular it is not required the state equation to be linear.

Instead of that we subsequently assume that problem (2.60) has a unique solution $u \in U_{ad}$ and there exists a corresponding adjoint state p at least in $L^2(\Omega)$ such that both objects satisfy the variational inequality

$$(\alpha(u - u_0) + p, v - u) \ge 0 \quad \forall\, v \in U_{ad}. \tag{2.61}$$

It easily can be shown that

$$u = P_{[u_a,u_b]}\left(-\frac{1}{\alpha}p + u_0\right) \qquad \text{a.e. in } \Omega \tag{2.62}$$

follows, where $P_{[u_a,u_b]} : L^2(\Omega) \to L^2(\Omega)$ denotes the usual orthogonal L^2-projection onto the space U_{ad}. We emphasize that this is a nonlinear equation only caused due to control constraints. When omitting those by setting $u_b = -u_a = \infty$ equation (2.62) simplifies to the dependence $u = -\frac{1}{\alpha}p + u_0$.

2.2.2. Finite element discretization and numerical realization. We follow the concept of "first discretize then optimize" and compare a classical discretization approach with variational discretization. We further investigate how these discretization schemes mimic the variational inequality (2.61) as well as the projection formula (2.62). As already mentioned in the introduction mathematical programs use the characterizing first order optimality conditions and hence also those variational inequalities for an efficient optimization. For applying a generalized NEWTON method onto the KKT-equations it is among others helpful to know the sensitivity of the variable u with respect to the adjoint state p and their corresponding discrete counterparts. We are going to investigate this sensitivities for both discretization schemes. Our observations are going to have important consequences onto the numerical realization.

2.2.2.1. *Classical discretization.* In many existing finite element codes at least the standard space of linear finite elements $P^1_{c,h}(\mathcal{T}_h)$ is contained. Therefore we focus our discussion onto this space. It is easily implementable if one naively discretizes $u_h, y_h \in Y_h := P^1_{c,h}(\mathcal{T}_h)$ in advance. We end up with the fully discrete finite dimensional optimal control problem

$$\begin{aligned} &\min_{u_h \in U^h_{ad}} J_h(u_h) = \|y_h - y_{0,h}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - u_{0,h}\|^2_U \\ &\text{subject to } y_h = \mathcal{G}_h(u_h), \end{aligned} \tag{2.63}$$

where $U^h_{ad} := \{v_h \in Y_h \,:\, u_a \le v_h \le u_b\}$ and $\mathcal{G}_h$ is a discrete approximation to the solution operator $\mathcal{G}$. The furthermore $u_{0,h}$ and $y_{0,h}$ are finite element functions in Y_h approximating the data u_0, y_0 such that

$$\|u_0 - u_{0,h}\|_{L^2(\Omega)} + \|y_0 - y_{0,h}\|_{L^2(\Omega)} = \mathcal{O}(h).$$

In analogue to Subsection 2.2.1 again we assume that problem (2.63) has a unique solution $u_h \in U^h_{ad}$ and there exists a corresponding adjoint state $p_h \in Y_h$ such that both objects satisfy the variational inequality

$$(\alpha(u_h - u_{0,h}) + p_h, v_h - u_h) \ge 0 \quad \forall\, v_h \in U^h_{ad}. \tag{2.64}$$

Recalling the notation from Section 1.2 (2.64) can be rewritten into

$$(\alpha(\mathbf{u} + \mathbf{u_0}) + \mathbf{p})^T \mathbf{M}(\mathbf{v} - \mathbf{u}) \ge 0 \quad \forall\, \mathbf{v} \in [u_a, u_b]^m \tag{2.65}$$

or equivalently into

$$\textstyle\sum_{i=1}^m \sum_{j \in \mathcal{N}_i} (\alpha(u_i - u_{0,i} + p_i) m_{ij}(v_j - u_j) \ge 0 \quad \forall\, \mathbf{v} \in [u_a, u_b]^m. \tag{2.66}$$

Remark 2.2.1. The discrete counterpart of the projection formula (2.62) is neither the first guess $u_h = P_{[u_a,u_b]}\left(-\frac{1}{\alpha}p_h + u_{0,h}\right)$, since this contradicts $u_h \in Y_h$ in general, nor the second guess $u_i = P_{[u_a,u_b]}(-\frac{1}{\alpha}p_i + u_{0,i})$ for all $i \in \{1, \dots, m\}$. If it would, we assume for conviction w.l.o.g. $u_1 = u_b < -\frac{1}{\alpha}p_1 + u_{0,1}$ and $u_i = -\frac{1}{\alpha}p_i + u_{0,i} \in (u_a, u_b)$ for $i \in \{2, \dots, m\}$. For a fixed $k \in \mathcal{N}_1 \setminus \{1\}$ we define $v_k := u_b$ and $v_j := u_j$ for all $j \in \{1, \dots, m\} \setminus \{k\}$. Then $\mathbf{v} \in [u_a, u_b]^m$ and by (2.66)

$$\underbrace{(\alpha(u_1 - u_{0,1}) + p_1)}_{<0} \underbrace{m_{1k}}_{>0} \underbrace{(v_k - u_k)}_{>0} \ge 0$$

contradicts the assumption $u_i = P_{[u_a,u_b]}(-\frac{1}{\alpha}p_i + u_{0,i})$ for all $i \in \{1, \dots, m\}$.

There is another way to express the dependence between the discrete control $\mathbf{u}$ and adjoint state $\mathbf{p}$. The orthogonal projection of $-\frac{1}{\alpha}p_h + u_{0,h}$ onto Y_h is the solution of a box constrained quadratic minimization problem, i.e.

$$\mathbf{u} = \operatorname*{argmin}_{\mathbf{v} \in [u_a,u_b]^m} \mathbf{v}^T\mathbf{M}\mathbf{v} - 2\left(-\frac{1}{\alpha}\mathbf{p} + \mathbf{u_0}\right)^T \mathbf{M}\mathbf{v} =: P^h_{[u_a,u_b]}\left(-\frac{1}{\alpha}\mathbf{p} + \mathbf{u_0}\right). \tag{2.67}$$

The Lagrangian $L : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ for the minimization problem (2.67) is given by

$$L(\mathbf{u}, \boldsymbol{\lambda}_\mathbf{b}, \boldsymbol{\lambda}_\mathbf{a}, \mathbf{p}) := \mathbf{u}^T\mathbf{M}\mathbf{u} - 2\left(-\frac{1}{\alpha}\mathbf{p} + \mathbf{u_0}\right)^T \mathbf{M}\mathbf{u} + \boldsymbol{\lambda}_\mathbf{b}^T(\mathbf{u} - \mathbf{u_b}) + \boldsymbol{\lambda}_\mathbf{a}^T(\mathbf{u_a} - \mathbf{u}),$$

where $\mathbf{u_a}, \mathbf{u_b} \in \mathbb{R}^m$ are the vector representations of $I_h u_a$ and $I_h u_b$. In order to implement a generalized NEWTON-method it is necessary to derive (2.67) for p_j. Therefore we state the theorem of sensitivity from [Fia83] adapted to our purpose as the following

Lemma 2.2.2. *Let $f, g_1, \dots, g_{2m} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ be two times continuously differentiable functions and $\hat{w} \in \mathbb{R}^m$ fixed. Let further $\hat{x} \in \mathbb{R}^m$ be a strictly regular local minimum of*

$$\min_{x \in \mathbb{R}^m} f(x, \hat{w}), \qquad s.t.\ g_i(x, \hat{w}) \le 0, \quad i = 1, \dots, 2m \tag{$NLP(\hat{w})$}$$

with corresponding Lagrangian $L(\hat{x}, \hat{\lambda}, \hat{w}) = f(\hat{x}, \hat{w}) + \hat{\lambda}^T g(\hat{x}, \hat{w})$ and Lagrange *multiplier $\hat{\lambda} \in \mathbb{R}^{2m}$. Then there exist neighborhoods $V_\epsilon(\hat{w})$ and $U_\delta(\hat{x}, \hat{\lambda})$, such that (NLP(w)) has a unique strictly regular local minimum $(x(w), \lambda(w)) \in U_\delta(\hat{x}, \hat{\lambda})$ for all $w \in V_\epsilon(\hat{w})$. Additionally $(x(w), \lambda(w))$ is continuously differentiable w.r.t. w with*

$$\begin{bmatrix} \frac{dx}{dw}(\hat{w}) \\ \frac{d\lambda}{dw}(\hat{w}) \end{bmatrix} = - \begin{bmatrix} \nabla^2_{xx} L & (g'_x)^T \\ \hat{\Lambda} g'_x & \hat{\Gamma} \end{bmatrix}^{-1} \begin{bmatrix} \nabla^2_{xw} L \\ \hat{\Lambda} g'_w \end{bmatrix}, \tag{2.68}$$

where $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_{2m})$, $\hat{\Gamma} = \operatorname{diag}(g_1, \dots, g_{2m})$. All functions and derivatives are evaluated at $(\hat{x}, \hat{\lambda}, \hat{w})$.

Before we apply the above lemma, we introduce subsets of indices corresponding to active and inactive box constraints.

Definition 2.2.3. For given $u_h \in U^h_{ad}$ we define the index sets

$$\begin{aligned}
\bullet &:= \{1, \dots, m\} \\
\circledcirc &:= \{i \in \bullet : u_i = u_a\} && \text{lower control active index set} \\
\circledast &:= \{i \in \bullet : u_i = u_b\} && \text{upper control active index set} \\
\circledcirc\!\!\circledast &:= \circledcirc \cup \circledast && \text{control active index set} \\
\circ &:= \bullet \setminus \circledcirc\!\!\circledast. && \text{control inactive index set}
\end{aligned}$$

The sets $\circledcirc, \circledast$ and $\circ$ are a disjoint decomposition of $\bullet = \{1, \dots, m\}$. Recalling Definition 1.2.12 of a blockwise split of a quadratic matrix, we are able to express the sensitivity in (2.67) with respect to p_i within the

Lemma 2.2.4. *Let $u_h \in U^h_{ad}$ and $p_h \in Y_h$ satisfy (2.64). Then*

$$\frac{d\mathbf{u}}{d\mathbf{p}} = -\frac{1}{\alpha} \begin{bmatrix} \mathbf{I}_\circ & \mathbf{M}_\circ^{-1} \mathbf{M}_{\circ\circledcirc\!\!\circledast} \\ \mathbf{0}_{\circledcirc\!\!\circledast\circ} & \mathbf{0}_{\circledcirc\!\!\circledast} \end{bmatrix}. \tag{2.69}$$

Proof. Applying Lemma 2.2.2 to (2.67), we obtain the sensitivities of $\mathbf{u}, \boldsymbol{\lambda}_\mathbf{b}$ and $\boldsymbol{\lambda}_\mathbf{a}$ with respect to $\mathbf{p}$ due to the solution of the following linear system:

$$\begin{bmatrix} \frac{d\mathbf{u}}{d\mathbf{p}} \\ \frac{d\boldsymbol{\lambda}_\mathbf{b}}{d\mathbf{p}} \\ \frac{d\boldsymbol{\lambda}_\mathbf{a}}{d\mathbf{p}} \end{bmatrix} = \begin{bmatrix} 2\mathbf{M} & \mathbf{I} & -\mathbf{I} \\ \operatorname{diag}(\boldsymbol{\lambda}_\mathbf{b}) & \operatorname{diag}(\mathbf{u} - \mathbf{u}_\mathbf{b}) & \mathbf{0} \\ -\operatorname{diag}(\boldsymbol{\lambda}_\mathbf{a}) & \mathbf{0} & \operatorname{diag}(\mathbf{u}_\mathbf{a} - \mathbf{u}) \end{bmatrix}^{-1} \begin{bmatrix} -\frac{2}{\alpha}\mathbf{M} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

All appearing matrix-blocks are in $\mathbb{R}^{m \times m}$. Fortunately we are able to proceed in computing the sensitivity $\frac{d\mathbf{u}}{d\mathbf{p}}$ due to the simple structure of the system-matrix. Let us decompose the inverse of the system-matrix into 9 blocks of size $m \times m$. Then

$$\begin{bmatrix} \mathbf{B} & \mathbf{B}_2 & \mathbf{B}_3 \\ \mathbf{B}_4 & \mathbf{B}_5 & \mathbf{B}_6 \\ \mathbf{B}_7 & \mathbf{B}_8 & \mathbf{B}_9 \end{bmatrix} \begin{bmatrix} 2\mathbf{M} & \mathbf{I} & -\mathbf{I} \\ \operatorname{diag}(\boldsymbol{\lambda}_\mathbf{b}) & \operatorname{diag}(\mathbf{u} - \mathbf{u}_\mathbf{b}) & \mathbf{0} \\ -\operatorname{diag}(\boldsymbol{\lambda}_\mathbf{a}) & \mathbf{0} & \operatorname{diag}(\mathbf{u}_\mathbf{a} - \mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{2.70}$$

Since we are only interested in $\frac{d\mathbf{u}}{d\mathbf{p}}$, we observe

$$\frac{d\mathbf{u}}{d\mathbf{p}} = -\frac{2}{\alpha}\mathbf{B}\mathbf{M}. \tag{2.71}$$

Now from (2.70) we obtain 3 characterizing equations

$$2\mathbf{B}\mathbf{M} + \mathbf{B}_2\mathrm{diag}(\boldsymbol{\lambda}_\mathbf{b}) - \mathbf{B}_3\mathrm{diag}(\boldsymbol{\lambda}_\mathbf{a}) = \mathbf{I} \tag{2.72a}$$

$$\mathbf{B} + \mathbf{B}_2\mathrm{diag}(\mathbf{u} - \mathbf{u}_\mathbf{b}) = \mathbf{0} \tag{2.72b}$$

$$-\mathbf{B} + \mathbf{B}_3\mathrm{diag}(\mathbf{u}_\mathbf{a} - \mathbf{u}) = \mathbf{0} \tag{2.72c}$$

to determine $\mathbf{B}$. Comparing the columns with respect to the upper active set $\oslash$ in (2.72b) tells us

$$\mathbf{B}_{\bullet\oslash} = -\mathbf{B}_{2\bullet}\mathrm{diag}(\mathbf{u} - \mathbf{u}_\mathbf{b})_{\bullet\oslash} = \mathbf{0}_{\bullet\oslash} = \mathbf{0}.$$

In the same way we obtain for (2.72c)

$$\mathbf{B}_{\bullet\circledcirc} = \mathbf{B}_{3\bullet}\mathrm{diag}(\mathbf{u}_\mathbf{a} - \mathbf{u})_{\bullet\circledcirc} = \mathbf{0}.$$

For determining $\mathbf{B}_{\bullet\circledcirc}$ we know that the LAGRANGE multipliers $\boldsymbol{\lambda}_\mathbf{b}$ and $\boldsymbol{\lambda}_\mathbf{a}$ vanish on the inactive set $\circledcirc$ due to strict complementarity. This yields for (2.72a)

$$2\mathbf{B}_\bullet\mathbf{M}_{\bullet\circledcirc} = \mathbf{I}_{\bullet\circledcirc} - \mathbf{B}_{2\bullet}\mathrm{diag}(\boldsymbol{\lambda}_\mathbf{b})_{\bullet\circledcirc} + \mathbf{B}_{3\bullet}\mathrm{diag}(\boldsymbol{\lambda}_\mathbf{a})_{\bullet\circledcirc} = \mathbf{I}_{\bullet\circledcirc}.$$

But we already know that $\mathbf{B}_{\bullet\circledast} = \mathbf{0}$ and hence (2.72a) simplifies further to

$$2\mathbf{B}_{\bullet\circledcirc}\mathbf{M}_{\circledcirc} = \mathbf{I}_{\bullet\circledcirc}.$$

Since $\mathbf{M}$ is positive definite, $\mathbf{M}_{\circledcirc}$ also is (see [HJ85]). This gives us $\mathbf{B}_{\circledast\circledcirc} = \mathbf{0}$ and $\mathbf{B}_{\circledcirc} = \frac{1}{2}\mathbf{M}_{\circledcirc}^{-1}$. Now with (2.71) we obtain

$$\frac{d\mathbf{u}}{d\mathbf{p}} = -\frac{1}{\alpha}\begin{bmatrix}\mathbf{M}_{\circledcirc}^{-1} & \mathbf{0}_{\circledcirc\circledast}\\ \mathbf{0}_{\circledast\circledcirc} & \mathbf{0}_{\circledast}\end{bmatrix}\begin{bmatrix}\mathbf{M}_{\circledcirc} & \mathbf{M}_{\circledcirc\circledast}\\ \mathbf{M}_{\circledast\circledcirc} & \mathbf{M}_{\circledast}\end{bmatrix} = -\frac{1}{\alpha}\begin{bmatrix}\mathbf{I}_{\circledcirc} & \mathbf{M}_{\circledcirc}^{-1}\mathbf{M}_{\circledcirc\circledast}\\ \mathbf{0}_{\circledast\circledcirc} & \mathbf{0}_{\circledast},\end{bmatrix}$$

which completes the proof. □

Remark 2.2.5. In the case where no box-constraints are active on the control, i.e. $\circledast = \varnothing$, (2.69) reduces to

$$\frac{d\mathbf{u}}{d\mathbf{p}} = -\frac{1}{\alpha}\mathbf{I}$$

as one would expect.

Remark 2.2.6. If one uses the lumped mass-matrix $\tilde{\mathbf{M}}$ instead in the above discussion beginning at (2.65), then indeed

$$u_i = P_{[u_a,u_b]}\left(-\frac{1}{\alpha}p_i + u_{0,i}\right) \quad \forall\, i \in \{1,\dots,m\} \tag{2.73}$$

immediately follows. Differentiating (2.73) w.r.t. p_j yields

$$\frac{d\mathbf{u}}{d\mathbf{p}} = -\frac{1}{\alpha}\begin{bmatrix}\mathbf{I}_{\circledcirc} & \mathbf{0}_{\circledcirc\circledast}\\ \mathbf{0}_{\circledast\circledcirc} & \mathbf{0}_{\circledast},\end{bmatrix} \tag{2.74}$$

which also is obtained by (2.69) since $\tilde{\mathbf{M}}$ is diagonal and hence $\tilde{\mathbf{M}}_{\circledcirc\circledast} = \mathbf{0}$.

Remark 2.2.7. The submatrix $\mathbf{M}_{\circledcirc\circledast} = \mathbf{M}_{\circledast\circledcirc}^T$ is almost the $\mathbf{0}_{\circledcirc\circledast}$-matrix. More precisely for $i \in \circledcirc$ and $j \in \circledast$ the entry m_{ij} is only then non-zero, if and only if $j \in \circledast \cap \mathcal{N}_i$ by (1.20). Therefore the matrix $\mathbf{M}_{\circledcirc\circledast}$ can be interpreted as a smearing interaction between those finite element functions having their support on the boundary from inactive to active sets.

Recalling the definition in (1.22) we conclude our discussion by

Theorem 2.2.8. *Let $u_h \in U_{ad}^h$ and $p_h \in Y_h$ satisfy* (2.64)*. Then*

$$\frac{d\hat{\mathbf{u}}}{d\mathbf{p}} = -\frac{1}{\alpha}\mathcal{M}, \tag{2.75}$$

with

$$\mathcal{M} := \begin{bmatrix} \mathbf{M}_{\mathbb{O}} & \mathbf{M}_{\mathbb{O}\circledcirc} \\ \mathbf{M}_{\circledcirc\mathbb{O}} & \mathbf{M}_{\circledcirc\mathbb{O}}\mathbf{M}_{\mathbb{O}}^{-1}\mathbf{M}_{\mathbb{O}\circledcirc} \end{bmatrix}. \tag{2.76}$$

Proof. We multiply equation (2.69) from left by the regular matrix $-\alpha\mathbf{M}$, which gives

$$-\alpha\mathbf{M}\frac{d\mathbf{u}}{d\mathbf{p}} = \begin{bmatrix} \mathbf{M}_{\mathbb{O}} & \mathbf{M}_{\mathbb{O}\circledcirc} \\ \mathbf{M}_{\circledcirc\mathbb{O}} & \mathbf{M}_{\circledcirc} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{\mathbb{O}} & \mathbf{M}_{\mathbb{O}}^{-1}\mathbf{M}_{\mathbb{O}\circledcirc} \\ \mathbf{0}_{\circledcirc\mathbb{O}} & \mathbf{0}_{\circledcirc} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{\mathbb{O}} & \mathbf{M}_{\mathbb{O}\circledcirc} \\ \mathbf{M}_{\circledcirc\mathbb{O}} & \mathbf{M}_{\circledcirc\mathbb{O}}\mathbf{M}_{\mathbb{O}}^{-1}\mathbf{M}_{\mathbb{O}\circledcirc} \end{bmatrix}.$$

□

Remark 2.2.9. The matrix $\mathcal{M} \in \mathbb{R}^{m\times m}$ is symmetric, but generally full in the block $\circledcirc \times \circledcirc$.

2.2.2.2. *Variational discretization.* We now discretize problem (2.60) in a minimal invasive way. The concept of variational discretization proposed in [Hin05] only discretizes the solution operator $\mathcal{G}$ of the state equation. Therefore primarily only the state y_h is an element of $Y_h := P_{c,h}^1(\mathcal{T}_h)$. The control u stays in function space U. We therefore consider the infinite dimensional optimal control problem

$$\begin{aligned} &\min_{u\in U_{ad}} J_h(u) = \frac{1}{2}\|y_h - y_{0,h}\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u - u_{0,h}\|_U^2 \\ &\text{subject to } y_h = \mathcal{G}_h(u). \end{aligned} \tag{2.77}$$

Again we assume that problem (2.77) has a unique solution $u_h \in U_{ad}$ and there exists a corresponding adjoint state $p_h \in Y_h$ such that both objects satisfy the variational inequality

$$(\alpha(u_h - u_{0,h}) + p_h, v - u_h) \geq 0 \quad \forall\, v \in U_{ad}. \tag{2.78}$$

Since in the above inequality the test space is U_{ad}, we are in the same setting as in (2.61) and hence the optimal control u_h can be characterized by the projection formula

$$u_h = P_{[u_a,u_b]}\left(-\frac{1}{\alpha}p_h + u_{0,h}\right). \tag{2.79}$$

Let us emphasize that the argument $-\frac{1}{\alpha}p_h + u_{0,h}$ to project is a finite element function in Y_h. So the optimal control u_h still has a lot of structure we are going to exploit.

Since for the solution finally the projection formula (2.79) between u_h and p_h has to hold, it is equivalent to determine the control active sets via the variable p_h. This is more convenient to implement since $p_h \in Y_h$. In analogy to Definition 2.2.3 we overload the meaning of the symbols therein by

Definition 2.2.10. For given $p_h \in Y_h$ we define the subsets

$$\begin{aligned} \bullet &:= \Omega \\ \circledcirc &:= \{x \in \Omega \,:\, -\tfrac{1}{\alpha}p_h(x) + u_{0,h}(x) \leq u_a\} && \text{lower control active set} \\ \circledcirc &:= \{x \in \Omega \,:\, u_b \leq -\tfrac{1}{\alpha}p_h(x) + u_{0,h}(x)\} && \text{upper control active set} \\ \circledcirc &:= \circledcirc \cup \circledcirc && \text{control active set} \\ \mathbb{O} &:= \bullet \setminus \circledcirc. && \text{control inactive set} \end{aligned}$$

The sets $\circledcirc, \circledcirc$ and $\circledcirc$ are a disjoint decomposition of $\bullet = \Omega$. The main difficulty in numerical realization consists in the fact, that the specific active sets are not known a priori. In contrast to the previous *blockwise* split of the mass-matrix $\mathbf{M}$ we further introduce an *additive* split within

Definition 2.2.11. For given $p_h \in Y_h$ we define $m \times m$-matrices

$$\begin{aligned}
\mathbf{M}_{\circledcirc} &:= \left[\int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m, \\
\mathbf{M}_{\circledcirc} &:= \left[\int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m, \\
\mathbf{M}_{\circledcirc} &:= \left[\int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m, \\
\mathbf{M}_{\circledcirc} &:= \mathbf{M}_{\circledcirc} + \mathbf{M}_{\circledcirc}.
\end{aligned}$$

Lemma 2.2.12. *For given $p_h \in Y_h$ there holds*

$$\mathbf{M} = \mathbf{M}_{\circledcirc} + \mathbf{M}_{\circledcirc} = \mathbf{M}_{\circledcirc} + \mathbf{M}_{\circledcirc} + \mathbf{M}_{\circledcirc}$$

as additive split of the mass matrix $\mathbf{M}$.

Proof. For the first equality we have

$$\mathbf{M} = \left[\int_{\bullet} \phi_i \phi_j\right]_{i,j=1}^m = \left[\int_{\circledcirc} \phi_i \phi_j + \int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m.$$

The second equality follows by definition. □

We have the following analogy to Theorem 2.2.8:

Theorem 2.2.13. *Let $u_h \in U_{ad}$ and $p_h \in Y_h$ satisfy* (2.79)*. Then*

$$\frac{d\hat{\mathbf{u}}}{d\mathbf{p}} = -\frac{1}{\alpha}\mathbf{M}_{\circledcirc}. \tag{2.80}$$

Proof. With $\mathbf{u_a} := [u_a]_{i=1}^m$, $\mathbf{u_b} := [u_b]_{i=1}^m$ and recalling (1.22), we obtain

$$\begin{aligned}
\hat{\mathbf{u}} &= \left[\int_\Omega u_h \phi_i\right]_{i=1}^m \\
&= \left[\int_{\bullet} P_{[u_a,u_b]}\left(-\tfrac{1}{\alpha}p_h + u_{0,h}\right)\phi_i\right]_{i=1}^m \\
&= \left[\int_{\circledcirc} u_a\phi_i + \int_{\circledcirc} u_b\phi_i + \int_{\circledcirc}\left(-\tfrac{1}{\alpha}p_h + u_{0,h}\right)\phi_i\right]_{i=1}^m \\
&= \mathbf{M}_{\circledcirc}\mathbf{u_a} + \mathbf{M}_{\circledcirc}\mathbf{u_b} + \mathbf{M}_{\circledcirc}\left(-\frac{1}{\alpha}\mathbf{p} + \mathbf{u_0}\right).
\end{aligned}$$

□

This result was in contrast to classical discretization easily and more smartly obtained. Variational discretization keeps the sparsity structure, since clearly in contrast to $\mathcal{M}$ the matrix $\mathbf{M}_{\circledcirc}$ is symmetric and sparse. It also mimics the analytical variational inequality (2.61) and is therefore structure exploiting. It sharply separates active from inactive parts. There is no smearing effect as is observed for the classical discretization concept.

Let us emphasize that assembling $\mathbf{M}_{\circledcirc}, \mathbf{M}_{\circledcirc}$ and $\mathbf{M}_{\circledcirc}$ can be done with a quit efficient algorithm and is not a huge computational drawback. For space dimension $d = 2$ this algorithm is implemented in Matlab by the routine `assem_mass`, whose code is exceptionally attached in Algorithm A.1. To get a little more insight, we define the auxiliary variable $\chi_h := -\frac{1}{\alpha}p_h + u_{0,h} \in Y_h$. Now the split mass matrix can be computed via the function call

$$\begin{bmatrix}\mathbf{M}_{\circledcirc} & \mathbf{M}_{\circledcirc} & \mathbf{M}_{\circledcirc}\end{bmatrix} = \texttt{assem\_mass}(\chi_h, u_a, u_b).$$

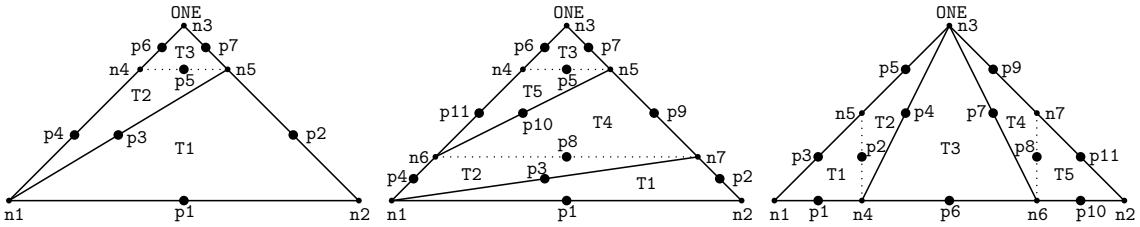


FIGURE 2.5. Quadrature nodes for the cases (2), (3) and (4).

The routine simultaneously separates 4 cases for each $T \in \mathcal{T}_h$ with respect to the activity behavior of $\chi_{h|T}$ in each of the 3 nodes `n1`, `n2` and `n3`. These cases are

(1) all nodes are lower active or
all nodes are upper active or
all nodes are inactive;
(2) exactly `ONE` active node or
exactly two lower active nodes (hence `ONE` inactive node) or
exactly two upper active nodes (hence `ONE` inactive node);
(3) two lower active nodes and `ONE` upper active node or
`ONE` lower active node and two upper active nodes;
(4) `ONE` inactive node and two differently active nodes.

W.l.o.g. we concentrate onto $\mathbf{M}_{\textcircled{1}}$. For $i, j \in \{1, \dots, m\}$ we can write

$$\int_{\textcircled{1}} \phi_i \phi_j = \sum_{T \in T_h} \int_{T \cap \textcircled{1}} \phi_i \phi_j = \sum_{T \in T_h} \int_T \mathbb{1}_{\textcircled{1}} \phi_i \phi_j$$

and observe that $\mathbb{1}_{\textcircled{1}} \phi_i \phi_j$ is a piecewisely quadratic function on T. Moreover there exists a partition of T into subtriangles `T1`, ..., `T5`, such that $\mathbb{1}_{\textcircled{1}} \phi_i \phi_j$ is indeed quadratic on every subtriangle. Now the final ingredient is the standard quadrature rule (2.81) for quadratic functions on triangles (see [GR05, Lemma 4.14]).

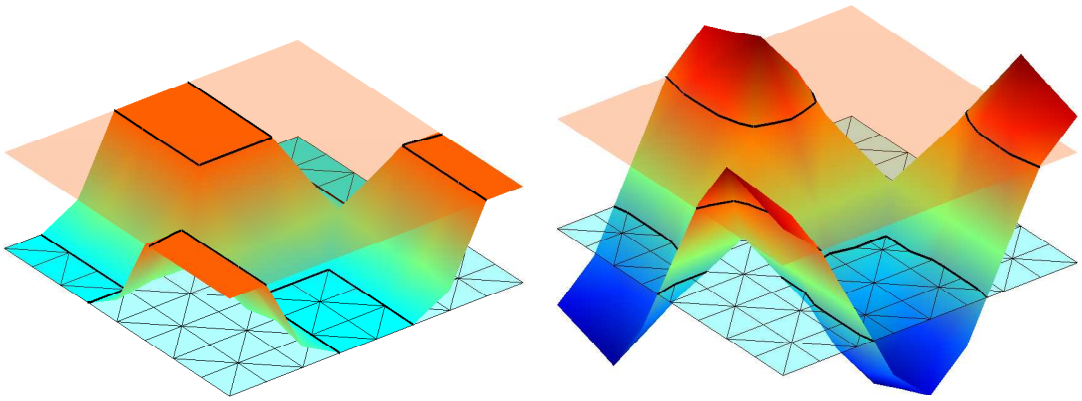
Lemma 2.2.14. *Let z_h be the quadratic function with values $z_{\boldsymbol{\alpha}}$ in the nodes $p^{\boldsymbol{\alpha}}$, $|\boldsymbol{\alpha}| = 2$ defined through*

$$z_h(p^{\boldsymbol{\alpha}}) = z_{\boldsymbol{\alpha}}, \qquad |\boldsymbol{\alpha}| = 2$$

over the triangle $K = \operatorname{conv}\{p^{100}, p^{010}, p^{001}\}$, where we have used the standard multi-index and barycentric coordinate notation. There holds

$$\int_K z_h(x) = \frac{1}{3}|K|(z_{110} + z_{011} + z_{101}). \tag{2.81}$$

The partitions of T as well as the quadrature nodes from Lemma 2.2.14 with respect to the nontrivial cases (2), (3) and (4) are depicted in Figure 2.5, where the boundaries of control active sets are displayed as dotted lines. Let us remark, that if one additionally assumes that the initially chosen mesh size h is sufficiently small such that lower and upper control activity does not occur within one triangle T than the routine `assem_mass` is even faster because of the absence of the cases (3) and (4). This is because the set of all triangles of $T \in \mathcal{T}_h$ is split into the four (possibly empty) cases. For each of these cases the terms $\int_{\mathtt{Tk}} \varphi_i \varphi_j$ ($\mathtt{k} \in \{1, \dots, 5\}$, $(i, j) \in \{(1,1), (1,2), (1,3), (2,2), (2,3), (3,3)\}$) are computed and sorted into their contribution to the local mass-matrices with respect to $\textcircled{1}, \textcircled{V}$ and $\textcircled{\Delta}$. In a final step, the global mass matrices $\mathbf{M}_{\textcircled{1}}$, $\mathbf{M}_{\textcircled{V}}$ and $\mathbf{M}_{\textcircled{\Delta}}$ are assembled from the local ones.



(a) Classical: $u_h \in U_{ad}^h$. (b) Variational: $u_h = P_{[u_a,u_b]}(-\frac{1}{\alpha}p_h) \in U_{ad}$.

FIGURE 2.6. Optimal controls u_h on a uniform mesh with $m = 81$.

2.2.3. Numerical experiment. The aim of this subsection is to visualize the advantages of variational discretization in terms of showing the sparsity structure from Theorems 2.2.8 and 2.2.13. Moreover we convince ourselves of the functionality of the routine `assem_mass`. The specific solution algorithm for the variational discretized problem (2.77) will be explained more generally in the next chapter in Section 3.1.4.2 where additional state constraints come into play. For the classical approach (2.63) we simply use the Matlab routine `quadprog`.

Let $\Omega = (0,1)^2$ and consider the optimal control problem (2.60), where for $u \in L^2(\Omega)$ the corresponding state $y = \mathcal{G}(u) \in H^1(\Omega)$ weakly solves the boundary value problem

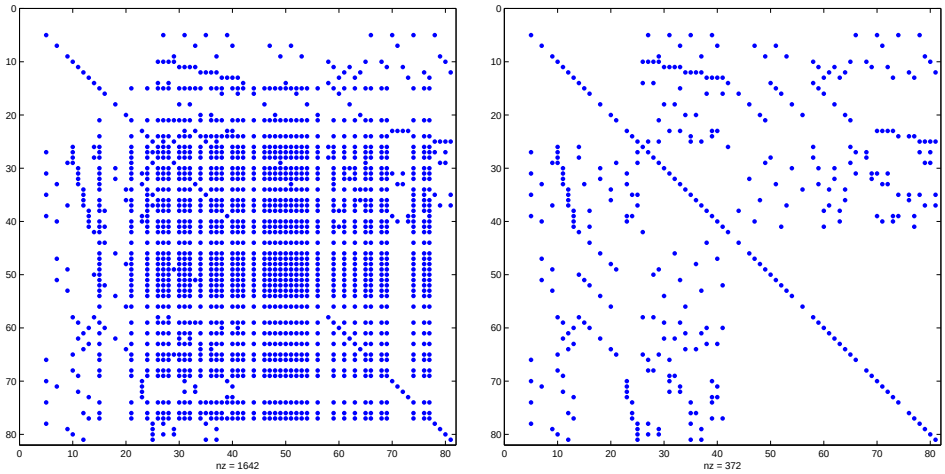
$$\begin{aligned} -\Delta y + y &= u \quad \text{in } \Omega \\ \partial_{\vec{\nu}} y &= 0 \quad \text{on } \Gamma. \end{aligned}$$

As further data we choose $\alpha = 10^{-3}$, $u_0 = 0$ and $y_0 = \cos(2\pi x_1)\sin(2\pi x_2)$. The control is constrained by the bounds $u_a = -0.5$ and $u_b = 0.7$. We solve this problem numerically for both approaches. Its solutions u_h are depicted in Figure 2.6 for both cases on a coarse mesh with $m = 81$ nodes. In Figure 2.6(b) the optimal variational control $u_h \in U_{ad}$ as well as $-\frac{1}{\alpha}p_h \in Y_h$, the bounds u_a, u_b and the numerical mesh can be seen. The control active sets are marked as solid lines and not necessarily follow edges of the triangulation. This is different to the classical approach depicted in Figure 2.6(a). Here the active sets for $u_h \in U_{ad}^h$ cannot escape its predefined structure due to the restrictive space U_{ad}^h.

Because for variational discretization the boundary of control active sets is resolved very well already on coarse grids, there is not that much the need to refine the mesh in this certain area compared to the classical approach. Here unnecessarily a lot of DOFs are required, to resolve these boundaries and to correct this unsuited concept.

Moreover let us briefly consider the sparsity structure of $\frac{d\hat{\mathbf{u}}}{d\mathbf{p}}$ depicted in Figure 2.7 as they are provided by the Theorems 2.2.8 and 2.2.13. Clearly for variational discretization $\mathbf{M}_{\textcircled{0}}$ is sparse, but for the classical approach we observe fill-in in $\mathcal{M}$.

We conclude that for control-constrained optimal control problems variational discretization is a tailored and structure exploiting concept.



(a) Classical discretization: $\mathcal{M}$. (b) Variational discretization: $\mathbf{M}_{\mathbb{O}}$.

FIGURE 2.7. Sparsity structure of $\frac{d\hat{\mathbf{u}}}{d\mathbf{p}}$ on a uniform mesh with $m = 81$.

CHAPTER 3

State constraints

3.0. Introduction

In this chapter we concentrate onto optimal control problems governed by PDEs with additional constraints on the state variable y. We recall from the introduction 0.3 the applications of optimal hyperthermia treatment planning, where lower and upper temperature bounds inside the human body have to be satisfied. Another important application is the optimal control of the Boussinesq-approximation of the Navier-Stokes equations in the chemical engineering field of crystal growing, where also state constraints in form of temperature bounds are present.

The first results concerning the existence of Lagrange multipliers for an elliptic optimal control problem under pointwise state constraints are proven by Casas in [Cas86, Cas93]. The challenging character of these problems roots in the fact that state constraints feature low regular Lagrange multipliers which are known to be Borel measures. A funded analysis about the regularity of these multipliers is also carried out in [BK03, Sch09b]. Their presence on the right hand side of the adjoint equation consequences the adjoint state p no longer to be in $H^1(\Omega)$ but only in $W^{1,s}(\Omega)$ for all $1 \leq s < \frac{d}{d-1}$. These facts complicate not only the analysis of such optimal control problems but also their numerical treatment as well. In addition, in the presence of control constraints, the solution may exhibit subsets of the underlying computational domain where both control and state are active simultaneously. Then the uniqueness of Lagrange multipliers cannot be guaranteed anymore. The statement of sufficient conditions for their uniqueness is subject in the work [Sha97].

The topic of state constraints is also well addressed in the books [HPUU09, IK08] and [Trö05]. Very popular are relaxation concepts for state constraints such as Lavrentiev, interior point and Moreau-Yosida regularization. The former one is investigated by e.g. Meyer, Rösch and Tröltzsch in [MRT06], and numerical analysis for this approach is presented by Cherednichenko and Rösch in [CR09], and by Cherednichenko, Krumbiegel and Rösch in [CKR08]. Hinze and Meyer in [HM08] present a uniform-in-parameter error analysis together with optimal parameter adjustment strategies for Lavrentiev regularization. Barrier methods applied to state constrained optimal control problems are considered by Schiela in [Sch09a]. For this approach he together with Hinze in [HS09] presents a uniform-in-parameter error analysis together with optimal parameter adjustment strategies. Numerical analysis for relaxation by penalization (see e.g. the work of Hintermüller and Kunisch [HK06a, HK06b]) including uniform-in-parameter error analysis and optimal parameter adjustment strategies is investigated by Hintermüller and Hinze in [HH09a]. Recently by Hintermüller and Kunisch in [HK09] a generalized Moreau-Yosida-based framework also applies for constraints on the gradient of the state.

In this chapter we focus onto distributed optimal control governed by linear elliptic PDEs. The main nonlinearity enters through the state constraints. We basically split this chapter into two parts.

First a priori analysis is carried out. After precisely stating the underlying optimal control problem in Section 3.1.1 and its discretization in Section 3.1.2, we further provide available literature concerning the finite element analysis and convergence rates separately in Section 3.1.3. Finally in Section 3.1.4 we explain the numerical solution of both the purely state constrained and the simultaneously control and state constrained optimal control problem.

In the second a posteriori part 3.2 of this chapter we address the issue of adaptive finite element methods, where adaption is with respect to a certain goal. In Section 3.2.1 we give an overview about the available literature concerning the so called *dual weighted residual* (DWR) approach and its extension to the presence of control and state constraints. In Section 3.2.2 we develop and investigate an a posteriori error estimator for the purely state constrained case. Some of these techniques also apply to the simultaneously control and state constrained case considered in Section 3.2.3.

3.1. A priori error analysis

3.1.1. Mathematical setting. Let Ω be a bounded domain in $\mathbb{R}^d\,(d = 2, 3)$ with either a polygonal convex or sufficiently smooth boundary $\partial\Omega$. We consider the general partial differential operator $\mathcal{A} : H^1(\Omega) \to H^1(\Omega)^*$ defined in (1.1) along with its formal adjoint operator $\mathcal{A}^*$ from (1.2). We subsequently assume the involved coefficients a_{ij}, b_i and c $(i, j = 1, \dots, d)$ to be sufficiently smooth functions on $\bar{\Omega}$ and that $\mathcal{A}$ is elliptic in the sense of Definition 1.1.1. We further suppose the corresponding bilinear form $a(\cdot,\cdot) : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}$ from (1.4) to be coercive on $H^1(\Omega)$ be means of satisfying (1.5). This can be attained by demanding a sufficient condition stated in Remark 1.1.2.

Let U be a Hilbert space and let the linear and bounded operator $B : U \to (H^1(\Omega))^*$ be given. We consider the homogeneous Neumann boundary value problem of finding $y \in H^1(\Omega)$ such that for given $u \in U$ and fixed function $f \in L^2(\Omega)$

$$\begin{aligned} \mathcal{A}y &= Bu + f && \text{in } \Omega \\ \partial_{\vec{\nu}_{\mathcal{A}}} y := \textstyle\sum_{i,j=1}^d a_{ij} y_{x_i} \nu_j &= 0 && \text{on } \partial\Omega \end{aligned} \tag{3.1}$$

holds. Here, $\vec{\nu}$ is again the unit outward normal to $\partial\Omega$. Rewriting problem (3.1) into finding $y \in H^1(\Omega)$ such that

$$a(y, \phi) = (Bu + f, \phi) \qquad \forall\, \phi \in H^1(\Omega) \tag{3.2}$$

holds, it follows by the Lax-Milgram lemma 1.1.3, that there exists a unique $y =: \mathcal{G}(Bu) \in H^1(\Omega)$. For our purpose we even have to impose a more regular range space for the operator B, namely that $B \in \mathcal{L}(U; L^2(\Omega))$. Then $\mathcal{G}(Bu)$ belongs to $H^2(\Omega)$ and the following estimate holds true

$$\|\mathcal{G}(Bu)\|_{H^2(\Omega)} \le C\|Bu\|_{L^2(\Omega)} \le C\|u\|_U,$$

with C being a constant depending on f and the domain Ω.

We now recall the general control and state constrained elliptic optimal control problem from [DGH07], which reads

$$\begin{aligned} &\min_{u\in U_{ad}} J(u) = \tfrac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u - u_0\|^2_U \\ &\text{subject to } y = \mathcal{G}(Bu) \text{ and } y_a(x) \le y(x) \le y_b(x) \text{ in } \Omega. \end{aligned} \tag{3.3}$$

Here, $U_{ad} \subseteq U$ denotes the set of admissible controls which is assumed to be a closed and convex subset of the Hilbert space U. Furthermore, we suppose that $\alpha > 0$ and that $y_0 \in H^1(\Omega)$, $u_0 \in U$, $y_a, y_b \in W^{2,\infty}(\Omega)$ are given. We impose a so called *Slater condition*:

Assumption 3.1.1. *There exists some $\tilde{u} \in U_{ad}$ such that*

$$y_a < \mathcal{G}(B\tilde{u}) < y_b \quad \text{in } \bar{\Omega}. \tag{3.4}$$

Since the state constraints form a convex set and the set of admissible controls is closed and convex it is not difficult to establish the existence of a unique solution $u \in U_{ad}$ to problem (3.3). The fact that its corresponding state $y := \mathcal{G}(Bu) \in H^2(\Omega)$ belongs to $C^0(\bar{\Omega})$ by an embedding theorem requires to introduce notation for its dual space in order to characterize the optimal solution. Below the space of RADON measures $\mathcal{M}(\bar{\Omega})$ is identified as the dual space of $C^0(\bar{\Omega})$ endowed with the norm

$$\|\mu\|_{\mathcal{M}(\bar{\Omega})} = \sup_{g \in C^0(\bar{\Omega}), |g| \leq 1} \int_{\bar{\Omega}} g \, \mathrm{d}\mu.$$

Moreover we define the dual pairing

$$\langle \mu, g \rangle := \langle \mu, g \rangle_{\mathcal{M}(\bar{\Omega}), C^0(\bar{\Omega})} := \int_{\bar{\Omega}} g \, \mathrm{d}\mu \quad \forall \, \mu \in \mathcal{M}(\bar{\Omega}) \, \forall \, g \in C^0(\bar{\Omega})$$

and

$$\mu \geq 0 \Longleftrightarrow \langle \mu, g \rangle \geq 0 \quad \forall \, g \in C^0(\bar{\Omega}) \text{ with } g \geq 0.$$

We are now ready to state the first order optimality conditions which can be found in [Cas86, Cas93].

Theorem 3.1.2. *The optimal control problem* (3.3) *has a unique solution $(y, u) \in H^2(\Omega) \times U_{ad}$. Moreover there exist $p \in W^{1,s}(\Omega)$ for all $1 \leq s < d/(d-1)$ and μ_a, $\mu_b \in \mathcal{M}(\bar{\Omega})$ which satisfy for all $\phi \in H^2(\Omega)$ with $\partial_{\vec{\nu}_{\mathcal{A}}} \phi|_{\partial\Omega} = 0$ the following optimality system*

$$y = \mathcal{G}(Bu), \tag{3.5a}$$

$$(p, \mathcal{A}\phi) = (y - y_0, \phi) + \langle \mu_b - \mu_a, \phi \rangle, \tag{3.5b}$$

$$(\alpha(u - u_0) + RB^*p, v - u)_U \geq 0 \quad \forall \, v \in U_{ad}, \tag{3.5c}$$

$$\mu_a \geq 0, \quad y \geq y_a, \quad \langle \mu_a, y - y_a \rangle = 0, \tag{3.5d}$$

$$\mu_b \geq 0, \quad y \leq y_b, \quad \langle \mu_b, y_b - y \rangle = 0. \tag{3.5e}$$

In the above theorem $R : U^* \to U$ denotes the inverse of the RIESZ isomorphism. Later on we are going to concentrate onto structure exploiting GALERKIN schemes for two scenarios of distributed optimal control problems, namely

3.1.1.1. *The purely state constrained problem.*

Problem 3.1.3 (purely state constrained)**.** Consider problem (3.3) with $B = id$, $U = L^2(\Omega) = U_{ad}$.

Under the assumption (see [GH08])

$$\overline{y}_a := \max_{x \in \bar{\Omega}} y_a(x) < \min_{x \in \bar{\Omega}} y_b(x) =: \underline{y}_b$$

our problem satisfies the Slater condition (3.4) with $\tilde{u} := \frac{c}{2}(\overline{y}_a + \underline{y}_b) \in L^2(\Omega)$ since then $y_a < \mathcal{G}(\tilde{u}) = \frac{1}{2}(\overline{y}_a + \underline{y}_b) < y_b$ in $\bar{\Omega}$. The absence of control constraints implies the LAGRANGE multipliers μ_a, μ_b as well as p to be unique. Moreover the variational inequality (3.5c) can be replaced by the equation

$$\alpha(u - u_0) + p = 0 \quad \text{in } L^2(\Omega). \tag{3.6}$$

3.1.1.2. *The control and state constrained problem.*

Problem 3.1.4 (control and state constrained)**.** Consider problem (3.3) with $B = id, U = L^2(\Omega)$ and $U_{ad} = \{u \in L^2(\Omega) : u_a \leq u \leq u_b\}$ for some fixed $u_a, u_b \in \mathbb{R}$ with $u_a < u_b$.

There exist LAGRANGE multipliers λ_a, $\lambda_b \in L^2(\Omega)$ for the control constraints such that the variational inequality (3.5c) can be replaced by

$$\begin{aligned} \alpha(u - u_0) + p + \lambda_b - \lambda_a &= 0 \quad \text{in } L^2(\Omega), \\ \lambda_a \geq 0, \quad u \geq u_a, \quad (\lambda_a, u - u_a) &= 0, \\ \lambda_b \geq 0, \quad u \leq u_b, \quad (\lambda_b, u_b - u) &= 0. \end{aligned}$$

Due to the circumstance that control and state active sets may intersect, the uniqueness of $p, \lambda_a, \lambda_b, \mu_a$ and μ_b can not be guaranteed anymore. We overcome this difficulty and the fact that μ_a, μ_b are measures in general by applying a Moreau-Yosida regularization technique as in [GT09]. This technique penalizes the state constraints $y_a \leq y \leq y_b$ by modifying the objective functional J. The regularized optimal control problem reads

$$\begin{aligned} &\min_{u \in U_{ad}} J^\gamma(u) := J(u) + \tfrac{\gamma}{2} \| \max(0, y_a - y) \|^2_{L^2(\Omega)} + \tfrac{\gamma}{2} \| \max(0, y - y_b) \|^2_{L^2(\Omega)} \\ &\text{subject to } y = \mathcal{G}(u), \end{aligned} \tag{3.7}$$

where $\gamma > 0$ denotes a regularization parameter tending to $+\infty$. The max-expressions in the regularized objective functional J^γ arise from regularizing the indicator function corresponding to the set of admissible states.
Notice that (3.7) is a purely control constrained optimal control problem that has a unique solution $(y^\gamma, u^\gamma) \in H^2(\Omega) \times U_{ad}$. Furthermore, under the Slater condition (3.4), we can prove the existence of LAGRANGE multipliers $(p^\gamma, \lambda_a^\gamma, \lambda_b^\gamma) \in L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega)$ using standard theory of mathematical programming in BANACH spaces [ZK79] such that for all $\phi \in H^2(\Omega)$ with $\partial_{\vec{\nu}_\mathcal{A}} \phi|_{\partial\Omega} = 0$

$$y^\gamma = \mathcal{G}(u^\gamma), \tag{3.8a}$$

$$(p^\gamma, \mathcal{A}\phi) = (y^\gamma - y_0, \phi) + (\mu_b^\gamma - \mu_a^\gamma, \phi)\,, \tag{3.8b}$$

$$\alpha(u^\gamma - u_0) + p^\gamma + \lambda_b^\gamma - \lambda_a^\gamma = 0, \tag{3.8c}$$

$$\lambda_a^\gamma \geq 0, \quad u^\gamma \geq u_a, \quad (\lambda_a^\gamma, u^\gamma - u_a) = 0, \tag{3.8d}$$

$$\lambda_b^\gamma \geq 0, \quad u^\gamma \leq u_b, \quad (\lambda_b^\gamma, u_b - u^\gamma) = 0, \tag{3.8e}$$

where

$$\mu_a^\gamma = \gamma \max(0, y_a - y^\gamma) \quad \text{and} \quad \mu_b^\gamma = \gamma \max(0, y^\gamma - y_b).$$

The convergence of the regularized primal-dual path $\gamma \mapsto (y^\gamma, u^\gamma, p^\gamma, \lambda_a^\gamma, \lambda_b^\gamma)$ is the purpose of the next result whose proof follows from the discussion in [HK09].

Theorem 3.1.5. *Let $\{(y^\gamma, u^\gamma, p^\gamma, \lambda_a^\gamma, \lambda_b^\gamma)\}_{\gamma>0}$ be a sequence of solutions of (3.8). Then there exists a subsequence still denoted by $\{(y^\gamma, u^\gamma, p^\gamma, \lambda_a^\gamma, \lambda_b^\gamma)\}_{\gamma>0}$ and $(p^\star, \lambda_a^\star,$*

$\lambda_b^\star, \mu_a^\star, \mu_b^\star) \in L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega) \times \mathcal{M}(\bar\Omega) \times \mathcal{M}(\bar\Omega)$ *such that*

$$\begin{aligned} y^\gamma &\to y \text{ in } C^0(\bar\Omega),\\ y^\gamma &\rightharpoonup y \text{ in } H^2(\Omega)^*,\\ u^\gamma &\rightharpoonup u \text{ in } L^2(\Omega),\\ \lambda_a^\gamma &\rightharpoonup \lambda_a^\star \text{ in } L^2(\Omega),\\ \lambda_b^\gamma &\rightharpoonup \lambda_b^\star \text{ in } L^2(\Omega),\\ \mu_a^\gamma &\rightharpoonup \mu_a^\star \text{ in } \mathcal{M}(\bar\Omega),\\ \mu_b^\gamma &\rightharpoonup \mu_b^\star \text{ in } \mathcal{M}(\bar\Omega), \end{aligned}$$

as $\gamma \to +\infty$, *with* $(y, u, p^\star, \lambda_a^\star, \lambda_b^\star, \mu_a^\star, \mu_b^\star)$ *being a solution to the optimality system* (3.5).

3.1.2. Finite element discretization. In this section we are going to apply variational discretization to problem (3.3). Since this is already worked out in [DGH07, DH07b] as well as in [HPUU09, Sec. 3.3.1.1] we go through this quickly.

Let $\mathcal{T}_h$ be a quasi-uniform triangulation of Ω with vertices $x_1, \dots, x_m$ and maximum mesh size h as already introduced in Section 1.2. We consider the space of linear finite elements $Y_h := P^1_{c,h}(\mathcal{T}_h)$ with LAGRANGE basis $\{\phi_i \in Y_h : i = 1, \dots, m\}$. If $\partial\Omega$ is not polygonal we allow an appropriate modification for boundary elements (compare also Section 2.1.2). In what follows it is convenient to introduce a discrete approximation of the operator $\mathcal{G}$. For a given function $v \in L^2(\Omega)$ we denote by $z_h = \mathcal{G}_h(v) \in Y_h$ the solution of the discrete NEUMANN problem

$$a(z_h, \phi_h) = (\phi_h + f, \phi_h) \quad \text{for all } \phi_h \in Y_h. \tag{3.9}$$

Problem (3.3) is now approximated by the following sequence of control problems depending on the mesh parameter h:

$$\begin{aligned} &\min_{u \in U_{ad}} J_h(u) = \tfrac{1}{2}\|y_h - y_0\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u - u_{0,h}\|^2_U\\ &\text{subject to } \; y_h = \mathcal{G}_h(Bu) \text{ and } y_a(x_j) \le y_h(x_j) \le y_b(x_j) \text{ for } j = 1, \dots, m. \end{aligned} \tag{3.10}$$

Here, $u_{0,h}$ denotes an approximation to u_0 which is assumed to satisfy

$$\|u_0 - u_{0,h}\|_U \le Ch. \tag{3.11}$$

Problem (3.10) represents a convex infinite-dimensional optimization problem of similar structure as problem (3.3), but with only finitely many equality and inequality constraints for the state, which define a convex set of admissible functions. Again we can apply [Cas93, Thm. 5.2] which yields

Theorem 3.1.6. *There exists* $h_0 > 0$ *such that for* $0 < h \le h_0$ *problem* (3.10) *has a unique solution* $u_h \in U_{ad}$ *with corresponding state* $y_h = \mathcal{G}_h(Bu_h) \in Y_h$. *Further there exist* $\mu_i^a, \mu_i^b \in \mathbb{R}$ $(i = 1, \dots, m)$ *and a function* $p_h \in Y_h$ *such that with* $\mu_{a,h} = \sum_{i=1}^m \mu_i^a \delta_{x_i}$, $\mu_{b,h} = \sum_{i=1}^m \mu_i^b \delta_{x_i}$ *and for all* $\phi_h \in Y_h$ *we have*

$$a(y_h, \phi_h) = (u_h + f, \phi_h), \tag{3.12a}$$

$$a(\phi_h, p_h) = (y_h - y_0, \phi_h) + \langle \mu_h^b - \mu_h^a, \phi_h \rangle, \tag{3.12b}$$

$$(\alpha(u_h - u_{0,h}) + RB^* p_h, v - u_h)_U \ge 0 \quad \forall\, v \in U_{ad}, \tag{3.12c}$$

$$\mu_{a,h} \ge 0, \quad y_h \ge I_h y_a, \quad \langle \mu_h^a, y_h - I_h y_a \rangle = 0, \tag{3.12d}$$

$$\mu_{b,h} \ge 0, \quad y_h \le I_h y_b, \quad \langle \mu_h^b, I_h y_b - y_h \rangle = 0. \tag{3.12e}$$

Here, δ_x denotes the DIRAC measure concentrated at x and $I_h : C^0(\bar{\Omega}) \to Y_h$ is the usual LAGRANGE interpolation operator. It is going to be useful to introduce also the Y_h-functions

$$I_h\mu_{a,h} := \sum_{i=1}^{m} \mu_i^a \phi_i \quad \text{and} \quad I_h\mu_{b,h} := \sum_{i=1}^{m} \mu_i^b \phi_i.$$

Problem 3.1.3 cont. The variational inequality (3.12c) can be replaced by

$$\alpha(u_h - u_{0,h}) + p_h = 0, \tag{3.13}$$

where $u_{0,h} \in Y_h$ is for instance the standard L^2-projection of u_0 onto Y_h. Hence the variational discrete optimal control u_h is itself an element of Y_h, i.e. the optimal discrete solution is discretized implicitly through the optimality condition of the discrete problem. Therefore in (3.10) $U_{ad} = U$ may be replaced by Y_h to obtain the same discrete solution u_h, which results in a finite–dimensional discrete optimization problem instead.

Problem 3.1.4 cont. We again apply variational discretization [Hin05] to problem (3.7) and consider therefore

(3.14)
$$\min_{u\in U_{ad}} J_h^\gamma(u) := J_h(u) + \tfrac{\gamma}{2}\| \max(0, I_h y_a - y_h)\|^2_{L^2(\Omega)} + \tfrac{\gamma}{2}\| \max(0, y_h - I_h y_b)\|^2_{L^2(\Omega)}$$
$$\text{subject to } y_h = \mathcal{G}_h(u).$$

The existence of a solution $u_h^\gamma \in U_{ad}$ of (3.14) as well as LAGRANGE multipliers again follows from standard arguments. The corresponding first order optimality system of (3.14) leads to the variationally discretized counterpart of (3.8). For all $\phi_h \in Y_h$ there holds

$$a(y_h^\gamma, \phi_h) = (u_h^\gamma + f, \phi_h), \tag{3.15a}$$
$$a(\phi_h, p_h^\gamma) = (y_h^\gamma - y_0, \phi_h) + (\mu_{b,h}^\gamma - \mu_{a,h}^\gamma, \phi_h), \tag{3.15b}$$
$$\alpha(u_h^\gamma - u_{0,h}) + p_h^\gamma + \lambda_{b,h}^\gamma - \lambda_{a,h}^\gamma = 0, \tag{3.15c}$$
$$\lambda_{a,h}^\gamma \geq 0, \quad u_h^\gamma \geq u_a, \quad (\lambda_{a,h}^\gamma, u_h^\gamma - u_a) = 0, \tag{3.15d}$$
$$\lambda_{b,h}^\gamma \geq 0, \quad u_h^\gamma \leq u_b, \quad (\lambda_{b,h}^\gamma, u_b - u_h^\gamma) = 0, \tag{3.15e}$$

where y_h^γ, $p_h^\gamma \in Y_h$ and u_h^γ, $\lambda_{a,h}^\gamma$, $\lambda_{b,h}^\gamma \in L^2(\Omega)$. The quantities $\mu_{a,h}^\gamma$ and $\mu_{b,h}^\gamma$ are given by

$$\mu_{a,h}^\gamma = \gamma \max(0, I_h y_a - y_h^\gamma) \quad \text{and} \quad \mu_{b,h}^\gamma = \gamma \max(0, y_h^\gamma - I_h y_b). \tag{3.16}$$

We mention here that (3.14) is a function space optimization problem and the optimal control u_h^γ is not lying in a finite element space in general. However, regarding (3.15), u_h^γ corresponds to the projection of a finite element quantity over the admissible set U_{ad}

$$u_h^\gamma = P_{[u_a,u_b]}\left(-\frac{1}{\alpha} p_h^\gamma + u_{0,h}\right). \tag{3.17}$$

This brings us back into the context of Section 2.2.2.2. Moreover due to the structure of $\mu_{a,h}^\gamma, \mu_{b,h}^\gamma$ in (3.16) also the state active set is not neccessarily resolved by the underlying mesh $\mathcal{T}_h$.

3.1.3. Available a priori error estimates. Finite element analysis for semilinear elliptic control problems in the presence of control and finitely many state constraints is presented by Casas in [Cas02] who proves convergence of a finite element approximation. In [Mey08] Meyer considers a fully discrete strategy to approximate an elliptic control problem with pointwise state and control constraints. A priori error estimates for a purely state constrained elliptic optimal control problem is derived by Deckelnick and Hinze in [DH07a]. Therein they consider optimal control of an homogeneous NEUMANN problem on a smooth bounded domain under pointwise state constraints. By variational discretization of the control and using piecewise linear finite element functions for the state they prove for space dimensions $d = 2, 3$

$$\|u - u_h\|_{L^2(\Omega)} + \|y - y_h\|_{H^1(\Omega)} = \mathcal{O}(h^{2-\frac{d}{2}-\varepsilon})$$

for arbitrary $\varepsilon > 0$. In [DH08] the same problem with additional pointwise state constraints is considered. While the state approximation stays in the space of linear finite elements, the corresponding optimal control is requested to be piecewise constant on every element of the domain partition. Compare also [HPUU09, Thm. 3.15], where

$$\|u - u_h\|_{L^2(\Omega)} + \|y - y_h\|_{H^1(\Omega)} = \begin{cases} \mathcal{O}(h|\log h|), & \text{if } d = 2, \\ \mathcal{O}(h^{\frac{1}{2}}), & \text{if } d = 3 \end{cases}$$

is proven.
By using results from [DH07b] a priori analysis for the variational discretization approach under the presence of control and state constraints is carried out in [Hin08] and [HPUU09]. With a general linear and bounded control operator $B : U \to H^1(\Omega)^*$ at the right hand side of the state equation [HPUU09, Thm. 3.14] reads

$$\|u - u_h\|_U + \|y - y_h\|_{H^1(\Omega)} = \begin{cases} \mathcal{O}(h^{\frac{1}{2}}), & \text{if } d = 2, \\ \mathcal{O}(h^{\frac{1}{4}}), & \text{if } d = 3. \end{cases}$$

If in addition $Bu \in W^{1,s}(\Omega)$ for some $s \in (1, \frac{d}{d-1})$ then

$$\|u - u_h\|_U + \|y - y_h\|_{H^1(\Omega)} = \mathcal{O}(h^{\frac{3}{2}-\frac{d}{2s}}|\log h|^{\frac{1}{2}}).$$

This error estimate also had been numerically validated by the author in [DGH07] on the unit square for $d = 2$ and $B = id$ for a purely pointwise state constrained problem taken from [DH07a, Ex. 4.1].

Under the additional assumptions that $u, u_h \in L^\infty(\Omega)$ with $\|u_h\|_{L^\infty(\Omega)} \leq C$ uniformly bounded in h there further holds for $d = 2, 3$ ([HPUU09, Cor. 3.3])

$$\|u - u_h\|_U + \|y - y_h\|_{H^1(\Omega)} = \mathcal{O}(h|\log h|).$$

Less is known in case of nonlinear state equations. Casas and Mateos in [CM02] consider a full finite element discretization of a semi-linear state-constrained optimal control problem and prove convergence of global optima of the discrete problems to a global optimum of the infinite dimensional problem. The results of [CM02] are remarkable since only low regularity of the nonlinearities is required, as the associated analysis is not based on optimality conditions. An approach to the same problem class using the first order optimality conditions is contained in the work [HM07] of Hinze and Meyer, where stability issues of state constrained semilinear elliptic control against perturbations are discussed and finite element discretization is considered as special class of perturbations. Recently Vierling in [Vie09] considers semilinear elliptic optimal control problems under control and state constraints.

Using variational discretization he proved $\mathcal{O}(h)$ convergence of the $L^2(\Omega)$-control error for a model problem in two space dimensions.

To the best of the authors knowledge there are only a few results on the analysis of parabolic optimal control problems with state constraints. The abstract framework for such problems can be found, e.g. in the book of Fattorini [Fat99, Chap. 10-11] and in the lecture notes [FF91] of Fattorini and Frankowska. Neumann boundary control problems with various constraints on the state, including integral constraints on the gradient of the state are analyzed in [Cas97]. In [NT09] Neitzel and Tröltzsch investigate Lavrentiev regularization of linear-quadratic problems and their convergence to the limit problem as the regularization parameter tends to zero. They consider the same approach for more general control problems including semilinear state equations and control constraints in [NT08]. De los Reyes, Merino, Rehberg, and Tröltzsch derive optimality conditions for elliptic and parabolic optimization problems with state constraints and controls taken from a suitably restricted control space in [dLRMRT08]. Recently in [DH09] Deckelnick and Hinze prove a priori error estimates for a parabolic state constrained problem for two and three space dimensions discretized besides others by piecewise constant time approximations. Lately in the work [MRV10] Meidner, Rannacher, and Vexler proved optimal a priori error estimates for a space-time finite element discretization of a linear parabolic control problem with state constraints pointwise in time.

3.1.4. Numerical realization. In this part of the manuscript we focus onto the development of structure exploiting numerical solution concepts for (3.3) and its variational discretized counterpart (3.10) for both Problem 3.1.3 and Problem 3.1.4.

3.1.4.1. *The purely state constrained problem.* The necessary and sufficient optimality conditions (3.1.6) for Problem 3.1.3 can be rewritten as

$$\begin{aligned}
\mathbf{A}\mathbf{y} &= \mathbf{M}(\mathbf{u}+\mathbf{f}),\\
\mathbf{A}^T\mathbf{p} &= \mathbf{M}(\mathbf{y}-\mathbf{y_0}) + \boldsymbol{\mu}_\mathbf{b} - \boldsymbol{\mu}_\mathbf{a},\\
\mathbf{p} + \alpha(\mathbf{u}-\mathbf{u_0}) &= \mathbf{0},\\
\boldsymbol{\mu}_\mathbf{a} \geq \mathbf{0}, \quad \mathbf{y} \geq \mathbf{y_a}, \quad (\mathbf{y}-\mathbf{y_a})^T\boldsymbol{\mu}_\mathbf{a} &= \mathbf{0},\\
\boldsymbol{\mu}_\mathbf{b} \geq \mathbf{0}, \quad \mathbf{y} \leq \mathbf{y_b}, \quad (\mathbf{y_b}-\mathbf{y})^T\boldsymbol{\mu}_\mathbf{b} &= \mathbf{0},
\end{aligned}$$

where we have used the matrices $\mathbf{M}$, $\mathbf{A}$ and the vector notation for the finite element functions $u_h, y_h, p_h, I_h\mu_{a,h}, I_h\mu_{b,h}, I_h y_a, I_h y_b \in P^1_{c,h}$ already introduced in Section 1.2. Furthermore the L^2–projections of the data f, u_0, y_0 into Y_h are represented by $\mathbf{u_0} = \mathbf{M}^{-1}\hat{\mathbf{u}}_\mathbf{0}$ respectively. We solve the above optimality system by an adapted Moreau-Yosida-based active set strategy taken from [BHHK00, p. 500]. Therefore we have the following

Definition 3.1.7. Let $\mathbf{y}^n, \boldsymbol{\mu}^n_\mathbf{a}, \boldsymbol{\mu}^n_\mathbf{b} \in \mathbb{R}^m$, $c_a, c_b > 0$ be given. Recalling $\bullet = \{1,\dots,m\}$ the active and inactive index sets for the state constraints are

$$\begin{aligned}
\circledcirc^n := \circledcirc &:= \{i \in \bullet : (\mathbf{y}^n - c_a\,\boldsymbol{\mu}^n_\mathbf{a})_i < \mathbf{y_a}_i\} && \text{lower state active index set}\\
\oslash^n := \oslash &:= \{i \in \bullet : (\mathbf{y}^n + c_b\,\boldsymbol{\mu}^n_\mathbf{b})_i > \mathbf{y_b}_i\} && \text{upper state active index set}\\
\circledast^n := \circledast &:= \circledcirc \cup \oslash && \text{state active index set}\\
\ominus^n := \ominus &:= \bullet \setminus \circledast. && \text{state inactive index set}
\end{aligned}$$

Subsequently we assume that $\circledcirc^n \cap \oslash^n = \varnothing$ holds. The modified Moreau-Yosida-based algorithm now reads

Algorithm 3.1.8 ([BHHK00, p. 500])**.**

(1) Initialization: choose $\mathbf{y}^0, \boldsymbol{\mu}_\mathbf{a}^0, \boldsymbol{\mu}_\mathbf{b}^0 \in \mathbb{R}^m$, $c_a, c_b > 0$, and set $n = 0$.
(2) Determine the subsets of active/inactive indices according to Definition 3.1.7.
(3) If $n \geq 1$, $\circledcirc^n = \circledcirc^{n-1}$ and $\circledast^n = \circledast^{n-1}$, then STOP; otherwise go to step (4).
(4) Find $(\mathbf{y}^n, \mathbf{u}^n, \mathbf{p}^n, \boldsymbol{\mu}_\mathbf{a}^n, \boldsymbol{\mu}_\mathbf{b}^n)$ as the solution to

$$\begin{aligned} \mathbf{A}\mathbf{y}^n &= \mathbf{M}(\mathbf{u}^n + \mathbf{f}), \\ \mathbf{A}^T\mathbf{p}^n &= \mathbf{M}(\mathbf{y}^n - \mathbf{y_0}) + \boldsymbol{\mu}_\mathbf{b}^n - \boldsymbol{\mu}_\mathbf{a}^n, \\ \mathbf{p}^n + \alpha(\mathbf{u}^n - \mathbf{u_0}) &= \mathbf{0}, \\ \mathbf{y}_i^n = \mathbf{y}_{\mathbf{a}i} \text{ for } i \in \circledcirc^n, \qquad & \boldsymbol{\mu}_{\mathbf{a}i}^n = 0 \text{ for } i \in \ominus^n \cup \circledast^n, \\ \mathbf{y}_i^n = \mathbf{y}_{\mathbf{b}i} \text{ for } i \in \circledast^n, \qquad & \boldsymbol{\mu}_{\mathbf{b}i}^n = 0 \text{ for } i \in \ominus^n \cup \circledcirc^n. \end{aligned}$$

(5) Set $n = n + 1$ and go to step (2).

For each iteration there are basically two ways to solve the system given in step (4), see also [NW06, Sec. 16.2]. One possibility is to directly solve an indefinite, symmetric sparse system with $m + \operatorname{card}(\ominus^n)$ unknowns. The other one is to iteratively solve by SCHUR complement a positive-definite, symmetric, dense system of size $\operatorname{card}(\ominus^n)$ with a *preconditioned conjugate gradient* (PCG)-algorithm. Both methods will be described in the following, while we again make use of a blockwise decomposition of a matrix as in Definition 1.2.12.

Direct approach. Since in our discussion $(\mathbf{A}^T)_{\ominus\bullet}$ would appear, it is more convenient to introduce $\hat{\mathbf{A}} := \mathbf{A}^T$. The solution of the linear system in step (4) of Algorithm 3.1.8 is obtained by solving

$$\begin{bmatrix} \mathbf{M} & -\mathbf{A}_{\bullet\ominus} \\ -\hat{\mathbf{A}}_{\ominus\bullet} & -\alpha^{-1}\mathbf{M}_\ominus \end{bmatrix} \begin{bmatrix} \mathbf{u}^n \\ \mathbf{y}_\ominus^n \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1^n \\ \alpha^{-1}\mathbf{r}_2^n \end{bmatrix}, \tag{3.18}$$

with

$$\begin{bmatrix} \mathbf{r}_1^n \\ \mathbf{r}_2^n \end{bmatrix} := \begin{bmatrix} \mathbf{A}_{\bullet\circledcirc}\mathbf{y}_{\mathbf{a}\circledcirc} + \mathbf{A}_{\bullet\circledast}\mathbf{y}_{\mathbf{b}\circledast} - \mathbf{M}\mathbf{f} \\ \mathbf{M}_{\ominus\circledcirc}\mathbf{y}_{\mathbf{a}\circledcirc} + \mathbf{M}_{\ominus\circledast}\mathbf{y}_{\mathbf{b}\circledast} - \mathbf{M}_{\ominus\bullet}\mathbf{y_0} - \alpha\hat{\mathbf{A}}_{\ominus\bullet}\mathbf{u_0} \end{bmatrix}.$$

The further quantities can be determined by

$$\mathbf{y}_{\circledcirc}^n = \mathbf{y}_{\mathbf{a}\circledcirc}, \tag{3.19a}$$

$$\mathbf{y}_{\circledast}^n = \mathbf{y}_{\mathbf{b}\circledast}, \tag{3.19b}$$

$$\mathbf{p}^n = -\alpha(\mathbf{u}^n - \mathbf{u_0}), \tag{3.19c}$$

$$\boldsymbol{\mu}_\mathbf{a}^n = \left(\mathbf{A}^T\mathbf{p}^n - \mathbf{M}(\mathbf{y}^n - \mathbf{y_0})\right)_-, \tag{3.19d}$$

$$\boldsymbol{\mu}_\mathbf{b}^n = \left(\mathbf{A}^T\mathbf{p}^n - \mathbf{M}(\mathbf{y}^n - \mathbf{y_0})\right)_+, \tag{3.19e}$$

where $(\mathbf{v})_\pm$ denotes the non-negative positive/negative part of each component of the vector $\mathbf{v}$. With

$$\mathbf{C} := \begin{bmatrix} \mathbf{M} & -\mathbf{A} \\ -\mathbf{A}^T & -\alpha^{-1}\mathbf{M} \end{bmatrix}$$

and $\tilde{\ominus}^n := \tilde{\ominus} := (\bullet, \ominus)$ the sparse symmetric indefinite system matrix in (3.18) can be written as $\mathbf{C}_{\tilde{\ominus}}$. A direct solution of (3.18) by a *symmetric indefinite factorization* of the form $\mathbf{C}_{\tilde{\ominus}} = \mathbf{L}_{\tilde{\ominus}}\mathbf{D}_{\tilde{\ominus}}\mathbf{L}_{\tilde{\ominus}}^T$ is most suited and works well for small numbers of unknowns. Here $\mathbf{L}_{\tilde{\ominus}}$ is a lower triangular matrix with ones on the diagonal and $\mathbf{D}_{\tilde{\ominus}}$ is block diagonal with diagonal blocks of dimension 1 or 2. In order to reduce

fill-in in the sparse factor $\mathbf{L}_{\tilde{\Theta}}$ usually permutations of rows and columns in $\mathbf{C}_{\tilde{\Theta}}$ are considered which we omit for our purpose.

In the literature sparse update strategies of the factors for rank-1 modifications and fixed size of the underlying linear system are available (see for instance [DH99]). These techniques are not offhand applicable since our system size usually changes its dimension from iterate n to $n+1$ due to the change of the inactive set. However considering active set strategies for PDE-constrained optimization problems under additional state constraints one observes a change of the inactive set in the domain from one iterate to the next basically near its boundary. Speaken in the context of equation (3.18) compared to the overall system size only a few new equations arise while just a few others disappear. Therefore it is desirable to efficiently update the factors by a routine

$$\left[\tilde{\mathbf{L}}_{\tilde{\Theta}^{n+1}} \quad \tilde{\mathbf{D}}_{\tilde{\Theta}^{n+1}}\right] = \texttt{LD\_update\_indexchange}(\mathbf{C}, \mathbf{L}_{\tilde{\Theta}^n}, \mathbf{D}_{\tilde{\Theta}^n}, \tilde{\Theta}^n, \tilde{\Theta}^{n+1}).$$

We are going to indicate how this functionality can be implemented for a similar update of a CHOLESKY-factor $\mathbf{R}_{\Theta^n}$ within the next paragraph and the appendix B. It goes without saying that once the new factors are obtained with low effort basically operating on the difference of both involved inactive sets, we expect to be much more efficient than recompute the factors in every iteration from scratch. In fact once the factors are determined for a guess of the inactive index set (for instance from a previous mesh level) the suggested routine should provide a massive speed up for the overall CPU-time.

Since one possibly attains memory bounds when saving $\mathbf{L}_{\tilde{\Theta}}$ for larger systems we even suggest a more structure-exploiting solution concept in the following.

Iterative approach. If one eliminates $\mathbf{u}^n$ in (3.18) we only need to solve the even smaller system

$$(\mathbf{M}_\Theta + \alpha \hat{\mathbf{A}}_{\Theta\bullet}\mathbf{M}^{-1}\mathbf{A}_{\bullet\Theta})\mathbf{y}^n_\Theta = \mathbf{r}_\Theta := -\mathbf{r}^n_2 - \alpha\hat{\mathbf{A}}_{\Theta\bullet}\mathbf{M}^{-1}\mathbf{r}^n_1 \tag{3.20}$$

on the inactive set Θ. The control-vector $\mathbf{u}^n$ is then given by

$$\mathbf{u}^n = \mathbf{M}^{-1}(\mathbf{r}^n_1 + \mathbf{A}_{\bullet\Theta}\mathbf{y}^n_\Theta).$$

Again the equations (3.19) can be used to compute the still missing quantities. The matrix $\mathbf{C}_\Theta := \mathbf{M}_\Theta + \alpha\hat{\mathbf{A}}_{\Theta\bullet}\mathbf{M}^{-1}\mathbf{A}_{\bullet\Theta}$ is symmetric, positive-definite and dense. In order to solve the system efficiently we are applying a PCG method. Since solving with the well-conditioned, positive-definite, symmetric matrix $\mathbf{M}$ is needed for providing $\mathbf{r}_\Theta$, for vector products with the matrix $\mathbf{C}_\Theta$ and for computing $\mathbf{u}^n$ in each iteration, $\mathbf{M}$ is factorized into $\mathbf{M} = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R}$ is an upper triangular sparse matrix. Then the inverse of $\mathbf{M}$ is given by $\mathbf{M}^{-1} = \mathbf{R}\mathbf{R}^T$. In order to solve systems with matrix $\mathbf{M}$ one alternatively could only compute an incomplete CHOLESKY-factorization of $\mathbf{M}$ and use these factors for a preconditioned conjugate gradient method as well.

Since $\mathbf{C}_\Theta$ is ill-conditioned for large m, we need a suitable preconditioner for $\mathbf{C}_\Theta$. Let therefore $\tilde{\mathbf{M}}$ denote the lumped mass-matrix given in (1.21). Because $\tilde{\mathbf{M}}^{-1}$ is a diagonal matrix we compute the sparse, symmetric and positive-definite matrix

$$\mathbf{P} := \mathbf{M} + \alpha\mathbf{A}^T\tilde{\mathbf{M}}^{-1}\mathbf{A}$$

only once and provide $\mathbf{P}_\Theta$ as preconditioner in every iteration. The computation of $\mathbf{P}$ is still expensive due to a sparse matrix–matrix product. Moreover the solution with $\mathbf{P}_\Theta$ in every iteration and every CG-iteration is the most expensive step. The

difficulty again consists in the changing dimensions of the inactive set $\ominus^n$ during the iteration. We therefore again suggest an efficient update routine

$$\tilde{\mathbf{R}}_{\ominus^{n+1}} = \texttt{R\_update\_indexchange}(\mathbf{P}, \mathbf{R}_{\ominus^n}, \ominus^n, \ominus^{n+1}), \tag{3.21}$$

where $\mathbf{R}_{\ominus^n}$ satisfies $\mathbf{P}_{\ominus^n} = \mathbf{R}_{\ominus^n}\mathbf{R}^T_{\ominus^n}$ respectively and $\mathbf{R}_{\ominus^n}, \tilde{\mathbf{R}}_{\ominus^{n+1}}$ are upper triangular sparse matrices. For a realization of such a tailored CHOLESKY-factor update in context to an active set strategy we refer to the appendix B.

A blockwise investigation with respect to the active and inactive index sets $\circledcirc$ and $\ominus$ leads to the identity

$$\mathbf{P}_\ominus = \mathbf{M}_\ominus + \alpha \left(\hat{\mathbf{A}}_\ominus \tilde{\mathbf{M}}^{-1}_\ominus \mathbf{A}_\ominus + \hat{\mathbf{A}}_{\ominus\circledcirc} \tilde{\mathbf{M}}^{-1}_\circledcirc \mathbf{A}_{\circledcirc\ominus} \right),$$

where we have used the fact that $\tilde{\mathbf{M}}^{-1}$ is diagonal and $\circledcirc \cap \ominus = \varnothing$ holds. Again attaining memory bounds can be compensated by using simpler preconditioners and calculating more CG-iterates. For small α the systems condition becomes better and obviously $\mathbf{M}_\ominus$ itself should be a good preconditioner. For large α then we suggest the simpler preconditioner

$$\tilde{\mathbf{P}}_\ominus := \alpha \hat{\mathbf{A}}_\ominus \tilde{\mathbf{M}}^{-1}_\ominus \mathbf{A}_\ominus.$$

Now there is no need to compute and save $\mathbf{P}$. The new overhead consists in solving with the matrix $\hat{\mathbf{A}}_\ominus$ and $\mathbf{A}_\ominus$. We want to emphasize that $\mathbf{A}$ has less then a fifth of non-zero-entries compared to $\mathbf{P}$. Therefore we suggest to provide the CHOLESKY-factors of the matrix $\mathbf{A}_\ominus$ in each iteration similar as in (3.21).

Further remarks. Looking at step (1) of Algorithm 3.1.8 one notices some degrees of freedom concerning the choice of $\mathbf{y}^0, \boldsymbol{\mu}^0_\mathbf{a}, \boldsymbol{\mu}^0_\mathbf{b} \in \mathbb{R}^m$ and $c_a, c_b > 0$. During a refinement process we interpolate the optimal control $\mathbf{u}$ and the multipliers $\boldsymbol{\mu}_\mathbf{a}$ and $\boldsymbol{\mu}_\mathbf{b}$ from the last mesh to the current one and take them as $\mathbf{u}^0, \boldsymbol{\mu}^0_\mathbf{a}$ and $\boldsymbol{\mu}^0_\mathbf{b}$. For $\mathbf{y}^0$ we compute $\mathbf{y}^0 = \mathbf{A}^{-1}\mathbf{M}(\mathbf{u}^0 + \mathbf{f})$. A good choice of $c_a, c_b > 0$ is crucial for fast termination of the active-set strategy. For too large c_a and c_b the algorithm oscillates or even cycles appear. For too small parameters needless iterates were computed. Until now no recipe is given us for the best choice of these parameters.

Furthermore we want to mention, that the PCG-tolerance for solving (3.20) is coupled to the termination criterion of the algorithm. Starting with $tol = 10^{-8}$ this is decreased by the factor 0.01 whenever the active sets do not change anymore until $tol < 10^{-14}$. For the case that systems with $\mathbf{M}$ are solved by PCG itself, the tolerance of $\frac{tol}{10}$ is taken.

For the factorization of $\mathbf{M}$ to all matrices and vectors a symmetric approximate minimum degree permutation is applied in order to reduce the number of nonzero elements and hence the memory costs.

3.1.4.2. *The control and state constrained problem.* Now we consider a structure exploiting solution concept for Problem 3.1.4 as it is also part of the work [GT09]. In what follows we extend the algorithm prescribed in [DH07b] to the regularized problem (3.14). The special structure of u^γ_h in (3.17) allows a matricial representation of (3.15) with the techniques from Section 2.2.2.2 and 3.1.4.1. More precisely with $p_h := p^\gamma_h$ in Definition 2.2.10 we obtain the control active subsets $\circledcirc, \circledcirc \subset \bullet = \Omega$. They provide the additive split from Definition 2.2.11 for the mass-matrix $\mathbf{M}$. Additionally, due to penalization of the state constraints and the appearance of the pointwise max-operator in (3.16) we introduce further domain subsets in

Definition 3.1.9. For given $y_h^\gamma \in Y_h$ we define the subsets

$$\begin{aligned}
\CIRCLE &:= \Omega \\
\circledcirc &:= \{x \in \CIRCLE : y_h^\gamma(x) \le I_h y_a(x)\} && \text{lower state active set} \\
\circledcirc &:= \{x \in \CIRCLE : y_h^\gamma(x) \ge I_h y_b(x)\} && \text{upper state active set} \\
\circledcirc &:= \circledcirc \cup \circledcirc && \text{state active set} \\
\ominus &:= \CIRCLE \setminus \circledcirc. && \text{state inactive set}
\end{aligned}$$

With this domain decomposition at hand we analogously can introduce

$$\mathbf{M}_{\circledcirc} := \left[\int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m \quad \text{and} \quad \mathbf{M}_{\circledcirc} := \left[\int_{\circledcirc} \phi_i \phi_j\right]_{i,j=1}^m$$

as in Definition 2.2.11. It is clear that these matrices can efficiently be assembled with the same routine `assem_mass` from Section 2.2.2.2.

We are ready to present the matricial form of (3.15)

$$\mathbf{A}\mathbf{y}^\gamma = \hat{\mathbf{u}}^\gamma + \mathbf{M}\mathbf{f}, \tag{3.22a}$$

$$\mathbf{A}^T\mathbf{p}^\gamma = \mathbf{M}(\mathbf{y}^\gamma - \mathbf{y_0}) + \gamma \mathbf{M}_{\circledcirc}(\mathbf{y}^\gamma - \mathbf{y_b}) - \gamma \mathbf{M}_{\circledcirc}(\mathbf{y_a} - \mathbf{y}^\gamma), \tag{3.22b}$$

$$\hat{\mathbf{u}}^\gamma = \mathbf{M}_{\circledcirc}\mathbf{u_a} + \mathbf{M}_{\circledcirc}\mathbf{u_b} + \mathbf{M}_{\ominus}\left(-\frac{1}{\alpha}\mathbf{p}^\gamma + \mathbf{u_0}\right). \tag{3.22c}$$

We reduce (3.22) to a nonlinear system in $\mathbf{x}^\gamma = [\mathbf{y}^\gamma; \mathbf{p}^\gamma]$ as the following

$$G^\gamma(\mathbf{x}^\gamma) := \begin{bmatrix} \mathbf{A}\mathbf{y}^\gamma - \mathbf{M}_{\circledcirc}\mathbf{u_a} - \mathbf{M}_{\circledcirc}\mathbf{u_b} - \mathbf{M}_{\ominus}(-\frac{1}{\alpha}\mathbf{p}^\gamma + \mathbf{u_0}) - \mathbf{M}\mathbf{f} \\ \mathbf{A}^T\mathbf{p}^\gamma - \mathbf{M}(\mathbf{y}^\gamma - \mathbf{y_0}) - \gamma \mathbf{M}_{\circledcirc}(\mathbf{y}^\gamma - \mathbf{y_b}) + \gamma \mathbf{M}_{\circledcirc}(\mathbf{y_a} - \mathbf{y}^\gamma) \end{bmatrix} = \mathbf{0}. \tag{3.23}$$

Notice that, due to the presence of max-operations, G^γ is not FRÉCHET-differentiable and a classical NEWTON method can not be applied to solve (3.23). Nevertheless, a generalized Jacobian can be defined for $G^\gamma(\mathbf{x})$ with $\mathbf{x} = [\mathbf{y}; \mathbf{p}] \in \mathbb{R}^{2m}$ by

$$DG^\gamma(\mathbf{x}) := \begin{bmatrix} \mathbf{A} & \frac{1}{\alpha}\mathbf{M}_{\ominus} \\ -(\mathbf{M} + \gamma \mathbf{M}_{\circledcirc} + \gamma \mathbf{M}_{\circledcirc}) & \mathbf{A}^T \end{bmatrix}.$$

Therefore to solve (3.23) we therefore perform semi-smooth NEWTON iterations (see for instance [QS93, Mif77, HIK03])

$$\mathbf{x}_{n+1} = \mathbf{x}_n - DG^\gamma(\mathbf{x}_n)^{-1} G^\gamma(\mathbf{x}_n) \quad \text{for } n = 0, 1, \ldots \tag{3.24}$$

until some stopping criterion is satisfied. With an approximate solution of $G^\gamma(\mathbf{x}^\gamma) = \mathbf{0}$ at hand we recover the $L^2(\Omega)$-function u_h^γ via (3.17).

Proposition 3.1.10. *The semi-smooth* NEWTON *iteration* (3.24) *is well defined. The sequence* $(\mathbf{x}_n)_{n\in\mathbb{N}}$ *generated by* (3.24) *converges to a solution* $\mathbf{x}^\gamma := [\mathbf{y}^\gamma; \mathbf{p}^\gamma]$ *of* (3.23) *provided that* $\|\mathbf{x}^\gamma - \mathbf{x}_0\|$ *is small enough. Here* $\|\cdot\|$ *denotes a norm in* $\mathbb{R}^d$ *(for instance* $\|\cdot\|_1$*) and the resulting induced matrix norm.*

PROOF. In order to show this proposition it suffices to prove that DG^γ has got an inverse which is bounded in some neighborhood of $\mathbf{x}^\gamma$.
For an arbitrary chosen $\mathbf{x} := [\mathbf{y}; \mathbf{p}] \in \mathbb{R}^{2m}$, we know that $\mathbf{C} := \mathbf{M} + \gamma \mathbf{M}_{\circledcirc} + \gamma \mathbf{M}_{\circledcirc}$ is symmetric and positive definite, $\mathbf{A}$ is positive definite and $\frac{1}{\alpha}\mathbf{M}_{\ominus}$ is symmetric positive semi-definite. A SCHUR complement of the matrix block $DG^\gamma(\mathbf{x})$ reads

$$\mathbf{S} := \mathbf{A} + \frac{1}{\alpha}\mathbf{M}_{\ominus}\mathbf{A}^{-T}\mathbf{C},$$

which can be written as

$$\mathbf{S} = \mathbf{A}(\mathbf{I} + \frac{1}{\alpha}\mathbf{A}^{-1}\mathbf{M}_{\ominus}\mathbf{A}^{-T}\mathbf{C}). \tag{3.25}$$

From ([HJ85, Thm. 7.6.3]) it follows that the product of a real symmetric positive definite matrix and a real symmetric positive semi-definite one is a positive semi-definite matrix (which is not necessary symmetric). Therefore $\mathbf{A}^{-1}\mathbf{M}_{\textcircled{1}}\mathbf{A}^{-T}\mathbf{C}$ is a positive semi-definite matrix and, from (3.25), $\mathbf{S}$ is invertible. Moreover, for a given $\mathbf{r} = [\mathbf{r}_1; \mathbf{r}_2] \in \mathbb{R}^{2m}$, the solution $\mathbf{d} = [\mathbf{d}_1; \mathbf{d}_2] \in \mathbb{R}^{2m}$ to the linear system

$$DG^{\gamma}(\mathbf{x})\mathbf{d} = \mathbf{r}$$

can be computed using

$$\mathbf{d}_1 = \mathbf{S}^{-1}\mathbf{r}_1 - \frac{1}{\alpha}\mathbf{S}^{-1}\mathbf{M}_{\textcircled{1}}\mathbf{A}^{-T}\mathbf{r}_2, \tag{3.26a}$$

$$\mathbf{d}_2 = \mathbf{A}^{-T}\mathbf{r}_2 + \mathbf{A}^{-T}\mathbf{C}\mathbf{d}_1, \tag{3.26b}$$

where

$$\mathbf{S}^{-1} = (\mathbf{I} + \frac{1}{\alpha}\mathbf{A}^{-1}\mathbf{M}_{\textcircled{1}}\mathbf{A}^{-T}\mathbf{C})^{-1}\mathbf{A}^{-1}.$$

Notice that

$$\|\mathbf{S}^{-1}\| \leq C\|\mathbf{A}^{-1}\|, \tag{3.27}$$

$$\max\left(\|\mathbf{M}_{\textcircled{1}}\|, \|\mathbf{M}_{\textcircled{2}}\|, \|\mathbf{M}_{\textcircled{3}}\|\right) \leq C\|\mathbf{M}\|, \tag{3.28}$$

with C being a generic positive constant not depending on $\mathbf{x}$. Consequently, from (3.26), (3.27), and (3.28) we infer that $\|DG^{\gamma}(\mathbf{x})^{-1}\|$ is bounded independently of $\mathbf{x}$ which completes the proof of this proposition. □

3.2. A posteriori error analysis

After the a priori analysis of state constrained optimal control problems together with the development of structure exploiting solution concepts for the underlying KKT-equations we now continue with its a posteriori analysis. At the first place we give an overview about available literature related to adaptive concepts for state and control constrained problems in Section 3.2.1. Secondly for the purely state constrained Problem 3.1.3 a goal-oriented a posteriori error estimator is developed in Section 3.2.2. Its extension towards control and state constraints is matter of issue in Section 3.2.3.

3.2.1. Extension of the dual weighted residual method. Let us briefly comment on adaptive approaches in PDE-constrained optimization. For problems with neither constraints on controls nor on states an excellent overview of the DWR approach is contained in [BR01] and in the book [BR03]. The main idea is to represent the error in a quantity of interest or goal E by a locally weighted sum

$$E(u) - E(u_h) = \sum_{T \in \mathcal{T}_h} \rho_T \omega_T.$$

Here roughly speaken ρ_T plays the role of a local residual interlinked to the involved equations from the optimization problem and ω_T are the local weights. The latter ones reflect the sensitivity of the local residual ρ_T to the overall error. The specialty in the above error representation is the absence of unknown constants and potences of the mesh-parameter h as they appear in residual based error estimators. Moreover it naturally turns out, that the weights ρ_T are residuals stemming from dual equations of the optimization problem. This explains the name for the DWR-method having its roots already in the papers [BR96] and [BKR00].

Residual-type estimators for elliptic optimal control problems also dealing with constraints on the control are discussed in [LY01] and [Sch06]. A posteriori analysis

of an adaptive algorithm for elliptic control problems with constraints on the control is presented in [HHIK08]. Let us further mention the recent work [KRS10] with a new convergence proof of AFEM for control constrained optimal control problems. A posteriori error estimators of residual-type for mixed control-state constrained problems are derived in [HK08]. Residual-type a posteriori error estimators for state constrained distributed optimal control problems for second order elliptic boundary value problems are discussed in [HK07].

An extension of the DWR concept to elliptic PDE-constrained optimization problems in the presence of control constraints is proposed in [HH08, VW08]. Goal-oriented adaptive approaches for elliptic PDE-constrained optimization problems in the presence of state constraints is the topic of the authors diploma thesis [Gün06]. This work developed further towards [GH08], which is to the best of the authors knowledge the first contribution concerning the extension of the DWR-method to state constraints. This paper basically builds the content of the next Section 3.2.2. In the meantime further literature [BV09, HH09b] of similar topic appeared.

Within the framework of function space algorithms, goal-oriented adaptive algorithms based on Lavrentiev regularization and interior point approaches are proposed in [HH09c] and [Wol08] respectively.

Let us emphasize the authors contribution in the work [SG09]. Therein an interior point method in function space for PDE-constrained optimal control problems with state constraints is considered. The emphasis is on the construction and analysis of an algorithm that integrates a NEWTON path-following method with adaptive grid refinement. The algorithm consists of three nested loops: a path-following scheme, a NEWTON corrector, and the approximate solution of an operator equation. The crucial point is that the two outer loops are performed inexactly in function space. Discretization only takes place in the innermost loop such that the discretization error (considered as perturbation in function space) of each NEWTON step is controlled by adaptive grid refinement. As a consequence the algorithm allows to perform most of the required NEWTON steps on coarse grids, such that the overall computational time is dominated by the last few steps.

The extension of the DWR-method for parabolic optimization problems is addressed in the works [MV07, SV08].

3.2.2. The purely state constrained problem. In this section we present the results from [GH08]. We develop a posteriori error estimators for the purely state constrained Problem 3.1.3. For their construction we extend the DWR concept to elliptic optimal control problems with state constraints, where the refinement goal consists in the construction of finite element meshes which allow to resolve well the value J of the cost functional as quantity of interest.

Until the end of this chapter we make the following

Assumption 3.2.1. $u_0 = u_{0,h}$.

As a consequence of this assumption it holds $J = J_h$. In order to distinguish between $J(u)$ and $J_h(u_h)$ later on, we dissociate ourselves from the reduced objective functional and write $J(y,u) := J(\mathcal{G}(u),u)$ and $J(y_h,u_h) := J_h(\mathcal{G}_h(u_h),u_h)$ instead. We mention here that the previous assumption is fulfilled by affine linear functions or, more precisely, by a piecewise linear function over the coarsest mesh in refinement processes which is not restrictive from practical point of view. Indeed, in contrast to y_0, u_0 is not a desired control but a background control. In many applications, it is corresponding to the result of trial and error experiments performed with a small

number of degrees of freedom. Including more general desired controls u_0 would lead to additional weighted data oscillation quantities $(u_0 - u_{0,h}, \cdot)$ in the following error representation. For residual type a posteriori estimators this was done in [HK08].

The main analytical result of this work consists in proving an error representation for the values of the cost functional J of the form

$$J(y,u) - J(y_h,u_h) = \frac{1}{2}\left(\rho^y(p - i_h p) + \rho^p(y - i_h y) + \langle \mu + \mu_h, y_h - y\rangle\right),$$

where ρ^p, ρ^y denote the dual and primal residual of the underlying PDE and i_h denotes an appropriate interpolation operator, compare (3.29). To anticipate discussion let us point out two basic facts of our approach;

- Under common assumptions no residual ρ^u associated to the optimality conditions (3.6),(3.13) appears in our approach. This is due to the fact that we do not discretize controls explicitly. This result remains valid in the case of additional control constraints, see also Section 3.2.3.
- Differences of multipliers do not occur in our concept. This is of particular importance for multipliers associated to state constraints, since these may be represented by measures. As a consequence there is no need to construct a computable approximation to μ which carries more information than μ_h. In fact we use $\mu \equiv \mu_h$ in a first numerical approach.

Next we specify the local error indicators and test their effectivity indices by means of a numerical example in Section 3.2.2.2.

3.2.2.1. *Local error indicators.* Let us abbreviate

$$\mu := \mu_b - \mu_a, \quad \mu_h := \mu_{b,h} - \mu_{a,h}.$$

Following [BR01] we introduce the dual, control and primal residual functionals determined by the discrete solution $y_h, u_h, p_h, \mu_{a,h}$ and $\mu_{b,h}$ of (3.12b)-(3.12e) by

$$\begin{aligned}
\rho^p(\cdot) &:= J_y(y_h,u_h)(\cdot) - a(\cdot,p_h) + \langle \mu_h, \cdot\rangle,\\
\rho^u(\cdot) &:= J_u(y_h,u_h)(\cdot) + (\cdot,p_h) \text{ and}\\
\rho^y(\cdot) &:= -a(y_h,\cdot) + (u_h,\cdot).
\end{aligned}$$

In addition we introduce the error stemming from the complementarity conditions (3.5d), (3.5e), (3.12d) and (3.12e), respectively by

$$e^\mu(y) := \langle \mu + \mu_h, y_h - y\rangle.$$

It follows from (3.12c) that $\rho^u(\cdot) \equiv 0$. This is due to the fact that we do not discretize the control, so that the discrete structure of the solution u_h of problem (3.10) is induced by the optimality condition (3.12c).

We are now in the position to prove the analogue to [Ran05, Thm. 1] for the state constrained case.

Theorem 3.2.2 (Compare [Ran05, Thm. 1] and [BR01])**.** *There holds the error representation*

$$J(y,u) - J(y_h,u_h) = \frac{1}{2}\rho^p(y - i_h y) + \frac{1}{2}\rho^y(p - i_h p) + \frac{1}{2}e^\mu(y) \tag{3.29}$$

with arbitrary quasi–interpolants $i_h y$ and $i_h p \in Y_h$.

Proof. It follows from (3.6) and (3.13) that

$$u_h - u = \frac{1}{\alpha}(p - p_h) \tag{3.30}$$

holds. This yields

$$\begin{aligned}
&2(J(y_h,u_h)-J(y,u))\\
&\quad=(y_h-y_0,y_h-i_hy)+(y_h-y_0,i_hy-y)+(y-y_0,y_h-y)\\
&\qquad+\alpha(u_h-u_0,\tfrac{1}{\alpha}(p-p_h))-\alpha(u-u_0,\tfrac{1}{\alpha}(p_h-p))\\
&\quad=J_y(y_h,u_h)(y_h-i_hy)+(y_h-y_0,i_hy-y)+J_y(y,u)(y_h-y)\\
&\qquad-(u_h,p_h-p)+2(u_0,p_h-p)-(u,p_h-p).
\end{aligned}$$

Since by (3.30), (3.9)

$$(u_0,p_h-p)=-(u_h,p-i_hp)-a(y_h,i_hp)+a(y,p_h)$$

holds, we obtain

$$\begin{aligned}
&2(J(y_h,u_h)-J(y,u))\\
&\quad=a(y_h-i_hy,p_h)-\langle\mu_h,y_h-i_hy\rangle+(y_h-y_0,i_hy-y)\\
&\qquad+a(y_h-y,p)-\langle\mu,y_h-y\rangle\\
&\qquad-(u_h,p_h-p)-2(u_h,p-i_hp)-2a(y_h,i_hp)+2a(y,p_h)-(u,p_h-p)\\
&\quad=a(y_h-i_hy,p_h)-\langle\mu_h,y-i_hy\rangle+(y_h-y_0,i_hy-y)+a(y_h-y,p)\\
&\qquad-(u_h,p_h-p)-2(u_h,p-i_hp)-2a(y_h,i_hp)+2a(y,p_h)-(u,p_h-p)\\
&\qquad-e^\mu(y)\\
&\quad=[a(y_h,p_h)-(u_h,p_h)]+a(y,p_h)-a(i_hy,p_h)-\langle\mu_h,y-i_hy\rangle\\
&\qquad-J_y(y_h,u_h)(y-i_hy)+[(u,p)-a(y,p)]+[(u_h,i_hp)-a(y_h,i_hp)]\\
&\qquad+[a(y,p_h)-(u,p_h)]+a(y_h,p)-a(y_h,i_hp)-(u_h,p-i_hp)-e^\mu(y),
\end{aligned}$$

where we have used

$$\langle\mu_h,y_h-i_hy\rangle+\langle\mu,y_h-y\rangle=\langle\mu_h,y-i_hy\rangle+e^\mu(y).$$

Since the terms within the squared brackets vanish, we finally obtain

$$\begin{aligned}
&2(J(y_h,u_h)-J(y,u))\\
&\quad=-J_y(y_h,u_h)(y-i_hy)+a(y-i_hy,p_h)-\langle\mu_h,y-i_hy\rangle\\
&\qquad+a(y_h,p-i_hp)-(u_h,p-i_hp)-e^\mu(y)\\
&\quad=-\rho^p(y-i_hy)-\rho^y(p-i_hp)-e^\mu(y).
\end{aligned}$$

□

Remark 3.2.3. We emphasize that no differences of the multipliers μ,μ_h appear in this error representation. We exploit this fact in the definition of the error estimators, since it now is meaningful to replace the continuous multiplier μ by their discrete counterpart μ_h. This idea is different from the one used in [VW08] to construct an a posteriori error estimator for control constrained optimization problems, and takes care of the fact that a better approximation to $\mu\in\mathcal{M}(\bar\Omega)$ than μ_h can hardly be constructed using only the values $\mu_1^{a,b},\dots,\mu_m^{a,b}$ of Theorem 3.1.6.

The goal now consists in deriving an a posteriori error representation of the form

$$J(y,u)-J(y_h,u_h)\approx\frac{1}{2}\sum_{T\in\mathcal{T}_h}\rho_T^p((y-i_hy)_{|T})+\rho_T^y((p-i_hp)_{|T})+e_T^\mu(y_{|T}),$$

and in a final step to replace continuous quantities by computable analogues. To begin with let us first consider $\rho^y(p-i_hp)$. It follows from the definition of the

bilinear form a that

$$\begin{aligned}\rho^y(p - i_h p) &= -a(y_h, p - i_h p) + (u_h, p - i_h p)\\ &= \sum_{T\in\mathcal{T}_h} \int_T \Big(\sum_{i,j=1}^{d} -a_{ij}(y_h)_{x_i}(p - i_h p)_{x_j}\\ &\quad - \sum_{i=1}^{d} b_i(y_h)_{x_i}(p - i_h p) - c\, y_h(p - i_h p) + u_h(p - i_h p)\Big),\end{aligned}$$

so that we may define

$$\begin{aligned}\rho_T^y((p - i_h p)_{|_T}) := \int_T \Big(\sum_{i,j=1}^{d} -a_{ij}(y_h)_{x_i}(p - i_h p)_{x_j} - \sum_{i=1}^{d} b_i(y_h)_{x_i}(p - i_h p)\\ - c\, y_h(p - i_h p) + u_h(p - i_h p)\Big).\end{aligned}$$

For $\rho^p(y - i_h y)$ the situation is more involved, since it contains the term $\langle \mu_h, y - i_h y\rangle$. We interpret this contribution as a quadrature rule of an integral of a certain function. Recalling the set $\mathcal{N}_i$ of neighboring indices from Remark 1.2.9 we set for $i = 1, \dots, m$

$$n_i := \operatorname{card}(\mathcal{N}_i) \in \mathbb{N}$$

and introduce the LAGRANGE interpolant $N_h \in Y_h$ by

$$0 < N_h := \sum_{i=1}^{m} n_i \phi_i.$$

Denoting by $x_j^T (j = 1, \dots, d+1)$ the finite element nodes of a simplex T and by μ_j^T the corresponding coefficients of μ_h we have

$$\langle \mu_h, y - i_h y\rangle = \sum_{i=1}^{m} \mu_i (y - i_h y)(x_i) = \sum_{T\in\mathcal{T}_h} \frac{|T|}{d+1} \sum_{j=1}^{d+1} \frac{d+1}{|T|} \frac{(y - i_h y)(x_j^T)\mu_j^T}{N_h(x_j^T)},$$

so that $\langle \mu_h, y - i_h y\rangle$ may be considered as the application of the quadrature rule

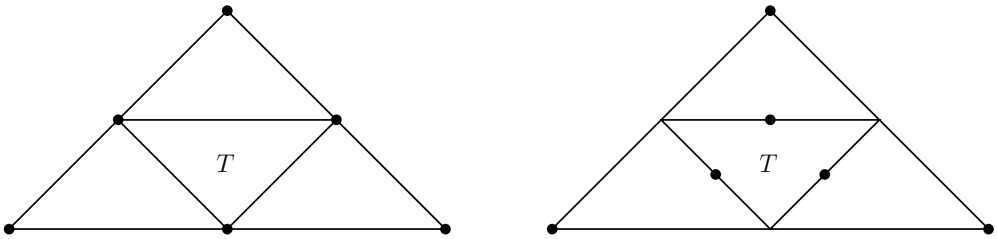
$$\int_T g(x)\,\mathrm{d}x \approx \frac{|T|}{d+1} \sum_{j=1}^{d+1} g(x_j^T) \tag{3.31}$$

to the expression

$$\sum_{T\in\mathcal{T}_h} \int_T \frac{d+1}{|T|} \frac{(y - i_h y) I_h \mu_h}{N_h}(x)\,\mathrm{d}x.$$

We use the quadrature rule (3.31) since the quadrature weights μ_j^T $(j = 1, \dots, d+1)$ are only given in the vertices of a simplex T. The previous considerations motivate to define the local adjoint residual by

$$\begin{aligned}&\rho_T^p((y - i_h y)_{|_T}) :=\\ &\int_T \Big(\sum_{i,j=1}^{d} -a_{ij}(y - i_h y)_{x_i}(p_h)_{x_j} - \sum_{i=1}^{d} b_i(y - i_h y)_{x_i}(p_h) - c(y - i_h y)p_h\Big)\\ &\qquad + \int_T (y_h - y_0)(y - i_h y) + \sum_{j=1}^{d+1} \frac{(y - i_h y)(x_j^T)(\mu_j^{b,T} - \mu_j^{a,T})}{N_h(x_j^T)}.\end{aligned}$$



(a) $i_{2h}^{(2)}$. (b) $i^{(2)}$.

FIGURE 3.1. Used sampling nodes for interpolant operators and $d = 2$.

Let us finally consider $e^{\mu}(y)$. Remark 3.2.3 motivates to approximate this term according to

$$\begin{aligned} e^{\mu}(y) &= \langle \mu + \mu_h, y_h - y \rangle = 2\langle \mu_h, y_h - y \rangle + \langle \mu - \mu_h, y_h - y \rangle \approx 2\langle \mu_h, y_h - y \rangle \\ &= 2\sum_{i=1}^{m} \mu_i (y_h - y)(x_i) = \sum_{T \in \mathcal{T}_h} \sum_{j=1}^{d+1} \frac{2\mu_j^T}{N_h(x_j^T)} (y_h - y)(x_j^T), \end{aligned}$$

where $\mu_i := \mu_i^b - \mu_i^a$ $(i = 1, \dots, m)$, and $\mu_j^T := \mu_j^{b,T} - \mu_j^{a,T}$ $(j = 1, \dots, d+1)$ denote the discrete multipliers in the element-wise renumbering. We note, that the error induced by this approximation has the form $\langle \mu - \mu_h, y_h - y \rangle$ and in essence is of the size of $\|y - y_h\|_{L^\infty(\Omega)}$ since $\|\mu_h\|_{\mathcal{M}(\bar{\Omega})}$ is uniformly bounded w.r.t. the gridsize h.

We now set

$$e_T^{\mu}(y_{|T}) := \sum_{j=1}^{d+1} \frac{2\mu_j^T}{N_h(x_j^T)} (y_h - y)(x_j^T).$$

In order to obtain computable local indicators for $d = 2$, we approximate $y - i_h y$ and $p - i_h p$ on every triangle T by $(i_{2h}^{(2)} y_h - y_h)_{|T}$ and $(i_{2h}^{(2)} p_h - p_h)_{|T}$ as suggested in [Ran05, Rem. 1]. Here, $i_{2h}^{(2)} y_h$ denotes a quadratic LAGRANGE interpolation of y_h on a coarser mesh using function values of y_h at element vertices (similarly for p_h). In detail the local interpolant $i_{2h}^{(2)} \phi_h$ for an arbitrary function $\phi_h \in Y_h$ on a triangle T is defined by

$$(i_{2h}^{(2)} \phi_h)(x_1, x_2) := a + bx_1 + cx_2 + dx_1 x_2 + ex_1^2 + fx_2^2, \quad (x_1, x_2)^T \in \Omega, \tag{3.32}$$

where the coefficients $a, b, c, d, e, f \in \mathbb{R}$ are obtained by the solution of a linear system demanding the exact interpolation in the sampling nodes shown in Figure 3.1(a). For approximating $(y_h - y)(x_j^T)$ we compute $(y_h - i^{(2)} y_h)(x_j^T)$. The quadratic interpolation operator $i^{(2)}$ differs from $i_{2h}^{(2)}$ in interpolating the function values of y_h in the midpoints of element edges. Its use is caused by the fact that our approximation to $e^{\mu}(y)$ relies on function evaluations in the finite element nodes x_i $(i = 1, \dots, m)$. If the interpolants $i_{2h}^{(2)} y_h$ and $i^{(2)} y_h$ violate the state constraints we use $\max(y_a, \min(y_b, i_{2h}^{(2)} y_h))$ and $\max(y_a, \min(y_b, i^{(2)} y_h))$, respectively instead.

Our error estimator finally takes the form

$$\eta := \frac{1}{2} \sum_{T \in \mathcal{T}_h} \rho_T^p((i_{2h}^{(2)} y_h - y_h)_{|T}) + \rho_T^y((i_{2h}^{(2)} p_h - p_h)_{|T}) + e_T^{\mu}((i^{(2)} y_h)_{|T}). \tag{3.33}$$

It turns out, that the direct cellwise error representation leads to typical oscillations of neighboring residuals ρ_T^p and $\rho_{T'}^p$ respectively with $T, T' \in \mathcal{T}_h, T \neq T', \overline{T} \cap \overline{T'} \neq \varnothing$.

Furthermore one observes that starting from the error representation (3.29) it is possible to avoid dealing with measures by the help of the dual state equation (3.5b). As the multipliers disappear, function evaluations in the nodes are not necessary and hence there is no need for two different heuristical interpolants anymore. In detail we have for $i_h y = y_h$:

$$\begin{aligned} J(y,u) - J(y_h,u_h) = \frac{1}{2}\left(J_y(y_h,u_h)(y-y_h) - a(y-y_h,p_h) + \langle \mu_h, y-y_h\rangle\right) \\ + \frac{1}{2}\left(-a(y_h,p-i_h p) + (u_h, p-i_h p)\right) + \frac{1}{2}\langle \mu+\mu_h, y_h - y\rangle. \end{aligned}$$

Summing up the multiplier parts and using the adjoint equation (3.5b), one obtains

$$\langle \mu_h, y-y_h\rangle + \langle \mu+\mu_h, y_h-y\rangle = \langle \mu, y_h - y\rangle = (y-y_0, y-y_h) - a(y-y_h,p).$$

Finally we have

$$\begin{aligned} 2(J(y,u) - J(y_h,u_h)) = J_y(y_h,u_h)(y-y_h) - a(y-y_h,p_h) \\ - a(y_h,p-i_h p) + (u_h, p-i_h p) + (y-y_0, y-y_h) - a(y-y_h,p). \end{aligned} \tag{3.34}$$

Following the lines of Remark 3.5 in [BR01] we split the above equation into a cellwise representation and integrate by parts. This gives rise to define

$$\begin{aligned} R^{y_h}_{|T} &= u_h - \mathcal{A} y_h \\ R^{p_h}_{|T} &= y_h - y_0 - \mathcal{A}^* p_h \\ R^{p}_{|T} &= y - y_0 - \mathcal{A}^* p \\ r^{y_h}_{|e} &= \begin{cases} \frac{1}{2}\vec{\nu}\cdot[\nabla y_h \cdot (a_{ij})], & \text{for } e \subset \partial T \setminus \partial\Omega \\ \vec{\nu}\cdot(\nabla y_h \cdot (a_{ij})), & \text{for } e \subset \partial\Omega \end{cases} \\ r^{p_h}_{|e} &= \begin{cases} \frac{1}{2}\vec{\nu}\cdot[(a_{ij})\nabla p_h], & \text{for } e \subset \partial T \setminus \partial\Omega \\ \vec{\nu}\cdot((a_{ij})\nabla p_h + p_h\vec{b}), & \text{for } e \subset \partial\Omega \end{cases} \\ r^{p}_{|e} &= \begin{cases} \frac{1}{2}\vec{\nu}\cdot[(a_{ij})\nabla p], & \text{for } e \subset \partial T \setminus \partial\Omega \\ \vec{\nu}\cdot((a_{ij})\nabla p + p\vec{b}), & \text{for } e \subset \partial\Omega \end{cases}, \end{aligned}$$

where $[\cdot]$ defines the jump across the inter-element edge e. Now equation (3.34) reads

$$\begin{aligned} 2(J(y,u) - J(y_h,u_h)) = \sum_{T\in\mathcal{T}_h} (y-y_h, R^{p_h}_{|T})_T - (y-y_h, r^{p_h}_{|\partial T})_{\partial T} \\ + (R^{y_h}_{|T}, p-i_h p)_T - (r^{y_h}_{|\partial T}, p-i_h p)_{\partial T} + (y-y_h, R^{p}_{|T})_T - (y-y_h, r^{p}_{|\partial T})_{\partial T}. \end{aligned} \tag{3.35}$$

In order to obtain a computable error estimator we apply the same technique as before with the difference that due to the missing multipliers the special interpolation operator $i^{(2)}$ is not needed anymore. We substitute $R^p_{|T}$ and $r^p_{|e}$ by

$$\begin{aligned} R^{i^{(2)}_{2h} p_h}_{|T} &= i^{(2)}_{2h} y_h - y_0 - \mathcal{A}^* i^{(2)}_{2h} p_h \\ r^{i^{(2)}_{2h} p_h}_{|e} &= \begin{cases} 0, & \text{for } e \subset \partial T \setminus \partial\Omega \\ \vec{\nu}\cdot((a_{ij})\nabla i^{(2)}_{2h} p_h + i^{(2)}_{2h} p_h \vec{b}), & \text{for } e \subset \partial\Omega \end{cases} \end{aligned}$$

and define

$$\tilde{\eta} := \frac{1}{2}\sum_{T\in\mathcal{T}_h} (i_{2h}^{(2)} y_h - y_h, R_{|T}^{p_h})_T - (i_{2h}^{(2)} y_h - y_h, r_{|\partial T}^{p_h})_{\partial T} + (R_{|T}^{y_h}, i_{2h}^{(2)} p_h - p_h)_T - (r_{|\partial T}^{y_h}, i_{2h}^{(2)} p_h - p_h)_{\partial T} + (i_{2h}^{(2)} y_h - y_h, R_{|T}^{i_{2h}^{(2)} p_h})_T - (i_{2h}^{(2)} y_h - y_h, r_{|\partial T}^{i_{2h}^{(2)} p_h})_{\partial T}. \tag{3.36}$$

In the following numerical example we investigate the effectivity index of an estimator η in terms of

$$I_{\text{eff}}^{\eta} := \frac{J(y,u) - J(y_h,u_h)}{\eta}. \tag{3.37}$$

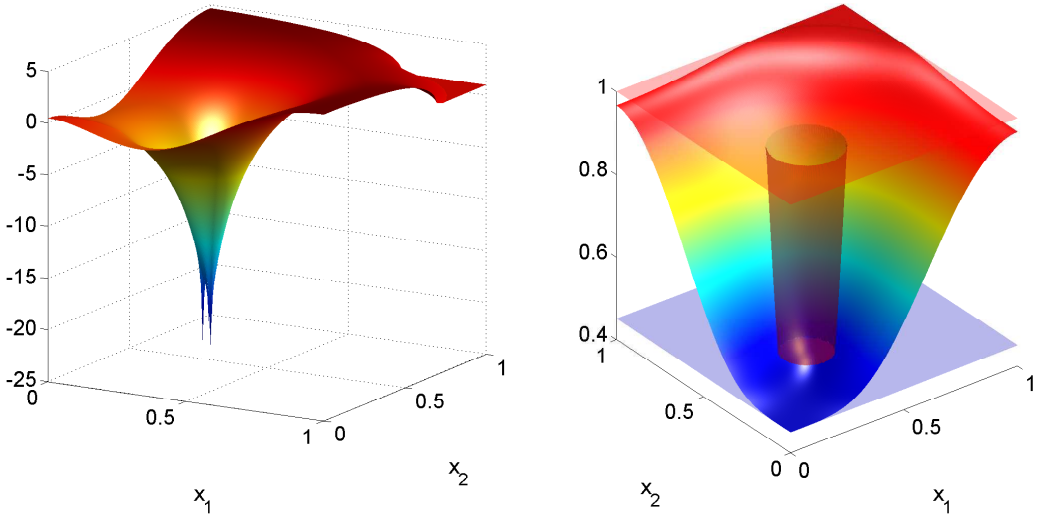
3.2.2.2. *Numerical experiment.* We set $d = 2$ and consider the domain $\Omega := (0,1)^2$ with the elliptic differential operator $\mathcal{A}$ defined by $a_{ij} = \delta_{ij}$, $b_i = 0$, $(i,j = 1,2)$, and $c = 1$. The regularization parameter in the cost functional J is set to $\alpha = 1$. The desired control and state functions u_0 and y_0 as well as the bounds y_a and y_b for the state are given by

$$\begin{aligned} &u_0(x) = 60, && y_0(x) = 0.5, \\ &y_a(x) = 0.45 \quad \text{and} && y_b(x) = \min\left(1, \max\left(0.5,\, 50\left|x - (0.3,0.3)^T\right|^2\right)\right) \end{aligned}$$

for every $x \in \bar{\Omega}$. The corresponding optimal control problem reads

$$\min_{u\in L^2(\Omega)} J(y,u) = \tfrac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|u - u_0\|^2_{L^2(\Omega)}$$
$$\text{s.t.} \quad \begin{aligned} -\Delta y + y &= u \quad \text{in} \quad \Omega \\ \partial_{\bar{\nu}} y &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \quad \text{and } y_a(x) \le y(x) \le y_b(x) \quad \forall x \in \bar{\Omega}.$$

In order to avoid specialties introduced by test problems admitting exact solutions we consider a fully generic test case by taking the numerical solution (y,u) obtained on an equidistant grid containing 1001^2 nodes as substitute for the exact solution, see Figure 3.2. The reference functional value $J^* := J(y,u)$ takes the value $J^* = 1759.04686$. The support of the corresponding multiplier μ is depicted in Figure 3.3. Let us note that it is a difficult task to determine J^* as accurate as possible. Therefore not only the errors in the objective but also the effectivity indices at the finest refinement levels should be treated with care. We start the numerical run on a uniform triangulation containing 484 nodes. On a mesh with 113569 nodes obtained by congruent refinement we obtain $J^* - J(y_h,u_h) \approx 0.00679$. Local refinement using the so called tolerance reduction strategy (see [BR96]) together with the estimator η leads to meshes where this value of the error already is reached with less then a quarter of unknowns. Specifically, for $m = 23216$ we already obtain $J^* - J(y_h,u_h) \approx 0.00469$. The development of the error in the objective is presented in Figure 3.4. The effectivity index of both estimators is documented in Table 3.2 and Table 3.3, where Table 3.1 contains the effectivities for global refinement. We observe that the estimators η and $\tilde{\eta}$ slightly underestimate the real error, but always have the same magnitude as the true error. Figure 3.5 shows two meshes obtained by the tolerance reduction strategy. These meshes clearly indicate that the largest errors in the numerical approximation have their origin in the square $[0.3,0.5]^2$. In this area the discrete multipliers take their largest values. Moreover, the quantity $i_{2h}^{(2)} p_h$ appearing in η and $\tilde{\eta}$ produces additional errors since p seems to have a singularity in this region. The observation $J^* > J(y_h,u_h)$ can be explained by fact that J^* is obtained from a discrete optimal control problem which contains more constraints



(a) Optimal control u_h. (b) $y_a \le y_h \le y_b$.

FIGURE 3.2. Solution on a uniform mesh with 1001^2 nodes.

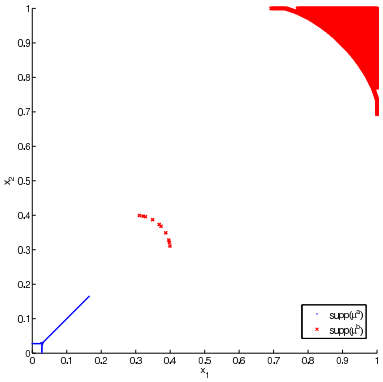


FIGURE 3.3. • $\mathrm{supp}(\mu_{a,h})$, × $\mathrm{supp}(\mu_{b,h})$

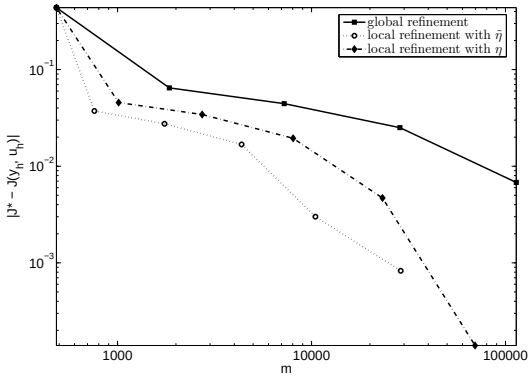


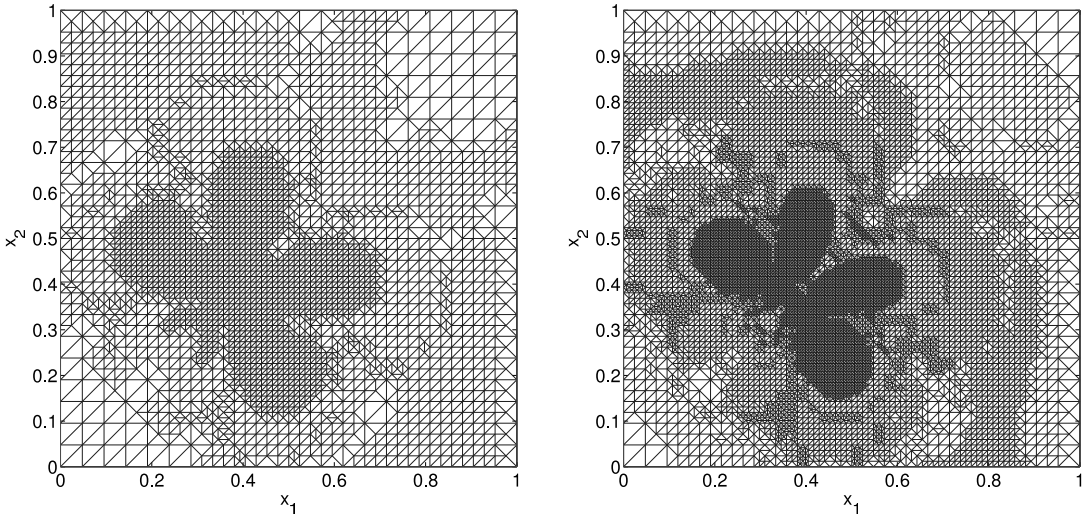
FIGURE 3.4. Error in the cost functional J.

i	$m = i^2$	$h = \frac{\sqrt{2}}{i-1}$	$h_{\min} = \frac{1}{i-1}$	$J^* - J(y_h, u_h)$	I_{eff}^{η}	$I_{\text{eff}}^{\tilde{\eta}}$
22	484	0.0673	0.0476	0.43808	-2.0	-29.3
43	1849	0.0337	0.0238	0.06467	-0.9	-3.2
85	7225	0.0168	0.0119	0.04445	-2.5	-9.6
169	28561	0.0084	0.0060	0.02506	-5.5	-71.2
337	113569	0.0042	0.0030	0.00679	-5.9	-29.1

TABLE 3.1. Mesh data, error and effectivity indices for global refinement.

then all other intermediate discrete optimal control problems. Let us note that on the other hand the interpolations used in the estimators η and $\tilde{\eta}$ are subject to the same set of constraints as the corresponding discrete state in the associated discrete optimal control problem. This may explain the sign of I_{eff} in all tables.

All solutions of the discrete optimization problems are computed with Matlab by a Moreau-Yosida-based active set strategy described in Algorithm 3.1.8.



(a) $m = 2730$. (b) $m = 8038$.

FIGURE 3.5. Locally refined meshes from η.

m	h	$h_{\min}$	$J^* - J(y_h, u_h)$	I_{eff}^{η}
484	0.0673	0.0476	0.43808	-2.0
1013	0.0673	0.0238	0.04559	-0.5
2730	0.0673	0.0119	0.03430	-1.1
8038	0.0673	0.0060	0.01945	-1.9
23216	0.0673	0.0030	0.00469	-1.3
69645	0.0673	0.0015	-0.00014	0.1

TABLE 3.2. Mesh data, error and effectivity index for local refinement with η.

m	h	$h_{\min}$	$J^* - J(y_h, u_h)$	$I_{\text{eff}}^{\tilde{\eta}}$
484	0.0673	0.0476	0.43808	-29.3
760	0.0673	0.0238	0.03734	-0.8
1747	0.0673	0.0119	0.02751	-1.3
4359	0.0673	0.0060	0.01679	-2.0
10471	0.0673	0.0030	0.00300	-0.8
28844	0.0673	0.0015	-0.00083	0.6

TABLE 3.3. Mesh data, error and effectivity index for local refinement with $\tilde{\eta}$.

3.2.3. The control and state constrained problem. We now extend the above techniques to the simultaneously control and state constrained Problem 3.1.4, which is matter of subject in the work [GT09]. For a fixed regularization parameter γ we develop a goal-oriented a posteriori error estimator within the next section for the regularized solutions of (3.7) and (3.14) respectively. We therefore derive a regularized extension of the error representation obtained in Theorem 3.2.2 to the

control and state constrained case. In particular no residual associated to the first order optimality condition with respect to the control appears in our approach. We mention here that we are not interested to the error involved by the regularization parameter. Our aim is rather performing a first attempt to understand the behavior of a goal-oriented based error estimate in connection with a Moreau-Yosida regularization. An overall error reduction which tie the regularization parameter with the current mesh size is subject of an ongoing research work.

Using the solution algorithm from Section 3.1.4.2 the performance of the overall adaptive solver is assessed by numerical examples in Section 3.2.3.2.

3.2.3.1. *Local error indicators.* To achieve high accuracies in an optimal fashion, we marry our regularization semi-smooth NEWTON solver with an adaptive mesh refinement process based on a goal-oriented approach. As quantity of interest we consider the objective functional J which is corresponding to the tracking part in the objective functional of the regularized optimal control problem (3.7). In this section we again assume $u_0 = u_{0,h}$.

Following Section 3.2.2.1 we define the following residuals

$$\begin{aligned}\rho^{p^\gamma}(\cdot) &:= J_y(y_h^\gamma, u_h^\gamma)(\cdot) - a(\cdot, p_h^\gamma) + (\mu_h^\gamma, \cdot)\\ \rho^{y^\gamma}(\cdot) &:= -a(y_h^\gamma, \cdot) + (u_h^\gamma + f, \cdot)\end{aligned}$$

with

$$\begin{aligned}\mu^\gamma &:= \gamma \max(0, y^\gamma - y_b) - \gamma \max(0, y_a - y^\gamma),\\ \mu_h^\gamma &:= \gamma \max(0, y_h^\gamma - I_h y_b) - \gamma \max(0, I_h y_a - y_h^\gamma).\end{aligned}$$

As $\gamma \to \infty$, μ^γ and μ_h^γ play the role of the measure LAGRANGE multipliers corresponding to state constraints in the limit problem (3.3). Moreover we abbreviate

$$\lambda^\gamma := \lambda_b^\gamma - \lambda_a^\gamma \quad \text{and} \quad \lambda_h^\gamma := \lambda_{b,h}^\gamma - \lambda_{a,h}^\gamma.$$

Theorem 3.2.4. *Let (u^γ, y^γ) and (u_h^γ, y_h^γ) be the solutions of the optimal control problems (3.7) and (3.14) with corresponding adjoint states p^γ, p_h^γ and multipliers associated to the control and state constraints $\lambda^\gamma, \lambda_h^\gamma, \mu^\gamma, \mu_h^\gamma$. Then*

$$\begin{aligned}(3.38)\quad &2(J(y^\gamma, u^\gamma) - J_h(y_h^\gamma, u_h^\gamma)) =\\ &\rho^{p^\gamma}(y^\gamma - i_h y^\gamma) + \rho^{y^\gamma}(p^\gamma - i_h p^\gamma) + (\mu^\gamma + \mu_h^\gamma, y_h^\gamma - y^\gamma) + (\lambda^\gamma + \lambda_h^\gamma, u_h^\gamma - u^\gamma).\end{aligned}$$

PROOF. For ease of exposition, we omit the upperscript γ in this proof for the quantities $y^\gamma, u^\gamma, p^\gamma, \lambda^\gamma, \mu^\gamma$ and their discrete counterparts. We have

$$\begin{aligned}&2(J(y,u) - J_h(y_h, u_h))\\ &= \alpha((u - u_0) + (u_h - u_0), (u - u_0) - (u_h - u_0))\\ &\quad + ((y - y_0) + (y_h - y_0), (y - y_0) - (y_h - y_0))\\ &= \alpha(u_h - u_0, u) + (-\alpha(u - u_0), u_h - u) + (-\alpha(u_h - u_0), u_h)\\ &\quad + (y_h - y_0, y) + a(y, p) - a(y_h, p_h) - a(y_h, p)\\ &\quad + (\mu_h, y) - (\mu, y) + (\mu_h, y_h) + (\mu, y_h)\\ &\quad - (\mu_h, y).\end{aligned}$$

For the last step, the adjoint equation was used 3 times and a zero was added. The last 4 terms can be summed up to $(\mu + \mu_h, y_h - y)$. The term $(y_h - y_0, y) + (\mu_h, y)$ already belongs to the dual residual, while $-a(y_h, p)$ belongs to the primal residual. The remaining both bilinear forms with a are expressed by using the both primal

equations. The furthermore $(p_h, u+f) - a(y, p_h) = 0$ is added to the equation. We obtain:

$$\begin{aligned}
&2(J(y,u) - J_h(y_h,u_h)) \\
=&- a(y_h,p) \\
&+ (y_h - y_0, y) + (\mu_h, y) && - a(y,p_h) \\
&+ (\mu+\mu_h, y_h - y) \\
&+ \alpha(u_h - u_0, u) + (-\alpha(u-u_0), u_h - u) + (-\alpha(u_h-u_0), u_h) \\
&+ (-p, -u-f) + (-p_h, u_h + f) \\
& && + (p_h, u+f) \\
&+ (u_h, p) - (p, u_h) \\
=&- a(y_h,p) + (u_h + f, p) \\
&- a(y,p_h) + (y_h - y_0, y) + (\mu_h, y) \\
&+ (\mu+\mu_h, y_h - y) \\
&+ (\alpha(u_h-u_0) + p_h, u) \\
&+ (-\alpha(u-u_0) - p, u_h - u) + (-\alpha(u_h-u_0) - p_h, u_h) \\
=&\rho^y(p) + \rho^p(y) + (\mu+\mu_h, y_h - y) \\
&+ \underbrace{(\alpha(u_h-u_0) + p_h + \lambda_h, u)}_{=0} - (\lambda_h, u) + (\lambda, u_h - u) + (\lambda_h, u_h) \\
=&\rho^y(p) + \rho^p(y) + (\mu+\mu_h, y_h - y) + (\lambda+\lambda_h, u_h - u).
\end{aligned}$$

Let us emphasize that in last intermediate step due to variational discretization the residual for the control vanishes. Because of Galerkin orthogonality of the error in the state and costate equation we could subtract arbitrary functions $i_h p$ and $i_h y \in Y_h$ within the residuals ρ^y and ρ^p and end up with the assertion. □

Let us now define the inner residuals

$$\begin{aligned}
R^{y_h^\gamma}_{|T} &:= u_h^\gamma + f - \mathcal{A} y_h^\gamma, \\
R^{p_h^\gamma}_{|T} &:= y_h^\gamma - y_0 - \mathcal{A}^* p_h^\gamma, \\
R^{p^\gamma}_{|T} &:= y^\gamma - y_0 - \mathcal{A}^* p^\gamma,
\end{aligned}$$

and the edge residuals

$$\begin{aligned}
r^{y_h^\gamma}_{|e} &:= \begin{cases} \frac{1}{2}\vec{\nu} \cdot [\nabla y_h^\gamma \cdot (a_{ij})], & e \subset \partial T \setminus \partial\Omega \\ \vec{\nu} \cdot (\nabla y_h^\gamma \cdot (a_{ij})), & e \subset \partial\Omega \end{cases}, \\
r^{p_h^\gamma}_{|e} &:= \begin{cases} \frac{1}{2}\vec{\nu} \cdot [(a_{ij}) \nabla p_h^\gamma], & e \subset \partial T \setminus \partial\Omega \\ \vec{\nu} \cdot ((a_{ij}) \nabla p_h^\gamma + p_h^\gamma \vec{b}), & e \subset \partial\Omega \end{cases}, \\
r^{p^\gamma}_{|e} &:= \begin{cases} \frac{1}{2}\vec{\nu} \cdot [(a_{ij}) \nabla p^\gamma], & e \subset \partial T \setminus \partial\Omega \\ \vec{\nu} \cdot ((a_{ij}) \nabla p^\gamma + p^\gamma \vec{b}), & e \subset \partial\Omega \end{cases}.
\end{aligned}$$

Here $[\cdot]$ denotes the jump across the inter-element edge e. Now by integration by parts we can localize the error representation (3.38) by

$$\begin{aligned}2(J(y^\gamma,u^\gamma)-J_h(y_h^\gamma,u_h^\gamma)) = &\sum_{T\in\mathcal{T}_h}(y^\gamma-y_h^\gamma,R_{|T}^{p_h^\gamma})_T-(y^\gamma-y_h^\gamma,r_{\partial T}^{p_h^\gamma})_{\partial T}\\ &+(R_{|T}^{y_h^\gamma},p^\gamma-i_hp^\gamma)_T-(r_{|\partial T}^{y_h^\gamma},p^\gamma-i_hp^\gamma)_{\partial T}\\ &+(y^\gamma-y_h^\gamma,R_{|T}^{p^\gamma})_T-(y^\gamma-y_h^\gamma,r_{|\partial T}^{p^\gamma})_{\partial T}\\ &+(\lambda^\gamma+\lambda_h^\gamma,u_h^\gamma-u^\gamma)_T.\end{aligned}$$

Since this localized sum still contains unknown quantities, we make use of a local higher order quadratic interpolant operator $i_{2h}^{(2)}: Y_h\to\mathbb{P}^2(T)$ for some $T\in\mathcal{T}_h$ as already introduced in (3.32) for $d=2$. The technique for computing $i_{2h}^{(2)}\phi_h$ for some $\phi_h\in Y_h$ can easily be carried over to three space dimensions. However this is supposed to be numerically expensive. In order to derive a computable estimator we now replace the unknown functions y^γ and p^γ in (3.38) by $i_{2h}^{(2)}y_h^\gamma$ and $i_{2h}^{(2)}p_h^\gamma$. Since $u^\gamma=P_{[u_a,u_b]}\left(-\frac{1}{\alpha}p^\gamma+u_0\right)$ holds, a reasonable locally computable approximation is

$$\tilde{u}^\gamma=P_{[u_a,u_b]}\left(-\frac{1}{\alpha}i_{2h}^{(2)}p_h^\gamma+u_0\right)$$

as already suggested in [VW08]. Similar for $\lambda^\gamma=-p^\gamma-\alpha(u^\gamma-u_0)$ we locally compute

$$\tilde{\lambda}^\gamma=-i_{2h}^{(2)}p_h^\gamma-\alpha\left(\tilde{u}^\gamma-u_0\right)$$

instead.

The estimator η^γ now reads

$$\eta^\gamma=\sum_{T\in\mathcal{T}_h}\eta_T^\gamma,$$

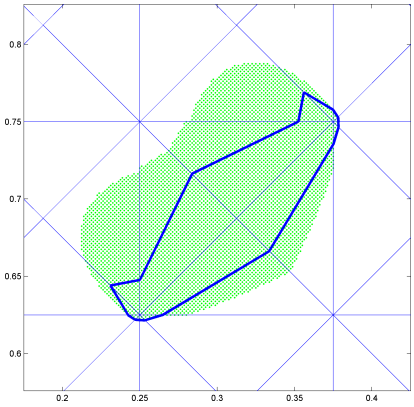
where

$$\begin{aligned}2\eta_T^\gamma=&(i_{2h}^{(2)}y_h^\gamma-y_h^\gamma,R_{|T}^{p_h^\gamma})_T-(i_{2h}^{(2)}y_h^\gamma-y_h^\gamma,r_{|\partial T}^{p_h^\gamma})_{\partial T}\\ &+(R_{|T}^{y_h^\gamma},i_{2h}^{(2)}p_h^\gamma-p_h^\gamma)_T-(r_{|\partial T}^{y_h^\gamma},i_{2h}^{(2)}p_h^\gamma-p_h^\gamma)_{\partial T}\\ &+(i_{2h}^{(2)}y_h^\gamma-y_h^\gamma,R_{|T}^{i_{2h}^{(2)}p_h^\gamma})_T-(i_{2h}^{(2)}y_h^\gamma-y_h^\gamma,r_{|\partial T}^{i_{2h}^{(2)}p_h^\gamma})_{\partial T}\\ &+(\tilde{\lambda}^\gamma+\lambda_h^\gamma,u_h^\gamma-\tilde{u}^\gamma)_T.\end{aligned}$$

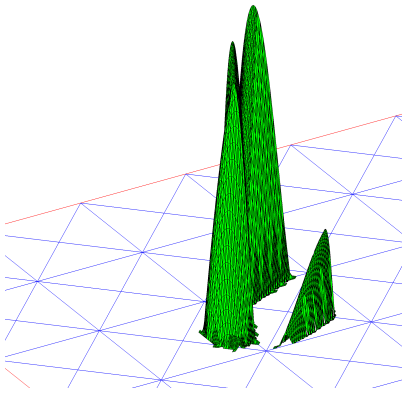
While for the other appearing quantities in η_T^γ quadrature rules of moderate order are suited, one has to take care for the last term

$$(\tilde{\lambda}^\gamma+\lambda_h^\gamma,u_h^\gamma-\tilde{u}^\gamma)_T=\int_T(\tilde{\lambda}^\gamma+\lambda_h^\gamma)(u_h^\gamma-\tilde{u}^\gamma). \tag{3.39}$$

The integrand has a support within the symmetric difference of the control active set of the variational discrete solution and the locally improved quantities. This also is depicted in Figure 3.6. One recognizes that $\tilde{u}^\gamma$ keeps the activity structure as u_h^γ has but smoothes the control active boundary towards the exact control active boundary. The kidney-shaped green area resolves the true control active set from Example 3.2.5 already very good even on a coarse mesh (compare also Figure 3.7(c)). Finally for computing (3.39) we just provide the integrand and a desired tolerance and apply an adaptive quadrature routine given in [Vog06, Algo. 31] for triangles containing the boundary of the control active set.



(a) Blue by u_h^γ, green by $\tilde{u}^\gamma$.



(b) Integrand $(\tilde{\lambda}^\gamma + \lambda_h^\gamma)(u_h^\gamma - \tilde{u}^\gamma)$.

FIGURE 3.6. Part of the u_a-active set for Example 3.2.5.

In order to study the efficiency of our implemented estimator, we define the effectivity index as

$$I_{\text{eff}} := \frac{J(y^\gamma, u^\gamma) - J_h(y_h^\gamma, u_h^\gamma)}{\eta^\gamma}.$$

Since the analytic solutions of the numerical examples are not known, we approximate $J(y^\gamma, u^\gamma)$ by $J_h(y_h^\gamma, u_h^\gamma)$ computed on a very fine mesh via the expression

$$\begin{aligned} J_h(y_h^\gamma, u_h^\gamma) = \frac{1}{2}\mathbf{y}^{\gamma T}\mathbf{M}\mathbf{y}^\gamma - \mathbf{y}^{\gamma T}\mathbf{M}\mathbf{y_0} + \frac{1}{2}\int_\Omega y_0^2 + \frac{1}{2\alpha}\mathbf{p}^{\gamma T}\mathbf{M}_{\textcircled{\scriptsize D}}\mathbf{p}^\gamma \\ + \frac{\alpha}{2}(\mathbf{u_a} - \mathbf{u_0})^T\mathbf{M}_{\textcircled{\scriptsize V}}(\mathbf{u_a} - \mathbf{u_0}) + \frac{\alpha}{2}(\mathbf{u_b} - \mathbf{u_0})^T\mathbf{M}_{\textcircled{\scriptsize A}}(\mathbf{u_b} - \mathbf{u_0}). \end{aligned} \tag{3.40}$$

3.2.3.2. *Numerical experiments.* Based on the previous error estimations and the semi-smooth NEWTON solvers described earlier, we design an adaptive finite element algorithm to solve (3.14). The algorithm consists in performing cycles of the form

$$\text{Solve} \Longrightarrow \text{Estimate} \Longrightarrow \text{Mark} \Longrightarrow \text{Refine}.$$

In the Mark step, elements are selected according to a bulk-type criterion [Dör96]. We select, for fixed specified $0 < \theta_i < 1$ ($i \in \{1, 2, 3\}$) the set $\mathcal{M} = \bigcup_{i=1}^3 \mathcal{M}_i \subset \mathcal{T}_h$, such that

$$\begin{aligned} \theta_1 \left| \sum_{T \in \mathcal{T}_h} \tau_{\mathring{T}} \right| &\leq \left| \sum_{T \in \mathcal{M}_1} \tau_{\mathring{T}} \right|, \\ \theta_2 \left| \sum_{T \in \mathcal{T}_h} \tau_{\partial T} \right| &\leq \left| \sum_{T \in \mathcal{M}_2} \tau_{\partial T} \right|, \\ \theta_3 \left| \sum_{T \in \mathcal{T}_h} \tau_{\lambda} \right| &\leq \left| \sum_{T \in \mathcal{M}_3} \tau_{\lambda} \right|, \end{aligned}$$

l	m	nt	h
1	81	128	0.17678
2	145	256	0.12500
3	289	512	0.08839
4	545	1024	0.06250
5	1089	2048	0.04419
6	2113	4096	0.03125
7	4225	8192	0.02210
8	8321	16384	0.01563
9	16641	32768	0.01105
10	33025	65536	0.00781
11	66049	131072	0.00552
12	131585	262144	0.00391
13	263169	524288	0.00276
14	525313	1048576	0.00195

TABLE 3.4. Mesh parameters for Example 3.2.5 (global refinement).

where the local quantities $\tau_{\mathring{T}}$, $\tau_{\partial T}$ and τ_λ are defined by

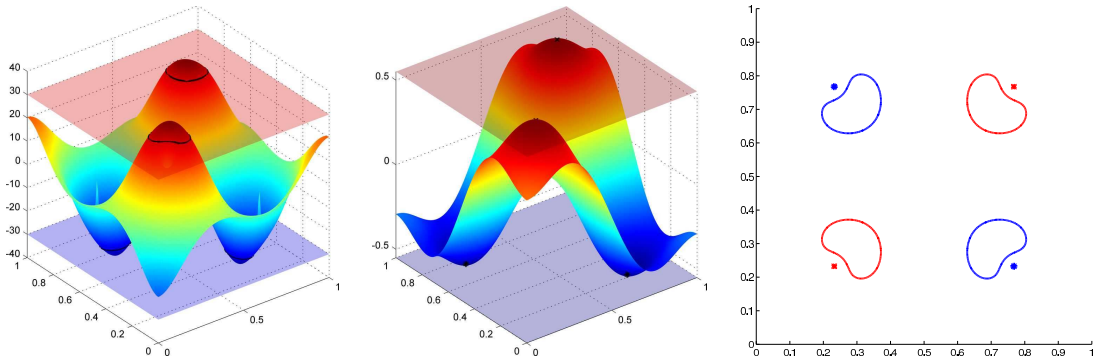
$$
\begin{aligned}
2\tau_{\mathring{T}} &:= (i_{2h}^{(2)} y_h^\gamma - y_h^\gamma, R_{|T}^{p_h^\gamma})_T + (R_{|T}^{y_h^\gamma}, i_{2h}^{(2)} p_h^\gamma - p_h^\gamma)_T + (i_{2h}^{(2)} y_h^\gamma - y_h^\gamma, R_{|T}^{i_{2h}^{(2)} p_h^\gamma})_T,\\
2\tau_{\partial T} &:= (i_{2h}^{(2)} y_h^\gamma - y_h^\gamma, r_{|\partial T}^{p_h^\gamma})_{\partial T} + (r_{|\partial T}^{y_h^\gamma}, i_{2h}^{(2)} p_h^\gamma - p_h^\gamma)_{\partial T} + (i_{2h}^{(2)} y_h^\gamma - y_h^\gamma, r_{|\partial T}^{i_{2h}^{(2)} p_h^\gamma})_{\partial T},\\
2\tau_\lambda &:= (\tilde\lambda^\gamma + \lambda_h^\gamma, u_h^\gamma - \tilde u^\gamma)_T.
\end{aligned}
$$

Flagging elements in such three separate steps has the advantage of properly handling possible scaling difference between jump, element and multiplier contributions in particular if the regularization parameter $\gamma \to \infty$. Once all the elements to be refined are marked, a new finer mesh is generated using the longest bisection rule implemented within the Matlab PDE toolbox. To assess the performance of the overall adaptive finite element algorithm we compare it with a uniform mesh refinement by monitoring values of the objective functional versus the numbers of degrees of freedom $N_{dof} := m$. Uniform refinement levels and the corresponding number of nodes m, number of triangles nt and grid size h are documented in Table 3.4.
In the sequel we provide the documentation for two numerical examples. For both examples, the analytic solution is not known, so for obtaining the efficiency index we compute a reference solution on the finest grid in Table 3.4 and hence an approximation of $J(y^\gamma, u^\gamma)$. The semi-smooth NEWTON solver converges generally in few iterations provided an appropriate update strategy is used for the regularization coefficient. In our experiments we use a simple continuation method. However more sophisticated techniques might be used (see for instance [HK06a]). We stop the semi-smooth NEWTON solver as soon as

$$\|G^\gamma(\mathbf{x}_n^\gamma)\|_2 \le \varepsilon_{\rm rel}\|G^\gamma(\mathbf{x}_0^\gamma)\|_2 + \varepsilon_{\rm abs}, \quad n = 1, \dots, n_{max},$$

for some user-specified maximum number of iterations n_{max} and tolerances $\varepsilon_{\rm rel}$ and $\varepsilon_{\rm abs}$. In our experiments we used $n_{max} = 100$. The absolute and relative tolerances are chosen more and more stringent as $\gamma \to \infty$ such that the final values are

$$\varepsilon_{\rm rel} = 10^{-12}, \quad \varepsilon_{\rm abs} = 10^{-8}.$$



(a) $u_a, -\frac{1}{\alpha}p_h^\gamma, u_b$. (b) $y_a \leq y_h^\gamma \leq y_b$. (c) Active sets.

FIGURE 3.7. Numerical solution for Example 3.2.5 and $l = 14$.

Example 3.2.5. As a first example we consider problem (3.3) with data

$$\Omega = (0,\,1)^2, \quad \mathcal{A} = -\Delta + id, \quad y_0 = \sin(2\pi x_1)\sin(2\pi x_2), \quad f = u_0 = 0,$$
$$u_a = -30, \quad u_b = 30, \quad y_a = -0.55, \quad y_b = 0.55, \quad \alpha = 10^{-4}.$$

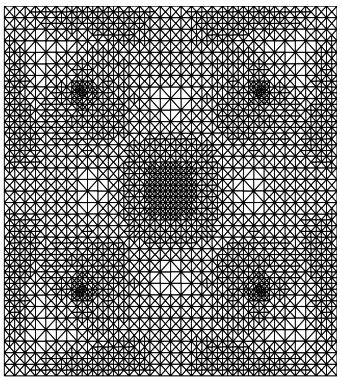
Its numerical solution in terms of $-\frac{1}{\alpha}p_h^\gamma$ as well as the optimal state y_h^γ is displayed in Figure 3.7 for $\gamma = 10^{14}$ on the mesh $l = 14$. The projection of $-\frac{1}{\alpha}p_h^\gamma$ onto $[u_a, u_b]$ corresponds to the optimal control u_h^γ which comprises together with y_h^γ our best approximation to the solution of (3.7). The boundaries of the control active sets are depicted as solid lines, while the state active sets are coded as star and cross markers. The color blue corresponds to the lower bound while the color red highlights the upper bound. Now by using the expression (3.40) we get $J(y^\gamma, u^\gamma) \approx 0.0375586175$. In Table 3.5 we depict the efficiency coefficient and the convergence history of the quantity of interest. Notice that the values of the efficiency coefficient are close to 1 which illustrate the good performance of our error estimator. A comparison between our adaptive finite element algorithm and a uniform mesh refinement in terms of number of degrees of freedom is reported in Figure 3.8(b). The adaptive refinement process performs well even though the benefit in this example is not big since the characteristic features of the optimal solution occupy an important area of the computational domain as illustrated by the adapted grid in Figure 3.8(a). Our motivation from including this example is to illustrate the variational discretization effect on the mesh refinement process. Indeed, regarding the shape of the control active set one would expect finest grids around the boundary of this set if a standard discretization for the control would have been used.

The author appends, that the suggested solution algorithm in the manuscript [DH07b] works for this particular example. Therein a generalized NEWTON method for the unregularized problem is investigated. Its practicability relies on the assumption that the appearing matrices are regular. This is the case for this particular example, because the control and state active sets do not intersect. But since the active sets are not known in advance, one cannot guarantee this property.

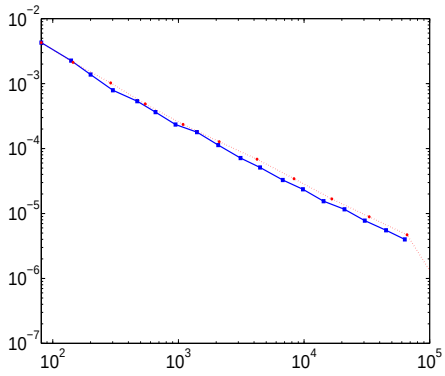
Example 3.2.6. In this example we set the computational domain to $\Omega = (-1, 1) \times (-1, 1)$ and $\mathcal{A} = -\Delta + id$. We take $\alpha = 10^{-3}$ and $u_0 = y_0 = (-3x_1^4 + 4x_1^3)\mathbb{1}_{[0,1]}(x_1)$, where $\mathbb{1}_A$ denotes the characteristic function of a set A. Furthermore we fix $f = (36x_1^2 - 24x_1)\mathbb{1}_{[0,1]}(x_1)$ and the bounds $0.1 \leq u \leq 2$, $0.1 \leq y \leq 2$. This data is chosen such that the optimal control and optimal state exhibit active sets whose intersection is not empty (see Figure 3.9). An approximation $J(y^\gamma, u^\gamma) \approx 0.0130624289$ of the

k	m	$J(y^\gamma, u^\gamma) - J_h(y_h^\gamma, u_h^\gamma)$	I_{eff}
1	81	$4.275 \cdot 10^{-3}$	1.622
2	140	$2.259 \cdot 10^{-3}$	1.543
3	200	$1.380 \cdot 10^{-3}$	1.390
4	301	$7.904 \cdot 10^{-4}$	1.119
5	470	$5.369 \cdot 10^{-4}$	1.176
6	657	$3.643 \cdot 10^{-4}$	1.269
7	948	$2.343 \cdot 10^{-4}$	1.127
8	1405	$1.790 \cdot 10^{-4}$	1.187
9	2075	$1.133 \cdot 10^{-4}$	1.227
10	3123	$7.148 \cdot 10^{-5}$	1.144
11	4469	$5.115 \cdot 10^{-5}$	1.137
12	6775	$3.281 \cdot 10^{-5}$	1.172
13	9799	$2.360 \cdot 10^{-5}$	1.165
14	14305	$1.546 \cdot 10^{-5}$	1.181
15	20977	$1.161 \cdot 10^{-5}$	1.186
16	30445	$7.763 \cdot 10^{-6}$	1.256
17	44958	$5.524 \cdot 10^{-6}$	1.289
18	63389	$3.996 \cdot 10^{-6}$	1.290

TABLE 3.5. Adaptive refinement for Example 3.2.5 (bulk criterion, $\theta_i = 0.6$).



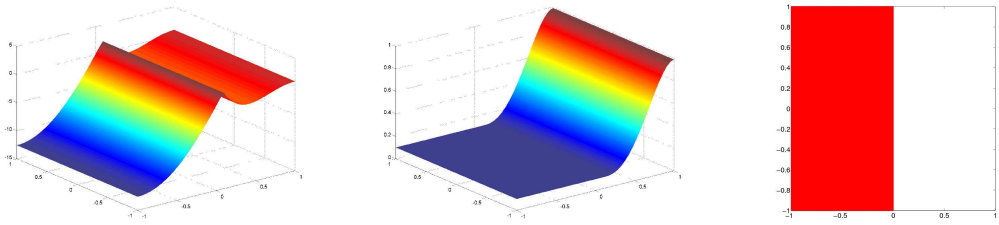
(a) Adaptive mesh for $k = 10$.



(b) Comparison of error decrement in the quantity of interest.

FIGURE 3.8. Example 3.2.5.

optimal quantity of interest is computed on the mesh level $l = 14$. We notice that the global refined meshes have the same topology as denoted in Table 3.4 for Example 3.2.5 but due to the enlarged domain have the doubled mesh parameter h. Figure 3.9 displays the corresponding state y_h^γ and the finite element quantity $-\frac{1}{\alpha}p_h^\gamma + u_{0,h}$. Throughout our computations we take $\gamma = 10^8$. The history of the efficiency indices as well as the convergence of the quantities of interest are reported in Table 3.6. As for the previous example we notice the high accuracy of our error estimator illustrated by the fact that the efficiency coefficient stays close to 1 during the adaptive procedure. The superiority of the performance of our adaptive algorithm over uniform mesh refinements is illustrated in Figure 3.10(b). We clearly observe

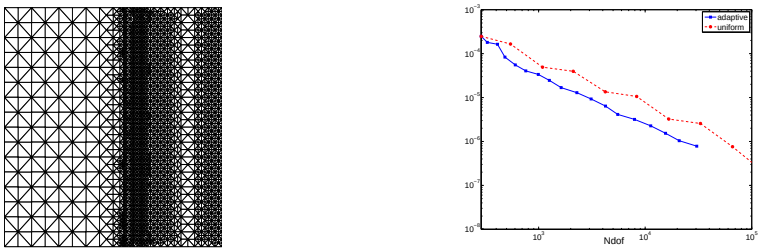


(a) $-\frac{1}{\alpha}p_h^\gamma + u_{0,h}$. (b) y_h^γ (c) Active sets intersection.

FIGURE 3.9. Numerical solution for Example 3.2.6 and $l = 14$.

k	m	$J(y^\gamma, u^\gamma) - J_h(y_h^\gamma, u_h^\gamma)$	I_{eff}
1	289	$2.482 \cdot 10^{-4}$	1.261
2	330	$1.805 \cdot 10^{-4}$	1.128
3	411	$1.635 \cdot 10^{-4}$	1.307
4	483	$8.344 \cdot 10^{-5}$	1.674
5	604	$5.544 \cdot 10^{-5}$	1.215
6	758	$4.051 \cdot 10^{-5}$	1.000
7	993	$3.370 \cdot 10^{-5}$	1.155
8	1261	$2.463 \cdot 10^{-5}$	1.198
9	1628	$1.684 \cdot 10^{-5}$	1.202
10	2287	$1.292 \cdot 10^{-5}$	1.140
11	3110	$9.290 \cdot 10^{-6}$	1.155
12	4242	$6.399 \cdot 10^{-6}$	1.167
13	5526	$4.136 \cdot 10^{-6}$	1.168
14	7942	$3.184 \cdot 10^{-6}$	1.109
15	11281	$2.268 \cdot 10^{-6}$	1.121
16	15531	$1.537 \cdot 10^{-6}$	1.144
17	20867	$1.041 \cdot 10^{-6}$	1.148
18	30498	$7.828 \cdot 10^{-7}$	1.095

TABLE 3.6. Adaptive refinement for Example 3.2.6 (bulk criterion, $\theta_i = 0.5$).



(a) Adaptive mesh for $k = 10$. (b) Comparison of error decrement in the quantity of interest.

FIGURE 3.10. Example 3.2.6.

in Figure 3.10(a) that the characteristic features of the solution are tracked in the adapted grid.

CHAPTER 4

Constraints on the gradient of the state

4.0. Introduction

Constraints on the gradient of the state play an important role in practical applications where solidification of melts forms a critical process. In order to accelerate the production it is highly desirable to speed up the cooling processes while avoiding damage of the products caused by large material stresses. Cooling frequently is described by systems of partial differential equations involving the temperature as a system variable, so that large (VON MISES) stresses in the optimization can be kept small by imposing pointwise bounds on the gradient of the temperature. Pointwise bounds on the gradient of the state in optimization in general deliver adjoint variables admitting low regularity only. This fact then necessitates the development of tailored discrete concepts which take into account the low regularity of adjoint variables and multipliers involved in the optimality conditions of the underlying optimization problem.

Let us briefly comment on related literature. In [CF93] Casas and Fernández investigate optimal control of semilinear elliptic PDEs with pointwise constraints on the gradient of the state. They provide a complete analysis including results on the structure and on the regularity of multipliers.

To the best of the authors knowledge the works [DGH09c] and [GH09] are the first contributions to finite element analysis for elliptic control problems with pointwise bounds on the gradient of the state. Both works separately study two different scenarios varying in assumptions on the domain, the regularization term of the control in the objective and used discretization techniques. Both scenarios complement one another and are also shortly introduced in [DGH08]. They build up the content for Subsection 4.1 and Subsection 4.2.

In the meantime the work [OW09] of Ortner and Wollner appeared which presents error bounds similar to ours, but derived by following techniques developed in [DH07a]. For a further discussion concerning constraints on the gradient of the state we also refer to [HPUU09, Sec. 3.3.2] and [Hin08].

The already announced paper [HK09] by Hintermüller and Kunisch also applies for optimal control problems dealing with constraints on the gradient of the state. Therein a general Moreau-Yosida-based framework is considered, which is also used to study a semismooth NEWTON algorithm in function space.

On the part of a posteriori error analysis for gradient constrained optimal control problems the only contribution known by the author is the work [Wol08] from Wollner. Therein goal-oriented error estimators for the approximate finite element solution combined with an interior point method are developed and investigated.

Let us give an outline of this chapter. We are interested in finite element analysis of the following control problem

$$\min_{u\in U_{ad}} J(u) = \frac{1}{2}\int_\Omega |y-y_0|^2 + \frac{\alpha}{\sigma}\int_\Omega |u|^\sigma \quad \text{subject to } y \text{ solves (4.2) and } \nabla y \in \vec{C}. \tag{4.1}$$

Here, $y_0 \in L^2(\Omega)$, $\alpha > 0$ and $\vec{C} := \{\vec{z} \in C^0(\bar\Omega)^2 \,:\, |\vec{z}(x)| \le \delta, x \in \bar\Omega\}$ are given for fixed $\delta > 0$.

First in Subsection 4.1 we consider

Scenario 4.0.1 ([DGH09c])**.** Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with a smooth boundary $\partial\Omega$. We further choose $\sigma := 2$, $\bar r := \infty$ and $U_{ad} := \{u \in L^2(\Omega) : u_a \le u \le u_b \text{ a.e. in } \Omega\}$, where $u_a < u_b$ are fixed constants. We consider the elliptic differential operator $\mathcal{A}$ from (1.1) with $b_i = 0$ $(i = 1, \dots, d)$ and subsequently assume the coefficients $a_{ij} = a_{ji}$ and $c \ge 0$ to be smooth functions on $\bar\Omega$.

Secondly in Subsection 4.2 we are concerned with

Scenario 4.0.2 ([GH09])**.** Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded, convex polyhedral domain with boundary $\partial\Omega$, whose inner dihedral angles at $\partial\Omega$ in the case $d = 3$ are assumed to be smaller than $\frac{3}{4}\pi$. This condition ensures the existence of some $\bar r > d$ such that we are allowed to choose $\bar r > \sigma := r > d$, $U_{ad} = L^r(\Omega)$ and $\mathcal{A} = -\Delta$.

From the above assumptions we infer that for a given $u \in L^r(\Omega)$ $(1 < r < \bar r)$ the elliptic boundary value problem

$$\begin{aligned} \mathcal{A}y &= u \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.2}$$

has a unique solution $y \in W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$.

We apply variational discretization of the controls combined with the lowest order Raviart–Thomas finite element approximations of a mixed formulation of the state equation. This in particular leads to piecewise constant approximations to the state and the adjoint state, respectively. The main result reads

$$\|u - u_h\|_{L^2(\Omega)} + \|y - y_h\|_{L^2(\Omega)} \le Ch^{\frac{1}{2}} |\log h|^{\frac{1}{2}}.$$

However, many existing finite element codes use finite elements based on conventional continuous piecewise polynomial Ansatz spaces. This is our motivation to provide numerical analysis for elliptic control problems with gradient constraints also for piecewise polynomial and continuous state approximations. In Subsection 4.2 we consider from [GH09] besides variational discretization also piecewise constant approximations of the controls. In both cases the state is discretized with standard piecewise linear, continuous finite elements. Our main result reads

$$\|u - u_h\|_{L^r(\Omega)} \le Ch^{\frac{1}{r}(1-\frac{d}{r})}, \text{ and } \|y - y_h\|_{L^2(\Omega)} \le Ch^{\frac{1}{2}(1-\frac{d}{r})},$$

for variational discretization as well as for piecewise constant control approximations. The presented finite element error estimates are confirmed by numerical experiments. Both approaches require to prove uniform bounds on the discrete multipliers associated to the discretized gradient constraints. Uniform error estimates of finite element approximations to elliptic equations then deliver the respective results.

4.1. Mixed finite element approximations for Scenario 4.0.1

We consider the elliptic optimal control problem (4.1) with control constraints and pointwise bounds on the gradient of the state under Scenario 4.0.1. We present a tailored finite element approximation to this optimal control problem, where the cost functional is approximated by a sequence of functionals which are obtained by discretizing the state equation with the help of the lowest order Raviart–Thomas mixed finite element. Pointwise bounds on the gradient variable are enforced in the elements of the triangulation. Controls are not discretized. Error bounds for control and state are obtained in two and three space dimensions. A numerical example confirms our analytical findings.

4.1.1. Mathematical setting. For $u \in L^r(\Omega)$ the unique solution

$$y \in W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$$

from boundary value problem (4.2) satisfies

$$\|y\|_{W^{2,r}(\Omega)} \le C\|u\|_{L^r(\Omega)}. \tag{4.3}$$

By tracing the dependence on r in the proof of the above inequality (see e.g. [ADN59] or [GT01, Chap. 9]) it is possible to prove that

$$\|y\|_{W^{2,r}(\Omega)} \le Cr\|u\|_{L^r(\Omega)}, \tag{4.4}$$

where C is independent of r, (compare also [JT81, Lem. 1.2] or [GN88, p. 17]). Since $U_{ad} \subset L^r(\Omega)$ for $r > d$ we have $y \in W^{2,r}(\Omega)$ and hence $\nabla y \in C^0(\bar\Omega)^d$ by a well–known embedding result. This ensures the validity of the pointwise gradient constraints in problem (4.1).
Finally we suppose that the following Slater condition holds:

$$\exists\, \tilde u \in U_{ad} \quad |\nabla \tilde y(x)| < \delta,\ x \in \bar\Omega \qquad \text{where } \tilde y \text{ solves (4.2) with } u = \tilde u. \tag{4.5}$$

Since $\tilde u$ is feasible for (4.1) we deduce from Theorem 3 in [CF93], that the above control problem has a unique solution $u \in U_{ad}$. From [CF93, Cor. 1] we deduce

Theorem 4.1.1. *An element $u \in U_{ad}$ is a solution of (4.1) if and only if there exist $\vec\mu \in \mathcal{M}(\bar\Omega)^d$ and $p \in L^t(\Omega)$ $(t < \frac{d}{d-1})$ such that*

$$\int_\Omega p\mathcal{A}\phi - \int_\Omega (y - y_0)\phi - \int_{\bar\Omega} \nabla\phi \cdot d\vec\mu = 0 \qquad \forall\, \phi \in W^{2,t'}(\Omega) \cap W_0^{1,t'}(\Omega) \tag{4.6a}$$

$$\int_\Omega (p + \alpha u)(v - u) \ge 0 \qquad \forall\, v \in U_{ad} \tag{4.6b}$$

$$\int_{\bar\Omega} (\vec z - \nabla y) \cdot d\vec\mu \le 0 \qquad \forall\, \vec z \in \vec C. \tag{4.6c}$$

Here, y is the solution of (4.2) and $\frac{1}{t} + \frac{1}{t'} = 1$.

Remark 4.1.2. Lemma 1 in [CF93] shows that the vector valued measure $\vec\mu$ appearing in Theorem 4.1.1 can be written in the form

$$\vec\mu = \frac{1}{\delta}\nabla y\, \mu,$$

where $\mu \in \mathcal{M}(\bar\Omega)$ is a nonnegative measure that is concentrated in the set $\{x \in \bar\Omega : |\nabla y(x)| = \delta\}$.

Our aim is to develop and analyze a finite element approximation of problem (4.1). We start by approximating the cost functional J by a sequence of functionals J_h where h is a mesh parameter related to a sequence of triangulations. Since p has very little regularity we propose to use a mixed finite element method based on the Raviart–Thomas element of lowest order. It is a specialty of our approach that it avoids explicit discretization of the controls. This procedure is motivated by the fact that the structure of the discrete analogue to (4.6b) already induces a discrete structure on the control through the discretization of the adjoint state p, compare Remark 4.1.6.

4.1.2. Finite element discretization. As already introduced in Section 1.2 equation (4.2) can be written in mixed formulation (1.9) with $f_u = u$ and $g_u = 0$. For $u \in L^2(\Omega)$ we denote its solution $(y, \vec{v}) \in L^2(\Omega) \times H(\mathrm{div}; \Omega)$ by $\mathcal{G}(u)$. Next, let $\mathcal{T}_h$ be a triangulation of Ω with maximum mesh size h. We suppose that $\bar{\Omega}$ is the union of the elements of $\mathcal{T}_h$; boundary elements are allowed to have one curved face. In addition, we assume that the triangulation is quasi-uniform in the sense of Definition 1.2.4. As already mentioned above we use a mixed finite element method based on the lowest order Raviart–Thomas element. Let therefore $\vec{V}_h := RT_0(\mathcal{T}_h)$ and $Y_h := P^0_{td,h}(\mathcal{T}_h)$. The variational formulation (1.9) gives rise to the discrete approximation of $\mathcal{G}$. For a given function $u \in L^2(\Omega)$ let $(y_h, \vec{v}_h) = \mathcal{G}_h(u) \in Y_h \times \vec{V}_h$ be the solution of (1.27). It is well–known ([BF91]) that the difference between $(y, \vec{v}) = \mathcal{G}(u)$ and $(y_h, \vec{v}_h) = \mathcal{G}_h(u)$ can be estimated as follows:

$$\begin{aligned} \|y - y_h\|_{L^2(\Omega)} + \|\vec{v} - \vec{v}_h\|_{L^2(\Omega)^d} &\le Ch\big(\|y\|_{H^1(\Omega)} + \|A\nabla y\|_{H^1(\Omega)^d}\big) \\ &\le Ch\|y\|_{H^2(\Omega)} \le Ch\|u\|_{L^2(\Omega)} \end{aligned} \tag{4.7}$$

by (4.3). In what follows it will be crucial to control the error between $\vec{v}$ and $\vec{v}_h$ in $L^\infty(\Omega)$.

Lemma 4.1.3. *Let $u \in L^\infty(\Omega)$ and $(y, \vec{v}) = \mathcal{G}(u)$ and $(y_h, \vec{v}_h) = \mathcal{G}_h(u)$. Then*

$$\|y - y_h\|_{L^\infty(\Omega)} + \|\vec{v} - \vec{v}_h\|_{L^\infty(\Omega)^d} \le Ch|\log h|\,\|u\|_{L^\infty(\Omega)}.$$

Proof. See [GN89, Cor. 5.5], where the result is proved for the model problem $a_{ij} = \delta_{ij}$ and $c = 0$, but it can be extended to the general case using techniques developed in [GN88]. □

Remark 4.1.4. More recently, localized pointwise error estimates for general second order elliptic equations on smooth domains were proved in [Dem04].

Next define

$$\vec{C}_h := \{\vec{c}_h : \bar{\Omega} \to \mathbb{R}^d \,:\, \vec{c}_{h|T} \text{ is constant and } |\vec{c}_{h|T}| \le \delta,\ T \in \mathcal{T}_h\}.$$

We approximate (4.1) by the following control problem depending on the mesh parameter h:

$$\begin{aligned} &\min_{u \in U_{ad}} J_h(u) := \frac{1}{2}\int_\Omega |y_h - y_0|^2 + \frac{\alpha}{2}\int_\Omega |u|^2 \\ &\text{subject to } \ (y_h, \vec{v}_h) = \mathcal{G}_h(u) \text{ and } \Big(\fint_T A^{-1}\vec{v}_h\Big)_{T \in \mathcal{T}_h} \in \vec{C}_h. \end{aligned} \tag{4.8}$$

Here, $\fint_T \cdot = \frac{1}{|T|}\int_T \cdot$. We note that the control is not discretized in (4.8) and that the state variable's gradient is only constrained on average on each cell. This problem represents a convex infinite–dimensional optimization problem of similar structure as problem (4.1), but with only finitely many constraints on the state.

Lemma 4.1.5. *There exists $h_0 > 0$ such that problem (4.8) has a unique solution $u_h \in U_{ad}$ for $0 < h \le h_0$. Furthermore, there are $\boldsymbol{\mu}_T \in \mathbb{R}^d, T \in \mathcal{T}_h$ and $(p_h, \vec{\chi}_h) \in Y_h \times \vec{V}_h$ such that with $(y_h, \vec{v}_h) = \mathcal{G}_h(u_h)$ we have*

$$\int_\Omega A^{-1}\vec{\chi}_h \cdot \vec{w}_h + \int_\Omega p_h \,\mathrm{div}\vec{w}_h + \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \fint_T A^{-1}\vec{w}_h = 0 \qquad \forall\, \vec{w}_h \in \vec{V}_h, \tag{4.9a}$$

$$\int_\Omega z_h \,\mathrm{div}\vec{\chi}_h - \int_\Omega c p_h\, z_h + \int_\Omega (y_h - y_0)\, z_h = 0 \qquad \forall\, z_h \in Y_h, \tag{4.9b}$$

$$\int_\Omega (p_h + \alpha u_h)(v - u_h) \ge 0 \qquad \forall\, v \in U_{ad}, \tag{4.9c}$$

$$\sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \big(\vec{c}_{h|T} - \fint_T A^{-1}\vec{v}_h\big) \le 0 \qquad \forall\, \vec{c}_h \in \vec{C}_h. \tag{4.9d}$$

Proof. We first prove that $\tilde{u}$ from (4.5) is feasible for (4.8). Let $\big(\tilde{y}, \vec{\tilde{v}}\big) = \mathcal{G}(\tilde{u})$ and $\big(\tilde{y}_h, \vec{\tilde{v}}_h\big) = \mathcal{G}_h(\tilde{u})$. For $T \in \mathcal{T}_h$ we deduce with the help of Lemma 4.1.3 and (4.5)

$$\begin{aligned}\left|\fint_T A^{-1}\vec{\tilde{v}}_h\right| &\le \left|\fint_T A^{-1}\left(\vec{\tilde{v}}_h - \vec{\tilde{v}}\right)\right| + \left|\fint_T A^{-1}\vec{\tilde{v}}\right| \\ &\le C\big\|\vec{\tilde{v}} - \vec{\tilde{v}}_h\big\|_{L^\infty(\Omega)^d} + \max_{x\in\bar{\Omega}} |\nabla\tilde{y}(x)| \\ &\le Ch|\log h| + \max_{x\in\bar{\Omega}} |\nabla\tilde{y}(x)| \le (1-\varepsilon)\delta,\end{aligned} \tag{4.10}$$

for some $\varepsilon > 0$ and $0 < h \le h_0$, so that $\big(\fint_T A^{-1}\vec{\tilde{v}}_h\big)_{T\in\mathcal{T}_h} \in \vec{C}_h$. The result now follows from [CF93, Thm. 7] with the choices $U = L^2(\Omega)$, $U_{ad} \subset U$ and $\vec{C}_h \subset \mathbb{R}^{nt} \times \mathbb{R}^d$, where nt is the number of elements in $\mathcal{T}_h$. □

Remark 4.1.6. We deduce from (4.9c) that $u_h = P_{[u_a,u_b]}\big(-\frac{1}{\alpha}p_h\big)$, where $P_{[u_a,u_b]}$ denotes the orthogonal projection in $L^2(\Omega)$ onto U_{ad}. Hence, the discrete solution is also a piecewise constant function.

Similarly to Remark 4.1.2 we have

Lemma 4.1.7. *Let $u_h \in U_{ad}$ denote the unique solution of (4.8) with corresponding state $(y_h, \vec{v}_h) = \mathcal{G}_h(u_h)$ and multiplier $(\boldsymbol{\mu}_T)_{T\in\mathcal{T}_h}$. Then there holds*

$$\boldsymbol{\mu}_T = |\boldsymbol{\mu}_T|\frac{1}{\delta}\fint_T A^{-1}\vec{v}_h, \quad T \in \mathcal{T}_h.$$

Proof. Fix $T \in \mathcal{T}_h$. The assertion is clear if $\boldsymbol{\mu}_T = 0$. Suppose that $\boldsymbol{\mu}_T \ne 0$ and define $\vec{c}_h : \bar{\Omega} \to \mathbb{R}^d$ by

$$\vec{c}_{h|\tilde{T}} := \begin{cases} \fint_{\tilde{T}} A^{-1}\vec{v}_h, & \tilde{T} \ne T, \\ \delta\frac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|}, & \tilde{T} = T. \end{cases}$$

Clearly, $\vec{c}_h \in \vec{C}_h$ so that (4.9d) implies

$$\boldsymbol{\mu}_T \cdot \left(\delta\frac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|} - \fint_T A^{-1}\vec{v}_h\right) \le 0,$$

and therefore since $\big(\fint_T A^{-1}\vec{v}_h\big)_{T\in\mathcal{T}_h} \in \vec{C}_h$

$$\delta|\boldsymbol{\mu}_T| \le \boldsymbol{\mu}_T \cdot \fint_T A^{-1}\vec{v}_h \le \delta|\boldsymbol{\mu}_T|.$$

Hence we obtain $\dfrac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|} = \dfrac{1}{\delta}\fint_T A^{-1}\vec{v}_h$ and the lemma is proved. □

As a consequence of Lemma 4.1.7 we immediately infer that

$$|\boldsymbol{\mu}_T| = \boldsymbol{\mu}_T \cdot \frac{1}{\delta}\fint_T A^{-1}\vec{v}_h, \qquad T \in \mathcal{T}_h. \tag{4.11}$$

We now use (4.11) in order to derive an important a–priori estimate.

Lemma 4.1.8. *Let $u_h \in U_{ad}$ be the optimal solution of (4.8) with corresponding state $(y_h, \vec{v}_h) \in Y_h \times \vec{V}_h$ and adjoint variables $(p_h, \vec{\chi}_h) \in Y_h \times \vec{V}_h$, $\boldsymbol{\mu}_T, T \in \mathcal{T}_h$. Then*

$$\|y_h\|_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| \le C \qquad \text{for all } 0 < h \le h_0.$$

Proof. Combining (4.11) with (4.10) we deduce

$$\boldsymbol{\mu}_T \cdot \fint_T A^{-1}\big(\vec{v}_h - \vec{\tilde{v}}_h\big) \ge \delta|\boldsymbol{\mu}_T| - (1-\varepsilon)\delta\,|\boldsymbol{\mu}_T| = \varepsilon\delta|\boldsymbol{\mu}_T|.$$

Choosing $\vec{w}_h = \vec{v}_h - \vec{\tilde{v}}_h$ in (4.9a) and using the symmetry of A as well as the definition of $\mathcal{G}_h$ we hence obtain

$$\begin{aligned}
\varepsilon\delta \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| &\le \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \fint_T A^{-1}\big(\vec{v}_h - \vec{\tilde{v}}_h\big) \\
&= -\int_\Omega A^{-1}\vec{\chi}_h \cdot \big(\vec{v}_h - \vec{\tilde{v}}_h\big) - \int_\Omega p_h \,\mathrm{div}(\vec{v}_h - \vec{\tilde{v}}_h) \\
&= \int_\Omega (y_h - \tilde{y}_h)\mathrm{div}\vec{\chi}_h - \int_\Omega c(y_h - \tilde{y}_h)p_h + \int_\Omega (u_h - \tilde{u})p_h.
\end{aligned}$$

If we use $z_h = y_h - \tilde{y}_h$ in (4.9b) and $v = \tilde{u}$ in (4.9c) we finally deduce

$$\begin{aligned}
\varepsilon\delta \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| &\le -\int_\Omega (y_h - y_0)(y_h - \tilde{y}_h) + \alpha \int_\Omega u_h(\tilde{u} - u_h) \\
&= -\int_\Omega y_h^2 + \int_\Omega y_h(y_0 + \tilde{y}_h) - \int_\Omega y_0\tilde{y}_h + \alpha \int_\Omega u_h(\tilde{u} - u_h) \\
&\le -\frac{1}{2}\int_\Omega y_h^2 - \frac{\alpha}{2}\int_\Omega u_h^2 + C\int_\Omega \big(y_0^2 + \tilde{y}_h^2 + \tilde{u}^2\big),
\end{aligned}$$

where we have used

$$y_h(y_0 + \tilde{y}_h) \le \frac{1}{2}y_h^2 + \frac{1}{2}(y_0 + \tilde{y}_h)^2, \quad y_0\tilde{y}_h \le \frac{1}{2}y_0^2 + \frac{1}{2}\tilde{y}_h^2$$

together with a similar estimate for $u_h(\tilde{u} - u_h)$. This gives the result. □

Remark 4.1.9. For the measure $\vec{\mu}_h \in \mathcal{M}(\bar{\Omega})^d$ defined by

$$\int_{\bar{\Omega}} \vec{f} \cdot d\vec{\mu}_h := \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \fint_T \vec{f}dx, \qquad \vec{f} \in C^0(\bar{\Omega})^d,$$

it follows immediately that

$$\|\vec{\mu}_h\|_{\mathcal{M}(\bar{\Omega})^d} \le C, \quad 0 < h \le h_0.$$

4.1.3. Error analysis.

Theorem 4.1.10. *Let u and u_h be the solutions of (4.1) and (4.8) with corresponding states y and y_h respectively. Then*

$$\|u-u_h\|_{L^2(\Omega)}+\|y-y_h\|_{L^2(\Omega)}\le Ch^{\frac12}|\log h|^{\frac12}$$

for all $0<h\le h_0$.

Proof. Inserting $v=u_h$ into (4.6b) and $v=u$ into (4.9c) we derive

$$\alpha\int_\Omega|u-u_h|^2\le\int_\Omega p(u_h-u)+\int_\Omega p_h(u-u_h)\equiv I+II. \tag{4.12}$$

In order to treat the first term we note that Lemma 4.1.3 with $u=u_h\in L^\infty(\Omega)$, $\big(y^h,\vec v^{\,h}\big)=\mathcal G(u_h)$, and $(y_h,\vec v_h)=\mathcal G_h(u_h)$ yields

$$\|\vec v^{\,h}-\vec v_h\|_{L^\infty(\Omega)^d}\le Ch|\log h|\,\|u_h\|_{L^\infty(\Omega)}. \tag{4.13}$$

Recalling (4.6a) we have

$$\begin{aligned}
I&=\int_\Omega p\big(\mathcal Ay^h-\mathcal Ay\big)\\
&=\int_\Omega(y-y_0)(y^h-y)+\int_{\bar\Omega}\big(\nabla y^h-\nabla y\big)\cdot d\vec\mu\\
&=\int_\Omega(y-y_0)(y^h-y)+\int_{\bar\Omega}\big(P_\delta(\nabla y^h)-\nabla y\big)\cdot d\vec\mu+\int_{\bar\Omega}\big(\nabla y^h-P_\delta(\nabla y^h)\big)\cdot d\vec\mu
\end{aligned}$$

where P_δ denotes the orthogonal projection onto $\bar B_\delta(0)=\{x\in\mathbb R^d\,:\,|x|\le\delta\}$. Note that

$$|P_\delta(x)-P_\delta(\tilde x)|\le|x-\tilde x|\qquad\forall\,x,\tilde x\in\mathbb R^d. \tag{4.14}$$

Since $x\mapsto P_\delta(\nabla y^h(x))\in\vec C$ we infer from (4.6c)

$$I\le\int_\Omega(y-y_0)(y^h-y)+\max_{x\in\bar\Omega}|\nabla y^h(x)-P_\delta(\nabla y^h(x))|\,\|\vec\mu\|_{\mathcal M(\bar\Omega)^d}. \tag{4.15}$$

Let $x\in\bar\Omega$, say $x\in T$ for some $T\in\mathcal T_h$. Since u_h is feasible for (4.8) we have that $⨍_T A^{-1}\vec v_h\in\bar B_\delta(0)$ so that (4.14) implies

$$\begin{aligned}
&\big|\nabla y^h(x)-P_\delta(\nabla y^h(x))\big|\\
&\qquad\le\left|\nabla y^h(x)-⨍_T A^{-1}\vec v_h\right|+\left|P_\delta(\nabla y^h(x))-P_\delta\left(⨍_T A^{-1}\vec v_h\right)\right|\\
&\qquad\le 2\left|\nabla y^h(x)-⨍_T A^{-1}\vec v_h\right|.
\end{aligned} \tag{4.16}$$

Using a well–known interpolation estimate (cf. [BS08, Cor. 4.4.7]), (4.4) and (4.13) we obtain

$$\begin{aligned}
\left|\nabla y^h(x)-⨍_T A^{-1}\vec v_h\right|&\le\left|\nabla y^h(x)-⨍_T\nabla y^h\right|+\left|⨍_T A^{-1}(\vec v^{\,h}-\vec v_h)\right|\\
&\le Ch^{1-\frac dr}\|\nabla y^h\|_{W^{1,r}(\Omega)^d}+C\|\vec v^{\,h}-\vec v_h\|_{L^\infty(\Omega)^d}\\
&\le Crh^{1-\frac dr}\|u_h\|_{L^r(\Omega)}+C\|\vec v^{\,h}-\vec v_h\|_{L^\infty(\Omega)^d}\\
&\le C\big(rh^{1-\frac dr}+h|\log h|\big)\|u_h\|_{L^\infty(\Omega)}
\end{aligned}$$

for $r > d$. Since for $0 < h < 1$

$$h^{1-\frac{d}{|\log h|}}|\log h| = \exp\left(\left(1 - \frac{d}{|\log h|}\right)\log h\right)|\log h| \\ = \exp(\log h)\exp(d)|\log h| = Ch|\log h|$$

with $C = \exp(d)$ holds we deduce, after choosing $r = |\log h|$ and recalling that $u_h \in U_{ad}$, that

$$\left|\nabla y^h(x) - \fint_T A^{-1}\vec{v}_h\right| \leq Ch|\log h|,$$

which combined with (4.15) and (4.16) yields

$$I \leq \int_\Omega (y - y_0)(y^h - y) + Ch|\log h|. \tag{4.17}$$

Next, let us introduce $\big(\hat{y}_h, \vec{\hat{v}}_h\big) := \mathcal{G}_h(u) \in Y_h \times \vec{V}_h$. Using (1.27b) and (4.9a) we infer for the second term

$$\begin{aligned} II &= -\int_\Omega p_h \mathrm{div}\big(\vec{\hat{v}}_h - \vec{v}_h\big) + \int_\Omega c\, p_h\big(\hat{y}_h - y_h\big) \\ &= \int_\Omega A^{-1}\vec{\chi}_h \cdot \big(\vec{\hat{v}}_h - \vec{v}_h\big) + \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \fint_T A^{-1}\big(\vec{\hat{v}}_h - \vec{v}_h\big) + \int_\Omega c\, p_h\big(\hat{y}_h - y_h\big) \\ &= \int_\Omega A^{-1}\vec{\chi}_h \cdot \big(\vec{\hat{v}}_h - \vec{v}_h\big) + \int_\Omega c\, p_h\big(\hat{y}_h - y_h\big) \\ &\quad + \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \Big(P_\delta\big(\fint_T A^{-1}\vec{\hat{v}}_h\big) - \fint_T A^{-1}\vec{v}_h\Big) \\ &\quad + \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot \Big(\fint_T A^{-1}\vec{\hat{v}}_h - P_\delta\big(\fint_T A^{-1}\vec{\hat{v}}_h\big)\Big). \end{aligned}$$

Since

$$\Big(P_\delta\big(\fint_T A^{-1}\vec{\hat{v}}_h\big)\Big)_{T\in\mathcal{T}_h} \in \vec{C}_h$$

we deduce from (4.9d) that

$$\begin{aligned} II \leq \int_\Omega A^{-1}\vec{\chi}_h \cdot \big(\vec{\hat{v}}_h - \vec{v}_h\big) + \int_\Omega c\, p_h\big(\hat{y}_h - y_h\big) \\ + \max_{T\in\mathcal{T}_h}\left|\fint_T A^{-1}\vec{\hat{v}}_h - P_\delta\left(\fint_T A^{-1}\vec{\hat{v}}_h\right)\right| \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T|. \end{aligned}$$

In order to estimate the last term we note that $\nabla y \in \vec{C}$ implies that $\big(\fint_T \nabla y\big)_{T\in\mathcal{T}_h} = \big(\fint_T A^{-1}\vec{v}\big)_{T\in\mathcal{T}_h} \in \vec{C}_h$ and hence again by Lemma 4.1.3, now with $(y, \vec{v}) = \mathcal{G}(u)$, $\big(\hat{y}_h, \vec{\hat{v}}_h\big) = \mathcal{G}_h(u)$,

$$\begin{aligned} \left|\fint_T A^{-1}\vec{\hat{v}}_h - P_\delta\big(\fint_T A^{-1}\vec{\hat{v}}_h\big)\right| &\leq \left|\fint_T A^{-1}(\vec{\hat{v}}_h - \vec{v})\right| + \left|P_\delta\big(\fint_T A^{-1}(\vec{\hat{v}}_h - \vec{v})\big)\right| \\ &\leq C\|\vec{\hat{v}}_h - \vec{v}\|_{L^\infty(\Omega)^d} \leq Ch|\log h|, \end{aligned}$$

which combined with Lemma 4.1.8 yields

$$II \leq \int_\Omega A^{-1}\vec{\chi}_h \cdot \big(\vec{\hat{v}}_h - \vec{v}_h\big) + \int_\Omega c\, p_h\big(\hat{y}_h - y_h\big) + Ch|\log h|.$$

The symmetry of A, (1.27a) and (4.9b) then give

$$\begin{aligned} II &\leq -\int_\Omega \big(\hat{y}_h - y_h\big)\,\mathrm{div}\vec{\chi}_h + \int_\Omega c\,p_h\big(\hat{y}_h - y_h\big) + Ch|\log h| \\ &= \int_\Omega (y_h - y_0)\big(\hat{y}_h - y_h\big) + Ch|\log h|. \end{aligned} \tag{4.18}$$

Inserting (4.17) and (4.18) into (4.12) we finally obtain

$$\begin{aligned} \alpha|u-u_h|^2 &\leq \int_\Omega (y-y_0)\big(y^h - y\big) + \int_\Omega (y_h - y_0)\big(\hat{y}_h - y_h\big) + Ch|\log h| \\ &= -\int_\Omega |y-y_h|^2 + \int_\Omega \big((y_0-y_h)(y-\hat{y}_h) + (y-y_0)(y^h-y_h)\big) + Ch|\log h| \\ &\leq -\int_\Omega |y-y_h|^2 + C\big(\|y-\hat{y}_h\|_{L^2(\Omega)} + \|y^h - y_h\|_{L^2(\Omega)}\big) + Ch|\log h| \\ &\leq -\int_\Omega |y-y_h|^2 + Ch\big(\|u\|_{L^2(\Omega)} + \|u_h\|_{L^2(\Omega)}\big) + Ch|\log h| \end{aligned}$$

in view of (4.7) and the result follows. □

4.1.4. Numerical experiment. In order to have an universal problem for both scenarios 4.0.1 and 4.0.2 at hand, we construct an example for admissible controls in

$$U_{ad} = \{u \in L^2(\Omega) \,:\, -2 \leq u \leq 2 \text{ a.e. in } \Omega\},$$

where the bounds on the control are not active in the analytical solution. This implies $p = -\alpha u$ by equality (4.6b).

Example 4.1.11. We consider (4.1) with the choices $\Omega = B_2(0) \subset \mathbb{R}^2$, $\alpha = 1$,

$$\vec{C} = \{\vec{z} \in C^0(\bar{\Omega})^2 \,:\, |\vec{z}(x)| \leq \frac{1}{2}, x \in \bar{\Omega}\}$$

as well as

$$y_0(x) := \begin{cases} \frac{1}{4} + \frac{1}{2}\ln 2 - \frac{1}{4}|x|^2, & 0 \leq |x| \leq 1, \\ \frac{1}{2}\ln 2 - \frac{1}{2}\ln|x|, & 1 < |x| \leq 2. \end{cases}$$

In order to construct a test example we allow an additional right hand side f in the state equation and replace (4.2) by

$$\begin{aligned} -\Delta y &= f + u \quad \text{in } \Omega \\ y &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

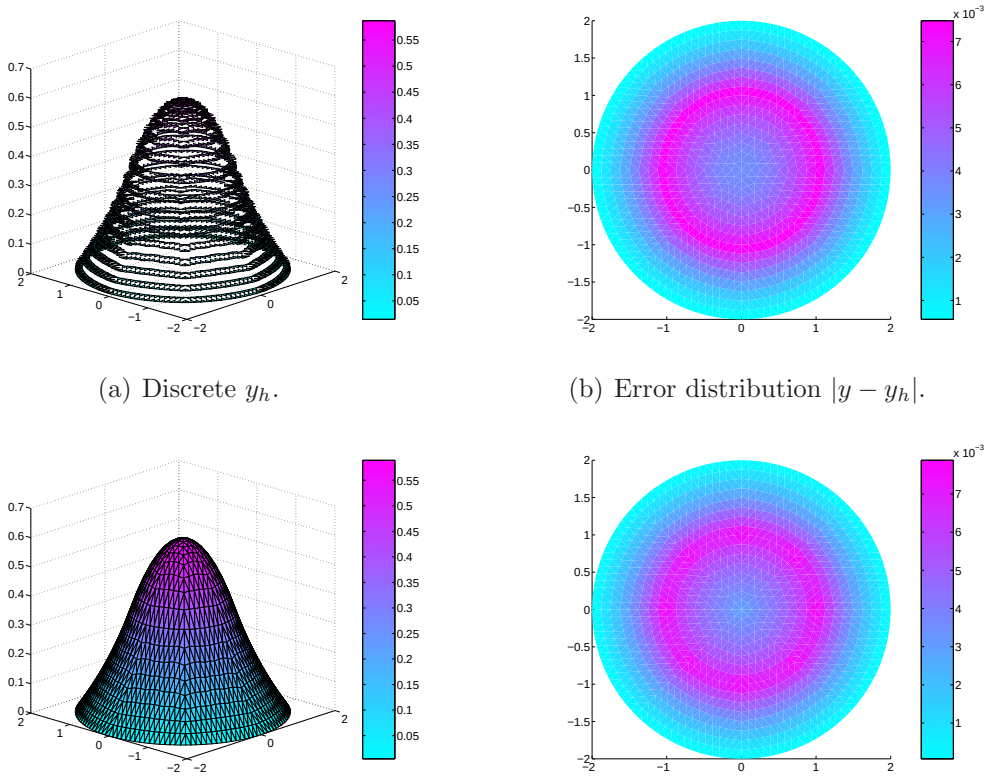
where

$$f(x) := \begin{cases} 2, & 0 \leq |x| \leq 1, \\ 0, & 1 < |x| \leq 2. \end{cases}$$

The optimization problem then has the unique solution

$$u(x) = \begin{cases} -1, & 0 \leq |x| \leq 1, \\ 0, & 1 < |x| \leq 2 \end{cases}$$

with corresponding state $y \equiv y_0$. The action of the measure $\vec{\mu}$ applied to a vectorfield $\vec{\phi} \in C^0(\bar{\Omega})^2$ is given by $\int_{\bar{\Omega}} \vec{\phi} \cdot d\vec{\mu} = -\int_{\partial B_1(0)} x \cdot \vec{\phi} dS$.



(a) Discrete y_h.

(b) Error distribution $|y - y_h|$.

(c) Post–processed y_h^P.

(d) Error distribution $|y - y_h^P|$.

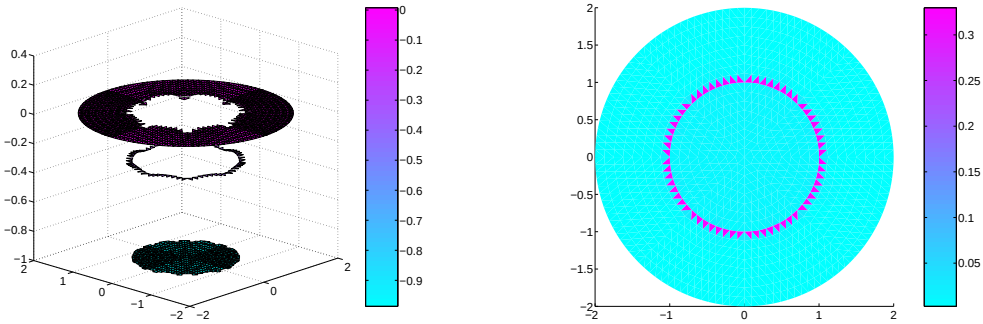
FIGURE 4.1. Optimal state.

For the numerical solution we use the routine `fmincon` contained in the Matlab optimization toolbox. The state equation was approximated with the help of the Matlab implementation of the lowest order Raviart–Thomas element provided by [BC05]. In Figures 4.1 – 4.4 we present the numerical approximations $y_h, y_h^P, u_h, \vec{v}_h$ and $\vec{\mu}_h$ on a grid containing $m = 1089$ gridpoints. Figure 4.4 clearly shows that the support of $\vec{\mu}_h$ is concentrated around $|x| = 1$. We mention that the used meshes have the same topology as the piecewise $\mathcal{O}(h^2)$ irregular meshes from Table 2.2 for triangulating $B_1(0)$. The only difference is that the mesh parameter h has doubled for $\Omega = B_2(\Omega)$.

In Table 4.1 we investigate the experimental orders of convergence (EOCs) for the error functionals

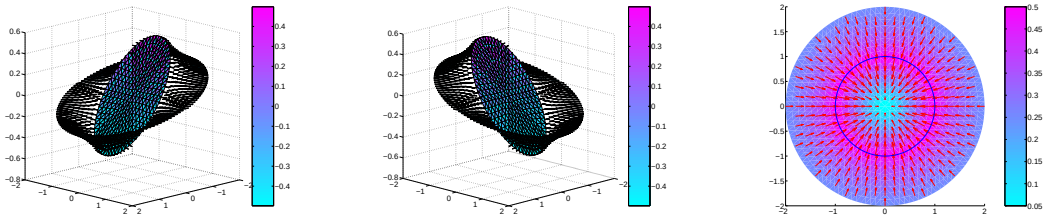
$$E_u(h) := \|u - u_h\|_{L^2(\Omega)}, \quad E_y(h) := \|y - y_h\|_{L^2(\Omega)}, \text{ and } E_y^P(h) := \|y - y_h^P\|_{L^2(\Omega)},$$

where the superscript P is assigned to the piecewise linearly post–processed state associated to u_h. It turns out that the controls show the behavior predicted by Theorem 4.1.10, whereas the L^2-norm of the state seems to converge linearly. The post–processed state shows the same order of convergence, but has a smaller error. In Table 4.1 we also display the values of $\sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T|$, where $(\boldsymbol{\mu}_T)_{T\in\mathcal{T}_h}$ is given by (4.11). These values are expected to converge to 2π as $h \to 0$, since this gives the value of μ applied to the function which is identically equal to 1 on $\bar{\Omega}$.



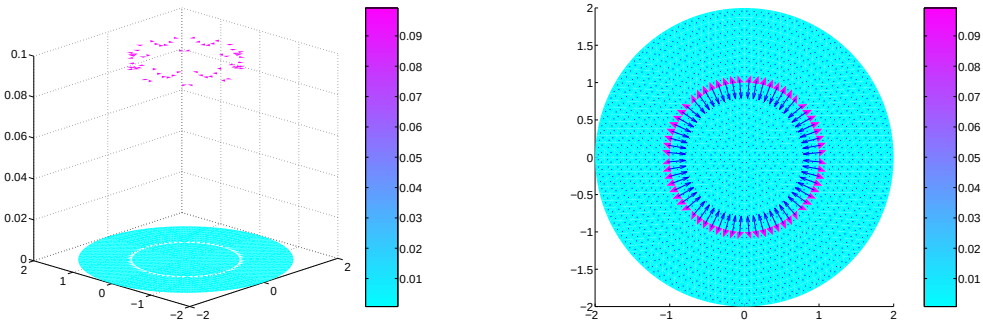
(a) Discrete u_h. (b) Error distribution $|u - u_h|$.

FIGURE 4.2. Optimal control.



(a) $(\vec{v}_h)_1$. (b) $(\vec{v}_h)_2$. (c) $\vec{v}_h$.

FIGURE 4.3. Discrete gradient approximation $\vec{v}_h$ to the state.



(a) $\left(|\boldsymbol{\mu}_T|\right)_{T\in\mathcal{T}_h}$. (b) $\vec{\mu}_h$.

FIGURE 4.4. Discrete measure.

j	h	$\sum_{T\in\mathcal{T}_h} \lvert\boldsymbol{\mu}_T\rvert$	$E_u(h)$	EOC	$E_y(h)$	EOC	$E_y^P(h)$	EOC
2	1.14214	5.024	0.729649	-	0.302178	-	0.137428	-
3	0.60439	5.891	0.389627	0.986	0.153204	1.067	0.068697	1.089
4	0.31017	6.138	0.275764	0.518	0.077299	1.025	0.032998	1.099
5	0.15703	6.222	0.196169	0.500	0.038752	1.014	0.015806	1.081

TABLE 4.1. Multiplier approximation, errors and EOCs for the controls, the state and the piecewise linearly post–processed state.

4.2. Variational discrete and piecewise constant control approximations for Scenario 4.0.2

The present work from [GH09] complements the discrete approach to elliptic optimal control problems with gradient constraints presented in Section 4.1 from [DGH09c]. We consider the elliptic optimal control problem (4.1) with pointwise bounds on the gradient of the state for Scenario 4.0.2. To guarantee the required regularity of the state we include the L^r-norm of the control in our cost functional with $r > d$, $(d = 2, 3)$. We investigate variational discretization of the control problem [Hin05] as well as piecewise constant approximations of the control. In both cases we use standard piecewise linear and continuous finite elements for the discretization of the state. Pointwise bounds on the gradient of the discrete state are enforced element-wise. Error bounds for control and state are obtained in two and three space dimensions depending on the value of r. In the presence of gradient constraints variational discretization of the controls automatically leads to globally continuous approximations of the controls, if globally continuous Ansatz functions for the state are used, see relation (4.29). This is certainly a drawback of the approach, since the optimal control and the associated adjoint state may develop jumps, as the example in Section 4.2.4 shows. Piecewise constant control approximations here seem to be the better choice. However, the approximation order in both cases is the same, and also the errors in the numerical experiments for both approaches are of similar size, see Tables 4.2, 4.3.

4.2.1. Mathematical setting. For the convenience of the reader we recall the assumptions made in Scenario 4.0.2. Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded, convex polyhedral domain with boundary $\partial\Omega$, whose inner dihedral angles at $\partial\Omega$ in the case $d = 3$ are assumed to be smaller than $\frac{3}{4}\pi$. We consider the differential operator $\mathcal{A} := -\Delta$ and associate to it the bilinear form

$$a(y, \phi) := \int_\Omega \nabla y \cdot \nabla \phi \qquad \forall y, \phi \in H^1(\Omega).$$

With the above assumptions we conclude that there exists some $\bar{r} > d$ such that for a given $u \in L^r(\Omega)$ $(1 < r \leq \bar{r})$ the elliptic boundary value problem

$$\begin{aligned} \mathcal{A}y &= u \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.19}$$

admits a unique solution $y = \mathcal{G}(u) \in W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$ (see [Dau92] for $d = 3$, and [Gri92] for $d = 2$). Furthermore,

$$\|y\|_{W^{2,r}(\Omega)} \leq C\|u\|_{L^r(\Omega)}$$

holds. Moreover, for $u \in W^{-1,r}(\Omega)$ we have $\mathcal{G}(u) \in W_0^{1,r}(\Omega)$ (see [Grö89] for $d = 2$, and [JK95] for $d = 3$) with

$$\|y\|_{W^{1,r}(\Omega)} \leq C\|u\|_{W^{-1,r}(\Omega)},$$

where the positive constant C is independent of u.

Remark 4.2.1. We think that our considerations also carry over to more general elliptic operators

$$\mathcal{A}y = -\sum_{i,j=1}^{d} \partial_{x_j}\big(a_{ij} y_{x_i}\big) + cy,$$

with sufficiently smooth coefficients a_{ij} and c, and to curved domains Ω with sufficiently smooth boundary.

If not specified otherwise, let $d < r < \infty$, $\alpha > 0$ and $y_0 \in L^2(\Omega)$ be given. We now consider the control problem

$$\min_{u \in L^r(\Omega)} J(u) = \frac{1}{2} \int_\Omega |y - y_0|^2 + \frac{\alpha}{r} \int_\Omega |u|^r \quad \text{subject to } y = \mathcal{G}(u) \text{ and } \nabla y \in \vec{C}. \tag{4.20}$$

Here,

$$\vec{C} = \{\vec{z} \in C^0(\bar{\Omega})^d \,:\, |\vec{z}(x)| \leq \delta, x \in \bar{\Omega}\}. \tag{4.21}$$

Since $r > d$ we have $y \in W^{2,r}(\Omega)$ and hence $\nabla y \in C^0(\bar{\Omega})^d$ by a well–known embedding result. We impose the following Slater condition:

$$\exists\, \tilde{u} \in L^r(\Omega) : \quad |\nabla \tilde{y}(x)| < \delta,\ x \in \bar{\Omega},\ \text{where } \tilde{y} = \mathcal{G}(\tilde{u}). \tag{4.22}$$

Since J is strictly convex and the set of admissible controls and states forms a closed and convex set, problem (4.20) admits a unique solution u with associated state $y = \mathcal{G}(u)$.

The KKT system of problem (4.20) is obtained with the help of [CF93, Cor. 1]. There holds

Theorem 4.2.2. *An element $u \in L^r(\Omega)$ is a solution of (4.20) if and only if there exist $\vec{\mu} \in \mathcal{M}(\bar{\Omega})^d$ and $p \in L^t(\Omega)$ $(t < \frac{d}{d-1})$ such that*

$$\int_\Omega p \mathcal{A}\phi - \int_\Omega (y - y_0)\phi - \int_{\bar{\Omega}} \nabla\phi \cdot d\vec{\mu} = 0 \qquad \forall\, \phi \in W^{2,t'}(\Omega) \cap W_0^{1,t'}(\Omega) \tag{4.23a}$$

$$p + \alpha |u|^{r-2} u = 0 \qquad \text{in } \Omega \tag{4.23b}$$

$$\int_{\bar{\Omega}} (\vec{z} - \nabla y) \cdot d\vec{\mu} \leq 0 \qquad \forall\, \vec{z} \in \vec{C}. \tag{4.23c}$$

Here, y is the solution of (4.19), $\frac{1}{t} + \frac{1}{t'} = 1$, and $\mathcal{M}(\bar{\Omega})$ denotes the space of regular Borel *measures.*

Again as in Remark 4.1.2 the vector valued measure $\vec{\mu}$ can be represented by some nonnegative $\mu \in \mathcal{M}(\bar{\Omega})$ such that

$$\vec{\mu} = \frac{1}{\delta} \nabla y\, \mu \tag{4.24}$$

holds.

4.2.2. Finite element discretization. We sketch an approach from Section 3.3.2 of the book [HPUU09] which uses classical piecewise linear, continuous approximations of the states.

Let $\mathcal{T}_h$ denote by Definition 1.2.4 a quasi-uniform triangulation of Ω with maximum mesh size h. We choose the space of linear finite elements $Y_h := P^1_{c,h}(\mathcal{T}_h)$ and let $Y_{h0} := Y_h \cap H^1_0(\Omega)$. Furthermore let us recall the definition of the discrete approximation of the operator $\mathcal{G}$. For a given function $v \in L^2(\Omega)$ we denote by $z_h = \mathcal{G}_h(v) \in Y_{h0}$ the solution of

$$a(z_h, \phi_h) = \int_\Omega v\phi_h \quad \text{for all } \phi_h \in Y_{h0}.$$

It is well–known that for all $v \in L^r(\Omega)$ by an embedding theorem the corresponding state $\mathcal{G}(v)$ is in $W^{1,\infty}(\Omega)$, where we recall $r > d$. Furthermore, using [GLRS09, (1.2)] and [BS08, (4.4.29)]

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\|_{W^{1,\infty}(\Omega)} \leq C \inf_{z_h \in Y_{h0}} \|\mathcal{G}(v) - z_h\|_{W^{1,\infty}(\Omega)} \leq C h^{1-\frac{d}{r}} \|\mathcal{G}(v)\|_{W^{2,r}(\Omega)}$$

$$\leq C h^{1-\frac{d}{r}} \|v\|_{L^r(\Omega)}. \tag{4.25}$$

For each $T \in \mathcal{T}_h$ let $\mathbf{z}_T \in \mathbb{R}^d$ denote constant vectors. We define

$$\vec{C}_h := \{\vec{z}_h : \Omega \to \mathbb{R}^d \,:\, \vec{z}_{h|T} = \mathbf{z}_T \text{ on } T \text{ and } |\vec{z}_{h|T}| \leq \delta,\, T \in \mathcal{T}_h\}.$$

4.2.2.1. *Variational discretization.* Let us first consider variational discretization of problem (4.20) which reads:

$$\begin{aligned} &\min_{u \in L^r(\Omega)} J_h(u) := \frac{1}{2}\int_\Omega |y_h - y_0|^2 + \frac{\alpha}{r}\int_\Omega |u|^r \\ &\text{subject to } \; y_h = \mathcal{G}_h(u) \text{ and } \nabla y_h \in \vec{C}_h. \end{aligned} \tag{4.26}$$

Now (4.25) implies that $\tilde{y}_h := \mathcal{G}_h(\tilde{u})$ satisfies the Slater condition

$$|\nabla \tilde{y}_h(x)| < \delta \text{ for all } x \in \bar{\Omega}, \tag{4.27}$$

and for $0 < h \leq h_0$ with $h_0 > 0$ small enough. This delivers

Lemma 4.2.3. *Problem* (4.26) *admits a unique solution* $u_h \in L^r(\Omega)$. *There exist* $\boldsymbol{\mu}_T \in \mathbb{R}^d$ *for all* $T \in \mathcal{T}_h$ *and* $p_h \in Y_{h0}$ *such that with* $y_h = \mathcal{G}_h(u_h)$ *we have*

$$a(\phi_h, p_h) - \int_\Omega (y_h - y_0)\phi_h - \sum_{T \in \mathcal{T}_h} \nabla \phi_{h|T} \cdot \boldsymbol{\mu}_T = 0 \qquad \forall\, \phi_h \in Y_{h0}, \tag{4.28a}$$

$$p_h + \alpha |u_h|^{r-2} u_h = 0 \qquad \text{in } \Omega, \tag{4.28b}$$

$$\sum_{T \in \mathcal{T}_h} \big(\mathbf{z}_T - \nabla y_{h|T}\big) \cdot \boldsymbol{\mu}_T \leq 0 \qquad \forall\, \vec{z}_h \in \vec{C}_h. \tag{4.28c}$$

In problem (4.26) we apply variational discretization of [Hin05]. From (4.28b) we infer for the discrete optimal control

$$u_h = -\alpha^{-\frac{1}{r-1}} |p_h|^{\frac{2-r}{r-1}} p_h. \tag{4.29}$$

In order to numerically implement the infinite dimensional optimal control problem (4.26) we essentially make use of this equation. The treatment of the nonlinear dependence between u_h and the finite element object p_h is strongly challenging. An explanation of the numerical solution algorithm is part of Section 4.2.4 and the Appendix C.

Furthermore, according to (4.24) we have the following analogon to Lemma 4.1.7 as representation of the discrete multipliers.

Lemma 4.2.4. *Let* u_h *denote the unique solution of* (4.26) *with corresponding state* $y_h = \mathcal{G}_h(u_h)$ *and multiplier* $(\boldsymbol{\mu}_T)_{T \in \mathcal{T}_h}$. *Then there holds*

$$\boldsymbol{\mu}_T = |\boldsymbol{\mu}_T| \frac{1}{\delta} \nabla y_{h|T} \text{ for all } T \in \mathcal{T}_h. \tag{4.30}$$

Proof. Fix $T \in \mathcal{T}_h$. The assertion is clear if $\boldsymbol{\mu}_T = 0$. Suppose that $\boldsymbol{\mu}_T \neq 0$ and define $\vec{z}_h : \bar{\Omega} \to \mathbb{R}^d$ by

$$\vec{z}_{h|\tilde{T}} := \begin{cases} \nabla y_{h|T}, & \tilde{T} \neq T, \\ \delta \frac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|}, & \tilde{T} = T. \end{cases}$$

Clearly, $\vec{z}_h \in \vec{C}_h$ so that (4.28c) implies

$$\boldsymbol{\mu}_T \cdot \left(\delta \frac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|} - \nabla y_{h|T} \right) \leq 0,$$

and therefore, since $\big(\nabla y_{h|T}\big)_{T\in\mathcal{T}_h} \in \vec{C}_h$,

$$\delta|\boldsymbol{\mu}_T| \leq \boldsymbol{\mu}_T \cdot \nabla y_{h|T} \leq \delta|\boldsymbol{\mu}_T|.$$

Hence we obtain $\dfrac{\boldsymbol{\mu}_T}{|\boldsymbol{\mu}_T|} = \dfrac{1}{\delta}\nabla y_{h|T}$ and the lemma is proved. □

As a consequence of Lemma 4.2.4 we immediately infer that

$$|\boldsymbol{\mu}_T| = \boldsymbol{\mu}_T \cdot \frac{1}{\delta}\nabla y_{h|T} \text{ for all } T \in \mathcal{T}_h. \tag{4.31}$$

We now use (4.31) in order to derive an important a priori estimate.

Lemma 4.2.5. *Let $u_h \in L^r(\Omega)$ be the optimal solution of (4.26) with corresponding state $y_h \in Y_{h0}$ and adjoint variables $p_h \in Y_{h0}$, $\boldsymbol{\mu}_T \in \mathbb{R}^d, T \in \mathcal{T}_h$. Then there exists $h_0 > 0$ such that*

$$\|y_h\|_{L^2(\Omega)} + \|u_h\|_{L^r(\Omega)} + \|p_h\|_{L^{\frac{r}{r-1}}(\Omega)} + \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| \leq C \qquad \textit{for all } 0 < h \leq h_0.$$

Proof. Combining (4.31) with (4.27) we deduce

$$\boldsymbol{\mu}_T \cdot (\nabla y_{h|T} - \nabla \tilde{y}_{h|T}) \geq \delta|\boldsymbol{\mu}_T| - (1-\varepsilon)\delta\,|\boldsymbol{\mu}_T| = \varepsilon\delta|\boldsymbol{\mu}_T|.$$

Choosing $\phi_h = y_h - \tilde{y}_h$ in (4.28a) and using the definition of $\mathcal{G}_h$ together with (4.28b) we hence obtain

$$\begin{aligned}
\varepsilon\delta \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| &\leq \sum_{T\in\mathcal{T}_h} \boldsymbol{\mu}_T \cdot (\nabla y_{h|T} - \nabla \tilde{y}_{h|T}) \\
&= a(y_h - \tilde{y}_h, p_h) - \int_\Omega (y_h - y_0)(y_h - \tilde{y}_h) \\
&= \int_\Omega (u_h - \tilde{u}) p_h - \int_\Omega (y_h - y_0)(y_h - \tilde{y}_h) \\
&= -\alpha \int_\Omega |u_h|^r + \alpha \int_\Omega |u_h|^{r-2} u_h \tilde{u} - \int_\Omega y_h^2 + \int_\Omega y_h(y_0 + \tilde{y}_h) - \int_\Omega y_0 \tilde{y}_h \\
&\leq -\alpha \int_\Omega |u_h|^r + \alpha \|u_h^{r-1}\|_{L^{\frac{r}{r-1}}(\Omega)} \|\tilde{u}\|_{L^r(\Omega)} - \frac{1}{2}\int_\Omega y_h^2 - y_0^2 - \tilde{y}_h^2 \\
&\leq -\frac{\alpha}{2} \int_\Omega |u_h|^r - \frac{1}{2}\int_\Omega |y_h|^2 + C\big(1 + \|y_0\|^2_{L^2(\Omega)} + \|\tilde{u}\|^r_{L^r(\Omega)}\big),
\end{aligned}$$

where we have used $y_h(y_0 + \tilde{y}_h) \leq \frac{1}{2}y_h^2 + \frac{1}{2}(y_0 + \tilde{y}_h)^2$. This implies the bounds on y_h, u_h and $\boldsymbol{\mu}_T$. The bound on p_h follows from (4.28b). □

Remark 4.2.6. For the measure $\vec{\mu}_h \in \mathcal{M}(\bar{\Omega})^d$ defined by

$$\int_{\bar{\Omega}} \vec{f} \cdot \, d\vec{\mu}_h := \sum_{T\in\mathcal{T}_h} \int_T \vec{f}\, dx \cdot \boldsymbol{\mu}_T \text{ for all } \vec{f} \in C^0(\bar{\Omega})^d,$$

it follows immediately that

$$\|\vec{\mu}_h\|_{\mathcal{M}(\bar{\Omega})^d} \leq C.$$

4.2.3. Error analysis. Now we are in the position to prove the following error estimates.

Theorem 4.2.7. *Let u and u_h be the solutions of* (4.20) *and* (4.26) *respectively. Then there exists $h_1 \le h_0$ such that*

$$\|y - y_h\|_{L^2(\Omega)} \le Ch^{\frac{1}{2}(1-\frac{d}{r})}, \text{ and } \|u - u_h\|_{L^r(\Omega)} \le Ch^{\frac{1}{r}(1-\frac{d}{r})}$$

for all $0 < h \le h_1$.

Proof. Let us introduce $y^h := \mathcal{G}(u_h) \in W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$, and $\hat{y}_h := \mathcal{G}_h(u)$. In view of Lemma 4.2.5 and (4.25) we have

$$\|y^h - y_h\|_{W^{1,\infty}(\Omega)} \le Ch^{1-\frac{d}{r}} \|u_h\|_{L^r(\Omega)} \le Ch^{1-\frac{d}{r}}. \tag{4.32}$$

Let us now turn to the actual error estimate. To begin, we recall that for $r \ge 2$ there exists $\theta_r > 0$ such that

$$(|a|^{r-2}a - |b|^{r-2}b)(a-b) \ge \theta_r |a-b|^r \qquad \forall\, a, b \in \mathbb{R}.$$

Hence, using (4.23b) and (4.28b),

$$\begin{aligned}\alpha\theta_r \int_\Omega |u - u_h|^r &\le \alpha \int_\Omega \big(|u|^{r-2}u - |u_h|^{r-2}u_h\big)(u - u_h) \\ &= \int_\Omega p(u_h - u) + \int_\Omega p_h(u - u_h) \equiv I + II.\end{aligned}$$

Recalling (4.23a) we have

$$\begin{aligned} I &= \int_\Omega p\big(\mathcal{A}y^h - \mathcal{A}y\big) \\ &= \int_\Omega (y - y_0)(y^h - y) + \int_{\bar\Omega} \big(\nabla y^h - \nabla y\big) \cdot d\vec{\mu} \\ &= \int_\Omega (y - y_0)(y^h - y) + \underbrace{\int_{\bar\Omega} \big(P_\delta(\nabla y^h) - \nabla y\big) \cdot d\vec{\mu}}_{\le 0} + \int_{\bar\Omega} \big(\nabla y^h - P_\delta(\nabla y^h)\big) \cdot d\vec{\mu} \end{aligned}$$

where P_δ denotes the orthogonal projection onto $\bar{B}_\delta(0) = \{x \in \mathbb{R}^d : |x| \le \delta\}$. Note that

$$|P_\delta(x) - P_\delta(\tilde{x})| \le |x - \tilde{x}| \qquad \forall\, x, \tilde{x} \in \mathbb{R}^d. \tag{4.33}$$

Since $x \mapsto P_\delta(\nabla y^h(x)) \in \vec{C}$ we infer from (4.23c)

$$I \le \int_\Omega (y - y_0)(y^h - y) + \max_{x \in \bar\Omega} |\nabla y^h(x) - P_\delta(\nabla y^h(x))| \, \|\vec{\mu}\|_{\mathcal{M}(\bar\Omega)^d}. \tag{4.34}$$

Let $x \in \bar\Omega$, say $x \in T$ for some $T \in \mathcal{T}_h$. Since u_h is feasible for (4.26) we have that $\nabla y_{h|T} \in \bar{B}_\delta(0)$ so that (4.33) together with (4.32) implies

$$\begin{aligned} \big|\nabla y^h(x) - P_\delta(\nabla y^h(x))\big| &\le \big|\nabla y^h(x) - \nabla y_{h|T}\big| + \big|P_\delta(\nabla y^h(x)) - P_\delta\big(\nabla y_{h|T}\big)\big| \\ &\le 2\big|\nabla y^h(x) - \nabla y_{h|T}\big| \le Ch^{1-\frac{d}{r}} \|u_h\|_{L^r(\Omega)}. \end{aligned} \tag{4.35}$$

Thus

$$I \le \int_\Omega (y - y_0)(y^h - y) + Ch^{1-\frac{d}{r}}. \tag{4.36}$$

Similarly,

$$
\begin{aligned}
II = a(\hat{y}_h - y_h, p_h) &= \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + \sum_{T\in\mathcal{T}_h} \big(\nabla \hat{y}_{h|T} - \nabla y_{h|T}\big)\cdot \boldsymbol{\mu}_T \\
&= \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + \sum_{T\in\mathcal{T}_h} \big(\nabla \hat{y}_{h|T} - P_\delta(\nabla \hat{y}_{h|T})\big)\cdot \boldsymbol{\mu}_T \\
&\quad + \underbrace{\sum_{T\in\mathcal{T}_h} \big(P_\delta(\nabla \hat{y}_{h|T}) - \nabla y_{h|T}\big)\cdot \boldsymbol{\mu}_T}_{\leq 0} \\
&\leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + \sum_{T\in\mathcal{T}_h} \big(\nabla \hat{y}_{h|T} - \nabla y(x_T)\big)\cdot \boldsymbol{\mu}_T \\
&\quad + \sum_{T\in\mathcal{T}_h} \big(P_\delta(\nabla y(x_T)) - P_\delta(\nabla \hat{y}_{h|T})\big)\cdot \boldsymbol{\mu}_T,
\end{aligned}
$$

where $x_T \in T$, so that $\big(\nabla y(x_T)\big)_{T\in\mathcal{T}_h} \in \vec{C}_h$. We infer from Lemma 4.2.5 and (4.25)

$$
\begin{aligned}
II &\leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + 2\max_{T\in\mathcal{T}_h} |\nabla \hat{y}_{h|T} - \nabla y(x_T)| \sum_{T\in\mathcal{T}_h} |\boldsymbol{\mu}_T| \\
&\leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + Ch^{1-\frac{d}{r}} \|u\|_{L^r(\Omega)}. \qquad (4.37)
\end{aligned}
$$

Combining I and II we finally obtain

$$
\begin{aligned}
\alpha\theta_r \int_\Omega |u - u_h|^r &\leq \int_\Omega (y - y_0)\big(y^h - y\big) + \int_\Omega (y_h - y_0)\big(\hat{y}_h - y_h\big) + Ch^{1-\frac{d}{r}} \\
&= -\int_\Omega |y - y_h|^2 + \int_\Omega \big((y_0 - y_h)(y - \hat{y}_h) + (y - y_0)(y^h - y_h)\big) \\
&\quad + Ch^{1-\frac{d}{r}} \\
&\leq -\int_\Omega |y - y_h|^2 + C\big(\|y - \hat{y}_h\|_{L^2(\Omega)} + \|y^h - y_h\|_{L^2(\Omega)}\big) + Ch^{1-\frac{d}{r}} \\
&\leq -\int_\Omega |y - y_h|^2 + Ch\big(\|u\|_{L^2(\Omega)} + \|u_h\|_{L^2(\Omega)}\big) + Ch^{1-\frac{d}{r}}
\end{aligned}
$$

and the result follows. □

Remark 4.2.8. Theorem 4.2.7 suggests to use the coupling $r = 2d$ to obtain the best convergence order for the control error. This would deliver errors of magnitude $\mathcal{O}(h^{1/8})$ for $d = 2$ and of magnitude $\mathcal{O}(h^{1/12})$ for $d = 3$. We note that our numerical results for $d = 2$ deliver $\mathcal{O}(h^{1/4})$. However, presently we are not able to prove this result for the control problems (4.20), (4.26).

4.2.3.1. *Piecewise constant controls.* Let us now consider the following optimal control problem with piecewise constant controls as discretization of problem (4.20);

$$
\begin{aligned}
&\min_{u_h\in U_h} J_h(u_h) := \frac{1}{2}\int_\Omega |y_h - y_0|^2 + \frac{\alpha}{r}\int_\Omega |u_h|^r \\
&\text{subject to } \; y_h = \mathcal{G}_h(u_h) \text{ and } \nabla y_h \in \vec{C}_h,
\end{aligned} \qquad (4.38)
$$

where $U_h := \{v_h \in L^r(\Omega) \,:\, v_{h|T} \in \mathbb{R} \text{ for all } T \in \mathcal{T}_h\}$. It is not difficult to prove that this problem admits a unique solution $u_h \in U_h$. Our finite element error analysis for

this problem is based on approximation properties of the orthogonal L^2-projection $Q_h : L^2(\Omega) \to U_h$ defined by

$$(Q_h v)(x) := \fint_T v = \frac{1}{|T|}\int_T v \text{ for all } v \in L^2(\Omega), x \in T.$$

For $v \in L^r(\Omega)$ we have the stability estimate

$$\|Q_h v\|_{L^r(\Omega)} = \left(\sum_{T\in\mathcal{T}_h} |T|^{1-r}\left|\int_T v\right|^r\right)^{\frac{1}{r}} \le \left(\sum_{T\in\mathcal{T}_h} |T|^{1-r}\left|\|1\cdot v\|_{L^1(T)}\right|^r\right)^{\frac{1}{r}} \le \left(\sum_{T\in\mathcal{T}_h}\|v\|^r_{L^r(T)}\right)^{\frac{1}{r}} = \|v\|_{L^r(\Omega)}, \tag{4.39}$$

and for $\phi \in W^{1,r}(\Omega)$ the approximation property

$$\|\phi - Q_h\phi\|_{L^r(\Omega)} \le Ch^l\|\phi\|_{W^{l,r}(\Omega)}, \quad 0 \le l \le 1, \tag{4.40}$$

holds, see [EG04, Prop. 1.135]. Furthermore,

$$\|\mathcal{G}(v) - \mathcal{G}_h(Q_h v)\|_{W^{1,\infty}(\Omega)} \le \|\mathcal{G}(v) - \mathcal{G}(Q_h v)\|_{W^{1,\infty}(\Omega)} + \|\mathcal{G}(Q_h v) - \mathcal{G}_h(Q_h v)\|_{W^{1,\infty}(\Omega)} \equiv I + II.$$

Now, for $v \in L^r(\Omega)$, by (4.25) and (4.39) there holds

$$II \le Ch^{1-\frac{d}{r}}\|Q_h v\|_{L^r(\Omega)} \le Ch^{1-\frac{d}{r}}\|v\|_{L^r(\Omega)}.$$

Furthermore

$$\|\nabla\mathcal{G}(v - Q_h v)\|_{L^\infty(\Omega)^d} \le C\|\nabla\mathcal{G}(v - Q_h v)\|^\beta_{L^r(\Omega)^d}|\nabla\mathcal{G}(v - Q_h v)|^{1-\beta}_{W^{1,r}(\Omega)^d} \le C\|v - Q_h v\|^\beta_{W^{-1,r}(\Omega)}\|v - Q_h v\|^{1-\beta}_{L^r(\Omega)},$$

where we have used the Lyapunov inequality ([Fri69, Thm. 10.1]) with $0 < \beta := 1 - \frac{d}{r} < 1$. Now, for $w \in W^{1,r'}(\Omega)$ with $\frac{1}{r} + \frac{1}{r'} = 1$ we have

$$\int_\Omega (v - Q_h v)w = \int_\Omega (v - Q_h v)(w - Q_h w) \le \|v - Q_h v\|_{L^r(\Omega)}\|w - Q_h w\|_{L^{r'}(\Omega)} \le Ch\|v - Q_h v\|_{L^r(\Omega)}\|w\|_{W^{1,r'}(\Omega)}.$$

This yields

$$\|v - Q_h v\|_{W^{-1,r}(\Omega)} = \sup_{0\neq w\in W^{1,r'}(\Omega)} \frac{\int_\Omega (v - Q_h v)w}{\|w\|_{W^{1,r'}(\Omega)}} \le Ch\|v\|_{L^r(\Omega)},$$

so that we obtain again by (4.40)

$$\|\nabla\mathcal{G}(v - Q_h v)\|_{L^\infty(\Omega)^d} \le Ch^{1-\frac{d}{r}}\|v\|_{L^r(\Omega)}.$$

Hence I can also be estimated by

$$I = \|\mathcal{G}(v - Q_h v)\|_{W^{1,\infty}(\Omega)} \le C\|\nabla\mathcal{G}(v - Q_h v)\|_{L^\infty(\Omega)^d} \le Ch^{1-\frac{d}{r}}\|v\|_{L^r(\Omega)}.$$

Finally we conclude

$$\|\mathcal{G}(v) - \mathcal{G}_h(Q_h v)\|_{W^{1,\infty}(\Omega)} \le Ch^{1-\frac{d}{r}}\|v\|_{L^r(\Omega)}. \tag{4.41}$$

Thus, with $v := \tilde{u} \in L^r(\Omega)$ we have that for $h > 0$ small enough the function $\tilde{y}_h := \mathcal{G}_h(Q_h v)$ satisfies the Slater condition (4.27). For the optimal control problem (4.38) the result of Lemma 4.2.3 is valid if we replace (4.28b) by

$$\int_\Omega (p_h + \alpha |u_h|^{r-2} u_h) v_h = 0 \qquad \forall\, v_h \in U_h. \tag{4.42}$$

Furthermore Lemma 4.2.4 holds accordingly and the analogon to Lemma 4.2.5 reads

Lemma 4.2.9. *Let $u_h \in U_h$ be the optimal solution of (4.38) with corresponding state $y_h \in Y_{h0}$ and adjoint variables $p_h \in Y_{h0}$, $\boldsymbol{\mu}_T, T \in \mathcal{T}_h$. Then there exists $h_0 > 0$ such that*

$$\|y_h\|_{L^2(\Omega)} + \|u_h\|_{L^r(\Omega)} + \sum_{T \in \mathcal{T}_h} |\boldsymbol{\mu}_T| \le C \qquad \text{for all } 0 < h \le h_0$$

holds.

Proof. Since $0 \le J_h(u_h) \le J_h(Q_h \tilde{u}) \le C$ uniformly in h we have

$$\|y_h\|_{L^2(\Omega)} + \|u_h\|_{L^r(\Omega)} \le C \quad \text{for all } 0 < h \le h_0.$$

We continue with the estimate

$$\begin{aligned}
\boldsymbol{\mu}_T \cdot (\nabla y_{h|T} - \nabla \tilde{y}_{h|T}) &= \delta |\boldsymbol{\mu}_T| - |\boldsymbol{\mu}_T| \frac{1}{\delta} \nabla y_{h|T} \cdot \nabla \tilde{y}_{h|T} \\
&\ge \delta |\boldsymbol{\mu}_T| - |\boldsymbol{\mu}_T| \, |\nabla \tilde{y}_{h|T}| \\
&\ge \delta |\boldsymbol{\mu}_T| - (\delta - \frac{\varepsilon}{4}) |\boldsymbol{\mu}_T| = \frac{\varepsilon}{4} |\boldsymbol{\mu}_T|,
\end{aligned}$$

for some $\varepsilon > 0$. Choosing $\phi_h = y_h - \tilde{y}_h$ in (4.28a) and using the definition of $\mathcal{G}_h$ together with (4.42) we hence obtain

$$\begin{aligned}
\frac{\varepsilon}{4} \sum_{T \in \mathcal{T}_h} |\boldsymbol{\mu}_T| &\le \sum_{T \in \mathcal{T}_h} \boldsymbol{\mu}_T \cdot (\nabla y_{h|T} - \nabla \tilde{y}_{h|T}) \\
&= a(y_h - \tilde{y}_h, p_h) - \int_\Omega (y_h - y_0)(y_h - \tilde{y}_h) \\
&= \int_\Omega (u_h - Q_h v) p_h - \int_\Omega y_h^2 + \int_\Omega y_h (y_0 + \tilde{y}_h) - \int_\Omega y_0 \tilde{y}_h \\
&\le -\alpha \int_\Omega |u_h|^{r-2} u_h (u_h - Q_h v) - \frac{1}{2} \int_\Omega y_h^2 + \frac{1}{2} \int_\Omega y_0^2 + \frac{1}{2} \int_\Omega \tilde{y}_h^2 \\
&\le -\alpha \int_\Omega |u_h|^r + \alpha \int_\Omega |u_h|^{r-2} u_h Q_h v + C \int_\Omega (y_0^2 + \tilde{y}_h^2) \\
&\le \alpha \|u_h^{r-1}\|_{L^{\frac{r}{r-1}}(\Omega)} \|Q_h v\|_{L^r(\Omega)} + C \int_\Omega (y_0^2 + \tilde{y}_h^2) \\
&= \alpha \|u_h\|_{L^r(\Omega)}^{r-1} \|Q_h v\|_{L^r(\Omega)} + C \int_\Omega (y_0^2 + \tilde{y}_h^2) \\
&\le C(\|Q_h v\|_{L^r(\Omega)} + \|y_0\|_{L^2(\Omega)}^2 + \|\tilde{y}_h\|_{L^2(\Omega)}^2),
\end{aligned}$$

where we again have used $y_h(y_0 + \tilde{y}_h) \le \frac{1}{2} y_h^2 + \frac{1}{2}(y_0 + \tilde{y}_h)^2$. This implies the bound on $\boldsymbol{\mu}_T$. □

Theorem 4.2.10. *Let u and u_h be the solutions of (4.20) and (4.38) respectively. Then there exists $h_1 \le h_0$ such that*

$$\|y - y_h\|_{L^2(\Omega)} \le C h^{\frac{1}{2}(1 - \frac{d}{r})}, \text{ and } \|u - u_h\|_{L^r(\Omega)} \le C h^{\frac{1}{r}(1 - \frac{d}{r})}$$

for all $0 < h \le h_1$.

PROOF. Let us introduce $y^h := \mathcal{G}(u_h) \in W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$, and $\hat{y}_h := \mathcal{G}_h(Q_h u)$. In view (4.25) we have

$$\|y^h - y_h\|_{W^{1,\infty}(\Omega)} \leq Ch^{1-\frac{d}{r}} \|u_h\|_{L^r(\Omega)} \leq Ch^{1-\frac{d}{r}}.$$

Let us now turn to the actual error estimate. Using (4.23b) and (4.42) we have

$$\begin{aligned}\alpha\theta_r \int_\Omega |u - u_h|^r &\leq \alpha \int_\Omega \big(|u|^{r-2}u - |u_h|^{r-2}u_h\big)(u - u_h) \\ &= \underbrace{\int_\Omega p(u_h - u)}_{=:I} + \underbrace{\int_\Omega p_h(Q_h u - u_h)}_{=:II} - \alpha \underbrace{\int_\Omega \underbrace{|u_h|^{r-2}u_h}_{\in U_h} \underbrace{(u - Q_h u)}_{\in U_h^\perp}}_{=0}.\end{aligned}$$

To estimate the terms I and II we follow the lines of the proof of Theorem 4.2.7 and obtain

$$I \leq \int_\Omega (y - y_0)(y^h - y) + Ch^{1-\frac{d}{r}}, \tag{4.43}$$

as well as

$$\begin{aligned}II &\leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + 2 \max_{T \in \mathcal{T}_h} |\nabla \hat{y}_{h|T} - \nabla y(x_T)| \sum_{T \in \mathcal{T}_h} |\boldsymbol{\mu}_T| \\ &\leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + C\|\nabla(\hat{y}_h - y)\|_{L^\infty(\Omega)^d}.\end{aligned}$$

As in inequality (4.41) with $v := u$ we estimate

$$\|\nabla(\hat{y}_h - y)\|_{L^\infty(\Omega)^d} = \|\nabla(\mathcal{G}_h(Q_h u) - \mathcal{G}(u))\|_{L^\infty(\Omega)^d} \leq Ch^{1-\frac{d}{r}}$$

and thus

$$II \leq \int_\Omega (y_h - y_0)(\hat{y}_h - y_h) + Ch^{1-\frac{d}{r}}.$$

Combining I and II we finally obtain

$$\alpha\theta_r \int_\Omega |u - u_h|^r + \int_\Omega |y - y_h|^2 \leq Ch\big(\|u\|_{L^2(\Omega)} + \|u_h\|_{L^2(\Omega)}\big) + Ch^{1-\frac{d}{r}}$$

and the result follows. □

4.2.4. Numerical experiments. We now consider the finite element approximation of problem (4.20) with the data from Example 4.1.11. The optimization problem then has the unique solution $u(x) = -\mathbb{1}_{B_1(0)}(x)$. We note that we obtain equality in (4.23b), i.e. $p = -u$ for all $r > d$. For all numerical computations we take $r = 4$.

4.2.4.1. *Variational discretization.* We solve problem (4.26), where we essentially make use of the structure of u_h in terms of equation (4.29). With our choices for the specific example the optimal control $u_h \in L^4(\Omega)$ with corresponding adjoint state $p_h \in P^1_{c,h}(\mathcal{T}_h)$ satisfies

$$u_h = -\alpha^{-\frac{1}{3}} |p_h|^{-\frac{2}{3}} p_h = -\alpha^{-\frac{1}{3}} \operatorname{sign}(p_h) |p_h|^{\frac{1}{3}}. \tag{4.44}$$

Now the idea is to minimize over the subset

$$U_h := \{u \in L^4(\Omega) \,:\, u \text{ satisfies (4.44) with } u_h = u \text{ for some } p_h \in P^1_{c,h}(\mathcal{T}_h)\}.$$

In order to use the Matlab optimization routine `fmincon` one has to provide the objective $J_h(u)$ for those elements $u \in U_h$, especially the term

$$\frac{\alpha}{4}\int_\Omega |u|^4 = \frac{\alpha^{\frac{2}{3}}}{4}\int_\Omega |p_h|^{\frac{4}{3}} = \frac{\alpha^{\frac{2}{3}}}{4}\sum_{T\in\mathcal{T}_h}\int_T |p_{h|T}|^{\frac{4}{3}}. \tag{4.45}$$

We transform the involved integrals over triangles onto the standard 2-simplex $\hat{T}$ in $\mathbb{R}^2$. For a triangle $T = \operatorname{conv hull}\{\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2\}$ in $\mathbb{R}^2$ there exists by (1.13) a diffeomorphism $\vec{F}_T : \hat{T} \to T$ with

$$\vec{F}_T(\hat{x}) = \mathbf{A}_T\hat{x} + \mathbf{a}_0,$$

where $\mathbf{A}_T \in \mathbb{R}^{2\times 2}$ is a regular matrix. For any integrable $f : T \to \mathbb{R}$ the transformation-theorem gives

$$\begin{aligned}\int_T f(x)\,dx = \int_{\vec{F}_T(\hat{T})} f(x)\,dx &= \int_{\hat{T}} f(\vec{F}_T(\hat{x}))|\det(D\vec{F}_T(\hat{x}))|\,d\hat{x}\\ &= |\det(\mathbf{A}_T)|\int_{\hat{T}} f(\vec{F}_T(\hat{x}))\,d\hat{x}\\ &= 2|T|\int_0^1\int_0^{1-\hat{x}_1} f(\vec{F}_T(\hat{x}_1,\hat{x}_2))\,d\hat{x}_2\,d\hat{x}_1.\end{aligned} \tag{4.46}$$

Let $\varphi_i : T \to \mathbb{R}$ for $i = 1, 2, 3$ be the local linear finite element basis functions corresponding to the vertex $\mathbf{a}_{i-1}$ on the element T. Their associated linear finite element basis functions on the standard simplex $\hat{T}$ are given by

$$\hat{\varphi}_1(\hat{x}_1,\hat{x}_2) = 1 - \hat{x}_1 - \hat{x}_2, \quad \hat{\varphi}_2(\hat{x}_1,\hat{x}_2) = \hat{x}_1, \quad \hat{\varphi}_3(\hat{x}_1,\hat{x}_2) = \hat{x}_2.$$

By setting $p_i := p_h(\mathbf{a}_{i-1}) \in \mathbb{R}$ for $i \in \{1, 2, 3\}$ we consider the local affine linear function $p : T \to \mathbb{R}$

$$p := p_{h|T} = p_1\varphi_1 + p_2\varphi_2 + p_3\varphi_3.$$

Obviously for $(\hat{x}_1, \hat{x}_2) \in \hat{T}$ one obtains

$$p(\vec{F}_T(\hat{x}_1,\hat{x}_2)) = p_1\hat{\varphi}_1(\hat{x}_1,\hat{x}_2) + p_2\hat{\varphi}_2(\hat{x}_1,\hat{x}_2) + p_3\hat{\varphi}_3(\hat{x}_1,\hat{x}_2) =: \hat{p}(\hat{x}_1,\hat{x}_2).$$

Finally for evaluating (4.45) we set $f := |p|^{\frac{4}{3}}$ in (4.46) and compute

$$\frac{1}{2|T|}\int_T f(x)\,dx = \int_{\hat{T}} f(\vec{F}_T(\hat{x}))\,d\hat{x} = \int_{\hat{T}} |\hat{p}(\hat{x})|^{\frac{4}{3}}\,d\hat{x}.$$

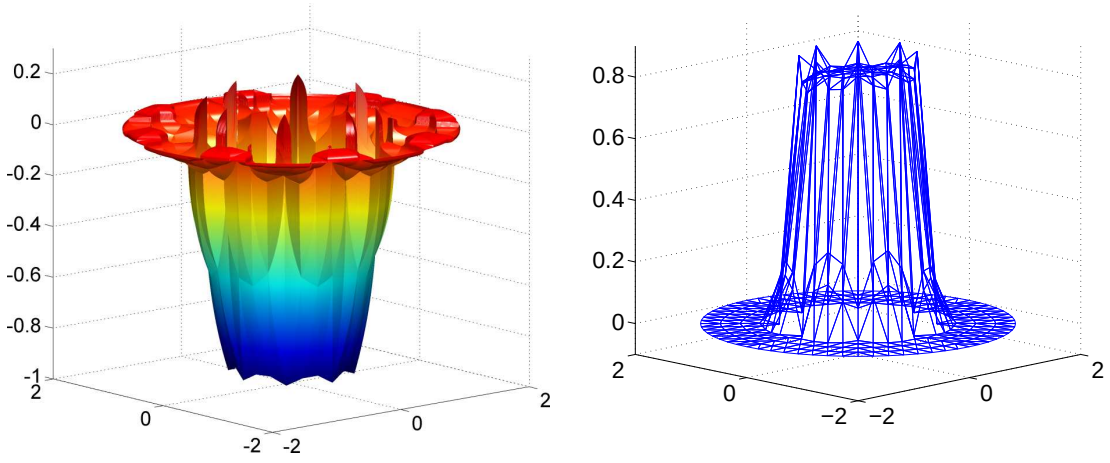
This integral can be computed analytically and is part of Lemma C.1 in the appendix. Moreover for evaluating the right-hand-side of the state equation one needs to assemble the vector

$$\hat{\mathbf{u}} = \left[\int_\Omega u\phi_j\right]_{j=1}^m = -\alpha^{-\frac{1}{3}}\left[\sum_{T\in\mathcal{T}_h}\int_T \operatorname{sign}(p_{h|T})|p_{h|T}|^{\frac{1}{3}}\phi_j\right]_{j=1}^m.$$

Therefore we need to know the integrals $\int_T \operatorname{sign}(p_{h|T})|p_{h|T}|^{\frac{1}{3}}\varphi_i$ for $i \in \{1, 2, 3\}$ involving the local basis functions φ_i on T. Again from the previous setting let $p := p_{h|T}$ and $f_i := \operatorname{sign}(p)|p|^{\frac{1}{3}}\varphi_i$ so that we need to compute

$$\frac{1}{2|T|}\int_T f_i(x)\,dx = \int_{\hat{T}} f_i(\vec{F}_T(\hat{x}))\,d\hat{x} = \int_{\hat{T}} \operatorname{sign}(\hat{p}(\hat{x}))|\hat{p}(\hat{x})|^{\frac{1}{3}}\hat{\varphi}_i(\hat{x})\,d\hat{x} =: I_i.$$

It turns out that also these integrals can be computed analytically. This is carried out in Lemma C.2 in the appendix. To speed up the numerical performance of a generalized Newton method for the control-reduced problem it is also required to have derivative information at hand. The difficult part is to derive the right hand



(a) Control u_h. (b) Adjoint state p_h.

FIGURE 4.5. Variational discretization.

nt	$\|u-u_h\|_{L^4(\Omega)}$	EOC	$\|u-u_h\|_{L^2(\Omega)}$	EOC	$\|y-y_h\|_{L^2(\Omega)}$	EOC
32	0.834633	-	1.360030	-	0.220346	-
128	0.588566	0.549	0.904770	0.640	0.797200	1.597
512	0.484191	0.293	0.582014	0.661	0.035210	1.225

TABLE 4.2. Errors and EOCs for variational discretization.

side of the state equation with respect to p_h. More precisely we locally end up to compute the numbers

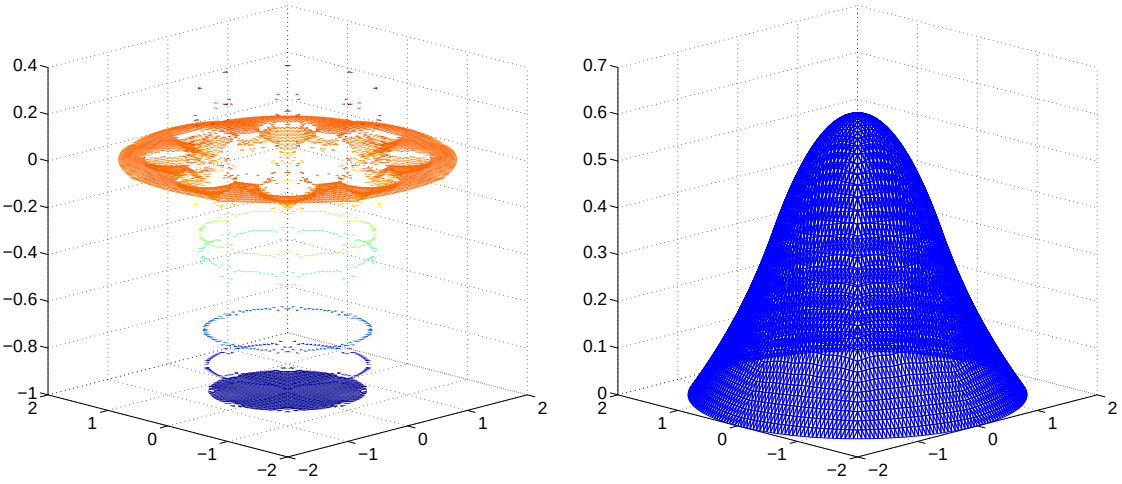
$$G_{ij} := \frac{\partial I_i}{\partial p_j},$$

which are explicitly stated in Lemma C.3.

Figure 4.5 illustrates the optimal solution u_h and corresponding adjoint state p_h on a mesh consisting of $nt = 512$ triangles. We note that due to relation (4.29) the variational control has to be a continuous function. The exact control however has a jump. We conclude that variational discretization combined with piecewise linear and continuous finite elements for the state approximation is not ideally suited to approximate control problems with gradient constraints on the state. To illustrate this fact we in Table 4.2 present some numerical computations for up to $nt = 512$ elements.

Led by the findings of [DGH09c] we think that variational discretization combined with the lowest order Raviart–Thomas finite element as state approximations in a mixed formulation of the state equation seems to be a more appropriate choice. However, many existing finite element codes use standard finite elements, so that there exists a demand in these approximation approaches also in optimization of elliptic PDEs in the presence of gradient constraints on the state. Therefore, in the present work we also investigate piecewise constant control approximations combined with piecewise linear, continuous approximations of the state.

4.2.4.2. *Piecewise constant controls.* We use piecewise constant, discontinuous Ansatz functions for the control u_h. For the numerical solution we use the routine `fmincon` contained in the Matlab optimization toolbox. The state equation



(a) Control u_h. (b) State y_h.

FIGURE 4.6. Piecewise constant controls.

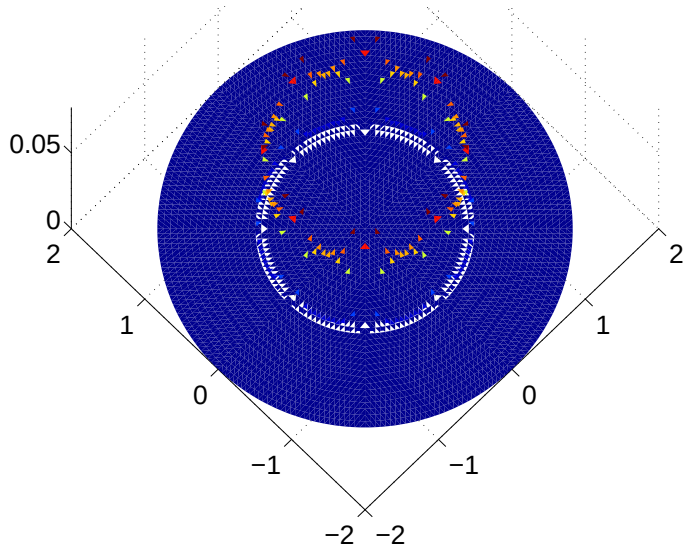


FIGURE 4.7. Discrete multiplier for piecewise constant controls.

is approximated with piecewise linear, continuous finite elements on quasi-uniform triangulations $\mathcal{T}_h$ of $B_2(0)$. The gradient constraints are required element-wise. The resulting discretized optimization problem then reads

$$\min_{u_h \in U_h} J_h(u_h) = \frac{1}{2}\|y_h - y_0\|^2_{L^2(\Omega)} + \frac{\alpha}{r}\|u_h\|^r_{L^r(\Omega)}$$
$$\text{subject to } \; y_h = \mathcal{G}_h(u_h) \text{ and } |\nabla y_{h|T}| \leq \delta = \frac{1}{2} \quad \forall\, T \in \mathcal{T}_h.$$

In Figures 4.6, 4.7 we present the numerical approximations u_h, y_h, and μ_h on a grid containing $nt = 8192$ triangles, where μ_h is obtained by $\vec{\mu}_h$ according to relation (4.31). Figure 4.7 clearly shows that the support of μ_h is concentrated at $|x| = 1$.

In Table 4.3 we document the experimental order of convergence. The controls show an approximation behavior which is slightly better than that predicted by Theorem 4.2.10. However, this may be caused by the fact that $\|u\|_{L^\infty(\Omega)}, \|u_h\|_{L^\infty(\Omega)} \leq C$

nt	$\|u-u_h\|_{L^4(\Omega)}$	EOC	$\|u-u_h\|_{L^2(\Omega)}$	EOC	$\|y-y_h\|_{L^2(\Omega)}$	EOC
32	0.834550	-	1.376190	-	0.230207	-
128	0.541825	0.679	0.845567	0.765	0.081135	1.639
512	0.457207	0.255	0.603292	0.506	0.032682	1.363
2048	0.363216	0.338	0.411190	0.563	0.013326	1.318
8192	0.295328	0.301	0.274811	0.587	0.005277	1.348

TABLE 4.3. Errors and EOCs for piecewise constant controls.

uniformly in h. The L^2-norm of the state seems to converge at least with linear order. This can be explained by the high regularity of the exact solution. In the second column of Table 4.5 we display the values of $\sum_{T\in\mathcal{T}_h}|\boldsymbol{\mu}_T|$. These values are expected to converge to 2π as $h\to 0$, since this gives the value of μ applied to the function which is identically equal to 1 on $\bar{\Omega}$.

In order to motivate the convergence behavior of $\|u-u_h\|_{L^2(\Omega)}$ we briefly consider

4.2.4.3. TYCHONOV *regularization.* Since $u\in L^r(\Omega)$ with $r>d\geq 2$ we may also penalize with the L^2-norm of the control. The corresponding optimal control problem reads

$$\begin{aligned}&\min_{u_h\in U_h} J_h(u_h)=\frac{1}{2}\|y_h-y_0\|^2_{L^2(\Omega)}+\frac{\alpha}{2}\|u_h\|^2_{L^2(\Omega)}+\frac{\alpha}{r}\|u_h\|^r_{L^r(\Omega)}\\ &\text{subject to }\ y_h=\mathcal{G}_h(u_h)\text{ and }|\nabla y_{h|T}|\leq\delta=\frac{1}{2}\quad\forall\,T\in\mathcal{T}_h.\end{aligned}$$

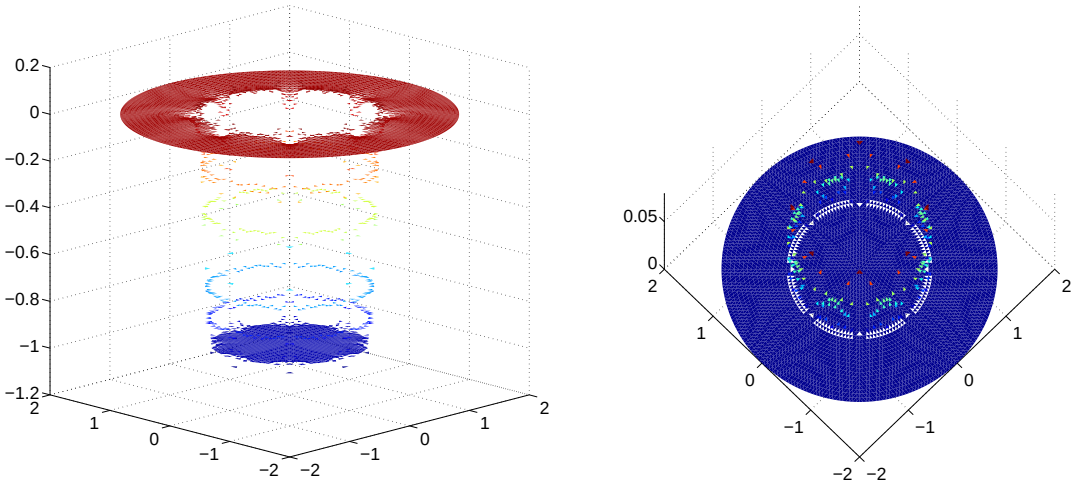
An analytic solution can be obtained by adapting the constants in our example. Since the variational equality for the control for this control problem reads

$$\int_\Omega (p_h+\alpha(u_h+|u_h|^{r-2}u_h))v_h=0\text{ for all }v_h\in U_h$$

we have a solution for the same data as before except for $\alpha=0.5$. An analysis along the lines of Theorems 4.2.7, 4.2.10 now shows that we also get

$$\|u-u_h\|_{L^2(\Omega)}\leq Ch^{\frac{1}{2}(1-\frac{d}{r})},$$

with $C=C(\|u\|_{L^r(\Omega)},\|u_h\|_{L^r(\Omega)})$. Since in the present example we have $u\in L^\infty(\Omega)$ and that $\|u_h\|_{L^\infty(\Omega)}$ is uniformly bounded in h we expect the error behavior $\|u-u_h\|_{L^2(\Omega)}\sim\mathcal{O}(h^{\frac{1}{2}-\varepsilon})$ for $h\to 0$. In Figure 4.8 we present the numerical approximations u_h and μ_h on a grid containing $nt=8192$ triangles. In Table 4.4 we investigate the experimental order of convergence for different error functionals. All convergence orders are in the same range as those obtained in the case without TYCHONOV regularization and piecewise constant controls. We observe that the control does not oscillate that much along $\partial B_1(0)$ as in the unregularized case.



(a) Control u_h. (b) Multiplier $(\boldsymbol{\mu}_T)_{T\in\mathcal{T}_h}$.

FIGURE 4.8. TYCHONOV regularization.

nt	$\|u-u_h\|_{L^4(\Omega)}$	EOC	$\|u-u_h\|_{L^2(\Omega)}$	EOC	$\|y-y_h\|_{L^2(\Omega)}$	EOC
32	0.863533	-	1.224542	-	0.383556	-
128	0.530078	0.767	0.772724	0.723	0.114305	1.902
512	0.425213	0.330	0.503372	0.642	0.049405	1.257
2048	0.352524	0.275	0.348416	0.541	0.021354	1.232
8192	0.289696	0.286	0.241345	0.534	0.009586	1.166

TABLE 4.4. Errors and EOCs for TYCHONOV regularization.

nt	$\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\mu}_T\|$	$\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\mu}_T\|$
32	0	0.923
128	2.498	3.657
512	4.217	4.958
2048	5.213	5.603
8192	5.740	5.940

TABLE 4.5. Multiplier approximations for piecewise constant controls (middle) and for TYCHONOV regularization (right).

CHAPTER 5

Summary and conclusions

Let us summarize our investigations in structure exploiting GALERKIN schemes for the problems of consideration within the last chapters. Throughout this manuscript we permanently apply the variational discretization concept proposed by Hinze in [Hin05]. Not explicitly discretizing the control allows for crosswisely testing variational inequalities from the analytic and variational discretized optimal control problems with their optimal solutions. This technique is a generally applicable and elegant starting point to derive a priori error estimates under the further use of finite element error estimates for the state equation.

In Chapter 2 we derive new results for optimal DIRICHLET boundary control problems with control constraints on smooth bounded domains in two and three space dimensions. The proven orders of convergence, which also are numerically confirmed, can even be increased in two space dimensions by additional assumptions onto the underlying mesh of computation. Another potency of variational discretization lies in the natural behaviour of control active sets as their boundary not necessarily needs to be resolved by edges from the computational grid. This of course leads to increased costs of implementation. However its practicability is figured out in the second part of Chapter 2. Moreover we introduce the useful notation of additive mass-matrix splitting and highlight that variational discretization keeps the sparsity structure in Jacobian matrices.

In Chapter 3 we focus onto elliptic optimal control problems, when pointwise constraints onto the state come into play. Since a priori error analysis even for variational discretization is already carried out in the cited literature, we concentrate onto a posteriori structure exploiting GALERKIN concepts. With the design of a goal-oriented error estimator we extend the dual weighted residual approach proposed in [BR96] by Becker and Rannacher to the state constrained case. Our approach avoids the appearance of additional control error terms in error representations. We even extend some of these techniques to an optimal control problem with additional control constraints. This setting numerically requires to introduce a regularization of the state constraints. By choosing a Moreau-Yosida penalization we construct computable a posteriori error estimators. Numerous numerical examples highlight the performance and effectiveness of the adaptive solution loop.

The concept of variational discretization is not restricted to conforming finite elements. In Chapter 4 we approximate the elliptic state equation by lowest order Raviart-Thomas elements in a mixed formulation. In this way of discretization the gradient of the state is represented by an own quantity and better reflects the gradient constraints, which are of main interest within this chapter. A priori error estimates for the control and the state are proven for two replenishing scenarios which distinguish in the control costs in the objective and the presence of additional control constraints. The experimental order of convergence measurements show better behaviour of the state variable then predicted. Moreover variational discretized finite elements for the second scenario with L^r-norm of the control in the objective

are numerically implemented. The arising formulas show, that easy numerical manageability of variational discretization relies on quadratic objectives, which is often the case in practice.

Finally structure exploiting GALERKIN methods are a powerful tool on both sides: On the a priori part to elegantly investigate and prove finite element error estimates and on the a posteriori part to save degrees of freedom due to problem adapted meshes and model reduction. But these methods are just one possible approach. Todays and future applications in terms of highly nonlinear, time dependent optimal control problems enforce us to efficiently combine these techniques with parallel computing, automatic differentiation or multigrid methods for instance.

APPENDIX A

Control constraints

The additive mass-matrix-splitting routine assem_mass

Algorithm A.1.

```
% split mass matrix into 3 parts w.r.t. CHI= -1/al*P + U0

function [M_tilde_IS, M_tilde_Aa, M_tilde_Ab]= assem_mass(CHI, ua, ub)
global Mesh
persistent int_T_ij i112 j233 Mat_np_nt_100 Mat_np_nt_010 Mat_np_nt_001

if size(int_T_ij, 2)~=Mesh.nt
  % assemble local mass matrix
  % \int_T \phi_i \phi_j                ij
  int_T_ij= [Mesh.area()./6           % 11
             Mesh.area()./12          % 12
             Mesh.area()./12          % 13
             Mesh.area()./6           % 22
             Mesh.area()./12          % 23
             Mesh.area()./6];         % 33

  i112= [Mesh.pn100(), Mesh.pn100(), Mesh.pn010()];
  j233= [Mesh.pn010(), Mesh.pn001(), Mesh.pn001()];
  Mat_np_nt_100= sparse(Mesh.pn100(), 1:Mesh.nt, true, Mesh.np, Mesh.nt);
  Mat_np_nt_010= sparse(Mesh.pn010(), 1:Mesh.nt, true, Mesh.np, Mesh.nt);
  Mat_np_nt_001= sparse(Mesh.pn001(), 1:Mesh.nt, true, Mesh.np, Mesh.nt);
end

mij_Aa= zeros(6, Mesh.nt);
mij_IS= zeros(6, Mesh.nt);
mij_Ab= zeros(6, Mesh.nt);

int_T1_ij= sparse(6, Mesh.nt);
int_T2_ij= sparse(6, Mesh.nt);
int_T3_ij= sparse(6, Mesh.nt);
int_T4_ij= sparse(6, Mesh.nt);
int_T5_ij= sparse(6, Mesh.nt);

CHI123_m101= 1.0.*(CHI(Mesh.t(1:3, :))>ub) - 1.0.*(CHI(Mesh.t(1:3, :))<ua);

%-------------------------------------------------------------------------
% 1st case: all active or all inactive
all_act_Aa= sum(     CHI123_m101) ==-3;
all_ina_IS= sum(abs(CHI123_m101))== 0;
all_act_Ab= sum(     CHI123_m101) == 3;

mij_Aa(:, all_act_Aa)= int_T_ij(:, all_act_Aa);
mij_IS(:, all_ina_IS)= int_T_ij(:, all_ina_IS);
mij_Ab(:, all_act_Ab)= int_T_ij(:, all_act_Ab);
%-------------------------------------------------------------------------
% 2nd case: exactly ONE active or          exactly TWO   active
%                                    (hence exactly ONE inactive)
ONE_act_Aa= (sum(abs(CHI123_m101))==1) & (sum(CHI123_m101)==-1);
```

```
ONE_act_Ab= (sum(abs(CHI123_m101))==1) & (sum(CHI123_m101)== 1);
TWO_act_Aa= (sum(abs(CHI123_m101))==2) & (sum(CHI123_m101)==-2);
TWO_act_Ab= (sum(abs(CHI123_m101))==2) & (sum(CHI123_m101)== 2);
ONE= ONE_act_Aa | ONE_act_Ab | TWO_act_Aa | TWO_act_Ab;

if any(ONE) % not empty 2nd case
  % ONE       triangle number with 2nd case      (1..nt)
  % ONE_loc   local ONE number on this triangle  (1,2,3)
  % uaub      ua active / ub active              (ua,ub)

  [j, k]= find(CHI123_m101.*repmat(ONE_act_Aa, 3, 1));
  ONE=                k';
  ONE_loc=            j';
  uaub=               ua*ones(1, length(j));

  [j, k]= find(CHI123_m101.*repmat(ONE_act_Ab, 3, 1));
  ONE=      [ONE      k'];
  ONE_loc=  [ONE_loc  j'];
  uaub=     [uaub     ub*ones(1, length(j))];

  [j, k]= find((CHI123_m101 + 1).*repmat(TWO_act_Aa, 3, 1));
  ONE=      [ONE      k'];
  ONE_loc=  [ONE_loc  j'];
  uaub=     [uaub     ua*ones(1, length(j))];

  [j, k]= find((CHI123_m101 - 1).*repmat(TWO_act_Ab, 3, 1));
  ONE=      [ONE      k'];
  ONE_loc=  [ONE_loc  j'];
  uaub=     [uaub     ub*ones(1, length(j))];

  n1= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE))); %
      successor of local ONE number
  n2= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE))); %
      succesuccessor of local ONE number
  n3= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc         , ONE))); %
                          local ONE number

  CHI_n1=   CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE)))';
  CHI_n2=   CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE)))';
  CHI_n3=   CHI(Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc         , ONE)))';

  n4= n1 + repmat((uaub - CHI_n1)./(CHI_n3 - CHI_n1), 2, 1).*(n3 - n1); %
      intersection of CHI between n1 and n3 with ua / ub
  n5= n2 + repmat((uaub - CHI_n2)./(CHI_n3 - CHI_n2), 2, 1).*(n3 - n2); %
      intersection of CHI between n2 and n3 with ua / ub

  T1= Ti_area(n1, n2, n5);
  T2= Ti_area(n1, n5, n4);
  T3= Ti_area(n4, n5, n3);

  pi= zeros(7, 2, length(ONE));
  pi(1, :, :)= 1/2*(n1 + n2);
  pi(2, :, :)= 1/2*(n2 + n5);
  pi(3, :, :)= 1/2*(n1 + n5);
  pi(4, :, :)= 1/2*(n1 + n4);
  pi(5, :, :)= 1/2*(n4 + n5);
  pi(6, :, :)= 1/2*(n3 + n4);
  pi(7, :, :)= 1/2*(n3 + n5);

  % pi in global barycentric coordinates pi1 pi2 pi3 (i= 1...7)
  p= zeros(7, 3, length(ONE));
  for j= 1:length(ONE)
    p(:, :, j)= ([Mesh.p(1:2, Mesh.t(1:3, ONE(j))')';
```

```
                ones(1, 3)]\[pi(1:7, 1:2, j)';
                                ones(1, 7)])';
  end

  % 11; 12; 13; 22; 23; 33
  int_T1_ij(:, ONE)= int_Tk_ij(T1, p, [1 2 3]);
  int_T2_ij(:, ONE)= int_Tk_ij(T2, p, [3 5 4]);
  int_T3_ij(:, ONE)= int_Tk_ij(T3, p, [5 7 6]);

  mij_Aa(:, ONE_act_Aa)= int_T3_ij(:, ONE_act_Aa);
  mij_IS(:, ONE_act_Aa)= int_T1_ij(:, ONE_act_Aa) + int_T2_ij(:, ONE_act_Aa);

  mij_IS(:, ONE_act_Ab)= int_T1_ij(:, ONE_act_Ab) + int_T2_ij(:, ONE_act_Ab);
  mij_Ab(:, ONE_act_Ab)= int_T3_ij(:, ONE_act_Ab);

  mij_Aa(:, TWO_act_Aa)= int_T1_ij(:, TWO_act_Aa) + int_T2_ij(:, TWO_act_Aa);
  mij_IS(:, TWO_act_Aa)= int_T3_ij(:, TWO_act_Aa);

  mij_IS(:, TWO_act_Ab)= int_T3_ij(:, TWO_act_Ab);
  mij_Ab(:, TWO_act_Ab)= int_T1_ij(:, TWO_act_Ab) + int_T2_ij(:, TWO_act_Ab);
end
% ------------------------------------------------------------------------

% 3rd case: three differently active (two active and ONE active)
ONE_act_Aa= (sum(abs(CHI123_m101))==3) & (sum(CHI123_m101)== 1);
ONE_act_Ab= (sum(abs(CHI123_m101))==3) & (sum(CHI123_m101)==-1);
ONE= ONE_act_Aa | ONE_act_Ab;
if any(ONE) % not empty 3rd case
  [j, k]= find((CHI123_m101==-1).*repmat(ONE_act_Aa, 3, 1));
  ONE=                k';
  ONE_loc=            j';
  uaub=               ua*ones(1, length(j));

  [j, k]= find((CHI123_m101==1).*repmat(ONE_act_Ab, 3, 1));
  ONE=     [ONE       k'];
  ONE_loc= [ONE_loc   j'];
  uaub=    [uaub      ub*ones(1, length(j))];

  n1= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE))); %
      successor of local ONE number
  n2= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE))); %
      succesuccessor of local ONE number
  n3= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc         , ONE))); %
                          local ONE number

  CHI_n1=   CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE)))';
  CHI_n2=   CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE)))';
  CHI_n3=   CHI(Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc         , ONE)))';

  n4= n1 + repmat((uaub  - CHI_n1)./(CHI_n3 - CHI_n1), 2, 1).*(n3 - n1); %
      intersection of CHI between n1 and n3 with ua / ub
  n5= n2 + repmat((uaub  - CHI_n2)./(CHI_n3 - CHI_n2), 2, 1).*(n3 - n2); %
      intersection of CHI between n2 and n3 with ua / ub

  uaub2= ua + ub - uaub; % opposite
  n6= n1 + repmat((uaub2 - CHI_n1)./(CHI_n3 - CHI_n1), 2, 1).*(n3 - n1); %
      intersection of CHI between n1 and n3 with ua / ub
  n7= n2 + repmat((uaub2 - CHI_n2)./(CHI_n3 - CHI_n2), 2, 1).*(n3 - n2); %
      intersection of CHI between n2 and n3 with ua / ub

  T1= Ti_area(n1, n2, n7);
  T2= Ti_area(n1, n7, n6);
  T3= Ti_area(n4, n5, n3);
```

```
  T4= Ti_area(n6, n7, n5);
  T5= Ti_area(n4, n6, n5);

  pi= zeros(11, 2, length(ONE));
  pi( 1, :, :)= 1/2*(n1 + n2);
  pi( 2, :, :)= 1/2*(n2 + n7);
  pi( 3, :, :)= 1/2*(n1 + n7);
  pi( 4, :, :)= 1/2*(n1 + n6);
  pi( 5, :, :)= 1/2*(n4 + n5);
  pi( 6, :, :)= 1/2*(n3 + n4);
  pi( 7, :, :)= 1/2*(n3 + n5);
  pi( 8, :, :)= 1/2*(n6 + n7);
  pi( 9, :, :)= 1/2*(n5 + n7);
  pi(10, :, :)= 1/2*(n5 + n6);
  pi(11, :, :)= 1/2*(n4 + n6);

  % pi in global barycentric coordinates pi1 pi2 pi3 (i= 1...11)
  p= zeros(11, 3, length(ONE));
  for j= 1:length(ONE)
    p(:, :, j)= ([Mesh.p(1:2, Mesh.t(1:3, ONE(j))');
                  ones(1, 3)]\[pi(1:11, 1:2, j)';
                               ones(1, 11)])';
  end

  % 11; 12; 13; 22; 23; 33
  int_T1_ij(:, ONE)= int_Tk_ij(T1, p, [1 2 3]);
  int_T2_ij(:, ONE)= int_Tk_ij(T2, p, [3 8 4]);
  int_T3_ij(:, ONE)= int_Tk_ij(T3, p, [5 7 6]);
  int_T4_ij(:, ONE)= int_Tk_ij(T4, p, [8 9 10]);
  int_T5_ij(:, ONE)= int_Tk_ij(T5, p, [5 11 10]);

  mij_Aa(:, ONE_act_Aa)= int_T3_ij(:, ONE_act_Aa);
  mij_IS(:, ONE_act_Aa)= int_T4_ij(:, ONE_act_Aa) + int_T5_ij(:, ONE_act_Aa);
  mij_Ab(:, ONE_act_Aa)= int_T1_ij(:, ONE_act_Aa) + int_T2_ij(:, ONE_act_Aa);

  mij_Aa(:, ONE_act_Ab)= int_T1_ij(:, ONE_act_Ab) + int_T2_ij(:, ONE_act_Ab);
  mij_IS(:, ONE_act_Ab)= int_T4_ij(:, ONE_act_Ab) + int_T5_ij(:, ONE_act_Ab);
  mij_Ab(:, ONE_act_Ab)= int_T3_ij(:, ONE_act_Ab);
end
% ------------------------------------------------------------------------

% 4th case: two differently active and ONE inactive
ONE= (sum(abs(CHI123_m101))==2) & (sum(CHI123_m101)== 0);
if any(ONE) % not empty 4th case
  [j, k]= find((CHI123_m101==0).*repmat(ONE, 3, 1));
  ONE=                k';
  ONE_loc=            j';
  uaub= (ub-ua)/2*(CHI123_m101(sub2ind([3, Mesh.nt], mod(ONE_loc, 3)+1, ONE)) + 1)
      + ua;

  ONE_T1_act_Aa= ONE(CHI123_m101(sub2ind([3, Mesh.nt], mod(ONE_loc, 3)+1,
      ONE))==-1);
  ONE_T1_act_Ab= ONE(CHI123_m101(sub2ind([3, Mesh.nt], mod(ONE_loc, 3)+1, ONE))==
      1);

  n1= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE))); %
      successor of local ONE number
  n2= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE))); %
      succesuccessor of local ONE number
  n3= Mesh.p(:, Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc         , ONE))); %
                        local ONE number

  CHI_n1=    CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc   , 3)+1, ONE)))';
```

```
  CHI_n2=   CHI(Mesh.t(sub2ind([4, Mesh.nt], mod(ONE_loc+1, 3)+1, ONE)))';
  CHI_n3=   CHI(Mesh.t(sub2ind([4, Mesh.nt],      ONE_loc      , ONE)))';

  n4= n1 + repmat((uaub  - CHI_n1)./(CHI_n2 - CHI_n1), 2, 1).*(n2 - n1); %
      intersection of CHI between n1 and n2 with ua / ub
  n5= n1 + repmat((uaub  - CHI_n1)./(CHI_n3 - CHI_n1), 2, 1).*(n3 - n1); %
      intersection of CHI between n1 and n3 with ua / ub

  uaub2= ua + ub - uaub; % opposite
  n6= n1 + repmat((uaub2 - CHI_n1)./(CHI_n2 - CHI_n1), 2, 1).*(n2 - n1); %
      intersection of CHI between n1 and n2 with ua / ub
  n7= n2 + repmat((uaub2 - CHI_n2)./(CHI_n3 - CHI_n2), 2, 1).*(n3 - n2); %
      intersection of CHI between n2 and n3 with ua / ub

  T1= Ti_area(n1, n4, n5);
  T2= Ti_area(n3, n5, n4);
  T3= Ti_area(n3, n4, n6);
  T4= Ti_area(n3, n6, n7);
  T5= Ti_area(n2, n7, n6);

  pi= zeros(11, 2, length(ONE));
  pi( 1, :, :)= 1/2*(n1 + n4);
  pi( 2, :, :)= 1/2*(n4 + n5);
  pi( 3, :, :)= 1/2*(n1 + n5);
  pi( 4, :, :)= 1/2*(n3 + n4);
  pi( 5, :, :)= 1/2*(n3 + n5);
  pi( 6, :, :)= 1/2*(n4 + n6);
  pi( 7, :, :)= 1/2*(n3 + n6);
  pi( 8, :, :)= 1/2*(n6 + n7);
  pi( 9, :, :)= 1/2*(n3 + n7);
  pi(10, :, :)= 1/2*(n2 + n6);
  pi(11, :, :)= 1/2*(n2 + n7);

  % pi in global barycentric coordinates pi1 pi2 pi3 (i= 1...11)
  p= zeros(11, 3, length(ONE));
  for j= 1:length(ONE)
    p(:, :, j)= ([Mesh.p(1:2, Mesh.t(1:3, ONE(j)))';
                  ones(1, 3)]\[pi(1:11, 1:2, j)';
                               ones(1, 11)])';
  end

  % 11; 12; 13; 22; 23; 33
  int_T1_ij(:, ONE)= int_Tk_ij(T1, p, [1 2 3]);
  int_T2_ij(:, ONE)= int_Tk_ij(T2, p, [2 4 5]);
  int_T3_ij(:, ONE)= int_Tk_ij(T3, p, [4 6 7]);
  int_T4_ij(:, ONE)= int_Tk_ij(T4, p, [7 8 9]);
  int_T5_ij(:, ONE)= int_Tk_ij(T5, p, [8 10 11]);

  mij_Aa(:, ONE_T1_act_Aa)= int_T1_ij(:, ONE_T1_act_Aa);
  mij_IS(:, ONE_T1_act_Aa)= int_T2_ij(:, ONE_T1_act_Aa) + int_T3_ij(:,
      ONE_T1_act_Aa) + int_T4_ij(:, ONE_T1_act_Aa);
  mij_Ab(:, ONE_T1_act_Aa)= int_T5_ij(:, ONE_T1_act_Aa);

  mij_Aa(:, ONE_T1_act_Ab)= int_T5_ij(:, ONE_T1_act_Ab);
  mij_IS(:, ONE_T1_act_Ab)= int_T2_ij(:, ONE_T1_act_Ab) + int_T3_ij(:,
      ONE_T1_act_Ab) + int_T4_ij(:, ONE_T1_act_Ab);
  mij_Ab(:, ONE_T1_act_Ab)= int_T1_ij(:, ONE_T1_act_Ab);
end
% ------------------------------------------------------------------------

% assemble global mass matrices
M_tilde_Aa= local2global(mij_Aa);
M_tilde_IS= local2global(mij_IS);
```

```
M_tilde_Ab= local2global(mij_Ab);


function ar= Ti_area(x1, x2, x3)
  ar= 0.5*(x1(1, :).*x2(2, :) - ...
           x1(2, :).*x2(1, :) + ...
           x2(1, :).*x3(2, :) - ...
           x2(2, :).*x3(1, :) + ...
           x3(1, :).*x1(2, :) - ...
           x3(2, :).*x1(1, :));
end

function int= int_Tk_ij(Tk, p, quad_pts)
  int= repmat(Tk./3, 6, 1).*[squeeze(sum(p(quad_pts, 1, :).*p(quad_pts, 1, :)))';
                             squeeze(sum(p(quad_pts, 1, :).*p(quad_pts, 2, :)))';
                             squeeze(sum(p(quad_pts, 1, :).*p(quad_pts, 3, :)))';
                             squeeze(sum(p(quad_pts, 2, :).*p(quad_pts, 2, :)))';
                             squeeze(sum(p(quad_pts, 2, :).*p(quad_pts, 3, :)))';
                             squeeze(sum(p(quad_pts, 3, :).*p(quad_pts, 3, :)))'];
end

function M= local2global(mij)
  M= sparse(i112, j233, [mij(2, :) mij(3, :) mij(5, :)], Mesh.np, Mesh.np);
  M= M + M';
  M= M + ...
     spdiags(Mat_np_nt_100*mij(1, :)' + ...
             Mat_np_nt_010*mij(4, :)' + ...
             Mat_np_nt_001*mij(6, :)', 0, Mesh.np, Mesh.np);
end

end
```

APPENDIX B

State constraints

A tailored Cholesky-factor update `R_update_indexchange`

Let $\mathbf{A} \in \mathbb{R}^{m\times m}$ be a symmetric, positive-definite matrix. Consider the index subsets $\mathbb{O}, \tilde{\mathbb{O}} \subset \bullet := \{1, \ldots, m\}$ and set $n := \mathrm{card}(\mathbb{O})$ and $\tilde{n} := \mathrm{card}(\tilde{\mathbb{O}})$. Since the principal minor $\mathbf{A}_{\mathbb{O}}$ is also symmetric and positive-definite (compare [HJ85]) there exists the upper triangular matrix $\mathbf{R}_{\mathbb{O}} \in \mathbb{R}^{n\times n}$ with $\mathbf{A}_{\mathbb{O}} = \mathbf{R}_{\mathbb{O}}\mathbf{R}_{\mathbb{O}}^T$. We assume that $\mathbf{R}_{\mathbb{O}}$ is already given to us.

We are looking for an efficient computation of the upper triangular CHOLESKY-factor $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ that satisfies $\mathbf{A}_{\tilde{\mathbb{O}}} = \tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}^T$, where we essentially make use of the already known factor $\mathbf{R}_{\mathbb{O}}$. Speaking in a functional computer language we aim to implement an efficient routine `R_update_indexchange`, which is called via

$$\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}} = \texttt{R\_update\_indexchange}(\mathbf{A}, \mathbf{R}_{\mathbb{O}}, \mathbb{O}, \tilde{\mathbb{O}}).$$

For $\mathbb{O} = \varnothing$ we of course have to compute the CHOLESKY-factor $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}$ from scratch. On the other hand $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}$ is the empty square matrix if $\tilde{\mathbb{O}} = \varnothing$. We therefore consider $\mathbb{O}, \tilde{\mathbb{O}} \neq \varnothing$. If $\mathbb{O} = \tilde{\mathbb{O}}$ again $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}} = \mathbf{R}_{\mathbb{O}}$ is trivial. Let us assume $\mathbb{O} \neq \tilde{\mathbb{O}}$ from now on. The new factor $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}$ can be obtained by considering loops over the two cases

(1) $k \in \mathbb{O} \setminus \tilde{\mathbb{O}}$ (k-th equation disappears),
(2) $k \in \tilde{\mathbb{O}} \setminus \mathbb{O}$ (k-th equation appears).

Case (1): $\{k\} = \mathbb{O} \setminus \tilde{\mathbb{O}}$. We emphasize the k-th row and column in $\mathbf{A}_{\mathbb{O}}$ and its factorization as follows:

$$\mathbf{A}_{\mathbb{O}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} & \mathbf{A}_{13} \\ \mathbf{a}_{12}^T & a & \mathbf{a}_{23}^T \\ \mathbf{A}_{13}^T & \mathbf{a}_{23} & \mathbf{A}_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{11}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{r}_{12}^T & r & \mathbf{0} \\ \mathbf{R}_{13}^T & \mathbf{r}_{23} & \mathbf{R}_{33}^T \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} & \mathbf{R}_{13} \\ \mathbf{0} & r & \mathbf{r}_{23}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{33} \end{bmatrix} = \mathbf{R}_{\mathbb{O}}^T\mathbf{R}_{\mathbb{O}}.$$

This k-th row and column disappears when considering $\mathbf{A}_{\tilde{\mathbb{O}}}$ and its factorization

$$\mathbf{A}_{\tilde{\mathbb{O}}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{13} \\ \\ \mathbf{A}_{13}^T & \mathbf{A}_{33} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^T & \mathbf{0} \\ \\ \tilde{\mathbf{R}}_{13}^T & \tilde{\mathbf{R}}_{33}^T \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{R}}_{11} & \tilde{\mathbf{R}}_{13} \\ \\ \mathbf{0} & \tilde{\mathbf{R}}_{33} \end{bmatrix} = \tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}^T\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}.$$

One immediately finds

$$\begin{aligned} \tilde{\mathbf{R}}_{11} &= \mathbf{R}_{11}, \\ \tilde{\mathbf{R}}_{13} &= \mathbf{R}_{13}, \\ \tilde{\mathbf{R}}_{33}^T\tilde{\mathbf{R}}_{33} &= \mathbf{R}_{33}^T\mathbf{R}_{33} + \mathbf{r}_{23}\mathbf{r}_{23}^T. \end{aligned}$$

The last equation tells us that the factor $\tilde{\mathbf{R}}_{33}$ is obtained by a rank-1 modification and can easily be computed for instance in Matlab via

$$\tilde{\mathbf{R}}_{33} = \texttt{cholupdate}(\mathbf{R}_{33}, \mathbf{r}_{23}, +).$$

Case (2): $\{k\} = \tilde{\mathbb{O}} \setminus \mathbb{O}$. We now emphasize the absence of the k-th row and column in $\mathbf{A}_{\mathbb{O}}$ and its factorization

$$\mathbf{A}_{\mathbb{O}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{13} \\ \\ \mathbf{A}_{13}^T & \mathbf{A}_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{11}^T & \mathbf{0} \\ \\ \mathbf{R}_{13}^T & \mathbf{R}_{33}^T \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{13} \\ \\ \mathbf{0} & \mathbf{R}_{33} \end{bmatrix} = \mathbf{R}_{\mathbb{O}}^T \mathbf{R}_{\mathbb{O}},$$

where in contrast to

$$\mathbf{A}_{\tilde{\mathbb{O}}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} & \mathbf{A}_{13} \\ \mathbf{a}_{12}^T & a & \mathbf{a}_{23}^T \\ \mathbf{A}_{13}^T & \mathbf{a}_{23} & \mathbf{A}_{33} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{R}}_{11}^T & \mathbf{0} & \mathbf{0} \\ \tilde{\mathbf{r}}_{12}^T & \tilde{r} & \mathbf{0} \\ \tilde{\mathbf{R}}_{13}^T & \tilde{\mathbf{r}}_{23} & \tilde{\mathbf{R}}_{33}^T \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{R}}_{11} & \tilde{\mathbf{r}}_{12} & \tilde{\mathbf{R}}_{13} \\ \mathbf{0} & \tilde{r} & \tilde{\mathbf{r}}_{23}^T \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{R}}_{33} \end{bmatrix} = \tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}^T \tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}$$

these parts appear. A blockwise comparison gives the equations

$$\begin{aligned} \tilde{\mathbf{R}}_{11} &= \mathbf{R}_{11}, \\ \tilde{\mathbf{r}}_{12} &= \mathbf{R}_{11}^{-T} \mathbf{a}_{12}, \\ \tilde{\mathbf{R}}_{13} &= \mathbf{R}_{13}, \\ \tilde{r} &= \sqrt{a - \tilde{\mathbf{r}}_{12}^T \tilde{\mathbf{r}}_{12}}, \\ \tilde{\mathbf{r}}_{23} &= \frac{1}{\tilde{r}} \left(\mathbf{a}_{32} - \mathbf{R}_{13}^T \tilde{\mathbf{r}}_{12} \right), \\ \tilde{\mathbf{R}}_{33}^T \tilde{\mathbf{R}}_{33} &= \mathbf{R}_{33}^T \mathbf{R}_{33} - \tilde{\mathbf{r}}_{23} \tilde{\mathbf{r}}_{23}^T. \end{aligned}$$

Again the last equation tells us that $\tilde{\mathbf{R}}_{33}$ is the outcome of a rank-1 update of $\mathbf{R}_{33}$ and can similarly be realized in Matlab via

$$\tilde{\mathbf{R}}_{33} = \texttt{cholupdate}(\mathbf{R}_{33}, \tilde{\mathbf{r}}_{23}, -).$$

Remark B.1. We indeed observe an efficient exploitation of the already computed factor $\mathbf{R}_{\mathbb{O}}$. The new matrix $\tilde{\mathbf{R}}_{\tilde{\mathbb{O}}}$ is obtained via loops over disappearing and appearing indices in each of which we have to perform rank-1 updates and possibly forward solves.

Remark B.2. Certainly similar considerations are required in order to develop tailored updates for a symmetric indefinite factorization, when the underlying matrix $\mathbf{A}$ is only indefinite and symmetric.

APPENDIX C

Constraints on the gradient of the state

Details for variational L^r-discretization

Let $\hat{T}$ denote the standard 2-simplex in $\mathbb{R}^2$. Further let $\hat{\varphi}_i : \hat{T} \to \mathbb{R}$ with

$$\begin{aligned}\hat{\varphi}_1(\hat{x}_1, \hat{x}_2) &= 1 - \hat{x}_1 - \hat{x}_2,\\ \hat{\varphi}_2(\hat{x}_1, \hat{x}_2) &= \hat{x}_1,\\ \hat{\varphi}_3(\hat{x}_1, \hat{x}_2) &= \hat{x}_2\end{aligned}$$

be the linear finite element basis functions on the standard simplex $\hat{T}$. For given numbers $p_1, p_2, p_3 \in \mathbb{R}$ we define the affine linear function $\hat{p} : \hat{T} \to \mathbb{R}$ by

$$\hat{p}(\hat{x}_1, \hat{x}_2) := p_1\hat{\varphi}_1(\hat{x}_1, \hat{x}_2) + p_2\hat{\varphi}_2(\hat{x}_1, \hat{x}_2) + p_3\hat{\varphi}_3(\hat{x}_1, \hat{x}_2).$$

For abbreviating often recurring cases we further introduce the symbols

$$\begin{aligned}c_{123} &: \text{for } p_1 = p_2 = p_3\\ c_{12,3} &: \text{for } p_1 = p_2 \neq p_3\\ c_{31,2} &: \text{for } p_3 = p_1 \neq p_2\\ c_{23,1} &: \text{for } p_2 = p_3 \neq p_1\\ c_{1,2,3} &: \text{otherwise.}\end{aligned}$$

Lemma C.1. *Let $p_1, p_2, p_3 \in \mathbb{R}$. There holds*

$$\int_{\hat{T}} |\hat{p}(\hat{x})|^{\frac{4}{3}}\, d\hat{x} = \begin{cases} \frac{1}{2}|p_1|^{\frac{4}{3}}, & c_{123}\\ \frac{3}{70}\frac{(7p_2-10p_3)p_2|p_2|^{\frac{4}{3}}+3|p_3|^{\frac{10}{3}}}{(p_2-p_3)^2}, & c_{12,3}\\ \frac{3}{70}\frac{(7p_1-10p_2)p_1|p_1|^{\frac{4}{3}}+3|p_2|^{\frac{10}{3}}}{(p_1-p_2)^2}, & c_{31,2}\\ \frac{3}{70}\frac{3|p_1|^{\frac{10}{3}}+(7p_3-10p_1)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^2}, & c_{23,1}\\ -\frac{9}{70}\frac{(p_2-p_3)|p_1|^{\frac{10}{3}}+(p_3-p_1)|p_2|^{\frac{10}{3}}+(p_1-p_2)|p_3|^{\frac{10}{3}}}{(p_1-p_2)(p_2-p_3)(p_3-p_1)}, & c_{1,2,3}. \end{cases}$$

Proof. For better readability we write $\hat{x} = (x, y)$ instead of $\hat{x} = (\hat{x}_1, \hat{x}_2)$. We compute

$$\int_{\hat{T}} |\hat{p}(\hat{x})|^{\frac{4}{3}}\, d\hat{x} = \int_0^1 \int_0^{1-x} |p_1(1-x-y) + p_2x + p_3y|^{\frac{4}{3}}\, dy\, dx$$

for the different cases. For $p_1 = p_2 = p_3$ the above term simplifies indeed towards

$$\int_0^1 \int_0^{1-x} |p_1|^{\frac{4}{3}}\, dy\, dx = \frac{1}{2}|p_1|^{\frac{4}{3}}.$$

Let $p_1 = p_2 \neq p_3$. The inner primitive is given by

$$\int |p_2(1-y) + p_3y|^{\frac{4}{3}}\, dy = \frac{3}{7}\frac{|p_2(1-y) + p_3y|^{\frac{7}{3}}}{(p_3 - p_2)\mathrm{sign}(p_2(1-y) + p_3y)} + c.$$

Hence,

$$\int_0^{1-x} |p_2(1-y)+p_3y|^{\frac{4}{3}}\,dy \\ = -\frac{3}{7}\frac{|p_2|^{\frac{7}{3}}}{(p_3-p_2)\mathrm{sign}(p_2)} + \frac{3}{7}\frac{|p_2x+p_3(1-x)|^{\frac{7}{3}}}{(p_3-p_2)\mathrm{sign}(p_2x+p_3(1-x))}.$$

The primitive of the last part is given by

$$\int \frac{|p_2x+p_3(1-x)|^{\frac{7}{3}}}{\mathrm{sign}(p_2x+p_3(1-x))} = -\frac{3}{10}\frac{|p_2x+p_3(1-x)|^{\frac{10}{3}}}{(p_3-p_2)} + c,$$

so that we infer

$$\begin{aligned}
&\int_0^1\int_0^{1-x} |p_2(1-y)+p_3y|^{\frac{4}{3}}\,dy\,dx \\
&= -\frac{3}{7}\frac{|p_2|^{\frac{7}{3}}}{(p_3-p_2)\mathrm{sign}(p_2)} + \frac{3}{7}\frac{1}{p_3-p_2}\left(\frac{3}{10}\frac{|p_3|^{\frac{10}{3}}}{p_3-p_2} - \frac{3}{10}\frac{|p_2|^{\frac{10}{3}}}{p_3-p_2}\right) \\
&= \frac{3}{70}\frac{1}{(p_3-p_2)^2}\left(-10|p_2|^{\frac{7}{3}}\mathrm{sign}(p_2)(p_3-p_2) + 3|p_3|^{\frac{10}{3}} - 3|p_2|^{\frac{10}{3}}\right) \\
&= \frac{3}{70}\frac{(7p_2-10p_3)p_2|p_2|^{\frac{4}{3}} + 3|p_3|^{\frac{10}{3}}}{(p_2-p_3)^2}.
\end{aligned}$$

The cases $p_3 = p_1 \neq p_2$ and $p_2 = p_3 \neq p_1$ are obtained by cyclic permutation. The case of mutually different p_i $(i = 1, 2, 3)$ can be computed with a similar technique. □

Lemma C.2. *Let $p_1, p_2, p_3 \in \mathbb{R}$. For*

$$I_i := \int_{\hat{T}} \mathrm{sign}(\hat{p}(\hat{x}))|\hat{p}(\hat{x})|^{\frac{1}{3}}\hat{\varphi}_i(\hat{x})\,d\hat{x}$$

there holds

$$I_1 = \begin{cases}
\frac{1}{6}\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}, & c_{123}\\
\frac{3}{280}\frac{\left(14p_2^2-5(8p_2-7p_3)p_3\right)|p_2|^{\frac{4}{3}}-9|p_3|^{\frac{10}{3}}}{(p_2-p_3)^3}, & c_{12,3}\\
\frac{3}{280}\frac{\left(14p_1^2-5(8p_1-7p_2)p_2\right)|p_1|^{\frac{4}{3}}-9|p_2|^{\frac{10}{3}}}{(p_1-p_2)^3}, & c_{31,2}\\
-\frac{9}{140}\frac{(2p_1-5p_3)p_1|p_1|^{\frac{4}{3}}+(5p_1-2p_3)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^3}, & c_{23,1}\\
-\frac{9}{280}\frac{\left(7(p_2+p_3)p_1-10p_2p_3-4p_1^2\right)(p_2-p_3)p_1|p_1|^{\frac{4}{3}}-3(p_3-p_1)^2|p_2|^{\frac{10}{3}}+3(p_1-p_2)^2|p_3|^{\frac{10}{3}}}{(p_1-p_2)^2(p_2-p_3)(p_3-p_1)^2}, & c_{1,2,3}
\end{cases}$$

$$I_2 = \begin{cases}
\frac{1}{6}\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}, & c_{123}\\
\frac{3}{280}\frac{\left(14p_2^2-5(8p_2-7p_3)p_3\right)|p_2|^{\frac{4}{3}}-9|p_3|^{\frac{10}{3}}}{(p_2-p_3)^3}, & c_{12,3}\\
-\frac{9}{140}\frac{(5p_2-2p_1)p_1|p_1|^{\frac{4}{3}}+(2p_2-5p_1)p_2|p_2|^{\frac{4}{3}}}{(p_1-p_2)^3}, & c_{31,2}\\
\frac{3}{280}\frac{-9|p_1|^{\frac{10}{3}}+\left(14p_3^2-5(8p_3-7p_1)p_1\right)|p_3|^{\frac{4}{3}}}{(p_3-p_1)^3}, & c_{23,1}\\
-\frac{9}{280}\frac{3(p_2-p_3)^2|p_1|^{\frac{10}{3}}+\left(7(p_1+p_3)p_2-10p_1p_3-4p_2^2\right)(p_3-p_1)p_2|p_2|^{\frac{4}{3}}-3(p_1-p_2)^2|p_3|^{\frac{10}{3}}}{(p_1-p_2)^2(p_2-p_3)^2(p_3-p_1)}, & c_{1,2,3}
\end{cases}$$

$$I_3 = \begin{cases} \frac{1}{6}\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}, & c_{123} \\ -\frac{9}{140}\frac{(5p_3-2p_2)p_2|p_2|^{\frac{4}{3}}+(2p_3-5p_2)p_3|p_3|^{\frac{4}{3}}}{(p_2-p_3)^3}, & c_{12,3} \\ \frac{3}{280}\frac{\left(14p_1^2-5(8p_1-7p_2)p_2\right)|p_1|^{\frac{4}{3}}-9|p_2|^{\frac{10}{3}}}{(p_1-p_2)^3}, & c_{31,2} \\ \frac{3}{280}\frac{-9|p_1|^{\frac{10}{3}}+\left(14p_3^2-5(8p_3-7p_1)p_1\right)|p_3|^{\frac{4}{3}}}{(p_3-p_1)^3}, & c_{23,1} \\ -\frac{9}{280}\frac{-3(p_2-p_3)^2|p_1|^{\frac{10}{3}}+3(p_3-p_1)^2|p_2|^{\frac{10}{3}}+\left(7(p_1+p_2)p_3-10p_1p_2-4p_3^2\right)(p_1-p_2)p_3|p_3|^{\frac{4}{3}}}{(p_1-p_2)(p_2-p_3)^2(p_3-p_1)^2}, & c_{1,2,3} \end{cases}$$

Proof. We only consider the second integral I_2 since the others are obtained by cyclic permutation. The first case $p_1 = p_2 = p_3$ easily gives

$$I_2 = \int_0^1 \int_0^{1-x} \mathrm{sign}(p_1)|p_1|^{\frac{1}{3}} x \, dy \, dx = \frac{1}{6}\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}.$$

Let us exemplarily consider the case $p_1 = p_2 \neq p_3$. We have

$$I_2 = \int_0^1 \int_0^{1-x} \mathrm{sign}(p_2(1-y)+p_3y)|p_2(1-y)+p_3y|^{\frac{1}{3}} x \, dy \, dx.$$

The inner primitive is given by

$$\int \mathrm{sign}(p_2(1-y)+p_3y)|p_2(1-y)+p_3y|^{\frac{1}{3}} \, dy = \frac{3}{4}\frac{|p_2(1-y)+p_3y|^{\frac{4}{3}}}{p_3-p_2} + c,$$

so that we continue

$$I_2 = \frac{3}{4(p_3-p_2)} \int_0^1 |p_2x+p_3(1-x)|^{\frac{4}{3}} x \, dx - \frac{3}{8}\frac{|p_2|^{\frac{4}{3}}}{p_3-p_2}.$$

Substituting $p_{23}(x) := p_2x+p_3(1-x)$ and applying partial integration the remaining primitive is given by

$$\begin{aligned} \int |p_{23}(x)|^{\frac{4}{3}} x \, dx &= \frac{3}{7(p_2-p_3)} \left(\mathrm{sign}(p_{23}(x))p_{23}(x)^{\frac{7}{3}} x - \int \mathrm{sign}(p_{23}(x))p_{23}(x)^{\frac{7}{3}} \right) \\ &= -\frac{3}{7(p_3-p_2)} \left(\mathrm{sign}(p_{23}(x))p_{23}(x)^{\frac{7}{3}} x - \frac{3}{10(p_2-p_3)} p_{23}(x)^{\frac{10}{3}} + c \right) \\ &= -\frac{3}{70}\frac{10(p_3-p_2)\mathrm{sign}(p_{23}(x))|p_{23}(x)|^{\frac{7}{3}} x + 3|p_{23}(x)|^{\frac{10}{3}}}{(p_3-p_2)^2} + c. \end{aligned}$$

Therefore we end up with

$$\begin{aligned} I_2 &= \frac{3}{4(p_3-p_2)} \left(-\frac{3}{70}\frac{10(p_3-p_2)\mathrm{sign}(p_2)|p_2|^{\frac{7}{3}} + 3|p_2|^{\frac{10}{3}} - 3|p_3|^{\frac{10}{3}}}{(p_3-p_2)^2} \right) - \frac{3}{8}\frac{|p_2|^{\frac{4}{3}}}{p_3-p_2} \\ &= \frac{3}{280}\frac{-30(p_3-p_2)|p_2|^{\frac{4}{3}}p_2 - 9|p_2|^{\frac{10}{3}} + 9|p_3|^{\frac{10}{3}} - 35(p_3-p_2)^2|p_2|^{\frac{4}{3}}}{(p_3-p_2)^3} \\ &= \frac{3}{280}\frac{(14p_2^2 - 5(8p_2-7p_3)p_3)\,|p_2|^{\frac{4}{3}} - 9|p_3|^{\frac{10}{3}}}{(p_2-p_3)^3}. \end{aligned}$$

The remaining cases are obtained by similar arguments. □

Lemma C.3. *Let $p_1, p_2, p_3 \in \mathbb{R}$ with $(p_1, p_2, p_3) \neq (0, 0, 0)$. For*

$$G_{ij} := \frac{\partial I_i}{\partial p_j}$$

there holds

$$G_{11} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{1}{280}\frac{(14p_2^3-60p_2^2p_3+105p_2p_3^2-140p_3^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}+81|p_3|^{\frac{10}{3}}}{(p_2-p_3)^4}, & c_{12,3} \\ \frac{1}{280}\frac{(14p_1^3-60p_1^2p_2+105p_1p_2^2-140p_2^3)\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}+81|p_2|^{\frac{10}{3}}}{(p_1-p_2)^4}, & c_{31,2} \\ \frac{3}{140}\frac{(2p_1^2-10p_1p_3+35p_3^2)|p_1|^{\frac{4}{3}}-3(10p_1-p_3)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{23,1} \\ -\frac{3}{140}\frac{\begin{array}{r}\big(2p_1^4-7(p_2+p_3)p_1^3+(14p_2^2+29p_2p_3+14p_3^2)p_1^2-40(p_2p_3^2+p_2^2p_3)p_1+35p_2^2p_3^2\big)(p_2-p_3)|p_1|^{\frac{4}{3}} \\ +9(p_3-p_1)^3|p_2|^{\frac{10}{3}} \\ +9(p_1-p_2)^3|p_3|^{\frac{10}{3}}\end{array}}{(p_1-p_2)^3(p_2-p_3)(p_3-p_1)^3}, & c_{1,2,3} \end{cases}$$

$$G_{22} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{1}{280}\frac{(14p_1^3-60p_3p_1^2+105p_3^2p_1-140p_3^3)\mathrm{sign}(p_1)|p_1|^{\frac{1}{3}}+81|p_3|^{\frac{10}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{3}{140}\frac{(2p_2^2-10p_2p_3+35p_3^2)|p_2|^{\frac{4}{3}}-3(10p_2-p_3)p_3|p_3|^{\frac{4}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{1}{280}\frac{81|p_1|^{\frac{10}{3}}+(14p_2^3-60p_1p_2^2+105p_1^2p_2-140p_1^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ -\frac{3}{140}\frac{\begin{array}{r}9(p_2-p_3)^3|p_1|^{\frac{10}{3}} \\ +\big(2p_2^4-7(p_3+p_1)p_2^3+(14p_1^2+29p_1p_3+14p_3^2)p_2^2-40(p_1p_3^2+p_1^2p_3)p_2+35p_1^2p_3^2\big)(p_3-p_1)|p_2|^{\frac{4}{3}} \\ +9(p_1-p_2)^3|p_3|^{\frac{10}{3}}\end{array}}{(p_1-p_2)^3(p_2-p_3)^3(p_3-p_1)}, & c_{1,2,3} \end{cases}$$

$$G_{33} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{3}{140}\frac{-3(10p_3-p_1)p_1|p_1|^{\frac{4}{3}}+(2p_3^2-10p_1p_3+35p_1^2)|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{1}{280}\frac{-(140p_2^3-105p_3p_2^2+60p_3^2p_2-14p_3^3)\mathrm{sign}(p_3)|p_3|^{\frac{1}{3}}+81|p_2|^{\frac{10}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{1}{280}\frac{81|p_1|^{\frac{10}{3}}-(140p_1^3-105p_2p_1^2+60p_2^2p_1-14p_2^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ -\frac{3}{140}\frac{\begin{array}{r}9(p_2-p_3)^3|p_1|^{\frac{10}{3}} \\ +9(p_3-p_1)^3|p_2|^{\frac{10}{3}} \\ +\big(2p_3^4-7(p_1+p_2)p_3^3+(14p_1^2+29p_1p_2+14p_2^2)p_3^2-40(p_1^2p_2+p_1p_2^2)p_3+35p_1^2p_2^2\big)(p_1-p_2)|p_3|^{\frac{4}{3}}\end{array}}{(p_1-p_2)(p_2-p_3)^3(p_3-p_1)^3}, & c_{1,2,3} \end{cases}$$

$$G_{12} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{1}{280}\frac{(14p_2^3-60p_2^2p_3+105p_2p_3^2-140p_3^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}+81|p_3|^{\frac{10}{3}}}{(p_2-p_3)^4}, & c_{12,3} \\ \frac{3}{280}\frac{3(p_2-10p_3)p_2|p_2|^{\frac{4}{3}}+(2p_3^2-10p_2p_3+35p_2^2)|p_3|^{\frac{4}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{3}{140}\frac{3(p_1-10p_2)p_1|p_1|^{\frac{4}{3}}+(2p_2^2-10p_1p_2+35p_1^2)|p_2|^{\frac{4}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}\big(p_1^2-(7p_2+4p_3)p_1+10p_2p_3\big)(p_2-p_3)^2p_1|p_1|^{\frac{4}{3}} \\ -\big(p_2^2-(7p_1+4p_3)p_2+10p_1p_3\big)(p_3-p_1)^2p_2|p_2|^{\frac{4}{3}} \\ +3(p_1-p_2)^3|p_3|^{\frac{10}{3}}\end{array}}{(p_1-p_2)^3(p_2-p_3)^2(p_3-p_1)^2}, & c_{1,2,3} \end{cases}$$

$$G_{21} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{1}{280}\frac{(14p_2^3-60p_2^2p_3+105p_2p_3^2-140p_3^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}+81|p_3|^{\frac{10}{3}}}{(p_2-p_3)^4}, & c_{12,3} \\ \frac{3}{140}\frac{3(p_2-10p_3)p_2|p_2|^{\frac{4}{3}}+(2p_3^2-10p_2p_3+35p_2^2)|p_3|^{\frac{4}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{3}{280}\frac{3(p_1-10p_2)p_1|p_1|^{\frac{4}{3}}+(2p_2^2-10p_1p_2+35p_1^2)|p_2|^{\frac{4}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}\big(p_1^2-(7p_2+4p_3)p_1+10p_2p_3\big)(p_2-p_3)^2p_1|p_1|^{\frac{4}{3}} \\ -\big(p_2^2-(7p_1+4p_3)p_2+10p_1p_3\big)(p_3-p_1)^2p_2|p_2|^{\frac{4}{3}} \\ +3(p_1-p_2)^3|p_3|^{\frac{10}{3}}\end{array}}{(p_1-p_2)^3(p_2-p_3)^2(p_3-p_1)^2}, & c_{1,2,3} \end{cases}$$

$$G_{23} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{3}{280}\frac{(2p_1^2-10p_1p_3+35p_3^2)|p_1|^{\frac{4}{3}}+3(p_3-10p_1)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{3}{140}\frac{3(p_2-10p_3)p_2|p_2|^{\frac{4}{3}}+(2p_3^2-10p_2p_3+35p_2^2)|p_3|^{\frac{4}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{1}{280}\frac{81|p_1|^{\frac{10}{3}}-(140p_1^3-105p_2p_1^2+60p_2^2p_1-14p_2^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}3(p_2-p_3)^3|p_1|^{\frac{10}{3}} \\ +\big(p_2^2-(7p_3+4p_1)p_2+10p_1p_3\big)(p_3-p_1)^2p_2|p_2|^{\frac{4}{3}} \\ -\big(p_3^2-(7p_2+4p_1)p_3+10p_1p_2\big)(p_1-p_2)^2p_3|p_3|^{\frac{4}{3}}\end{array}}{(p_1-p_2)^2(p_2-p_3)^3(p_3-p_1)^2}, & c_{1,2,3} \end{cases}$$

$$G_{32} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{3}{140}\frac{(2p_1^2-10p_1p_3+35p_3^2)|p_1|^{\frac{4}{3}}+3(p_3-10p_1)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{3}{280}\frac{3(p_2-10p_3)p_2|p_2|^{\frac{4}{3}}+(2p_3^2-10p_2p_3+35p_2^2)|p_3|^{\frac{4}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{1}{280}\frac{81|p_1|^{\frac{10}{3}}-(140p_1^3-105p_2p_1^2+60p_2^2p_1-14p_2^3)\mathrm{sign}(p_2)|p_2|^{\frac{1}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}3(p_2-p_3)^3|p_1|^{\frac{10}{3}} \\ +\big(p_2^2-(7p_3+4p_1)p_2+10p_1p_3\big)(p_3-p_1)^2p_2|p_2|^{\frac{4}{3}} \\ -\big(p_3^2-(7p_2+4p_1)p_3+10p_1p_2\big)(p_1-p_2)^2p_3|p_3|^{\frac{4}{3}}\end{array}}{(p_1-p_2)^2(p_2-p_3)^3(p_3-p_1)^2}, & c_{1,2,3} \end{cases}$$

$$G_{31} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{3}{140}\frac{(2p_1^2-10p_1p_3+35p_3^2)|p_1|^{\frac{4}{3}}+3(p_3-10p_1)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{1}{280}\frac{81|p_2|^{\frac{10}{3}}-(140p_2^3-105p_3p_2^2+60p_3^2p_2-14p_3^3)\mathrm{sign}(p_3)|p_3|^{\frac{1}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{3}{280}\frac{3(p_1-10p_2)p_1|p_1|^{\frac{4}{3}}+(2p_2^2-10p_1p_2+35p_1^2)|p_2|^{\frac{4}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}-\big(p_1^2-(7p_3+4p_2)p_1+10p_2p_3\big)(p_2-p_3)^2p_1|p_1|^{\frac{4}{3}} \\ +3(p_3-p_1)^3|p_2|^{\frac{10}{3}} \\ +\big(p_3^2-(7p_1+4p_2)p_3+10p_1p_2\big)(p_1-p_2)^2p_3|p_3|^{\frac{4}{3}}\end{array}}{(p_1-p_2)^2(p_2-p_3)^2(p_3-p_1)^3}, & c_{1,2,3} \end{cases}$$

$$G_{13} = \begin{cases} \frac{1}{18}|p_1|^{-\frac{2}{3}}, & c_{123} \\ \frac{3}{280}\frac{(2p_1^2-10p_1p_3+35p_3^2)|p_1|^{\frac{4}{3}}+3(p_3-10p_1)p_3|p_3|^{\frac{4}{3}}}{(p_3-p_1)^4}, & c_{12,3} \\ \frac{1}{280}\frac{81|p_2|^{\frac{10}{3}}-(140p_2^3-105p_3p_2^2+60p_3^2p_2-14p_3^3)\mathrm{sign}(p_3)|p_3|^{\frac{1}{3}}}{(p_2-p_3)^4}, & c_{31,2} \\ \frac{3}{140}\frac{3(p_1-10p_2)p_1|p_1|^{\frac{4}{3}}+(2p_2^2-10p_1p_2+35p_1^2)|p_2|^{\frac{4}{3}}}{(p_1-p_2)^4}, & c_{23,1} \\ \frac{9}{280}\frac{\begin{array}{r}-\left(p_1^2-(7p_3+4p_2)p_1+10p_2p_3\right)(p_2-p_3)^2p_1|p_1|^{\frac{4}{3}} \\ +3(p_3-p_1)^3|p_2|^{\frac{10}{3}} \\ +\left(p_3^2-(7p_1+4p_2)p_3+10p_1p_2\right)(p_1-p_2)^2p_3|p_3|^{\frac{4}{3}}\end{array}}{(p_1-p_2)^2(p_2-p_3)^2(p_3-p_1)^3}, & c_{1,2,3} \end{cases}$$

The proof is omitted since the same techniques apply as already used to prove Lemma C.1 and C.2. These formulas are also checked by the computer algebra software Maple.

Bibliography

[ACT02] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Comput Optim Appl*, 23(2):201–229, 2002.

[ADN59] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Commun. Pure Appl. Math.*, 12:623–727, 1959.

[Alt06] H.W. Alt. *Linear functional analysis. An application oriented introduction. (Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung.) 5th revised ed.* Berlin: Springer, 2006.

[AO00] M. Ainsworth and J.T. Oden. *A posteriori error estimation in finite element analysis.* Pure and Applied Mathematics. A Wiley-Interscience Series of Texts, Monographs, and Tracts. Chichester: John Wiley & Sons, Inc., 2000.

[ARS09] T. Apel, A. Rösch, and D. Sirch. l^∞-error estimates on graded meshes with application to optimal control. *SIAM J. Control Optim.*, 48(3):1771–1796, 2009.

[ARW06] T. Apel, A. Rösch, and G. Winkler. Discretization error estimates for an optimal control problem in a nonconvex domain. In *Numerical mathematics and advanced applications. Proceedings of ENUMATH 2005*, pages 299–307, Berlin: Springer, 2006.

[BC05] C. Bahriawati and C. Carstensen. Three Matlab implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control. *Comput. Methods Appl. Math.*, 5(4):333–361, 2005.

[BC08] S. Bartels and C. Carstensen. A convergent adaptive finite element method for an optimal design problem. *Numer. Math.*, 108(3):359–385, 2008.

[BDD04] P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219–268, 2004.

[Ber89] C. Bernardi. Optimal finite-element interpolation on curved domains. *SIAM J. Numer. Anal.*, 26(5):1212–1240, 1989.

[Ber04] M. Berggren. Approximations of very weak solutions to boundary-value problems. *SIAM J. Numer. Anal.*, 42(2):860–877, 2004.

[BF91] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods.* Springer Series in Computational Mathematics. 15. New York: Springer, 1991.

[BGHvBW03] L.T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders. *Large-scale PDE-constrained optimization.* Lecture Notes in Computational Science and Engineering, 30. Springer, 2003.

[BHHK00] 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(2):495–521, 2000.

[BK94] J.H. Bramble and J.T. King. A robust finite element method for nonhomogeneous Dirichlet problems in domains with curved boundaries. *Math. Comput.*, 63(207):1–17, 1994.

[BK03] M. Bergounioux and K. Kunisch. On the structure of Lagrange multipliers for state-constrained optimal control problems. *Syst. Control Lett.*, 48(3-4):169–176, 2003.

[BKR00] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concept. *SIAM J. Control Optim.*, 39(1):113–132, 2000.

[BR78] I. Babuška and W.C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978.

[BR96] R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math.*, 4(4):237–264, 1996.

[BR01] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[BR03] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics, ETH Zürich. Basel: Birkhäuser, 2003.

[Bra07] D. Braess. *Finite elements. Theory, fast solvers and applications in elasticity theory. (Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.) 4th revised and extended ed.* Berlin: Springer, 2007.

[BS08] S.C. Brenner and R.L. Scott. *The mathematical theory of finite element methods. 3rd ed.* Texts in Applied Mathematics 15. New York: Springer, 2008.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[BV09] O. Benedix and B. Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Comput Optim Appl*, 44(1):3–25, 2009.

[BW85] R.E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comput.*, 44:283–301, 1985.

[BX03] R.E. Bank and J. Xu. Asymptotically exact a posteriori error estimators. I: Grids with superconvergence. *SIAM J. Numer. Anal.*, 41(6):2294–2312, 2003.

[Cas85] E. Casas. L^2 estimates for the finite element method for the Dirichlet problem with singular data. *Numer. Math.*, 47:627–632, 1985.

[Cas86] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24:1309–1318, 1986.

[Cas93] E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM J. Control Optim.*, 31(4):993–1006, 1993.

[Cas97] E. Casas. Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations. *SIAM J. Control Optim.*, 35(4):1297–1327, 1997.

[Cas02] E. Casas. Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints. *ESAIM, Control Optim. Calc. Var.*, 8:345–374, 2002.

[CF93] E. Casas and L.A. Fernández. Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state. *Appl. Math. Optimization*, 27(1):35–56, 1993.

[CGM10] E. Casas, A. Günther, and M. Mateos. A paradox in the approximation of Dirichlet control problems in curved domains. Technical report, submitted, 2010.

[Cia80] P.G. Ciarlet. *The finite element method for elliptic problems.* Studies in Mathematics and its Applications, Vol. 4. Amsterdam: North-Holland Publishing Company, 1980.

[CKR08] S. Cherednichenko, K. Krumbiegel, and A. Rösch. Error estimates for the Lavrentiev regularization of elliptic optimal control problems. *Inverse Probl.*, 24(5):Article ID 055003, 2008.

[CM02] E. Casas and M. Mateos. Uniform convergence of the FEM. Applications to state constrained control problems. *Comput. Appl. Math.*, 21(1):67–100, 2002.

[CM08] E. Casas and M. Mateos. Error estimates for the numerical approximation of Neumann control problems. *Comput Optim Appl*, 39(3):265–295, 2008.

[CR06] E. Casas and J.-P. Raymond. Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations. *SIAM J. Control Optim.*, 45(5):1586–1611, 2006.

[CR09] S. Cherednichenko and A. Rösch. Error estimates for the discretization of elliptic control problems with pointwise control and state constraints. *Comput Optim Appl*, 44(1):27–55, 2009.

[CS10] E. Casas and J. Sokolowski. Approximation of boundary control problems on curved domains. *SIAM J. Control Optim.*, 48(6):3746–3780, 2010.

[CT03] E. Casas and F. Tröltzsch. Error estimates for linear-quadratic elliptic control problems. In *Analysis and optimization of differential systems. IFIP TC7/WG 7.2*, pages 89–100, Boston, MA: Kluwer Academic Publishers, 2003.

[Dau92] M. Dauge. Neumann and mixed problems on curvilinear polyhedra. *Integral Equations Oper. Theory*, 15(2):227–261, 1992.

[Dem04] A. Demlow. Localized pointwise error estimates for mixed finite element methods. *Math. Comput.*, 73(248):1623–1653, 2004.

[DGH07] K. Deckelnick, A. Günther, and M. Hinze. Numerical analysis and algorithms in control and state constrained optimization with pdes. In *PAMM*, volume 7, pages 1060503–1060504, 2007.

[DGH08] K. Deckelnick, A. Günther, and M. Hinze. Discrete concepts for elliptic optimal control problems with constraints on the gradient of the state. In *PAMM*, volume 8, pages 10863–10864, 2008.

[DGH09a] K. Deckelnick, A. Günther, and M. Hinze. A priori estimates for optimal Dirichlet boundary control problems. In *PAMM*, volume 9, pages 611–612, 2009.

[DGH09b] K. Deckelnick, A. Günther, and M. Hinze. Finite element approximation of Dirichlet boundary control for elliptic pdes on two- and three-dimensional curved domains. *SIAM J. Control Optim.*, 48(4):2798–2819, 2009.

[DGH09c] K. Deckelnick, A. Günther, and M. Hinze. Finite element approximation of elliptic control problems with constraints on the gradient. *Numer. Math.*, 111(3):335–350, 2009.

[DH99] T.A. Davis and W.W. Hager. Modifying a sparse Cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 20(3):606–627, 1999.

[DH07a] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM J. Numer. Anal.*, 45(5):1937–1953, 2007.

[DH07b] K. Deckelnick and M. Hinze. A finite element approximation to elliptic control problems in the presence of control and state constraints. Technical Report Preprint HBAM2007-01, Hamburger Beiträge zur Angewandten Mathematik, Universität Hamburg, 2007.

[DH08] K. Deckelnick and M. Hinze. Numerical analysis of a control and state constrained elliptic control problem with piecewise constant control approximations. In K. Kunisch, G. Of, and O. Steinbach, editors, *Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2007*, pages 597–604. Springer, 2008.

[DH09] K. Deckelnick and M. Hinze. Variational discretization of parabolic control problems in the presence of pointwise state constraints. *Journal of Computational Mathematics*, 2009. accepted.

[dLRMRT08] J.C. de Los Reyes, P. Merino, J. Rehberg, and F. Tröltzsch. Optimality conditions for state-constrained PDE control problems with time-dependent controls. *Control Cybern.*, 37(1):5–38, 2008.

[Dör96] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.

[EG04] A. Ern and J.-L. Guermond. *Theory and practice of finite elements.* Applied Mathematical Sciences 159. New York: Springer, 2004.

[Eva98] L.C. Evans. *Partial differential equations.* Graduate Studies in Mathematics. 19. Providence, RI: American Mathematical Society (AMS), 1998.

[Fal73] R.S. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.*, 44(1):28–47, 1973.

[Fat99] H.O. Fattorini. *Infinite dimensional optimization and control theory.* Encyclopedia of Mathematics and Its Applications. 62. Cambridge: Cambridge University Press., 1999.

[FF91] H.O. Fattorini and H. Frankowska. Infinite dimensional control problems with state constraints. In *Lect. Notes Control Inf. Sci.*, volume 154, pages 52–62, 1991.

[FGH98] A.V. Fursikov, M.D. Gunzburger, and L.S. Hou. Boundary value problems and optimal boundary control for the Navier-Stokes system: The two-dimensional case. *SIAM J. Control Optim.*, 36(3):852–894, 1998.

[Fia83] A.V. Fiacco. *Introduction to sensitivity and stability analysis in nonlinear programming.* Mathematics in Science and Engineering, Vol. 165. New York - London: Academic Press, 1983.

[Fri69] A. Friedman. *Partial differential equations.* New York etc.: Holt, Rinehart and Winston, Inc., 1969.

[Gev79] T. Geveci. On the approximation of the solution of an optimal control problem governed by an elliptic equation. *RAIRO, Anal. Numér.*, 13(4):313–328, 1979.

[GH08] A. Günther and M. Hinze. A posteriori error control of a state constrained elliptic control problem. *J. Numer. Math.*, 16(4):307–322, 2008.

[GH09] A. Günther and M. Hinze. Elliptic control problems with gradient constraints - variational discrete versus piecewise constant controls. *Comput Optim Appl*, 2009. DOI: 10.1007/s10589-009-9308-8.

[GK99] C. Geiger and C. Kanzow. *Numerical methods for solution of unconstrained optimization problems. (Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.)*. Berlin: Springer, 1999.

[GK02] C. Geiger and C. Kanzow. *Theory and numerics of constrained problems of optimization. (Theorie und Numerik restringierter Optimierungsaufgaben.)*. Berlin: Springer, 2002.

[GLRS09] J. Guzmán, D. Leykekhman, J. Rossmann, and A.H. Schatz. Hölder estimates for Green's functions on convex polyhedral domains and their applications to finite element methods. *Numer. Math.*, 112(2):221–243, 2009.

[GN88] L. Gastaldi and R.H. Nochetto. On L^∞-accuracy of mixed finite element methods for second order elliptic problems. *Mat. Aplic. Comp.*, 7(1):13–39, 1988.

[GN89] L. Gastaldi and R.H. Nochetto. Sharp maximum norm error estimates for general mixed finite element approximations to second order elliptic equations. *RAIRO, Modélisation Math. Anal. Numér.*, 23(1):103–128, 1989.

[GR05] C. Großmann and H.-G. Roos. *The numerical analysis of partial differential equations. (Numerische Behandlung partieller Differentialgleichungen.) 3rd revised and expanded ed.* Teubner Studienbücher Mathematik. Wiesbaden: Teubner, 2005.

[Gri92] P. Grisvard. Singularities in boundary value problems. *Recherches en Mathématiques Appliquées (Research in Applied Mathematics), Masson, Paris*, 22, 1992.

[Grö89] K. Gröger. A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. *Math. Ann.*, 283(4):679–687, 1989.

[GT01] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order. Reprint of the 1998 ed.* Classics in Mathematics. Berlin: Springer, 2001.

[GT09] A. Günther and M.H. Tber. A goal-oriented adaptive Moreau-Yosida algorithm for control- and state-constrained elliptic control problems. Preprint SPP1253-089, DFG Schwerpunktprogramm 1253, 2009.

[Gün06] A. Günther. Gittersteuerung bei zustandsrestringierten Optimalsteuerungsproblemen. Diploma thesis. TU Dresden, 2006.

[Hac10] W. Hackbusch. *Elliptic differential equations: theory and numerical treatment.* Springer Series in Computational Mathematics 18. Dordrecht: Springer, 2010.

[HH08] M. Hintermüller and R.H.W. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM J. Control Optim.*, 47(4):1721–1743, 2008.

[HH09a] M. Hintermüller and M. Hinze. Moreau–Yosida regularization in state constrained elliptic control problems: Error estimates and parameter adjustment. *SIAM J. Numer. Anal.*, 47(3):1666–1683, 2009.

[HH09b] M. Hintermüller and R.H.W. Hoppe. Goal-oriented adaptivity in pointwise state constrained optimal control of partial differential equations. Technical Report Preprint Nr. 2009-16, Institut für Mathematik, Universität Augsburg, 2009.

[HH09c] M. Hintermüller and R.H.W. Hoppe. Goal oriented mesh adaptivity for mixed control-state constrained elliptic optimal control problems. In W. Fitzgibbon et al., editor, *Int. Conf. on Sci. Comput. in Simulation, Optimization and Control*, Jyväskylä, Finland, 2009.

[HHIK08] 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(3):540–560, 2008.

[HIK03] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.

[Hin05] M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comput Optim Appl*, 30(1):45–61, 2005.

[Hin08] M. Hinze. A priori and a posteriori error control for elliptic control problems with pointwise constraints. *Oberwolfach Rep.*, 5(1):613–615, 2008. Optimal Control of Coupled Systems of PDE, Report No. 13/2008.

[HJ85] R.A. Horn and C.R. Johnson. *Matrix analysis.* Cambridge etc.: Cambridge University Press, 1985.

[HK04] M. Hinze and K. Kunisch. Second order methods for boundary control of the instationary Navier-Stokes system. *ZAMM*, 84(3):171–187, 2004.

[HK06a] M. Hintermüller and K. Kunisch. Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[HK06b] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, 17(1):159–187, 2006.

[HK07] R.H.W. Hoppe and M. Kieweg. A posteriori error estimation of finite element approximations of pointwise state constrained distributed control problems. Technical Report Preprint Nr. 2007-16, Institut für Mathematik, Universität Augsburg, 2007.

[HK08] R.H.W. Hoppe and M. Kieweg. Adaptive finite element methods for mixed control-state constrained optimal control problems for elliptic boundary value problems. *Comput Optim Appl*, 2008. 10.1007/s10589-008-9195-4.

[HK09] M. Hintermüller and K. Kunisch. Pde-constrained optimization subject to pointwise constraints on the control, the state and its derivative. *SIAM J. Optim.*, 20(3):1133–1156, 2009.

[HM07] M. Hinze and C. Meyer. Stability of infinite dimensional control problems with pointwise state constraints. Technical Report No. 1236, Weierstraß-Institut für Angewandte Analysis und Stochastik, 2007.

[HM08] M. Hinze and C. Meyer. Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems. *Comput Optim Appl*, 2008. DOI 10.1007/s10589-008-9198-1.

[HM09] M. Hinze and U. Matthes. A note on variational discretization of elliptic Neumann boundary control. *Control Cybern.*, 38:577–591, 2009.

[HPUU09] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints.* Mathematical Modelling: Theory and Applications 23. Dordrecht: Springer, 2009.

[HS09] 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*, 2009. DOI 10.1007/s10589-009-9278-x.

[HV09] M. Hinze and M. Vierling. Variational discretization and semi-smooth Newton methods; implementation, convergence and globalization in pde constrained optimization with control constraints. Technical Report Preprint HBAM2009-13, Hamburger Beiträge zur Angewandten Mathematik, Universität Hamburg, 2009.

[IK08] K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications.* Advances in Design and Control 15. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2008.

[IKP06] K. Ito, K. Kunisch, and G.H. Peichl. Variational approach to shape derivatives for a class of Bernoulli problems. *J. Math. Anal. Appl.*, 314(1):126–149, 2006.

[JK95] D. Jerison and C.E. Kenig. The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Funct. Anal.*, 130(1):161–219, 1995.

[Jos07] J. Jost. *Partial Differential Equations. 2nd ed.* Graduate Texts in Mathematics 214. New York: Springer, 2007.

[JT81] C. Johnson and V. Thomée. Error estimates for some mixed finite element methods for parabolic type problems. *RAIRO, Anal. Numér.*, 15:41–78, 1981.

[KRS10] K. Kohls, A. Rösch, and K.G. Siebert. A posteriori error estimators for control constrained optimal control problems. Technical Report No. SM-DU-711, Universität Duisburg-Essen, 2010.

[KV07] K. Kunisch and B. Vexler. Constrained Dirichlet boundary control in L^2 for a class of evolution equations. *SIAM J. Control Optim.*, 46(5):1726–1753, 2007.

[Lio71] J.L. Lions. *Optimal control of systems governed by partial differential equations.* Die Grundlehren der mathematischen Wissenschaften. Band 170. Berlin: Springer, 1971.

[LM72] J.L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Die Grundlehren der mathematischen Wissenschaften. Band 181. Berlin: Springer, 1972.

[LY01] W. Liu and N. Yan. A posteriori error estimates for distributed convex optimal control problems. *Adv. Comput. Math.*, 15(1):285–309, 2001.

[Mal82] K. Malanowski. Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems. *Appl. Math. Optimization*, 8(1):69–95, 1982.

[Mey08] C. Meyer. Error estimates for the finite element approximation of an elliptic control problem with pointwise constraints on the state and the control. *Control Cybern.*, 37:51–85, 2008.

[Mif77] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, 15:959–972, 1977.

[MNS00] P. Morin, R.H. Nochetto, and K.G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488, 2000.

[MR04] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. *SIAM J. Control Optim.*, 43(3):970–985, 2004.

[MR06] C. Meyer and A. Rösch. L^∞-estimates for approximated optimal control problems. *SIAM J. Control Optim.*, 44(5):1636–1649, 2006.

[MRT06] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Comput Optim Appl*, 33(2-3):209–228, 2006.

[MRV08] S. May, R. Rannacher, and B. Vexler. Error analysis of a finite element approximation of elliptic Dirichlet boundary control problems. Preprint 05/2008, Lehrstuhl für Numerische Mathematik, Universität Heidelberg, 2008.

[MRV10] D. Meidner, R. Rannacher, and B. Vexler. A priori error estimates for finite element discretizations of parabolic optimization problems with pointwise state constraints in time. Preprint SPP1253-098, DFG Schwerpunktprogramm 1253, 2010.

[MV07] D. Meidner and B. Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):116–142, 2007.

[MV08a] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I: Problems without control constraints. *SIAM J. Control Optim.*, 47(3):1150–1177, 2008.

[MV08b] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. II: Problems with control constraints. *SIAM J. Control Optim.*, 47(3):1301–1329, 2008.

[NST06] P. Neittaanmaki, J. Sprekels, and D. Tiba. *Optimization of elliptic systems. Theory and applications.* Springer Monographs in Mathematics. New York, NY: Springer, 2006.

[NT08] I. Neitzel and F. Tröltzsch. On convergence of regularization methods for nonlinear parabolic optimal control problems with control and state constraints. *Control Cybern.*, 37(4):1045–1064, 2008.

[NT09] I. Neitzel and F. Tröltzsch. On regularization methods for the numerical solution of parabolic control problems with pointwise state constraints. *ESAIM, Control Optim. Calc. Var.*, 15(2):426–453, 2009.

[NW06] J. Nocedal and S.J. Wright. *Numerical optimization. 2nd ed.* Springer Series in Operations Research and Financial Engineering. New York, NY: Springer, 2006.

[OPS10] G. Of, T.X. Phan, and O. Steinbach. Boundary element methods for Dirichlet boundary control problems. *Math. Methods Appl. Sci.*, 2010. accepted.

[OW09] C. Ortner and W. Wollner. A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state. Preprint SPP1253-071, DFG Schwerpunktprogramm 1253, 2009.

[QS93] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Program., Ser. A*, 58(3):353–367, 1993.

[Ran05] R. Rannacher. Adaptive solution of pde-constrained optimal control problems. In W. Liu, editor, *The 2nd Intern. Conf. on Scientific Computing and Partial Differential Equations (SCPDE05)*, 2005.

[Rös06] A. Rösch. Error estimates for linear-quadratic control problems with control constraints. *Optim. Methods Softw.*, 21(1):121–134, 2006.

[RT77] P.A. Raviart and J.M. Thomas. A mixed finite element method for second-order elliptic problems. Math. Aspects Finite Elem. Meth., Lect. Notes Math. 606, 292–315, 1977.

[Sch06] A. Schiela. *The Control Reduced Interior Point Method - A Function Space Oriented Algorithmic Approach.* PhD thesis, Verlag Dr. Hut, München, 2006.

[Sch09a] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.*, 20(2):1002–1031, 2009.

[Sch09b] A. Schiela. State constrained optimal control problems with states of low regularity. *SIAM J. Control Optim.*, 48(4):2407–2432, 2009.

[SF73] G. Strang and G.J. Fix. *An analysis of the finite element method.* Prentice-Hall Series in Automatic Computation. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973.

[SG09] A. Schiela and A. Günther. Interior point methods in function space for state constraints - inexact Newton and adaptivity. Preprint SPP1253-08-06, DFG Schwerpunktprogramm 1253, 2009.

[Sha97] A. Shapiro. On uniqueness of Lagrange multipliers in optimization problems subject to cone constraints. *SIAM J. Optim.*, 7(2):508–518, 1997.

[SV08] M. Schmich and B. Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM J. Sci. Comput.*, 30(1):369–393, 2008.

[SZ90] L.R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comput.*, 54(190):483–493, 1990.

[Trö05] F. Tröltzsch. *Optimal control of partial differential equations. Theory, procedures, and applications. (Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen.)*. Wiesbaden: Vieweg, 2005.

[Ver96] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Wiley-Teubner Series Advances in Numerical Mathematics. Chichester: John Wiley & Sons. Stuttgart: B. G. Teubner, 1996.

[Vex07] B. Vexler. Finite element approximation of elliptic Dirichlet optimal control problems. *Numer. Funct. Anal. Optim.*, 28(7-8):957–973, 2007.

[Vie09] M. Vierling. An error estimate for the variational discretization of semilinear optimal control problems in the presence of pointwise control and state constraints. Technical Report Preprint HBAM2009-11, Hamburger Beiträge zur Angewandten Mathematik, Universität Hamburg, 2009.

[Vog06] W. Vogt. Adaptive Verfahren zur numerischen Quadratur und Kubatur. Technical Report Preprint No. M 1/06, IfMath TU Ilmenau, 2006.

[VW08] B. Vexler and W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.*, 47(1):509–534, 2008.

[Wol08] W. Wollner. A posteriori error estimates for a finite element discretization of interior point methods for an elliptic optimization problem with state constraints. *Comput Optim Appl*, 2008. 10.1007/s10589-008-9209-2.

[ZK79] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optimization*, 5:49–62, 1979.

Name: Andreas Günther
Titel: Structure exploiting Galerkin schemes for optimal control of PDEs with constraints on the involved variables
Jahr der Drucklegung: 2010

Zusammenfassung

Gegenstand dieser Arbeit ist die Untersuchung strukturausnutzender Galerkin Methoden für die Optimierung elliptischer partieller Differentialgleichungen. Die wesentliche Nichtlinearität der Probleme kommt durch Hinzunahme von Schranken an die Kontrolle, den Zustand und dessen Gradienten zum Tragen. Die betonte Struktur-Spezifik äußert sich zum einen durch konsequente Anwendung des variationellen Diskretisierungskonzeptes für die Steuerung nach [Hin05]. Diese Technik ermöglicht eine elegante und fundierte a priori Fehleranalyse für die diskretisierten Optimierungsprobleme. Zum anderen ermöglicht dieser minimal-invasive Ansatz die Vermeidung von Steuerungsfehlertermen in a posteriori Fehlerschätzern. Mit Hilfe eines solchen Werkzeuges werden ferner durch adaptive Verfeinerung problemangepasste Finite Elemente-Räume gefunden. Zahlreiche numerische Experimente untermauern einerseits bewiesene a priori Fehlerabschätzungen, andererseits die Robustheit zielorientierter Fehlerschätzer und den durch Modellreduktion resultierenden Performancegewinn.

Angelehnt an [DGH09b] werden in Kapitel 2 optimale Randsteuerungsprobleme unter Kontrollschranken auf glatt berandeten 2- und 3-dimensionalen Gebieten behandelt. Erstmalig werden Konvergenzordnungen für allgemeine quasi-uniforme Gitter bewiesen. Für den 2d-Fall kann unter speziellen Gittervoraussetzungen und Anwendung eines Superkonvenz-Lemmas sogar ein verbessertes Resultat gezeigt werden. Diese Ergebnisse werden ferner in zahlreichen numerischen Studien anhand analytischer Beispiele verifiziert. Auf Seiten der beschränkten, verteilten Steuerung werden nützliche Notationen zur variationellen Diskretisierung eingeführt und deren Vorteilhaftigkeit auch numerisch gezeigt.

Kapitel 3 widmet sich Optimalsteuerungsproblemen mit zusätzlichen Schranken an den Zustand. Nach ausführlicher Diskussion bereits verfügbarer a priori Fehlerabschätzungen liegt der Schwerpunkt im Entwurf und der Analyse von zielorientierten adaptiven Konzepten. Bei den zugrunde liegenden diskretisierten Problemen wird sowohl der unregularisierte Ansatz als auch Moreau-Yosida-Penalisierung verfolgt. Unter alleiniger Verwendung der numerischen Lösungen werden wie in [GH08] und [GT09] auswertbare Fehlerschätzer zur zielgenauen Darstellung des Kostenfunktionales entwickelt. Dazu werden numerische Experimente zur Effizienzmessung der Schätzer durchgeführt.

Abschließend werden in Kapitel 4 Schranken an den Gradienten des Zustandes betrachtet. Die Regularitätstheorie erfordert die separate Untersuchung zweier Szenarien. Zum einen werden nach [DGH09c] für ein rein quadratisches Zielfunktional unter Hinzunahme von Kontrollschranken erstmalig Konvergenzaussagen bewiesen. Zum anderen werden diese Abschätzungen wie in [GH09] durch den verbleibenden Fall einer L^r-Regularisierung der unbeschränkten Kontrolle ergänzt. Experimentelle Konvergenzraten werden auch hier für beide Szenarien gemessen.

Lebenslauf

Andreas Günther

06.05.1982	geboren in Bautzen
1988 - 1992	Teiloberschule Puschwitz zur POS Neschwitz
1992 - 2000	Schiller-Gymnasium in Bautzen
2000 - 2001	Zivildienst bei der ev.-luth. Kirchgemeinde Königswartha
10/2001 - 11/2006	Studium an der Technischen Universität Dresden Studiengang Technomathematik
09/2005 - 01/2006	Auslandssemester an der Strathclyde University in Glasgow
11/2006	Diplommathematiker (Technomathematik)
12/2006 - 01/2010	Wissenschaftlicher Mitarbeiter des DFG-SPP 1253 *Optimierung mit partiellen Differentialgleichungen* am Department Mathematik der Universität Hamburg Bereich Optimierung und Approximation
03/2010 -	Wissenschaftlicher Mitarbeiter des DFG Matheon-F2 *Atlas-based 3D Image Segmentation* am Konrad-Zuse-Zentrum für Informationstechnik Berlin, Numerical Analysis and Modelling