Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

vorgelegt von
Diplom-Mathematiker Winnifried Wollner
aus Henstedt-Ulzburg

Tag der mündlichen Prüfung: 22.07.2010

Adaptive Methods for PDE-based Optimal Control with Pointwise Inequality Constraints

19. April 2010

Gutachter: Prof. Dr. Rolf Rannacher

Gutachter: Prof. Dr. Hans Georg Bock

Abstract

This work is devoted to the development of efficient numerical methods for a certain class of PDE-based optimization problems. The optimization is constraint by an elliptic PDE. In addition to prior work in this context pointwise inequality constraints on the control and state variable are considered. These problems are infinite dimensional and their solution can in general not be obtained exactly. Instead the solution of such problems means to find an approximate solution. This is done by (approximately) solving for some set of first order necessary optimality conditions. Hence an efficient algorithm has to find such an approximate solution with as little effort as possible while still being accurate enough for whatever the goal of the computation is.

The work at hand contributes to this goal by deriving a posteriori error estimates with respect to a given functional. These estimates are required for two purposes, first, to generate efficient meshes for the solution of the PDEs required in the process of solving the necessary conditions. Second, to choose several parameters that occur in order to regularize the problems at hand in such a way that the regularization error is both small enough, to obtain a 'good result', and yet large enough to have 'easy to solve' problems.

These a posteriori estimators are supplemented with a priori estimates in several cases where non have been available in the literature for the problem class under consideration.

Finally, all theory and all heuristics will be substantiated with several numerical examples of different complexity.

Zusammenfassung

Ziel der Arbeit ist es effiziente numerische Verfahren zur Lösung von PDE basierten Optimierungsproblemen zu entwickeln. Hierbei betrachten wir als Nebenbedingung eine elliptische PDE sowie im Unterschied zu früheren Arbeiten zusätzliche (Ungleichungs-) Beschränkungen an die Kontroll- und Zustandsvariablen. Es handelt sich bei diesen Problemen um unendlichdimensionale Optimierungsprobleme, so dass die Lösung im Allgemeinen nicht exakt bestimmt werden kann. Stattdessen wird eine Approximation bestimmt. Diese erhält man durch die (approximative) Lösung geeigneter Systeme notwendiger Optimalitätsbedingungen. Ein effizienter Algorithmus hat die Aufgabe, eine solche approximative Lösung mit so wenig Aufwand wie möglich und dennoch hinreichend genau zu bestimmen.

Die vorliegende Arbeit leistet hierzu einen Beitrag indem a posteriori Fehlerschätzer, bezüglich eines gegebenen Funktionals, hergeleitet werden. Diese werden aus zwei Gründen benötigt. Zum Einen, um sparsame Gitter für die Lösung der auftretenden PDEs zu erzeugen. Zum Anderen, um diverse Parameter zur Regularisierung des Problems derart zu steuern, dass einerseits der Regularisierungsfehler „klein genug" ist und andererseits die Probleme noch immer „einfach zu lösen" sind.

Ferner werden die a posteriori Schätzer durch a priori Fehleranalysen ergänzt sofern solche für die betrachtete Problemklasse noch nicht in der Literatur verfügbar waren.

Schließlich werden die theoretischen Resultate und die verwendeten Heuristiken durch mehrere Beispiele unterschiedlicher Komplexität untermauert.

Contents

1 Introduction

This work is devoted to the development of efficient numerical methods for solving optimization problems subject to an elliptic PDE constraint and additional (inequality) constraints on the control variable and zero or first-order constraints on the state variable. Since these problems exhibit very rough adjoint variables their solution usually requires some kind of regularization.

To state things precise, we will consider optimization problems that are constrained by an elliptic partial differential equation

$$A(q,u) = f.$$

If the control is given by external forces the operator A is typically given in the form

$$A(q,u) = \bar{A}(u) - B(q)$$

with a (nonlinear) elliptic operator $\bar{A}$ and a (usually linear) control operator B. On the other hand, for 'control by the coefficients' this simple splitting will not suffice. For computational purposes it is more convenient to rewrite this equation in a variational form, with a suitable trial space V to be specified later on. It reads as follows

$$a(q,u)(\varphi) = (f,\varphi) \quad \forall\, \varphi \in V.$$

The target of the optimization is to minimize a given cost functional $J(q,u)$, e.g., a tracking type functional

$$J(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{r}\|q\|_{L^r}^r$$

for some $r \geq 2$. The emphasis of this thesis is the consideration of additional constraints on the control and state variable. Therefore let Q^{ad} be a closed convex set and g a functional. Then, we require the control and state variable to fulfill

$$q \in Q^{\mathrm{ad}}, \quad g(u, \nabla u) \leq 0.$$

A typical choice of Q^{ad} are 'box-constraints', e.g., let $a < b \in \hat{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ be given. We set

$$Q^{\mathrm{ad}} = \{\, q \mid a \leq q(x) \leq b \text{ a.e.} \,\}.$$

As state constraints we will consider zero-order state constraints, e.g.,

$$g(u, \nabla u) = g(u) = u - \psi$$

or first-order state constraints, e.g.,

$$g(u, \nabla u) = g(\nabla u) = |\nabla u|^2 - \psi.$$

To summarize: The general problem under consideration is given as

$$\text{Minimize } J(q,u)$$
$$\text{subject to } \begin{cases} A(q,u) = f, \\ q \in Q^{\mathrm{ad}}, \\ g(u,\nabla u) \leq 0. \end{cases}$$

In order to understand the problems being addressed here, we remark that in many cases the constraint $g(u,\nabla u) \leq 0$ is formulated as a pointwise inequality in a space of continuous functions. This leads to several problems which will be discussed in this thesis.

First of all, it is necessary that $g(u,\nabla u)$ lies in a space for which the inequality is meaningful. Hence the Nemytskii operator g should map $(u,\nabla u)$ onto a continuous function. This is a problem, particularly if one considers problems on polygonal or polyhedral domains. In addition, to obtain convergence rates for the (necessary) discretization, it is useful to know the regularity of the solutions that are approximated. Secondly, if we consider the inequality in a space of continuous functions, the Lagrange multiplier associated to $g(u,\nabla u) \leq 0$ is expected to be a measure. As we do not consider it sensible to discretize a space of measures, we will consider regularizations of the general problem. Thirdly, it is natural to ask what the error coming from the discretization and, if present, the regularization is. Especially, we have to ask, whether the sequence of discretized and regularized problems converges towards the solution at all. This leads to a priori convergence estimates. Albeit recent publications have been paying a lot of attention to control and zero-order state constraints, the case of first-order state constraints has hardly been tackled.

Finally, we will come to the question what the 'best possible' or 'most efficient' choice for the discretization and regularization is. This will lead to a posteriori error estimation with respect to the 'goal' of the computation.

The aim of this work is manyfold, as can be seen from the questions above. We start by discussing existence and regularity in the context of first-order constraints for a model problem. Then we will derive convergence estimates for the discretization of the state and control variable. We proceed to consider the regularized problems in function spaces, show existence, necessary conditions, and convergence towards the solution of the non-regularized problem. Finally, we will derive a posteriori error estimates with respect to given goal functionals. The estimates will be separated, so that we are able to balance the contributions to the global error arising from regularization and discretization. Naturally, all results will be substantiated by numerical examples.

In what follows, we will summarize the contents of this thesis.

We will start by recalling some well known results in Chapter 2 to fix our notation and to precisely state the problem class under consideration. After that we will continue by answering the questions above.

Existence, Uniqueness, Regularity In Chapter 3 we will discuss two fundamental questions for all computations.

First we consider the problem of first-order state constraints on non-smooth domains, e.g., domains that do not posses a $C^{1,1}$ boundary. The major problem in this context is that the 'solution operator' which maps control to state variables does not map an arbitrary 'regular' control onto a state which is sufficiently regular to impose the first-order constraint.

Despite this, typical applications require domains that have polygonal boundaries. In order to show that there is still a unique solution, which we can compute using the 'usual' first-order conditions, we consider a simple elliptic model PDE, e.g., $A(q,u) = -\Delta u - q$ and $g(u, \nabla u) = |\nabla u|^2 - \psi$.

For this we show that given a bounded polygonal domain $\Omega \subset \mathbb{R}^2$ and a Slater point, there exists a unique solution $\overline{u}$ to the model problem. In addition, the solution satisfies the regularity $\overline{u} \in W^{2,t} \cap H_0^1$ for some $t > 2$.

One immediately obtains that there exists $\overline{\mu} \in C(\overline{\Omega})^*$ and a function $\overline{z} \in L^{t'}(\Omega)$ such that the following holds

$$\begin{aligned}
(\nabla \overline{u}, \nabla \varphi) &= (\overline{q}, \varphi) && \forall\ \varphi \in H_0^1,\\
\langle -\Delta \varphi\,, \overline{z}\,\rangle_{Z^* \times Z} &= J_u'(\overline{q}, \overline{u})(\varphi) + \langle\, \overline{\mu}\,, 2\nabla \overline{u}\, \nabla \varphi\,\rangle_{C^* \times C} && \forall\ \varphi \in W^{2,t} \cap H_0^1,\\
J_q'(\overline{q})(\delta q - \overline{q}) &\geq -(\delta q - \overline{q}, \overline{z}) && \forall\, \delta q \in Q^{\mathrm{ad}} \cap I,\\
\langle\, \overline{\mu}\,, \varphi\,\rangle_{C^* \times C} &\leq 0 && \forall\ \varphi \in C(\overline{\Omega}), \varphi \leq 0,\\
\langle\, \overline{\mu}\,, |\nabla \overline{u}|^2 - \psi\,\rangle_{C^* \times C} &= 0.
\end{aligned}$$

Here I denotes the image of $W^{2,t} \cap H_0^1$ under Δ. This is almost the same regularity condition present in the case of a smooth boundary, see, e.g., (Casas and Bonnans [35], Casas and Fernández [36]), however, it differs due to the fact that one has to exclude a certain finite dimensional subspace corresponding to singular solutions of the state equation. We consider the problem where $\Omega \subset \mathbb{R}^3$, too, but there appears to be a gap, meaning there exist certain polyhedral domains, where a solution exists, but necessary conditions could not be derived by the method proposed in this thesis.

After having shown that the existence results in the literature extend to the case of non-smooth domains, we will consider the second question relevant for numerical approximation. Namely the question of regularity of the optimal solution. We will return to the case of a smooth boundary and show that the adjoint state $\overline{z}$ is not only in the space $L^{r'}$ but actually in the space $W^{1-2/r-\varepsilon,r'}$ for any $\varepsilon > 0$. From this we can finally deduce that the optimal control satisfies

$$\overline{q} \in W^{\frac{1}{r-1}(1-2/r-\varepsilon),r}(\Omega).$$

A priori error estimates Having derived regularity for the optimal solution, we turn to the error introduced by a discretization of problems with first-order state constraints in Chapter 4. We discretize the state using continuous, piecewise linear, bilinear, or trilinear test functions.

The control is discretized using either discontinuous piecewise constant functions or using the same finite elements used for the state variable to obtain (for piecewise constants)

$$\begin{aligned}\|\overline{u} - \overline{u}_h\|_{H^1} &\leq ch^{\frac{1}{r}(1-2/r)},\\ \|\overline{u} - \overline{u}_h\|_{L^2} &\leq ch^{\frac{1}{2}(1-2/r)},\\ \|\overline{q} - \overline{q}_h\|_{L^r} &\leq ch^{\frac{1}{r}(1-2/r)}.\end{aligned}$$

It should be noted that these estimates are suboptimal concerning possible best approximation of both state and control, however, our numerical evidence shows that the estimate for the control is sharp.

Regularization As it is computationally expensive, and not always possible, to solve the discrete state constraint problem, it appears to be mandatory to use 'regularization' techniques in order to approximate the solution sufficiently well, while still retaining a solvable problem. In Chapter 5 we will consider methods to regularize the problem under consideration. For zero-order state constraints there are three methods lively discussed in literature, namely quadratic penalty methods, also referred to as 'Moreau-Yosida' regularization, secondly barrier methods, and thirdly 'Lavrentiev' regularization.

As in the beginning of writing this thesis none of these methods had been discussed for first-order state constraints, we will study barrier methods, in great detail, in a function space setting. This appears natural, since this is what is needed to have that 'regularize-discretize' yields the same solutions as 'discretize-regularize'.

We first show that, under some standard assumptions, the barrier problem has a unique solution $(\overline{q}_\gamma, \overline{u}_\gamma)$ for every sufficiently small barrier parameter $\mu = 1/\gamma$. Then we derive necessary optimality conditions. Finally, we obtain that the barrier solution converges towards the solution $(\overline{q}, \overline{u})$ of the state constraint problem with the following rates

$$\begin{aligned}\|\overline{u} - \overline{u}_\gamma\|_{W^{2,r}} &\leq c\mu^{\frac{1}{r}},\\ \|\overline{u} - \overline{u}_\gamma\|_{L^2} &\leq c\mu^{\frac{1}{2}},\\ \|\overline{q} - \overline{q}_\gamma\|_{L^r} &\leq c\mu^{\frac{1}{r}}.\end{aligned}$$

A posteriori error estimates In Chapter 6 we will finally consider a posteriori error estimates for optimization problems with state constraints. With the previous results at hand, it is natural to ask the question: 'What is *the best* coupling between the parameters γ and h?'. This is directly related to the question of estimating the error introduced by both parameters. Once both errors can be estimated it is reasonable to choose both parameters in such a way, that the errors coming from both parameters are of the same size, in order to have minimal computational effort.

First of all, we note that the regularized problem is an optimization problem without inequality constraints on the state variable. Hence we derive a posteriori error estimates for the discretization error in the case of control constraints. Then the only question remaining is that of how to estimate the regularization error. To answer this, we will discuss the estimation

of the error in the cost functional. Here we will see that the error is approximately given by an integral of the state constraint mapping with an appropriately defined Lagrange multiplier μ_γ, e.g.,

$$|J(q,u) - J(q_\gamma, u_\gamma)| \approx \frac{1}{2}\Big(g(u_\gamma, \nabla u_\gamma), \mu_\gamma\Big)$$

for the case of a barrier method without control constraints. We will also consider the case of control and state constraints, as well as error estimates for the penalty approach.

Algorithmic details We proceed in Chapter 7 by describing the solution algorithm used for the computation of the considered examples. In Section 7.1 we will describe the active-set method used to solve the control constraint problems that arise when solving the regularized discrete problems. In Section 7.2 we will describe our overall solution algorithm for the state constraint problems and give some details on its implementation.

Applications and Outlook Finally, we will summarize the results and give an indication on remaining questions and possible further development.

2 Foundations

The main purpose of this chapter is to fix the notation used in this thesis. We will start by briefly recalling some basics from convex analysis in Section 2.1. Then we continue and state the optimization problems under consideration in a precise framework in Section 2.2. This will be done in such a way, that we can apply standard results from optimization theory to obtain existence of solutions and optimality conditions. Hence there is nothing surprising to anyone familiar with the the monographs of (Clarke [47], Ekeland and Témam [59], Fursikov [63], Ito and Kunisch [93], Lions [100], Tröltzsch [141]) except for the specific choice of the spaces in order to apply the general theory to our situation and the explicit treatment of first-order state constraints. As it is not the aim of this thesis to summarize the most general results concerning existence and first and second-order conditions, we will make some convenient assumptions that allow for a short survey on these results.

2.1 Basic Notation

In all that follows, let $\Omega \subset \mathbb{R}^n$ be an open bounded domain with Lipschitz boundary $\Gamma := \partial\Omega$, see, e.g. (Adams and Fournier [1], Chapter 4).

We adopt standard notation, see (Adams and Fournier [1], Wloka [155]) for the usual Lebesgue and Sobolev spaces, e.g., $W^{m,p}(\Omega)$ denotes the space of functions whose weak derivatives up to order m are in $L^p(\Omega)$. Sometimes we will require non-integer differentiability, in this case we will write $W^{s,p}(\Omega)$ instead. These spaces can either be defined by interpolation using Besov spaces, see, e.g., (Triebel [140]), or equivalently using completion with respect to norms of certain difference quotients ($p \neq \infty$).

The space $L^2(\Omega)$ is a Hilbert space, its scalar product is denoted by $(\,\cdot\,,\cdot\,)$ and the corresponding norm by $\|\cdot\|$. All other scalar products and norms will be given an appropriate index, e.g., $(\,\cdot\,,\cdot\,)_V$ for the scalar product on a Hilbert space V. As it is common, we write $H^m(\Omega)$ for $W^{m,2}(\Omega)$. If we are concerned with vector valued function spaces, we indicate this by adding the image space to the definition, e.g., $C(\overline{\Omega};\mathbb{R}^d)$ for continuous functions on $\overline{\Omega}$ with values in $\mathbb{R}^d$.

Let B be a real Banach space. We denote its topological dual by B^* and the duality pairing by $\langle\,\cdot\,,\cdot\,\rangle_{B^*\times B}$.

Let C be a convex subset of B. We define the *dual* or *polar cone* by

$$C^+ := \{\, b^* \in B^* \mid \langle\, b^*\,, c\,\rangle_{B^*\times B} \leq 0 \;\forall\, c \in C \,\}\,. \tag{2.1}$$

In addition for $b \in B$, we define the *conical hull* of $C \setminus \{b\}$ by

$$C(b) := \{\, a(c-b) \mid c \in C, a \geq 0, a \in \mathbb{R} \,\}. \tag{2.2}$$

Let B_1 and B_2 be Banach spaces and let $F\colon B_1 \to B_2;\ b \mapsto F(b)$ be directional differentiable, then we denote its directional derivative at $b \in B_1$ in direction $\delta b \in B_1$ by

$$F'_b(b)(\delta b) := \lim_{\varepsilon \to 0} \frac{F(b+\varepsilon\delta b) - F(b)}{\varepsilon} \in B_2.$$

Further if F is Gâteaux differentiable, we denote the corresponding Gâteaux-derivative by

$$F'_b(b) \in \mathcal{L}(B_1, B_2)$$

where $\mathcal{L}(B_1, B_2)$ denotes, as usual, the space of bounded linear operators from B_1 into B_2. Finally, if $t \geq 1$ is a given number we set $t' = \frac{t}{t-1}$ with the usual meaning if $t = 1$ or $t = \infty$.

2.2 Abstract Optimization Problem

Let V be a Hilbert space, such that $V \subset L^2(\Omega)$ with a dense compact embedding, e.g., $V = H^1(\Omega)$. Let further W, Q be reflexive Banach spaces with $W \subset V$ dense and continuous. Further let $Q^{\mathrm{ad}} \subset Q$ be closed, convex, and non-empty.

For the abstract differential operator $A\colon Q^{\mathrm{ad}} \times V \to V^*$, we define its variational form by setting

$$a(q,u)(\varphi) := \langle\, A(q,u)\,, \varphi \,\rangle_{V^* \times V}. \tag{2.3}$$

We can now state our partial differential equation. For given $f \in V^*$, $q \in Q^{\mathrm{ad}}$ find $u \in V$ such that

$$a(q,u)(\varphi) = (f, \varphi) \quad \forall\, \varphi \in V. \tag{2.4}$$

In the following we assume that (2.4) has a unique solution u_q, such that the mapping $(q, f) \mapsto u_q$ is at least once continuously differentiable.

Remark 2.1. It's clear that non-variational boundary conditions such as non-homogeneous Dirichlet values for u can not be obtained in this fashion. However, we don't want to complicate the notation by searching u in an affine space $V + \hat{u}$ since the modifications are straightforward.

As it is in general not suitable to consider (2.4) under minimal regularity assumptions, we will make the following assumption on the regularity of the differential operator. Assume that there is a subspace $Z^* \subset V^*$ such that for $q \in Q^{\mathrm{ad}}$, $f \in Z^*$ any solution u_q of (2.4) satisfies the additional regularity $u_q \in W$.

We continue by defining the state constraints. For this reason, we define the compact set $\Omega^C \subset \overline{\Omega}$ with non empty interior. We assume that the mapping $g\colon u \mapsto g(u, \nabla u)$ is C^2 from W into $G := C(\Omega^C)$. We employ the usual ordering to state the constraint $g(u, \nabla u) \leq 0$. In order to shorten notation, we define the derivative of this mapping by

$$g'(u, \nabla u)(\delta u) := \frac{d}{d\varepsilon} g(u + \varepsilon\delta u, \nabla(u\ + \varepsilon\delta u))\Big|_{\varepsilon=0} = g'_u(u, \nabla u)(\delta u) + g'_{\nabla u}(u, \nabla u)(\nabla \delta u).$$

Remark 2.2. The case of finitely many state constraints can be handled in the same framework. It is obtained by setting $\Omega^C \subset \{1, \dots n\}$ for some n. In this case $G := C(\Omega^C) = \mathbb{R}^n$ and we identify its dual G^* with G.

In order to state necessary conditions, and to apply Newton's method it is convenient—although not necessary, see (Chen, Nashed, and Qi [44], Hintermüller, Ito, and Kunisch [85], Ulbrich [144])—to require that $A\colon Q^{\mathrm{ad}} \times V \to V^*$ and $A\colon Q^{\mathrm{ad}} \times W \to Z^*$ are of class C^2 and for any $q \in Q^{\mathrm{ad}}$ and $u \in W$ the partial derivative with respect to u is an isomorphism between the following pairs of spaces

$$\begin{aligned} A'_u(q,u) &: V \to V^*, \\ A'_u(q,u) &: W \to Z^*. \end{aligned}$$

Finally, we consider the cost functional $J\colon Q \times V \mapsto \mathbb{R}$ to be C^2, weakly lower semi-continuous, coercive, and bounded from below.

We can now state our abstract optimization problem as

$$\text{Minimize } J(q,u) \tag{2.5a}$$

$$\text{subject to } \begin{cases} a(q,u)(\varphi) = f(\varphi) \quad \forall\, \varphi \in V, \\ (q,u) \in Q^{\mathrm{ad}} \times V, \\ g(u, \nabla u) \leq 0. \end{cases} \tag{2.5b}$$

It should be noted that there are two ways to proceed, one is to consider the tuple (q,u) as two independent variables of the optimization problem (2.5), see, e.g., (Fursikov [63]). The other one, that we will pursue in the fashion of (Lions [100]) is to consider only q as a variable and associate u_q by means of (2.4). For a more recent survey on this see (Tröltzsch [141]). We define the *solution operator* S by

$$a(q, S(q))(\varphi) = (f, \varphi) \quad \forall\, \varphi \in V. \tag{2.6}$$

We assume differentiability of the control to state mapping $S\colon Q \to W$. Then we obtain that the derivative $S'(q)(\delta q)$ exists, and satisfies

$$\langle\, A'_u(q, S(q)) S'(q)(\delta q)\,, \varphi \,\rangle_{Z^*,Z} = -\langle\, A'_q(q, S(q))\delta q\,, \varphi\,\rangle_{Z^*,Z} \quad \forall\, \varphi \in Z. \tag{2.7}$$

Now we can state the reduced optimization problem by

$$\text{Minimize } \hat{J}(q) := J(q, S(q)) \tag{2.8a}$$

$$\text{subject to } \begin{cases} \hat{g}(q) := g(S(q), \nabla S(q)) \leq 0, \\ q \in Q^{\mathrm{ad}}. \end{cases} \tag{2.8b}$$

Finally, we note that the last condition can be equivalently expressed as

$$\hat{g}(q) \in G^{\mathrm{ad}},$$

where $G^{\mathrm{ad}} := \{ v \in G \mid v \leq 0 \}$ is the cone of non positive functions. We assume here, that the mapping $\hat{g}$ is weakly sequential continuous.

For the analysis of the reduced problem (2.8) it is convenient to assume that the reduced cost functional $\hat{J}$ is coercive and weakly lower semicontinuous.

Remark 2.3. Note that this is not implied by the assumptions on the cost functional J on $Q \times V$, they are required for simplicity only, e.g., if Q^{ad} is bounded there is no need for growth conditions on the cost functional.

Existence There are several theorems available that ensure existence of solutions to minimization problems. However if applied to PDEs one has to carefully check whether all premises are fulfilled. For a discussion in the case of semi-linear elliptic equations with control constraints see (Tröltzsch [141], Chapter 4), in the case of state constraints see, e.g., (Casas and Bonnans [35]) and (Casas and Fernández [36]) for first-order state constraints. For quasi linear equation with control constraints see (Casas and Tröltzsch [41]).

Let us assume that (2.8) has at least one feasible point, that is, we assume that the set

$$Q^{\mathrm{feas}} := \{ \delta q \in Q^{\mathrm{ad}} \mid \hat{g}(\delta q) \in G^{\mathrm{ad}} \}$$

is not empty.

Now using standard arguments, cf., (Dacorogna [49]), it is simple to show existence of an optimal solution.

Theorem 2.1. *Let $Q^{feas} \neq \emptyset$, then there exists a solution to* (2.8).

Proof. Take a minimizing sequence $q_k \in Q^{\mathrm{feas}}$. By coercivity of $\hat{J}$ the sequence is bounded and hence, possibly taking a subsequence, q_k is weakly convergent to some $q \in Q^{\mathrm{ad}}$, as Q^{ad} is weakly sequentially closed. Further the limit q lies in Q^{feas} as $\hat{g}$ is weakly sequential continuous. Finally, by lower semi continuity of $\hat{J}$, we obtain that $\hat{J}(q) \leq \min_{\delta q \in Q^{\mathrm{feas}}} \hat{J}(\delta q)$. □

Necessary Conditions In order to apply standard calculus it is convenient to assume that a solution $\overline{q}$ to problem (2.8) satisfies a regularity condition (sometimes called *Slater condition* or *constraint qualification*) namely that the solution is a regular point.

Definition 2.1. (Regular point) An element $q \in Q^{\mathrm{feas}} := \{ \delta q \in Q^{\mathrm{ad}} \mid \hat{g}(\delta q) \in G^{\mathrm{ad}} \}$ is called regular if

$$\hat{g}'(q)Q^{\mathrm{ad}}(q) - G^{\mathrm{ad}}(\hat{g}(q)) = G. \tag{2.9}$$

This definition goes back to (Mäurer and Zowe [105], Zowe and Kurcyusz [159]).

Remark 2.4. This definition is equivalent to

$$0 \in \operatorname{int} \{ \hat{g}'(q)(Q^{\mathrm{ad}} - q) - G^{\mathrm{ad}} + \hat{g}(q) \}$$

where int denotes the interior of the set see, e.g., (Ito and Kunisch [93]) for a proof.

From Definition 2.1 one readily deduces the existence of a Lagrange multiplier.

Theorem 2.2. *Let $\overline{q}$ be a solution to* (2.8)*. Further, let $\overline{q}$ be regular, then there exists a Lagrange multiplier $\overline{\mu} \in G^*$. This means, that*

$$\begin{aligned} \hat{J}'(\overline{q})(\varphi - \overline{q}) + \langle\, \overline{\mu}\,, \hat{g}'(\overline{q})(\varphi - \overline{q})\,\rangle_{G^*\times G} \geq 0 \;\; \forall\, \varphi \in Q^{ad}, \quad & (e.g.\;\; \hat{J}'(\overline{q}) + \overline{\mu} \circ \hat{g}'(\overline{q}) \in -Q^{ad}(\overline{q})) \\ \langle\, \overline{\mu}\,, \varphi\,\rangle_{G^*\times G} \leq 0 \;\; \forall\, \varphi \in G^{ad}, \quad & (e.g.\;\; \overline{\mu} \in G^{ad^+}) \\ \langle\, \overline{\mu}\,, \hat{g}(\overline{q})\,\rangle_{G^*\times G} = 0. & \end{aligned}$$

For a proof see (Ito and Kunisch [93], Theorem 1.6). Now defining $\overline{z} \in Z$ as a solution to

$$\langle\, A'_u(\overline{q},\overline{u})\varphi\,, \overline{z}\,\rangle_{Z^*\times Z} = J'_u(\overline{q},\overline{u})(\varphi) + \langle\, \overline{\mu}\,, g'(\overline{u}, \nabla\overline{u})(\varphi)\,\rangle_{G^*\times G} \quad \forall \varphi \in W$$

(which exists due to our assumptions on $A'_u(\overline{q}, u)$) we obtain using (2.7)

Theorem 2.3. *Let $\overline{q} \in Q^{ad}$ be a solution to* (2.8)*—or* (2.5)*—. Further let $\overline{q}$ be regular in the sense of Definition 2.1, then there exists $u \in W$, $z \in Z$, $\mu \in M(\Omega^C)$ such that:*

$$a(\overline{q},\overline{u})(\varphi) = f(\varphi) \qquad \forall\;\; \varphi \in V, \tag{2.10a}$$

$$\langle\, A'_u(\overline{q},\overline{u})\varphi\,, z\,\rangle_{Z^*\times Z} = J'_u(\overline{q},\overline{u})(\varphi) + \langle\, \overline{\mu}\,, g'(\overline{u},\nabla\overline{u})(\varphi)\,\rangle_{G^*\times G} \quad \forall\;\; \varphi \in W, \tag{2.10b}$$

$$J'_q(\overline{q},\overline{u})(\delta q - \overline{q}) \geq \langle\, A'_q(\overline{q},\overline{u})(\delta q - \overline{q})\,, z\,\rangle_{Z^*\times Z} \qquad \forall\, \delta q \in Q^{ad}, \tag{2.10c}$$

$$\langle\, \overline{\mu}\,, \varphi\,\rangle_{G^*\times G} \leq 0 \qquad \forall\;\; \varphi \in G;\; \varphi \leq 0, \tag{2.10d}$$

$$\langle\, \overline{\mu}\,, \hat{g}(\overline{q})\,\rangle_{G^*\times G} = 0. \tag{2.10e}$$

This is the form that can be found in several publications. For instance (Bergounioux [17], Casas [34], Casas and Bonnans [35]) showed that this is true for state constraints with distributed control or (Bergounioux [18]) for boundary control. For first-order state constraints the existence of the multipliers can be found in (Casas and Fernández [36]). Recently similar results were obtained in the case of state constraints with discontinuous states by (Schiela [131]).

Naturally there are also necessary second-order conditions possible, see, e.g, (Casas and Tröltzsch [39]).

One should note that for mixed control-state constraints, e.g., g depends also on the control variable, it can be shown that the multiplier μ in Theorem 2.3 is actually in L^2, see (Tröltzsch [142]). Also in the case of state constraints the multiplier μ is not an arbitrary measure, but can in fact be split into a measure part on the boundary of the active set, and a L^2 function in the interior of the active set, see (Bergounioux and Kunisch [21]).

Sufficient Conditions In general the necessary conditions from the previous section are not sufficient. Hence one needs some sufficient conditions. As there can not be a local condition to ensure that a point q is a global minimizer these conditions usually only ensure that the point q is a local minimizer. The main idea is to show that the problem at hand is convex in a suitable neighborhood of a point q that satisfies the necessary condition (2.10). As the

necessary conditions are sufficient for convex problems one immediately obtains that q is a local minimizer in the given neighborhood.

In order to state this conditions in a compact form we define the Lagrangian

$$\begin{aligned}\mathcal{L}\colon Q \times W \times Z &\to \mathbb{R},\\ (q,u,z) &\mapsto J(q,u) + f(z) - \langle\, A(q,u)\,, z \,\rangle_{Z^*\times Z}.\end{aligned}$$

Then the following holds

Theorem 2.4 (Sufficient optimality condition)**.** *Let $\overline{\xi} = (\overline{q},\overline{u},\overline{z},\overline{\mu}) \in Q^{ad} \times W \times Z \times G^*$ satisfy the first-order necessary condition* (2.10) *of optimization problem* (2.5)*. Moreover, let $z \mapsto \langle\, A_u'(\overline{q},\overline{u})(\cdot)\,, z\,\rangle_{Z^*\times Z} : Z \to W^*$ be surjective. If there exists $\rho > 0$ such that*

$$(\delta q\,,\, \delta u) \begin{bmatrix} \mathcal{L}_{qq}''(\overline{\xi})(\cdot,\cdot) & \mathcal{L}_{qu}''(\overline{\xi})(\cdot,\cdot) \\ \mathcal{L}_{uq}''(\overline{\xi})(\cdot,\cdot) & \mathcal{L}_{uu}''(\overline{\xi})(\cdot,\cdot) \end{bmatrix} \begin{pmatrix} \delta q \\ \delta u \end{pmatrix} \geq \rho \left(\|\delta u\|_W^2 + \|\delta q\|_Q^2 \right) \tag{2.11}$$

holds for all $(\delta q, \delta u)$ satisfying the linear (tangent) PDE (2.7) *then $(\overline{q},\overline{u})$ is a (strict) local solution to the optimization problem* (2.5)*.*

We refer to (Mäurer and Zowe [105]) for the proof.

Such conditions were derived for problems with semilinear elliptic equations subject to state or mixed control-state constraints, see, e.g., (Casas and Mateos [37], Casas, Tröltzsch, and Unger [42], Meyer and Tröltzsch [114]).

In general Fréchet differentiability and coercivity of the second derivative can not be shown in the same spaces, in this case one has to deal with the well known 'two-norm' discrepancy, see, e.g, (Tröltzsch [141]).

2.3 Discretization of the State Constraint

As we will later use finite elements for the discretization of the equation (2.4) we need some remarks concerning the state constraint g. Let V_h be a finite element space, for the precise definition of which we refer to Section 4.2. In order to work in a convenient framework we assume that the 'same' operator g also defines a mapping $g\colon V_h \to L^\infty(\Omega^C)$.

We will come back to this in Chapter 6, where it will be important that the solution to (2.5) remains the same regardless of whether we consider the constraint $g(u, \nabla u)) \leq 0$ with respect to the ordering in $C(\Omega^C)$ or $L^\infty(\Omega^C)$.

2.4 Examples

Example with First Order Constraints Let $\Omega \subset \mathbb{R}^2$ be a bounded convex polygonal domain. We consider the following model problem

$$\begin{aligned} \text{Minimize } J(q,u) &= \frac{1}{2}\| u - u^d \|^2 + \frac{\alpha}{r}\| q \|_{L^r}^r \\ \text{subject to } &\begin{cases} (\nabla u, \nabla\varphi) = (q,\varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ |\nabla u|^2 \le \psi & \text{in } \overline{\Omega}, \\ a \le q \le b & \text{a.e. in } \Omega. \end{cases} \end{aligned} \tag{2.12}$$

Where $\alpha, \psi > 0$, $-\infty \le a < b \le \infty$ are given numbers such that at least $r > 2$ or $a, b \in \mathbb{R}$.

It is well known, cf., (Grisvard [73]), that there is $t > n = 2$ such that $-\Delta$ is an isomorphism from $W^{2,t}(\Omega) \cap H_0^1(\Omega)$ into $L^t(\Omega)$. Hence setting $W = W^{2,t}(\Omega) \cap H_0^1(\Omega)$, we have the required regularity of the state equation and its derivatives with respect to u. The mapping $g(u, \nabla u) = |\nabla u|^2 - \psi$ is differentiable from $W \subset C^1(\overline{\Omega})$ into $C(\overline{\Omega})$.

Now, noting that the 'control-to-state' mapping $L^t(\Omega) \to C(\overline{\Omega})$, $q \mapsto u_q$ defined by

$$(\nabla u, \nabla\varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega)$$

is compact, we obtain that the reduced state constraint mapping $\hat{g}$ is weakly sequential continuous. Hence, by Theorem 2.1, there exists a unique solution $(\overline{q}, \overline{u})$ to this problem.

By definition of a regular point, see Remark 2.4, we obtain that the solution is regular, provided that

$$0 \in \operatorname{int}\{\, \hat{g}'(\overline{q})(Q^{\mathrm{ad}} - \overline{q}) - G^{\mathrm{ad}} + \hat{g}(\overline{q}) \,\}.$$

This means there exists ε such that for any $v \in C(\overline{\Omega})$ with $\|v\|_\infty \le \varepsilon$ there exists $(\delta q, \delta v) \in Q^{\mathrm{ad}} \times G^{\mathrm{ad}}$ such that

$$v = 2\nabla\overline{u} \cdot \nabla S(\delta q) - 2|\nabla\overline{u}|^2 - \delta v + |\nabla\overline{u}|^2 - \psi. \tag{2.13}$$

Now, assume that there is a Slater point, e.g., a point $\hat{q} \in Q^{\mathrm{ad}}$ such that

$$g(S\hat{q}, \nabla S\hat{q}) = |\nabla S\hat{q}|^2 - \psi < 0. \tag{2.14}$$

We obtain that

$$\begin{aligned} &2\nabla\overline{u} \cdot \nabla S\hat{q} - 2|\nabla\overline{u}|^2 + |\nabla\overline{u}|^2 - \psi \\ &\quad \le |\nabla\overline{u}|^2 + |\nabla S\hat{q}|^2 - 2|\nabla\overline{u}|^2 + |\nabla\overline{u}|^2 - \psi \\ &\quad \le |\nabla S\hat{q}|^2 - \psi < 0 \end{aligned}$$

and as $\overline{\Omega}$ is compact there is $\varepsilon > 0$ such that

$$2\nabla\overline{u} \cdot \nabla S\hat{q} - 2|\nabla\overline{u}|^2 + |\nabla\overline{u}|^2 - \psi \le -\varepsilon.$$

Hence we have that

$$0 < v - 2\nabla\overline{u} \cdot \nabla S\hat{q} + |\nabla\overline{u}|^2 + \psi$$

provided $\|v\|_\infty \leq \varepsilon$. Which gives that there is $\delta v \in G^{\mathrm{ad}}$ such that $(\hat{q}, \delta v)$ fulfills (2.13).

We see that if there is a Slater point (2.14) the solution is always regular, and Theorem 2.3 gives

Theorem 2.5. *Let $(\overline{q}, \overline{u}) \in L^r(\Omega) \times H_0^1(\Omega) \cap W^{2,t}(\Omega)$ be the solution to (2.12), with some $t > n$ and $n < r < \infty$ or $a, b \in \mathbb{R}$. Assume that there is a point $\hat{q}$ satisfying (2.13), then there exists $\overline{z} \in L^s(\Omega)$ for all $s < \frac{n}{n-1}$ and a measure $\overline{\mu}$ with support contained in $\overline{\Omega}$, such that*

$$
\begin{aligned}
(\nabla \overline{u}, \nabla \varphi) &= (\overline{q}, \varphi) && \forall\ \varphi \in H_0^1(\Omega), && \text{(2.15a)}\\
(\overline{z}, -\Delta \varphi) &= (\overline{u} - u^d, \varphi) + \langle \overline{\mu}, \nabla \varphi \nabla \overline{u} \rangle_{C^* \times C} && \forall\ \varphi \in H_0^1(\Omega) \cap W^{2,r}(\Omega), && \text{(2.15b)}\\
\langle \overline{\mu}, \varphi \nabla \overline{u} \rangle_{C^* \times C} &\leq \langle \overline{\mu}, |\nabla \overline{u}|^2 \rangle_{C^* \times C} && \forall\ \varphi \in C(\overline{\Omega}, \mathbb{R}^d),\ |\varphi| \leq \psi, && \text{(2.15c)}\\
\alpha(|\overline{q}|^{r-2}\overline{q}, \delta q - \overline{q}) &\geq (-\overline{z}, \delta q - \overline{q}) && \forall\, \delta q \in Q^{ad}. && \text{(2.15d)}
\end{aligned}
$$

This is not exactly the form stated in Theorem 2.3, but can be derived from this, see (Casas and Fernández [36]).

Example with Zero Order Constraints

$$
\text{Minimize } J(q,u) = \frac{1}{2}\| u - u^d \|^2 + \frac{\alpha}{2}\| q \|^2
$$
$$
\text{subject to } \begin{cases} (\nabla u, \nabla \varphi) + (u, \varphi) = (q, \varphi)_{\partial\Omega} & \forall\, \varphi \in H^1(\Omega), \\ u \leq \psi & \text{in } \overline{\Omega}, \\ a \leq q \leq b & \text{a.e. in } \Omega. \end{cases} \tag{2.16}
$$

Where $\alpha, \psi > 0$, $-\infty \leq a < b \leq \infty$ are given numbers. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. If $\partial\Omega$ is smooth it is clear that for any $q \in L^2(\partial\Omega)$ the solution to

$$
(\nabla u, \nabla \varphi) + (u, \varphi) = (q, \varphi)_{\partial\Omega} \quad \forall\, \varphi \in H^1(\Omega)
$$

is contained in $H^{3/2}(\Omega)$, see, e.g., (Lions and Magenes [101], Section 7.3). Hence we obtain from Sobolev embedding that $u \in W^{1,t}(\Omega) \subset C^0(\overline{\Omega})$ for some $t > 2$.

Hence by setting $V = H^1(\Omega)$, $W = W^{1,t}(\Omega)$, we obtain that the mapping $g(u, \nabla u) = u - \psi$ is continuous (and differentiable) from W into $C^0(\overline{\Omega})$. In fact it is also compact and hence the mapping $\hat{g}$ is weakly sequential continuous. By assumption $\psi > 0$ and hence $\hat{q} \equiv 0$ is a slater point, e.g., $\hat{u} = u_{\hat{q}} \equiv 0$ fulfills $g(\hat{u}, \nabla \hat{u}) < 0$. Following the same line of arguments as in the previous example, we obtain that every feasible point is regular, and hence the following theorem holds

Theorem 2.6. *Let $(\overline{q}, \overline{u}) \in L^2(\Omega) \times W_0^{1,t}(\Omega)$ for some $t > 2$ be the solution to (2.16). Then there exists $\overline{z} \in W^{1,t'}(\Omega)$ for $t' = \frac{t}{t-1}$ and a measure $\overline{\mu}$ with support contained in $\bar{\Omega}$, such that*

$$(\nabla \overline{u}, \nabla \varphi) = (\overline{q}, \varphi) \qquad \forall \ \ \varphi \in H_0^1(\Omega), \tag{2.17a}$$

$$(\nabla \overline{z}, \nabla \varphi) = (\overline{u} - u^d, \varphi) + \langle \overline{\mu}, \varphi \rangle_{C^* \times C} \qquad \forall \ \ \varphi \in W_0^{1,t}(\Omega), \tag{2.17b}$$

$$\alpha(\overline{q}, \delta q - \overline{q}) \geq (-\overline{z}, \delta q - \overline{q}) \qquad \forall\, \delta q \ \in Q^{ad}, \tag{2.17c}$$

$$\langle \overline{\mu}, \varphi \rangle_{C^* \times C} \leq 0 \qquad \forall \ \ \varphi \in C(\overline{\Omega}),\ \varphi \leq 0, \tag{2.17d}$$

$$\langle \overline{\mu}, g(\overline{u}) \rangle_{C^* \times C} = 0. \tag{2.17e}$$

3 Existence and Regularity

Throughout this chapter we will consider the most simple setting for first-order state constraints. We begin by describing two model problems, each of them having unique features. First we will consider the case of pure first-order state constraints. There we have to use a stronger regularization in order to assure existence of a solution. Secondly we will introduce additional bounds on the control variable.

Problem without Control Constraints

$$\begin{aligned} \text{Minimize } J(q,u) &= \frac{1}{2}\| u - u^d \|^2 + \frac{\alpha}{r}\| q \|_{L^r}^r \\ \text{subject to } &\begin{cases} (\nabla u, \nabla \varphi) = (q, \varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ |\nabla u|^2 \leq \psi & \text{a.e. in } \overline{\Omega}. \end{cases} \end{aligned} \tag{3.1}$$

Where $\alpha, \psi > 0$, $r > n$ are given numbers.

Problem with Control Constraints

$$\begin{aligned} \text{Minimize } J(q,u) &= \frac{1}{2}\| u - u^d \|^2 + \frac{\alpha}{2}\| q \|^2 \\ \text{subject to } &\begin{cases} (\nabla u, \nabla \varphi) = (q, \varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ |\nabla u|^2 \leq \psi & \text{a.e. in } \overline{\Omega}, \\ a \leq q \leq b & \text{a.e. in } \Omega. \end{cases} \end{aligned} \tag{3.2}$$

Where $a, b \in \mathbb{R}$, $a < b$ and $\alpha, \psi > 0$ are given numbers..

Both examples fit into the framework of Chapter 2, see Section 2.4 for details.

3.1 Results on Non-Smooth Domains

As already stated in Section 2.4, some smoothness requirements on the domain are sufficient to obtain that there exists a (unique) solution and additional Lagrange multiplier. We will now consider the, more realistic case, that the domain is not smooth, but is bounded by finitely many smooth $(n-1)$-dimensional manifolds.

This will lead to some difficulty because the control-to-state mapping is no longer a map from L^r into $W^{2,r}$ due to the singularities arising from corners and edges. To the authors knowledge this setting has not been considered in the literature prior to this thesis.

3.1.1 Existence

We start the discussion by showing that both problems (3.1) and (3.2) posses a solution.

Theorem 3.1. *Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ be a polygonal or polyhedral domain. Then problem* (3.1) *has a unique solution $(\overline{q}, \overline{u}) \in L^r(\Omega) \times W^{2,t}(\Omega) \cap H_0^1(\Omega)$ for some $t > 2$ depending only on the angles in the corners and the edges of the domain.*

If in addition (3.2) *has at least one feasible control, e.g., such that the control is in Q^{ad} and the corresponding state satisfy the inequality constraints, then* (3.2) *has a unique solution $(q, u) \in L^\infty(\Omega) \times W^{2,t}(\Omega) \cap H_0^1(\Omega)$. Where $t > 2$ depends only on the angles in the corners and the edges of the domain.*

Before we are able to proof Theorem 3.1 we will require a short lemma.

Lemma 3.2. *Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ be a polygonal or polyhedral domain. Further, let $f \in L^p(\Omega)$ for some $p \geq 2$. If the solution u of*

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega)$$

satisfies $u \in W^{1,\infty}(\Omega)$ then $u \in W^{2,t}(\Omega)$ for some $t \in [2, p]$. Moreover if $p > 2$ then $t \in (2, p]$ is possible. The value of t can be determined by knowledge of the angles in the corners, edges and vertices of the domain.

Proof of Lemma 3.2. The proof is based on well known singular behavior of the solution near the corners and edges, cf., (Grisvard [71]) for the 2d case. The 3d case was considered in (Dauge [50], Grisvard [73]) in Hilbert spaces, its extension to the non-hilbertian case can be found in (Grisvard [74]). The idea of the proof is as follows, the solution to the state equation can be split into a regular part that exhibits the regularity introduced by the right-hand side f and a singular part corresponding to the non-smooth boundary. By the bound on the gradient of the solution one obtains, that the singular part may not exist.

We begin with a discussion in 2d, e.g., $n = 2$. Let $\mathcal{C}$ be the (finite) set of corners of the domain. For a corner $c \in \mathcal{C}$ we denote the interior angle by ω_c, and we denote polar coordinates with respect to the corner c by (ρ_c, θ_c). Then assuming that $f \in L^t$ for some $t \in (2, p]$ and $\frac{2\omega_c}{\pi t'} \notin \mathbb{N}$ there exist numbers $C_{c,j}$ such that the solution u to

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega)$$

satisfies

$$u - \sum_{c \in \mathcal{C}} \sum_{\substack{j=1 \\ \frac{j\pi}{\omega_c} \neq 1}}^{j < \frac{2\omega_c}{\pi t'}} C_{c,j} \eta_c(\rho_c) \rho_c^{\frac{j\pi}{\omega_c}} \sin\left(\frac{j\pi}{\omega_c}\theta_c\right) \in W^{2,t}(\Omega)$$

with suitable cutoff functions η_c, cf., (Grisvard [71], Theorem 4.4.3.7).

To proceed further, note that if $t > 2 = n$, e.g., $W^{2,t}(\Omega) \subset C^1(\overline{\Omega})$, we have that $t' < 2$, and $t' \to 2$ as $t \to 2$. Let us assume first that $\omega_c \neq 2\pi$, then by considering $t > 2$ small enough $\frac{2\omega_c}{\pi t'} < 2$ and the second sum in the singular expansion contains at most the value $j = 1$. If

$\omega_c = 2\pi$, then $t > 2$ small enough implies $\frac{2\omega_c}{\pi t'} < 3$, and as the case $j = 2$ is prohibited by the condition $\frac{j\pi}{\omega_c} \neq 1$ we obtain again that the second sum in the singular expansion contains at most the value $j = 1$.

Now we will discuss the behavior of the derivative of the singular solutions. We obtain for all $c \in \mathcal{C}$ that $\frac{j\pi}{\omega_c} < \frac{2}{t'}$. Here we have to distinguish two cases:

First assume that $\frac{j\pi}{\omega_c} < 1$ then the first derivative of $\rho_c^{\frac{j\pi}{\omega_c}}$ becomes unbounded and the assumption $u \in W^{1,\infty}(\Omega)$ implies $C_{c,j} = 0$.

Second if $\frac{j\pi}{\omega_c} > 1$, then by reducing $t > 2$ even further, we obtain that $\frac{2}{t'} < \frac{j\pi}{\omega_c}$, and hence this case doesn't exist anymore.

It is clear the the same argument remains true if $t = t' = 2$.

Summing up we obtain that, provided t is sufficiently small,

$$\sum_{c\in\mathcal{C}} \sum_{j=1}^{j<\frac{2\omega_c}{\pi t'}} C_{c,j}\eta_c(\rho_c)\rho_c^{\frac{j\pi}{\omega_c}} \sin\left(\frac{j\pi}{\omega_c}\theta_c\right) \in W^{1,\infty}(\Omega)$$

if and only if all the singular coefficients fulfill $C_{c,j} = 0$ and hence that $u \in W^{2,t}(\Omega)$ for some $t \in [2, p]$ sufficiently small, and if $p > 2$ we can actually choose $t > 2$.

We now turn our attention to the case $n = 3$. Here we will have to consider contributions by vertices and edges. Therefore we denote the set of vertices on $\partial\Omega$ by $\mathcal{V}$ and the set of edges by $\mathcal{E}$.

We will begin by considering a vertex $v \in \mathcal{V}$. Here we introduce spherical coordinates $(\rho_v, \theta_v, \varphi_v)$. Let now be B_v a sufficiently small ball around v, let $G_v = \partial B_v \cap \Omega$ then let $w_{j,v}(\theta_v, \varphi_v)$ be the sequence of eigenfunctions of the Laplace-Beltrami operator with homogeneous Dirichlet boundary conditions on G_v and $\lambda_{j,v}$ be the corresponding eigenvalues. Then, assuming $\lambda_{j,v} \neq \left(\frac{3}{t} - 2\right)\left(\frac{3}{t} - 3\right)$ for all j, the corresponding singular expansion reads

$$\sum_{\lambda_{j,v}<\left(\frac{3}{t}-2\right)\left(\frac{3}{t}-3\right)} C_{j,v}\eta_v(\rho)\rho_v^{\beta_{j,v}-\frac{1}{2}} w_{j,v}(\theta_v, \varphi_v).$$

Where $\beta_{j,v}$ is given as $\beta_{j,v} = \sqrt{\left(\frac{3}{t} - 1\right)^2 + \lambda_{j,v}}$, see (Grisvard [74], Theorem 4.6). First let $p > 3$ then for $t > 3$ we obtain that for $t \to 3$ the upper bound on $\lambda_{j,v}$ converges to two. Hence $\beta_{j,v} \leq \sqrt{2} + \varepsilon < 1.5$ and we obtain that the first derivative is not bounded for t sufficiently small.

We remark that we can choose $t \in (3, p]$ independent of the angles, as the only requirement is $\beta_{j,v} < 1.5$ (Although one may obtain the same for larger t by using information on $\lambda_{j,v}$.).

If $p < 3$ we can use the same argument, as we only remove summands in the singular expansion.

We now consider the contributions from an edge $e \in \mathcal{E}$. We denote its interior angle by ω_e and introduce cylindrical coordinates (ρ_e, θ_e, z_e) with respect to the edge e.

Then we obtain, see (Dauge [50], Section 17.D) and (Grisvard [73], Theorem 2.5.11) for $t = 2$, or (Grisvard [72], Theorem 4.1) and (Grisvard [74], Section 7) for $t > 2$, that there exist functions $q_{j,e} \in W^{2/t'-j\pi/\omega_e,t}(\mathbb{R}_+)$, such that the singular part of the solution is of the form

$$\sum_{\substack{j=1 \\ \frac{j\pi}{\omega_c} \neq 1}}^{j<\frac{2\omega_c}{\pi t'}} (G_j(\rho_e, z_e) * q_{j,e})\, \varphi_{j,e}(\rho_e, \theta_e)$$

where $\varphi_{j,e}(\rho_e, \theta_e) := \eta_e(\rho_e)\rho_e^{\frac{j\pi}{\omega_e}} \sin\left(\frac{j\pi}{\omega_e}\theta_e\right)$ is the same function as in the 2d case, and

$$(G_j(\rho_e, z_e) * q_{j,e}) = \int_0^\infty G_j(\rho_e, s, z_e) * q_{j,e}(s)\, ds.$$

For $\frac{j\pi}{\omega_e} > 1 - \frac{2}{t}$ we set

$$G_j(\rho_e, s, z_e) := \frac{\rho_e}{\pi} \frac{2z_e s}{(\rho_e^2 + (z_e - s)^2)(\rho_e^2 + (z_e + s)^2)}$$

and for $\frac{j\pi}{\omega_e} \leq 1 - \frac{2}{t}$ we set

$$G_j(\rho_e, s, z_e) := \frac{2\rho_e^3}{\pi} \left(\frac{1}{(\rho_e^2 + (z_e - s)^2)^2} - \frac{1}{(\rho_e^2 + (z_e + s)^2)^2} \right).$$

We obtain that the second case $\frac{j\pi}{\omega_e} \leq 1 - \frac{2}{t}$ does not exist for sufficiently small $t > 3$, as this implies $j \leq \frac{\omega_e}{\pi}(1 - \frac{2}{t}) \leq 2(1 - \frac{2}{t}) \to \frac{2}{3}$. Hence we only have to deal with the first case.

We proceed exactly as in the 2d case. Let $t > 2$ sufficiently small, and $\frac{2\omega_e}{\pi t'} \notin \mathbb{N}$, then in the above sum only $j = 1$ appears, and the first derivative of $\varphi_{j,e}(\rho_e, \theta_e)$ is unbounded for $\rho_{j,e} \to 0$. Then noting that for arbitrary $\rho_e > 0$ we may interchange differentiation and integration by standard theorems, see, e.g., (Amann and Escher [4], Theorem 3.18), and $\lim_{\rho_e \to 0}(G_j(\rho_e, z_e) * q_{j,e}) = q_{j,e}(z)$ we obtain that as in the 2d case boundedness of the first derivative of $(G_j(\rho_e, z_e) * q_{j,e})\, \varphi_{j,e}$ implies $q_{j,e} \equiv 0$.

Combining edges and vertices, see (Grisvard [74], Section 7.2) or (Dauge [50], Section 17.D) we obtain the assertion. □

Remark 3.1. In addition to the result of Lemma 3.2, we remark that provided certain (countably many) critical values of t are avoided the operator $-\Delta$ is closed from $W^{2,t}(\Omega) \cap H_0^1(\Omega)$ into $L^t(\Omega)$ and the image $I \subset L^t(\Omega)$ is of finite codimension. Especially the operator is closed for $t = 2$.

To see this we first consider the case $n = 2$, then the result is obtained by (Grisvard [71], Theorem 4.3.2.4) for $t > 2$ under the condition $\frac{2\omega_c}{\pi t'} \notin \mathbb{N}$ for all interior angles ω_c. The case $t = 2$ is covered by (Grisvard [71], Theorem 4.3.1.4).

The case $n = 3$ and $t = 2$ is covered by (Dauge [51], Corollary 3.10), for the case $t > 2$ see (Grisvard [70], Theorem 5.8)[1].

[1] I would like to acknowledge the support of M. Dauge for an e-mail giving the same result, before I was able to find a citable source.

In particular, this implies that there exists a constant C such that for any $f \in I$ and corresponding solution $u \in W^{2,t}(\Omega) \cap H_0^1(\Omega)$ the following holds

$$\|u\|_{2,t} \leq C\|f\|_t.$$

Proof of Theorem 3.1. We begin by noting that, by assumption, there is at least one feasible control $q_0 \in Q^{\text{feas}}$. From Lemma 3.2 we obtain that for any $q \in Q^{\text{feas}}$ the corresponding state $u_q \in W^{2,t}(\Omega) \cap H_0^1(\Omega)$ for some $t > 2$ and the mapping $q \mapsto u_q$ is continuous.

Hence if $n = 2$ we can apply Theorem 2.1 by noting that the embedding $W^{2,t}(\Omega) \mapsto C^1(\overline{\Omega})$ is compact, and hence $\hat{g}\colon q \mapsto |\nabla u|^2 - \psi$ is weakly sequential continuous.

If $n = 3$ this is no longer the case. However, analog to the proof of Theorem 2.1 taking a minimizing sequence q_k in Q^{feas} we obtain boundedness of the sequence q_k in $L^t(\Omega)$. Hence there exists a weakly convergent subsequence w.l.o.g again denoted by q_k. By continuity of the solution operator, see Remark 3.1, and possibly selecting a further subsequence the states u_{q_k} converge weakly in $W^{2,t}(\Omega) \cap H_0^1(\Omega)$, and by compactness converge strongly in $W_0^{1,t}(\Omega)$. In particular the sequence $|\nabla u_{q_k}|^2$ converges in $L^1(\Omega)$. Hence possibly selecting another subsequence we obtain that $|\nabla u_{q_k}|^2$ converges pointwise almost everywhere. Thus Q^{feas} is closed with respect to weak convergence. Hence any weak limit $\overline{q}$ of the sequence q_k is feasible and a solution to the problem by weakly lower semi continuity of $\hat{J}$. □

3.1.2 Necessary Conditions

After having established the existence of a solution we will consider the system of first-order necessary conditions. We will not discuss second-order sufficient conditions as the problems at hand are convex. Hence the necessary conditions are already sufficient.

Lemma 3.3. *Let $(\overline{q}, \overline{u})$ be the solution of* (3.1) *or* (3.2)*. If it is a solution to* (3.2) *assume that there is a strictly feasible control $\hat{q}$, e.g., the corresponding state $\hat{u}$ satisfies $|\nabla \hat{u}|^2 < \psi$.*

Assume that t obtained in Lemma 3.2 is larger than n and that Δ is closed from $W^{2,t}(\Omega) \cap H_0^1(\Omega) \to L^t(\Omega)$.

Then there exists a measure $\overline{\mu} \in C(\overline{\Omega})^*$ such that the following holds:

$$\begin{aligned} \hat{J}'(\overline{q})(\varphi - \overline{q}) + \langle\, \overline{\mu}\,, \hat{g}'(\overline{q})(\varphi - \overline{q})\,\rangle_{C^* \times C} \geq 0 \quad & \forall\, \varphi \in Q^{ad} \cap I, \\ \langle\, \overline{\mu}\,, \varphi\,\rangle_{C^* \times C} \leq 0 \quad & \forall\, \varphi \in C(\overline{\Omega}), \varphi \leq 0, \\ \langle\, \overline{\mu}\,, |\nabla \overline{u}|^2 - \psi\,\rangle_{C^* \times C} = 0, & \end{aligned} \tag{3.3}$$

where I denotes the image of $W^{2,t}(\Omega) \cap H_0^1(\Omega)$ under Δ.

Proof. We note that the image I of $W^{2,t}(\Omega) \cap H_0^1(\Omega)$ under Δ is closed in $L^t(\Omega)$. Hence $I \cap L^p(\Omega)$ is closed in $L^p(\Omega)$ $(p \geq t)$, too. This means, it is sufficient to consider the optimization problem on the smaller space $Q = I \cap L^p(\Omega)$. Then $\hat{g}$ is differentiable on Q by construction. As in Section 6.2.1.3 we obtain that the solution $\overline{q}$ is a regular point.

Applying Theorem 2.2 yields the desired result. □

Remark 3.2. The result is almost identical to the smooth case, however, we had to consider $Q = I \cap L^p(\Omega)$ and hence we get a non local constraint into the admissible set $Q^{\mathrm{ad}} \cap I$.

In order to obtain a system similar to Theorem 2.3, where the influence of the equality and inequality constraints are separated, we have to consider the adjoint equation.

$$\langle -\Delta\varphi\,, \overline{z}\,\rangle_{Z^*\times Z} = J_u'(\overline{q},\overline{u})(\varphi) + \langle\,\overline{\mu}\,, g'(\overline{u},\nabla\overline{u})(\varphi)\,\rangle_{C^*\times C} \quad \forall \varphi \in W \tag{3.4}$$

where we have to find suitable spaces Z and W. As we like to use $\overline{u}$ as a test function we consider $W = W^{2,t}(\Omega) \cap H_0^1(\Omega)$ where t is given by Lemma 3.2. We need to consider the solvability of the adjoint equation.

To see this, we first note that the equation

$$(\nabla\varphi, \nabla z_0) = J_u'(\overline{q},\overline{u})(\varphi) \quad \forall \varphi \in H_0^1(\Omega)$$

possesses a solution $z_0 \in H_0^1(\Omega)$. Hence it is sufficient to consider solvability of the equation

$$\langle -\Delta\varphi\,, z_1\,\rangle_{Z^*\times Z} = \langle\,\overline{\mu}\,, 2\nabla\overline{u}\nabla\varphi\,\rangle_{C^*\times C} \quad \forall \varphi \in W. \tag{3.5}$$

It is clear that the right-hand side is an element of $\left(W^{2,t}(\Omega) \cap H_0^1(\Omega)\right)^*$ for any $t > n$. As $-\Delta\colon W^{2,t}(\Omega) \cap H_0^1(\Omega) \to I \subset L^t(\Omega)$ is an isomorphism. Hence the same holds true for $-\Delta^*\colon I^* \to \left(W^{2,t}(\Omega) \cap H_0^1(\Omega)\right)^*$. Setting $I^\perp = \{v \in L^{t'}(\Omega)\,|\,(v,q) = 0\ \forall\, q \in I\}$ we have $I^* \cong L^{t'}(\Omega)/I^\perp$ because I is closed in $L^t(\Omega)$, see, e.g., (Werner [154], Theorem III.1.10).

By choosing $Z = L^{t'}(\Omega)$ there exists a solution z_1 to (3.5) which is uniquely determined modulo $I^\perp$. Hence $\overline{z} = z_0 + z_1 \in L^{t'}(\Omega)$ is a solution to (3.4).

By combining this with Lemma 3.3 we get the following

Theorem 3.4. *Under the assumptions of Lemma 3.3 for a solution $(\overline{q},\overline{u})$ of (3.1) or (3.2), there exists a measure $\overline{\mu} \in C(\overline{\Omega})^*$ and a function $\overline{z} \in L^{t'}(\Omega)$ such that:*

$$\begin{aligned}
(\nabla\overline{u}, \nabla\varphi) &= (\overline{q},\varphi) && \forall\ \varphi \in H_0^1(\Omega),\\
\langle -\Delta\varphi\,, \overline{z}\,\rangle_{Z^*\times Z} &= J_u'(\overline{q},\overline{u})(\varphi) + \langle\,\overline{\mu}\,, g'(\overline{u},\nabla\overline{u})(\varphi)\,\rangle_{C^*\times C} && \forall\ \varphi \in W,\\
J_q'(\overline{q})(\delta q - \overline{q}) &\geq -\langle\,\delta q - \overline{q}\,, \overline{z}\,\rangle_{Z^*\times Z} && \forall\, \delta q \in Q^{ad} \cap I,\\
\langle\,\overline{\mu}\,, \varphi\,\rangle_{C^*\times C} &\leq 0 && \forall\ \varphi \in C(\overline{\Omega}), \varphi \leq 0,\\
\langle\,\overline{\mu}\,, |\nabla\overline{u}|^2 - \psi\,\rangle_{C^*\times C} &= 0.
\end{aligned} \tag{3.6}$$

We remark that z_1 being determined only modulo $I^\perp$ doesn't affect the variational inequality

$$J_q'(\overline{q})(\delta q - \overline{q}) \geq -\langle\,\delta q - \overline{q}\,, \overline{z}\,\rangle_{Z^*\times Z} \quad \forall\, \delta q \in Q^{\mathrm{ad}} \cap I$$

because the test functions are chosen from I.

Remark 3.3. We note that, in the case $n = 3$, there is a gap between the existence Theorem 3.1 and the necessary conditions Lemma 3.3 and Theorem 3.4 because there are certain angles for which we could not obtain $W^{2,t}$ regularity of the solution for $t > 3$. The problem in the proof of the necessary conditions is that this implies that the mapping $\hat{g}$ is not differentiable in a neighborhood of $\overline{q}$.

3.2 Regularity

We will now discuss another issue of importance, namely that of regularity in the context of first-order constraints.

We recall that the necessary condition (2.10) for problem (2.12) takes the explicit form (2.15). In particular, the adjoint state $\overline{z}$ is contained in any $L^s(\Omega)$ with $s < \frac{n}{n-1}$, see (Casas and Fernández [36]). If this would be best possible, then this would automatically limit the possibility to obtain higher regularity for the control variable by bootstrapping arguments, because the control and the adjoint state are linked by the (algebraic) equation (2.15d). For instance if $\overline{z}$ has no derivatives, then in general $\overline{q}$ has none either.

We will now show that there is in fact some additional regularity for this problem. Parts of this proof are already published in (Ortner and Wollner [120]). We will add here the case of pointwise bounds on the control variable, and conclude with a regularity result on non-smooth domains.

The Case of a Smooth Domain It will be crucial for our analysis that there exists $t > n$ such that $-\Delta$ is $W^{2,t}$-regular, e.g. for $q \in L^t(\Omega)$ the weak solution $u \in H_0^1(\Omega)$ to

$$(\nabla u, \nabla \varphi) = (q, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega)$$

belongs in fact to $W^{2,t}(\Omega)$, and $\|u\|_{W^{2,t}} \leq c\|q\|_{L^t}$. If the boundary is of class $C^{1,1}$ this is obtained by classical regularity theory for any t, this dates back to (Agmon, Douglis, and Nirenberg [2]) for a more recent exposition see (Gilbarg and Trudinger [67]).

If the domain is polygonal or polyhedral the existence of such a t requires additional conditions on the domain. If $n = 2$ there is such t provided the domain is convex see (Grisvard [73], Thm. 4.4.3.7). If $n = 3$ then one needs to assume in addition, that the angle between any two faces of Ω is bounded strictly above by $\frac{3}{4}\pi$ (Dauge [51], Cor. 3.7). Let now be $t_{\max} > n$ be defined such that $-\Delta$ is $W^{2,t}$-regular for any $t \in (n, t_{\max})$.

From the necessary optimality conditions (2.15) we can in a first step derive additional regularity for the adjoint state $\overline{z}$. To do so we will employ the K-Method of interpolation (although any other method would do fine). Hence we define fractional-order Sobolev spaces $W^{s,p}$ by Besov spaces $B^s_{p,p}$. For details on this see, e.g., (Triebel [140], Definition 4.2.1) or (Adams and Fournier [1], Chapter 7).

Lemma 3.5. *The solution $\overline{z}$ of* (2.15b) *belongs to $W^{1-n/t-\varepsilon,t'}(\Omega)$ for every $\varepsilon > 0$ and $t \in (n, t_{max})$, where we define t' as usual by $\frac{1}{t} + \frac{1}{t'} = 1$.*

Proof. Let $\varepsilon > 0$ be given, then

$$\langle \nabla\varphi \nabla\overline{u}, \mu\rangle_{C,C^*} \leq \|\overline{u}\|_{C^1}\|\overline{\mu}\|_{C^*}\|\varphi\|_{C^1} \leq C\|\varphi\|_{W^{1+n/t+\varepsilon,t}}$$

by standard embedding theorems (Triebel [140], Theorem 4.6.1). Hence, the right-hand side of (2.15b) is an element of $W^{-1-n/t-\varepsilon,t'}(\Omega)$.

By definition of $t_{\max}$ we have that

$$\begin{aligned} A_1 &:= -\Delta \colon W_0^{1,t}(\Omega) \to W^{-1,t}(\Omega), \quad \text{and} \\ A_2 &:= -\Delta \colon W^{2,t}(\Omega) \cap W_0^{1,t}(\Omega) \to L^t(\Omega) \end{aligned}$$

are isomorphisms. Hence, the adjoint operators

$$\begin{aligned} A_1^* &\colon W_0^{1,t'}(\Omega) \to W^{-1,t'}(\Omega), \quad \text{and} \\ A_2^* &\colon L^{t'}(\Omega) \to W^{-2,t'}(\Omega) \end{aligned}$$

are isomorphisms as well. By interpolation we obtain (Triebel [140], Theorem 4.6.1, Theorem 4.8.2), that

$$W^{-1-n/t-\varepsilon,t'}(\Omega) = (W^{-1,t'}(\Omega), W^{-2,t'}(\Omega))_{n/t+\varepsilon,t'},$$

and hence that (Triebel [140], Theorem 1.3.3)

$$\overline{z} \in (W_0^{1,t'}(\Omega), L^{t'}(\Omega))_{n/t+\varepsilon,t'} = W^{1-n/t-\varepsilon,t'}(\Omega).$$

This concludes the proof. □

Setting $\Phi(g) = \operatorname{sign}(g)|g|^{1/(r-1)}$ it follows from (2.15d) that

$$\overline{q} = \max\bigl(a, \min\bigl(b, \Phi(\frac{-1}{\alpha}\overline{z})\bigr)\bigr) \tag{3.7}$$

almost everywhere. Hence we can deduce regularity of $\overline{q}$ from regularity of $\overline{z}$.

Lemma 3.6. *Let $f \in W^{s,t'}(\Omega)$ with $s < 1$ and let $r \geq 2$, then*

$$\operatorname{sign}(f)|f|^{\frac{1}{r-1}} \in W^{\frac{s}{r-1},t'(r-1)}(\Omega).$$

Proof. The result follows from the fact that the function $\Phi(g) = \operatorname{sign}(g)|g|^{\alpha}$ belongs to $C^{0,\alpha}(\mathbb{R})$, more precisely, that it satisfies the Hölder condition

$$\sup_{\substack{g_1,g_2 \in \mathbb{R} \\ g_1 \neq g_2}} \frac{|\Phi(g_1) - \Phi(g_2)|}{|g_1 - g_2|^{\alpha}} \leq 2. \tag{3.8}$$

The stated result follows if we show, setting $\alpha = 1/(r-1)$ in the definition of Φ, that $\Phi \circ f \in W^{\frac{s}{r-1},t'(r-1)}(\Omega)$. To this end, we need to show that $\Phi \circ f \in L^{t'(r-1)}(\Omega)$ (this is easy to see), and that the semi-norm

$$|\Phi \circ f|^{t'(r-1)}_{\frac{s}{r-1},t'(r-1)} = \int_\Omega \int_\Omega \frac{\bigl|\Phi(f(x)) - \Phi(f(y))\bigr|^{t'(r-1)}}{|x-y|^{n+\frac{s}{r-1}t'(r-1)}}\, dx\, dy$$

is finite. Using the Hölder-condition (3.8) we estimate

$$|\Phi \circ f|^{t'(r-1)}_{W^{\frac{s}{r-1},t'(r-1)}} \leq 2^{t'(r-1)} \int_\Omega \int_\Omega \frac{|f(x) - f(y)|^{\frac{t'(r-1)}{r-1}}}{|x-y|^{n+s\frac{t'(r-1)}{r-1}}} dx dy = 2^{t'(r-1)} |f|^{t'}_{s,t'},$$

which is finite due to our assumption that $f \in W^{s,t'}(\Omega)$. □

We can now formulate the desired regularity result for the optimal control $\overline{q}$.

Corollary 3.7. *For any $\varepsilon > 0$ the optimal control $\overline{q}$ given by (2.12) belongs to the space $W^{\gamma,p}$, where $\gamma = (1 - n/t - \varepsilon)/(r-1)$ and $p = t'(r-1)$ for any $t \in (n, t_{max}]$.*

Proof. Recall from Lemma 3.5 that $z \in W^{1-n/t-\varepsilon,t'}(\Omega)$. Applying Lemma 3.6 with $f = \frac{-1}{\alpha} z$, and $s = 1 - n/t - \varepsilon$, together with (3.7) and Lipschitz continuity of the max and min function in $W^{\frac{s}{r-1},t'(r-1)}(\Omega)$, see (Kinderlehrer and Stampacchia [95], Thm. II.A.1), we obtain that $\overline{q} \in W^{\frac{1}{r-1}(1-n/t-\varepsilon),t'(r-1)}(\Omega)$ which establishes the stated regularity. □

Remark 3.4. We note that Corollary 3.7 shows that the convergence orders obtained in (Deckelnick, Günther, and Hinze [55], Günther and Hinze [76], Hinze, Pinnau, Ulbrich, and Ulbrich [90]), namely $O(h^{\frac{1-n/t}{r}})$, are not of optimal order for the control variable with respect to the given regularity. However we will see in Section 4.4 that they are apparently sharp. This means the discretization doesn't have a quasi best-approximation property for the control.

The Case of a Non-Smooth Domain For the sake of brevity, we restrict ourself to the case $n = 2$, $Q^{\mathrm{ad}} = Q$. Then the premises of Theorem 3.4 are always met. Let t be given by Theorem 3.4.

Lemma 3.8. *Any solution $\overline{z}$ of (3.4) belongs to $W^{1-n/t-\varepsilon,t'}(\Omega)$ for every $\varepsilon > 0$ where we define t' as usual by $\frac{1}{t} + \frac{1}{t'} = 1$.*

Proof. The proof is almost identical to Lemma 3.5. We assume for simplicity that $t < 3$, then we obtain from (Dauge [51], Remark 3.11) that

$$A_1 := -\Delta \colon W_0^{1,t}(\Omega) \to W^{-1,t}(\Omega),$$

is an isomorphism. By assumption on t we have in addition, that

$$A_2 := -\Delta \colon W^{2,t}(\Omega) \cap W_0^{1,t}(\Omega) \to I \subset L^t(\Omega),$$

is an isomorphism. Hence the adjoint operators

$$\begin{aligned} A_1^* &\colon\ W_0^{1,t'}(\Omega) \to W^{-1,t'}(\Omega), \quad \text{and} \\ A_2^* &\colon\ I^* \to (W^{2,t}(\Omega) \cap W_0^{1,t}(\Omega))^* \end{aligned}$$

are isomorphisms, too. We note that, as in Section 3.1.2, $I^* \cong L^{t'}(\Omega)/I^\perp$. Especially by selecting an arbitrary element $s \in I^\perp$ and using $L^{t'}(\Omega) \cong I^* \oplus I^\perp$ we can lift an element $z \in I^*$ to $L^{t'}(\Omega)$ by setting $l_s(z) = z + s$. Hence the 'inverse' mapping $l_s \circ (A_2^*)^{-1} \colon (W^{2,t}(\Omega) \cap H_0^1(\Omega))^* \to L^{t'}(\Omega)$ is continuous.

By interpolation for the continuous operators $(A_1^*)^{-1}$ and $l_s \circ (A_2^*)^{-1}$ we obtain, that

$$\overline{z} \in (W_0^{1,t'}(\Omega), L^{t'}(\Omega))_{n/t+\varepsilon,t'}.$$

This proofs the assertion. □

Then we obtain from the reduced gradient

$$J_q'(\overline{q})(\delta q) = -\langle\, \delta q\,, \overline{z}\,\rangle_{Z^*\times Z} \quad \forall\, \delta q \in Q^{\mathrm{ad}} \cap I$$

that

$$\alpha|\overline{q}|^{r-2}\overline{q} = \overline{z} + s$$

with some $s \in I^{\perp}$, compare Theorem 3.4. In the case $n = 2$ we can explicitly state a basis for the space of dual singular functions $I^{\perp}$, see, e.g., (Blum and Dobrowolski [26]) and obtain, that $s \in W_0^{1-n/t-\varepsilon,t'}(\Omega)$ for any $\varepsilon > 0$.

Hence we obtain

Corollary 3.9. *For any $\varepsilon > 0$ the optimal control $\overline{q}$ given by (2.12) belongs to the space $W^{\gamma,p}$, where $\gamma = (1 - n/t - \varepsilon)/(r-1)$ and $p = t'(r-1)$.*

Proof. From the regularity

$$\alpha|\overline{q}|^{r-2}\overline{q} = \overline{z} + s \in W_0^{1-n/t-\varepsilon,t'}(\Omega)$$

we obtain the desired result from Lemma 3.6. □

3.3 Existence with L^2-regularization

In this section we consider a similar problem to (3.1). The main difference is, that the regularization is too weak to obtain a solution in $C^1(\overline{\Omega})$ even on a smooth domain. Namely we consider:

$$\begin{aligned} \text{Minimize } J(q,u) = \frac{1}{2}\| u - u^d \|^2 + \frac{\alpha}{2}\| q \|^2 & \\ (\nabla u, \nabla\varphi) = (q,\varphi) \qquad & \forall\, \varphi \in H_0^1(\Omega), \\ |\nabla u|^2 \le \psi \qquad & \text{a.e. in } \overline{\Omega}, \end{aligned} \tag{3.9}$$

with given $\alpha, \psi > 0$.

Theorem 3.10. *Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ be a polygonal or polyhedral domain. Then problem (3.9) has a unique solution $(q,u) \in L^2(\Omega) \times H^2(\Omega) \cap H_0^1(\Omega)$.*

Proof. The proof is analogous to Theorem 3.1, by noting that Lemma 3.2 remains true for $p = 2$. □

As in the case of certain non-smooth domains the problem in showing necessary conditions lies in the fact, that the mapping $\hat{g}$ is not differentiable in a neighbourhood of $\overline{q}$, compare Remark 3.3.

4 A Priori Error Estimates

We will now turn our attention towards the discretization of problem (2.5).

First estimates for the simpler case of pure control-constraints, e.g. $g \equiv 0$, were already obtained for distributed controls by (Falk [62], Geveci [66]) a nice overview including Neumann control can be found in (Malanowski [104]). All these results where obtained for a linear state equation and using a piecewise constant discretization of the control space and continuous piecewise linear elements for the state. These results have been extended to semilinear equations in (Arada, Casas, and Tröltzsch [5]) and (Casas, Mateos, and Tröltzsch [43]) for the case of Neumann control. In the case of Dirichlet control there has up to now only been an analysis for pure equality constraint problems in (May, Rannacher, and Vexler [106]) for polygonal domains or (Deckelnick, Günther, and Hinze [56]) for domains with curved boundaries.

In a next step, the results have been extended to the case of a continuous piecewise linear approximation of the control space in (Casas and Tröltzsch [40], Rösch [127]). Also results for the convergence in L^∞ have been derived by (Meyer and Rösch [113]). Finally, it could be shown that certain post-processing could enhance the convergence order of the control variable (Meyer and Rösch [112]) for a scalar state equation or (Rösch and Vexler [128]) for the stokes equation.

A so called 'variational discretization' was introduced by (Hinze [88]) where the discretization for the control variable is implicitly given by the adjoint state via (2.10c).

In all cases the obtained convergence orders are optimal with respect to ansatz space and regularity of the solution.

In the case of pointwise (zero-order) state constraints, e.g., $g = u - \psi$, (Casas and Mateos [38]) showed convergence for semilinear equations but without rate, in (Deckelnick and Hinze [53]) convergence rates for the variational discretization followed. For piecewise constant control approximations rates were obtained in (Deckelnick and Hinze [52]) and the case of piecewise linear controls is discussed in (Meyer [111]).

These results yield optimal rates with respect to the control variable, the rates for the state however are not optimal. The first optimal rate for the state was obtained in (Merino, Tröltzsch, and Vexler [110]) for the case of finitely many controls.

For first-order state constraints, e.g., $g = |\nabla u|^2 - \psi$, (Deckelnick, Günther, and Hinze [54]) showed convergence for a variational discretization in combination with a mixed discretization of the state equation and additional control constraints. In (Günther and Hinze [76]) this is extended to the case of piecewise constant controls, see also (Deckelnick et al. [55], Hinze

et al. [90]). One should note that the results depend on boundedness results for the discrete multiplier for the state constraint.

We will recall the results published in (Ortner and Wollner [120]) which yield the same results without using discrete Lagrange multipliers. Further we will obtain convergence for a piecewise bi- or trilinear control discretization.

To outline the structure of the following proofs, let $V_h \subset V$ be a finite element space. Then we can discretize the state equation using the discrete space V_h. Then, as assumed in Section 2.3, the state constraint g is well-defined as a mapping from V_h into $L^\infty(\Omega^C) = G_h$. If the control variable is searched for in a finite dimensional space, we can show convergence at this point. Otherwise this serves as an intermediate problem and we may discretize the space for the control variable using some possible different finite element space. Alternatively, one can discretize the control variable implicitly by the set of first-order necessary conditions. In either case, we denote the space from which we take the control variable by Q_h (even if $Q_h = Q$) and set $Q_h^{\mathrm{ad}} := Q^{\mathrm{ad}} \cap Q_h$ then we obtain the following discretized problem:

$$\begin{aligned} &\text{Minimize } \ J(q_h, u_h) \\ &\text{subject to } \begin{cases} (q_h, u_h) \in Q_h^{\mathrm{ad}} \times V_h, \\ g(u_h, \nabla u_h) \leq 0 \qquad \text{in } G_h, \\ a(q_h, u_h)(\varphi_h) = f(\varphi_h) \ \ \forall\ \varphi_h \in V_h. \end{cases} \end{aligned} \tag{4.1}$$

4.1 Problem Formulation

Let Ω be a convex polygonal (or polyhedral) domain in $\mathbb{R}^n$, $n \leq 3$. We consider the linear elliptic PDE

$$-\Delta u = Bq \quad \text{in } \Omega, \tag{4.2a}$$
$$u = 0 \qquad \text{on } \partial\Omega, \tag{4.2b}$$

where $B\colon Q^{\mathrm{ad}} \to L^t(\Omega)$, $t \in (n, \infty)$, is a linear continuous operator, and Q is a reflexive Banach space which we specify below.

It will be crucial for our analysis that there exists $t > n$ such that (4.2) is $W^{2,t}$-regular (the weak solution $u \in V = H_0^1(\Omega)$ belongs in fact to $W = W^{2,t}(\Omega)$). If $n = 2$ then this result follows from (Grisvard [73], Thm. 4.4.3.7). If $n = 3$ then one needs to assume, in addition, that the angle between any two faces of Ω is bounded strictly above by $\frac{3}{4}\pi$ (Dauge [51], Cor. 3.7). We assume throughout that this is satisfied, that is, if $Bq \in L^t(\Omega)$ then

$$u \in W^{2,t}(\Omega) \cap W_0^{1,t}(\Omega) \subset C^{1,1-n/t}(\overline{\Omega}),$$

and that there exist constants c, c' such that

$$\|u\|_{C^{1,1-n/t}} \leq c\|u\|_{2,t} \leq c'\|Bq\|_{L^t}. \tag{4.3}$$

In terms of Section 3.2 we assume a ‘smooth’ domain.

After these preliminary remarks we pose the following optimal control problem:

$$\text{Minimize } J(q,u) := \frac{1}{2}\|u-u^d\|_{L^2}^2 + R(q), \tag{4.4a}$$

$$\text{subject to } \begin{cases} (q,u) \in Q \times V, \\ (q,u) \text{ satisfies (4.2)}, \\ |\nabla u|^2 \le \psi \text{ in } \overline{\Omega}, \end{cases} \tag{4.4b}$$

where $u^d \in L^2(\Omega)$ is fixed, and the regularization function R depends on Q. We consider two possible situations:

(Q.1) $Q = \mathbb{R}^d$, for some $d \in \mathbb{N}$, $R(q) = \frac{\alpha}{2}\|q\|_{\ell^2}^2$, and $B : \mathbb{R}^d \to L^r(\Omega)$ is an arbitrary linear continuous operator.

(Q.2) $Q = L^r(\Omega)$, $R(q) = \frac{\alpha}{r}\|q\|_{L^r}^r$, and $B = \mathrm{Id}$, $r > n$.

The case (Q.1) corresponds to the –more realistic– case of a finite dimensional control that has distributed influence on the solution variable. The second case (Q.2) is important in inverse problems, e.g., when the 'control' is in fact an unknown volume force that one tries to recover from a set of measurements.

Either of the assumptions (Q.1) or (Q.2) guarantee strong convexity of the optimal control problem, that is, the following Clarkson-type inequality holds for $w_1 = (q_1, u_1)$, $w_2 = (q_2, u_2) \in Q \times V$

$$\frac{1}{2}\|\frac{1}{2}(u_1-u_2)\|_{L^2}^2 + R(\frac{1}{2}(q_1-q_2)) + J(\frac{1}{2}(w_1+w_2)) \le \frac{1}{2}J(w_1) + \frac{1}{2}J(w_2). \tag{4.5}$$

The inequality for the q-variable is simply Clarkson's inequality the inequality for the u-variable is obtained by applying Clarkson's inequality to $u_1 - u^d$ and $u_2 - u^d$. Using (4.5) and the discussion of Example (2.12) one can show that there exists a unique solution $(\overline{q}, \overline{u}) \in Q^{\mathrm{ad}} \times V$ to (4.4).

We will later use the fact that both $R\colon Q \to \mathbb{R}$ and $J\colon Q \times V \to \mathbb{R}$ are twice differentiable, and denote the first derivatives, respectively, by R' and J', for example,

$$\langle\, R'(q)\,, \delta q \,\rangle_{Q^*\times Q} := \lim_{\varepsilon \to 0} \frac{R(\|q+\varepsilon\delta q\|_Q) - R(\|q\|_Q)}{\varepsilon}.$$

We note that, in the case (Q.1) $R'(q) = \alpha q$, while in the case (Q.2), the first derivative takes the form $R'(q) = \alpha|q|^{r-2}q$.

4.2 Discretization

For the discretization of (4.4), we assume that we are given a family $(\mathcal{T}_h)_{h\in(0,1]}$ of triangulations, consisting of triangles or quadrilaterals in 2d, and of tetrahedra or hexahedra in 3d, which are *affine-equivalent* to their respective reference elements, such that $\mathrm{diam}(T) \le h$ for all $T \in \mathcal{T}_h$, $h \in (0,1]$. We assume throughout that the family is quasi-uniform in the sense of

(Brenner and Scott [32], Def. 4.4.13), that is, there exists $\rho > 0$ such that, for each $T \in \mathcal{T}_h$ and $h \in (0, 1]$ there exists a ball $B_T \subset T$ such that $\operatorname{diam}(B_T) \geq \rho h$.

We define the discrete state space V_h as the space of continuous piecewise linear (or bi-, or tri-linear) functions with respect to the mesh $\mathcal{T}_h$. For fixed $q \in Q^{\mathrm{ad}}$, the semi-discretized state equation then reads: Find $u_h \in V_h$

$$(\nabla u_h, \nabla \varphi_h) = (Bq, \varphi_h) \quad \forall\, \varphi_h \in V_h. \tag{4.6}$$

The corresponding semi-discretized optimal control problem becomes

$$\text{Minimize } J(q_h, u_h) := \frac{1}{2}\|u_h - u^d\|_{L^2}^2 + R(q_h), \tag{4.7a}$$

$$\text{subject to } \begin{cases} (q_h, u_h) \in Q \times V_h, \\ (q_h, u_h) \text{ satisfies (4.6)}, \\ |\nabla u_h|^2 \leq \psi \text{ a.e. in } \overline{\Omega}. \end{cases} \tag{4.7b}$$

In the case (Q.2), i.e., $Q = L^r(\Omega)$, we also need to discretize the control space Q. We consider two different discretizations: either $Q^h = Q^h_{(0)}$ or $Q^h = Q^h_{(1)}$, where

$$Q^h_{(0)} = \{\, q_h \in Q \,|\, q_h \text{ is p.w. constant w.r.t. } \mathcal{T}_h \,\}, \quad \text{and}$$
$$Q^h_{(1)} = \{\, q_h \in C(\bar{\Omega}) \,|\, q_h \text{ is p.w. (bi-/tri-)linear w.r.t. } \mathcal{T}_h \,\}.$$

Our analysis applies to both choices, however, we will see that for the choice $Q^h = Q^h_{(0)}$ it yields a better convergence rate. The choice $Q^h = Q^h_{(1)}$ was not previously considered in the literature.

The fully discretized optimal control problem reads

$$\text{Minimize } J(q_h^h, u_h^h) := \frac{1}{2}\|u_h^h - u^d\|_{L^2}^2 + R(q_h^h), \tag{4.8a}$$

$$\text{subject to } \begin{cases} (q_h^h, u_h^h) \in Q^h \times V_h, \\ (q_h^h, u_h^h) \text{ satisfies (4.6)}, \\ |\nabla u_h^h|^2 \leq \psi \text{ in } \overline{\Omega}. \end{cases} \tag{4.8b}$$

We remark that the restrictions we imposed on the family $(\mathcal{T}_h)_{h\in(0,1]}$ ensure that the usual interpolation error results, best approximation results, and inverse estimates hold (Brenner and Scott [32], Sec. 4.4 and 4.5). In particular, it follows that the Ritz projection is stable in $W^{1,\infty}(\Omega)$, that is, there exists $c \in \mathbb{R}$ such that if $u \in W_0^{1,\infty}(\Omega)$, and if $u_h \in V_h$ satisfies

$$(\nabla u_h, \nabla \varphi) = (\nabla u, \nabla \varphi) \quad \forall \varphi \in V_h, \quad \text{then} \quad \|\nabla u_h\|_\infty \leq c\|\nabla u\|_\infty; \tag{4.9}$$

see (Rannacher and Scott [121]) for simplicial meshes and (Brenner and Scott [32], Thm. 8.1.11 and Ex. 8.x.1) for the general case.

Finally, we define $\Pi_h \colon L^1(\Omega) \to Q^h$ to be the natural extension of the L^2-projection operator, that is, for $u \in L^1(\Omega)$, we define $\Pi_h u \in Q^h$ via

$$(\Pi_h q, \varphi) = (q, \varphi) \qquad \forall \varphi \in Q^h. \tag{4.10}$$

It is shown in (Douglas, Dupont, and Wahlbin [57]) that Π_h is stable as an operator from $L^p(\Omega)$ to $L^p(\Omega)$, for any $p \in [1, \infty]$, that is, there exist constants c_p, independent of h, such that

$$\|\Pi_h q\|_{L^p} \leq c_p \|q\|_{L^p} \qquad \forall q \in L^p(\Omega). \tag{4.11}$$

4.3 A Priori Estimates

4.3.1 State Discretization

First we consider the case, where only the state space is discretized. This is reasonable if the control space is finite dimensional (i.e., case (Q.1)), or if we use the 'variational discretization' concept discussed in (Hinze [88]). In general this is an intermediate step that gives us preliminary insights into the convergence behavior of the discretization. The results that we will obtain are essentially the same as those in (Deckelnick et al. [54, 55], Günther and Hinze [76], Hinze et al. [90]), however, we do not require bounds on the discrete adjoint variables in our analysis.

Theorem 4.1. *Let $(\overline{q}, \overline{u}) \in Q \times V$ be the solution to (4.4), and $(\overline{q}_h, \overline{u}_h) \in Q \times V_h$ be the solution to the semi-discretized problem (4.7). Then there exists a constant C, independent of h, such that*

$$|J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)| \leq C h^{1-n/t}. \tag{4.12}$$

Proof. Instead of using the criticality conditions for solutions, the idea of the proof is to construct discrete and continuous competitors for which the error in the cost functional can be estimated immediately.

We begin by investigating the solutions $\overline{u}$ of (4.2) and its Ritz projection u_h which are, respectively, given by

$$\begin{aligned} (\nabla \overline{u}, \nabla \varphi) &= (B\overline{q}, \varphi) \qquad \forall\, \varphi \in V, \qquad \text{and} \\ (\nabla u_h, \nabla \varphi) &= (B\overline{q}, \varphi) \qquad \forall\, \varphi \in V_h. \end{aligned}$$

The difficulty is that, possibly, $|\nabla u_h| \not\leq \psi$. However, using the stability of the Ritz projection in $W^{1,\infty}(\Omega)$ (Rannacher and Scott [121]), we can see that the constraint on the gradient is *almost* satisfied. Namely, in view of the regularity estimate (4.3), it follows from (Rannacher and Scott [121], Eq. (1.7)) that

$$\|\nabla \overline{u} - \nabla u_h\|_{L^\infty} \leq c h^{1-n/t} \|\overline{u}\|_{C^{1,1-n/t}} \leq c h^{1-n/t} \|B\overline{q}\|_{L^t}.$$

From this, we derive the bound

$$|\nabla u_h(x)| \leq |\nabla \overline{u}(x)| + c h^{1-n/t} \|B\overline{q}\|_{L^t} \qquad \text{for a.e. } x \in \Omega.$$

Setting $\beta = 1 - n/t$ and $\tilde{c}\psi \geq c\|B\overline{q}\|_{L^t}$, it follows that

$$(1 - \tilde{c}h^\beta)|\nabla u_h| \leq (1 - \tilde{c}h^\beta)\psi + (1 - \tilde{c}h^\beta) c h^\beta \|B\overline{q}\|_{L^t} < \psi \quad \text{a.e. in } \overline{\Omega} \quad \forall h \in (0, 1].$$

Thus, we find that the sequence

$$(\tilde{q}_h, \tilde{u}_h) := ((1 - \tilde{c}h^\beta)\overline{q}, (1 - \tilde{c}h^\beta)u_h) \tag{4.13}$$

is feasible for (4.7) and that the following estimates hold:

$$\|B\overline{q} - B\tilde{q}_h\|_{L^r} \leq c\|\overline{q} - \tilde{q}_h\|_Q \leq ch^\beta\|\overline{q}\|_Q, \quad \text{and}$$
$$\|\overline{u} - \tilde{u}_h\|_{1,\infty} \leq \|\overline{u} - u_h\|_{1,\infty} + \|u_h - \tilde{u}_h\|_{1,\infty} \leq ch^\beta\|B\overline{q}\|_{L^r} + ch^\beta\|u_h\|_{1,\infty}.$$

Using again the $W^{1,\infty}$-stability of the Ritz projection, we obtain

$$\|\overline{u} - \tilde{u}_h\|_{1,\infty} \leq ch^\beta\|B\overline{q}\|_{L^r} \leq ch^\beta\|\overline{q}\|_Q.$$

Differentiability of the cost functional implies local Lipschitz continuity, and therefore we can deduce that

$$|J(\overline{q}, \overline{u}) - J(\tilde{q}_h, \tilde{u}_h)| \leq ch^\beta.$$

Since $(\tilde{u}_h, \tilde{q}_h)$ is an admissible pair for (4.7), the relation $J(\overline{q}_h, \overline{u}_h) \leq J(\tilde{q}_h, \tilde{u}_h)$ is satisfied, and hence

$$J(\overline{q}_h, \overline{u}_h) - J(\overline{q}, \overline{u}) \leq J(\tilde{q}_h, \tilde{u}_h) - J(\overline{q}, \overline{u}) \leq ch^\beta.$$

It should be mentioned that the last inequality already implies that $R(\overline{q}_h)$ is uniformly bounded for $h \in (0, 1]$, and hence there exists c independent of h such that $\|\overline{q}_h\|_Q \leq c$.

To obtain the reverse inequality, we start from $(\overline{q}_h, \overline{u}_h)$ and, using precisely the same arguments, construct $(\hat{q}, \hat{u})$ that are feasible for the exact problem (4.4) (note though, that $\hat{q}, \hat{u}$ do depend on h) and satisfy

$$|J(\overline{q}_h, \overline{u}_h) - J(\hat{q}, \hat{u})| \leq ch^\beta.$$

In summary, we obtain

$$-ch^\beta \leq J(\overline{q}, \overline{u}) - J(\tilde{q}_h, \tilde{u}_h) \leq J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h) \leq J(\hat{q}, \hat{u}) - J(\overline{q}_h, \overline{u}_h) \leq ch^\beta,$$

which concludes the proof of the theorem. □

From the error estimate on the objective functional, we can derive an estimate for the control and for the state.

Corollary 4.2. *Let $(\overline{q}, \overline{u}) \in Q \times V$ be the solution of (4.4), and let $(\overline{q}_h, \overline{u}_h) \in Q \times V_h$ be the solution of (4.7), then*

$$\|\overline{q} - \overline{q}_h\|_Q \leq ch^{\frac{1-n/t}{a}} \quad \text{and} \quad \|\overline{u} - \overline{u}_h\|_{L^2} \leq ch^{\frac{1-n/t}{2}},$$

where $a = 2$ in the case (Q.1) and $a = r$ in the case (Q.2).

Proof. Let $(\tilde{q}_h, \tilde{u}_h)$ be defined by (4.13), then it follows that

$$\begin{aligned} \|\overline{q} - \overline{q}_h\|_Q &\leq \|\overline{q} - \tilde{q}_h\|_Q + \|\tilde{q}_h - \overline{q}_h\|_Q \leq ch^{1-n/t} + \|\tilde{q}_h - \overline{q}_h\|_Q, \\ \|\overline{u} - \overline{u}_h\|_{L^2} &\leq \|\overline{u} - \tilde{u}_h\|_{L^2} + \|\tilde{u}_h - \overline{u}_h\|_{L^2} \leq ch^{1-n/t} + \|\tilde{u}_h - \overline{u}_h\|_{L^2}. \end{aligned} \tag{4.14}$$

To bound the remaining terms on the right-hand side, we apply Clarkson's inequality (4.5), which gives

$$\frac{1}{2}\Big\|\frac{1}{2}(\tilde{u}_h-\overline{u}_h)\Big\|_{L^2}^2+R\Big(\frac{1}{2}(\tilde{q}_h-\overline{q}_h)\Big) \\ \leq\frac{1}{2}J(\tilde{q}_h,\tilde{u}_h)+\frac{1}{2}J(\overline{q}_h,\overline{u}_h)-J\Big(\frac{1}{2}(\tilde{q}_h+\overline{q}_h),\frac{1}{2}(\tilde{u}_h+\overline{u}_h)\Big).$$

Since $(\tilde{q}_h,\tilde{u}_h)$ is feasible for (4.7) it follows that

$$J(\overline{q}_h,\overline{u}_h)\leq J\Big(\frac{1}{2}(\tilde{q}_h+\overline{q}_h),\frac{1}{2}(\tilde{u}_h+\overline{u}_h)\Big),$$

and hence, using Theorem 4.1,

$$\frac{1}{2}\Big\|\frac{1}{2}(\tilde{u}_h-\overline{u}_h)\Big\|_{L^2}^2+R\Big(\frac{1}{2}(\tilde{q}_h-\overline{q}_h)\Big)\leq\frac{1}{2}J(\tilde{q}_h,\tilde{u}_h)-\frac{1}{2}J(\overline{q}_h,\overline{u}_h)\leq ch^{1-n/t}.$$

This establishes the assertion. □

4.3.2 Control Discretization

We are now concerned with the error introduced by a discretization of the control space Q. We assume, from now on, that (Q.2) holds, that is, $B=\mathrm{Id}$ and $R(q)=\frac{1}{r}\|q\|_{L^r}^r$, and hence $R'(q)=|q|^{r-2}q$. Note that in the other case there is no point in discretizing Q.

Our analysis is based on the regularity result for the optimal control given in Corollary 3.7.

Theorem 4.3. *Let $(\bar{u},\bar{q})$ be the solution of (4.4), $r=\frac{r}{r-1}$, and let (Q.2) be fulfilled, then there exist constants $\gamma,\gamma'>0$ such that $\gamma+\gamma'\geq 1-n/t$, with*

$$\overline{q}\in W_0^{\gamma,r}\Omega\qquad\text{and}\qquad R'(\overline{q})\in W_0^{\gamma',r'}(\Omega). \tag{4.15}$$

Proof. From Lemma 3.5 we have that $R'(\overline{q})=-\frac{-1}{\alpha}\overline{z}\in W^{1-n/t-\varepsilon,t'}(\Omega)\subset W^{1-n/t-\varepsilon,r'}(\Omega)$. And hence from Corollary 3.7 we get that $\overline{q}\in W^{\gamma,p}$ with $\gamma=(1-n/t-\varepsilon)/(r-1)$ which shows the assertion. □

Although this regularity result is somewhat technical, and our proof uses information about the continuous adjoint system, we note that Theorem 4.5 only requires the regularity result itself which could, alternatively, be formulated as an assumption.

Before we state our main result, we first deduce an approximation property from the regularity result in Theorem 4.3.

Corollary 4.4. *There exists a constant c, independent of h such that*

$$\|\overline{q}-\Pi_h\overline{q}\|_{L^r}\leq ch^{\gamma}\quad\text{and}\quad\|R'(\overline{q})-\Pi_hR'(\overline{q})\|_{L^{r'}}\leq ch^{\gamma'}. \tag{4.16}$$

Proof. Stability of Π_h in L^p (compare (4.11)) provides, for example focusing on $\Pi_h\overline{q}$,

$$\|\overline{q} - \Pi_h\overline{q}\|_{L^r} \leq (1 + \|\Pi_h\|_{L(L^r,L^r)}) \inf_{q_h \in Q^h} \|\overline{q} - q_h\|_{L^r}.$$

Choosing a suitable quasi-interpolation operator for q_h, for example the Clément operator, gives the desired result. □

We are now ready to prove our main result.

Theorem 4.5. *Let $(\overline{q}, \overline{u}) \in Q \times V$ be the solution of (4.4) and $(\overline{q}_h^h, \overline{u}_h^h) \in Q^h \times V_h$ be the solution of (4.8), then*

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}_h^h, \overline{u}_h^h)\right| \leq \begin{cases} Ch^{\min(2\gamma, 1-n/t)}, & \text{if } Q^h = Q^h_{(1)}, \\ Ch^{1-n/t}, & \text{if } Q^h = Q^h_{(0)}. \end{cases} \tag{4.17}$$

Proof. As above, we set $\beta = 1 - n/t$ throughout this proof. Let $(\overline{q}^h, \overline{u}^h) \in Q^h \times V$ be the solution of the following auxiliary problem where only the control variable is discretized:

$$\text{Minimize } J(q^h, u^h) := \frac{1}{2}\|u^h - u^d\|_{L^2}^2 + R(q^h), \tag{4.18a}$$

$$\text{subject to } \begin{cases} (q^h, u^h) \in Q^h \times V, \\ (q^h, u^h) \text{ satisfies (4.2)}, \\ |\nabla u^h|^2 \ \leq \psi \text{ in } \overline{\Omega}. \end{cases} \tag{4.18b}$$

We will first show that

$$\left|J(\overline{q}, \overline{u}) - J(\overline{q}^h, \overline{u}^h)\right| \leq Ch^{\min(2\gamma,\beta)}. \tag{4.19}$$

Once this is established, we can repeat the proof of Theorem 4.1 verbatim to show that

$$|J(\overline{q}^h, \overline{u}^h) - J(\overline{q}_h^h, \overline{u}_h^h)| \leq Ch^{\beta}.$$

This is possible since all constants in this proof would only depend on the regularity of the triangulation and on $\|\overline{q}^h\|_{L^r}$. The fact that $\|\overline{q}^h\|_{L^r}$ remains bounded, as $h \to 0$, is immediately deduced from the fact that $J(\overline{q}^h, \overline{u}^h)$ converges and is therefore bounded itself. Combining the two estimates gives the desired result.

To establish (4.19) we proceed along the lines of the proof of Theorem 4.1 as well. Let $q^h = \Pi_h\overline{q}$ and let $u^h \in V$ solve the state equation (4.2) with right-hand side $q = q^h$, then, using our regularity assumptions on the state equation, Corollary 4.4, and the continuous embedding $W^{1,\infty}(\Omega) \subset W^{1+d/t+\varepsilon,t}(\Omega)$ for all $\varepsilon > 0$, we can estimate

$$\|u^h - \overline{u}\|_{1,\infty} \leq c\|u^h - \overline{u}\|_{1+n/t+\varepsilon,t} \leq c\|q^h - \overline{q}\|_{-1+n/t+\varepsilon,t} \leq ch^{1-n/t-\varepsilon+\gamma}. \tag{4.20}$$

Choosing $\varepsilon \leq \gamma$, we obtain $\|u^h - \overline{u}\|_{1,\infty} \leq ch^{\beta}$. Thus, setting

$$(\tilde{q}^h, \tilde{u}^h) = (1 - \tilde{c}h^{\beta})(q^h, u^h),$$

for $\tilde{c}$ sufficiently large, gives an admissible pair for (4.18) and we obtain

$$0 \leq J(\overline{q}^h, \overline{u}^h) - J(\overline{q}, \overline{u}) \leq J(\tilde{q}^h, \tilde{u}^h) - J(\overline{q}, \overline{u}).$$

Since J is differentiable as a mapping from $L^r(\Omega) \to L^{r'}(\Omega)$ (hence locally Lipschitz) it follows that

$$|J(\tilde{q}^h, \tilde{u}^h) - J(q^h, u^h)| \leq ch^\beta,$$

hence, we only need to bound the term $J(q^h, u^h) - J(\overline{q}, \overline{u})$ from above. Using convexity of J, we can estimate

$$\begin{aligned} J(\overline{q}, \overline{u}) &\geq J(q^h, u^h) + \langle J'(q^h, u^h), (\overline{q} - q^h, \overline{u} - u^h)\rangle \\ &= J(q^h, u^h) + \langle R'(q^h), \overline{q} - q^h\rangle + (u^h - u^d, \overline{u} - u^h) \end{aligned}$$

The term $(u^h - u^d, \overline{u} - u^h)$ is easily bounded by ch^β, using (4.20). In summary, we obtain

$$0 \leq J(\overline{q}^h, \overline{u}^h) - J(\overline{q}, \overline{u}) \leq J(\tilde{q}^h, \tilde{u}^h) - J(\overline{q}, \overline{u}) \leq \langle R'(q^h), q^h - \overline{q}\rangle + ch^\beta. \tag{4.21}$$

Up to this point, the proof is entirely independent of the choice of the control discretization.

If $Q^h = Q^h_{(0)}$ is the space of piecewise constant functions then $R'(q^h) = |q^h|^{r-2}q^h$ also belongs to Q^h, and since $q^h = \Pi_h \overline{q}$ it follows that $\langle R'(q^h), q^h - \overline{q}\rangle = 0$. This concludes the proof of (4.17) for the case $Q^h = Q^h_{(0)}$. We note that for precisely the same reason, namely that $R'(q^h) \in Q^h$, the analysis in (Günther and Hinze [76]) did not require regularity of the optimal control.

If $Q^h = Q^h_{(1)}$ is the space of linear (or bi- or tri-linear) functions then this argument is not valid. Instead, we estimate

$$\begin{aligned} \langle R'(q^h), q^h - \overline{q}\rangle &= \langle R'(q^h) - R'(\overline{q}), q^h - \overline{q}\rangle + \langle R'(\overline{q}), q^h - \overline{q}\rangle \\ &= \langle R'(q^h) - R'(\overline{q}), q^h - \overline{q}\rangle + \langle R'(\overline{q}) - \Pi_h R'(\overline{q}), q^h - \overline{q}\rangle \\ &\leq (\|R'(q^h) - R'(\overline{q})\|_{L^{r'}} + \|R'(\overline{q}) - \Pi_h R'(\overline{q})\|_{L^{r'}})\|q^h - \overline{q}\|_{L^r}. \end{aligned}$$

Using the fact that R' is differentiable (hence locally Lipschitz continuous) as well as Corollary 4.4, we finally obtain

$$\langle R'(q^h), q^h - \overline{q}\rangle \leq c(h^\gamma + h^{\gamma'})h^\gamma.$$

Since $\gamma + \gamma' \geq \beta$, we obtain the convergence rate $O(h^{\min(2\gamma,\beta)})$. This concludes the proof of the theorem. □

As before, the error estimate on the objective functional provides an error estimate for the primal variables.

Corollary 4.6. *Let $(\overline{q}, \overline{u}) \in Q \times V$ be the solution of (4.4), and let $(\overline{q}^h_h, \overline{u}^h_h) \in Q^h \times V_h$ be the solution of (4.8), then*

$$\|\overline{q} - \overline{q}^h_h\|_Q \leq ch^{a/r} \quad \text{and} \quad \|\overline{u} - \overline{u}^h_h\|_{L^2} \leq ch^{a/2},$$

where r is defined in (Q.2), and where $a = 1 - n/t$ if $Q^h = Q^h_{(0)}$, or $a = \min(2\gamma, 1 - n/t)$ if $Q^h = Q^h_{(1)}$.

Proof. We set $\overline{w} = (\overline{q}, \overline{u})$, and so forth. We split the error

$$\overline{w}_h^h - \overline{w} = (\overline{w}_h^h - \overline{w}^h) + (\overline{w}^h - \overline{w}),$$

where $\overline{w}^h$ is the solution of the auxiliary problem (4.18). The first contribution, $(\overline{w}_h^h - \overline{w}^h)$, can be estimated precisely as in the proof of Corollary 4.2, yielding

$$\|\overline{u}_h^h - \overline{u}^h\|_{L^2} \leq ch^{\frac{1-n/t}{2}} \quad \text{and} \quad \|\overline{q}_h^h - \overline{q}^h\|_{L^2} \leq ch^{\frac{1-n/t}{r}}.$$

To estimate $(\overline{w}^h - \overline{w})$ we employ again Clarkson's inequality (4.5) and get

$$\frac{1}{2}\Big\|\frac{1}{2}(\overline{u}^h - \overline{u})\Big\|_{L^2}^2 + R\Big(\frac{1}{2}(\overline{q}^h - \overline{q})\Big) \leq \frac{1}{2}J(\overline{w}^h) + \frac{1}{2}J(\overline{w}) - J\Big(\frac{1}{2}(\overline{w}^h + \overline{w})\Big).$$

Since $\overline{w}^h$ is admissible for the full problem (4.4), we have $J(\frac{1}{2}(\overline{w}^h + \overline{w})) \geq J(\overline{w})$ which gives

$$J(\frac{1}{2}(\overline{w}^h - \overline{w})) \leq \frac{1}{2}\big(J(\overline{w}^h) - J(\overline{w})\big) \leq ch^{\min(2\gamma, 1-n/t)},$$

where we also used (4.19). □

Remark 4.1. The analysis in the appendix shows that possible choices for the constants γ, γ' appearing in Theorem 4.3 and in the subsequence results are

$$\gamma = \frac{1 - n/t - \varepsilon}{r - 1} \quad \text{and} \quad \gamma' = 1 - n/t,$$

for any $\varepsilon > 0$. We have deliberately not included these explicit formulas in the convergence results above since we have no reason to believe that these estimates are optimal.

We note, however, that $2\gamma = \frac{2}{r-1}(1 - n/t - \varepsilon)$. Thus, if $n = 2$ then choosing $r < 3$ allows us to recover the rate $1 - n/t$. If $n = 3$ then choosing $r = 3 + \varepsilon$ gives $2\gamma = 1 - n/t - \varepsilon'$ for some $\varepsilon' > 0$ which tends to zero as $\varepsilon \to 0$.

4.4 Numerical Results

Here we will demonstrate our findings on a numerical example.

The computations in this section were done using the finite element toolkit Gascoigne (Gascoigne [65]) and the optimization toolbox RoDoBo (RoDoBo [126]). The computations were done using a barrier method of order six, see Section 5.1 for details. In order to generate the results a small barrier parameter $\gamma = 10^{-6}$ was chosen. In these examples this was sufficient to have dominant discretization error on all meshes under consideration.

Example with First Order Constraints an Known Solution The example is a slight modification from (Deckelnick et al. [54]) with known solution. The original problem reads as follows:

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2$$
$$\text{subject to } \begin{cases} (\nabla u, \nabla\varphi) = (q+f,\varphi) & \forall\,\varphi \in H_0^1(\Omega), \\ -2 \le q \le 2 & \text{a.e. in } \Omega, \\ |\nabla u|^2 \le 0.25 & \text{in } \overline{\Omega}, \\ (q,u) \in L^2(\Omega)\times H_0^1(\Omega). \end{cases}$$

Where $\alpha = 1$, the domain $\Omega = \{x \in \mathbb{R}^2 \,|\, |x| < 2\}$ and the data of the problem is

$$f = \begin{cases} 2 & |x| \le 1, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$u^d = \begin{cases} 0.25 + 0.5\ln(2) - 0.25|x|^2 & |x| \le 1, \\ 0.5\log(2) - 0.5\ln(|x|) & \text{otherwise}. \end{cases}$$

The exact solution satisfies $\overline{u} = u^d$ and $\overline{q} = \begin{cases} -1 & |x| \le 1, \\ 0 & \text{otherwise}, \end{cases}$ and the functional value is given as $J(\overline{q},\overline{u}) = \frac{\pi}{2}$.

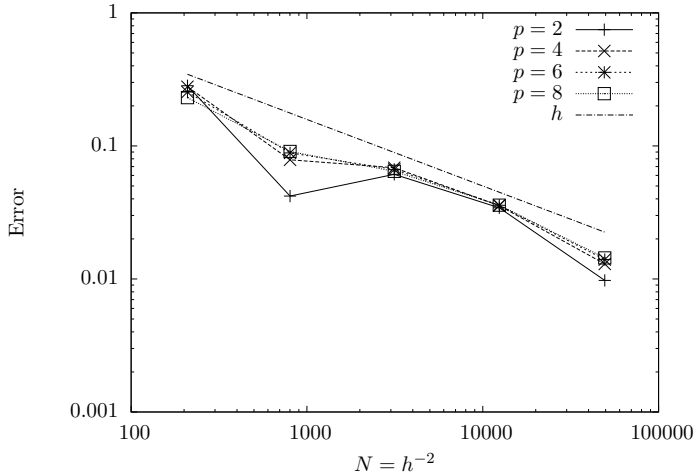


Figure 4.1: Error in J_p

From the corresponding KKT-System on immediately gets that the same solution also solves

$$\text{Minimize } J_p(q,u) := \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{p}\|q\|_{L^p}^p$$
$$\text{subject to } \begin{cases} (\nabla u, \nabla\varphi) = (q+f,\varphi) & \forall\,\varphi \in H_0^1(\Omega), \\ -2 \le q \le 2 & \text{a.e. in } \Omega, \\ |\nabla u|^2 \le 0.25 & \text{in } \overline{\Omega}, \\ (q,u) \in L^p(\Omega)\times H_0^1(\Omega), \end{cases}$$

for any $p \geq 2$. The optimal functional value is $J_p(\overline{q}, \overline{u}) = \frac{\pi}{p}$.

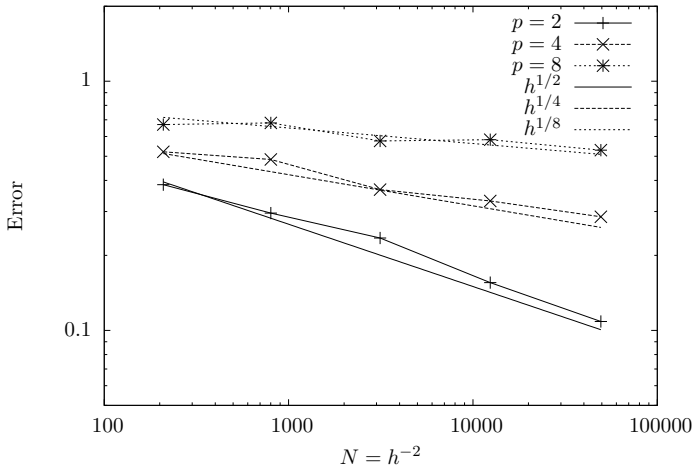


Figure 4.2: $\|\overline{q} - \overline{q}_h^h\|_{L^p}$ on a sequence of globally refined meshes

As it appears to be the most interesting case, we have considered Q_1 Finite Elements for the discretization of the control and the state variable. Then, considering that we know the exact solution $\overline{q}$, we can choose γ in Corollary 4.6 to be $\gamma = \frac{1}{2}$. In addition, $\overline{u}$ in $W^{2,t}(\Omega)$ for any $t \in [2, \infty]$ hence $\beta = 1 - n/t$ can be chosen as 1, see (Brenner and Scott [32], Corollary 8.1.12).

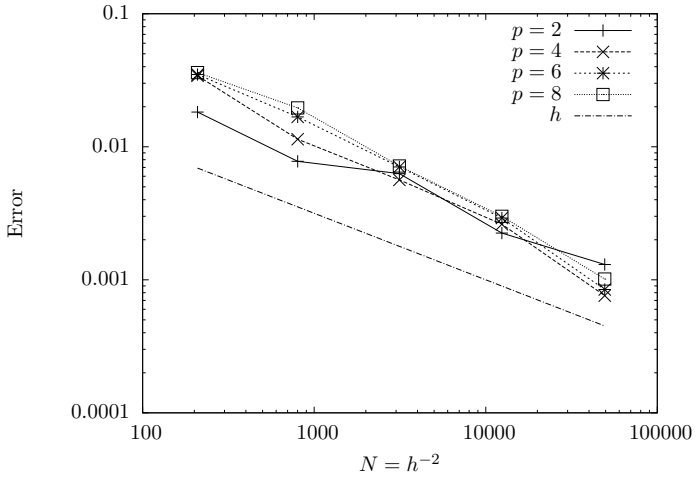


Figure 4.3: $\|\overline{u} - \overline{u}_h^h\|_{L^2}$ on a sequence of globally refined meshes

Hence by Theorem 4.5 we expect that the cost functional converges of order h independent of p which can also be seen in Figure 4.1, where the behavior of the absolute error in J_p under mesh refinement is depicted, and N denotes the number of nodes in the mesh. The element size h and the number of nodes N relate like $N \approx h^{-2}$. Then by Corollary 4.6 the

convergence of the control variable in L^p should be of order $h^{1/p}$ for the minimization of J_p. This can be seen from Figure 4.2 where this rate is recovered very well by this example.

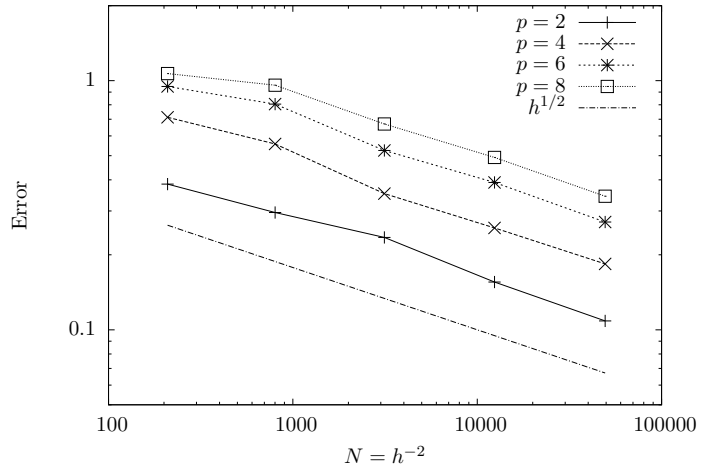


Figure 4.4: $\|\overline{q} - \overline{q}_h^h\|_{L^2}$ on a sequence of globally refined meshes

In addition, from Corollary 4.6, we would expect that the state variable is converging with order $h^{1/2}$ in L^2. Here however we can see from Figure 4.3 that the estimate for the state variable is apparently better than the predicted order. To explain this we note, that the error in the state variable can not only be obtained by convexity of J_p in Corollary 4.6, but also by the error in the control variable. Hence higher convergence rates of the control in weaker norms might account for this behavior. That this may be the case is indicated by the following numerical evidence Figure 4.4. Where we can see, that the control converges in fact with order $h^{1/2}$ independent of p in the L^2-norm.

5 Algorithms for State Constraints

In the previous chapter we obtained that there is actually a sequence of solutions to (4.1) which is converging to (2.5) as $h \to 0$. The next step is hence to consider algorithms that are capable of solving the given problems (4.1). We note that taking any algorithm from finite dimensional optimization, see, e.g., (Nocedal and Wright [119]), will obviously give the required solution, but may in general not be of optimal complexity as all dimensions are not really fixed quantities but tend to infinity. With this several constants in the complexity estimates, arising for instance from norm equivalence, may blow up.

The main point in showing some mesh independent behavior of the optimization algorithm is usually to show that in fact the algorithm could be 'applied' to the continuous problem.

For first-order state constraints a variety of methods have been proposed on the continuous level. Ranging from rather specialized methods, for instance (Mossino [117]) who proposed to solve a dual problem which in certain cases no longer has state constraints, to general purpose methods like penalty and barrier methods.

Later augmented Lagrangian techniques where used in (Bergounioux [18]) and further refined in (Bergounioux and Kunisch [19]) where Uzawa type methods were used to solve the resulting saddle point problems. The development of augmented Lagrangian methods lead to Moreau-Yosida Regularization (Bergounioux, Haddou, Hintermüller, and Kunisch [23], Hintermüller and Kunisch [81, 82]) for state constraints giving rise to a regularized primal-dual-active-set method for its solution. In terms of finite dimensional optimization it is a quadratic penalty function for the state constraints.

The direct application of a primal-dual-active-set method is possible in certain situations (Bergounioux and Kunisch [20]), however as it may be difficult or even impossible to find a control that actually generates a state with given active set we will not consider this approach here, although we will use it for the case of control constraints (Bergounioux, Ito, and Kunisch [22], Kunisch and Rösch [98]) where it is equivalent to a semi-smooth newton method (Hintermüller et al. [85]). For the same reason SQP-methods for state constrained problems will in general not be applicable, although, in certain situations, they can be analyzed in Banach spaces, see (Arada, Raymond, and Tröltzsch [6]).

Inspired by the fast convergence of the prima-dual-active-set method for control constraints so called Lavrentiev regularization for state constraints where proposed in (Meyer, Rösch, and Tröltzsch [115]). Here the state constraints are replaced by certain mixed control-state constraints which then lead to a control-constraint problem with a degenerate equation. This has been further analyzed in (Meyer, Prüfert, and Tröltzsch [116]) and in (Cherednichenko and Rösch [45]) for stability with respect to perturbations. In (Hintermüller, Tröltzsch, and Yousept [87]) mesh-independent convergence was shown. However as this method requires

that both state constraint and control space fit together, it is a rather specialized method which already introduces complications when the control is acting on the boundary, see e.g., (Tröltzsch and Yousept [143]).

An approach that is based on the idea of considering the optimization problem as a free boundary problem between the active and inactive set has been proposed in (Hintermüller and Ring [84]).

The other classical method for the solution of inequality constraint optimization problems are barrier methods. They are rather well understood in the case of control-constraints. For a primal method analysis has been done in (Weiser and Deuflhard [152]) for which superlinear convergence could be shown (Schiela and Weiser [134]), see also (Weiser, Gänzler, and Schiela [153]) for a path following algorithm. Similar results could be obtained for primal-dual methods (Ulbrich and Ulbrich [145], Weiser [151]).

When concerned with state constraints there as been some results concerning zero-order constraints for a primal barrier function by (Schiela and Weiser [133]). Later, this has been extended to rational barrier functions instead of the usual logarithmic barrier in (Schiela [129, 130]). In (Schiela [132]) a damping step was introduced, to improve convergence of Newton's method in comparison with the usual step length determination.

In this thesis we will use penalty and barrier methods for the solution of the state constrained optimization problem as they appear to be both relatively easy to implement and on the other hand sufficiently versatile to be applied to several situations.

As in the course of writing this thesis none of these methods where analyzed for first-order state constraints we will derive convergence of a primal barrier method for such constraints. Parts of these results are already published in (Schiela and Wollner [135]). We will not discuss the case of the quadratic penalty as this has been developed simultaneously by (Hintermüller and Kunisch [83]).

5.1 Barrier Methods for First Order State Constraints

The results that are shown here have been published in (Schiela and Wollner [135]). We remark that, in order to be in correspondence with usual notation used for barrier methods throughout this section we denote the barrier parameter by μ and to avoid confusion with the notation of measures we denote all measures by m.

5.1.1 Preliminaries

Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ and $\Omega^C \subseteq \Omega$ be a closed subset with non empty interior. In addition to the definitions of Section 2.2, we define the space of states U as a closed subspace of $C^1(\Omega^C) \times L^2(\Omega \setminus \Omega^C)$, which is clearly a Banach space, and let $W \subset U$ be a dense subspace of U. Consider $W = W^{2,t}(\Omega) \subset U = C^1(\Omega^C) \times L^2(\Omega \setminus \Omega^C)$ with $t > n$ for an example.

We specify the abstract equation (2.4) to be the following abstract linear partial differential equation in Z^*:

$$Au = Bq \tag{5.1}$$

where we require the following properties:

Assumption 5.1. Assume that $A\colon U \supset \operatorname{dom} A = W \to Z^*$ is densely defined and possesses a bounded inverse. Further let $B\colon Q \to Z^*$ be a continuous operator.

We will see later, in Lemma 5.1, that continuous invertibility of A is equivalent to closedness and bijectivity. The distinction between the state space U and the domain of definition W of A allows us to consider our optimal control problem in a convenient topological framework (the topology of U), while being able to model differential operators by A, which are only defined on a dense subspace W.

To define an optimal control problem, we specify an objective functional J with some basic regularity assumptions:

Assumption 5.2. Let $J = J_1 + J_2$. We assume that $J_1\colon U \to \mathbb{R}$ and $J_2\colon Q \to \mathbb{R}$ are lower semi-continuous, convex and Gâteaux differentiable. In addition, let J_1 be bounded from below and J_2 be strictly convex. Assume that the derivatives are uniformly bounded on bounded sets. This means that there exists a continuous $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\|J_1'(u)\|_{U^*} \leq g(\|u\|_U)$ and $\|J_2'(q)\|_{Q^*} \leq g(\|q\|_Q)$.

We now consider the following minimization problem

$$\text{Minimize } J(q,u) = J_1(u) + J_2(q), \tag{5.2a}$$

$$\text{subject to } \begin{cases} Au = Bq, \\ (q,u) \in Q^{\mathrm{ad}} \times W, \\ |\nabla u(x)|^2 \leq \psi(x) \quad \text{on } \Omega^C, \end{cases} \tag{5.2b}$$

where $\psi \in C(\Omega^C)$ with $\psi \geq \delta > 0$.

In order to ensure that there exists a solution we require that the following assumption holds:

Assumption 5.3. We assume that at least one of the following holds:

(1) Q^{ad} is bounded in Q.

(2) J_2 is coercive on Q.

For the discussion of interior point methods for the gradient constraint we require an additional property, which is of Slater type

Assumption 5.4. Assume there exists a feasible control $\breve{q} \in Q^{\mathrm{ad}}$, such that the corresponding state $\breve{u}$ given by $A\breve{u} = B\breve{q}$ is strictly feasible, that is, $|\nabla \breve{u}|^2 < \psi$.

We state the following basic continuity result, whose proof can be found, e.g., in (Schiela [130], Lemma A.1).

Lemma 5.1. *Let U be a Banach space. An operator $A\colon U \supset W \to Z^*$ has a continuous inverse if and only if A is closed and bijective.*

If Assumption 5.1 holds, then there exists a continuous "control-to-state" mapping

$$S\colon Q \to U, \quad S := A^{-1}B.$$

Using the Assumptions 5.1, 5.2, 5.3 and 5.4 it follows by standard arguments (coercivity, weak sequential compactness, convexity), see also Theorem 2.1, that (5.2) admits a unique solution $(\overline{q}, \overline{u}) \in Q^{\mathrm{ad}} \times W$.

For the discussion of the adjoint operator A^* of A we exploit density of W in U and reflexivity of Z. A^* possesses a domain of definition $\operatorname{dom} A^*$, given by

$$\operatorname{dom} A^* = \{z \in Z \,|\, \exists\, c_z \,:\, \langle Au, z\rangle \le c_z \|u\|_U \quad \forall\, u \in \operatorname{dom} A = W\}.$$

Because W is dense in U for each $z \in \operatorname{dom} A^*$ the linear functional $\langle A\cdot, z\rangle$ has a unique continuous extension to a functional on the whole space U. This defines a linear operator $A^*\colon Z \supset \operatorname{dom} A^* \to U^*$ and it holds

$$\langle u, A^* z\rangle = \langle Au, z\rangle \quad \forall\, u \in \operatorname{dom} A, z \in \operatorname{dom} A^*.$$

Lemma 5.2. *The operator A^* defined above has a continuous inverse, and it holds*

$$\left(A^{-1}\right)^* = (A^*)^{-1}. \tag{5.3}$$

Proof. Since Z^* is complete and A is surjective, we can apply (Goldberg [68], Theorem II.3.13), which states that A^* has a bounded inverse under these conditions. Hence, both $(A^{-1})^*$ and $(A^*)^{-1}$ exist, and by (Goldberg [68], Theorem II.3.9) they are equal. □

Examples Illustrating the Setting Let us apply our abstract framework to optimal control problems with PDEs. First we consider two variants of modeling an elliptic partial differential operator of second order: via the strong form and via the weak form. It will turn out that the strong form yields a more convenient representation of A^* and is thus preferable.

Example 5.1. [Second-Order Elliptic PDE in Strong Form] Let $\Omega^C = \overline{\Omega} \subset \mathbb{R}^n$, $U = C^1(\overline{\Omega}) \cap H_0^1(\Omega)$, $r > n$, and $Z = L^{r'}(\Omega)$ with $\frac{1}{r} + \frac{1}{r'} = 1$. Consider $A = -\Delta$ as a mapping from $\operatorname{dom} A = W = W^{2,r}(\Omega) \cap H_0^1(\Omega)$ to $L^r(\Omega)$. This means that A is a differential operator in strong form. We can write this as integral equation in the following form:

$$\langle Au, z\rangle = \int_\Omega -\Delta u z \, dx \quad \forall\, u \in W, z \in Z.$$

Assume that the boundary of $\Omega \subset \mathbb{R}^n$ is either of class $C^{1,1}$ or that $\Omega \subset \mathbb{R}^n$ with $n = 2$ is convex and has a polygonal boundary. Then there exists r with $n < r < \infty$ such that A is an isomorphism from W onto Z^*, see, e.g., (Gilbarg and Trudinger [67], Theorem 9.15) for the case of a $C^{1,1}$ boundary or (Grisvard [71]) for the polygonal case. In particular, A has a continuous inverse from Z^* onto W. By Sobolev embedding W is continuously embedded

into U and thus A^{-1} can also be defined as a continuous mapping from Z^* into U. Because W is dense in U the requirements on A from Assumption 5.1 are fulfilled.

A simple choice for the control space is $Q^{\mathrm{ad}} = Q = L^r(\Omega) = Z^*$. Then $B = \mathrm{Id}$ is a continuous operator. This corresponds to distributed control. As a second setting for the control we may consider $Q = \mathbb{R}^m$ and $f_i \in L^r(\Omega)$, $i = 1 \dots m$. Then the operator B defined by $Bq = \sum_{i=1}^m f_i q_i$ satisfies Assumption 5.1 on B.

In the case of distributed control a simple cost functional might be

$$J(q,u) = J_1(u) + J_2(q) = \frac{1}{2}\|u - u^d\|^2_{L^2(\Omega)} + \frac{1}{r}\|q\|^r_{L^r(\Omega)}.$$

with given $u^d \in L^2$, $r > n$. It is easily seen that J_2 is coercive on Q. Thus Assumption 5.3 is satisfied. By simple calculations Assumption 5.2 on J is verified.

Since the gradient bound ψ is assumed to be strictly positive, taking $\breve{q} = 0$ yields the required Slater condition from Assumption 5.4.

The adjoint operator $A^*\colon Z \supset \operatorname{dom} A^* \to U^*$ can be interpreted as a very weak form of the Laplace operator, i.e.

$$\langle u, A^* z\rangle = \langle Au, z\rangle = \int_\Omega -\Delta u z \, dx \quad \forall\, u \in W, z \in \operatorname{dom} A^*.$$

Lemma 5.2 already yields the continuous invertibility of A^*.

Example 5.2. [Second-Order Elliptic PDE in Weak Form] Let us discuss an alternative approach to Example 5.1: the weak form of the "same" elliptic operator. Usually one defines the differential operator $A = -\Delta\colon H^1_0(\Omega) \to H^{-1}(\Omega)$ by:

$$\langle Au, z\rangle = \int_\Omega \nabla u^T \nabla z \, dx \quad \forall\, z \in H^1_0(\Omega).$$

Our aim is to redefine the spaces for this operator such that Assumption 5.1 holds. To this end we have to restrict the image space from $H^{-1}(\Omega)$ to $L^{r'}(\Omega)^*$. Then the space W is given by

$$W = \left\{ u \in H^1_0 \,\middle|\, \int_\Omega \nabla u^T \nabla z \, dx \le c_u \|z\|_{L^{r'}} \quad \forall\, z \in H^1_0(\Omega) \right\}.$$

Observe that the integral in this expression is not defined for all $z \in L^{r'}$, but only for $z \in H^1(\Omega)$. However, if $u \in W$ then by definition of W it follows, that Au has a unique continuous extension to an element of $L^{r'}(\Omega)^*$. It is given canonically by

$$\langle Au, z\rangle = \lim_{\substack{z_k \in H^1_0, \\ z_k \to z \text{ in } L^{r'}}} \langle Au, z_k\rangle. \tag{5.4}$$

Under the same regularity assumptions as in Example 5.1 we obtain that $W \subset C^1(\overline{\Omega})$ and $\|u\|_{C^1} \le c\|Au\|_{(L^{r'})^*}$, thus Assumption 5.1 is fulfilled.

In spite of the complicated representation of A via (5.4), we may represent the equation $Au = f$ conveniently in the form

$$\int_\Omega \nabla u^T \nabla \varphi \, dx = \int_\Omega f\varphi \, dx \quad \forall \varphi \in H_0^1(\Omega) \tag{5.5}$$

via density.

However, since the linear functional Au is defined in $L^{r'}(\Omega)^*$ by continuous extension (5.4), the representation of the adjoint operator A^* is quite cumbersome. It is given by

$$\langle u, A^* z\rangle = \lim_{\substack{z_k \in H_0^1, \\ z_k \to z \text{ in } L^{r'}}} \langle Au, z_k\rangle = \lim_{\substack{z_k \in H_0^1, \\ z_k \to z \text{ in } L^{r'}}} \int_\Omega \nabla u^T \nabla z_k \, dx.$$

and has to be used in the adjoint PDE. In contrast to the weak formulation of the primal equation (5.5), where the limit formulation for the *test functions* can be dropped by density, now the limit formulation applies to elements of the *ansatz space*, and thus cannot be neglected. Continuous invertibility of A^*, which follows from our abstract considerations only applies to its correct representation. A naive formulation of the adjoint PDE would yield wrong results. This is the reason why we prefer the strong formulation for optimal control problems of second-order equations with gradient bounds.

Example 5.3. [Fourth-Order Elliptic PDE] As a different example we consider once again $\Omega^C = \overline{\Omega}$ but choose different spaces. Let $U = \{v \in C^1(\overline{\Omega}) \mid v(x) = |\nabla v(x)| = 0 \; \forall x \in \partial\Omega\}$, $Z = W_0^{2,r'}(\Omega)$. We consider the biharmonic operator $A = \Delta^2$ as a mapping from $\operatorname{dom} A = W = W_0^{2,r}(\Omega)$ to $Z^* = W^{-2,r}(\Omega)$ with $\frac{1}{r} + \frac{1}{r'} = 1$.

Assume that the domain $\Omega \subset \mathbb{R}^2$ is convex with polygonal boundary, then it is well known (Blum and Rannacher [27], Theorem 2) that A has a continuous inverse from Z^* onto W if $r > 2$ is chosen small enough. As it has already been remarked for $2 < r < \infty$ the embedding from W into U exists and is dense.

Note that in this case both dual and primal operator can be represented by

$$\langle Au, z\rangle = \langle u, A^* z\rangle = \int_\Omega \Delta u \Delta z \, dx \quad \forall \, u \in W_0^{2,r}(\Omega), \, z \in W_0^{2,r'}(\Omega).$$

By the choice $Q = L^2(\Omega)$ with B the embedding from L^2 into $W^{-2,r}$ we see that Assumption 5.1 is fulfilled.

5.1.2 Barrier Functional and its Subdifferentiability

In this section we are concerned with the analysis of barrier functionals for the problem under consideration. We proceed as in (Schiela [130]):

Definition 5.1. For $\kappa \geq 1$ and $\mu > 0$ we define barrier functions l of order κ by

$$\begin{aligned} l(v;\mu;\kappa) \;&:\; \mathbb{R}_+ \to \overline{\mathbb{R}}, \\ l(v;\mu;\kappa) &:= \begin{cases} -\mu \ln(v) & \kappa = 1, \\ \frac{\mu^\kappa}{(\kappa-1)v^{\kappa-1}} & \kappa > 1. \end{cases} \end{aligned}$$

We extend their domain of definition to $\mathbb{R}$ by setting $l(v;\mu;\kappa) = \infty$ for $x \leq 0$. We denote the pointwise derivative of $l(v;\mu;\kappa)$ by $l'(v;\mu;\kappa)$ if $v > 0$. This yields

$$l'(v;\mu;\kappa) = \frac{-\mu^\kappa}{v^\kappa}.$$

With this we define a barrier functional b for the constraint $v \geq 0$ by:

$$\begin{aligned} b(\,\cdot\,;\mu;\kappa)\colon C(\Omega^C) &\to \overline{\mathbb{R}}, \\ v &\mapsto \int_{\Omega^C} l(v(x);\mu;\kappa)\; dx. \end{aligned}$$

Its formal derivative $b'(v,\mu;\kappa) \in C(\Omega^C)^*$, is defined as

$$\langle b'(v;\mu;\kappa), \delta v\rangle := \int_{\Omega^C} l'(v(x);\mu;\kappa)\delta v(x)\; dx$$

if the right-hand side exists.

Obviously, if $0 < \varepsilon \leq v \in C(\Omega^C)$, then b is differentiable with respect to v, and b' is the Fréchet derivative of b. If $v(x) = 0$, for some $x \in C(\Omega^C)$, then the situation is more involved, and techniques of sub-differential calculus have to be applied, for a recent survey on this field see (Borwein and Zhu [30]).

In contrast to the case of state constraints, we may not use $\psi = 0$ to ease notation. This is due to the fact that in this case $u = 0$ would be the only admissible state. Therefore we introduce the following shifted barrier functional:

Definition 5.2. We define the barrier functional for the constraint $|\nabla u|^2 \leq \psi$ on a compact set $\Omega^C \subseteq \overline{\Omega}$ by

$$\begin{aligned} b_\psi(\,\cdot\,;\mu;\kappa)\colon C^1(\Omega^C) &\to \overline{\mathbb{R}}, \\ u &\mapsto b_\psi(u;\mu;\kappa) := b(\psi - |\nabla u|^2;\mu;\kappa). \end{aligned} \tag{5.6}$$

In several cases we are only interested in a barrier functional of a fixed given order κ, and sometimes even for only one fixed value of μ, in those cases we write $b(\cdot;\mu)$ or even $b(\cdot)$ if no confusion can occur.

Lemma 5.3. *The barrier functional b_ψ defined in* (5.6) *is well defined, convex, and lower-semicontinuous.*

Proof. By (Schiela [130], Proposition 4.3) the outer function $b(\cdot;\mu;\kappa)$ is well defined and lower semi-continuous. Since the inner function $\psi - |\nabla u|^2$ is well defined and continuous on U, the composition of both functions is well defined and lower semi-continuous.

Moreover, we know that $b(\cdot;\mu;\kappa)$ is convex and monotonically decreasing. Further, the mapping $T(u) := \psi - |\nabla u|^2$ is pointwise concave. With these properties we can proof convexity of $b_\psi = b \circ T$ by the following computation which holds for every x in Ω^C:

$$l\big(T(\lambda u + (1-\lambda)\tilde{u})\big)(x) \leq l\big(\lambda T(u) + (1-\lambda)T(\tilde{u})\big)(x) \leq \lambda l\big(T(u)\big)(x) + (1-\lambda) l\big(T(\tilde{u})\big)(x).$$

By monotonicity of the integral we obtain that b_ψ is convex. □

We approach subdifferentiability of $b_\psi = b \circ (\psi - |\nabla \cdot|^2)$ via the following chain rule.

Lemma 5.4. *Let U, V be Banach spaces, $f : V \to \overline{\mathbb{R}}$ be a convex, lower-semicontinuous function, and $T : U \to V$ a continuously differentiable mapping with first derivative T'. Assume that the composite mapping $f \circ T$ is also convex.*

Let u be given and let $T'(u)$ be bounded. Assume that there is $\breve{u} \in U$, such that f is bounded above in a neighbourhood of $T(u) + T'(u)\breve{u}$. Then

$$\partial(f \circ T)(u) = (T'(u))^* \partial f(T(u)). \tag{5.7}$$

Proof. This is a slight extension of the well known chain rule of convex analysis (cf., (Ekeland and Témam [58], Prop. I.5.7)), which is, however, hard to find in the literature. We thus derive this result from a more general theorem from non-smooth analysis due to Clarke and Rockafellar (cf., (Clarke [47], Thm. 2.9.9) or (Rockafellar [124], Thm. 3)). Although the construction of the corresponding generalized differential is rather complicated in general, it reduces to the convex subdifferential in the case of convex functions (cf., (Rockafellar [125], Thm. 5)).

First of all, we may assume that $f(T(u))$ is finite. Otherwise, $\partial(f \circ T)(u) = \partial(f(T(u)) = \emptyset$ holds trivially, because $\partial g(u) := \emptyset$ in case $g(u) = +\infty$ for every convex function g.

Otherwise we may argue as in (Rockafellar [124], Cor. 1), which shows that the chain rule (Rockafellar [124], Thm. 3) can be applied to show our assertion under the additional assumption that T is linear. However, inspection of its (short) proof shows that the same argumentation is still true in the case that T is "strictly differentiable" at u and $f \circ T$ is convex, as long as $\breve{u}$ exists that satisfies our assumptions. Now, the Corollary subsequent to (Clarke [47], Prop. 2.2.1) asserts that "strict differentiability" is implied by continuous differentiability, and our assertion follows. □

Remark 5.1. Lemma 5.3 and Lemma 5.4 are also useful in the context of pointwise state constraints of the form $g(u(x), x) \leq 0$, if g is convex and differentiable in u.

With the help of this lemma we can now characterize the subdifferential for barrier functionals with respect to gradient bounds in terms of the known subdifferential of a barrier functional in $C(\Omega^C)$, see (Schiela [130]).

Proposition 5.5. *Assume that $\psi \geq \delta > 0$. Define*

$$\begin{aligned} b_\psi \,:\, & C^1(\Omega^C) \to \overline{\mathbb{R}} \\ & u \mapsto b(\psi - |\nabla u|^2) \end{aligned}$$

as in Definition 5.2. Then the subdifferential $\partial b_\psi(u)$ has the following representation:

$$\partial b_\psi(u) = (-2\nabla u^T \nabla)^* \partial b(\psi - |\nabla u|^2). \tag{5.8}$$

This means, $\tilde{m} \in \partial b_\psi(u)$, if and only if there is $m \in \partial b(\psi - |\nabla u|^2)$, such that

$$\langle \delta u, \tilde{m} \rangle_{C^1(\Omega^C), C^1(\Omega^C)^*} = -2\langle \nabla u^T \nabla \delta u, m \rangle_{C(\Omega^C), C(\Omega^C)^*} \quad \forall \delta u \in C^1(\Omega^C).$$

If u is strictly feasible, then $m = b'(\psi - |\nabla u|^2)$.

Proof. Let $T\colon C^1(\Omega^C) \to C(\Omega^C)$ be defined by $T(u) := \psi - |\nabla u|^2$. Obviously, the mapping $\psi - |\nabla u|^2\colon C^1(\Omega^C) \to C(\Omega^C)$ is continuously differentiable with bounded derivative $(T'(u)\delta u)(x) = -2(\nabla u(x))^T \nabla \delta u(x)$.

We are going to apply Lemma 5.4 to the function $b_\psi\colon U \to \overline{\mathbb{R}}$, $b_\psi(u) = b \circ T$. By (Schiela [130], Lemma 3.2), b is convex and lower semi-continuous and by Lemma 5.3 b_ψ is convex, too. Setting $\breve{u} := -0.5u$, we have $T'(u)\breve{u} = |\nabla u|^2$, and $\tilde{v} := T(u) + T'(u)\breve{u} = \psi$. Since $\psi \geq \delta > 0$, b is bounded from above in a $C(\Omega^C)$-neighbourhood of $\tilde{v}$. Hence, Lemma 5.4 can be applied and yields our representation formula (5.8). Finally, (Schiela [130], Prop. 3.5) shows that $\partial b(v) = \{b'(v)\}$ if v is strictly feasible. □

The barrier functional b_ψ can also be analyzed on closed subspaces $\tilde{U}$ of $C^1(\Omega^C)$. To this end let $E : \tilde{U} \to C^1(\Omega^C)$ be the continuous embedding operator. Then its adjoint $E^* : C^1(\Omega^C)^* \to \tilde{U}^*$ is the restriction operator for linear functionals. If $\breve{u}$ in Assumption 5.4 can be chosen from $\tilde{U}$, then the chain-rule of convex analysis applied to $b_\psi \circ E$ yields a characterization of the subdifferential of the restriction of b_ψ to $\tilde{U}$ as restriction of the subdifferential:

$$\partial(b_\psi \circ E)(u) = E^* \partial b_\psi(Eu).$$

Closed subspaces of $C^1(\Omega^C)$ may for example be spaces that incorporate Dirichlet boundary conditions on $\Omega^C \cap \Omega$ or finite dimensional subspaces.

5.1.3 Minimizers of Barrier Problems

With the preparations made in the previous sections we will now show that there exists a unique solution for the barrier problem, and later on some first-order necessary conditions that are fulfilled by these.

Theorem 5.6. *(Existence of Solutions to Barrier Problems)*
Let Assumption 5.1—Assumption 5.4 be fulfilled. Then the Problem

$$\begin{aligned} &\text{Minimize } J_\mu(q,u) := J(q,u) + b_\psi(u;\mu), \\ &\text{subject to } \begin{cases} Au = Bq, \\ (q,u) \in Q^{ad} \times W, \end{cases} \end{aligned} \tag{5.9}$$

admits a unique minimizer (q_μ, u_μ)*. Moreover* u_μ *is strictly feasible almost everywhere in* Ω^C*.*

Proof. The proof is almost analog to the one for Theorem 2.1 as the possible value ∞ of J doesn't complicate the proof.

By Assumption 5.4 $J_\mu(\breve{q}, \breve{u}) < \infty$. Further, J_μ is bounded from below by Assumption 5.3, by the required lower bound for J_1, and because b_ψ is bounded from below, since ψ is bounded above.

Taking a minimizing sequence $(q_k, u_k) = (q_k, Sq_k)$ (recall that $S = A^{-1}B$ is continuous by Lemma 5.1), we obtain from Assumption 5.3 that w.l.o.g. q_k converges weakly to some $q_\mu \in Q^{\mathrm{ad}}$. From Lemma 5.1 together with Assumption 5.1 we obtain that w.l.o.g. the sequence u_k converges to u_μ weakly in W where u fulfills equation (5.1). From lower semi-continuity of J and b_ψ (compare Lemma 5.3), we obtain that the limit (q_μ, u_μ) solves (5.9) and since $J_\mu(q_\mu, u_\mu) < \infty$ it follows that u is strictly feasible almost everywhere in Ω_C.

Furthermore, the limit (q_μ, u_μ) is unique, since J is strictly convex with respect to the control variable, and the mapping $q_\mu \mapsto u_\mu$ is injective. □

The next theorem shows that the regularity of the solutions doesn't degenerate as $\mu \to 0$:

Theorem 5.7. *Let Assumption 5.1—Assumption 5.4 be fulfilled. Then for every* $\mu_0 > 0$ *the solutions* $(q_\mu, u_\mu) \in Q \times W$ *of* (5.9) *are uniformly bounded on* $(0, \mu_0]$*.*

Proof. First note that due to Lemma 5.1 in combination with Assumption 5.1 it is sufficient to show that q_μ is uniformly bounded. To see this we note that, cf., (Schiela [130]),

$$J_\mu(q_\mu, u_\mu) \leq J_\mu(q_{\mu_0}, u_{\mu_0}) \leq J_{\mu_0}(q_{\mu_0}, u_{\mu_0}).$$

From $J(q_\mu, u_\mu) \leq J_\mu(q_\mu, u_\mu)$ together with Assumption 5.3 we obtain, that q_μ is bounded, which concludes the proof. □

Usually, if $W \subset C^1(\Omega^C)$ the state satisfies the additional regularity $W \subset C^{1,\beta}(\Omega^C) \subset C^1(\Omega^C)$. This means the gradients are even Hölder continuous of order β. Then we obtain for a sufficiently high order κ of the barrier method that the state is in fact strictly feasible everywhere in Ω_C, as the following theorem shows.

Theorem 5.8. *Let* $\Omega_C \subset \mathbb{R}^d$ *be compact satisfying a cone property (see (Adams and Fournier [1], Def. 4.6)) and for some* $\beta \in (0,1)$ *let* $\psi \in C^{0,\beta}(\Omega_C)$ *be given. Let Assumption 5.1—Assumption 5.4 be satisfied. If the state has the additional regularity* $u_\mu \in C^{1,\beta}(\Omega_C)$*, then for* $\kappa - 1 > \frac{n}{\beta}$ *the state* u_μ *is strictly feasible in* Ω_C*.*

Proof. By Theorem 5.6 we obtain $0 \leq \psi - |\nabla u_\mu|^2 \in C^{0,\beta}(\Omega_C)$. From (Schiela [130], Lemma 7.1) we obtain that therefore $(\psi - |\nabla u_\mu|^2)^{-1} \in C(\Omega_C)$ which concludes the proof. □

We are now prepared to derive first-order necessary conditions for the minimizer of the barrier problem (5.9).

Theorem 5.9. *Let the Assumption 5.1—Assumption 5.4 be fulfilled. Then $(q_\mu, u_\mu) \in Q^{ad} \times U$ is a solution to (5.9) if and only if there exist $m_\mu \in \partial b(\psi - |\nabla u_\mu|^2) \subset C(\Omega^C)^*$ and $z_\mu \in Z$, $q_\mu^* \in Q^*$ such that the following holds:*

$$Au_\mu = Bq_\mu \qquad \text{in } Z^* \tag{5.10a}$$

$$A^* z_\mu = J_1'(u_\mu) + (-2(\nabla u_\mu)^T \nabla)^* m_\mu \qquad \text{in } U^* \tag{5.10b}$$

$$J_2'(q_\mu) = -B^* z_\mu - q_\mu^* \qquad \text{in } Q^* \tag{5.10c}$$

$$\langle q - q_\mu, q_\mu^* \rangle \leq 0 \qquad \forall\, q \in Q^{ad} \tag{5.10d}$$

Proof. We consider the following minimization problem where we omit the dependence on the parameter μ:

$$\min_{q \in Q} F(q) = \chi_{Q^{\text{ad}}}(q) + \hat{J}_\mu(q) := \chi_{Q^{\text{ad}}}(q) + J_\mu(q, Sq) \tag{5.11}$$

where $\chi_{Q^{\text{ad}}}$ is the indicator function for the admissible set of the controls, and S is the control to state mapping defined by (5.1). Clearly $(q_\mu, u_\mu) = (q_\mu, Sq_\mu)$ is a solution to (5.9) if and only if q_μ is a solution to (5.11), which is in turn equivalent to $0 \in \partial F(q_\mu)$. In order to utilize this we will split the subdifferential by the sum-rule of convex analysis:

$$\partial F(q_\mu) = \partial(\chi_{Q^{\text{ad}}})(q_\mu) + \partial \hat{J}_\mu(q_\mu). \tag{5.12}$$

For its application note that Assumption 5.4 asserts the existence of a point

$$\breve{q} \in \operatorname{dom} \chi_{Q^{\text{ad}}} \cap \operatorname{dom} \hat{J}_\mu$$

such that $\hat{J}_\mu$ is continuous in $\breve{q}$. In addition the function $\chi_{Q^{\text{ad}}}$ is convex and lower semicontinuous, thus it coincides with its Γ-*regularization* (Ekeland and Témam [58], Chapter I, Prop. 3.1). We can therefore apply the sum-rule of convex analysis, cf., (Ekeland and Témam [58], Chapter I, Prop. 5.6) to obtain (5.12).

Since $\hat{J}$ is continuous in q_μ we obtain by the same argument that:

$$\partial \hat{J}_\mu(q_\mu) = \partial \hat{J}(q_\mu) + \partial(b_\psi \circ S)(q_\mu)$$

where we recall the definition $b_\psi(u) = b(\psi - |\nabla u|^2)$. Now, we note that

$$\hat{J}(q) = J \circ (1, S)(q)$$

with the linear mapping

$$(1, S) \colon Q \to Q \times U, \qquad q \mapsto (q, Sq).$$

Together with Assumption 5.4 we are able to apply the linear chain rule and obtain

$$\partial \hat{J}(q_\mu) = (1, S^*)\partial J(q_\mu, u_\mu),$$
$$\partial(b_\psi \circ S)(q_\mu) = S^* \partial b_\psi(S q_\mu).$$

Inserting the representation for the subdifferential of the barrier function b_ψ in Proposition 5.5 our computations have shown so far that

$$0 \in \partial(\chi_{Q^{\mathrm{ad}}})(q_\mu) + (1, S^*)\partial J(q_\mu, u_\mu) + S^*(-2(\nabla u_\mu)^T \nabla)^* \partial b(\psi - |\nabla u_\mu|^2) \tag{5.13}$$

is equivalent to (q_μ, u_μ) being a solution to (5.9). Since the cost functional is differentiable we obtain, cf., (Ekeland and Témam [58], Chapter I, Prop. 5.3):

$$\partial J(q_\mu, u_\mu) = \{J_1'(u_\mu) + J_2'(q_\mu)\}.$$

Equation (5.13) means there exist $q_\mu^* \in \partial \chi_{Q^{\mathrm{ad}}}(q_\mu)$, and $m_\mu \in \partial b(\psi - |\nabla u_\mu|^2)$ such that

$$0 = q_\mu^* + J_2'(q_\mu) + S^*\big(J_1'(u_\mu) + (-2(\nabla u_\mu)^T \nabla)^* m_\mu\big) \in Q^*. \tag{5.14}$$

Note that $S^* = (A^{-1}B)^* = B^*(A^{-1})^* = B^*(A^*)^{-1}$, where $A^*\colon Z \supset \operatorname{dom} A^* \to U^*$ is well defined with continuous inverse due to Lemma 5.2. Define

$$z_\mu = (A^*)^{-1}\big(J_1'(u_\mu) + (-2(\nabla u_\mu)^T \nabla)^* m_\mu\big). \tag{5.15}$$

Then $z_\mu \in \operatorname{dom} A^* \subset Z$ and satisfies (5.10b) by definition. Equation (5.10c) now follows from (5.14). Further note that q_μ^* fulfills, see, e.g., (Ekeland and Témam [58], Chapter I, Prop. 5.1),

$$\sup_{q \in Q^{\mathrm{ad}}} \langle q, q_\mu^* \rangle = \langle q_\mu, q_\mu^* \rangle \tag{5.16}$$

which is equivalent to (5.10d). □

Example 5.4. Let us apply our abstract results to Example 5.1 in the case of distributed control ($B = \mathrm{Id}$). Using the notation from there the first-order optimality conditions have the following form. Let (q_μ, u_μ) be a solution to (5.9), then there exists $z_\mu \in Z$, $m_\mu \in \partial b(\psi - |\nabla u_\mu|^2; \mu)$ such that:

$$\int_\Omega -\Delta u_\mu \, \varphi \, dx = \int_\Omega q_\mu \varphi \, dx \qquad \forall\, \varphi \in Z, \tag{5.17a}$$
$$\int_\Omega -\Delta \varphi \, z_\mu \, dx = \int_\Omega (u_\mu - u^d)\varphi \, dx - 2\int_{\overline{\Omega}} (\nabla u_\mu)^T \nabla \varphi \, dm_\mu \qquad \forall\, \varphi \in W, \tag{5.17b}$$
$$|q_\mu|^{r-2} q_\mu = -z_\mu \qquad \text{a.e. in } \Omega. \tag{5.17c}$$

For a discussion of the first two equations and in particular the representation of A and A^* we refer to Example 5.1. The barrier gradient m_μ is an element of $\partial b(u_\mu; \mu; \kappa)$, and a measure in general. If u_μ is strictly feasible, which can usually be guaranteed a priori by a proper choice of the order κ, then $m_\mu = b'(y; \mu; \kappa)$ and thus a function, cf., (Schiela [130], Prop. 4.6).

Equation (5.17c) holds pointwise almost everywhere since it holds in L^r. The multiplier q_μ^* does not appear due to the fact that $Q^{\mathrm{ad}} = Q$.

After having studied the necessary optimality conditions we will now discuss the behavior of the dual variables. The hard part is showing the boundedness of the measure obtained from the subdifferential of the barrier functional.

Theorem 5.10. *Let the assumptions of Theorem 5.9 be fulfilled. Then for each $\mu_0 > 0$*

$$\sup_{\mu\in(0,\mu_0]} \|m_\mu\|_{C(\Omega^C)^*} \leq C.$$

Proof. Let (q_μ, u_μ) be the solution to (5.9) and $(\breve{q}, \breve{u})$ be a Slater point, e.g., let $\psi - |\nabla\breve{u}|^2 \geq \tau > 0$. Then, following (Schiela [130]), we multiply (5.10b) with $\delta u = u_\mu - \breve{u}$ and (5.10c) with $\delta q = q_\mu - \breve{q}$ and obtain

$$\begin{aligned} 0 &= \langle \delta u, -A^* z_\mu + J_1'(u_\mu) + (-2(\nabla u_\mu)^T\nabla)^* m_\mu\rangle + \langle \delta q, J_2'(q_\mu) + B^* z_\mu + q_\mu^*\rangle \\ &= \langle \delta u, J_1'(u_\mu) + (-2(\nabla u_\mu)^T\nabla)^* m_\mu\rangle + \langle \delta q, J_2'(q_\mu) + q_\mu^*\rangle + \langle A\delta u - B\delta q, -z_\mu\rangle. \end{aligned}$$

As $(\delta q, \delta u)$ fulfills the state equation (5.1) this simplifies to

$$0 = \langle \delta u, J_1'(u_\mu)\rangle + \langle \delta q, J_2'(q_\mu)\rangle - 2\langle (\nabla u_\mu)^T\nabla\delta u, m_\mu\rangle + \langle \delta q, q_\mu^*\rangle. \tag{5.18}$$

From the uniform boundedness of the primal variable, see Theorem 5.7 together with Assumption 5.2, we obtain that

$$|\langle \delta u, J_1'(u_\mu)\rangle + \langle \delta q, J_2'(q_\mu)\rangle| \leq C$$

with a constant C independent of μ. Inserting this estimate into (5.18) yields

$$|-2\langle (\nabla u_\mu)^T\nabla\delta u, m_\mu\rangle + \langle \delta q, q_\mu^*\rangle| \leq C. \tag{5.19}$$

We would like to split this into the sum of the absolute values. To do so we will show that both terms have essentially the same sign. First, we now define the 'almost' active set

$$\mathcal{A} = \{x \in \Omega^C \,|\, \psi - |\nabla u_\mu|^2 \leq 0.5\tau\}.$$

This is motivated by the fact, see (Schiela [130], Corollary 3.6), that

$$|\langle (\nabla u_\mu)^T\nabla\delta u, m_\mu|_{\Omega^C\setminus\mathcal{A}}\rangle| \leq \|m_\mu\|_{L^1(\Omega^C\setminus\mathcal{A})}\|(\nabla u_\mu)^T\nabla\delta u\|_{L^\infty} \leq C. \tag{5.20}$$

Thus it remains to take a look at the behavior of $\langle m_\mu|_{\mathcal{A}}, (\nabla u_\mu)^T\nabla\delta u\rangle$. We will now show that $0 < c \leq (\nabla u_\mu)^T\nabla\delta u$ holds on $\mathcal{A}$. For this we apply Young's inequality and obtain

$$2|(\nabla u_\mu)^T\nabla\breve{u}| \leq |\nabla u_\mu|^2 + |\nabla\breve{u}|^2 \leq |\nabla u_\mu|^2 + \psi - \tau$$

leading to the following pointwise estimate on $\mathcal{A}$:

$$0.25\tau \leq 0.5(|\nabla u_\mu|^2 - \psi) + 0.5\tau \leq -(\nabla u_\mu)^T\nabla\breve{u} \leq |\nabla u_\mu|^2 - (\nabla u_\mu)^T\nabla\breve{u} \leq (\nabla u_\mu)^T\nabla\delta u.$$

From (Schiela [130], Prop. 4.6) we obtain that $m_\mu \leq 0$ as a measure thus leading to

$$-2\langle (\nabla u_\mu)^T\nabla\delta u, m_\mu|_{\mathcal{A}}\rangle \geq 0.$$

Now we take a look on (5.10d) to see that $\langle q_\mu - \breve{q}, q_\mu^* \rangle \geq 0$. Together with (5.20) we obtain from (5.19) that

$$|\langle (\nabla u_\mu)^T \nabla \delta u, m_\mu|_{\mathcal{A}} \rangle| \leq C.$$

Finally, we note that due to $m_\mu \leq 0$ the following holds:

$$\langle (\nabla u_\mu)^T \nabla \delta u, m_\mu|_{\mathcal{A}} \rangle \leq \min_{\mathcal{A}} \big((\nabla u_\mu)^T \nabla \delta u \big) \langle 1, m_\mu|_{\mathcal{A}} \rangle \leq -\frac{\tau}{4} \|m_\mu\|_{C(\mathcal{A})^*}.$$

This implies

$$\|m_\mu\|_{C(\mathcal{A})^*} \leq \frac{4}{\tau} |\langle (\nabla u_\mu)^T \nabla \delta u, m_\mu|_{\mathcal{A}} \rangle| \leq C$$

and completes the proof. □

Corollary 5.11. *Under the Assumption 5.1—Assumption 5.4 the following holds for every given $\mu_0 > 0$:*

$$\sup_{\mu \in (0,\mu_0]} \|z_\mu\|_Z \leq C,$$
$$\sup_{\mu \in (0,\mu_0]} \|q_\mu^*\|_{Q^*} \leq C.$$

Proof. First we note that the right-hand side of (5.10b) is bounded due to Assumption 5.2, boundedness of u_μ, m_μ, and continuity of $((\nabla u_\mu)^T \nabla)^* \colon C(\Omega^C)^* \to U^*$. The bound for z_μ follows from the boundedness of the right-hand side of (5.10b) and continuity of $(A^*)^{-1}$. The bound for q_μ^* then follows from the bound on z_μ and q_μ using (5.10c) and Assumption 5.2 and continuity of B^*. □

5.1.4 Properties of the Central Path

We will now show convergence of the cost functional with rate μ.

Theorem 5.12. *Let Assumption 5.1—Assumption 5.4 be fulfilled, and (q_μ, u_μ) be a solution of the barrier problem (5.9) for $\mu > 0$. Then the following holds for the minimizer $(\overline{q}, \overline{u})$ of (5.2):*

$$J(q_\mu, u_\mu) \leq J(\overline{q}, \overline{u}) + C\mu. \tag{5.21}$$

Proof. The proof follows the lines of (Schiela [130], Lemma 6.1), however since we consider nonlinear constraints on the gradient of the states we have to modify the argumentation concerning the multiplier coming from the subdifferential of the barrier functional.

From the proof of Theorem 5.9 together with the relation

$$\partial b(\psi - |\nabla u_\mu|^2; \mu; \kappa) = \mu^\kappa \partial b(\psi - |\nabla u_\mu|^2; 1; \kappa),$$

cf., (Ekeland and Témam [58], Chaper I, (5.21)), we obtain that there exists $m \in \partial b(\psi - |\nabla u_\mu|^2; 1)$ and $\varphi \in \partial \chi_{Q^{\mathrm{ad}}}(q_\mu) + \partial \hat{J}(q_\mu) = \partial(\chi_{Q^{\mathrm{ad}}} + \hat{J})(q_\mu)$ such that:

$$\varphi - 2\mu^\kappa S^*((\nabla u_\mu)^T \nabla)^* m = 0.$$

This shows that

$$2\mu^\kappa S^*((\nabla u_\mu)^T\nabla)^* m \in \partial(\chi_{Q^{\mathrm{ad}}} + \hat{J})(q_\mu).$$

From convexity of $\chi_{Q^{\mathrm{ad}}} + \hat{J}$ we obtain that for every $l \in \partial(\chi_{Q^{\mathrm{ad}}} + \hat{J})(q_\mu)$ the following holds:

$$\hat{J}(q_\mu) \le \hat{J}(\overline{q}) + \langle l, q_\mu - \overline{q}\rangle.$$

Applied to $2\mu^\kappa S^*((\nabla u_\mu)^T\nabla)^* m$ we obtain:

$$J(q_\mu, u_\mu) \le J(\overline{q}, \overline{u}) + 2\mu^\kappa \langle m, (\nabla u_\mu)^T\nabla(u_\mu - \overline{u})\rangle.$$

Since b is monotonically decreasing, the measure m is negative, cf., (Schiela [130], Prop. 4.6). Thus we can estimate further

$$2\mu^\kappa \langle m, (\nabla u_\mu)^T\nabla(u_\mu - \overline{u})\rangle \le 2\mu^\kappa \langle m|_{\Omega_S}, (\nabla u_\mu)^T\nabla(u_\mu - \overline{u})\rangle$$

where we define $\Omega_S := \{x \in \Omega^C \,|\, (\nabla u_\mu)^T\nabla(u_\mu - \overline{u}) < 0\}$. From Cauchy-Schwarz inequality it follows that $|\nabla u_\mu(x)| < |\nabla \overline{u}(x)| \le \psi(x)$ on Ω_S and thus $\Omega_S \subset \{x \in \Omega^C \,|\, |\nabla u_\mu|^2 < \psi\}$. Hence we obtain from (Schiela [130], Prop. 4.6.)

$$2\mu^\kappa \langle m|_{\Omega_S}, \nabla u_\mu \nabla(u_\mu - \overline{u})\rangle = -2\int_{\Omega_S} \frac{\mu^\kappa}{(\psi - |\nabla u_\mu|^2)^\kappa}(\nabla u_\mu)^T\nabla(u_\mu - \overline{u})\, dx.$$

From $(\nabla u_\mu)^T\nabla\overline{u} \le |\nabla u_\mu|\,|\nabla\overline{u}| \le \psi$ we see that

$$\frac{-(\nabla u_\mu)^T\nabla(u_\mu - \overline{u})}{\psi - |\nabla u_\mu|^2} = \frac{(\nabla u_\mu)^T\nabla\overline{u} - |\nabla u_\mu|^2}{\psi - |\nabla u_\mu|^2} \le 1$$

and thus

$$2\mu^\kappa \langle m|_{\Omega_S}, (\nabla u_\mu)^T\nabla(u_\mu - \overline{u})\rangle \le 2\mu \int_{\Omega_S} \frac{\mu^{\kappa-1}}{(\psi - |\nabla u_\mu|^2)^{\kappa-1}}\, dx. \tag{5.22}$$

From Theorem 5.10 and boundedness of the domain Ω^C we obtain for the function $f := \mu/(\psi - |\nabla u_\mu|^2)$ that

$$\|f^{\kappa-1}\|_{L^1(\Omega^C)}^{1/(\kappa-1)} = \|f\|_{L^{\kappa-1}(\Omega^C)} \le C\,\|f\|_{L^\kappa(\Omega^C)} = C\,\|f^\kappa\|_{L^1(\Omega^C)}^{1/\kappa} \le C.$$

Thus the integral on the right-hand side of (5.22) is bounded independent of μ. Hence the assertion follows. □

Theorem 5.13. *Let $\mu > 0$, (q_μ, u_μ) be a solution to the barrier problem (5.9) and $(\overline{q}, \overline{u})$ be the solution to the minimization problem (5.2). Further assume that there exist $c > 0$, $p \ge 2$ and a norm $\|\cdot\|$ such that*

$$c\|q_1 - q_2\|^p \le J_2(q_1) + J_2(q_2) - 2J_2\left(\frac{q_1 + q_2}{2}\right).$$

Then the following estimate holds:

$$\|q_\mu - \overline{q}\| = O(\mu^{1/p}). \tag{5.23}$$

Proof. By assumption and convexity of J_1 the following proves the assertion

$$\begin{aligned} c\|q_\mu - \overline{q}\|^p &\leq J_2(q_\mu) + J_2(\overline{q}) - 2J_2\left(\frac{q_\mu + \overline{q}}{2}\right) \\ &\leq J(q_\mu, u_\mu) + J(\overline{q}, \overline{u}) - 2J((q_\mu + \overline{q})/2, (u_\mu + \overline{u})/2) \\ &\leq J(q_\mu, u_\mu) + J(\overline{q}, \overline{u}) - 2J(\overline{q}, \overline{u}) = O(\mu). \end{aligned}$$

□

Remark 5.2. By an analogous assumption on J_1 a similar result for the state u_μ can be obtained. In addition, if $\|\cdot\|$ is stronger than $\|\cdot\|_Q$ the convergence of u_μ in U (with the same rate $O(\mu^{1/p})$) follows by continuity of S.

Example 5.5. We finally return to Example 5.1. We apply the Clarkson inequality (Clarkson [48], Theorem 2 (3)) for L^r-spaces with $r > 2$, which yields

$$\|\frac{f-g}{2}\|_{L^r}^r \leq \frac{1}{2}\|f\|_{L^r}^r + \frac{1}{2}\|g\|_{L^r}^r - \|\frac{f+g}{2}\|_{L^r}^r$$

from this we see that $\|q\|_{L^r}^r$ matches the assumption of Theorem 5.13 with $p = r$.

With the same techniques as in Theorem 5.12 it is possible to show for $\mu_0 > \mu > 0$ that $J_\mu(q_{\mu_0}, u_{\mu_0}) \leq J_\mu(q_\mu, u_\mu) + C(\mu_0 - \mu)$. Then continuity of the central path follows via Theorem 5.13.

5.1.5 Numerical Results

Here we will demonstrate our findings on three numerical examples, corresponding to Example 5.1 and Example 5.3. First we will discuss an example already considered in the literature with a second order PDE. As the convergence rate in this example exceeds our expectations we consider another example for this setting, but this time without a constructed solution. Finally, we will consider a generic optimal control problem with a fourth order PDE. The results are computed using the Finite Element Toolkit Gascoigne (Gascoigne [65]) and the Optimization Toolbox RoDoBo (RoDoBo [126]). In all examples we choose the order of the barrier method $\kappa = 6$.

Example with Second Order PDE First we will consider an example corresponding to Example 5.1. For this purpose we consider an example from (Deckelnick et al. [54]) with known solution. The problem reads as follows:

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2$$

$$\text{subject to } \begin{cases} (\nabla u, \nabla \varphi) = (q + f, \varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ -2 \leq q \leq 2 & \text{a.e. in } \Omega, \\ |\nabla u|^2 \leq 0.25 & \text{in } \overline{\Omega}, \\ u \in H_0^1(\Omega). \end{cases}$$

Where $\alpha = 1$, the domain $\Omega = \{x \in \mathbb{R}^2 \,|\, |x| < 2\}$ and the data of the problem is

$$f = \begin{cases} 2 & |x| \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$u^d = \begin{cases} 0.25 + 0.5\ln(2) - 0.25|x|^2 & |x| \leq 1, \\ 0.5\log(2) - 0.5\ln(|x|) & \text{otherwise}. \end{cases}$$

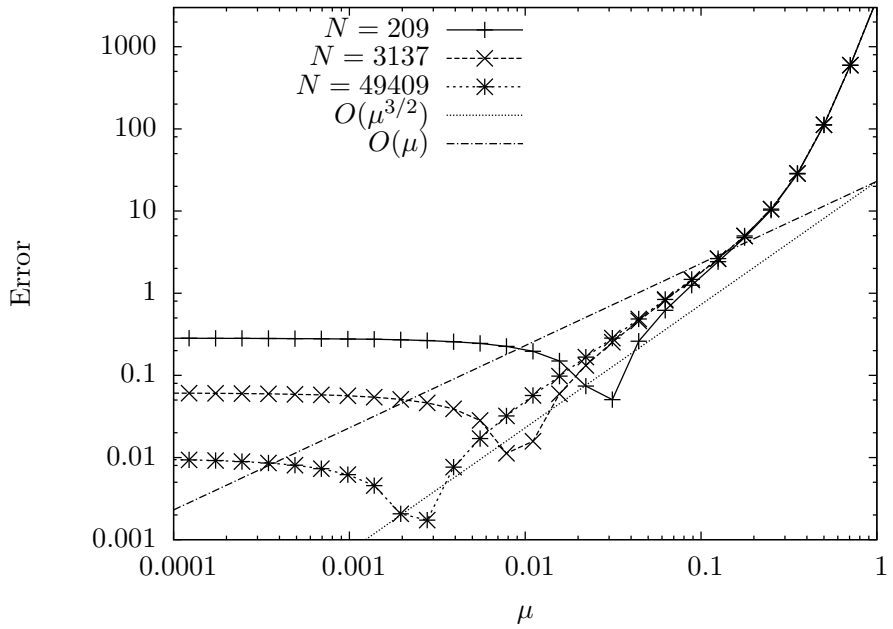


Figure 5.1: Error in the cost functional vs. barrier parameter μ on different meshes

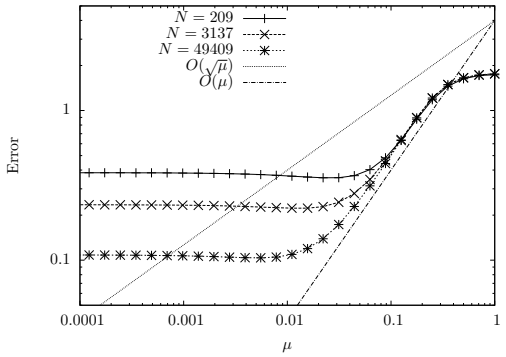
The exact solution satisfies $\overline{u} = u^d$,

$$\overline{q} = \begin{cases} -1 & |x| \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

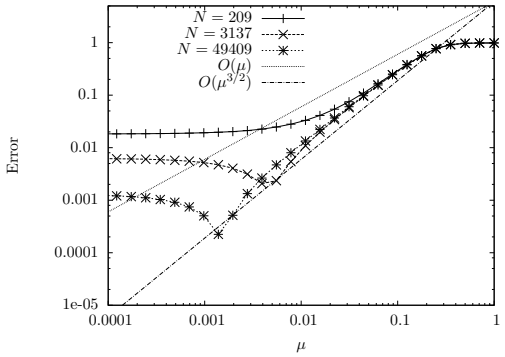
and the functional value is given as $J(\overline{q}, \overline{u}) = \frac{\pi}{2}$.

For the computation we have chosen an initial $\mu = 1.0$ and then successively reduced μ by $\sqrt{2}$ until $\mu < 10^{-4}$. The barrier subproblems were solved by a Newton's method in the control space which has been globalized using a line-search technique, as provided by RoDoBo. In our test problems, strictly feasible starting values were easy to obtain by taking $\breve{q} = -f$.

In Figure 5.1 we have depicted the convergence of the functional value. Here we can see, that after an initial phase the functional value is converging with an approximate order $O(\mu^{3/2})$ before it stabilizes at the value of the discretization error. The intermediate kink in the transition between regularization and discretization error is due to cancellation between the two error components.



(a) L^2-Error of the control variable vs. barrier parameter μ



(b) L^2-Error of the state variable vs. barrier parameter μ

Figure 5.2: Convergence behavior of the primal variables on different meshes

In Figure 5.2 we can see the convergence behavior of the primal variables. We see that the control variable is in fact converging with order μ instead of the predicted $\sqrt{\mu}$. The state variable is converging with approximately the same speed as the functional value, namely of order $O(\mu^{3/2})$, where we can see once again the cancellation in the transition between regularization and discretization error.

This rate of convergence exceeds our theoretical findings. In order to determine whether this is an exceptional case, caused by the specific construction of the example, or if our theory can be refined we consider an other example with a more generic structure.

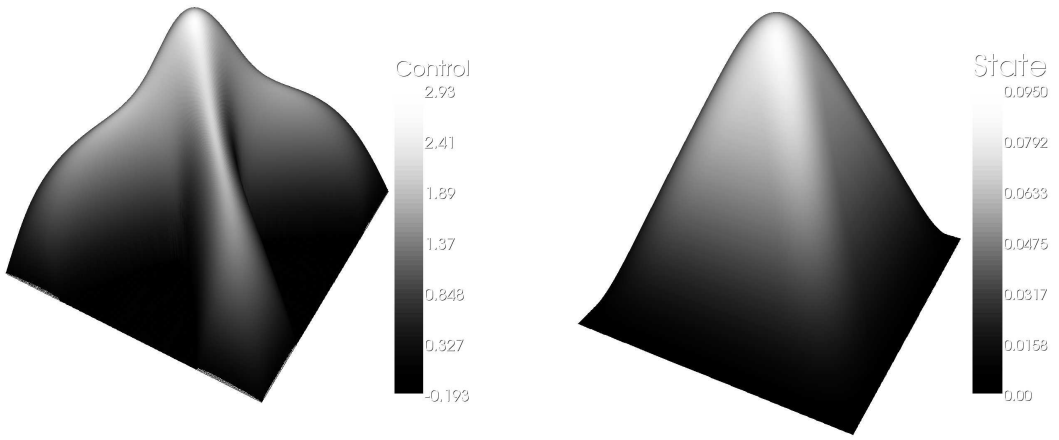
A Second Example with Unknown Solution In order to have a more generic example we remove the untypical inactive bound imposed in the previous example. The problem reads as follows:

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{3}\|q\|^3_{L^3(\Omega)}$$
$$\text{subject to } \begin{cases} (\nabla u, \nabla\varphi) = (q,\varphi) & \forall\,\varphi \in H^1_0(\Omega), \\ |\nabla u|^2 \le 0.05 & \text{in } \overline{\Omega}, \\ u \in H^1_0(\Omega). \end{cases}$$

We choose the desired state $u^d = \sin(\pi x)\sin(\pi y)$ and remark that u^d is infeasible with respect to the state constraint, e.g., in contrast to the previous example $\overline{u} - u^d \neq 0$. The Tikhonov parameter is chosen as $\alpha = 10^{-3}$, and the domain is given as $\Omega = (0,1)^2$.

In order to obtain a functional value for comparison we used a global uniform refinement with a total of 66049 vertices and $\mu = 5 \cdot 10^{-6}$.

In order to get an impression of the solution we depict the solution variables in Figure 5.3. We see that the state in Figure 5.3b is almost a pyramid, especially it has very flat surfaces. Correspondingly the control in Figure 5.3a exhibits very step gradients, coming from the measure on the boundary of the active set of the state constraint.



(a) Control (b) State

Figure 5.3: Optimal State and Control

We are now considering the convergence behavior of the cost functional for several values of μ. As in the preceeding example we begin with a moderate value of $\mu = 0.1$ and then successively reduce μ by a factor of $\sqrt{2}$. The computation is stopped on each mesh once the discretization error is reached. The results are depicted in Figure 5.4. We can immediately

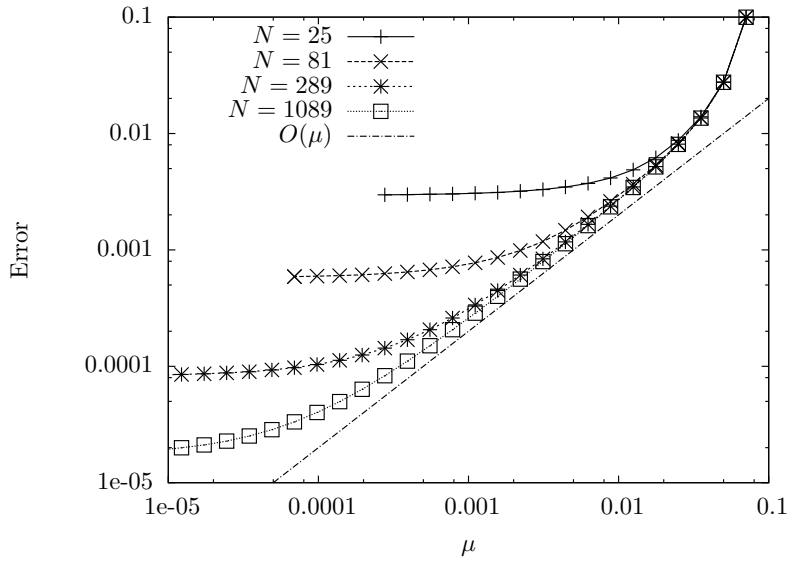


Figure 5.4: Error in the cost functional vs. barrier parameter μ on different meshes

see, that the convergence in the cost functional is as predicted by our analysis of order μ. Hence we conclude that the rates observed in the previous example are not the typical behavior but rather due to the specific problem structure. However we remark that the observed convergence order in the previous example also necessitates an a posteriori estimate of the actual error, as the a priori given convergence rate might be non-optimal.

Example with Fourth Order PDE We will now consider an example corresponding to Example 5.3. Hence we consider the following optimization problem

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2$$
$$\text{subject to } \begin{cases} (\Delta u, \Delta\varphi) = (q,\varphi) & \forall\, \varphi \in H_0^2(\Omega), \\ |\nabla u|^2 \le 0.04 & \text{in } \overline{\Omega}, \\ u \in H_0^2(\Omega). \end{cases}$$

We choose $\alpha = 10^{-3}$, the domain $\Omega = (-1,1)^2 \subset \mathbb{R}^2$ and

$$u^d = (x^2-1)^2(y^2-1)^2.$$

For the discretization of the state equation we consider a mixed finite element method, e.g., we consider $\sigma := \nabla u$ as an independent variable. Its continuous formulation is for given $q \in L^2(\Omega)$: Find $(\sigma, u) \in H^1(\Omega) \times H_0^1(\Omega)$ such that:

$$\begin{aligned} (\sigma,\varphi) + (\nabla u, \nabla\varphi) &= 0 && \forall\, \varphi \in H^1(\Omega) \\ (\nabla\sigma, \nabla\varphi) &= (q,\varphi) && \forall\, \varphi \in H_0^1(\Omega) \end{aligned}$$

which is discretized using conforming Q^1 finite elements. For details on this discretization see (Ciarlet and Raviart [46]) and (Reichmann [123]) for an implementation. In Figure 5.5

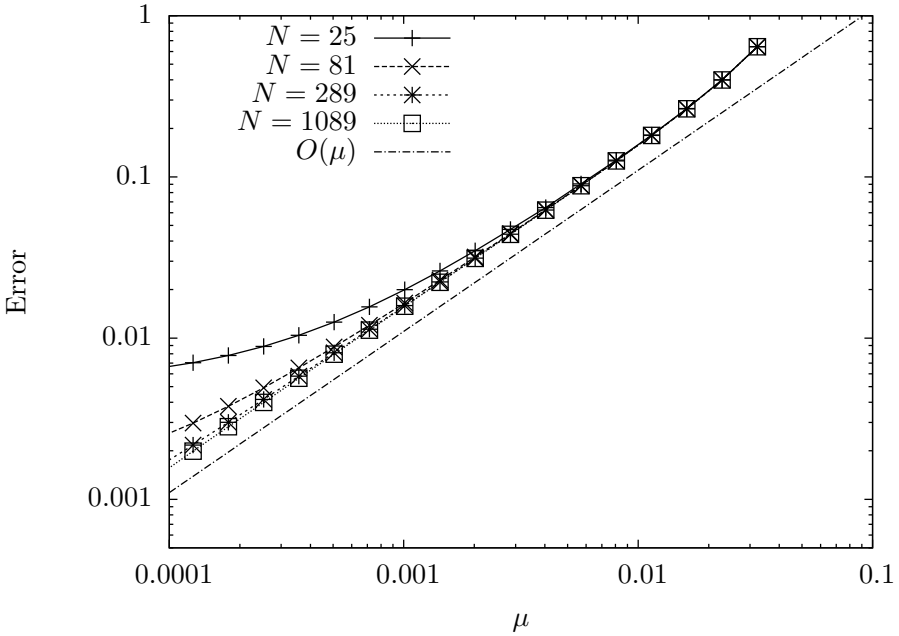


Figure 5.5: Error in the cost functional vs. barrier parameter μ on different meshes

we made a series of computations on different globally refined meshes, where N denotes the number of nodes in the mesh. For these computations the barrier parameter μ was initialized as 0.03 on each mesh and then successively decreased by a factor of $\sqrt{2}$ until it reached a

value lower than 10^{-4}. The choice of the initial μ was motivated by the previous example where an initial phase with very fast convergence was observed.

We can clearly see the predicted order of convergence of the cost functional. As in this example the exact solution is unknown we used a reference value obtained on a mesh with 10^6 nodes and a value $\mu = 10^{-6}$. The approximate functional value is 0.286619. Here we can clearly see the predicted oder of convergence namely $O(\mu)$.

5.2 Formal KKT-Conditions for Solution to Regularized Problems

In the previous Section 5.1 we have derived a convergence analysis and necessary optimality conditions for a minimization problem with first-order state constraints. We will now proceed and formally state the general first-order necessary conditions for approximate solutions to (2.5) using both barrier or penalty methods.

In order to take care of only on parameter we choose $\gamma > 0$ for both barrier and penalty methods where in both cases we consider $\gamma \to \infty$ to be the case corresponding to (2.5). Hence we have $\mu = 1/\gamma$ in the notation of Section 5.1.

We begin by stating the approximate problems to (2.5). To this end we define the barrier term $B_\gamma(\,\cdot\,)\colon W \to \mathbb{R} \cup \infty$ and the penalty term $P_\gamma\colon W \to \mathbb{R}$ using the following definition:

$$B_\gamma(u) := b(-g(u, \nabla u); 1/\gamma) \tag{5.24a}$$

$$P_\gamma(u) := \frac{\gamma}{2} \|(g(u, \nabla u))^+\|^2_{L^2(\Omega^C)} \tag{5.24b}$$

With this we are able to consider the following abstract regularized problems where we search for a pair $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q^{\mathrm{ad}} \times V$ which satisfies (2.4) and is a solution to either the barrier problem

$$\text{Minimize } J_\gamma(q, u) = J(q, u) + B_\gamma(u) \tag{5.25}$$

or the penalty problem

$$\text{Minimize } J_\gamma(q, u) = J(q, u) + P_\gamma(u). \tag{5.26}$$

Then we can formally state the corresponding first-order necessary conditions.

We begin with the barrier problem where we will assume for convenience that the barrier solution $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q^{\mathrm{ad}} \times V$ of (5.25) is strictly feasible, e.g., $g(u, \nabla u) < 0$, then there exists $\overline{\mu}_\gamma \in \{-l'(-g(\overline{u}_\gamma, \nabla \overline{u}_\gamma); 1/\gamma)\}$, $\overline{z}_\gamma \in V \cap Z$ such that

$$a(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) = (f, \varphi) \qquad \forall\ \varphi \in V, \tag{5.27a}$$

$$a'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi, \overline{z}_\gamma) = J'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) + (g'(\overline{u}_\gamma, \nabla \overline{u}_\gamma)(\varphi), \overline{\mu}_\gamma)_{\Omega^C} \qquad \forall\ \varphi \in V, \tag{5.27b}$$

$$J'_q(\overline{q}_\gamma, \overline{u}_\gamma)(\delta q - \overline{q}_\gamma) \geq a_q(\overline{q}_\gamma, \overline{u}_\gamma)(\delta q - \overline{q}_\gamma, \overline{z}_\gamma) \qquad \forall\, \delta q \in Q^{\mathrm{ad}}. \tag{5.27c}$$

Where l is given by Definition 5.1. The proof for linear equations and zero-order state constraints can for instance be found in (Schiela [130]) for first-order state constraints in Section 5.1 or (Schiela and Wollner [135]).

Let us now consider a solution $(\overline{q}_\gamma, \overline{u}_\gamma) \in Q^{\mathrm{ad}} \times V$ of (5.26). Then there exists $\overline{\mu}_\gamma = \gamma g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)^+$ and $\overline{z}_\gamma \in V \cap Z$ such that

$$a(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) = (f, \varphi) \qquad \forall\ \varphi \in V, \tag{5.28a}$$

$$a'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi, \overline{z}_\gamma) = J'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) + (g'(\overline{u}_\gamma, \nabla \overline{u}_\gamma)(\varphi), \overline{\mu}_\gamma)_{\Omega^C} \qquad \forall\ \varphi \in V, \tag{5.28b}$$

$$J'_q(\overline{q}_\gamma, \overline{u}_\gamma)(\delta q - \overline{q}_\gamma) \geq a_q(\overline{q}_\gamma, \overline{u}_\gamma)(\delta q - \overline{q}_\gamma, \overline{z}_\gamma) \qquad \forall\, \delta q \in Q^{\mathrm{ad}}, \tag{5.28c}$$

which has been shown for linear equations and zero and first-order constraints in (Hintermüller and Kunisch [83]).

In order to have that this is a reasonable regularization we assume that the solutions $(\overline{q}_\gamma, \overline{u}_\gamma)$ to (5.25) and (5.26) converge to the solution $(\overline{q}, \overline{u})$ of (2.5) if $\gamma \to \infty$. This has been shown in the previously mentioned articles (Hintermüller and Kunisch [83], Schiela [130], Schiela and Wollner [135]) for a linear 'control to state' mapping.

5.3 Discretization

In this section we discuss finite element discretization of the optimization problem (5.25) or (5.26).

To keep the following sections simple we restrain ourself to the case of problems where H^1-conforming finite elements are satisfactory. However the ideas can be adapted to other problems.

Let $\mathcal{T}_h$ be a triangulation (mesh) of the computational domain Ω consisting of closed elements K which are either triangles or quadrilaterals. The straight parts which make up the boundary ∂K of a cell K are called *faces*. The mesh parameter h is defined as a cell-wise constant function by setting $h\big|_K = h_K$ and h_K is the diameter of K. The mesh $\mathcal{T}_h$ is assumed to be shape regular. In order to ease the mesh refinement we allow the cells to have nodes, which lie on midpoints of faces of neighboring cells. But at most one of such *hanging nodes* is permitted per face.

On the mesh $\mathcal{T}_h$ we define a finite element space $V_h \subset V$ consisting of linear or bilinear shape functions, see, e.g., (Eriksson, Estep, Hansbo, and Johnson [61]) or (Brenner and Scott [32]). The case of hanging nodes requires some additional remarks. There are no degrees of freedom corresponding to these irregular nodes and therefore the value of the finite element function is determined by point-wise interpolation. This implies continuity and therefore global conformity.

For the discretization of the optimization problem (5.25) or (5.26) we introduce an additional finite dimensional subspace $Q_h \subset Q$ of the control space. Depending on the concrete situation there are different possible ways to choose the space Q_h. It is reasonable to set $Q_h = Q$ if Q is finite dimensional. In the case where the control variable is a distributed function on the computational domain Ω, i.e., $Q = L^2(\Omega)$, one may choose Q_h analog to V_h or consider Q_h as a space of cell-wise constant functions on the mesh $\mathcal{T}_h$. On the other hand if the control is acting on the boundary, i.e., $Q = L^2(\partial\Omega)$, we choose Q_h as traces of functions in V_h or as face-wise constant functions.

We denote a basis of Q_h by

$$\mathcal{B} = \{\psi_i\}, \text{ with } \psi_i \geq 0, \sum_i \psi_i = 1, \max_{x\in\omega} \psi_i(x) = 1. \tag{5.29}$$

The discrete admissible set Q_h^{ad} is defined as:

$$Q_h^{\text{ad}} = Q_h \cap Q^{\text{ad}},$$

where we assume that $Q_h^{\text{ad}} \neq \emptyset$. and the discretized optimization problem is formulated as follows:

$$\text{Minimize } J_\gamma(q_h, u_h), \quad u_h \in V_h,\, q_h \in Q_h^{\text{ad}}, \tag{5.30}$$

subject to

$$a(q_h, u_h)(v_h) = f(v_h) \quad \forall v_h \in V_h. \tag{5.31}$$

Here J_γ is defined either as (5.25) or (5.26). Assuming that first-order necessary conditions can be obtained they read as follows:

Let $(\overline{q}_\gamma^h, \overline{u}_\gamma^h) \in Q_h^{\text{ad}} \times V_h$ be a solution to (5.30), then there exists $\overline{z}_\gamma^h \in V_h$ and $\overline{\mu}_\gamma^h$ such that

$$a(\overline{q}_\gamma^h, \overline{u}_\gamma^h)(\varphi) = (f, \varphi^h) \qquad \forall\ \varphi^h \in V_h, \tag{5.32a}$$

$$a'_u(\overline{q}_\gamma^h, \overline{u}_\gamma^h)(\varphi^h, \overline{z}_\gamma^h) = J'_u(\overline{q}_\gamma^h, \overline{u}_\gamma^h)(\varphi^h) + (g'(\overline{u}_\gamma^h, \nabla\overline{u}_\gamma^h)(\varphi^h), \overline{\mu}_\gamma^h)_{\Omega^C} \qquad \forall\ \varphi^h \in V_h, \tag{5.32b}$$

$$J'_q(\overline{q}_\gamma^h, \overline{u}_\gamma^h)(\delta q^h - \overline{q}_\gamma^h) \geq a_q(\overline{q}_\gamma^h, \overline{u}_\gamma^h)(\delta q^h - \overline{q}_\gamma^h, \overline{z}_\gamma^h) \qquad \forall\, \delta q^h \in Q_h^{\text{ad}}. \tag{5.32c}$$

In the case of a barrier formulation (5.25) it holds

$$\overline{\mu}_\gamma^h = -l'(-g(\overline{u}_\gamma^h, \nabla\overline{u}_\gamma^h); 1/\gamma)$$

whereas in the case of a penalty formulation (5.26) the Lagrange multiplier is given by

$$\overline{\mu}_\gamma^h = \gamma g(u_\gamma^h, \nabla u_\gamma^h)^+.$$

For a priori error analysis with these choices for the discrete spaces we refer to the references at the beginning of Chapter 4. All of them deal with a direct discretization of the state constraints, e.g., they consider the limiting case $\gamma \to \infty$.

As the solution for $\gamma \to \infty$ tends to lead to ill-conditioned linear systems in the process of the computation of a descend direction, see, e.g., (Lootsma [103], Murray [118]) this seems to be unreasonable to compute. A way to overcome this difficulty is in balancing mesh size h and regularization parameter γ such that the error contributions are equilibrated. To do so one can either use a priori information (Hintermüller and Hinze [79], Hinze and Schiela [89]) or more efficiently using an a posteriori error estimate, see (Wollner [157]) or Chapter 6.

6 A Posteriori Error Estimates

In this chapter we will discuss a posteriori error estimation for control and state constrained problems.

As we have seen in the previous Chapter 5 we can conveniently consider control constrained problems for the adaptive procedure, as long as we can guarantee that the error introduced by the elimination of the state constraint using barrier or penalty methods is sufficiently small.

There are mainly two approaches to adaptive finite elements, one that is concerned with the a posteriori estimation of the discretization error in natural norms, which is well developed in the context of partial differential equations, see for instance the surveys (Ainsworth and Oden [3], Babuška and Strouboulis [7], Verfürth [146]). In articles (Gaevskaya, Hoppe, Iliash, and Kieweg [64], Hintermüller, Hoppe, Iliash, and Kieweg [86], Hoppe, Iliash, Iyyunni, and Sweilam [92], Li, Liu, Ma, and Tang [99], Liu and Yan [102]) the authors provide a posteriori error estimates for elliptic optimal control problems with distributed or Neumann control subject to box constraints. In (Gaevskaya et al. [64]) convergence of an adaptive algorithm for a control constrained optimal control problem is shown. For state constraint optimal control and a posteriori error estimation for norms of the solution some analysis has been done in (Hoppe and Kieweg [91]).

Secondly error estimation and mesh adaptation can also be guided by the error in a given functional the so called 'quantity of interest' going back to (Becker and Rannacher [11, 12], Eriksson, Estep, Hansbo, and Johnson [60]). The concept was extended to variational inequalities in (Blum and Suttmeier [28, 29], Suttmeier [139]). The application to PDE constrained optimal control was outlined in (Becker, Kapp, and Rannacher [15], Kapp [94]). For a survey of these results see (Bangert and Rannacher [8], Becker and Rannacher [12]). The method has been further extended to parameter identification problems in (Vexler [147]) and to non-stationary PDE in (Schmich [136], Schmich and Vexler [137]). In (Meidner [107], Meidner and Vexler [108]) this is extended to optimal control with parabolic PDE. Only recently the consideration of stationary optimal control problems subject to inequality constraints started. The case of pointwise control constraints is considered in (Hintermüller and Hoppe [80], Vexler and Wollner [148], Wollner [156]). In (Becker [10]) these techniques are used explicitly to estimate the error in the control with respect to its natural norm, as the error in the natural norm can be bounded by the error in the cost functional. For state constraints some recent work has been done simultaneously by (Benedix and Vexler [16], Günther and Hinze [75], Wollner [157]).

The rest of this chapter proceeds as follows. First we will discuss how to estimate the error with respect to a given quantity of interest in the case of control constraints. The results are already published in (Vexler and Wollner [148], Wollner [156]).

Then we will consider estimating the error introduced by the elimination of the state constraint. We begin with the discussion in the case of inactive control constraints for the error in the cost functional. This work is already published in (Wollner [157, 158]) for the case of a barrier functional. We will extend the estimates to the case of active control constraints for the barrier approach. Finally, we consider the regularization error estimate for the penalty approach.

These two estimates can then be combined to balance the contribution coming from discretization and regularization. This can also be considered in the spirit of estimating the iteration error of the path-following method for the state constraints, see, e.g., (Meidner, Rannacher, and Vihharev [109]) for the case of the error in the multigrid cycle.

6.1 Control Constraints

Here we will recall the results concerning the error estimation in the case of control constraints. This section contains results that have already been published in (Vexler and Wollner [148]).

To simplify notation we will only consider the case where Q is a Hilbert space, e.g., $Q = L^2(\omega)$ on some set ω. Typically, ω is a subset of the computational domain Ω or a subset of its boundary $\partial\Omega$. The case of finite dimensional controls is realized by choosing $\omega = \{1, 2, \dots, n\}$ resulting in $Q \cong \mathbb{R}^n$.

We consider the cost functional to be given in the form

$$J(q, u) = J_1(u) + \frac{\alpha}{2}\|q\|^2 , \tag{6.1}$$

where J_1 is a four times directionally differentiable operator on V and $\alpha > 0$. Let the admissible set Q^{ad} be given through box constraints on q, i.e.

$$Q^{\mathrm{ad}} = \{\, q \in Q \,|\, a \le q(x) \le b \text{ a.e. on } \omega \,\}, \tag{6.2}$$

with bounds $a, b \in \mathbb{R} \cup \{\pm\infty\}$ and $a < b$.

Now, we are able to restate the optimization problem as:

$$\text{Minimize } J(q, u) , \quad u \in V,\, q \in Q^{\mathrm{ad}} , \quad \text{subject to (2.4).} \tag{6.3}$$

To shorten notation, we introduce the space $\mathcal{X}$ and the admissible set $\mathcal{X}^{\mathrm{ad}}$ by:

$$\mathcal{X} = Q \times V \times V, \tag{6.4}$$

$$\mathcal{X}^{\mathrm{ad}} = Q^{\mathrm{ad}} \times V \times V. \tag{6.5}$$

In addition, we shall write $\xi = (q, u, z)$ for a vector in $\mathcal{X}$ or $\mathcal{X}^{\mathrm{ad}}$.

To shorten notation, we introduce the Lagrangian $\mathcal{L}\colon \mathcal{X} \to \mathbb{R}$ as follows:

$$\mathcal{L}(\xi) = J_1(u) + \frac{\alpha}{2}\|q\|^2 + f(z) - a(q, u)(z).$$

With this we can restate the abstract first-order necessary optimality condition (2.10). Let $(\overline{q}, \overline{u}) \in Q^{\mathrm{ad}} \times V$ be a solution to (6.3), then the following hold:
There exists $\overline{z} \in V$ such that the triple $\overline{\xi} = (\overline{q}, \overline{u}, \overline{z}) \in \mathcal{X}^{\mathrm{ad}}$ satisfies

$$\begin{aligned} \mathcal{L}'_u(\overline{\xi})(\delta u) &= 0 \quad \forall\, \delta u \in V, && \text{(6.6a)} \\ \mathcal{L}'_q(\overline{\xi})(\delta q - \overline{q}) &\geq 0 \quad \forall\, \delta q \in Q^{\mathrm{ad}}, && \text{(6.6b)} \\ \mathcal{L}'_z(\overline{\xi})(\delta z) &= 0 \quad \forall\, \delta z \in V. && \text{(6.6c)} \end{aligned}$$

This system can be stated explicitly in the following form:

$$\begin{aligned} J'_1(\overline{u})(\delta u) - a'_u(\overline{q}, \overline{u})(\delta u, \overline{z}) &= 0 \quad \forall\, \delta u \in V, && \text{(6.7a)} \\ \alpha(\overline{q}, \delta q - \overline{q}) - a'_q(\overline{q}, \overline{u})(\delta q - \overline{q}, \overline{z}) &\geq 0 \quad \forall\, \delta q \in Q^{\mathrm{ad}}, && \text{(6.7b)} \\ f(\delta z) - a(\overline{q}, \overline{u})(\delta z) &= 0 \quad \forall\, \delta z \in V. && \text{(6.7c)} \end{aligned}$$

We introduce a projection operator $\mathcal{P}_{Q^{\mathrm{ad}}} \colon Q \to Q^{\mathrm{ad}}$ by:

$$\mathcal{P}_{Q^{\mathrm{ad}}}(p) = \max\left(a, \min(p, b)\right)$$

pointwise almost everywhere. This allows us to rewrite variational inequality (6.7b), see, e.g., (Tröltzsch [141]), as:

$$\overline{q} = \mathcal{P}_{Q^{\mathrm{ad}}}\left(\frac{1}{\alpha} a'_q(\overline{q}, \overline{u})(\cdot, \overline{z})\right), \tag{6.8}$$

where $a'_q(\overline{q}, \overline{u})(\cdot, \overline{z})$ is understood as a Riesz representative of a linear functional on Q.

Remark 6.1. If one would consider $Q = L^r$ instead of $Q = L^2$ one would have to consider (3.7) instead of (6.8).

For a solution $(\overline{q}, \overline{u})$ of (6.3) we introduce active sets ω_- and ω_+ as follows:

$$\begin{aligned} \omega_- &= \{\, x \in \omega \,|\, \overline{q}(x) = a \,\}, && \text{(6.9)} \\ \omega_+ &= \{\, x \in \omega \,|\, \overline{q}(x) = b \,\}. && \text{(6.10)} \end{aligned}$$

Let $\overline{\xi} \in \mathcal{X}$ be a solution to (6.6), then we introduce an additional Lagrange multiplier $\overline{\mu} \in Q$ by the following identification:

$$(\overline{\mu}, \delta q) = -\alpha(\overline{q}, \delta q) + a'_q(\overline{q}, \overline{u})(\delta q, \overline{z}) = -\mathcal{L}'_q(\overline{\xi})(\delta q) \quad \forall \delta q \in Q\,. \tag{6.11}$$

The variational inequality (6.7b) or the projection formula (6.8) are known to be equivalent to the following conditions:

$$\begin{aligned} \overline{\mu}(x) &\leq 0 \quad \text{a.e. on } \omega_-\,, && \text{(6.12a)} \\ \overline{\mu}(x) &\geq 0 \quad \text{a.e. on } \omega_+\,, && \text{(6.12b)} \\ \overline{\mu}(x) &= 0 \quad \text{a.e. on } \omega \setminus (\omega_- \cup \omega_+)\,. && \text{(6.12c)} \end{aligned}$$

For the solution of problem (6.3) we refer to Section 7.1.

As we will encounter some trouble with the variational inequality in the necessary optimality condition (6.7) due to missing Galerkin orthogonality, we consider in addition the full Lagrangian $\tilde{\mathcal{L}}\colon \mathcal{X} \times Q \times Q \to \mathbb{R}$ which is given by:

$$\tilde{\mathcal{L}}(\chi) = \mathcal{L}(\xi) + (\,\mu^-\,, a - q\,)_Q + (\,\mu^+\,, q - b\,)_Q,$$

with $\chi = (\xi, \mu^-, \mu^+) = (q, u, z, \mu^-, \mu^+) \in \mathcal{X} \times Q \times Q$ where μ^- and μ^+ denote the variables corresponding to Lagrange multipliers for the inequality constraints. To shorten notation we introduce the abbreviation

$$\mathcal{Y} = \mathcal{X} \times Q \times Q. \tag{6.13}$$

Using the subspaces

$$Q_- = \{\, r \in Q \,|\, r = 0 \text{ a.e. on } \omega \setminus \omega_- \,\},$$
$$Q_+ = \{\, r \in Q \,|\, r = 0 \text{ a.e. on } \omega \setminus \omega_+ \,\},$$

we introduce

$$\mathcal{Y}^{\mathrm{ad}} = \mathcal{X}^{\mathrm{ad}} \times Q_- \times Q_+, \tag{6.14}$$
$$\tilde{\mathcal{Y}}^{\mathrm{ad}} = \mathcal{X} \times Q_- \times Q_+, \tag{6.15}$$

and see immediately that the following equality holds for arbitrary elements $\chi \in \overline{q} + Q \setminus (Q_- \cup Q_+) \times V \times V \times Q_- \times Q_+ \subset \tilde{\mathcal{Y}}^{\mathrm{ad}}$:

$$\mathcal{L}(\xi) = \tilde{\mathcal{L}}(\chi). \tag{6.16}$$

This means, both Lagrangians coincide, for all functions such that the active set of the control is larger then the one used to define Q_- and Q_+.

We can rewrite the first-order necessary optimality condition for $(\overline{q}, \overline{u}) \in Q^{\mathrm{ad}} \times V$ equivalently as follows, cf., (Tröltzsch [141]):
There exists $\overline{z} \in V$, $\overline{\mu}^- \in Q_-$, $\overline{\mu}^+ \in Q_+$ such that the following conditions hold for $\overline{\chi} = (\overline{q}, \overline{u}, \overline{z}, \overline{\mu}^-, \overline{\mu}^+) \in \mathcal{Y}^{\mathrm{ad}}$

$$\tilde{\mathcal{L}}'_u(\overline{\chi})(\delta u) = 0 \quad \forall \delta u \in V, \tag{6.17a}$$
$$\tilde{\mathcal{L}}'_q(\overline{\chi})(\delta q) = 0 \quad \forall \delta q \in Q, \tag{6.17b}$$
$$\tilde{\mathcal{L}}'_z(\overline{\chi})(\delta z) = 0 \quad \forall \delta z \in V, \tag{6.17c}$$
$$\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\delta \mu^-) = 0 \quad \forall \delta \mu^- \in Q_-, \tag{6.17d}$$
$$\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\delta \mu^+) = 0 \quad \forall \delta \mu^+ \in Q_+, \tag{6.17e}$$
$$\overline{\mu}^+, \overline{\mu}^- \geq 0 \quad \text{a.e. on } \omega. \tag{6.17f}$$

It is easy to verify that the Lagrange multipliers $\overline{\mu}^+$ and $\overline{\mu}^-$ are given as the positive and negative part of the Lagrange multiplier $\overline{\mu}$ from (6.11), cf., (Tröltzsch [141]).

Note that the equations (6.17d), (6.17e) are equivalent to the complementarity conditions

$$\overline{\mu}^-(a - \overline{q}) = \overline{\mu}^+(\overline{q} - b) = 0 \quad \text{a.e. on } \omega, \tag{6.18}$$

6.1.1 Discretization

For the choice of the discrete spaces we refer to Section 5.3. The discretized problem then reads as follows:

$$\text{Minimize } J_1(u_h) + \frac{\alpha}{2}\|q_h\|_Q^2, \quad u_h \in V_h,\, q_h \in Q_h^{\text{ad}}, \tag{6.19}$$

subject to

$$a(q_h, u_h)(v_h) = f(v_h) \quad \forall v_h \in V_h. \tag{6.20}$$

We introduce the discretized versions of (6.4) and (6.5) by

$$\mathcal{X}_h = Q_h \times V_h \times V_h, \tag{6.21}$$
$$\mathcal{X}_h^{\text{ad}} = Q_h^{\text{ad}} \times V_h \times V_h, \tag{6.22}$$

and denote a vector from these sets by $\xi_h = (q_h, u_h, z_h)$. The optimality system for the discretized optimization problem is formulated as follows: Let $(\overline{q}_h, \overline{u}_h) \in Q_h^{\text{ad}} \times V_h$ be a solution to (6.19) subject to (6.20), then there exists $\overline{z}_h \in V_h$ such that

$$J_1'(\overline{u}_h)(\delta u_h) - a_u'(\overline{q}_h, \overline{u}_h)(\delta u_h, \overline{z}_h) = 0 \quad \forall \delta u_h \in V_h, \tag{6.23a}$$
$$\alpha(\overline{q}_h, \delta q_h - \overline{q}_h) - a_q'(\overline{q}_h, \overline{u}_h)(\delta q_h - \overline{q}_h, \overline{z}_h) \geq 0 \quad \forall \delta q_h \in Q_h^{\text{ad}}, \tag{6.23b}$$
$$f(\delta z_h) - a(\overline{q}_h, \overline{u}_h)(\delta z_h) = 0 \quad \forall \delta z_h \in V_h. \tag{6.23c}$$

In order to formulate the analog system to (6.17a)—(6.17f) we introduce discrete active sets $\omega_{-,h}$ and $\omega_{+,h}$ for a solution $(\overline{q}_h, \overline{u}_h)$ to (6.19), (6.20) by:

$$\omega_{-,h} = \{\, x \in \omega \,|\, \overline{q}_h(x) = a \,\}, \tag{6.24}$$
$$\omega_{+,h} = \{\, x \in \omega \,|\, \overline{q}_h(x) = b \,\}, \tag{6.25}$$

and define a Lagrange multiplier $\overline{\mu}_h \in Q_h$ via:

$$(\overline{\mu}_h, \delta q_h) = -\mathcal{L}_q'(\overline{q}_h, \overline{u}_h, \overline{z}_h)(\delta q_h) \quad \forall\, \delta q_h \in Q_h. \tag{6.26}$$

Moreover, we introduce $\overline{\mu}_h^- \in Q_h$ and $\overline{\mu}_h^+ \in Q_h$ by:

$$\overline{\mu}_h^+ - \overline{\mu}_h^- = \overline{\mu}_h, \;\; (\,\overline{\mu}_h^-\,, \psi_i\,)_Q \geq 0, \;\; (\,\overline{\mu}_h^+\,, \psi_i\,)_Q \geq 0 \quad \forall \psi_i \in \mathcal{B}, \tag{6.27}$$

$$(\,\overline{\mu}_h^-\,, a - \overline{q}_h\,)_Q = (\,\overline{\mu}_h^+\,, \overline{q}_h - b\,)_Q = 0, \tag{6.28}$$

by which $\overline{\mu}_h^{\pm}$ are uniquely determined.

Remark 6.2. This definition corresponds to the Lagrange multipliers obtained for the inequality constraints if the discrete optimization problem (6.19), (6.20) is considered as finite dimensional optimization problem for $q_h = \sum_i q_i \psi_i \in Q_h$ with the restrictions:

$$a \leq q_i \leq b \quad \forall i.$$

Note that due to the choice of the basis $\mathcal{B}$ in (5.29) this is equivalent to $a \leq q_h(x) \leq b$ for all $x \in \omega$. Utilizing this fact, the discrete active sets $\omega_{-,h}$, $\omega_{+,h}$ are completely determined by the values of the coordinate vector of q_h. In particular they consist only of whole cells, edges and nodes.

To obtain the complementarity conditions with respect to the $Q = L^2(\omega)$-inner product (6.28) one requires

$$(\mu_h^+, \psi_i) = 0 \text{ if } q_i < b\,, \quad \text{and} \quad (\mu_h^-, \psi_i) = 0 \text{ if } q_i > a\,.$$

We now define the discretized versions of (6.13), (6.15) and (6.14) by:

$$\mathcal{Y}_h = \mathcal{X}_h \times Q_h \times Q_h, \tag{6.29}$$
$$\mathcal{Y}_h^{\text{ad}} = \mathcal{X}_h^{\text{ad}} \times Q_{-,h} \times Q_{+,h}, \tag{6.30}$$
$$\tilde{\mathcal{Y}}_h^{\text{ad}} = \mathcal{X}_h \times Q_{-,h} \times Q_{+,h}, \tag{6.31}$$

where

$$Q_{-,h} = \{\, r \in Q_h \,|\, r(x) = 0 \text{ a.e. on } \omega \setminus \omega_h^- \,\},$$
$$Q_{+,h} = \{\, r \in Q_h \,|\, r(x) = 0 \text{ a.e. on } \omega \setminus \omega_h^+ \,\}.$$

A vector from these spaces will be abbreviated by $\chi_h = (q_h, u_h, z_h, \mu_h^-, \mu_h^+)$.

Using the above definitions we have the first-order necessary optimality condition for a solution $(\overline{q}_h, \overline{u}_h) \in Q_h^{\text{ad}} \times V_h$ of (6.19), (6.20). Namely there exists $\overline{z}_h \in V_h$, $\overline{\mu}_h^- \in Q_{-,h}$, $\overline{\mu}_h^+ \in Q_{+,h}$ such that for $\overline{\chi}_h = (\overline{q}_h, \overline{u}_h, \overline{z}_h, \overline{\mu}_h^-, \overline{\mu}_h^+) \in \mathcal{Y}_h^{\text{ad}}$ the following conditions hold:

$$\tilde{\mathcal{L}}'_u(\overline{\chi}_h)(\delta u) = 0 \quad \forall \delta u \in V_h, \tag{6.32a}$$
$$\tilde{\mathcal{L}}'_q(\overline{\chi}_h)(\delta q) = 0 \quad \forall \delta q \in Q_h, \tag{6.32b}$$
$$\tilde{\mathcal{L}}'_z(\overline{\chi}_h)(\delta z) = 0 \quad \forall \delta z \in V_h, \tag{6.32c}$$
$$\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi}_h)(\delta \mu^-) = 0 \quad \forall \delta \mu^- \in Q_{-,h}, \tag{6.32d}$$
$$\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi}_h)(\delta \mu^+) = 0 \quad \forall \delta \mu^+ \in Q_{+,h}, \tag{6.32e}$$
$$\left.\begin{aligned} (\,\overline{\mu}_h^-\,, \psi_i\,)_Q &\geq 0 \quad \forall \psi_i \in \mathcal{B}, \\ (\,\overline{\mu}_h^+\,, \psi_i\,)_Q &\geq 0 \quad \forall \psi_i \in \mathcal{B}, \\ \overline{\mu}_h^+ - \overline{\mu}_h^- &= \overline{\mu}_h. \end{aligned}\right\} \tag{6.32f}$$

Here again (6.32d), (6.32e) is equivalent to the complementarity condition

$$(\,\overline{\mu}_h^-\,, a - \overline{q}_h\,)_Q = (\,\overline{\mu}_h^+\,, \overline{q}_h - b\,)_Q = 0. \tag{6.33}$$

6.1.2 A Posteriori Error Estimation

The aim of this section is to derive a posteriori error estimates for the error with respect to the cost functional and to an arbitrary quantity of interest. These error estimates extend the results from (Becker and Rannacher [12], Becker and Vexler [13, 14], Becker et al. [15])

to the case of optimization problems with control constraints. The provided estimators will be used within the following adaptive algorithm for error control and mesh refinement: We start on a coarse mesh, solve the discretized optimization problem and evaluate the error estimator. Thereafter we refine the current mesh using local information obtained from the error estimator, allowing for efficient reduction of the discretization error with respect to the quantity of interest. This procedure is iterated until the value of the error estimator is below a given tolerance, see, e.g., (Becker and Vexler [13]) for a detailed description of this algorithm.

The section is structured as follows: First we will derive two a posteriori error estimators for the error with respect to the cost functional. The first one is based on the first-order necessary condition (6.7) which involves a variational inequality, the second estimator uses the information obtained from the Lagrange multipliers for the inequality constraints. Both estimators can be evaluated in terms of the solution to the discretized optimization problem (6.19), (6.20). Then we will proceed with the error estimator with respect to an arbitrary quantity of interest, which requires the solution to an auxiliary linear-quadratic optimization problem. Even though the idea behind the estimators remains unchanged, the later estimators require a more technical discussion.

Throughout this section we shall denote a solution to the optimization problem (6.3) by $(\overline{q}, \overline{u})$, the corresponding solution to the optimality system (6.6) by $\overline{\xi} = (\overline{q}, \overline{u}, \overline{z}) \in \mathcal{X}^{\mathrm{ad}}$,and its discrete counterpart (6.23) by $\overline{\xi}_h = (\overline{q}_h, \overline{u}_h, \overline{z}_h) \in \mathcal{X}_h^{\mathrm{ad}}$. The corresponding solution to (6.17) and its discrete counterpart (6.32) will be abbreviated by $\overline{\chi} = (\overline{q}, \overline{u}, \overline{z}, \overline{\mu}^-, \overline{\mu}^+) \in \mathcal{Y}^{\mathrm{ad}}$ and $\overline{\chi}_h = (\overline{q}_h, \overline{u}_h, \overline{z}_h, \overline{\mu}_h^-, \overline{\mu}_h^+) \in \mathcal{Y}_h^{\mathrm{ad}}$.

6.1.2.1 Error in the Cost Functional

For the derivation of the error estimator with respect to the cost functional, we introduce the residual functionals $\rho_u(\xi_h)(\cdot)$, $\rho_z(\xi_h)(\cdot) \in V^*$ and $\rho_q(\xi_h)(\cdot) \in Q^*$ by

$$\rho_u(\xi_h)(\cdot) = f(\cdot) - a(q_h, u_h)(\cdot), \tag{6.34}$$

$$\rho_z(\xi_h)(\cdot) = J_1'(u_h)(\cdot) - a_u'(q_h, u_h)(\cdot, z_h), \tag{6.35}$$

$$\rho_q(\xi_h)(\cdot) = \alpha(q_h, \cdot) - a_q'(u_h, q_h)(\cdot, z_h). \tag{6.36}$$

The following theorem is an extension of the result from (Becker and Rannacher [12]).

Theorem 6.1. *Let $\overline{\xi} \in \mathcal{X}^{ad}$ be a solution to the first-order necessary system* (6.6) *and $\overline{\xi}_h \in \mathcal{X}_h^{ad}$ be its Galerkin approximation* (6.23)*. Then the following estimate holds:*

$$J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h) \leq \frac{1}{2}\rho_u(\overline{\xi}_h)(\overline{z} - \tilde{z}_h) + \frac{1}{2}\rho_z(\overline{\xi}_h)(\overline{u} - \tilde{u}_h) + \frac{1}{2}\rho_q(\overline{\xi}_h)(\overline{q} - q_h) + R_1, \tag{6.37}$$

where $\tilde{u}_h, \tilde{z}_h \in V_h$ are arbitrarily chosen and R_1 is a remainder term given by:

$$R_1 = \frac{1}{2}\int_0^1 \mathcal{L}'''(\overline{\xi}_h + s(\overline{\xi} - \overline{\xi}_h))(\overline{\xi} - \overline{\xi}_h, \overline{\xi} - \overline{\xi}_h, \overline{\xi} - \overline{\xi}_h)s(s-1)\,ds. \tag{6.38}$$

Proof. From optimality system (6.6a)—(6.6c) we obtain that

$$J(\overline{q},\overline{u}) = \mathcal{L}(\overline{\xi}).$$

A similar equality holds on the discrete level. Therefore we have:

$$J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h) = \mathcal{L}(\overline{\xi}) - \mathcal{L}(\overline{\xi}_h) = \int_0^1 \mathcal{L}'(\overline{\xi}_h + s(\overline{\xi}-\overline{\xi}_h))(\overline{\xi}-\overline{\xi}_h)\,ds.$$

We approximate this integral by the trapezoidal rule and obtain:

$$J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h) = \frac{1}{2}\mathcal{L}'(\overline{\xi})(\overline{\xi}-\overline{\xi}_h) + \frac{1}{2}\mathcal{L}'(\overline{\xi}_h)(\overline{\xi}-\overline{\xi}_h) + R_1, \tag{6.39}$$

with the remainder term R_1 as in (6.38). For the first term we have:

$$\mathcal{L}'(\overline{\xi})(\overline{\xi}-\overline{\xi}_h) \;= \mathcal{L}'_u(\overline{\xi})(\overline{u}-\overline{u}_h) + \mathcal{L}'_z(\overline{\xi})(\overline{z}-\overline{z}_h) + \mathcal{L}'_q(\overline{\xi})(\overline{q}-\overline{q}_h).$$

Using optimality system (6.6a)—(6.6c) and the fact that $\overline{q}_h \in Q_h^{\mathrm{ad}} \subset Q^{\mathrm{ad}}$, we deduce:

$$\mathcal{L}'(\overline{\xi})(\overline{\xi}-\overline{\xi}_h) = -\mathcal{L}'_q(\overline{\xi})(\overline{q}_h-\overline{q}) \le 0.$$

Rewriting the second term in (6.39) we obtain:

$$\mathcal{L}'(\overline{\xi}_h)(\overline{\xi}-\overline{\xi}_h) = \rho_u(\overline{\xi}_h)(\overline{z}-\overline{z}_h) + \rho_z(\overline{\xi}_h)(\overline{u}-\overline{u}_h) + \rho_q(\overline{\xi}_h)(\overline{q}-\overline{q}_h).$$

Due to the Galerkin orthogonality for the state and adjoint equations, we have for arbitrary $\tilde{u}_h, \tilde{z}_h \in V_h$

$$\rho_u(\overline{\xi}_h)(\overline{z}-\overline{z}_h) = \rho_u(\overline{\xi}_h)(\overline{z}-\tilde{z}_h) \quad \text{and} \quad \rho_z(\overline{\xi}_h)(\overline{u}-\overline{u}_h) = \rho_z(\overline{\xi}_h)(\overline{u}-\tilde{u}_h).$$

This completes the proof. □

Remark 6.3. We note that, in contrast to the terms involving the residuals of state and the adjoint equations, the error $\overline{q}-\overline{q}_h$ in the term $\rho_q(\overline{\xi}_h)(\overline{q}-\overline{q}_h)$ in (6.37) can not be replaced by $\overline{q}-\tilde{q}_h$ with an arbitrary $\tilde{q}_h \in Q_h^{\mathrm{ad}}$. This fact is caused by the control constraints. However we may replace $\rho_q(\overline{\xi}_h)(\overline{q}-\overline{q}_h)$ by $\rho_q(\overline{\xi}_h)(\overline{q}-\overline{q}_h+\tilde{q}_h)$ with arbitrary $\tilde{q}_h$ fulfilling $\operatorname{supp}(\tilde{q}_h) \subset \omega \setminus (\omega_{-,h} \cup \omega_{+,h})$ due to the structure of $\rho_q(\overline{\xi}_h)(\cdot)$. A similar structure is obtained for the case of error estimation in the context of variational inequalities, see, e.g., (Blum and Suttmeier [29]).

In order to use the estimate from the theorem above for computable error estimation we proceed as follows: First we choose $\tilde{u}_h = i_h\overline{u}$, $\tilde{z}_h = i_h\overline{z}$, with an interpolation operator $i_h\colon V \to V_h$. Then we have to approximate the corresponding interpolation errors $\overline{u}-i_h\overline{u}$ and $\overline{z}-i_h\overline{z}$. There are several heuristic techniques to do this, see for instance (Bangert and Rannacher [8], Becker and Rannacher [12], Becker and Vexler [13]). We assume to have an operator $\pi\colon V_h \to \tilde{V}_h$, with $\tilde{V}_h \neq V_h$, such that $\overline{u}-\pi\overline{u}_h$ has a better local asymptotical behavior as $\overline{u}-i_h\overline{u}$. Then we approximate:

$$\rho_u(\overline{\xi}_h)(\overline{z}-i_h\overline{z}) \approx \rho_u(\overline{\xi}_h)(\pi\overline{z}_h-\overline{z}_h) \quad \text{and} \quad \rho_z(\overline{\xi}_h)(\overline{u}-i_h\overline{u}) \approx \rho_z(\overline{\xi}_h)(\pi\overline{u}_h-\overline{u}_h).$$

Such an operator can be constructed for example by the interpolation of the computed bilinear finite element solution in the space of biquadratic finite elements on patches of cells. For this operator the improved approximation property relies on local smoothness of $\overline{u}$ and super-convergence properties of the approximation $\overline{u}_h$. The use of such 'local higher-order approximation' is observed to work very successfully in the context of a posteriori error estimation, see, e.g., (Bangert and Rannacher [8], Becker and Rannacher [12], Becker and Vexler [13]).

The approximation of the term $\rho_q(\overline{\xi}_h)(\overline{q} - \overline{q}_h)$ requires more care. In contrast to the state $\overline{u}$ and the adjoint state $\overline{z}$, the control variable $\overline{q}$ can generally not be approximated by 'local higher-order approximation', for the following reasons:

- In the case of finite dimensional control space Q, there is no 'patch-like' structure that allows for 'local higher-order approximation'.

- If $\overline{q}$ is a distributed control, it typically does not possess sufficient smoothness (due to the inequality constraints) for the improved approximation property.

Therefore we suggest another approximation of $\rho_q(\overline{\xi}_h)(\overline{q} - \overline{q}_h)$ based on the projection formula (6.8). To this end we introduce $\pi^q \in Q_h \to Q^{\mathrm{ad}}$ by:

$$\pi^q \overline{q}_h = \mathcal{P}_{Q^{\mathrm{ad}}}\Big(\frac{1}{\alpha} a_q'(\overline{q}_h, \pi\overline{u}_h)(\cdot, \pi\overline{z}_h)\Big). \tag{6.40}$$

In some cases one can show better approximation behavior of $\overline{q} - \pi^q\overline{q}_h$ in comparison with $\overline{q} - \overline{q}_h$, see (Meyer and Rösch [112]) and (Hinze [88]) for similar considerations in the context of a priori error analysis.

This construction results in the following computable a posteriori error estimator:

$$\eta_h^{(1)} = \frac{1}{2}\big(\rho_u(\overline{\xi}_h)(\pi\overline{z}_h - \overline{z}_h) + \rho_z(\overline{\xi}_h)(\pi\overline{u}_h - \overline{u}_h) + \rho_q(\overline{\xi}_h)(\pi^q\overline{q}_h - \overline{q}_h)\big).$$

Remark 6.4. In order to use this error estimator as an indicator for mesh refinement, we have to localize it to cell-wise or node-wise contributions. A direct localization of the terms like $\rho_u(\overline{\xi}_h)(\pi\overline{z}_h - \overline{z}_h)$ leads, in general, to local contributions of wrong order (overestimation) due to oscillatory behavior of the residual terms (Carstensen and Verfürth [33]). To overcome this, one may integrate the residual terms by part, see, e.g., (Becker and Rannacher [12]), or use a filtering operator, see (Vexler [147]) for details.

We should note, that (6.37) does not provide an estimate for the absolute value of $J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h)$, which is due to the inequality sign in (6.37). In the next section we will overcome this difficulty utilizing the alternative optimality system (6.17a)—(6.17f).

6.1.2.2 Error in the Cost Functional Reviewed

In order to derive an error estimator for the absolute value of $J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h)$ we introduce the following additional residual functionals $\tilde{\rho}_q(\chi_h)(\cdot)$, $\tilde{\rho}_{\mu^-}(\chi_h)(\cdot)$, $\tilde{\rho}_{\mu^+}(\chi_h)(\cdot) \in Q^*$ by:

$$\tilde{\rho}_q(\chi_h)(\cdot) = \alpha(\, q_h\,,\cdot\,)_Q - a_q'(q_h,u_h)(\cdot,z_h) + (\,\mu_h^+ - \mu_h^-\,,\cdot\,)_Q, \tag{6.41}$$

$$\tilde{\rho}_{\mu^-}(\chi_h)(\cdot) = (\,\cdot\,, a - q_h\,)_Q, \tag{6.42}$$

$$\tilde{\rho}_{\mu^+}(\chi_h)(\cdot) = (\,\cdot\,, q_h - b\,)_Q. \tag{6.43}$$

In the sequel the last two residual functional will also be evaluated in the point χ where they read as follows:

$$\tilde{\rho}_{\mu^-}(\chi)(\cdot) = (\,\cdot\,, a - q\,)_Q, \quad \tilde{\rho}_{\mu^+}(\chi)(\cdot) = (\,\cdot\,, q - b\,)_Q.$$

Analog to Theorem 6.1 we obtain:

Theorem 6.2. *Let $\overline{\chi} \in \mathcal{Y}^{ad}$ be a solution to the first-order necessary condition* (6.17a)—(6.17f) *and $\overline{\chi}_h \in \mathcal{Y}_h^{ad}$ be its Galerkin approximation* (6.32a)—(6.33)*. Then the following estimate holds:*

$$\begin{aligned} J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h) = {} & \frac{1}{2}\rho_u(\overline{\chi}_h)(\overline{z} - \tilde{z}_h) + \frac{1}{2}\rho_z(\overline{\chi}_h)(\overline{u} - \tilde{u}_h) + \frac{1}{2}\tilde{\rho}_q(\overline{\chi}_h)(\overline{q} - \tilde{q}_h) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\overline{\mu}^- - \tilde{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\overline{\mu}^+ - \tilde{\mu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\overline{\chi})(\tilde{\mu}^- - \overline{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\overline{\chi})(\tilde{\mu}^+ - \overline{\mu}_h^+) + R_2 \end{aligned} \tag{6.44}$$

where $\tilde{u}_h, \tilde{z}_h \in V_h$, $\tilde{q}_h \in Q_h$, $\tilde{\mu}_h^- \in Q_{-,h}$, $\tilde{\mu}_h^+ \in Q_{+,h}$, $\tilde{\mu}^- \in Q_-$, $\tilde{\mu}^+ \in Q_+$ are arbitrarily chosen and R_2 is a remainder term given by:

$$R_2 = \frac{1}{2}\int_0^1 \tilde{\mathcal{L}}'''(\overline{\chi}_h + s(\overline{\chi} - \overline{\chi}_h))(\overline{\chi} - \overline{\chi}_h, \overline{\chi} - \overline{\chi}_h, \overline{\chi} - \overline{\chi}_h)s(s-1)\,ds. \tag{6.45}$$

Proof. From (6.16) and optimality system (6.7a)—(6.7c) we obtain

$$J(\overline{q},\overline{u}) = \mathcal{L}(\overline{\xi}) = \tilde{\mathcal{L}}(\overline{\chi}).$$

The analog result holds on the discrete level. We therefore have:

$$J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h) = \tilde{\mathcal{L}}(\overline{\chi}) - \tilde{\mathcal{L}}(\overline{\chi}_h) = \int_0^1 \tilde{\mathcal{L}}'(\overline{\chi}_h + s(\overline{\chi} - \overline{\chi}_h))(\overline{\chi} - \overline{\chi}_h)\,ds.$$

As in the proof of Theorem 6.1 we approximate this integral by the trapezoidal rule and obtain:

$$J(\overline{q},\overline{u}) - J(\overline{q}_h,\overline{u}_h) = \frac{1}{2}\tilde{\mathcal{L}}'(\overline{\chi})(\overline{\chi} - \overline{\chi}_h) + \frac{1}{2}\tilde{\mathcal{L}}'(\overline{\chi}_h)(\overline{\chi} - \overline{\chi}_h) + R_2 \tag{6.46}$$

with the remainder term R_2 as in (6.45). For the first term we have:

$$\begin{aligned}\tilde{\mathcal{L}}'(\overline{\chi})(\overline{\chi}-\overline{\chi}_h) \;=\; &\tilde{\mathcal{L}}'_u(\overline{\chi})(\overline{u}-\overline{u}_h)+\tilde{\mathcal{L}}'_z(\overline{\chi})(\overline{z}-\overline{z}_h)+\tilde{\mathcal{L}}'_q(\overline{\chi})(\overline{q}-\overline{q}_h)\\ &+\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\overline{\mu}^- - \overline{\mu}_h^-)+\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\overline{\mu}^+ - \overline{\mu}_h^+).\end{aligned}$$

Using optimality system (6.17a)—(6.17f) we deduce:

$$\tilde{\mathcal{L}}'(\overline{\chi})(\overline{\chi}-\overline{\chi}_h)=\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\overline{\mu}^- - \overline{\mu}_h^-)+\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\overline{\mu}^+ - \overline{\mu}_h^+).$$

From (6.17d) and (6.17e) together with linearity of $\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\cdot)$ and $\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\cdot)$ we obtain that for arbitrary $\tilde{\mu}^- \in Q_-$ and $\tilde{\mu}^+ \in Q_+$

$$\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\overline{\mu}^- - \overline{\mu}_h^-)=\tilde{\mathcal{L}}'_{\mu^-}(\overline{\chi})(\tilde{\mu}^- - \overline{\mu}_h^-),\quad \tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\overline{\mu}^+ - \overline{\mu}_h^+)=\tilde{\mathcal{L}}'_{\mu^+}(\overline{\chi})(\tilde{\mu}^+ - \overline{\mu}_h^+)$$

holds, thus we obtain:

$$\tilde{\mathcal{L}}'(\overline{\chi})(\overline{\chi}-\overline{\chi}_h)=\tilde{\rho}_{\mu^-}(\overline{\chi})(\tilde{\mu}^- - \overline{\mu}_h^-)+\tilde{\rho}_{\mu^+}(\overline{\chi})(\tilde{\mu}^+ - \overline{\mu}_h^+).$$

Rewriting the second term in (6.46) we obtain:

$$\begin{aligned}\tilde{\mathcal{L}}'(\overline{\chi}_h)(\overline{\chi}-\overline{\chi}_h) \;=\; &\rho_u(\overline{\chi}_h)(\overline{u}-\overline{u}_h)+\rho_z(\overline{\chi}_h)(\overline{z}-\overline{z}_h)+\tilde{\rho}_q(\overline{\chi}_h)(\overline{q}-\overline{q}_h)\\ &+\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\overline{\mu}^- - \overline{\mu}_h^-)+\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\overline{\mu}^+ - \overline{\mu}_h^+),\end{aligned}$$

where we can use linearity of the residual functionals in the second argument and (6.32a)—(6.32c) to obtain the following equalities:

$$\rho_u(\overline{\chi}_h)(\overline{u}-\overline{u}_h) \;=\; \rho_u(\overline{\chi}_h)(\overline{u}-\tilde{u}_h), \tag{6.47}$$

$$\rho_z(\overline{\chi}_h)(\overline{z}-\overline{z}_h) \;=\; \rho_z(\overline{\chi}_h)(\overline{z}-\tilde{z}_h), \tag{6.48}$$

$$\tilde{\rho}_q(\overline{\chi}_h)(\overline{q}-\overline{q}_h) \;=\; \tilde{\rho}_q(\overline{\chi}_h)(\overline{q}-\tilde{q}_h), \tag{6.49}$$

for arbitrary $\tilde{u}_h, \tilde{z}_h \in V_h$, $\tilde{q}_h \in Q_h$. Additionally, we gain from (6.32d) and (6.32e) that

$$\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\overline{\mu}^- - \overline{\mu}_h^-) \;=\; \tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\overline{\mu}^- - \tilde{\mu}_h^-), \tag{6.50}$$

$$\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\overline{\mu}^+ - \overline{\mu}_h^+) \;=\; \tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\overline{\mu}^+ - \tilde{\mu}_h^+), \tag{6.51}$$

holds for arbitrary $\tilde{\mu}_h^- \in Q_{-,h}$ and $\tilde{\mu}_h^+ \in Q_{+,h}$. This completes the proof. □

To gain a computable error estimator we proceed as in the previous section. In order to deal with the new residual functionals we utilize (6.11) and construct an approximation for $\overline{\mu}$ by

$$\tilde{\mu} = -\alpha\pi^q\overline{q}_h + a'_q(\pi^q\overline{q}_h, \pi u_h)(\cdot, \pi z_h) \tag{6.52}$$

where $\pi^q\overline{q}_h$ is given by (6.40). This leads to a computable a posteriori error estimator:

$$\begin{aligned}\eta_h^{(2)} = \frac{1}{2}\big(&\rho_u(\overline{\chi}_h)(\pi\overline{z}_h-\overline{z}_h)+\rho_z(\overline{\chi}_h)(\pi u_h-\overline{u}_h)+\tilde{\rho}_q(\overline{\chi}_h)(\pi^q\overline{q}_h-\overline{q}_h)\\ &+\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\tilde{\mu}^- - \overline{\mu}_h^-)+\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\tilde{\mu}^+ - \overline{\mu}_h^+)\\ &+\tilde{\rho}_{\mu^-}(\tilde{\chi})(\tilde{\mu}^- - \overline{\mu}_h^-)+\tilde{\rho}_{\mu^+}(\tilde{\chi})(\tilde{\mu}^+ - \overline{\mu}_h^+)\big).\end{aligned}$$

Here $\tilde{\chi}$ is an abbreviation for $(\pi^q\overline{q}_h, \pi\overline{u}_h, \pi\overline{z}_h, \tilde{\mu}^-, \tilde{\mu}^+)$.

Remark 6.5. We note that the a posteriori error estimates derived in Theorem 6.1 and Theorem 6.2 coincide if the control constraints are inactive, e.g., if $Q^{\mathrm{ad}} = Q$.

6.1.2.3 Error in the Quantity of Interest

The aim of this section is the derivation of an error estimator for the error

$$I(\overline{q},\overline{u}) - I(\overline{q}_h,\overline{u}_h) \tag{6.53}$$

with a given functional $I\colon Q \times V \to \mathbb{R}$ describing the quantity of interest. We require I to be three times directional differentiable. To this end we consider an additional Lagrangian $\mathcal{M}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ defined by

$$\mathcal{M}(\chi)(\psi) = I(q,u) + \tilde{\mathcal{L}}'(\chi)(\psi), \tag{6.54}$$

where we abbreviate $\chi = (q,u,z,\mu^-,\mu^+)$ and $\psi = (p,v,y,\nu^-,\nu^+)$. Here (p,v,y,ν^-,ν^+) will be variables dual to (q,u,z,μ^-,μ^+). Note that for the solution $\overline{\chi}$ to the optimality system (6.17a)—(6.17f) of the optimization problem (6.3) the identity

$$\mathcal{M}(\overline{\chi})(\psi) = I(\overline{q},\overline{u}) \tag{6.55}$$

holds for all $\psi \in \tilde{\mathcal{Y}}^{\mathrm{ad}}$. To proceed as in the proof of Theorem 6.2 it remains to find $\overline{\psi} \in \tilde{\mathcal{Y}}^{\mathrm{ad}}$ such that $(\overline{\chi},\overline{\psi})$ is a stationary point of $\mathcal{M}$ on $\tilde{\mathcal{Y}}^{\mathrm{ad}} \times \tilde{\mathcal{Y}}^{\mathrm{ad}}$.

Therefore we consider the auxiliary (linear-quadratic) optimization problem:

$$\text{Minimize } K(\overline{\chi},p,v), \quad p \in P^{\mathrm{ad}},\, v \in V, \tag{6.56}$$

$$\text{subject to } \tilde{\mathcal{L}}''_{uz}(\overline{\chi})(v,\varphi) + \tilde{\mathcal{L}}''_{qz}(\overline{\chi})(p,\varphi) = 0 \quad \forall\, \varphi \in V, \tag{6.57}$$

for given $\overline{\chi} \in \mathcal{Y}$. The admissible set P^{ad} is given as

$$P^{\mathrm{ad}} = \{\, p \in Q \,|\, p_-(x) \le p(x) \le p_+(x) \text{ a.e. on } \omega \,\}, \tag{6.58}$$

with the bounds

$$p_-(x) \;=\; \begin{cases} 0 & \overline{\mu}(x) \neq 0 \text{ or } \overline{q}(x) = a, \\ -\infty & \text{else}, \end{cases}$$

$$p_+(x) \;=\; \begin{cases} 0 & \overline{\mu}(x) \neq 0 \text{ or } \overline{q}(x) = b, \\ +\infty & \text{else}, \end{cases}$$

and the cost functional $K : \mathcal{Y} \times Q \times V \to \mathbb{R}$ is defined via:

$$K(\overline{\chi},p,v) = I'_u(\overline{q},\overline{u})(v) + I'_q(\overline{q},\overline{u})(p) + \tilde{\mathcal{L}}''_{uq}(\overline{\chi})(v,p) + \frac{1}{2}\tilde{\mathcal{L}}''_{uu}(\overline{\chi})(v,v) + \frac{1}{2}\tilde{\mathcal{L}}''_{qq}(\overline{\chi})(p,p). \tag{6.59}$$

We introduce the following abbreviation for later use:

$$\bar{\mathcal{Y}}^{\mathrm{ad}} \;=\; P^{\mathrm{ad}} \times V \times V \times P_- \times P_+. \tag{6.60}$$

Where we define P_- and P_+ analog to the spaces Q_- and Q_+ as:

$$P_- = \{\, r \in Q \,|\, r = 0 \text{ a.e. on } \omega \setminus \{\, \overline{p} = p_- \,\} \,\},$$
$$P_+ = \{\, r \in Q \,|\, r = 0 \text{ a.e. on } \omega \setminus \{\, \overline{p} = p_+ \,\} \,\}.$$

Remark 6.6. Consideration of the auxiliary optimization problem (6.56), (6.57) is motivated by the unconstrained case $Q^{\mathrm{ad}} = Q$. There the stationary point of $\mathcal{M}$ is given as solution to (6.56), (6.57) with $P^{\mathrm{ad}} = Q$. A similar linear-quadratic optimization problem is considered in (Griesse and Vexler [69]) in the context of sensitivity analysis.

Remark 6.7. If we assume that the second-order sufficient condition from Theorem 2.4 holds, the linear-quadratic optimization problem (6.56) possesses a solution. This is the case as the quadratic part $\tilde{\mathcal{L}}''_{uq}(\overline{\chi})(v,p) + \frac{1}{2}\tilde{\mathcal{L}}''_{uu}(\overline{\chi})(v,v) + \frac{1}{2}\tilde{\mathcal{L}}''_{qq}(\overline{\chi})(p,p)$ of $K(p,v)$ is positive definite (see (2.11)) for all solutions to the linear equation (2.7) which is exactly the same as (6.57).

We introduce an auxiliary Lagrangian $\mathcal{N}\colon \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ for (6.56), (6.57) by

$$\mathcal{N}(\overline{\chi}, p, v, y) = K(\overline{\chi}, p, v) + \tilde{\mathcal{L}}''_{uz}(\overline{\chi})(v, y) + \tilde{\mathcal{L}}''_{qz}(\overline{\chi})(p, y). \tag{6.61}$$

For a solution $(\overline{p}, \overline{v})$ to (6.56), (6.57) the following first-order necessary condition holds: There exists $\overline{y} \in V$ such that:

$$\begin{aligned}
\mathcal{N}'_y(\overline{\chi}, \overline{p}, \overline{v}, \overline{y})(\delta y) &= 0 \quad \forall \delta y \in V, && \text{(6.62a)}\\
\mathcal{N}'_v(\overline{\chi}, \overline{p}, \overline{v}, \overline{y})(\delta v) &= 0 \quad \forall \delta v \in V, && \text{(6.62b)}\\
\mathcal{N}'_p(\overline{\chi}, \overline{p}, \overline{v}, \overline{y})(\delta p - \overline{p}) &\geq 0 \quad \forall \delta p \in P^{\mathrm{ad}}, && \text{(6.62c)}
\end{aligned}$$

or if written more explicitly:

$$\begin{aligned}
\tilde{\mathcal{L}}''_{uz}(\overline{\chi})(\overline{v}, \delta y) + \tilde{\mathcal{L}}''_{qz}(\overline{\chi})(\overline{p}, \delta y) &= 0 \quad \forall \delta y \in V, && \text{(6.63a)}\\
I'_u(\overline{q}, \overline{u})(\delta v) + \tilde{\mathcal{L}}''_{uq}(\overline{\chi})(\delta v, \overline{p}) + \tilde{\mathcal{L}}''_{uu}(\overline{\chi})(\delta v, \overline{v}) + \tilde{\mathcal{L}}''_{uz}(\overline{\chi})(\delta v, \overline{y}) &= 0 \quad \forall \delta v \in V, && \text{(6.63b)}\\
I'_q(\overline{q}, \overline{u})(\delta p) + \tilde{\mathcal{L}}''_{uq}(\overline{\chi})(\overline{v}, \delta p) + \tilde{\mathcal{L}}''_{qq}(\overline{\chi})(\delta p, \overline{p}) + \tilde{\mathcal{L}}''_{qz}(\overline{\chi})(\delta p, \overline{y}) &\geq 0 \quad \forall \delta p \in P^{\mathrm{ad}} - \overline{p}. && \text{(6.63c)}
\end{aligned}$$

Again, we can introduce the full Lagrangian $\tilde{\mathcal{N}}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ by

$$\tilde{\mathcal{N}}(\chi, \psi) = \mathcal{N}(\chi, p, v, y) + (\,\nu^-\,, p_- - p\,)_Q + (\,\nu^+\,, p - p_+\,)_Q. \tag{6.64}$$

As in (6.17a)—(6.17f) we can rewrite the necessary optimality condition for $\overline{\psi} \in \bar{\mathcal{Y}}^{\mathrm{ad}}$ as

$$\begin{aligned}
\tilde{\mathcal{N}}'_v(\overline{\chi}, \overline{\psi})(\delta v) &= 0 \quad \forall \delta v \in V, && \text{(6.65a)}\\
\tilde{\mathcal{N}}'_p(\overline{\chi}, \overline{\psi})(\delta p) &= 0 \quad \forall \delta p \in Q, && \text{(6.65b)}\\
\tilde{\mathcal{N}}'_y(\overline{\chi}, \overline{\psi})(\delta y) &= 0 \quad \forall \delta y \in V, && \text{(6.65c)}\\
\tilde{\mathcal{N}}'_{\nu^-}(\overline{\chi}, \overline{\psi})(\delta \nu^-) &= 0 \quad \forall \delta \nu^- \in P_-, && \text{(6.65d)}\\
\tilde{\mathcal{N}}'_{\nu^+}(\overline{\chi}, \overline{\psi})(\delta \nu^+) &= 0 \quad \forall \delta \nu^+ \in P_+, && \text{(6.65e)}\\
\overline{\nu}^+ - \overline{\nu}^- = \overline{\nu}, \quad \overline{\nu}^-(p_- - \overline{p}) &= \overline{\nu}^+(\overline{p} - p_+) = 0 \quad \text{a.e. on } \omega, && \text{(6.65f)}\\
\operatorname{supp} \nu^+ \subseteq \omega \setminus \{\, x \in \omega \,|\, \overline{q} = q_- \text{ and } \overline{\mu} \neq 0 \,\}, \quad \overline{\nu}^+ &\geq 0 \quad \text{a.e. where } \overline{\mu} = 0, && \text{(6.65g)}\\
\operatorname{supp} \nu^- \subseteq \omega \setminus \{\, x \in \omega \,|\, \overline{q} = q_+ \text{ and } \overline{\mu} \neq 0 \,\}, \quad \overline{\nu}^- &\geq 0 \quad \text{a.e. where } \overline{\mu} = 0. && \text{(6.65h)}
\end{aligned}$$

To be more clear, $\overline{\nu}^-$ and $\overline{\nu}^+$ are given by the following relations depending on $\overline{\nu} = -\mathcal{N}_p'(\overline{\chi}, \overline{p}, \overline{v}, \overline{y})(\cdot)$:

$$\overline{\nu}^+(x) = \begin{cases} \overline{\nu} & \overline{q}(x) = b \text{ and } \overline{\mu}(x) \neq 0, \\ 0 & \overline{q}(x) = a \text{ and } \overline{\mu}(x) \neq 0, \\ \max(0, \overline{\nu}) & \text{else}, \end{cases}$$

$$\overline{\nu}^-(x) = \begin{cases} \overline{\nu} & \overline{q}(x) = a \text{ and } \overline{\mu}(x) \neq 0, \\ 0 & \overline{q}(x) = b \text{ and } \overline{\mu}(x) \neq 0, \\ \max(0, -\overline{\nu}) & \text{else}. \end{cases}$$

By this construction we obtain that $\overline{\nu}^- \in Q_- \cap P_-$ and $\overline{\nu}^+ \in Q_+ \cap P_+$, in particular (6.55) holds for $\overline{\psi} = (\overline{p}, \overline{v}, \overline{y}, \overline{\nu}^-, \overline{\nu}^+$. However, $\{\, \overline{p} = p_- \,\} \subset \omega_- \cup \omega_+$ and $\{\, \overline{p} = p_+ \,\} \subset \omega_- \cup \omega_+$. Hence that $Q_- \subset P_\pm$ and $Q_+ \subset P_\pm$ and hence functions from Q_- and Q_+ are valid test functions in (6.65d) and (6.65d). In general, $\nu^\pm$ are not feasible for (6.17d) or (6.17e) respectively. However, if we assume strict complementarity, e.g., the set $\{q = a \,\text{and}\, \mu = 0\} \cup \{q = b \,\text{and}\, \mu = 0\}$ have zero measure, we obtain that $\nu^- \in Q_-$ and $\nu^+ \in Q_+$.

Remark 6.8. It should be noted that we use the convention $\pm\infty \cdot 0 = 0$ in (6.64), (6.65f) to ease notation. The same convention will be used throughout this section.

Remark 6.9. The condition (6.65g) arises naturally, as $\overline{\nu}^+$ is the Lagrange multiplier which corresponds to the equality and inequality constraints for $\overline{p}$ that are induced by the active upper control bound b. Similarly (6.65h) arises from the active lower control bound a.

We discretize using the discretized admissible set

$$P_h^{\text{ad}} = \{\, p \in Q_h \,|\, p_{h,-}(x) \leq p(x) \leq p_{h,-}(x) \text{ a.e. on } \omega \,\}, \tag{6.66}$$

with the bounds

$$p_{h,-}(x) = \begin{cases} 0 & \mu_h(x) \neq 0 \text{ or } q_h(x) = a, \\ -\infty & \text{else}, \end{cases}$$

$$p_{h,+}(x) = \begin{cases} 0 & \mu_h(x) \neq 0 \text{ or } q_h(x) = b, \\ +\infty & \text{else}. \end{cases}$$

We introduce

$$\bar{\mathcal{Y}}_h^{\text{ad}} = P_h^{\text{ad}} \times V_h \times V_h \times P_{-,h} \times P_{+,h} \tag{6.67}$$

to shorten notation. Where $P_{\pm,h}$ is defined analog to $Q_{\pm,h}$ as

$$P_{-,h} = \{\, r \in Q_h \,|\, r(x) = 0 \text{ a.e. on } \omega \setminus \{\overline{p}_h = p_-\} \,\},$$
$$P_{+,h} = \{\, r \in Q_h \,|\, r(x) = 0 \text{ a.e. on } \omega \setminus \{\overline{p}_h = p_+\} \,\}.$$

Then the following first-order condition holds with the discretized full Lagrangian

$$\tilde{\mathcal{N}}_h(\chi, \psi) = \mathcal{N}(\chi, p, v, y) + (\, \nu^- \,, p_{h,-} - p \,)_Q + (\, \nu^+ \,, p - p_{h,+} \,)_Q$$

$\tilde{\mathcal{N}}_h \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Let $\overline{p_h} \in P_h^{\mathrm{ad}}$ and $\overline{v}_h \in V_h$ be a solution to (6.56)—(6.57) There exists $\overline{y}_h \in V_h$, $\overline{\nu}_h^+, \overline{\nu}_h^- \in Q_h$ such that for $\overline{\psi}_h = (\overline{p}_h, \overline{v}_h, \overline{y}_h, \overline{\nu}_h^-, \overline{\nu}_h^+) \in \bar{\mathcal{Y}}_h^{\mathrm{ad}}$ the following holds:

$$\tilde{\mathcal{N}}'_{h,v}(\overline{\chi}_h, \overline{\psi}_h)(\delta v) = 0 \quad \forall \delta v \in V_h, \tag{6.68a}$$
$$\tilde{\mathcal{N}}'_{h,p}(\overline{\chi}_h, \overline{\psi}_h)(\delta p) = 0 \quad \forall \delta p \in Q_h, \tag{6.68b}$$
$$\tilde{\mathcal{N}}'_{h,y}(\overline{\chi}_h, \overline{\psi}_h)(\delta y) = 0 \quad \forall \delta y \in V_h, \tag{6.68c}$$
$$\tilde{\mathcal{N}}'_{h,\nu^-}(\overline{\chi}_h, \overline{\psi}_h)(\delta \nu^-) = 0 \quad \forall \delta \nu^- \in P_{-,h}, \tag{6.68d}$$
$$\tilde{\mathcal{N}}'_{h,\nu^+}(\overline{\chi}_h, \overline{\psi}_h)(\delta \nu^+) = 0 \quad \forall \delta \nu^+ \in P_{+,h}. \tag{6.68e}$$
$$\overline{\nu}_h^+ - \overline{\nu}_h^- = \overline{\nu}_h, \quad (\,\overline{\nu}_h^-\,, p_{h,-} - \overline{p}_h\,)_Q = (\,\overline{\nu}_h^+\,, \overline{p}_h - p_{h,+}\,)_Q = 0, \tag{6.68f}$$
$$(\overline{\nu}_h^+, \psi_i) = 0 \ \forall i : (\overline{\mu}_h, \psi_i) \neq 0 \text{ and } \overline{q}_i = q_-, \tag{6.68g}$$
$$(\overline{\nu}_h^+, \psi_i) \geq 0 \ \forall i : (\overline{\mu}_h, \psi_i) = 0, \tag{6.68h}$$
$$(\overline{\nu}_h^-, \psi_i) = 0 \ \forall i : (\overline{\mu}_h, \psi_i) \neq 0 \text{ and } \overline{q}_i = q_+, \tag{6.68i}$$
$$(\overline{\nu}_h^-, \psi_i) \geq 0 \ \forall i : (\overline{\mu}_h, \psi_i) = 0, \tag{6.68j}$$

For the error estimator with respect to the quantity of interest we introduce the residual functionals $\tilde{\rho}_v(\chi_h, \psi_h)(\cdot)$, $\tilde{\rho}_y(\chi_h, \psi_h)(\cdot) \in V^*$ and $\tilde{\rho}_p(\chi_h, \psi_h)(\cdot)$, $\tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\cdot)$, $\tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\cdot) \in Q^*$ by

$$\tilde{\rho}_v(\chi_h, \psi_h)(\cdot) = \tilde{\mathcal{L}}''_{zu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{zq}(\chi_h)(\cdot, p_h), \tag{6.69}$$
$$\tilde{\rho}_y(\chi_h, \psi_h)(\cdot) = I'_u(q_h, u_h)(\cdot) + \tilde{\mathcal{L}}''_{uu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{uz}(\chi_h)(\cdot, y_h) + \tilde{\mathcal{L}}''_{uq}(\chi_h)(\cdot, p_h), \tag{6.70}$$
$$\tilde{\rho}_p(\chi_h, \psi_h)(\cdot) = I'_q(q_h, u_h)(\cdot) + \tilde{\mathcal{L}}''_{qu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{qz}(\chi_h)(\cdot, y_h) + \tilde{\mathcal{L}}''_{qq}(\chi_h)(\cdot, p_h) + (\,\cdot\,, \nu_h\,)_Q, \tag{6.71}$$
$$\tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\cdot) = -(\,\cdot\,, p_h\,)_Q, \tag{6.72}$$
$$\tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\cdot) = (\,\cdot\,, p_h\,)_Q, \tag{6.73}$$

in addition to the already defined residual functionals (6.34)—(6.43). Again the last two residual functionals also have to be evaluated in the point (χ, ψ) where they read as follows:

$$\tilde{\rho}_{\nu^-}(\chi, \psi)(\cdot) = -(\,\cdot\,, p\,)_Q, \quad \tilde{\rho}_{\nu^+}(\chi, \psi)(\cdot) = (\,\cdot\,, p\,)_Q.$$

Theorem 6.3. *Let $\overline{\chi} \in \mathcal{Y}^{ad}$ be a solution to the necessary optimality condition* (6.17) *and $\overline{\chi}_h \in \mathcal{Y}_h^{ad}$ be its Galerkin approximation* (6.32)*. In addition let $\overline{\psi} \in \bar{\mathcal{Y}}^{ad}$ be a solution to the necessary optimality condition* (6.65) *of the auxiliary optimization problem* (6.56), (6.57) *and*

$\overline{\psi}_h \in \bar{\mathcal{Y}}_h^{ad}$ *be its discrete approximation* (6.68)*. Then the following estimate holds:*

$$\begin{aligned} I(\overline{q},\overline{u}) - I(\overline{q}_h,\overline{u}_h) = {} & \frac{1}{2}\rho_u(\overline{\chi}_h)(\overline{y}-\tilde{y}_h) + \frac{1}{2}\rho_z(\overline{\chi}_h)(\overline{v}-\tilde{v}_h) + \frac{1}{2}\tilde{\rho}_q(\overline{\chi}_h)(\overline{p}-\tilde{p}_h) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\overline{\nu}^- - \tilde{\nu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\overline{\nu}^+ - \tilde{\nu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_v(\overline{\chi}_h,\overline{\psi}_h)(\overline{z}-\tilde{z}_h) + \frac{1}{2}\tilde{\rho}_y(\overline{\chi}_h,\overline{\psi}_h)(\overline{u}-\tilde{u}_h) + \frac{1}{2}\tilde{\rho}_p(\overline{\chi}_h,\overline{\psi}_h)(\overline{q}-\tilde{q}_h) \\ & + \frac{1}{2}\tilde{\rho}_{\nu^-}(\overline{\chi}_h,\overline{\psi}_h)(\overline{\mu}^- - \tilde{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\nu^+}(\overline{\chi}_h,\overline{\psi}_h)(\overline{\mu}^+ - \tilde{\mu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\overline{\chi})(\tilde{\nu}^- - \overline{\nu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\overline{\chi})(\tilde{\nu}^+ - \overline{\nu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_{\nu^-}(\overline{\chi},\overline{\psi})(\tilde{\mu}^- - \overline{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\nu^+}(\overline{\chi},\overline{\psi})(\tilde{\mu}^+ - \overline{\mu}_h^+) + R_3, \end{aligned} \tag{6.74}$$

where $\tilde{u}_h,\ \tilde{v}_h,\ \tilde{z}_h,\ \tilde{y}_h \in V_h,\ \tilde{q}_h,\ \tilde{p}_h \in Q_h,\ \tilde{\mu}_h^-, \tilde{\nu}_h^- \in Q_{-,h},\ \tilde{\mu}_h^+, \tilde{\nu}_h^+ \in Q_{+,h}$*, as well as* $\tilde{\mu}^-, \tilde{\nu}^- \in Q_-,\ \tilde{\mu}^+, \tilde{\nu}^+ \in Q_+$*, are arbitrarily chosen and* R_3 *is a remainder term given by*

$$R_3 = \frac{1}{2}\int_0^1 \mathcal{M}'''((\overline{\chi}_h,\overline{\psi}_h)+se)(e,e,e)s(s-1)\,ds, \tag{6.75}$$

with $e = (\overline{\chi}-\overline{\chi}_h, \overline{\psi}-\overline{\psi}_h)$*.*

Proof. From (6.55) and the analog discrete result we obtain

$$I(\overline{q},\overline{u}) - I(\overline{q}_h,\overline{u}_h) = \mathcal{M}(\overline{\chi},\overline{\psi}) - \mathcal{M}(\overline{\chi}_h,\overline{\psi}_h) = \int_0^1 \mathcal{M}'((\overline{\chi}_h,\overline{\psi}_h)+se)(e)\,ds.$$

Approximation by the trapezoidal rule gives:

$$I(\overline{q},\overline{u}) - I(\overline{q}_h,\overline{u}_h) = \frac{1}{2}\mathcal{M}'(\overline{\chi},\overline{\psi})(e) + \frac{1}{2}\mathcal{M}'(\overline{\chi}_h,\overline{\psi}_h)(e) + R_3 \tag{6.76}$$

with the remainder term R_3 as in (6.75). For the first term we have:

$$\begin{aligned} \mathcal{M}'(\overline{\chi},\overline{\psi})(e) \;=\; & \mathcal{M}'_u(\overline{\chi},\overline{\psi})(\overline{u}-\overline{u}_h) + \mathcal{M}'_v(\overline{\chi},\overline{\psi})(\overline{v}-\overline{v}_h) \\ & + \mathcal{M}'_z(\overline{\chi},\overline{\psi})(\overline{z}-\overline{z}_h) + \mathcal{M}'_y(\overline{\chi},\overline{\psi})(\overline{y}-\overline{y}_h) \\ & + \mathcal{M}'_q(\overline{\chi},\overline{\psi})(\overline{q}-\overline{q}_h) + \mathcal{M}'_p(\overline{\chi},\overline{\psi})(\overline{p}-\overline{p}_h) \\ & + \mathcal{M}'_{\mu^-}(\overline{\chi},\overline{\psi})(\overline{\mu}^- - \overline{\mu}_h^-) + \mathcal{M}'_{\nu^-}(\overline{\chi},\overline{\psi})(\overline{\nu}^- - \overline{\nu}_h^-) \\ & + \mathcal{M}'_{\mu^+}(\overline{\chi},\overline{\psi})(\overline{\mu}^+ - \overline{\mu}_h^+) + \mathcal{M}'_{\nu^+}(\overline{\chi},\overline{\psi})(\overline{\nu}^+ - \overline{\nu}_h^+). \end{aligned}$$

Using the identities:

$$\begin{aligned} \mathcal{M}'_u(\overline{\chi},\overline{\psi})(\cdot) &= \tilde{\mathcal{N}}'_v(\overline{\chi},\overline{p},\overline{v},\overline{y})(\cdot), \quad \mathcal{M}'_v(\overline{\chi},\overline{\psi})(\cdot) = \tilde{\mathcal{L}}'_u(\overline{\chi})(\cdot), \\ \mathcal{M}'_z(\overline{\chi},\overline{\psi})(\cdot) &= \tilde{\mathcal{N}}'_y(\overline{\chi},\overline{p},\overline{v},\overline{y})(\cdot), \quad \mathcal{M}'_y(\overline{\chi},\overline{\psi})(\cdot) = \tilde{\mathcal{L}}'_z(\overline{\chi})(\cdot), \\ \mathcal{M}'_q(\overline{\chi},\overline{\psi})(\cdot) &= \tilde{\mathcal{N}}'_p(\overline{\chi},\overline{p},\overline{v},\overline{y})(\cdot), \quad \mathcal{M}'_p(\overline{\chi},\overline{\psi})(\cdot) = \tilde{\mathcal{L}}'_q(\overline{\chi})(\cdot), \end{aligned}$$

we see that the first six terms on the right-hand side vanish due to (6.17a)—(6.17c) and (6.65a)—(6.65c). Furthermore we see from (6.17d), (6.17e) that with arbitrary $\tilde{\mu}^-$, $\tilde{\nu}^- \in Q_-$, and $\tilde{\mu}^+$, $\tilde{\nu}^+ \in Q_+$ the following identities hold:

$$\mathcal{M}'_{\mu^-}(\overline{\chi},\overline{\psi})(\overline{\mu}^- - \overline{\mu}_h^-) = \mathcal{M}'_{\mu^-}(\overline{\chi},\overline{\psi})(\tilde{\mu}^- - \overline{\mu}_h^-) = \tilde{\rho}_{\nu^-}(\overline{\chi},\overline{\psi})(\tilde{\mu}^- - \overline{\mu}_h^-), \tag{6.77}$$

$$\mathcal{M}'_{\nu^-}(\overline{\chi},\overline{\psi})(\overline{\nu}^- - \overline{\nu}_h^-) = \mathcal{M}'_{\nu^-}(\overline{\chi},\overline{\psi})(\tilde{\nu}^- - \overline{\nu}_h^-) = \tilde{\rho}_{\mu^-}(\overline{\chi})(\tilde{\nu}^- - \overline{\nu}_h^-), \tag{6.78}$$

$$\mathcal{M}'_{\mu^+}(\overline{\chi},\overline{\psi})(\overline{\mu}^+ - \overline{\mu}_h^+) = \mathcal{M}'_{\mu^+}(\overline{\chi},\overline{\psi})(\tilde{\mu}^+ - \overline{\mu}_h^+) = \tilde{\rho}_{\nu^+}(\overline{\chi},\overline{\psi})(\tilde{\mu}^+ - \overline{\mu}_h^+), \tag{6.79}$$

$$\mathcal{M}'_{\nu^+}(\overline{\chi},\overline{\psi})(\overline{\nu}^+ - \overline{\nu}_h^+) = \mathcal{M}'_{\nu^+}(\overline{\chi},\overline{\psi})(\tilde{\nu}^+ - \overline{\nu}_h^+) = \tilde{\rho}_{\mu^+}(\overline{\chi})(\tilde{\nu}^+ - \overline{\nu}_h^+). \tag{6.80}$$

Thus we obtain:

$$\begin{aligned}\mathcal{M}'(\overline{\chi},\overline{\psi})(e) \;=\; & \tilde{\rho}_{\mu^-}(\overline{\chi})(\tilde{\nu}^- - \overline{\nu}_h^-) + \tilde{\rho}_{\mu^+}(\overline{\chi})(\tilde{\nu}^+ - \overline{\nu}_h^+) \\ & + \tilde{\rho}_{\nu^-}(\overline{\chi},\overline{\psi})(\tilde{\mu}^- - \overline{\mu}_h^-) + \tilde{\rho}_{\nu^+}(\overline{\chi},\overline{\psi})(\tilde{\mu}^+ - \overline{\mu}_h^+).\end{aligned}$$

For the second term we obtain from (6.32a)—(6.32e) and (6.65a)—(6.65e) that

$$\mathcal{M}'(\overline{\chi}_h,\overline{\psi}_h)(e) = \mathcal{M}'(\overline{\chi}_h,\overline{\psi}_h)(\overline{\chi} - \tilde{\chi}_h, \overline{\psi} - \tilde{\psi}_h)$$

for each $\tilde{\chi}_h, \tilde{\psi}_h \in \tilde{\mathcal{Y}}_h^{\mathrm{ad}}$ which completes the proof. □

Remark 6.10. Note that in the case $I = J$ the solution $(\overline{p},\overline{v},\overline{y})$ to (6.62) is given by $(0,0,\overline{z})$, which can be seen after some calculations. Using this, one obtains that for $I = J$ the estimates in Theorem 6.2 and Theorem 6.3 coincide.

We define the projection onto the admissible set by

$$\mathcal{P}_{P_h^{\mathrm{ad}}}(p) = \max\bigl(p_{h,-}, \min(p, p_{h,+})\bigr).$$

To obtain a computable error estimator we introduce $\tilde{p} \in P^{\mathrm{ad}}$ as approximation to $\overline{p}$ by

$$\tilde{p} = \mathcal{P}_{P_h^{\mathrm{ad}}}\biggl(\frac{1}{\alpha}(a'_q()(\cdot,\pi\overline{y}_h) + a''_{qu}()(\cdot,\pi\overline{v}_h,\pi\overline{z}_h) + a''_{qq}()(\cdot,\overline{p}_h,\pi\overline{z}_h) - I'_q(\pi^q\overline{q}_h,\pi\overline{u}_h)(\cdot))\biggr), \tag{6.81}$$

where () is an abbreviation for $(\pi^q\overline{q}_h,\pi\overline{u}_h)$, and $\tilde{\nu}$ is introduced as approximation to $\overline{\nu}$ by

$$\tilde{\nu} = -\alpha\tilde{p} + a'_q()(\cdot,\pi\overline{y}_h) + a''_{qu}()(\cdot,\pi\overline{v}_h,\pi\overline{z}_h) + a''_{qq}()(\cdot,\overline{p}_h,\pi\overline{z}_h) - I'_q(\pi^q\overline{q}_h,\pi\overline{u}_h)(\cdot), \tag{6.82}$$

which is analog to the construction of the approximations $\pi^q\overline{q}_h$ and $\tilde{\mu}$ in (6.40) and (6.52).

Using these approximations we obtain the following computable error estimator:

$$\begin{aligned}\eta_{\mathrm{QI}} \;=\; & \frac{1}{2}\rho_u(\overline{\chi}_h)(\pi y - \overline{y}_h) + \frac{1}{2}\rho_z(\overline{\chi}_h)(\pi v - \overline{v}_h) + \frac{1}{2}\tilde{\rho}_q(\overline{\chi}_h)(\tilde{p} - \overline{p}_h) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\overline{\chi}_h)(\tilde{\nu}^- - \overline{\nu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\overline{\chi}_h)(\tilde{\nu}^+ - \overline{\nu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_v(\overline{\chi}_h,\overline{\psi}_h)(\pi z - \overline{z}_h) + \frac{1}{2}\tilde{\rho}_y(\overline{\chi}_h,\overline{\psi}_h)(\pi u - \overline{u}_h) + \frac{1}{2}\tilde{\rho}_p(\overline{\chi}_h,\overline{\psi}_h)(\pi^q\overline{q}_h - \overline{q}_h) \\ & + \frac{1}{2}\tilde{\rho}_{\nu^-}(\overline{\chi}_h,\overline{\psi}_h)(\tilde{\mu}^- - \overline{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\nu^+}(\overline{\chi}_h,\overline{\psi}_h)(\tilde{\mu}^+ - \overline{\mu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_{\mu^-}(\tilde{\chi})(\tilde{\nu}^- - \overline{\nu}_h^-) + \frac{1}{2}\tilde{\rho}_{\mu^+}(\tilde{\chi})(\tilde{\nu}^+ - \overline{\nu}_h^+) \\ & + \frac{1}{2}\tilde{\rho}_{\nu^-}(\tilde{\chi},\tilde{\psi})(\tilde{\mu}^- - \overline{\mu}_h^-) + \frac{1}{2}\tilde{\rho}_{\nu^+}(\tilde{\chi},\tilde{\psi})(\tilde{\mu}^+ - \overline{\mu}_h^+),\end{aligned}$$

where $\tilde{\chi} = (\pi^q\overline{q}_h, \pi u, \pi z, \tilde{\mu}^-, \tilde{\mu}^+)$ and $\tilde{\psi} = (\tilde{p}, \pi v, \pi y, \tilde{\nu}^-, \tilde{\nu}^+)$.

Remark 6.11. We would like to point out that in case of strict complementarity, e.g., if the set

$$\{\, x \in \omega \mid \overline{q}(x) = a \text{ or } \overline{q}(x) = b\} \setminus \{x \in \omega \mid \overline{\mu}(x) \neq 0\,\}$$

has zero measure, the auxiliary problem (6.56), (6.57) does not involve inequality constraints for the controls. In that case the set P^{ad} is not only convex but in fact a real subspace of Q.

Remark 6.12. The constrained linear-quadratic optimization problem (6.56), (6.57) can be solved using primal-dual active set strategy. In the case of strict complementarity the algorithm will converge in one step due to the fact that P^{ad} is a linear subspace of Q is this case.

Remark 6.13. Due to the definition of P^{ad} (6.58), the solution $\overline{p} \in Q$ of auxiliary optimization problem (6.56)—(6.57) is usually discontinuous. Therefore, a cell-wise constant discretization of the control space Q seems to be more suitable than a discretization with continuous trial functions if the error with respect to a quantity of interested is estimated.

6.1.3 Numerical Results

In this section we discuss two numerical examples illustrating the behavior of our method. For both examples we use bilinear (H^1-conforming) finite elements for the discretization of the state variable and cell-wise constant discretization of the control space. The optimization problems are solved by primal-dual-active-set strategy that will be sketched in Section 7.1.

All examples have been computed using the optimization library RoDoBo (RoDoBo [126]) and the finite element toolkit Gascoigne (Gascoigne [65]).

Example on a Domain with a Hole We consider the following nonlinear optimization problem:

$$\text{Minimize} \quad \frac{1}{2}\|u - u^d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|q\|^2_{L^2(\Omega)}, \qquad u \in V,\ q \in Q^{\mathrm{ad}}, \tag{6.83}$$

subject to

$$\begin{aligned} -\Delta u + 30\,u^3 + u &= f + q && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{6.84}$$

where $\Omega = \omega = (0,1)^2 \setminus [0.4, 0.6]^2$, $V = H^1_0(\Omega)$, $Q = L^2(\Omega)$, and the admissible set Q^{ad} is given by

$$Q^{\mathrm{ad}} = \{\, q \in Q \mid -7 \leq q(x) \leq 20 \text{ a.e. on } \Omega\,\}.$$

The desired state u^d and the right-hand side f are defined as

$$u^d(x) = x_1 \cdot x_2, \quad f(x) = (x_1 - 0.5)^{-2}(x_2 - 0.5)^{-2}.$$

The regularization parameter is chosen $\alpha = 10^{-4}$. We note that the state equation (6.84) is a monotone semi-linear equation, which possesses a unique solution $u \in V$ for each $q \in Q$. The proof of the existence of a global solution as well as derivation of necessary and sufficient optimality conditions for the corresponding optimization problem (6.83)—(6.84) can be found, e.g., in (Tröltzsch [141]).

Table 6.1: Effectivity indices

(a) Random refinement

N	$I_{\text{eff}}(\eta_h^{(1)})$	$I_{\text{eff}}(\eta_h^{(2)})$	$I_{\text{eff}}(\eta_{\text{QI}})$
432	1.1	1.1	1.2
906	1.1	1.1	1.1
2328	1.3	1.2	2.3
5752	1.2	1.2	1.4
13872	1.3	1.3	1.5
33964	1.3	1.3	1.4
83832	1.2	1.2	1.5

(b) Refinement according to η_{QI}

N	$I_{\text{eff}}(\eta_h^{(1)})$	$I_{\text{eff}}(\eta_h^{(2)})$	$I_{\text{eff}}(\eta_{\text{QI}})$
432	1.1	1.1	1.1
824	1.1	1.1	1.4
1692	1.0	1.0	0.3
3992	1.0	1.0	0.2
11396	1.0	1.0	0.5
30604	1.0	1.0	1.0
80354	1.0	1.0	1.3

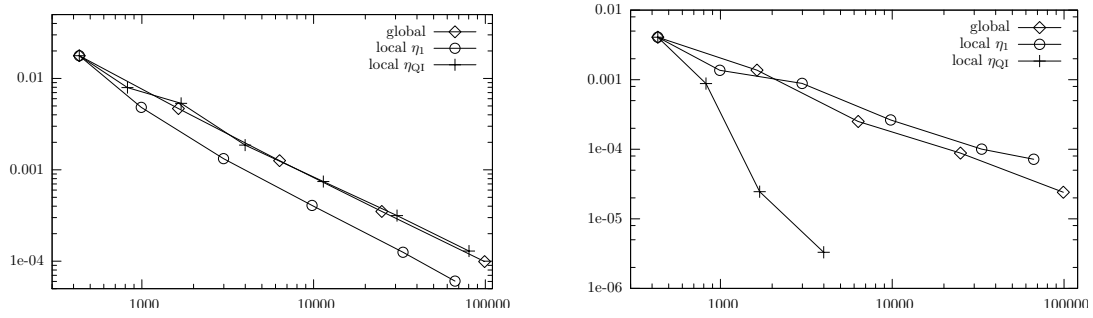
In Section 6.1.2 we derived two different error estimators for the error with respect to the cost functional and one error estimator with respect to a quantity of interest. In this example, we choose the quantity of interest as

$$I(q,u) = \frac{1}{2}\int_{(0.7,0.8)^2} |\nabla u(x)|^2\,dx \; + \int_{(0.2,0.3)^2} q(x)\,dx. \tag{6.85}$$

In order to check the quality of the error estimators, we define the following effectivity indices:

$$I_{\text{eff}}(\eta_h^{(1)}) = \frac{J(u) - J(u_h)}{\eta_h^{(1)}}, \quad I_{\text{eff}}(\eta_h^{(2)}) = \frac{J(u) - J(u_h)}{\eta_h^{(2)}}, \quad I_{\text{eff}}(\eta_{\text{QI}}) = \frac{I(q,u) - I(q_h,u_h)}{\eta_{\text{QI}}}\,. \tag{6.86}$$

In Table 6.1 these effectivity indices are listed for different types of mesh refinement: random refinement and refinement based on the error estimator η_{QI} for the quantity of interest.



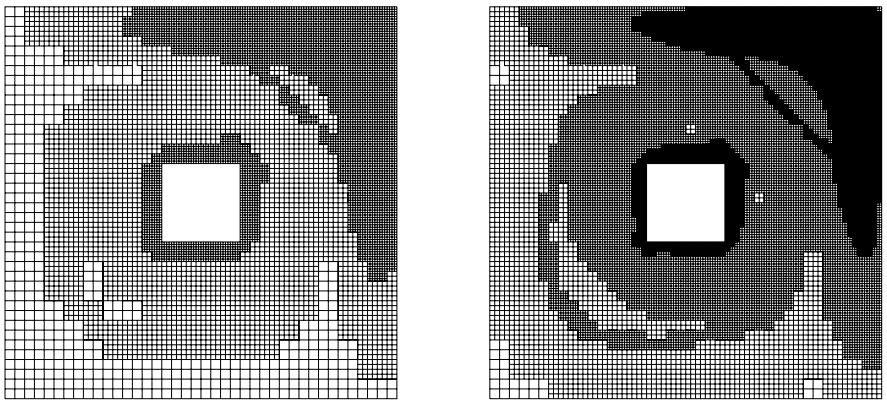
(a) Error in J (b) Error in I

Figure 6.1: Discretization error for different refinement criteria

We observe that the error estimators provide quantitative information about the discretization error. We note that the results for $\eta_h^{(1)}$ and $\eta_h^{(2)}$ are very close to each other in this example, cf., Remark 6.5.

In addition, our results show that the local mesh refinement based on error estimators derived above leads to substantial saving in degrees of freedom for achieving a given level of the discretization error. In Figure 6.1 the dependence of discretization error on the number of degrees of freedom is shown for different refinement criteria: global (uniform) refinement, refinement based on the error estimator $\eta_h^{(1)}$ for the cost functional, and refinement based on the error estimator η_{QI} for the quantity of interest. In Figure 6.1a the error with respect to the cost functional (6.83) and in Figure 6.1b the error with respect to the quantity of interest (6.85) are considered, respectively.

We observe the best behavior of error with respect to the cost functional if the mesh is refined based on $\eta_h^{(1)}$ and the best behavior of error with respect to the quantity of interest for the refinement based on η_{QI}.



(a) Mesh 3 from η_1

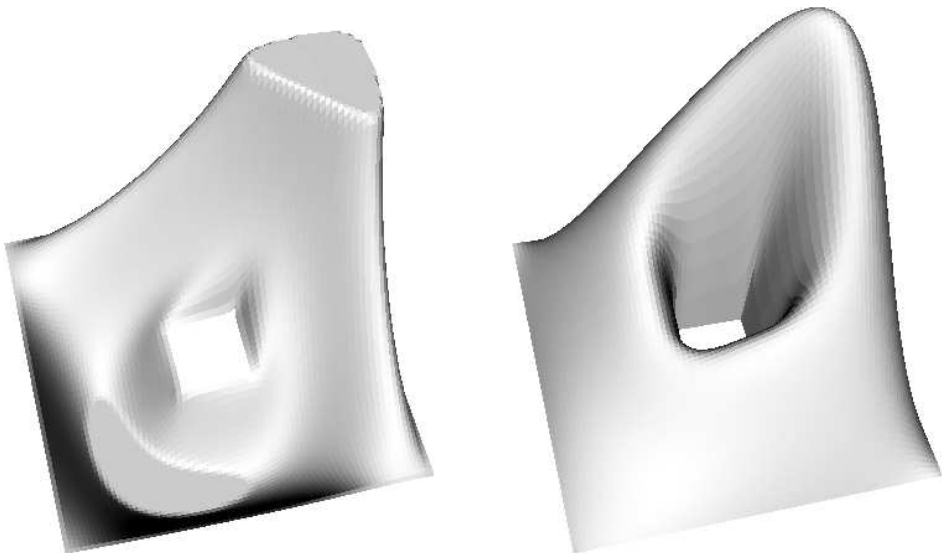
(b) Mesh 4 from η_1

(c) Mesh 4 from η_{QI}

(d) Mesh 5 from η_{QI}

Figure 6.2: Locally refined meshes

In Figure 6.2 a series of meshes generated according to the information obtained from the error estimators is shown. The corresponding optimal control $\overline{q}$ and corresponding state $\overline{u}$ are shown in Figure 6.3.



(a) Optimal control (b) State

Figure 6.3: Solution

Example with Bilinear State Equation Our second example is motivated by a parameter identification problem. The minimization problem is given by:

$$\text{Minimize} \quad \frac{1}{2}\|u - u^d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|q\|^2_{L^2(\Omega)}, \qquad u \in V,\ q \in Q^{\text{ad}}, \tag{6.87}$$

subject to

$$\begin{aligned} -\Delta u + qu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{6.88}$$

where $\Omega = \omega = (0, 0.5) \times (0, 1) \cup (0, 1) \times (0.5, 1)$, $V = H^1_0(\Omega)$, $Q = L^2(\Omega)$, and the admissible set Q^{ad} is given by

$$Q^{\text{ad}} = \{\, q \in Q \,|\, 0 \le q(x) \le 0.3 \text{ a.e. on } \Omega \,\}.$$

The desired state u^d and the right-hand side f are defined as

$$u^d(x) = \frac{1}{8\pi^2} \sin(2\pi x_1) \sin(2\pi x_2), \quad f(x) = 1.$$

The regularization parameter is chosen $\alpha = 10^{-4}$. Note that for any given $q \in Q^{\text{ad}}$ the state equation (6.88) possesses a unique solution $u \in V$ due to $q \ge 0$. For an a priori error analysis see, (Kröner [96], Kröner and Vexler [97]).

We are interested in the error in the unknown parameter, thus we choose

$$I(q, u) = \int_{\Omega_O} q(x)\, dx,$$

where $\Omega_O = (0, 0.25) \times (0.75, 1)$.

In Table 6.2 the effectivity indices, defined as in (6.86), are listed for different types of mesh refinement: global (uniform) refinement, random refinement, refinement based on the error estimator $\eta_h^{(1)}$ for the cost functional, and refinement based on the error estimator η_{QI} for the quantity of interest. As in the first example we observe that the error estimators provide quantitative information on the discretization errors. From Figure 6.4a, where the

Table 6.2: Effectivity indices

(a) Global refinement

N	$I_{\mathrm{eff}}(\eta_h^{(1)})$	$I_{\mathrm{eff}}(\eta_h^{(2)})$	$I_{\mathrm{eff}}(\eta_{\mathrm{QI}})$
65	1.2	1.2	2.0
225	1.3	1.2	1.9
833	1.4	1.4	1.5
3201	1.5	1.5	1.7

(b) Refinement according to η_1

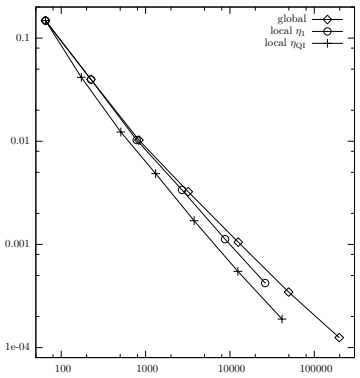
N	$I_{\mathrm{eff}}(\eta_h^{(1)})$	$I_{\mathrm{eff}}(\eta_h^{(2)})$	$I_{\mathrm{eff}}(\eta_{\mathrm{QI}})$
65	1.2	1.2	2.0
225	1.3	1.3	1.9
785	1.4	1.4	1.6
2705	1.5	1.5	1.7

(c) Random refinement

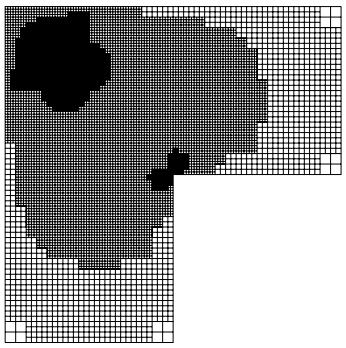
N	$I_{\mathrm{eff}}(\eta_h^{(1)})$	$I_{\mathrm{eff}}(\eta_h^{(2)})$	$I_{\mathrm{eff}}(\eta_{\mathrm{QI}})$
65	1.2	1.2	2.0
141	1.2	1.2	2.0
307	1.2	1.2	0.5
763	1.4	1.4	2.0

(d) Refinement according to η_{QI}

N	$I_{\mathrm{eff}}(\eta_h^{(1)})$	$I_{\mathrm{eff}}(\eta_h^{(2)})$	$I_{\mathrm{eff}}(\eta_{\mathrm{QI}})$
65	1.2	1.2	2.0
173	1.2	1.2	1.8
509	1.2	1.2	1.3
1317	1.2	1.2	1.3



(a) Error in I



(b) Mesh 5 from η_{QI}

Figure 6.4: Discretization error and mesh

discretization error with respect to the quantity of interest is plotted for different refinement criteria, we again observe that the local mesh refinement based on the appropriate error estimator leads to a certain saving in degrees of freedom for achieving a given tolerance for the discretization error. A typical mesh generated using the information obtained from η_{QI} is shown in Figure 6.4b.

6.2 Regularization Error for State Constraints

We will now consider the estimation of the regularization error. First with respect to the cost functional, to this end we will recall the case of a barrier regularization without control constraints, which has already been published in (Wollner [157]). Furthermore, we will take a more detailed look onto the behavior of the error estimators derived there. We extend the analysis of (Wollner [157]) to the case of active control constraints. Finally, we will consider estimates for the quadratic penalty method.

6.2.1 Estimates for the Cost Functional

6.2.1.1 Barrier Regularization without Control Constraints

A posteriori error estimation In this section we derive a posteriori estimates for the regularization error as well as for the discretization error with respect to the cost functional $J(q,u)$. Unfortunately, neither a solution to (2.10) provides feasible test functions for (5.32) nor is a solution to (5.32) feasible for (2.10). Therefore we split the estimation into two parts:

$$\begin{aligned} J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h,\overline{u}_\gamma^h) &= (J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma)) + (J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) - J_\gamma(\overline{q}_\gamma^h,\overline{u}_\gamma^h)) \\ &\approx \eta_{\text{hom}} + \eta_{\text{disc}}. \end{aligned}$$

Thus we estimate the error in the cost functional between the solution $(\overline{q},\overline{u})$ of (2.10) and the solution $(\overline{q}_\gamma,\overline{u}_\gamma)$ of the barrier problem (5.27), and then the discretization error between the solution of (5.27) and the solution $(\overline{q}_\gamma^h,\overline{u}_\gamma^h)$ to its discretization (5.32). To have a simple representation we assume that $Q^{\text{ad}} = Q$.

Homotopy error In order to estimate the error introduced by the homotopy parameter γ we define the Lagrangian $\mathcal{M}\colon Q \times W \times L^{t'}(\Omega) \times M(\Omega_C) \to \mathbb{R}$ by

$$\mathcal{M}(q,u,z,\mu) = J(q,u) + (f,\varphi) - a(q,u)(z) + \int_{\Omega_C} g(u,\nabla u)\, d\mu \tag{6.89}$$

where we consider $a(q,u)(z) := \langle\, A(q,u)\,, z\,\rangle_{Z^*\times Z}$ which for arguments $z \in V$ coincides with the definition of (2.3).

We can now formulate the following

Theorem 6.4. *Let $\overline{\xi} = (\overline{q},\overline{u},\overline{z},\overline{\mu})$ be a solution to the first-order necessary system* (2.10) *and let $\overline{\xi}_\gamma = (\overline{q}_\gamma,\overline{u}_\gamma,\overline{z}_\gamma,\overline{\mu}_\gamma)$ a solution to the first-order necessary system* (5.27) *of the barrier problem with sufficiently high order to obtain strictly feasible states. Then the following estimate holds:*

$$\begin{aligned} J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) = \frac{1}{2}\int_{\Omega_C} (g(\overline{u}_\gamma,\nabla\overline{u}_\gamma) - g(\overline{u},\nabla\overline{u}))\, d\overline{\mu}_\gamma \\ - B_\gamma(\overline{u}_\gamma) + \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma,\nabla\overline{u}_\gamma)\, d\overline{\mu} + \mathcal{R}_{hom}, \end{aligned} \tag{6.90}$$

with a remainder term $\mathcal{R}_{hom}$ given by:

$$\mathcal{R}_{hom} = \frac{1}{2}\int_0^1 \mathcal{M}'''(\overline{\xi}_\gamma + s(\overline{\xi} - \overline{\xi}_\gamma))(\overline{\xi} - \overline{\xi}_\gamma, \overline{\xi} - \overline{\xi}_\gamma, \overline{\xi} - \overline{\xi}_\gamma)s(s-1)\,ds. \tag{6.91}$$

Proof. From (2.10e) we conclude that the support of $\overline{\mu}$ is contained in the set

$$\mathcal{A} = \{\, x \in \Omega_C \mid g(\overline{u}, \nabla\overline{u})(x) = 0 \,\}.$$

Using this and (2.4), noting that $V \subset Z$ is dense, we obtain

$$J(\overline{q}, \overline{u}) = \mathcal{M}(\overline{q}, \overline{u}, \overline{z}, \overline{\mu}) = \mathcal{M}(\overline{\xi}).$$

Unfortunately, we do not have a complementarity condition for the solution to (5.27) thus utilizing (2.4) we obtain

$$\begin{aligned} J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) &= J(\overline{q}_\gamma, \overline{u}_\gamma) + B_\gamma(\overline{u}_\gamma) \\ &= \mathcal{M}(\overline{q}_\gamma, \overline{u}_\gamma, \overline{z}_\gamma, \overline{\mu}_\gamma) \\ &\quad - \int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\overline{\mu}_\gamma\,dx + B_\gamma(\overline{u}_\gamma) \\ &= \mathcal{M}(\overline{\xi}_\gamma) - \int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\overline{\mu}_\gamma\,dx + B_\gamma(\overline{u}_\gamma). \end{aligned}$$

Now we estimate the difference between the values of the Lagrangian $\mathcal{M}$ using the trapezoidal rule to evaluate the integral and obtain

$$\mathcal{M}(\overline{\xi}) - \mathcal{M}(\overline{\xi}_\gamma) = \int_0^1 \mathcal{M}'(\overline{\xi}_\gamma + s(\overline{\xi} - \overline{\xi}_\gamma))(\overline{\xi} - \overline{\xi}_\gamma)\,ds \tag{6.92}$$

$$= \frac{1}{2}\mathcal{M}'(\overline{\xi})(\overline{\xi} - \overline{\xi}_\gamma) + \frac{1}{2}\mathcal{M}'(\overline{\xi}_\gamma)(\overline{\xi} - \overline{\xi}_\gamma) + \mathcal{R}_{\text{hom}} \tag{6.93}$$

with the remainder $\mathcal{R}_{\text{hom}}$ as given in (6.91).

First we discuss $\mathcal{M}'(\overline{\xi})(\overline{\xi} - \overline{\xi}_\gamma)$, for that we consider the following functionals:

$$\begin{aligned} \mathcal{M}'_u(\overline{\xi})(\varphi) &= J'_u(\overline{q}, \overline{u})(\varphi) - a'_u(\overline{q}, \overline{u})(\varphi, \overline{z}) + \int_{\Omega_C} g'(\overline{u}, \nabla\overline{u})(\varphi)\,d\overline{\mu}, \\ \mathcal{M}'_z(\overline{\xi})(\varphi) &= (f, \varphi) - a(\overline{q}, \overline{u})(\varphi), \\ \mathcal{M}'_q(\overline{\xi})(\varphi) &= J'_q(\overline{q}, \overline{u})(\varphi) - a'_q(\overline{q}, \overline{u})(\varphi, \overline{z}), \\ \mathcal{M}'_\mu(\overline{\xi})(\varphi) &= \int_{\Omega_C} g(\overline{u}, \nabla\overline{u})\,d\varphi. \end{aligned}$$

Using (2.4) we obtain for the primal residual $\mathcal{M}'_z(\overline{\xi})(\overline{z} - \overline{z}_\gamma) = 0$, from (2.10b) we get for the adjoint residual $\mathcal{M}'_u(\overline{\xi})(\overline{u} - \overline{u}_\gamma) = 0$, and from (2.10c) together with the assumption $Q^{\text{ad}} = Q$ we deduce $\mathcal{M}'_q(\overline{\xi})(\overline{q} - \overline{q}_\gamma) = 0$ thus

$$\mathcal{M}'(\overline{\xi})(\overline{\xi} - \overline{\xi}_\gamma) = \mathcal{M}'_\mu(\overline{\xi})(\overline{\mu} - \overline{\mu}_\gamma).$$

We use complementarity to get

$$\mathcal{M}'(\overline{\xi})(\overline{\xi} - \overline{\xi}_\gamma) = \mathcal{M}'_\mu(\overline{\xi})(-\overline{\mu}_\gamma) = \int_{\Omega_C} -g(\overline{u}, \nabla\overline{u})\, d\overline{\mu}_\gamma.$$

Now we take a closer look on $\mathcal{M}'(\overline{\xi}_\gamma)(\overline{\xi} - \overline{\xi}_\gamma)$. Here we have

$$\begin{aligned}
\mathcal{M}'_u(\overline{\xi}_\gamma)(\varphi) &= J'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) - a'_u(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi, \overline{z}_\gamma) + \int_{\Omega_C} g'(\overline{u}_\gamma, \nabla\overline{u}_\gamma)(\varphi)\, d\overline{\mu}_\gamma, \\
\mathcal{M}'_z(\overline{\xi}_\gamma)(\varphi) &= (f, \varphi) - a(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi), \\
\mathcal{M}'_q(\overline{\xi}_\gamma)(\varphi) &= J'_q(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi) - a'_q(\overline{q}_\gamma, \overline{u}_\gamma)(\varphi, \overline{z}_\gamma), \\
\mathcal{M}'_\mu(\overline{\xi}_\gamma)(\varphi) &= \int_{\Omega_C} g(\overline{u}_\gamma, \overline{u}_\gamma)\, d\varphi.
\end{aligned}$$

As the solution $\overline{u}_\gamma$ to (2.4) satisfies the additional regularity $\overline{u}_\gamma \in W$ we obtain that $\mathcal{M}'_u(\overline{\xi}_\gamma)(\overline{u} - \overline{u}_\gamma) = 0$. Similarly, we obtain that $\mathcal{M}'_z(\overline{\xi}_\gamma)(\overline{z} - \overline{z}_\gamma) = 0$ and from (2.10c) that $\mathcal{M}'_q(\overline{\xi}_\gamma)(\overline{u} - \overline{u}_\gamma) = 0$. Hence we conclude that

$$\begin{aligned}
\mathcal{M}'(\overline{\xi}_\gamma)(\overline{\xi} - \overline{\xi}_\gamma) &= \mathcal{M}'_\mu(\overline{\xi}_\gamma)(\overline{\mu} - \overline{\mu}_\gamma) \\
&= \int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\, d\overline{\mu} - \int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\, d\overline{\mu}_\gamma.
\end{aligned}$$

Summing up all terms, we finally get:

$$\begin{aligned}
J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) = &-\frac{1}{2}\int_{\Omega_C} \left(g(\overline{u}, \nabla\overline{u}) + g(\overline{u}_\gamma, \overline{u}_\gamma)\right) d\overline{\mu}_\gamma \\
&+ \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\, d\overline{\mu} \\
&+ \int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\overline{\mu}_\gamma\, dx - B_\gamma(\overline{u}_\gamma) + \mathcal{R}_{\text{hom}} \\
= &\frac{1}{2}\int_{\Omega_C} \left(g(\overline{u}_\gamma, \overline{u}_\gamma) - g(\overline{u}, \nabla\overline{u})\right) d\overline{\mu}_\gamma \\
&- B_\gamma(\overline{u}_\gamma) + \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma, \nabla\overline{u}_\gamma)\, d\overline{\mu} + \mathcal{R}_{\text{hom}}.
\end{aligned}$$

This concludes the proof. □

We now have to obtain a computable estimate from this error identity. We suggest two possible methods to do this.

Complementarity driven estimation Using $-g(u, \nabla u) \geq 0$ together with the definition of $\overline{\mu}_\gamma$ we see that

$$-\int_{\Omega_C} g(\overline{u}, \nabla\overline{u})\, d\overline{\mu}_\gamma \geq 0.$$

Using the fact that $(\overline{q}_\gamma, \overline{u}_\gamma)$ are feasible for (2.10) we obtain

$$0 \geq J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \geq \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma - B_\gamma(\overline{u}_\gamma) + \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu} + \mathcal{R}_{\mathrm{hom}}.$$

Now we assume that $\int_{\Omega_C} g(\overline{u}_\gamma, \overline{u}_\gamma)\, d\overline{\mu} \approx \int_{\Omega_C} g(\overline{u}_\gamma, \overline{u}_\gamma)\, d\overline{\mu}_\gamma$. This is reasonable as $\overline{\mu}_\gamma$ converges weakly* to $\overline{\mu}$. A discussion on this in the case of pointwise state constraints can be found in (Schiela [130]).

We finally suggest to take the best approximation for $\overline{u}_\gamma$ available. Thus the computable estimate reads as:

$$0 \geq J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \gtrsim \eta^{(1)}_{\mathrm{hom}} = (g(\overline{u}^h_\gamma, \nabla \overline{u}^h_\gamma), \overline{\mu}^h_\gamma)_{\Omega_C} - B_\gamma(\overline{u}^h_\gamma). \tag{6.94}$$

As we neglected $-\int_{\Omega_C} g(\overline{u}, \nabla \overline{u})\, d\overline{\mu}_\gamma$ due to its sign we can expect this to overestimate the real error. Further, we remark that this feature requires feasibility of the solution $\overline{u}_\gamma$ which is not given in the case of penalty methods. Therefore we consider an alternative variant.

Convergence driven estimation The other suggested variant uses the idea that $\overline{u}_\gamma \to \overline{u}$ as $\frac{1}{\gamma} \to 0$, thus

$$\int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma) - g(\overline{u}, \nabla \overline{u})\, d\overline{\mu}_\gamma \to 0.$$

Hence we neglect the term $\int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma) - g(\overline{u}, \nabla \overline{u})\, d\overline{\mu}_\gamma$ in the error representation and approximate the multiplier $\overline{\mu}$ with $\overline{\mu}_\gamma$. Then using our discrete approximation $\overline{u}^h_\gamma$ to $\overline{u}_\gamma$ we obtain:

$$J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \approx \eta^{(2)}_{\mathrm{hom}} = \frac{1}{2}(g(\overline{u}^h_\gamma, \nabla \overline{u}^h_\gamma), \overline{\mu}^h_\gamma)_{\Omega_C} - B_\gamma(\overline{u}^h_\gamma). \tag{6.95}$$

Remark 6.14. This might seem unreasonable as $J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \to 0$, e.g., neglecting a term due to its convergence to zero might lead to different convergence rates of $J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma)$ and $\eta^{(2)}_{\mathrm{hom}}$ thus spoiling the estimates we obtain. This can not appear in this case, as we see by comparing our estimates that they vary only in the constant in front of the term $(g(\overline{u}^h_\gamma, \nabla \overline{u}^h_\gamma), \overline{\mu}^h_\gamma)_{\Omega_C}$.

Discretization error In order to estimate the discretization error in the value of the functional J_γ, defined in (5.25) or (5.26), one can use the estimates derived in Section 6.1. We remark that the assumptions on g made in Section 2.3 are important in order to use the same operator g for the definition of feasible states.

For convenience we restate the result using the notation involving the regularization parameter:

Theorem 6.5. *Let $\overline{\xi}_\gamma = (\overline{q}_\gamma, \overline{u}_\gamma, \overline{z}_\gamma) \in Q \times V \times V$ be a solution to the first-order necessary system (5.27) of the barrier problem with strictly feasible $\overline{u}_\gamma$ and let $\overline{\xi}_\gamma^h = (\overline{q}_\gamma^h, \overline{u}_\gamma^h, \overline{z}_\gamma^h) \in Q_h \times V_h \times V_h$ be the solution to its discretization (5.32). Then the following estimate holds:*

$$\begin{aligned} J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h) = \frac{1}{2}\rho_u(\overline{\xi}_\gamma^h)(\overline{z}_\gamma - \tilde{z}^h) + \frac{1}{2}\rho_z(\overline{\xi}_\gamma^h)(\overline{u}_\gamma - \tilde{u}^h) \\ + \frac{1}{2}\rho_q(\overline{\xi}_\gamma^h)(\overline{q}_\gamma - \tilde{q}^h) + \mathcal{R}_{disc}, \end{aligned} \tag{6.96}$$

with arbitrary $(\tilde{q}^h, \tilde{u}^h, \tilde{z}^h) \in Q^h \times V^h \times V^h$ and a remainder term $\mathcal{R}_{disc}$ given by:

$$\mathcal{R}_{disc} = \frac{1}{2}\int_0^1 \mathcal{L}'''(\overline{\xi}_\gamma^h + s(\overline{\xi}_\gamma - \overline{\xi}_\gamma^h))(\overline{\xi}_\gamma - \overline{\xi}_\gamma^h, \overline{\xi}_\gamma - \overline{\xi}_\gamma^h, \overline{\xi}_\gamma - \overline{\xi}_\gamma^h)s(s-1)\,ds. \tag{6.97}$$

By neglecting the remainder we obtain the computable estimate

$$\begin{aligned} J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h) &\approx \rho_u(\overline{\xi}_\gamma^h)(\pi\overline{z}_\gamma^h - \overline{z}_\gamma^h) + \rho_z(\overline{\xi}_\gamma^h)(\pi\overline{u}_\gamma^h - \overline{u}_\gamma^h) \\ &\quad + \rho_q(\overline{\xi}_\gamma^h)(\pi^q\overline{q}_\gamma^h - \overline{q}_\gamma^h) \\ &= \eta_{\text{disc}}. \end{aligned}$$

An adaptive algorithm The remaining question is how to steer the values of γ and h. As we are interested in computing the value of the cost functional, it is not sensible to have γ or h too small in comparison to the other, especially as this makes the underlying problems harder to solve. Instead, we try to choose both parameters in such a way, that the errors introduced by the parameters are equilibrated. To do this, we choose both parameters such that the error estimators are of approximately the same size, e.g.,:

$$|\eta_{\text{hom}}| \approx |\eta_{\text{disc}}|.$$

Thus we arrive at the following algorithm:

Algorithm 6.1 A Simple Adaptive Algorithm

```
Initialize TOL, h, γ, c
repeat
    Solve problem (5.30)
    if |η_hom| > c|η_disc| then
        Increase γ (reduce 1/γ)
    else
        Refine mesh according to η_disc
    end if
until |η_hom| + |η_disc| < TOL
```

Remark 6.15. In some problems it might occur, that $\eta_{\text{disc}}\eta_{\text{hom}} < 0$, e.g., the errors introduced by both parameters have different sign. If in these cases $|\eta_{\text{hom}}| \approx |\eta_{\text{disc}}|$ we may expect that

$|\eta_{\text{hom}} + \eta_{\text{disc}}| \ll |\eta_{\text{hom}}| + |\eta_{\text{disc}}|$. Thus we would loose reliability of the estimate if we would simply take $\eta = \eta_{\text{hom}} + \eta_{\text{disc}}$. Therefore we suggested to use

$$|J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)| \lesssim |\eta_{\text{hom}}| + |\eta_{\text{disc}}|$$

to obtain a more reliable estimate.

We will show later, in our numerical examples, that the changing sign doesn't occur only in the error estimation but can also be observed in the error $|J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|$, e.g., it is not caused by our approximate quantities η, but rather the estimates are good enough to capture this behavior. Because of this we can not expect any estimation, near the zero of $|J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|$ or $\eta_{\text{hom}} + \eta_{\text{disc}}$, respectively, to be both reliable and efficient.

6.2.1.2 Illustration of the Results for Two Specific Types of Constraints

In this section we are concerned to give two examples of constraints that can be treated in the framework we prepared in Section 6.2.1.1. For this we consider pointwise constraints on the state and pointwise constraints on the gradient of the state. For these choices the assumptions in Section 2.2 are fulfilled, see Section 2.4. Here we write down the a posteriori estimates from Section 6.2.1.1. Note that we won't discuss the case of finitely many state constraints in detail. However for these cases one may find the first-order necessary conditions for instance in (Casas and Bonnans [35], Casas and Tröltzsch [39]).

State constraints Here we choose pointwise bounds on the state hence we consider an example similar to the example of Section 2.4. Our optimization problem (2.5) takes the form

$$\text{Minimize } J(q,u) := \frac{1}{2}\|u - u^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|q\|_{L^2(\Omega)}^2, \tag{6.98a}$$

$$\text{such that } \begin{cases} (\nabla u, \nabla \varphi) = (q, \varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ u - \psi \leq 0 & \text{on } \Omega_C, \end{cases} \tag{6.98b}$$

where Ω_C is a compact subset of $\overline{\Omega}$ and $\psi \in \mathbb{R}$. The mapping g is defined by

$$g(u, \nabla u) = u - \psi.$$

Remark 6.16. The restriction to constant bounds for the state is crucial in the following sense. For our computation we have to replace ψ by a finite dimensional approximation ψ_h, either due to interpolation as a finite element function $\psi_h \in V^h$ or by numerical integration. Our discrete problem would then use the mapping $g_h(u, \nabla u) = u - \psi_h$. This is covered by our analysis only if $\psi_h = \psi$.

If we restrict the dimension of the domain $\Omega \subset \mathbb{R}^n$ to $n = 2, 3$ and assume that the domain is either convex polygonal or has a smooth boundary, then we have for $W = H^2(\Omega) \cap H_0^1(\Omega)$ that the mapping $g\colon W \to C(\overline{\Omega})$ is well-defined and continuous by a well known embedding theorem. In addition, we see that g is also continuous if interpreted as mapping $g\colon V_h \to L^\infty(\Omega_C)$. Thus our assumptions on g are verified. For the rest of the well-posedness we refer to Section 2.4.

For the convenience of the reader, we write down the a posteriori estimates derived in Section 6.2.1.1 for this setting:

$$J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) = \frac{\gamma^{-\kappa}}{2}\int_{\Omega_C}\frac{(\overline{u}_\gamma-\overline{u})}{(\psi-\overline{u}_\gamma)^\kappa}\,dx + \frac{1}{2}\int_{\Omega_C}(\psi-\overline{u}_\gamma)\,d\overline{\mu} - B_\gamma(\overline{u}_\gamma) + \mathcal{R}_{\mathrm{hom}}.$$

Here $\kappa \geq 1$ is the order of the barrier function. The residual ρ_z from (6.96) of the adjoint equation takes the form

$$\rho_z(\overline{\xi}_\gamma^h)(\,\cdot\,) = (\overline{u}_\gamma^h - u^d,\,\cdot\,) - \gamma^{-\kappa}(\psi-\overline{u}_\gamma^h)^{-\kappa},\,\cdot\,)_{\Omega_C} - (\nabla\cdot,\nabla\overline{z}_\gamma^h).$$

Gradient constraints Here we choose pointwise bounds on the gradient of the state and the optimization problem takes the form

$$\text{Minimize } J(q,u) := \frac{1}{2}\|u-u^d\|^2_{L^2(\Omega)} + \frac{\alpha}{r}\|q\|^r_{L^r(\Omega)}, \tag{6.99a}$$

$$\text{such that } \begin{cases} (\nabla u,\nabla\varphi) = (q,\varphi) & \forall\,\varphi\in H^1_0(\Omega), \\ |\nabla u|^2 - \psi \leq 0 & \text{on } \Omega_C, \end{cases} \tag{6.99b}$$

where once again Ω_C is a compact subset of $\overline{\Omega}$, $r > n$ and $\psi > 0$ a constant. The mapping g is defined as in Section 2.4 by

$$g(u,\nabla u) = |\nabla u|^2 - \psi.$$

Here we need $W \subset C^1(\Omega_C)$ to obtain that g is differentiable as mapping $g\colon W \to C(\Omega_C)$. Therefore we have to consider $W = W^{2,t}(\Omega)\cap W^{1,t}_0(\Omega)$ with $t > n$ ($\Omega\subset\mathbb{R}^n$). In addition we see that the interpretation of $g\colon V_h \to L^\infty(\Omega_C)$ makes sense.

However to obtain that the control to state mapping maps $L^t(\Omega)$ into W we have to require that either $n = 2$ and Ω is convex polygonal or that the boundary of Ω is sufficiently smooth, see Section 4.1 for details.

For the convenience of the reader, we write down the a posteriori estimates derived in Section 6.2.1.1 for this setting:

$$J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) = \frac{\gamma^{-\kappa}}{2}\int_{\Omega_C}\frac{(|\nabla\overline{u}_\gamma|^2-|\nabla\overline{u}|^2)}{(\psi-|\nabla\overline{u}_\gamma|^2)^\kappa)}\,dx + \frac{1}{2}\int_{\Omega_C}(\psi-|\nabla\overline{u}_\gamma|^2)\,d\overline{\mu} - B_\gamma(\overline{u}_\gamma) + \mathcal{R}_{\mathrm{hom}}.$$

Here $\kappa \geq 1$ is the order of the barrier function. The residual ρ_z from (6.96) of the adjoint equation takes the form

$$\rho_z(\overline{\xi}_\gamma^h)(\,\cdot\,) = (\overline{u}_\gamma^h - u^d,\,\cdot\,) + 2\gamma^{-\kappa}((\psi-|\nabla\overline{u}_\gamma^h|^2)^{-\kappa}\nabla\overline{u}_\gamma^h,\nabla\,\cdot\,)_{\Omega_C} - (\nabla\cdot,\nabla\overline{z}_\gamma^h).$$

6.2.1.3 Numerical Results

In this section we demonstrate our findings for two example configurations taken from other publications. All computations where made using the software packages RoDoBo (RoDoBo [126]) and Gascoigne (Gascoigne [65]). The Visualizations where obtained using VisuSimple (Visusimple [150]). In both examples bilinear finite elements were used for the discretization of the space for the state and control variable.

State Constraints Here we consider an example taken from (Günther and Hinze [75]). There the following problem was considered:

$$\text{Minimize } J(q,u) := \frac{1}{2}\|u-0.5\|^2_{L^2(\Omega)} + \frac{1}{2}\|q-60\|^2_{L^2(\Omega)},$$
$$\text{such that } \begin{cases} (\nabla u, \nabla\varphi) + (u,\varphi) = (q,\varphi) & \forall\, \varphi \in H^1(\Omega), \\ (q,u) \in L^2(\Omega) \times H^1(\Omega), & \\ 0.45 \le u(x) \le \psi(x) & \forall\, x \in \overline{\Omega}, \end{cases}$$

on the domain $\Omega = (0,1)^2 \subset \mathbb{R}^2$ with the upper bound

$$\psi(x) = \min\bigl(1, \max\bigl(0.5, 50((x_1-0.3)^2 + (x_2-0.3)^2)\bigr)\bigr).$$

An approximation $J^* \approx 1759.04733$ for the optimal value of the cost functional is given in (Günther and Hinze [75]) where it was obtained on a equidistant mesh with 557^2 nodes.

In Figure 6.5 we show a computed approximation of the state, the control, and the approximated multiplier on the mesh in Figure 6.5d.

In our computations we have chosen $\kappa = 4$ as order of our barrier function.

Remark 6.17. It should be noted that this example doesn't fit into our framework because the upper bound for u is not constant, see the discussion in Section 6.2.1.2. Here we neglect the error introduced by the discretization of the upper bound and see that we still get satisfactory results for this example.

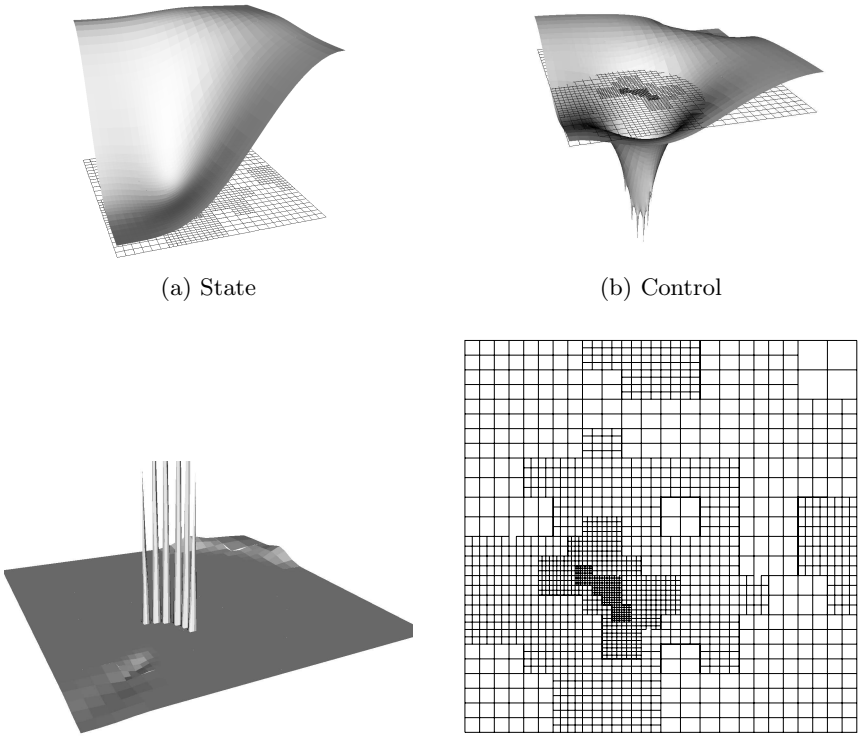
Table 6.3: Comparison of effectivity indices for the homotopy error

(a) Global refinement with $\eta^{(1)}_{\text{hom}}$

	N		
$1/\gamma$	625	2401	9409
$3 \cdot 10^{-1}$	0.36	0.36	0.36
$1 \cdot 10^{-1}$	0.65	0.62	0.65
$3 \cdot 10^{-2}$	0.11	0.21	0.64
$1 \cdot 10^{-2}$	1.89	1.82	0.29
$3 \cdot 10^{-3}$	5.82	7.12	3.29
$1 \cdot 10^{-3}$	10.5	16.2	10.0

(b) Global refinement with $\eta^{(2)}_{\text{hom}}$

	N		
$1/\gamma$	625	2401	9409
$3 \cdot 10^{-1}$	0.48	0.47	0.48
$1 \cdot 10^{-1}$	0.87	0.83	0.87
$3 \cdot 10^{-2}$	0.15	0.25	0.85
$1 \cdot 10^{-2}$	2.36	2.35	0.38
$3 \cdot 10^{-3}$	6.74	8.74	4.19
$1 \cdot 10^{-3}$	11.1	18.3	12.0



(a) State (b) Control

(c) Multiplier approximation (d) Mesh

Figure 6.5: Computed solutions and corresponding mesh

We start the discussion of this example with a comparison of the effectivity of the two proposed variants to estimate the error introduced by the barrier parameter γ. The effectivity indices

$$I_{\text{eff}} = \frac{|J^* - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|}{|\eta_{\text{disc}}| + |\eta_{\text{hom}}|}$$

for the choice $\eta_{\text{hom}} = \eta_{\text{hom}}^{(1)}$ can be found in Table 6.3a whereas those for $\eta_{\text{hom}} = \eta_{\text{hom}}^{(2)}$ can be found in Table 6.3b where they are depicted for a sequence of globally refined meshes.

Here we can see the expected behavior of the indices, especially for $\eta_{\text{hom}}^{(1)}$ we do not get effectivities near one. Note that for the smallest values of $1/\gamma$ both indicators give almost the same value, which is due to the fact, that the error in the cost functional is then dominated by the discretization error. For dominant discretization error we see, that the effectivity indices become rather large. This can be explained by the fact, that we used non-constant bounds. Thus we had to use an interpolation of the bounds which leads to the problem, that the discrete solutions are no longer feasible with respect to the continuous bounds. To substantiate this we introduce the following transformation:

$$v = \frac{\psi - u}{\psi - 0.45}.$$

Then v solves the problem:

$$\text{Minimize } J(q,u) := \frac{1}{2}\|(0.45-\psi)v+\psi-0.5\|^2_{L^2(\Omega)} + \frac{1}{2}\|q-60\|^2_{L^2(\Omega)}$$
$$\text{such that } \begin{cases} -\Delta((0.45-\psi)v+\psi) + ((0.45-\psi)v+\psi) = q \ \text{ in } \Omega, \\ \partial_n v = 0 \ \text{ on } \partial\Omega, \\ 0 \le v(x) \le 1 \ \ \forall\, x \in \overline{\Omega}. \end{cases}$$

The results for this computation are depicted in Table 6.4. Here we can see that the effectivities

Table 6.4: Comparison of effectivity indices for the homotopy error

(a) Transformed example with $\eta^{(1)}_{\text{hom}}$

	N		
$1/\gamma$	625	2401	9409
$3\cdot10^{-1}$	0.29	0.52	0.71
$1\cdot10^{-1}$	0.00	0.26	0.57
$3\cdot10^{-2}$	1.24	1.47	0.10
$1\cdot10^{-2}$	1.50	2.23	0.96
$3\cdot10^{-3}$	1.58	2.55	1.50
$1\cdot10^{-3}$	1.60	2.65	1.70

(b) Transformed example with $\eta^{(2)}_{\text{hom}}$

	N		
$1/\gamma$	625	2401	9409
$3\cdot10^{-1}$	0.36	0.67	0.94
$1\cdot10^{-1}$	0.00	0.31	0.72
$3\cdot10^{-2}$	1.29	1.60	0.12
$1\cdot10^{-2}$	1.52	2.30	1.03
$3\cdot10^{-3}$	1.58	2.58	1.54
$1\cdot10^{-3}$	1.60	2.66	1.71

are far better. However, even in the case of non-constant bounds we obtain remarkably good results if local refinement is used. This can be seen in the following Table 6.5. The estimation of the error introduced by the homotopy methods is better if $\eta^{(2)}_{\text{hom}}$ is used. Hence we employed this estimator in the following results.

Reasons for the very small effectivity indices for the choices $1/\gamma = 1\cdot10^{-1}$ and $1/\gamma = 3\cdot10^{-2}$ will be discussed in the next example.

Table 6.5: Effectivity indices for locally refined meshes

(a) Local refinement balanced with $\eta^{(2)}_{\text{hom}}$

N	I_{eff}	$1/\gamma$	$\lvert J^* - J_\gamma(\overline{q}^h_\gamma, \overline{u}^h_\gamma)\rvert$
169	0.48	$2\cdot10^{-2}$	$1.5\cdot10^{-1}$
281	3.83	$4\cdot10^{-3}$	$1.9\cdot10^{-1}$
401	1.27	$8\cdot10^{-3}$	$1.1\cdot10^{-1}$
1057	1.56	$4\cdot10^{-3}$	$4.7\cdot10^{-2}$
1981	0.65	$3\cdot10^{-3}$	$1.6\cdot10^{-2}$

(b) Local refinement for $1/\gamma \to 0$

N	I_{eff}	$\lvert J^* - J_\gamma(\overline{q}^h_\gamma, \overline{u}^h_\gamma)\rvert$
169	1.98	$3.4\cdot10^{-1}$
269	7.80	$2.2\cdot10^{-1}$
401	3.45	$1.7\cdot10^{-1}$
1045	3.90	$6.7\cdot10^{-2}$
1749	2.09	$3.2\cdot10^{-2}$

In Table 6.5a we see the behavior of the effectivity index for a sequence of locally refined meshes. Here γ was chosen in such a way that the discretization error is of the same size as the error introduced by the barrier parameter γ. Furthermore, we see that the influence of the

bad estimation for the discretization error, as seen in Table 6.3a and Table 6.3b, doesn't have great influence on the estimation for the values of γ obtained from our balancing strategy. In addition, the value of γ to have equilibrated error contributions is given as well as the total error obtained from discretization and barrier method is shown. As a comparison we show in Table 6.5b the values obtained for local refinement where $1/\gamma$ was driven towards zero as an estimate for the error obtained without a barrier method. For reasons unknown to the author these values do not correspond to those shown in (Günther and Hinze [75]), in fact, we can reach comparable errors with only about a quarter of the unknowns required in (Günther and Hinze [75]). This is even though the computed values for uniformly refined meshes are in good correspondence.

We now compare the development of the estimated errors in the cost functional in Figure 6.6. Here we compared global and local refinement for γ chosen to equilibrate the error contribu-

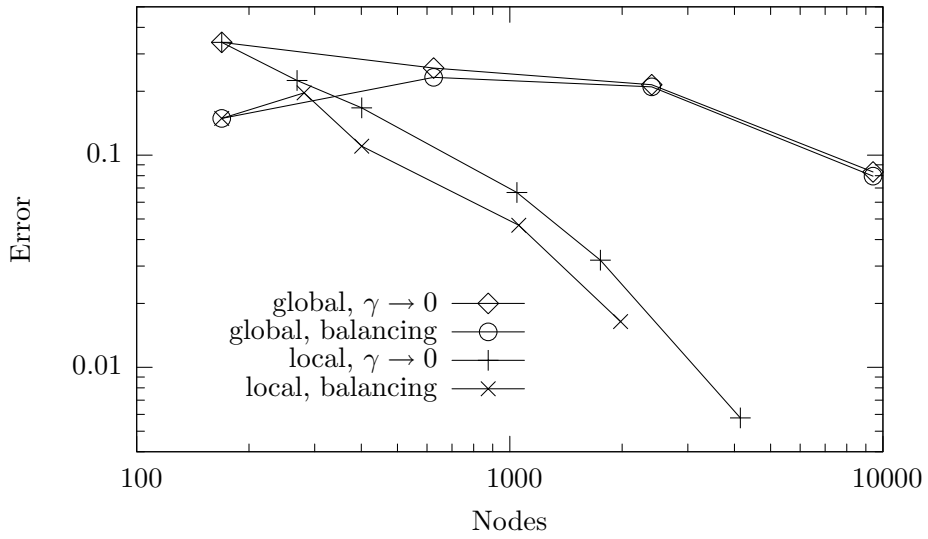


Figure 6.6: Error in J for different refinement strategies

tions and for the choice $1/\gamma \to 0$. The choice $1/\gamma \to 0$ is made to simulate the computation without 'regularization'. In that case we stopped the computation once the value of J_γ was unchanged in the first four digits by changing γ.

Especially we can see that there is no real difference in the error between the 'non-regularized' discretization simulated by $1/\gamma \to 0$ and the error obtained for γ chosen to equilibrate the error contributions, except for the fact that the solutions in the latter case are easier to compute.

Gradient Constraints Here we consider the example given in (Deckelnick et al. [54]). The problem reads as follows:

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2_{L^2(\Omega)} + \frac{1}{2}\|q\|^2_{L^2(\Omega)}$$
$$\text{such that } \begin{cases} (\nabla u, \nabla\varphi) = (f+q,\varphi) & \forall\,\varphi \in H^1_0(\Omega), \\ -2 \le q \le 2 & \text{a.e. in } \Omega, \\ \frac{1}{4} - |\nabla u(x)|^2 \ge 0 & \forall\, x \in \overline{\Omega}, \end{cases}$$

on the domain $\Omega = \{\, x \in \mathbb{R}^2 \mid |x| < 2 \,\}$. The desired state is given as

$$u^d(x) = \begin{cases} \frac{1}{4} + \frac{1}{2}\ln 2 - \frac{1}{4}|x|^2, & |x| \le 1, \\ \frac{1}{2}\ln 2 - \frac{1}{2}\ln|x|, & \text{otherwise}, \end{cases}$$

and the right-hand side

$$f(x) = \begin{cases} 2, & |x| \le 1, \\ 0, & \text{otherwise}, \end{cases}$$

this problem admits the unique solution

$$\overline{u} = u^d \quad \text{and} \quad \overline{q} = \begin{cases} -1, & |x| \le 1, \\ 0, & \text{otherwise}. \end{cases}$$

The optimal value of the cost functional is $\frac{\pi}{2}$ and in addition the control constraints are inactive at the solution $(\overline{u},\overline{q})$. In order to show the behavior of the cost functional on locally refined meshes in more detail Figure 6.8 has been recomputed in order to show the results on finer meshes. Unfortunately, the quadrature rules used for the different parts in the original computation where removed in the meantime, and hence the total error on a fixed mesh is slightly different to (Wollner [157]). The recomputation used tensor product two-point gauss formulas, using a sum of barrier functions of orders $\kappa = 2,\dots,6$. The original material was computed with a barrier function of order $\kappa = 6$.

We note that the introduction of the admissible set for the controls is necessary to ensure existence of a solution following standard arguments. They are inactive at the optimal solution hence our estimates are applicable.

In Figure 6.7 we can see that the error in the value of the cost functional has indeed a sign change. This verifies Remark 6.15. Especially we must expect that the effectivity index

$$I_{\text{eff}} = \frac{|0.5\pi - J\gamma(\overline{q}^h_\gamma, \overline{u}^h_\gamma)|}{|\eta_{\text{disc}}| + |\eta_{\text{hom}}|}$$

can go to zero for some values of γ.

In Table 6.6a the effectivity indices for $\eta_{\text{hom}} = \eta^{(1)}_{\text{hom}}$ and in Table 6.6b the effectivity indices for $\eta_{\text{hom}} = \eta^{(2)}_{\text{hom}}$ are shown. First of all we can see that for some value of γ the effectivity index is rather small, e.g., less than 0.1, which is in accordance with our observation, that $0.5\pi - J_\gamma(\overline{q}^h_\gamma, \overline{u}^h_\gamma)$ is zero for an appropriate choice of γ. Furthermore, we can see, that for

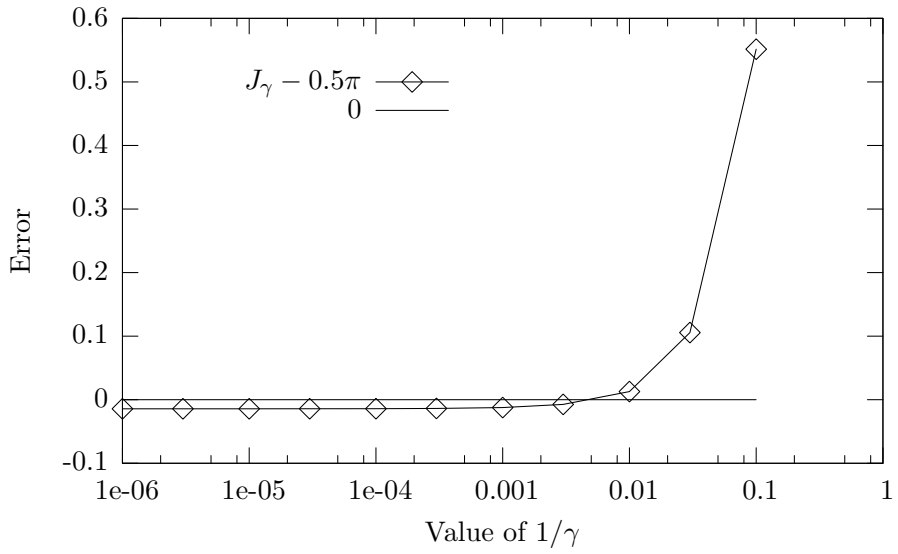


Figure 6.7: Error in J for different γ

Table 6.6: Comparison of effectivity indices for the homotopy error

(a) Global refinement with $\eta_{\text{hom}}^{(1)}$

	N		
$1/\gamma$	801	3137	12417
$1 \cdot 10^{-1}$	0.74	0.80	0.83
$3 \cdot 10^{-2}$	0.43	0.62	0.81
$1 \cdot 10^{-2}$	0.07	0.09	0.31
$3 \cdot 10^{-3}$	0.25	0.79	0.29
$1 \cdot 10^{-3}$	0.31	1.10	0.55

(b) Global refinement with $\eta_{\text{hom}}^{(2)}$

	N		
$1/\gamma$	801	3137	12417
$1 \cdot 10^{-1}$	0.97	1.07	1.12
$3 \cdot 10^{-2}$	0.52	0.79	1.07
$1 \cdot 10^{-2}$	0.08	0.11	0.37
$3 \cdot 10^{-3}$	0.26	0.83	0.32
$1 \cdot 10^{-3}$	0.31	1.12	0.56

large values of $1/\gamma$ the estimation is of less good quality. This is due to the fact, that $\overline{u}_\gamma$ is still a bad approximation to $\overline{u}$. These values do not change with grid refinement, hence we assume that this effect is not caused by the discretization. In addition, we note that in contrast to the previous example the effectivity indices are of moderate size for dominant discretization error.

In Table 6.7a the effectivity indices are shown for a locally refined mesh where γ was chosen in order to balance the error contributions. Additionally, for each mesh size the value of $1/\gamma$ obtained in the iteration, rounded to one decimal, is shown. Finally, in Table 6.7b the effectivity indices are shown for a locally refined mesh where $1/\gamma \to 0$ was taken to simulate the results one would obtain if the optimal control problem was discretized without further regularization.

A comparison of the development of the error in the cost functional is depicted in Figure 6.8. Here we compare global and local refinement as well as the choice of γ to balance the error

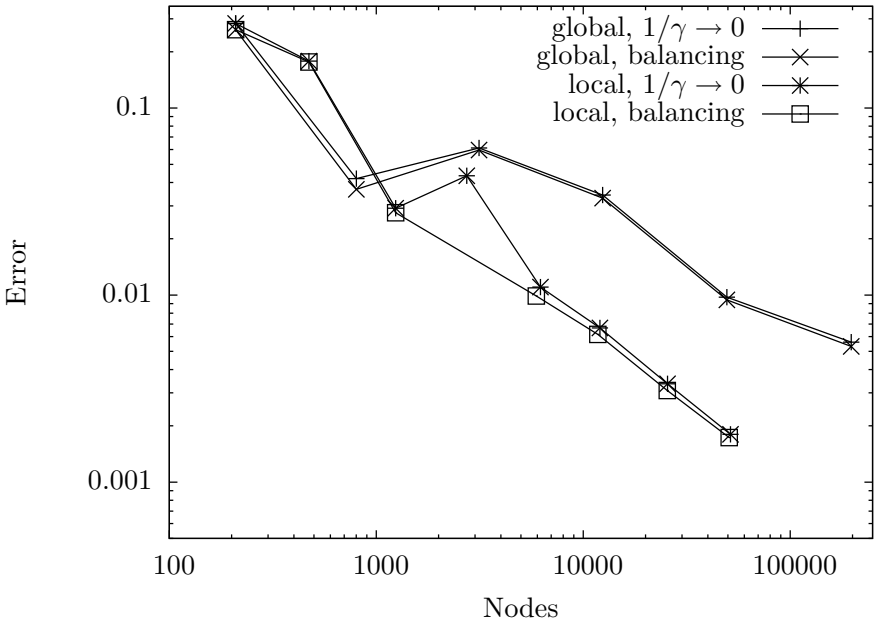


Figure 6.8: Error in J for different refinement strategies

contributions with $1/\gamma \to 0$. In order to avoid problems due to cancellation we considered both error contributions balanced as soon as $|\eta_{\text{disc}}| \approx 10|\eta_{\text{hom}}|$. This eliminates the oscillating behavior of the error seen in the original source (Wollner [157]). We can see now that on globally refined meshes both error obtained by balancing the error as well as from considering the limit case coincide well on sufficiently fine meshes. In the case of local refinement, the same behavior can be seen. The local refinement indicators lead to almost identical meshes in both cases. In addition we note that in order to reach an error of 0.01 locally refined meshes can do so using roughly one tenth of the degrees of freedom required by global refinement.

The computed state $\overline{u}_\gamma^h$, the control $\overline{q}_\gamma^h$, and the approximation to the multiplier for the state

Table 6.7: Effectivity indices for locally refined meshes

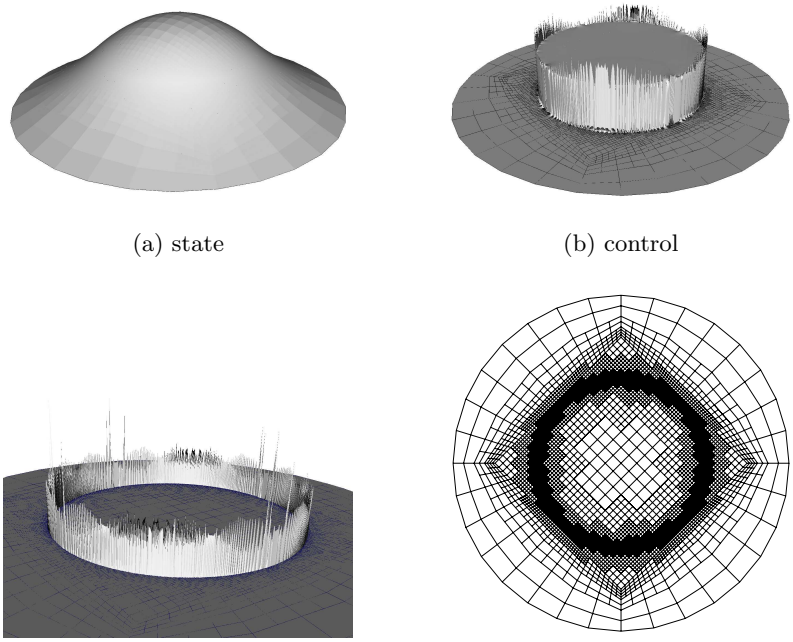
(a) Local refinement balanced with $\eta_{\text{hom}}^{(2)}$

N	I_{eff}	$1/\gamma$
169	0.33	$4 \cdot 10^{-2}$
427	0.41	$2 \cdot 10^{-2}$
1153	0.11	$8 \cdot 10^{-3}$
2709	0.14	$5 \cdot 10^{-3}$
6161	0.13	$3 \cdot 10^{-3}$
12885	0.33	$3 \cdot 10^{-3}$

(b) Local refinement for $1/\gamma \to 0$

N	I_{eff}
200	0.30
493	0.30
1201	1.37
2809	0.76
6121	0.95
12745	1.02

constraints obtained by local refinement with γ chosen to balance the error contributions are depicted in Figure 6.9 together with the mesh on which they were obtained.



(a) state (b) control

(c) multiplier approximation (d) mesh

Figure 6.9: Computed solutions and corresponding mesh

6.2.1.4 The Influence of the Approximations to the Estimate

We will now consider the influence of the approximations, made in the preceeding Section 6.2.1.1, in greater detail. We begin with a discussion of the approximations in η_{hom}, then turn our attention to to the discretization error estimate η_{disc}.

Approximations in the Regularization Error We remark that the approximation η_{hom} is not independent from the discretization due to the use of the discrete variables $\overline{u}_\gamma^h$ and $\overline{\mu}_\gamma^h$. Hence we will first take a look on the behavior of our estimator under mesh refinement. As both estimators only differ by a constant we consider $\eta_{\text{hom}} = \eta_{\text{hom}}^{(2)}$ throughout this section.

To do so we reconsider the example with gradient state constraints from Section 6.2.1.3. We will now consider the size of the regularization estimator on a sequence of globally refined meshes. The computations are done using a barrier function of order $\kappa = 6$. The discretization is done using Q_1 finite elements for both the state and control variable. The integrals are evaluated using tensor product two-point Gauss formulas and a four-point Gauss-Lobatto formula for the barrier function. In order to have a fine grained impression of the behavior

Table 6.8: Homotopy error estimate on various meshes. Values below the discretization error are on gray background.

	N				
γ	209	801	3137	12417	49409
$2.0 \cdot 10^{0}$	$1.4 \cdot 10^{+2}$	$1.4 \cdot 10^{+2}$	$1.4 \cdot 10^{+2}$	$1.4 \cdot 10^{+2}$	$1.4 \cdot 10^{+2}$
$2.8 \cdot 10^{0}$	$2.0 \cdot 10^{+1}$	$2.0 \cdot 10^{+1}$	$2.0 \cdot 10^{+1}$	$2.0 \cdot 10^{+1}$	$2.0 \cdot 10^{+1}$
$4.0 \cdot 10^{0}$	$4.3 \cdot 10^{+0}$	$4.4 \cdot 10^{+0}$	$4.4 \cdot 10^{+0}$	$4.4 \cdot 10^{+0}$	$4.4 \cdot 10^{+0}$
$5.7 \cdot 10^{0}$	$1.6 \cdot 10^{+0}$	$1.6 \cdot 10^{+0}$	$1.6 \cdot 10^{+0}$	$1.6 \cdot 10^{+0}$	$1.6 \cdot 10^{+0}$
$8.0 \cdot 10^{0}$	$7.4 \cdot 10^{-1}$	$7.6 \cdot 10^{-1}$	$7.5 \cdot 10^{-1}$	$7.4 \cdot 10^{-1}$	$7.4 \cdot 10^{-1}$
$1.1 \cdot 10^{1}$	$3.8 \cdot 10^{-1}$	$4.0 \cdot 10^{-1}$	$4.0 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$
$1.6 \cdot 10^{1}$	$2.1 \cdot 10^{-1}$	$2.3 \cdot 10^{-1}$	$2.3 \cdot 10^{-1}$	$2.3 \cdot 10^{-1}$	$2.3 \cdot 10^{-1}$
$2.3 \cdot 10^{1}$	$1.1 \cdot 10^{-1}$	$1.3 \cdot 10^{-1}$	$1.4 \cdot 10^{-1}$	$1.4 \cdot 10^{-1}$	$1.4 \cdot 10^{-1}$
$3.2 \cdot 10^{1}$	$6.4 \cdot 10^{-2}$	$8.1 \cdot 10^{-2}$	$8.7 \cdot 10^{-2}$	$8.9 \cdot 10^{-2}$	$8.8 \cdot 10^{-2}$
$4.5 \cdot 10^{1}$	$3.9 \cdot 10^{-2}$	$4.9 \cdot 10^{-2}$	$5.3 \cdot 10^{-2}$	$5.6 \cdot 10^{-2}$	$5.6 \cdot 10^{-2}$
$6.4 \cdot 10^{1}$	$2.4 \cdot 10^{-2}$	$3.0 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$	$3.7 \cdot 10^{-2}$
$9.0 \cdot 10^{1}$	$1.6 \cdot 10^{-2}$	$1.9 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$
$1.3 \cdot 10^{2}$	$1.1 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$
$1.8 \cdot 10^{2}$	$7.8 \cdot 10^{-3}$	$8.0 \cdot 10^{-3}$	$8.4 \cdot 10^{-3}$	$8.8 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$
$2.6 \cdot 10^{2}$	$5.4 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$	$5.6 \cdot 10^{-3}$	$5.8 \cdot 10^{-3}$	$5.9 \cdot 10^{-3}$
$3.6 \cdot 10^{2}$	$3.8 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$

we computed the values on each mesh beginning with the parameter $\gamma = 2$ and increasing this value in each successive iteration by a factor of $\sqrt{2}$.

The results for the estimator (6.95) are depicted in Table 6.8. We continued the computation on each mesh until $\gamma = 2^{15} \approx 3 \cdot 10^4$ at this time the first two digits of the error in the cost functional remain fixed with respect to $\gamma \to \infty$. This value was taken as reference value for the discretization error in the cost functional. In Table 6.8 all values of η_{hom} below this value are printed on gray background.

It can be seen that as long as the error estimator η_{hom} is not too small compared to the discretization error the estimates remain unchanged under mesh refinement. Hence the choice of the quantity (6.95) as error indicator for the regularization error is reasonable, in the sense that it remains constant as long as the discretization error is smaller than this quantity.

The next question to be addressed is whether the quantity (6.95) actually is a good approximation on the continuous level, e.g., whether the assumptions on the convergence of certain quantities are met in the parameter range we are considering and η_{hom} is in fact a good approximation of $J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma)$.

In Table 6.9 we consider the fraction $I_{\text{eff}} := \frac{|J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|}{|\eta_{\text{hom}}|}$. We can see that for small values of γ the estimators are too large which can be explained by the fact that the approximation $\overline{u} \approx \overline{u}_\gamma$ is not good enough yet. However since we only get an overestimation this doesn't conflict with our aim to balance the error contributions such that the discretization error is at least as large as the regularization error.

Table 6.9: Efficiency of η_{hom} on various meshes. Values below the discretization error are on gray background.

	N				
γ	209	801	3137	12417	49409
$2.0 \cdot 10^0$	0.3	0.3	0.3	0.3	0.3
$2.8 \cdot 10^0$	0.4	0.4	0.4	0.4	0.4
$4.0 \cdot 10^0$	0.6	0.7	0.7	0.7	0.7
$5.7 \cdot 10^0$	0.8	0.9	0.9	0.9	0.9
$8.0 \cdot 10^0$	0.8	1.1	1.0	1.0	1.1
$1.1 \cdot 10^1$	0.7	1.2	1.1	1.1	1.2
$1.6 \cdot 10^1$	0.4	1.2	1.1	1.2	1.3
$2.3 \cdot 10^1$	0.1	1.3	1.1	1.2	1.3
$3.2 \cdot 10^1$	1.0	1.4	0.9	1.1	1.3
$4.5 \cdot 10^1$	2.4	1.6	0.8	1.0	1.3
$6.4 \cdot 10^1$	4.5	1.8	0.5	0.8	1.3
$9.0 \cdot 10^1$	7.4	2.1	0.1	0.5	1.2
$1.3 \cdot 10^2$	11.6	2.6	0.6	0.0	1.2
$1.8 \cdot 10^2$	17.4	3.2	1.7	0.7	1.1
$2.6 \cdot 10^2$	25.6	4.0	3.3	1.9	1.0
$3.6 \cdot 10^2$	37.1	5.1	5.7	3.5	0.9

Then we see that for $\gamma > 5$ the estimate is very good with effectivity around one. The loss of effectivity once the level of discretization error is reached can be explained by the fact, that we took $J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)$ instead of $J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma)$ for comparison. This indicates that the use of the approximated quantity η_{hom} as an indicator for the regularization error is reasonable.

In addition, as the values below the discretization error are at least approximately of constant size, see Table 6.8, we can also expect these values to give a good prediction on the size of the regularization error for these values.

Now we turn our attention towards the estimate for the discretization error η_{disc} given by Theorem 6.5.

Approximations in the Discretization Error In order to study the discretization error estimator η_{disc} we consider the quantity $I_{\text{eff}} = \frac{|J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|}{|\eta_{\text{disc}}|}$ for various (fixed) values of γ under mesh refinement. The results are shown in Table 6.10. We mention that although we are interested in the quantity $|J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|$ it is reasonable to consider the quantity $|J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)|$ as long as the regularization error is small enough compared to the influence of the discretization. We can see that the efficiency index is not as stable as the one for the regularization error. In particular we notice, that the efficiency is oscillating between over and underestimation of the discretization error.

Table 6.10: Efficiency of η_{disc}. Values above the regularization error are on gray background.

	N					
γ	209	801	3137	12417	49409	197121
$5.1 \cdot 10^2$	1.3	0.3	2.1	1.2	0.4	1.1
$4.1 \cdot 10^3$	1.4	0.3	3.0	1.7	0.5	1.0
$2.3 \cdot 10^4$	1.4	0.3	3.1	1.8	0.7	1.3

However, we note that we are in fact neglecting the influence of the quadrature error obtained by the use of a Gauss-Lobatto formula for the integration of the discontinuous right-hand side as well as the barrier functional. To substantiate this, we will reconsider the above example using a summed midpoint rule, where each mesh element is split into 2^8 subelements for the integration. The results are depicted in Table 6.11. We can see that the efficiency index is

Table 6.11: Efficiency of η_{disc} for a summed quadrature rule. Values above the regularization error are on gray background.

	N			
γ	209	801	3137	12417
$5.1 \cdot 10^2$	0.5	0.5	1.5	0.7
$4.1 \cdot 10^3$	0.5	0.5	1.8	0.9
$2.3 \cdot 10^4$	0.5	0.5	1.9	1.0

better compared to Table 6.10. In order to understand what terms are relevant for the large influence of the integration error we will consider an other example in Section 6.2.1.6 where only the integration error for the barrier functional will be important.

6.2.1.5 Barrier Regularization with Control Constraints

Now we consider the general case with $Q^{\text{ad}} \neq Q$. Then in contrast to the proof of Theorem 6.4 we can no longer assume that $\mathcal{M}_q'(\overline{q}, \overline{u}, \overline{z}, \overline{\mu}) = 0$ with $\mathcal{M}$ given by (6.89). But instead by virtue of the necessary conditions (2.10), especially (2.10c) we have that

$$\mathcal{M}_q'(\overline{q}, \overline{u}, \overline{z}, \overline{\mu})(\delta q - \overline{q}) \geq 0 \quad \forall\, \delta q \in Q^{\text{ad}}.$$

The analogous result holds for the solution to (5.27). Hence we introduce the control residual

$$\rho_q(\xi)(\delta q) := \mathcal{M}_q'(\xi)(\delta q) = J_q'(q,u)(\delta q) - a_q'(q,u)(\delta q, z).$$

Then we obtain the following:

Theorem 6.6. *Let $\overline{\xi} = (\overline{q}, \overline{u}, \overline{z}, \overline{\mu})$ be a solution to the first-order necessary system* (2.10) *and let $\overline{\xi}_\gamma = (\overline{q}_\gamma, \overline{u}_\gamma, \overline{z}_\gamma, \overline{\mu}_\gamma)$ a solution to the first-order necessary system* (5.27) *of the barrier*

problem with sufficiently high order to obtain strictly feasible states. Then the following estimate holds:

$$\begin{aligned} J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) = \frac{1}{2}\int_{\Omega_C} \left(g(\overline{u}_\gamma,\nabla\overline{u}_\gamma) - g(\overline{u},\nabla\overline{u})\right) d\overline{\mu}_\gamma \\ - B_\gamma(\overline{u}_\gamma) + \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma,\nabla\overline{u}_\gamma)\, d\overline{\mu} \\ + \frac{1}{2}\rho_q(\overline{\xi})(\overline{q}-\overline{q}_\gamma) + \frac{1}{2}\rho_q(\overline{\xi}_\gamma)(\overline{q}-\overline{q}_\gamma) \\ + \mathcal{R}_{hom}, \end{aligned} \tag{6.100}$$

with a remainder term $\mathcal{R}_{hom}$ given by:

$$\mathcal{R}_{hom} = \frac{1}{2}\int_0^1 \mathcal{M}'''(\overline{\xi}_\gamma + s(\overline{\xi}-\overline{\xi}_\gamma))(\overline{\xi}-\overline{\xi}_\gamma,\overline{\xi}-\overline{\xi}_\gamma,\overline{\xi}-\overline{\xi}_\gamma)s(s-1)\,ds. \tag{6.101}$$

Proof. The proof is identical to the one for Theorem 6.4. But as $\mathcal{M}'_q \neq 0$ we obtain

$$\begin{aligned} \mathcal{M}'(\overline{\xi})(\overline{\xi}-\overline{\xi}_\gamma) &= \mathcal{M}'_\mu(\overline{\xi})(\overline{\mu}-\overline{\mu}_\gamma) + \mathcal{M}'_q(\overline{\xi})(\overline{q}-\overline{q}_\gamma), \\ \mathcal{M}'(\overline{\xi}_\gamma)(\overline{\xi}-\overline{\xi}_\gamma) &= \mathcal{M}'_\mu(\overline{\xi}_\gamma)(\overline{\mu}-\overline{\mu}_\gamma) + \mathcal{M}'_q(\overline{\xi}_\gamma)(\overline{q}-\overline{q}_\gamma) \end{aligned}$$

□

We see that, in addition, to the terms already present in Theorem 6.4 we have to estimate the two additional terms ρ_q. We remark that the simple estimate $\rho_q(\overline{\xi})(\overline{q}-\overline{q}_\gamma) \leq C\|\overline{q}-\overline{q}_\gamma\|_Q$ is possible, but is a large overestimation, as $\|\overline{q}-\overline{q}_\gamma\|_Q$ converges slower than the error in the functional, see Theorem 5.13.

To proceed, let us assume that $Q = L^2(\Omega)$, and that $J_q(q,u)(\cdot) = \alpha(q,\cdot)$ for some $\alpha > 0$.

To obtain a reasonable estimator we define the Lagrange multiplier for the control constraints as in (6.11) by

$$\mu_Q(\overline{\xi}_\gamma) = \mathcal{M}'_q(\overline{\xi}_\gamma)(\,\cdot\,)$$

as a Riesz representative. Then the following holds:

$$\mathcal{M}'_q(\overline{\xi}_\gamma)(\overline{q}-\overline{q}_\gamma) = \langle\, \mu_Q(\overline{\xi}_\gamma)\,, \overline{q}-\overline{q}_\gamma \,\rangle_{Q^*\times Q}.$$

As $\mu_Q \neq 0$ only on the active sets for the control constraints, we immediately obtain, that $\mathcal{M}'_q(\overline{\xi}_\gamma)(\overline{q}-\overline{q}_\gamma) = 0$ if the active sets coincide.

The multiplier μ_Q can be computed conveniently by using the available discrete multiplier. However in order to estimate $\overline{q}-\overline{q}_\gamma$ one may not use extrapolation of the discrete solutions for two different values of γ. This is because the dependence of elements $K \in \mathcal{T}_h$ to be either active or inactive not continuous with respect to γ on any given mesh. For instance, if a small step in γ is used, then the discrete active sets may coincide, hence the estimator will assume that no error is present. To circumvent this problem we propose to use projection formula (6.40).

The estimator then takes the following form for two given values $\gamma_1 < \gamma_2$. Determine the active sets

$$\mathcal{A}_\gamma^h = \{\, x \in \Omega \,|\, \pi^q \overline{q}_{\gamma_1}^h = a \text{ or } \pi^q \overline{q}_{\gamma_1}^h = b \,\},$$
$$\mathcal{A}^h = \{\, x \in \Omega \,|\, \tilde{q} = a \text{ or } \tilde{q} = b \,\},$$

where $\tilde{q}$ is computed by the projection formula (6.40) $\tilde{q} = \mathcal{P}_{Q^{\text{ad}}}\Big(\frac{1}{\alpha} a_q(\hat{q}, \hat{u})(\cdot, \hat{z})\Big)$. Here the variables denoted by $\hat{}$ are computed using linear extrapolation, e.g.,

$$\hat{z} = -\frac{1}{\gamma_1} \frac{\overline{z}_{\gamma_1}^h - \overline{z}_{\gamma_2}^h}{1/\gamma_1 - 1/\gamma_2}.$$

Then we consider the following estimator

$$\eta_{\text{CC}} = \eta_{\text{hom}} + \int_{\mathcal{A}_\gamma^h \cup \mathcal{A}^h} \mu_Q(\overline{\xi}_\gamma^h)(\overline{q}_{\gamma_1}^h - \overline{q}_{\gamma_2}^h)\, dx$$

as an estimator for the regularization error in the case of control constraints.

6.2.1.6 Numerical Results

Once again we consider a numerical example in order to show our findings. The state variable is discretized using continuous Q_1 finite elements while the control is discretized using P_0 discontinuous finite elements on the same triangulation. For the regularization we used a sum of barrier functions of order $\kappa = 2, \ldots, 6$. All computations where made using the software packages RoDoBo (RoDoBo [126]) and Gascoigne (Gascoigne [65]). The visualization is done using Visit (Visit [149]).

We consider the cube $(0,1)^3 \subset \mathbb{R}^3$, and the following optimization problem:

$$\text{Minimize } J(q,u) := \frac{1}{2}\|u - u^d\|_{L^2(\Omega)}^2 + \frac{10^{-3}}{2}\|q\|_{L^2(\Omega)}^2,$$
$$\text{such that } \begin{cases} (\nabla u, \nabla \varphi) = (q, \varphi) & \forall\, \varphi \in H_0^1(\Omega), \\ (q,u) \in L^2(\Omega) \times H_0^1(\Omega), & \\ -40 \le q(x) \le 40 & \forall\, x \in \Omega, \\ -1 \le u(x) & \forall\, x \in \overline{\Omega}. \end{cases}$$

The desired state is chosen as $u^d(x) = -5\sin(\pi r)$ where $r = \sqrt(|x|^2)$ is the euclidian distance to the origin. Especially it should be noted that u^d is infeasible for the state constraint.

As we do not have an analytic solution for this example we computed the values of $J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)$ on globally refined meshes of maximal 274625 vertices and $\gamma = 1.8 \cdot 10^3$ and extrapolated the value of $J(\overline{q}, \overline{u}) \approx 4.2827$. The choice of γ is such that the discretization error is dominant.

We begin by reconsidering the discretization error estimate as announced at the end of Section 6.2.1.4. Therefore we consider the error estimates for both discretization error $\eta_{\text{disc}} = \eta_h^{(1)}$

from Section 6.1.2.1 and the estimator for the regularization error η_{CC} proposed in Section 6.2.1.5 in Table 6.12. We immediately see that the effectivity index is almost one. Both in the case of dominant discretization as well as dominant regularization error. First of all this implies that the regularization error gives a good estimate for the regularization error contribution. Second we see that the integration error coming from the barrier functional is not critical for the discretization error estimate.

Table 6.12: Effectivity of $\eta_{\mathrm{disc}} + \eta_{\mathrm{CC}}$. Values with dominant regularization error are on gray background.

	N			
γ	125	729	4913	35937
$1.0 \cdot 10^1$	1.2	1.0	0.9	0.9
$5.6 \cdot 10^1$	1.2	1.1	1.0	1.0
$2.3 \cdot 10^2$	1.2	1.1	1.0	1.0
$1.3 \cdot 10^3$	1.2	1.1	1.0	1.0

In Figure 6.11 the solution on a mesh with 274625 vertices and a value $\gamma = 1.8 \cdot 10^3$ is shown. In order to give an impression of the solution variables, the intersection of the three dimensional domain with surfaces given by a constant x_3 value is shown. The left column shows the active-set for the control constraints. Black indicates an active lower bound while white indicates an active upper bound. In the second column of Figure 6.11 the discrete approximation to the Lagrange multiplier $\overline{\mu}$ is depicted, where black color indicates approximate zero values. The next two columns show the control and state variable.

In Figure 6.10 the behavior of the error under mesh refinement is shown. We can see that by using local mesh refinement approximately a factor of two is gained in terms of degrees

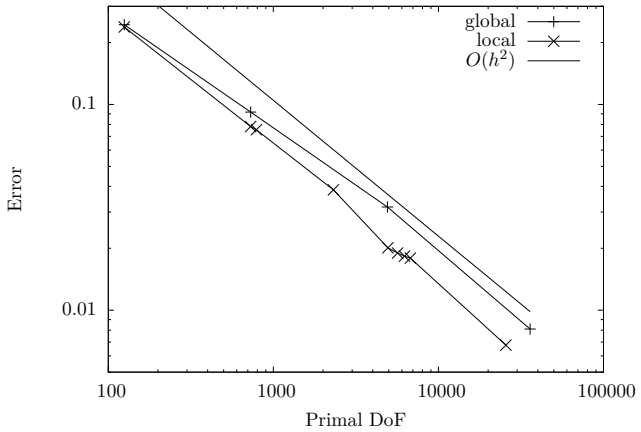
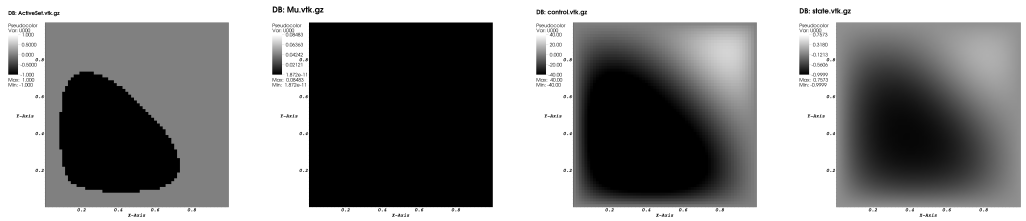
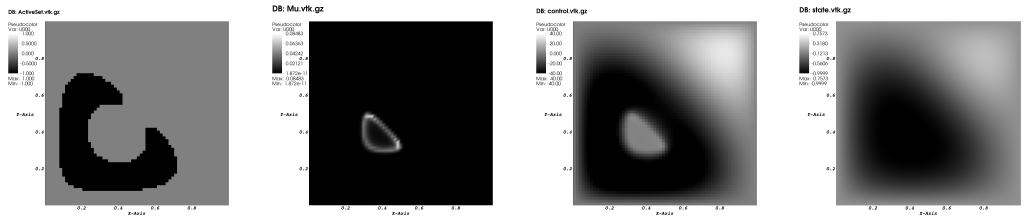


Figure 6.10: Comparison of the errors in the functional for different refinement strategies.

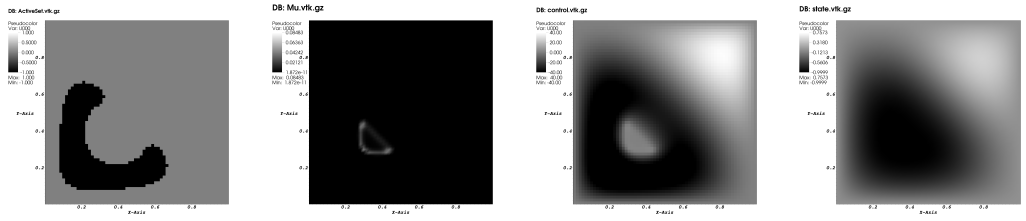
of freedom (DoF) for the state variable. That we could not obtain larger savings can be explained by the fact that the error under global refinement is already converging with order h^{-2}.



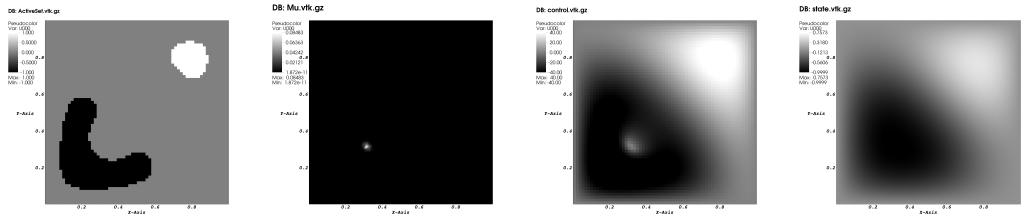
(a) Active-Set $z = 0.2$ (b) Multiplier $z = 0.2$ (c) Control $z = 0.2$ (d) State $z = 0.2$



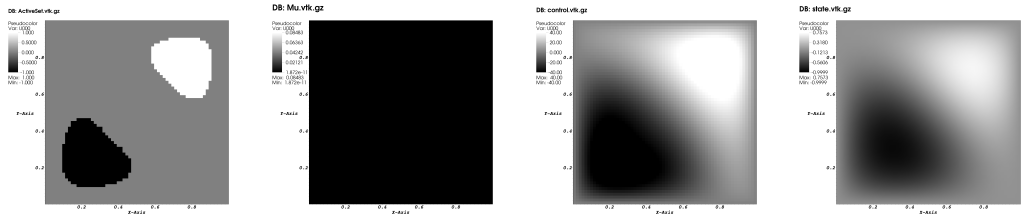
(e) Active-Set $z = 0.3$ (f) Multiplier $z = 0.3$ (g) Control $z = 0.3$ (h) State $z = 0.3$



(i) Active-Set $z = 0.4$ (j) Multiplier $z = 0.4$ (k) Control $z = 0.4$ (l) State-$z = 0.4$



(m) Active-Set $z = 0.5$ (n) Multiplier $z = 0.5$ (o) Control $z = 0.5$ (p) State $z = 0.5$



(q) Active-Set $z = 0.6$ (r) Multiplier $z = 0.6$ (s) Control $z = 0.6$ (t) State $z = 0.6$

Figure 6.11: Active-set for the control constraints, Lagrange multiplier $\overline{\mu}$ for the state constraints, control, and state variable. (Black - small values, white - large values)

6.2.1.7 Penalty Regularization

We now consider the case of a penalty method (5.26), as in Section 6.2.1.1, we consider the case without control constraints.

We obtain the following error representation:

Theorem 6.7. *Let $\overline{\xi} = (\overline{q}, \overline{u}, \overline{z}, \overline{\mu})$ be a solution to the first-order necessary system* (2.10) *and let $\overline{\xi}_\gamma = (\overline{q}_\gamma, \overline{u}_\gamma, \overline{z}_\gamma, \overline{\mu}_\gamma)$ a solution to the first-order necessary system* (5.28) *for the penalty problem. Then the following identity holds:*

$$\begin{aligned} J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) = \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu} \\ - \frac{1}{2} \int_{\Omega_C} g(\overline{u}, \nabla \overline{u})\, d\overline{\mu}_\gamma + \mathcal{R}_{hom}, \end{aligned} \tag{6.102}$$

with a remainder term $\mathcal{R}_{hom}$ given by:

$$\mathcal{R}_{hom} = \frac{1}{2} \int_0^1 \mathcal{M}'''(\overline{\xi}_\gamma + s(\overline{\xi} - \overline{\xi}_\gamma))(\overline{\xi} - \overline{\xi}_\gamma, \overline{\xi} - \overline{\xi}_\gamma, \overline{\xi} - \overline{\xi}_\gamma) s(s-1)\, ds. \tag{6.103}$$

Here $\mathcal{M}$ is defined exactly as in (6.89).

Proof. As in the proof of Theorem 6.4 we obtain that $J(\overline{q}, \overline{u}) = \mathcal{M}(\overline{\xi})$. By definition of $\overline{\mu}_\gamma$, see (5.28), we obtain further

$$\begin{aligned} J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) &= J(\overline{q}_\gamma, \overline{u}_\gamma) + \frac{\gamma}{2} \| g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)^+ \|^2_{\Omega_C} \\ &= J(\overline{q}_\gamma, \overline{u}_\gamma) + \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma \\ &= \mathcal{M}(\overline{\xi}_\gamma) - \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma. \end{aligned}$$

With this we can proceed exactly as in the proof of Theorem 6.4 to obtain

$$J(\overline{q}, \overline{u}) - J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) = \frac{1}{2} \mathcal{M}'_\mu(\overline{\xi})(\overline{\mu} - \overline{\mu}_\gamma) + \frac{1}{2} \mathcal{M}'_\mu(\overline{\xi}_\gamma)(\overline{\mu} - \overline{\mu}_\gamma) + \frac{1}{2} \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma + \mathcal{R}.$$

Now we consider the first term on the right-hand side, and see using complementarity (2.10e)

$$\mathcal{M}'_\mu(\overline{\xi})(\overline{\mu} - \overline{\mu}_\gamma) = \langle\, \overline{\mu} - \overline{\mu}_\gamma\, , g(\overline{u}, \nabla \overline{u})\, \rangle_{C^* \times C} = -\langle\, \overline{\mu}_\gamma\, , g(\overline{u}, \nabla \overline{u})\, \rangle_{C^* \times C}.$$

For the second and third term we obtain

$$\begin{aligned} \mathcal{M}'_\mu(\overline{\xi}_\gamma)(\overline{\mu} - \overline{\mu}_\gamma) + \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma &= \langle\, \overline{\mu} - \overline{\mu}_\gamma\, , g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, \rangle_{C^* \times C} + \int_{\Omega_C} g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, d\overline{\mu}_\gamma \\ &= \langle\, \overline{\mu}\, , g(\overline{u}_\gamma, \nabla \overline{u}_\gamma)\, \rangle_{C^* \times C} \end{aligned}$$

and hence the assertion follows. □

Using the fact that $0 \leq J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma)$ we obtain from the first-order necessary conditions

$$0 \leq J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma,\overline{u}_\gamma) \leq \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma,\nabla\overline{u}_\gamma) - g(\overline{u},\nabla\overline{u})\, d(\overline{\mu}+\overline{\mu}_\gamma).$$

We proceed heuristically by assuming that $g(\overline{u},\nabla\overline{u}) \approx 0$ on the support of $\overline{\mu}+\overline{\mu}_\gamma$, to obtain

$$\frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma,\nabla\overline{u}_\gamma) - g(\overline{u},\nabla\overline{u})\, d(\overline{\mu}+\overline{\mu}_\gamma) \approx \frac{1}{2}\int_{\Omega_C} g(\overline{u}_\gamma,\nabla\overline{u}_\gamma)\, d(\overline{\mu}+\overline{\mu}_\gamma) \approx \int_{\Omega_C} g(\overline{u}_\gamma)\, d\overline{\mu}_\gamma.$$

As an estimate we hence propose to use

$$\eta_{\text{reg}} = \int_{\Omega_C} g(\overline{u}_\gamma^h,\nabla\overline{u}_\gamma^h)\, d\overline{\mu}_\gamma^h$$

where $\overline{u}_\gamma^h$ is the solution to the Galerkin discretization (5.32).

The derivation of the corresponding discretization error is rather straight forward. Hence we do not derive it here separately. In the case of zero-order constraints it can be found in (Günther and Tber [77]). The only difficulty is in the fact, that $q(\,\cdot\,)^+$ is only directional differentiable. The estimator itself has the same form as it would have without the non differentiability.

Numerical examples We will consider the two numerical examples from Section 6.2.1.2 to illustrate that this approximation gives reasonable results. Once again the computations are done using the software packages RoDoBo (RoDoBo [126]) and Gascoigne (Gascoigne [65]).

Gradient Constraints Here we once again consider the problem from (Deckelnick et al. [54]), e.g.,

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2_{L^2(\Omega)} + \frac{1}{2}\|q\|^2_{L^2(\Omega)}$$
$$\text{subject to } \begin{cases} (\nabla u,\nabla\varphi) = (f+q,\varphi) & \forall\,\varphi\in H^1_0(\Omega),\\ \frac{1}{4} - |\nabla u(x)|^2 \geq 0 & \forall\, x\in\overline{\Omega},\\ q\in Q_{\text{ad}}. \end{cases}$$

The problem with active sets is now, that for Q_1 finite elements the gradient is not constant on each mesh element. This means that tracking the active set is difficult on the discrete level. To overcome this, and the fact that ∇u_h is discontinuous, we introduce an additional variable $w = |\nabla u|^2$. Hence we consider

$$\text{Minimize } J(q,u) = \frac{1}{2}\|u-u^d\|^2_{L^2(\Omega)} + \frac{1}{2}\|q\|^2_{L^2(\Omega)}$$
$$\text{subject to } \begin{cases} (\nabla u,\nabla\varphi) = (f+q,\varphi) & \forall\,\varphi\in H^1_0(\Omega),\\ (w,\varphi) = (|\nabla u|^2,\varphi) & \forall\,\varphi\in L^2(\Omega),\\ \frac{1}{4} - w \geq 0 & \forall\, x\in\overline{\Omega},\\ q\in Q_{\text{ad}}. \end{cases}$$

We begin by a consideration of the convergence behavior of the cost functional. In Figure 6.12 we have depicted the convergence of the quantity $J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)$. We can clearly see that

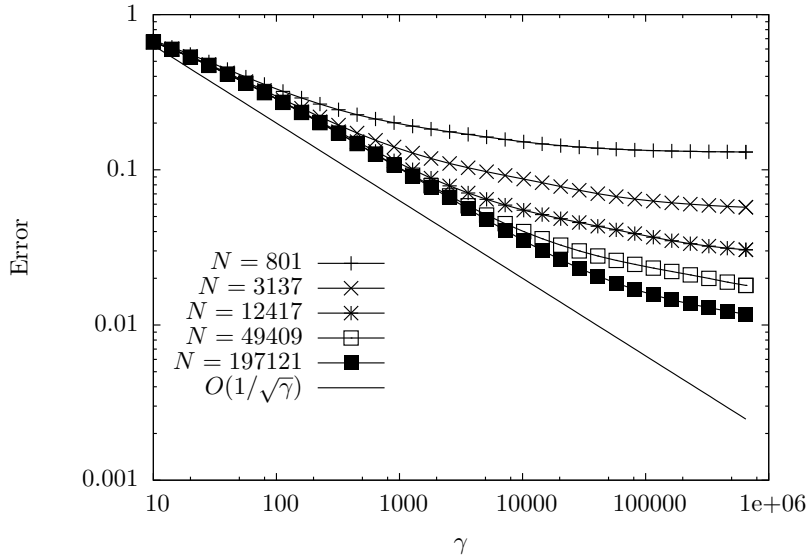


Figure 6.12: Convergence behavior of $J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)$

the functional converges with order $O(\gamma^{-1/2})$. By transferring the results of (Hintermüller and Hinze [79]) to the case of first-order state constraints a convergence order between $O(\gamma^{-1/2})$ and $O(\gamma^{-1})$ had to be expected. We remark that this directly shows that an a priori choice of the relation between h and γ is difficult, as the convergence behavior is not known a priori, but has to be found during the computation.

We will now consider the same question as in Section 6.2.1.4, namely whether the proposed method of estimating the regularization error is sufficiently accurate. We will consider a range of parameters of γ between 10 and 10000. In this range we will be able to see the behavior of the regularization estimate in the vicinity of the equilibrium of regularization and discretization error, see Figure 6.12.

Table 6.13: Efficiency of $\eta_{\text{disc}} + \eta_{\text{reg}}$ on various meshes. Values with $|\eta_{\text{reg}}|$ below the discretization error are on gray background.

	γ								
N	$2 \cdot 10^1$	$4 \cdot 10^1$	$8 \cdot 10^1$	$2 \cdot 10^2$	$3 \cdot 10^2$	$6 \cdot 10^2$	$1 \cdot 10^3$	$3 \cdot 10^3$	$5 \cdot 10^3$
801	1.5	1.4	1.4	1.5	1.7	2.0	2.3	2.5	2.6
3137	1.5	1.3	1.3	1.2	1.3	1.4	1.7	1.9	2.1
12417	1.4	1.3	1.2	1.2	1.1	1.1	1.2	1.3	1.4
49409	1.4	1.3	1.2	1.1	1.1	1.1	1.1	1.1	1.2
197121	1.5	1.3	1.2	1.1	1.1	1.1	1.1	1.1	1.1

In Table 6.13 we have depicted the effectivity index

$$I_{\text{eff}} = \frac{|J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h,\overline{u}_\gamma^h)|}{|\eta_{\text{disc}}| + |\eta_{\text{reg}}|}$$

on different meshes for various choices of γ. The sequence of γ was obtained by starting from $\gamma_0 = 10$ and then successively increasing γ by a factor of $\sqrt{2}$. The results clearly show that the estimate η_{reg} a good estimate for the influence of the regularization error, although it is slightly underestimating the real error.

Finally, we will have a short look on the interplay between the discretization error estimate and the regularization error estimate. For this we consider the behavior for both indicators separately on globally refined meshes with 12417 and 49409 vertices. The results are depicted in Figure 6.13. Here we can see that the discretization error indicators are growing towards a

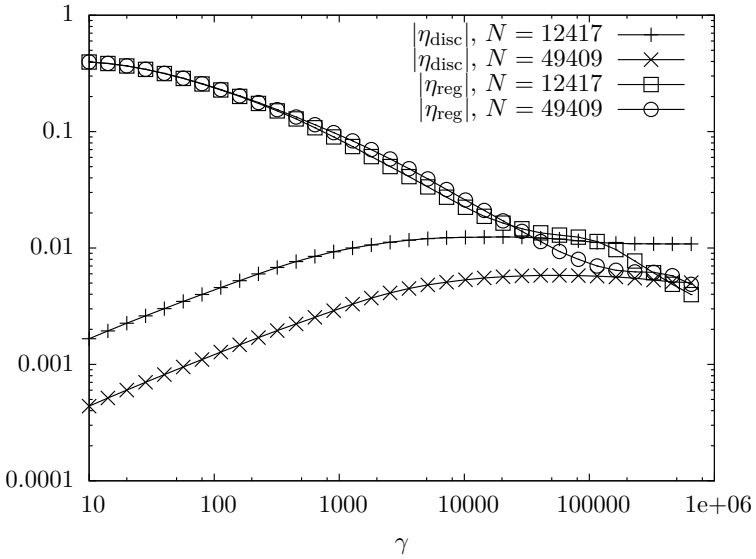


Figure 6.13: Convergence behavior of the error indicators

limit for $\gamma \to \infty$. This is exactly what we must expect, as a solution to the limiting problem exists and should be harder to approximate by a discretization due to the measure in the right-hand side of the adjoint equation. Next, we obtain that for $|\eta_{\text{reg}}| \gg |\eta_{\text{disc}}|$ the indicator for the regularization is almost unchanged under mesh refinement. Hence it makes sense to call $|\eta_{\text{reg}}|$ an estimate for the regularization error. However when $|\eta_{\text{reg}}| \approx |\eta_{\text{disc}}|$ the indicator remains stagnant for a short range of γ values, before they are again almost identical. This behavior indicates that when balancing the contributions of both indicators one should not try to have $|\eta_{\text{reg}}| \ll |\eta_{\text{disc}}|$ in order to obtain an efficient algorithm. From Figure 6.14 we can clearly see, that both strategies of balancing either balancing the error contributions as well as letting $\gamma \to \infty$ lead to comparable results concerning the error. However, the computational costs for the balancing strategy are far less.

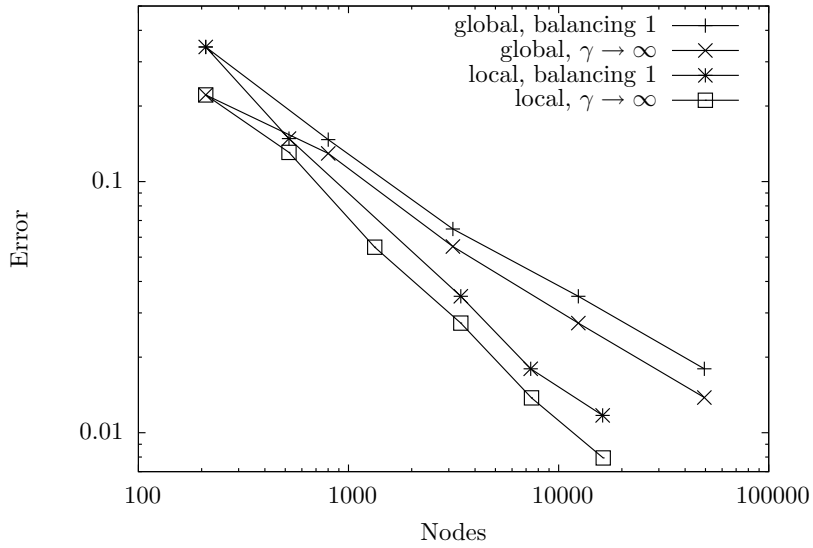


Figure 6.14: Convergence behavior of the error indicators

State Constraints We return to the example which we derived from (Günther and Hinze [75]) by transformation to constant bounds

$$\text{Minimize } J(q,u) := \frac{1}{2}\|(0.45-\psi)v+\psi-0.5\|^2_{L^2(\Omega)} + \frac{1}{2}\|q-60\|^2_{L^2(\Omega)}$$

$$\text{subject to } \begin{cases} -\Delta((0.45-\psi)v+\psi) + ((0.45-\psi)v+\psi) = q \quad \text{in } \Omega, \\ \partial_n v = 0 \quad \text{on } \partial\Omega, \\ 0 \le v(x) \le 1 \quad \forall\, x \in \overline{\Omega}, \\ q \in L^2(\Omega). \end{cases}$$

As in the previous example we begin the discussion by comparing the convergence of the functional value $J(\overline{q},\overline{u}) - J_\gamma(\overline{q}^h_\gamma, \overline{u}^h_\gamma)$ in Figure 6.15. We immediately see that we have a small range of γ values where we have a convergence of $O(\gamma^{-1/2})$. However, in contrast to the previous example the values here are not yet constant under mesh refinement, e.g., we have a discretization influence on the error with dominant regularization error. This is probably caused by the fact, that the coefficients in the equation and cost functional are not integrated exactly but with a tensor product four-point Gauss-Lobatto quadrature formula.

Then there is, on each mesh, a rather long transition zone between dominant regularization error and the level of the discretization error. This, once again, confirms our conclusion at the end of the previous example that it is not advisable to attempt to stir γ and h such that $\eta_{\text{reg}} \ll \eta_{\text{disc}}$.

We now turn to the evaluation of the quality of the estimators. To this end we consider in Table 6.14 the effectivity index of our estimate. As in the previous example we can see that the effectivities are almost one and not changing when the regularization error is

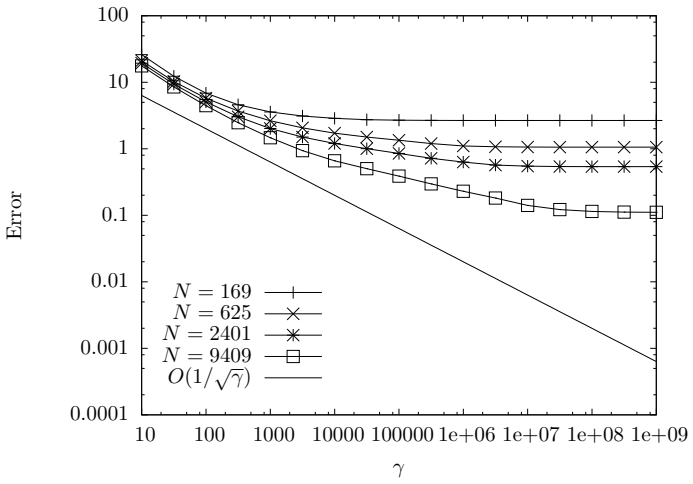


Figure 6.15: Convergence behavior of $J(\overline{q},\overline{u}) - J_\gamma(\overline{q}_\gamma^h, \overline{u}_\gamma^h)$

dominant. From this we can already obtain, that the use of the regularization error estimate is apparently reasonable.

Table 6.14: Efficiency of $\eta_{\text{disc}} + \eta_{\text{reg}}$ on various meshes. Values with $|\eta_{\text{reg}}|$ below the discretization error are on gray background.

	γ								
N	$1 \cdot 10^1$	$1 \cdot 10^2$	$3 \cdot 10^2$	$1 \cdot 10^3$	$3 \cdot 10^3$	$1 \cdot 10^4$	$3 \cdot 10^4$	$1 \cdot 10^5$	$1 \cdot 10^6$
169	0.7	1.1	1.6	2.4	3.3	5.0	8.2	13.8	22.9
625	0.7	1.0	1.1	1.2	1.3	1.4	1.4	1.5	1.5
2401	0.7	0.9	1.1	1.3	1.5	1.7	1.9	1.9	1.8
9409	0.7	0.9	1.0	1.1	1.2	1.4	1.5	1.3	1.2
37249	0.7	0.9	0.9	1.0	1.0	1.0	1.0	0.9	0.4

However, it is not yet clear, how to account for the visible change in the error for dominant regularization error under mesh refinement seen in Figure 6.15. To this end we consider the behavior of both discretization and regularization error indicator on to consecutive meshes and different values of γ. The results are depicted in Figure 6.16. Here we can see first of all, that the values of γ where both indicators are equilibrated coincide with the values where the error is almost on discretization accuracy. This means that in the region of interest both indicators are accurate. As in the previous example we see, that near equilibration the regularization error indicator becomes stagnant for a while, indicating as before that it is not efficient to achieve $\eta_{\text{reg}} \ll \eta_{\text{disc}}$.

As already mentioned, at the end of Section 6.2.1.4, we neglected the influence of the quadrature rule, or more precisely the fact that we do not integrate the coefficients in the equation and the cost functional exactly. As in Section 6.2.1.4 we are considering here a

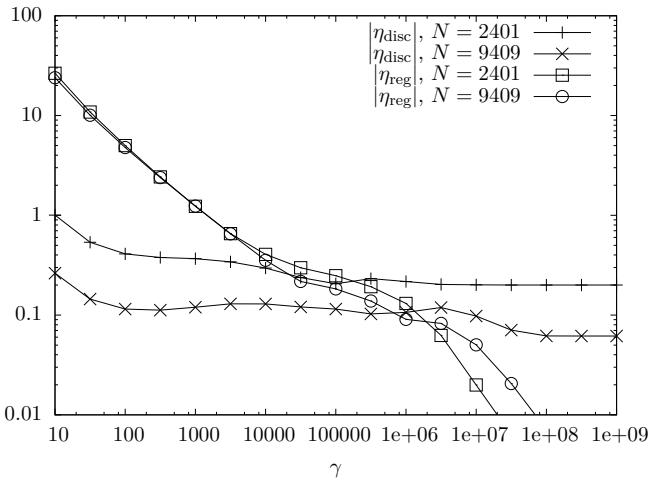


Figure 6.16: Convergence behavior of the error indicators

problem, where the quadrature is not exact for the problem data—here the coefficients of the equation—.

In order to substantiate this we reconsider the example this time using a summed midpoint rule where each element is split into 2^8 subelements for the integration. The results are shown in Table 6.15. By comparing Table 6.15 to Table 6.14 we immediately see that the use of a

Table 6.15: Efficiency of $\eta_{\text{disc}} + \eta_{\text{reg}}$ on various meshes for summed quadrature. Values with $|\eta_{\text{reg}}|$ below the discretization error are on gray background.

	γ								
N	$1 \cdot 10^1$	$1 \cdot 10^2$	$3 \cdot 10^2$	$1 \cdot 10^3$	$3 \cdot 10^3$	$1 \cdot 10^4$	$3 \cdot 10^4$	$1 \cdot 10^5$	$1 \cdot 10^6$
169	0.7	1.3	2.0	3.1	5.4	4.8	4.0	3.5	3.0
625	0.7	0.9	1.1	1.2	1.2	1.3	1.3	1.4	1.5
2401	0.7	0.9	1.0	1.1	1.2	1.2	1.3	1.3	1.3
9409	0.7	0.9	1.0	1.0	1.1	1.2	1.2	1.1	1.0

more accurate quadrature formula for the solution of the equations leads to a more accurate estimation of the error using the regularization and discretization error estimators derived in this thesis.

7 Algorithmic Aspects

7.1 Control Constraints

For the solution of the control constraint problem (6.3), we apply a nonlinear primal-dual-active-set strategy, see, e.g., (Bergounioux et al. [22], Kunisch and Rösch [98]). In the following, we sketch the corresponding algorithm on the continuous level. We assume the control to be from some $L^p(\omega)$ and $Q^{\mathrm{ad}} = \{\, q \in L^p(\omega) \,|\, a \le q(x) \le b \text{ for almost all } x \in \omega \,\}$.

Nonlinear primal-dual active set strategy

1. Choose initial guess q^0, μ^0 and $c > 0$ and set $n = 1$
2. While not converged
3. Determine the active sets ω_+^n and ω_-^n

$$\omega_-^n = \{x \in \omega \,|\, q^{n-1}(x) + \mu^{n-1}(x)/c - a \le 0\}$$
$$\omega_+^n = \{x \in \omega \,|\, q^{n-1}(x) + \mu^{n-1}(x)/c - b \ge 0\}$$

4. Solve the equality-constrained optimization problem

$$\text{Minimize} \quad J(q^n, u^n), \quad u^n \in V,\, q^n \in Q,$$

   subject to (2.4) and

$$q^n(x) = a \text{ on } \omega_-^n\,, \qquad q^n(x) = b \text{ on } \omega_+^n\,.$$

5. Set

$$\mu^n = -J_q'(q^n, u^n)(\cdot) + a_q'(q^n, u^n)(\cdot, z^n)$$

   with adjoint variable z^n.

6. Set $n = n + 1$ and go to 2.

The algorithm above is known to be globally convergent for certain classes of optimal control problems, see, e.g., (Bergounioux et al. [22], Kunisch and Rösch [98]). Moreover local superlinear convergence can be shown, see, e.g., (Hintermüller et al. [85]).

Concerning the practical realization, the active sets in step 3 can be determined in a finite number of points in ω. There are two cases to distinguish.

1. If the control is finite dimensional, e.g., $\omega = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$. Then step 3 is as in the continuous case.

2. If the control is infinite dimensional, e.g., $\omega \subset \Omega$ or $\omega \subset \partial\Omega$. Then the control is discretized using either P_0 or Q_1 finite elements. And we can determine the active set by comparing the components of the coordinate vectors with respect to the usual nodal basis.
   Hence in the P_0-case we compare the bound with the cellwise constant values of q and μ. In the Q_1-case we compare the values associated with the vertices of the grid to the bounds.

Convergence in step 2 can be determined conveniently from agreement of the active sets in two consecutive iterations.

Remark 7.1. In order to correctly determine μ^n in step 5 one has to choose a scalar product for the space Q such that the representation of μ^n can be computed.

In the practical realization in RoDoBo (RoDoBo [126]), the equality-constrained optimization problem in step 4 is solved by Newton's method on the control space without assembling the Hessian, for details, see (Meidner [107]).

This approach has the advantage that it is applicable once the state equation is solvable, e.g., whenever a good solver for the state equation is available. However, it has the drawback of being slow compared with a multigrid method directly applied to the KKT-system. Due to the saddle point structure of the KKT-System good multigrid methods are unfortunately not readily available, but are a field of active research even in the case of pure PDE constraint optimization, see, e.g., (Bauer [9], Biros and Ghattas [24, 25], Rees, Dollar, and Wathen [122], Schöberl and Zulehner [138]) or (Borzi and Schulz [31]) for a survey. Preconditioners for the solution of the KKT-System in PDE-based optimal control with regularized state constraints using a CG-method are considered in (Herzog and Sachs [78]). Preconditioning in the case of optimal control of the reduced problem is still open to further research.

7.2 State Constraints

For the solution of the discrete state constraint problems we employed both penalty and barrier methods for the computations in this thesis. As we introduced a common parameter $\gamma \to \infty$ for both problems, the algorithm used for the solution of these problems reads as follows:

Solution of the state constrained problem

1. Choose initial $\gamma^0 > 0$, mesh size distribution $h^0 > 0$, $\tilde{q}^0 \in Q_{h^0}$ and set $n = 0$.
2. Solve the control-constrained optimization problem

$$\text{Minimize} \quad J_{\gamma^n}(q^n, u^n), \quad u^n \in V_{h^n},\, q^n \in Q_{h^n}^{\text{ad}},$$

   subject to (5.31), with initial guess $\tilde{q}^n$ using the algorithm from Section 7.1.
3. While not converged, set $n = n + 1$
4. Determine h^n, γ^n, and $\tilde{q}^n$ and go to 2.

In the following, we will give details on steps 2 and 4 of the algorithm. We do not discuss step 1 and 3 here, as the only problem in step 1 is that for barrier methods the initial guess $\tilde{q}^0$ has to be chosen such that the corresponding state is feasible with respect to the state constraints because we solve the problems in the control space. In our applications such an initial guess was always straight forward to find, e.g., $\tilde{q}^0 \equiv 0$ was usually good enough. Finally, the convergence in step 3 depends on the goal of the computation, hence there can not be an all purpose stopping criterion.

Step 2. Solving the subproblems

Barrier Methods In this case we remark that although in the limit $h \to 0$ the barrier solution is, in general, only strictly feasible almost everywhere, see also Theorems 5.6 and 5.8. On the discrete level the solution is always strictly feasible, and hence J_γ is differentiable in a neighbourhood of the discrete solution. Hence one can directly apply Newton's method as proposed in Section 7.1. However in order to avoid leaving the region of strict feasibility one has to take care during the the Newton update, e.g., whenever the new iterate would be infeasible on has to perform damping in order to remain in the feasible region. Here we used a simple line-search method to ensure feasibility of the iterates. This may lead to several damping steps during the iteration, hence more efficient damping strategies may be required. For a KKT-based solver a special modification has been proposed recently in (Schiela [132]).

Furthermore, we remark that higher derivatives of the barrier function are large near the boundary of the feasible region. If the integration is done exact one would always retain some distance to the boundary. Hence special care has to be taken when selecting the quadrature formulas for the integration, e.g., when using Gauss quadrature formulas one may even obtain infeasible solutions.

We used either summed quadrature formulas, or Gauss-Lobatto formulas, the latter have the advantage of including the nodal values of the solution, which are on Cartesian meshes the points where the distance between the (constant) bound and the finite element function is

smallest. In our experience Gauss-Lobatto formulas gave equally good results as summed quadrature formulas while being much faster. A discussion on the effect of the quadrature rule in the case of zero-order constraints can be found in (Hinze and Schiela [89]).

Penalty Methods In the case of penalty methods the penalty term $\|g(u_h, \nabla u_h)^+\|^2$ is not twice differentiable, however, it is Newton differentiable, see, e.g., (Hintermüller and Kunisch [83]). The Newton derivative can be easily calculated noting that given an iterate u the penalty term takes the form

$$\|g(u, \nabla u)^+\|^2 = \|g(u, \nabla u)\|^2_{\mathcal{A}(u)},$$

where $\mathcal{A}(u) = \{\, x \in \Omega \,|\, g(u, \nabla u) > 0 \,\}$ is the 'active-set' of the state constraint. Then the Newton derivative is obtained by differentiating the integrand of the second term while fixing the set $A(u)$. For more details on the algorithm we refer to (Hintermüller and Kunisch [81, 82]).

In the discretization of this one has to be careful how to define the active set. In contrast to the case of control constraints the active set may not be determined by the value of $g(u, \nabla u)$ in the vertices of the triangulation. It has to be determined during integration of the penalty term and its derivatives. Hence a sufficiently accurate quadrature rule has to be employed during the integration in order to avoid problems during the Newton iteration. However, as exact integration is costly we employed Gauss-Lobatto formulas for our computation, although this does introduce additional newton steps due to bad linearization, it is much faster than the use of summed quadrature rules on the elements with a jump in the indicator function of $\mathcal{A}(u)$. In addition we introduced an additional variable $w = g(u, \nabla u)$ for the constraint if $g(u_h, \nabla u_h) \notin V_h$

Step 4. Preparing the next iteration In the case $h^n = h^{n-1}$ we can simply choose a value of γ^n and use $\tilde{q}^n = q^{n-1}$ as an initial point. Although it is advisable not to choose $\gamma^n \gg \gamma^{n-1}$ in order to have a sufficiently good starting value for the next Newton iteration. For the choice of the parameters see, e.g., (Hintermüller and Kunisch [81, 82], Schiela [132], Weiser and Deuflhard [152]).

The case of $h^n \neq h^{n-1}$ one is tempted to use $\tilde{q}^n = Iq^{n-1}$ with some simple operator $I\colon Q_{h^{n-1}} \to Q_{h^n}$, e.g., nodal interpolation. This is fine in the case of a penalty approach, however in the case of a barrier method the corresponding state $\tilde{u}^n \in Q_{h^n}$ may violate the feasibility constraint. In the examples considered in this thesis there was always a strictly feasible point $\breve{q} \in V_{h^n}$ known, such that $\tilde{q}^n := \lambda\breve{q} + (1-\lambda)Iq^{n-1}$ is strictly feasible for some $\lambda \in [0, 1)$ which can be determined by a line-search procedure.

8 Conclusions and Outlook

In this thesis we considered elliptic PDE-based optimal control problems subject to pointwise inequality constraints for the control and state variable. We considered questions of existence of solutions subject to first-order state constraints on non-smooth domains. We derived a new regularity result for the solution of such problems under first-order state constraints. Based on this we derived a priori error estimates for the discretization error caused by the finite element approximation of the problem. We proceeded with the derivation of error estimates with respect to a possible regularization of the problem. Finally, we derived a posteriori error estimates for both regularization and discretization error. Numerically the estimates have shown to be separated, so that we are able to balance the contributions arising from regularization and discretization to the global error. We will, in the sequel, recall what has been achieved, and what possible extensions of the work presented here would be.

Existence on non-smooth domains & regularity In view of applications, that typically require domains whose boundary is only piecewise smooth, we showed for a simple model problem with first-order state constraints that there still exists a solution to the minimization problem, even if the control to state mapping is not sufficiently regular to pose the state constraint on the whole image of the control space. The method employed is directly transferable to more complex equations, e.g., those of elasticity. However, with respect to first-order necessary optimality conditions there appears to be a gap between existence of a solution and existence of corresponding Lagrange multipliers which might be of further interest. Especially first-order necessary conditions are of interest, as several algorithms try to compute solutions to these conditions. In addition it is of interest with respect to a posteriori error estimation.

A priori error estimates In the field of a priori analysis, we have shown convergence of the discretization error for a direct discretization of the optimization problem with first-order state constraints by finite elements. The results give the same convergence orders that have been recently obtained by other authors, but the arguments in the proof presented here are far more elementary. Furthermore, we extended the convergence theory to the case of a bilinear control discretization. The analysis is directly transferable to any finite element discretization of the state equation that is stable with respect to $W^{1,\infty}$. In particular this shows that convergence is available for any reasonable discretization. However, the orders of convergence obtained here and those in the literature are not optimal with respect to best-approximation in view of the regularity of the control. However, numerical evidence shows, that the derived orders of convergence are sharp with respect to the control. This indicates that by a direct discretization of the optimization problem one does, in general, not have a

quasi best-approximation property for the control variable. Hence an interesting question for further research would be a discretization scheme which obtains optimal convergence with respect to the regularity of the continuous solution.

Regularization methods Concerning the solution of the state constrained optimization problem, we analyzed a barrier method for first-order state constraints. We were able to show convergence (with rates) of the method applied to the continuous problem. Due to the general statement of the results it is also applicable to the discretized problems. Hence by combination of the results for the discretization error for the state constrained problem, one can easily deduce convergence of the overall algorithm combining regularization and discretization. However, in the case of first-order state constraints there are still open questions concerning the step length selection and damping rules to obtain a very fast algorithm.

A posteriori error estimates Finally, concerning the main issue of the work: We were the first to have derived a posteriori error estimates for both the error in the cost functional as well as in an arbitrary quantity of interest for PDE-based optimization problems subject to pointwise control constraints. Then this work contains pioneering work for such problems with pointwise first or zero-order state constraints. We derived estimates for the regularization error due to both barrier or penalty methods applied to the state constraint. For the remaining regularized problem one can apply the discretization error estimates derived earlier in this work. We showed that for several model problems both error indicators are sufficiently well separated to allow for a balancing strategy of both error components. However, the extension of the regularization error estimate towards an arbitrary target quantity remains a very interesting open problem. The problem in applying the techniques derived in this thesis lies in the fact that first-order necessary conditions for the continuous dual problem associated to the quantity of interest are not known in general. In addition, the application of adaptive quadrature or an appropriate estimator for the quadrature error in the solution process might be advantageous as indicated by some of the numerical examples. Moreover, the application of the derived estimates to real world applications will be a very interesting perspective.

Acknowledgments

I would like to express my gratitude to my supervisor Prof. Dr. h.c. Rolf Rannacher for suggesting this interesting subject. In addition, I would like to thank him for his ongoing support, the possibility to present my work at several conferences and workshops, and of course the possibility to gain experience in both research and teaching.

In addition, I would like to thank all my collaboration partners, especially Christoph Ortner, Anton Schiela and Boris Vexler with whom I had the chance to work on joint papers that contributed to the content of this thesis. I would also like to thank all my other colleagues from the numerical analysis group and elsewhere on this planet with whom I had several interesting discussions.

This whole thesis would not have been possible without the software I have been using. Therefore I would like to thank the teams of Gascoigne, RoDoBo and VisuSimple. Further, I would like to thank all the people involved in the development of LaTeX, Gnuplot, GCC, and Emacs.

Further, I would like to thank Katrin Kohoutek, my friends, and family for their encouragement and ongoing support in all its forms.

Last but not least, I gratefully acknowledge the funding received by the German Research Foundation (DFG) through the priority program (SPP) 1253 'Optimization with Partial Differential Equations', the guest and travel funds of the international graduate college 710 'Complex Processes: Modeling, Simulation and Optimization' and the Heidelberg graduate school 'Mathematical and Computational Methods for the Sciences'.

Bibliography

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] Shmuel Agmon, Avron Douglis, and Louis Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Comm. Pure Appl. Math.*, 12(4):623–727, 1959.

[3] Mark Ainsworth and John Tinsley Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[4] Herbert Amann and Joachim Escher. *Analysis III*. Birkhäuser, second 2008.

[5] Nadir Arada, Eduardo Casas, and Fredi Tröltzsch. Error estimates for a semilinear elliptic control problem. *Comput. Optim. Appl.*, 23:201–229, 2002.

[6] Nadir Arada, Jean-Pierre Raymond, and Fredi Tröltzsch. On an augmented Lagrangian SQP method for a class of optimal control problems in banach spaces. *Comput. Optim. Appl.*, 22:369–398, 2002.

[7] Ivo Babuška and Theofanis Strouboulis. *The Finite Element Method and its Reliability*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 2001.

[8] Wolfgang Bangert and Rolf Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, Boston, Berlin, 1. edition, 2003. ISBN 3-7643-7009-2.

[9] Steffen Bauer. *Einfache Mehrgitterverfahren für Optimalsteuerungsprobleme elliptischer partieller Differentialgleichungen*. Diplomarbeit, Institut für Angewandte Mathematik, Universität Heidelberg, 2009.

[10] Roland Becker. Estimating the control error in discretized PDE-constraint optimization. *J. Numer. Math.*, 14(3):163–185, 2006.

[11] Roland Becker and Rolf 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.

[12] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation. In A. Iserles, editor, *Acta Numerica 2001*, pages 1–102. Cambridge University Press, 2001.

[13] Roland Becker and Boris Vexler. A posteriori error estimation for finite element discretization of parameter identification problems. *Numer. Math.*, 96:435–459, 2004.

[14] Roland Becker and Boris Vexler. Mesh refinement and numerical sensitivity analysis for parameter calibration of partial differential equations. *J. Comp. Physics*, 206(1): 95–110, 2005.

[15] Roland Becker, Hartmut Kapp, and Rolf Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concept. *SIAM J. Control Optim.*, 39(1):113–132, 2000.

[16] Olaf Benedix and Boris Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Comput. Optim. Appl.*, 44(1):3–25, 2009.

[17] Maïtine Bergounioux. A penalization method for optimal control of elliptic problems with state constraints. *SIAM J. Control Optim.*, 30(2):305–323, 1992.

[18] Maïtine Bergounioux. On boundary state constrained control problems. *Numer. Funct. Anal. Optim.*, 14(5&6):515–543, 1993.

[19] Maïtine Bergounioux and Karl Kunisch. Augmented Lagrangian techniques for elliptic state constrained optimal control problems. *SIAM J. Control Optim.*, 35(5):1524–1543, 1997.

[20] Maïtine Bergounioux and Karl Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Comp. Optim. Appl.*, 22:193–224, 2002.

[21] Maïtine Bergounioux and Karl Kunisch. On the structure of Lagrange multipliers for state-constrained optimal control problems. *Systems Control Lett.*, 48:169–176, 2003.

[22] Maïtine Bergounioux, Kazufumi Ito, and Karl Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37(4):1176–1194, 1999.

[23] Maïtine Bergounioux, Mounir Haddou, Michael Hintermüller, and Karl Kunisch. A comparison of interior point methods and a Moreau-Yosida based active set strategy for constrained optimal control problems. *SIAM J. Optim.*, 11(2):495–521, 2000.

[24] George Biros and Omar Ghattas. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. Part I: The Krylov-Schur solver. *SIAM J. Sci. Comput.*, 27(2):687–713, 2005.

[25] George Biros and Omar Ghattas. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. Part II: The Lagrange-Newton solver and its application to optimal control of steady viscous flows. *SIAM J. Sci. Comput.*, 27(2):714–739, 2005.

[26] Heribert Blum and Manfred Dobrowolski. On finite element methods for elliptic equations on domains with corners. *Computing*, 28:53–63, 1982.

[27] Heribert Blum and Rolf Rannacher. On the boundary value problem of the biharmonic operator on domains with angular corners. *Math. Meth. in the Appl. Sci.*, 2(2):556–581, 1980.

[28] Heribert Blum and Franz-Theo Suttmeier. An adaptive finite element discretization for a simplified signorini problem. *Calcolo*, 37(2):65–77, 2000.

[29] Heribert Blum and Franz-Theo Suttmeier. Weighted error estimates for finite element solutions of variational inequalities. *Computing*, 65(2):119–134, 2000.

[30] Jonathan Michael Borwein and Qiji J. Zhu. A survey of subdifferential calculus with applications. *Nonlinear Anal.*, 38(6):687–773, 1999.

[31] Alfio Borzi and Volker Schulz. Multigrid methods for pde optimization. *SIAM Review*, 51(2):361–395, 2009.

[32] Susanne C. Brenner and Larkin Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, New York, 3. edition, 2008.

[33] Carsten Carstensen and Rüdiger Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 36(5):1571–1587, 1999.

[34] Eduardo Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.

[35] Eduardo Casas and Joseph Frédéric Bonnans. Contrôle de systèmes elliptiques semilinéares comportant des contraintes sur l'état. In H. Brezzis and J.L. Lions, editors, *Nonlinear Partial Differential Equations and their Applications 8*, pages 69–86. Longman, New York, 1988.

[36] Eduardo Casas and Luis Alberto Fernández. Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state. *Appl. Math. Optim.*, 27:35–56, 1993.

[37] Eduardo Casas and Mariano Mateos. Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints. *SIAM J. Control Optim.*, 40(5):1431–1454, 2002.

[38] Eduardo Casas and Mariano Mateos. Uniform convergence of the FEM. Applications to state constrained control problems. *Comput. Appl. Math.*, 21(1):67–100, 2002.

[39] Eduardo Casas and Fredi Tröltzsch. Second-order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations. *Appl. Math. Optim.*, 39:211–227, 1999.

[40] Eduardo Casas and Fredi Tröltzsch. Error estimates for linear-quadratic elliptic control problems. In *Analysis and Optimization of Differential Systems (Constanta, 2002)*, pages 89–100. Kluwer Acad. Publ., Boston, MA, 2003.

[41] Eduardo Casas and Fredi Tröltzsch. First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations. *SIAM J. Control Optim.*, 48(2):688–718, 2009.

[42] Eduardo Casas, Fredi Tröltzsch, and Andreas Unger. Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations. *SIAM J. Control Optim.*, 38(5):1369–1391, 2000.

[43] Eduardo Casas, Mariano Mateos, and Fredi Tröltzsch. Error estimates for the numerical approximation of boundary semilinear elliptic control problems. *Comp. Optim. Appl.*, 31(2):193–220, 2005.

[44] Xiaojun Chen, Zuhair Nashed, and Liqun Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J, Numer. Anal.*, 38(4): 1200–1216, 2000.

[45] Svetlana Cherednichenko and Arnd Rösch. Error estimates for the regularization of optimal control problems with pointwise control and state constraints. *Z. Anal. Anwendungen*, 27(2):195–212, 2008.

[46] Philippe G. Ciarlet and Pierre-Arnaud Raviart. A mixed finite element method for the biharmonic equation. In *Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974)*, pages 125–145. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[47] Frank H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[48] James A. Clarkson. Uniformly convex spaces. *Trans. Amer. Math. Soc.*, 40(3):396–414, 1936.

[49] Bernard Dacorogna. *Direct Methods in the Calculus of Variations*, volume 78 of *Applied Mathematical Sciences*. Springer, second edition, 2008.

[50] Monique Dauge. *Elliptic Boundary Value Problems on Corner Domains*, volume 1341 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988.

[51] Monique Dauge. Neumann and mixed problems on curvelinear polyhedra. *Integral Equations Operator Theory*, 15(2):227–261, 1992.

[52] Klaus Deckelnick and Michael Hinze. Numerical analysis of a control and state constrained elliptic control problem with piecewise constant control approximations. In *Numerical Mathematics and Advanced Applications*, 2007.

[53] Klaus Deckelnick and Michael Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM J. Numer. Anal.*, 45(5):1937–1953, 2007.

[54] Klaus Deckelnick, Andreas Günther, and Michael Hinze. Finite element approximation of elliptic control problems with constraints on the gradient. *Numer. Math.*, 111: 335–350, 2008. doi: 10.1007/s00211-008-0185-3.

[55] Klaus Deckelnick, Andreas Günther, and Michael Hinze. Discrete concepts for elliptic optimal control problems with constraints on the gradient of the state. In *PAMM*, volume 8, pages 10873–10874. WILEY-VCH Verlag, 2008.

[56] Klaus Deckelnick, Andreas Günther, and Michael Hinze. Finite element approximation of dirichlet boundary control for elliptic PDEs on two- and three-dimensional curved domains. *SIAM J. Control Optim.*, 48(2):2798–2819, 2009.

[57] Jim Douglas, Jr., Todd Dupont, and Lars Wahlbin. The stability in L^q of the L^2-projection into finite element function spaces. *Numer. Math.*, 23:193–197, 1974/75. ISSN 0029-599X.

[58] Ivar Ekeland and Roger Témam. *Convex Analysis and Variational Problems*, volume 1 of *Studies in Mathematics and its Applications*. North-Holland Publishing Company, Amsterdam - Oxford, 1972.

[59] Ivar Ekeland and Roger Témam. *Convex Analysis and Variational Problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition, 1999.

[60] Kenneth Eriksson, Don Estep, Peter Hansbo, and Claes Johnson. Introduction to adaptive methods for differential equations. In A. Iserles, editor, *Acta Numerica 1995*, pages 105–158. Cambridge University Press, 1995.

[61] Kenneth Eriksson, Don Estep, Peter Hansbo, and Claes Johnson. *Computational Differential Equations*. Cambridge University Press, Cambridge, 1996.

[62] Richard S. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.*, 44:28–47, 1973.

[63] Andrei Vladimirovich Fursikov. *Optimal Control of Distributed Systems. Theory and Applications*, volume 187 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000. Translated from the 1999 Russian original by Tamara Rozhkovskaya.

[64] Alexandra Gaevskaya, Roland H. W. Hoppe, Yuri Iliash, and Michael Kieweg. Convergence analysis of an adaptive finite element method for distributed control problems with control constraints. In *Control of Coupled Partial Differential Equations*, International Series of Numerical Mathematics. Birkhäuser, 2007.

[65] Gascoigne. The finite element toolkit Gascoigne. URL `http://www.gascoigne.uni-hd.de`.

[66] Tunc 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.

[67] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Grundlehren der mathematischen Wissenschaften*. Springer, revised 3. edition, 2001.

[68] Seymour Goldberg. *Unbounded Linear Operators*. McGraw-Hill Book Company, 1966.

[69] Roland Griesse and Boris Vexler. Numerical sensitivity analysis for the quantity of interest in PDE-constrained optimization. *SIAM J. Sci. Comput.*, 29(1):22–48, 2007.

[70] Pierre Grisvard. Behavior of the solution of an elliptic boundary value problem in a polygonal or polyhedral domain. In *Numerical solution of partial differential equations, III (Proc. Third Sympos. (SYNSPADE), Univ. Maryland, College Park, Md., 1975)*, pages 207–274. Academic Press, 1976.

[71] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Monographs and studies in Mathematics. Pitman, Boston, 1. edition, 1985. ISBN 0-273-08647-2.

[72] Pierre Grisvard. Edge behavior of the solution of an elliptic problem. *Math. Nachr.*, 132:281–299, 1987.

[73] Pierre Grisvard. *Singularities in boundary value problems*, volume 22 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics]*. Masson, Paris, 1992.

[74] Pierre Grisvard. Singular behavior of elliptic problems in non hilbertian sobolev spaces. *J. Math. Pures Appl.*, 74:3–33, 1995.

[75] Andreas Günther and Michael Hinze. A-posteriori error control of a state constrained elliptic control problem. *J. Numer. Math.*, 16:307–322, 2008.

[76] Andreas Günther and Michael Hinze. Elliptic control problems with gradient constraints - variational discrete versus piecewise constant controls. Preprint SPP1253-08-07, DFG Priority Program 1253, 2009.

[77] Andreas Günther and Moulay Hicham Tber. A goal-oriented adaptive moreau-yosida algorithm for control- and state-constrained elliptic control problems. Preprint SPP1253-089, DFG Priority Program 1253, 2009.

[78] Roland Herzog and Ekkehard Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. Preprint SPP1253-088, DFG Priority Program 1253, 2009.

[79] Michael Hintermüller and Michael Hinze. Moreau-Yosida regularization in state constrained elliptic control problems: Error estimates and parameter adjustment. *SIAM J. Numer. Anal.*, 47(SPP1253-08-04):1666–1683, 2008. doi: 10.1137/080718735.

[80] Michael Hintermüller and Ronald H. W. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM J. Control Optim.*, 47(4):1721–1743, 2008.

[81] Michael Hintermüller and Karl Kunisch. Feasible and non-interior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45 (4):1198–1221, 2006.

[82] Michael Hintermüller and Karl Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, 17(1):159–187, 2006.

[83] Michael Hintermüller and Karl Kunisch. PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. *SIAM J. Optim.*, 20 (3):1133–1156, 2009.

[84] Michael Hintermüller and Wolfgang Ring. A level set approach for the solution of a state-constrained optimal control problem. *Numer. Math.*, 98:135–166, 2004.

[85] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth newton method. *SIAM J. Optim*, 13(3):865–888, 2003.

[86] Michael Hintermüller, Ronald H. W. Hoppe, Yuri Iliash, and Michael Kieweg. An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. *ESIAM Control Optim. Calc. Var.*, 14: 540–560, 2008.

[87] Michael Hintermüller, Fredi Tröltzsch, and Irwin Yousept. Mesh-independence of semismooth newton methods for Lavrentiev-regularized state constrained nonlinear optimal control problems. *Numer. Math.*, 108:571–603, 2008.

[88] Michael Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comp. Optim. Appl.*, 30(1):45–61, 2005.

[89] Michael Hinze and Anton 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.

[90] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.

[91] Ronald H. W. Hoppe and Michael Kieweg. A posteriori error estimation of finite element approximations of pointwise state constrained distributed control problems. *J. Numer. Math.*, 17(3):219–244, 2009.

[92] Ronald H. W. Hoppe, Yuri Iliash, Chakradhar Iyyunni, and Nasser H. Sweilam. A posteriori error estimates for adaptive finite element discretizations of boundary control problems. *J. Numer. Math.*, 14(1):57–82, 2006.

[93] Kazufumi Ito and Karl Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[94] Hartmut Kapp. *Adaptive Finite Elements for Optimization in Partial Differential Equations*. PhD thesis, Mathematisch-Naturwissenschaftilche Gesamtfakultät, Universität Heidelberg, 2000.

[95] David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and their Applications*. Classics in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1. edition, 2000.

[96] Axel Kröner. *A Priori Error Estimations for Finite Element Discretization of an Elliptic Optimal Control Problem with a Bilinear State Equation*. Diplomarbeit, Institut für Angewandte Mathematik, Universität Heidelberg, 2007.

[97] Axel Kröner and Boris Vexler. A priori error estimates for elliptic optimal control problems with bilinear state equation. *J. Comput. Appl. Math.*, 230(2):781–802, 2009.

[98] Karl Kunisch and Arnd Rösch. Primal-dual active set strategy for a general class of constrained optimal control problems. *SIAM J. Optim.*, 13(2):321–334, 2002.

[99] Ruo Li, Wenbin Liu, Heping Ma, and Tao Tang. Adaptive finite element approximation for distributed elliptic optimal control problems. *SIAM J. Control Optim.*, 41(5): 1321–1349, 2002.

[100] Jacques-Louis Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Die Grundlehren der mathematischen Wissenschaften. Springer, Berlin – Heidelberg – New York, 1. edition, 1971. ISBN 3-540-05115-5.

[101] Jacques-Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications. Vol. I.* Number 181 in Die Grundlehren der mathematischen Wissenschaften. Springer-Verlag, New York, 1972. Translated from the French by P. Kenneth.

[102] Wenbin Liu and Ningning Yan. A posteriori error estimates for distributed convex optimal control problems. *Adv. Comput. Math.*, 15(1):285–309, 2001.

[103] Freerk Auke Lootsma. Hessian matrices of penalty functions for solving constrained-optimization problems. *Philips Res. Repts.*, 24:322–330, 1969.

[104] Kazimierz Malanowski. Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems. *Appl. Math. Optim.*, 8(1):69–95, 1982.

[105] Helmut Mäurer and Jochem Zowe. First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems. *Math. Program.*, 16:98–110, 1979.

[106] Sandra May, Rolf Rannacher, and Boris Vexler. Error analysis for a finite element approximation of elliptic dirichlet boundary control problems. In K. Kunisch et al., editor, *Proc. ENUMATH-2007, Graz*, pages 637–644. Springer, 2008.

[107] Dominik Meidner. *Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems*. PhD thesis, Mathematisch-Naturwissenschaftilche Gesamtfakultät, Universität Heidelberg, 2007.

[108] Dominik Meidner and Boris Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):116–142, 2007.

[109] Dominik Meidner, Rolf Rannacher, and Jevgeni Vihharev. Goal-oriented error control of the iterative solution of finite element equations. *J. Numer. Math.*, 17:143–172, 2009.

[110] Pedro Merino, Fredi Tröltzsch, and Boris Vexler. Error estimates for the finite element approximation of a semilinear elliptic control problem with state constraints and finite dimensional control space. Preprint SPP1253-24-04, DFG Priority Program 1253, 2008.

[111] Christian Meyer. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. *Control Cybernet.*, 37: 51–85, 2008.

[112] Christian Meyer and Arnd Rösch. Superconvergence properties of optimal control problems. *SIAM J. Control Optim.*, 43(3):970–985, 2004.

[113] Christian Meyer and Arnd Rösch. L^∞-estimates for approximated optimal control problems. *SIAM J. Control Optim.*, 44(5):1636–1649, 2005.

[114] Christian Meyer and Fredi Tröltzsch. On an elliptic optimal control problem with pointwise mixed control-state constraints. *Lecture Notes in Econom. and Math. Systems*, 563:187–204, 2006.

[115] Christian Meyer, Arnd Rösch, and Fredi Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Comput. Optim. Appl.*, 33:209–228, 2006.

[116] Christian Meyer, Uwe Prüfert, and Fredi Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optim. Meth. Software*, 22:871–899, 2007.

[117] Jacqueline Mossino. An application of duality to distributed optimal control problems with constraints on the control and the state. *J. Math. Anal. Appl.*, 50:223–242, 1975.

[118] Walter Murray. Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions. *J. Optim. Theory Appl.*, 7(3): 189–196, 1971.

[119] Jorge Nocedal and Stephen Joseph Wright. *Numerical Optimization*. Springer, 1999.

[120] Christoph Ortner and Winnifried Wollner. A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state. Preprint SPP1253-071, DFG Priority Program 1253, 2009.

[121] Rolf Rannacher and Ridgway Scott. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comp.*, 38(158):437–445, 1982.

[122] Tyrone Rees, H. Sue Dollar, and Andrew J. Wathen. Optimal solvers for pde-constrained optimization. *SIAM J. Sci. Comput.*, 32(1):271–298, 2010.

[123] Uwe Reichmann. *Das Ciarlet-Raviart-Verfahren zur Lösung der biharmonischen Gleichung*. Diplomarbeit, Institut für Angewandte Mathematik, Universität Heidelberg, 1993.

[124] Ralph Tyrell Rockafellar. Directionally lipschitzian functions and subdifferential calculus. *Proc. London Math. Soc.*, 39(3):331–355, 1979.

[125] Ralph Tyrrell Rockafellar. Generalized directional derivatives and subgradients of nonconvex functions. *Can. J. Math.*, 32(2):257–280, 1980.

[126] RoDoBo. RoDoBo: A C++ library for optimization with stationary and nonstationary PDEs. URL `http://www.rodobo.uni-hd.de`.

[127] Arnd Rösch. Error estimates for linear-quadratic control problems with control constraints. *Optim. Methods Softw.*, 21(1):121–134, 2006.

[128] Arnd Rösch and Boris Vexler. Optimal control of the stokes equations: A priori error analysis for finite element discretization with postprocessing. *SIAM J. Numer. Anal.*, 44(5):1903–1020, 2006.

[129] Anton Schiela. Convergence of the control reduced interior point method for PDE constrained optimal control with state constraints. Preprint 06-16, ZIB Report, 2006.

[130] Anton Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.*, 20(2):1002–1031, 2009.

[131] Anton Schiela. Optimality conditions for convex state constrained optimal control problems with discontinuous states. Preprint 07-35, ZIB Report, 2007.

[132] Anton Schiela. An interior point method in function space for the efficient solution of state constrained optimal control problems. Preprint 07-44, ZIB Report, 2008.

[133] Anton Schiela and Martin Weiser. Function space interior point methods for PDE constrained optimization. In *PAMM*, volume 4, pages 43–46, 2004.

[134] Anton Schiela and Martin Weiser. Superlinear convergence of the control reduced interior point method for PDE constrained optimization. *Comput. Optim. Appl.*, 39(3): 369–393, 2008.

[135] Anton Schiela and Winnifried Wollner. Barrier methods for optimal control problems with convex nonlinear gradient state constraints. Preprint SPP1253-23-03, DFG Priority Program 1253, 2008.

[136] Michael Schmich. *Adaptive Finite Element Methods for Compution Nonstationary Incompressible Flow*. PhD thesis, Mathematisch-Naturwissenschaftilche Gesamtfakultät, Universität Heidelberg, 2009.

[137] Michael Schmich and Boris Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM J. Sci. Comput.*, 30(1): 369–393, 2008.

[138] Joachim Schöberl and Walter Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773, 2007.

[139] Franz-Theo Suttmeier. *Adaptive Finite Element Approximation of Problems in Elasto-Plasticity Theory*. PhD thesis, Mathematisch-Naturwissenschaftilche Gesamtfakultät, Universität Heidelberg, 1996.

[140] Hans Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. Johann Ambrosius Barth Verlag; Heidelberg, Leipzig, 2., rev. and enl. edition, 1995.

[141] Fredi Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen*. Vieweg, 1. edition, 2005. ISBN 3-528-03224-3.

[142] Fredi Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM J. Optim.*, 15(2):616–634, 2005.

[143] Fredi Tröltzsch and Irvin Yousept. A regularization method for the numerical solution of elliptic boundary control problems with pointwise state constraints. *Comput. Optim. Appl.*, 42(1):43–66, 2009.

[144] Michael Ulbrich. Semismooth newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–841, 2003.

[145] Michael Ulbrich and Stefan Ulbrich. Primal-dual interior point methods for PDE-constrained optimization. *Math. Program.*, 117(1-2):435–485, 2009.

[146] Rüdiger Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley/Teubner, New York-Stuttgart, 1996.

[147] Boris Vexler. *Adaptive Finite Element Methods for Parameter Identification Problems*. PhD thesis, Mathematisch-Naturwissenschaftilche Gesamtfakultät, Universität Heidelberg, 2004.

[148] Boris Vexler and Winnifried Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.*, 47(1):509–534, 2008.

[149] Visit. VISIT: A free interactive parallel visualization and graphical analysis tool. URL `https://wci.llnl.gov/codes/visit/`.

[150] Visusimple. VISUSIMPLE: A VTK-based visualization software. URL `http://visusimple.uni-hd.de/`.

[151] Martin Weiser. Interior point methods in function space. *SIAM J. Control Optim.*, 44 (5):1766–1786, 2005.

[152] Martin Weiser and Peter Deuflhard. Inexact central path following algorithms for optimal control problems. *SIAM J. Control Optim.*, 46(3):792–815, 2007.

[153] Martin Weiser, Tobias Gänzler, and Anton Schiela. A control reduced primal interior point method for PDE constrained optimization. *Comput. Optim. Appl.*, 41(1):127–145, 2008.

[154] Dirk Werner. *Funktionalanalysis*. Springer, Berlin – Heidelberg – New York, 5., erw. edition, 2005. ISBN 3-540-21381-3.

[155] Joseph Wloka. *Partielle Differentialgleichungen. Sobolewräume und Randwertaufgabe*. Teubner, Leipzig, 1. edition, 1982.

[156] W. Wollner. Adaptive finite element methods for optimal control problems with control constraints. In *Oberwolfach Rep.*, volume 4, pages 1728–1731. Eur. Math. Soc. EMS Publ. House, 2007. Abstracts from the workshop held June 10–16, 2007, Organized by Rolf Rannacher, Endre Süli and Rüdiger Verfürth.

[157] Winnifried 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. doi: 10.1007/s10589-008-9209-2.

[158] Winnifried Wollner. Adaptive FEM for PDE constrained optimization with pointwise constraints on the gradient of the state. In *PAMM*, volume 8, pages 10873–10874. WILEY-VCH Verlag, 2008. doi: 10.1002/pamm.200810873.

[159] Jochem Zowe and Stanislaw Kurcyusz. Regularity and stability for the mathematical programming problem in banach spaces. *Appl. Math. Optim.*, 5:49–62, 1979.