Optimization with partial differential equations

Vorlesungsskriptum im SoSe 2016 — Stand March 1, 2021

Prof. Dr. Winnifried Wollner
e-mail: wollner@mathematik.tu-darmstadt.de
web: www2.mathematik.tu-darmstadt.de/~wollner

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Mathematik

Preface

These notes accompany the lecture "Optimization with partial differential equations" in Summer 2021, 2016 at TU Darmstadt. They are based upon my lectures on "Optimization of Complex Systems" held at University of Hamburg. They are made available to participants for their private educational use. Please do not distribute them any further.

Be advised that these notes can not replace your own notes taken during the lecture, because neither will these notes contain any sketches nor will all comments from the lecture be included. In addition, these notes may differ from the lecture at some times.

The author gratefully acknowledges the help of Roland Herzog, Christian Meyer and Anton Schiela who made their own lecture notes on PDE-constrained optimization available to me.

This lecture will contain content of the three scripts [18, 25, 32] as well as the books [22, 23, 36].

If you wish to broaden some of the topics of this lecture here are some helpful references (although there may be many more)

Optimization with PDE constraints [22, 23, 36].

Finite Elements [3, 4, 8, 35].

PDE [16, 41].

Optimization [28, 39].

Further reading Calculus of Variations [11]; Sobolev-spaces [1]; Functional analysis [30].

You may send comments and corrections to

wollner@mathematik.tu-darmstadt.de

I will gratefully incorporate these into the next version of the script

Winnifried Wollner

Darmstadt, March 2021.

Contents

1 Introduction

Introduction

In this lecture, we are interested in the analysis of optimization problems governed by partial differential equations (PDEs). That is, we consider a Banach space X and a functional $J : X \to \mathbb{R}$ which we aim to minimize. In addition, we have some constraints on the variables, i.e., there is some subset X^{ad} of admissible values. Altogether we aim at solving

$$\begin{aligned} J(x) &\to \min \\ \text{s.t. } x &\in X^{\mathrm{ad}}. \end{aligned}$$

Sometimes, the set of variables can be separated into a set of 'control' or input variables Q and a set of 'state' variables U. Then we can split $X = Q \times U$. The feasible set can be decomposed accordingly. First there will be some relation between the control and state variable. Thereto, we assume to have an operator

$$A \colon Q \times U \to Z^*$$

with some reflexive Banach space Z^*. Let the relation between $q \in Q$ and $u \in U$ be given such that

$$(q,u) \in X^{\mathrm{ad}} \to A(q,u) = 0.$$

We may encounter additional constraints, to incorporate these, we assume that $Q^{\mathrm{ad}} \subset Q$ is a closed and convex subset. Further, we have a Banach space K and mapping $g : Q \times U \to K$ with a closed convex cone $K^{\mathrm{ad}} \subset K$ such that we can define the set

$$X^{\mathrm{ad}} = \{(q,u) \in Q \times U \,|\, q \in Q^{\mathrm{ad}},\, g(q,u) \in K^{\mathrm{ad}},\, A(q,u) = 0\}.$$

(Sometimes one also writes $g(q,u) \geq_{K^{\mathrm{ad}}} 0$ or $g(q,u) \geq 0$ to denote the inclusion $g(q,u) \in K^{\mathrm{ad}}$.) Then our problem can be restated as

$$\begin{aligned} J(q,u) &\to \min \\ \text{s.t. } q &\in Q^{\mathrm{ad}}, \\ A(q,u) &= 0, \\ g(q,u) &\in K^{\mathrm{ad}}. \end{aligned}$$

Before we proceed with the abstract notation, we will start with some motivating examples.

1.1 Motivation

1.1.1 Optimal Heating

A prototypical example for optimization processes with partial differential equations would be the heating/cooling of an object.

The (stationary) temperature distribution of a body Ω with a heat source q on the boundary Γ is given by the following simplified relation **draw a sketch**

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega, \\ \partial_n u &= \gamma(q-u) && \text{on } \Gamma, \end{aligned} \tag{1.1}$$

with some parameter $\gamma > 0$.

Let us assume, we know a desired temperature profile u^d—This is sometimes a reasonable assumption, for instance hardening of steel depends heavily on the temperature distribution—then one could aim to minimize the distance of the temperature u to the desired profile using

$$\frac{1}{2}\int_\Omega (u(x)-u^d(x))^2\,\mathrm{d}x = \frac{1}{2}\|u-u^d\|^2.$$

Unfortunately, this is usually not a well-posed problem (no minimizers exist or they don't depend continuously on the problem data). To resolve this issue, several possibilities exist. One may add an additional term to the cost functional, e.g., a Tikhonov regularization,

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

with some parameter $\alpha > 0$. This is sometimes justified as so called 'control-costs'.

Additionally, some constraints may be imposed, e.g., the control on the boundary may not exceed certain bounds $q_{\min} < q_{\max}$, yielding so called 'box-constraints'

$$Q^{\mathrm{ad}} := \{q \,|\, q_{\min} \le q(x) \le q_{\max} \text{ on } \Gamma\}.$$

Further, state constraints may be required, e.g., if the temperature has to remain in certain regions $u_{\min} < u_{\max}$, for example to prevent melting/freezing. Then one has the additional constraint

$$u_{\min} \le u \le u_{\max}$$

or equivalently $g(q,u) = (u-u_{\min}, u_{\max}-u) \ge 0$ in $C(\overline{\Omega})^2$. Finally, it may also be important to avoid large gradients of the temperature because this would induce large stresses and may destroy the object. Hence, additional constraints of the type

$$|\nabla u| \le u_{\mathrm{grad}}$$

with some number $u_{\mathrm{grad}} > 0$ may be necessary.

Of course in real applications, things may get more involved, e.g., in order to model time dependent processes, it is usually necessary to consider nonstationary equations. In our case this would give

$$\begin{aligned} \partial_t u - \Delta u &= 0 && \text{in } \Omega, \\ \partial_n u &= \gamma(q-u) && \text{on } \Gamma. \end{aligned}$$

In addition, in front of the Laplace operator a coefficient may occur yielding

$$\begin{aligned} \partial_t u - \nabla \cdot (a \nabla u) &= 0 && \text{in } \Omega, \\ \partial_n u &= \gamma(q-u) && \text{on } \Gamma. \end{aligned}$$

with a coefficient depending possibly on $x \in \Omega$ as well as on the temperature u yielding a quasilinear equation. A typical example for this would be living tissue in planing of medical procedures (e.g., hypothermia), which changes its properties drastically with changing temperature.

Of course different heating methods yield different conditions. For example, if heating is done by radiative transfer, the boundary condition changes to the Stefan-Boltzmann condition, e.g.,

$$\partial_n u = \gamma(q^4 - u^4) \quad \text{on } \Gamma.$$

1.1.2 Variable Thickness Sheets

A second possible goal is the consideration of controls acting not via the right-hand side; but in the coefficients of the problem.

A typical example is that of a variable thickness sheet. Let $\Gamma \supset \Gamma_D \cup \Gamma_T$ with both $|\Gamma_D|, |\Gamma_T| \neq 0$. Then, given a thickness distribution $0 < \rho_{\min} \leq \rho \leq \rho_{\max}$ of the sheet, the displacement **draw sketch**

$$u \in H^1_D(\Omega;\mathbb{R}^2) = \{v : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2 \,|\, \nabla v \in L^2(\Omega;\mathbb{R}^{2\times 2}),\ v \in L^2(\Omega;\mathbb{R}^2),\ v = 0 \text{ on } \Gamma_D\}$$

of the sheet under given loads $t \in L^2(\Gamma_T;\mathbb{R}^2)$ is given as the unique solution to the problem

$$(\rho \nabla u, \nabla \varphi) = (t, \varphi)_{\Gamma_T} \quad \forall \varphi \in H^1_D(\Omega;\mathbb{R}^2).$$

The minimum compliance problem for the variable thickness sheet is given as

$$\begin{aligned} \min_{\rho}\ & (t,u) \\ & \text{s.t } (\rho \nabla u, \nabla \varphi) = (t,\varphi)_{\Gamma_T} \quad \forall \varphi \in H^1_D(\Omega;\mathbb{R}^2), \\ & \rho_{\min} \leq \rho \leq \rho_{\max}, \\ & (1,\rho) \leq V_{\max}, \end{aligned}$$

with given bounds $V_{\max}, \rho_{\min}, \rho_{\max} \geq 0$.

1.2 What we know from finite dimensional optimization

And what is different for PDE-constrained optimization?

To give a simple example, we consider a linear-quadratic problem in a finite-dimensional setting. Let $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a given as

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

where $\alpha \geq 0$, $u^d \in \mathbb{R}^m$, and $\|\cdot\|$ denotes the euclidean vector norm. Further, let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times n}$ be given matrices. Assume A to be invertible. Then, we consider the problem

$$\begin{aligned} &\min J(q,u) \\ &\text{s.t. } Au = Bq, \\ &\quad q \in Q^{\text{ad}} = \{q \in \mathbb{R}^n \,|\, q_{\min} \leq q_i \leq q_{\max}\}, \end{aligned}$$

with some numbers $q_{\min} < q_{\max} \in \mathbb{R} \cup \{\pm\infty\}$.

With these preparations, it is clear that, we can define the solution operator

$$S = A^{-1}B$$

and eliminate the state variable u from the optimization problem. Then with the reduced cost functional $j(q) := J(q, Sq)$ we can equivalently solve

$$\begin{aligned} &\min j(q) \\ &\text{s.t. } q \in Q^{\text{ad}}. \end{aligned} \tag{1.2}$$

Definition 1.2.1. $\overline{q} \in Q^{\text{ad}}$ is called a local solution of (1.2) if there exists a neighborhood $\mathcal{Q}$ of $\overline{q}$ such that

$$j(\overline{q}) \leq j(q) \quad \forall q \in Q^{\text{ad}} \cap \mathcal{Q}.$$

It is called a global solution of (1.2) if this holds with $\mathcal{Q} = \mathbb{R}^n$.

Remark 1.2.2. Formally, this looks a lot like the problem in Section 1.1.1. But first of all in the PDE-context the functions u and q will, in general, not be defined in some finite dimensional space, but rather in some infinite dimensional space. We will see, in the sequel, that this will have some influence on the analysis.

However, in order to (approximately) solve these problems, we will have to discretize to end up with something finite dimensional; which we can then solve (approximately) with a computer. After discretization of the PDE the problems will look exactly the same as they are finite dimensional, but with some important aspects.

- m is usually (very) large, e.g., $m = 100\,000$ or more.
- A is usually a sparse matrix, but has condition increasing with m, e.g., for the example in Section 1.1.1 and $\Omega \subset \mathbb{R}^2$ we have for reasonable discretizations $\text{cond}(A) \approx m$.

- S may be smaller ($(m \times n)$ depending on the controls) but will be a dense matrix! For example, in Section 1.1.1 and $\Omega \subset \mathbb{R}^2$ it is reasonable to assume $n \approx \sqrt{m}$, then there are $O(m^{3/2})$ entries to store (and compute) for S. In contrast to the storage of $O(m)$ entries in the sparse matrices A and B.
- There is a lot of structure in the problem that can (and should) be used.

We will pause here to give a short outlook on what we will need to do in the following weeks.

- Analyze optimization problems in function spaces.
- Derive necessary optimality conditions.
- Analyze the discretization of such problems (convergence of the discrete solutions to the continuous limit).
- Design solution algorithms.

1.2.1 Existence of solutions

We will show existence of solutions in a little more abstract setting for (1.2) without assuming the special form chosen there.

Theorem 1.2.3. *Let j be continuous on $Q^{ad} \subset \mathbb{R}^n$. Further let Q^{ad} be nonempty and closed. Then if either*

1. *Q^{ad} is bounded or*
2. *There exists some $q_0 \in Q^{ad}$, and $r > 0$ such that $j(q) \geq j(q_0)$ for all $\|q - q_0\|_{l^2} > r$.*

there exists at least one global solution to (1.2).

Proof. In case 1: We have that Q^{ad} is compact by the theorem of Bolzano-Weierstrass. Hence the continuous function j attains its minimum on Q^{ad} due to the extreme value theorem.

In case 2: It is clearly sufficient to search for a minimizer on the ball $\overline{B_r(q_0)} = \{q \in \mathbb{R}^n \,|\, \|q - q_0\| \leq r\}$. Then again the set $Q^{\text{ad}} \cap \overline{B_r(q_0)}$ is compact. □

Remark 1.2.4. (Counterexample) We remark here, that the above proof cannot be used in the context of PDE-constrained optimization, because, in general, the set Q^{ad} will be contained in an infinite dimensional space. It is known from functional analysis that a Banach space X is finite dimensional if and only if the set $\{x \in X \,|\, \|x\| \leq 1\}$ is compact.

As a prototypical example what may happen in the infinite dimensional case, consider the problem

$$\min_{u\in C[-1,1]} J(u) := \int_{-1}^{1} (u(x)-u^d(x))^2\,\mathrm{d}x = \|u-u^d\|^2_{L^2(-1,1)} = \|u-u^d\|^2$$
$$\text{s.t.} -1 \le u \le 1.$$

with

$$u^d(x) = \begin{cases} -1 & x < 0, \\ 1 & x \ge 0. \end{cases}$$

Clearly $u \mapsto J(u) \ge 0$ is continuous and the set $\{u \in C[-1,1] \,|\, -1 \le u \le 1\}$ is bounded. But the sequence

$$u_k(x) = \begin{cases} -1 & x < -1/k, \\ kx & -1/k \le x \le 1/k, \\ 1 & x \ge 1/k, \end{cases}$$

satisfies

$$J(u_k) = \int_{-1/k}^{1/k} (1-k|x|)^2\,\mathrm{d}x \le \int_{-1/k}^{1/k} \mathrm{d}x \to 0 \quad (k\to\infty).$$

Hence it holds

$$\inf_{u\in C[-1,1]} J(u) = 0$$

but $J(\overline{u}) = 0$ if and only if $\overline{u} = u^d \notin C[-1,1]$.

In particular, we will need some additional tools for the analysis of such problems. We can already see that if instead of functions $u \in C[-1,1]$ we would have taken functions in $L^2[-1,1]$ (functions that are square integrable) the counterexample would have failed.

Hence, we may expect that the right choice of the space where minimizers are searched for is crucial.

1.2.2 Lavrentiev Phenomenon

It should be noted that, in general, not only the existence of a minimizer is depending on the choice of the spaces. Even the infimum values of the functional may depend on the chosen spaces. This is obvious, of course, since on any one-dimensional subspace not containing a minimizing sequence for the functional this is true. But infact, such defects can be a lot more subtle as it is demonstrated by the so called Lavrentiev phenomenon, [24] which stated that certain minimal functional values can not be approximated by Lipschitz continuous functions.

We illustrate this with the following example. Consider the functional

$$J(u) = \int_0^1 (u(t)^3 - t)^2 u'(t)^6\,dt$$

then it holds (with the boundary conditions $u(0)=0$, $u(1)=1$)

$$\min_{u\in C^1(0,1)} J(u) = 0 < \inf_{u\in C^{0,1}(0,1)} J(u).$$

To see this, we note that $\sqrt[3]{t} \in C^1(0,1)$ but not in $C^{0,1}(0,1)$ and $J(\sqrt[3]{\cdot}) = 0$. Now, let $u \in C^{0,1}(0,1)$ be arbitrary and let $v = \sqrt[3]{t}/2$. Then there exists some $t_0 \in (0,1)$ such that

$$u(t) \le v(t) \quad \forall t \in [0,t_0]$$

and $u(t_0) = v(t_0)$ due to the regularity (and boundary values) of u and v. **draw sketch** Hence for $t \in [0,t_0]$ and $\xi \in \mathbb{R}$ it holds ($|u^3 - t| \ge |v^3 - t|$)

$$(u^3(t)-t)^2\xi^6 \ge (v^3(t)-t)^2\xi^6 = \left(\frac{\sqrt[3]{t^3}}{2^3} - t\right)^2 \xi^6 = \frac{7^2}{8^2} t^2 \xi^6.$$

Further, by fundamental theorem of calculus and Hölder inequality, we get

$$\begin{aligned}
\frac{\sqrt[3]{t_0}}{2} = v(t_0) = u(t_0) &= \int_0^{t_0} u'(t)\,dt = \int_0^{t_0} t^{-1/3}(t^{1/3}u'(t))\,dt \\
&\le \Big(\int_0^{t_0} t^{-2/5}\,dt\Big)^{5/6}\Big(\int_0^{t_0} t^2 u'(t)^6\,dt\Big)^{1/6} \\
&\le \Big(\frac{5}{3}t^{3/5}\Big|_0^{t_0}\Big)^{5/6}\Big(\int_0^{t_0} t^2 u'(t)^6\,dt\Big)^{1/6} \\
&= \left(\frac{5}{3}\right)^{5/6} t_0^{1/2}\Big(\int_0^{t_0} t^2 u'(t)^6\,dt\Big)^{1/6}.
\end{aligned}$$

Hence we obtain

$$\int_0^{t_0} t^2 u'(t)^6\,dt \ge \left(\frac{t_0^{1/3}}{2} t_0^{-1/2}\left(\frac{3}{5}\right)^{5/6}\right)^6 = \frac{1}{2^6 t_0}\left(\frac{3}{5}\right)^5.$$

Combining this we conclude that for any $u \in C^{0,1}(0,1)$ it holds

$$\begin{aligned}
J(u) &\ge \int_0^{t_0} (u(t)^3 - t)^2 u'(t)^6\,dt \\
&\ge \int_0^{t_0} (v(t)^3 - t)^2 u'(t)^6\,dt \\
&\ge \frac{7^2}{8^2}\int_0^{t_0} t^2 u'(t)^6\,dt \\
&\ge \frac{7^2}{8^2}\frac{3^5}{5^5}\frac{1}{2^6 t_0} \ge \frac{7^2}{8^2}\frac{3^5}{5^5}\frac{1}{2^6} > 0.
\end{aligned}$$

Which shows the claim.

This shows in particular, that if we would like to approximate the minimizers of the above functional in $C^1(0,1)$ then standard (conforming) finite elements will not be enough. This motivates why we want to consider convergence of discrete solutions in this lecture.

1.2.3 Necessary optimality conditions

We continue with the analysis of (1.2). To this end, we define the directional derivative of j in a point q and direction δq as

$$j'(q;dq) = \lim_{t \downarrow 0} \frac{j(q+t\delta q) - j(q)}{t}.$$

Theorem 1.2.5. *Let $\overline{q} \in Q^{ad}$ be a local solution to (1.2) with j differentiable and Q^{ad} convex. Then it holds*

$$j'(\overline{q}; q - \overline{q}) \geq 0 \quad \forall q \in Q^{ad}. \tag{1.3}$$

Proof. The assertion follows considering difference quotients. □

We will now rewrite this for our given functional $j(\overline{q}) = \frac{1}{2}\|\overline{u} - u^d\|^2 + \frac{\alpha}{2}\|\overline{q}\|^2$ where we define $\overline{u} = S\overline{q}$. Then it holds

$$j'(\overline{q}; \delta q) = (S^*(S\overline{q} - u^d) + \alpha\overline{q}, \delta q).$$

(S^* is the adjoint matrix to S) In particular, we see that $j'(\overline{q}; \cdot)$ defines a linear operator $j'(\overline{q}; \cdot) \colon \mathbb{R}^n \to \mathbb{R}$, hence one sometimes writes $j'(\overline{q})\delta q$ instead of $j'(\overline{q}; \delta q)$. Further, by choice of the l^2 scalar product, we can define the gradient $\nabla j(\overline{q}) \in \mathbb{R}^n$ of j by

$$(\nabla j(\overline{q}), \delta q) = j'(\overline{q})\delta q \quad \forall \delta q \in \mathbb{R}^n$$

and hence $\nabla j(\overline{q}) = S^*(S\overline{q} - u^d) + \alpha\overline{q}$. Thus the variational inequality (1.3) becomes

$$(S^*(\overline{u} - u^d) + \alpha\overline{q}, q - \overline{q}) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}$$

or for a general functional J

$$(S^*\nabla_u J(\overline{q}, \overline{u}) + \nabla_q J(\overline{q}, \overline{u}), q - \overline{q}) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}$$

by the chain rule (Exercise).

Remark 1.2.6. Note that the derivative $j'(\overline{q})$ is a linear operator $j'(\overline{q}) \colon \mathbb{R}^n \to \mathbb{R}$ (hence $j'(\overline{q}) \in \mathbb{R}^{1\times n}$ is a row vector after choosing a basis) while $\nabla j(\overline{q})$ is a (column) vector in $\mathbb{R}^n$ and depends on on the chosen scalar product.

Remark 1.2.7. We note that the choice of the scalar product in order to define the gradient ∇j gives some freedom. In the finite dimensional case, however, all scalar products induce the same topology, meaning that if $\nabla j \to 0$ in any scalar product, then this holds for all possible scalar products. In the infinite dimensional case this is no longer true. In particular, a sequence $\nabla j(q_n)$ may converge to zero using one scalar product, while in another the sequence is bounded away from zero or is even diverging. This is why we need to carefully choose the topology in the later chapters.

As a prototypical example consider the sequence of functions u_n on $[0,1]$ defined by

$$u_n(x) = \begin{cases} 1 - nx & x < \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

where it holds

$$\begin{aligned} \|u_n\|_1 &= \int_0^1 |u_n(x)|\,\mathrm{d}x \to 0, \\ \|u_n\|_\infty &= \operatorname{ess\,sup}_{x\in[0,1]} |u_n(x)| = 1, \\ \|u_n\|_{1,\infty} &= \operatorname{ess\,sup}_{x\in[0,1]} \Big(|u_n'(x)| + |u_n(x)| \Big) \to \infty. \end{aligned}$$

Adjoint State

We will now have a closer look at the first term, namely

$$\overline{z} = A^{*-1} \nabla_u J(\overline{q}, \overline{u}).$$

We call z the adjoint variable (or adjoint state) which is given by the adjoint equation

$$A^* \overline{z} = \nabla_u J(\overline{q}, \overline{u}).$$

Then the necessary optimality condition (1.3) becomes

$$(B^* \overline{z} + \nabla_q J(\overline{q}, \overline{u}), q - \overline{q}) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}$$

or

$$(B^* \overline{z} + \alpha \overline{q}, q - \overline{q}) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}$$

with the definition

$$A\overline{u} = B\overline{q}, \quad A^* \overline{z} = \nabla_u J(\overline{q}, \overline{u}).$$

Remark 1.2.8. We remark, that this is exactly what one would obtain by an application of the Lagrange calculus to the Lagrange function

$$\mathscr{L}(q, u, z) = J(q, u) - (Au - Bq, z).$$

In particular, the adjoint state is the Lagrange multiplier for the equality constraint $Au = Bq$.

Remark 1.2.9. We know from linear algebra (and functional analysis) that an operator $A: X \to Y$ between two vector spaces defines an adjoint operator $A^*: Y^* \to X^*$. Now if X or Y are infinite dimensional in general $X^* \not\cong X$ or $Y^* \not\cong Y$. Typical examples are $L^p(0,1)$ of functions whose p-th power is integrable whose dual space can be identified with $L^{p'}(0,1)$ for any $1 \leq p < \infty$ and $\frac{1}{p} + \frac{1}{p'} = 1$. Another typical example is the space $C[0,1]$ where we can identify the dual space with the space $M[0,1]$ of regular Borel measures (e.g., Dirac measure).

Again, this demonstrates how vital the choice of the topology may be in the infinite dimensional setting.

The Variational Inequality

Finally, we want to discuss the variational inequality (1.3) for the special (but common) case that the admissible set is given by box constraints, e.g., $Q^{\text{ad}} = \{q \in \mathbb{R}^n \,|\, q_{\min} \le q_i \le q_{\max}\}$. Then with our preparations we can rewrite (1.3) as

$$(B^*\overline{z} + \alpha\overline{q}, \overline{q}) \le (B^*\overline{z} + \alpha\overline{q}, q) \quad \forall q \in Q^{\text{ad}}.$$

This means that $\overline{q}$ solves the following minimization problem

$$\min_{q \in Q^{\text{ad}}} (B^*\overline{z} + \alpha\overline{q}, q) = \min_{q \in Q^{\text{ad}}} \sum_{i=1}^{n} \left(B^*\overline{z} + \alpha\overline{q}\right)_i q_i.$$

Now, all summands are independent (i.e., the function is separable), and we can determine q_i by minimizing the i-th summand. This gives

$$\overline{q}_i = \begin{cases} q_{\max} & \left(B^*\overline{z} + \alpha\overline{q}\right)_i < 0, \\ q_{\min} & \left(B^*\overline{z} + \alpha\overline{q}\right)_i > 0, \end{cases}$$

in the case $\left(B^*\overline{z} + \alpha\overline{q}\right)_i = 0$ we can solve this equation for $\overline{q}_i$ and obtain

$$\overline{q}_i = -\frac{1}{\alpha}(B^*\overline{z})_i.$$

In particular, the control $\overline{q}$ is given by

$$\overline{q} = \max(q_{\min}, \min(q_{\max}, -1/\alpha(B^*\overline{z}))) = \max(q_{\min}, \min(q_{\max}, -1/\alpha(S^*(S\overline{q} - u^d)))$$

which can be used to design algorithms (semi-smooth Newton).

Remark 1.2.10. Again the resulting optimality conditions can also be obtained by Lagrange calculus, applied to the Lagrange function

$$\mathscr{L}(q, u, z, \mu_{\min}, \mu_{\max}) = J(q, u) - (Au - Bq, z) + (q_{\min} - q, \mu_{\min}) + (q - q_{\max}, \mu_{\max})$$

which yields the KKT-Conditions

$$\begin{aligned} \nabla_q \mathscr{L}(q, u, z, \mu_{\min}, \mu_{\max}) &= 0, \\ \nabla_u \mathscr{L}(q, u, z, \mu_{\min}, \mu_{\max}) &= 0, \\ \nabla_z \mathscr{L}(q, u, z, \mu_{\min}, \mu_{\max}) &= 0, \\ \mu_{\min} \ge 0, \quad q_{\min} - q \le 0, \quad (q_{\min} - q, \mu_{\min}) &= 0, \\ \mu_{\max} \ge 0, \quad q - q_{\max} \le 0, \quad (q - q_{\max}, \mu_{\max}) &= 0. \end{aligned}$$

It is then easy to see, that

$$\begin{aligned} \mu_{\min} &= \left(B\overline{z} + \nabla J_q(\overline{q}, \overline{u})\right)^+, \\ \mu_{\max} &= \left(B\overline{z} + \nabla J_q(\overline{q}, \overline{u})\right)^-, \end{aligned}$$

with $(f)^+ = \max(0, f)$ and $(f)^- = -\min(0, f)$ and the system is equivalent to (1.3).

2 Functional Analytic Background

2.1 Normed Linear Spaces

Definition 2.1.1. A $\mathbb{R}$-linear space V is called normed linear space if there exists a map $\|\cdot\| = \|\cdot\|_V : V \to \mathbb{R}_{\geq 0}$ which satisfies:

1. $\|v\| = 0$ if and only if $v = 0$.
2. $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.
3. $\|\lambda v\| = |\lambda| \|v\|$ for all $\lambda \in \mathbb{R}$ and $v \in V$.

Definition 2.1.2. A sequence $v_k \in V$ in a normed linear space V is called Cauchy sequence if for any given ε there exists some $n_0 \in \mathbb{N}$ such that

$$\|v_k - v_l\| \leq \varepsilon \quad \forall\, k, l \geq n_0.$$

Definition 2.1.3. A normed linear space V is called Banach space if it is complete, i.e., any Cauchy sequence in V has a limit in V.

Definition 2.1.4. A Banach space V whose norm is induced by a scalar product is called a Hilbert space. We recall, a scalar product is a positive definite bilinear form, $(\cdot,\cdot)_V = (\cdot,\cdot) : V \times V \to \mathbb{R}$, e.g.,

- $(v, v) \geq 0$ for all $v \in V$ and $(v, v) = 0$ if and only if $v = 0$.
- $(v, w) = (w, v)$ for all $v, w \in V$.
- $(\lambda v_1 + v_2, w) = \lambda (v_1, w) + (v_2, w)$ for all $\lambda \in \mathbb{R}$, $v_1, v_2, w \in V$.

The scalar product induces a norm by the following definition

$$\|v\| = \sqrt{(v, v)}.$$

Example 2.1.5. An example for an infinite dimensional Banach space is given by the space $C[0,1]$ of continuous functions on the interval $[0,1]$ equipped with the norm

$$\|f\|_\infty = \max_{x \in [0,1]} |f(x)|.$$

Example 2.1.6. An example for an infinite dimensional Hilbert space is given by the space l_2 of square summable sequences, e.g., $(x_i)_{i=0}^{\infty} \in l_2$ if and only if $\sum_{i=0}^{\infty} x_i^2 < \infty$. The scalar product is given by

$$(x, y) = \sum_{i=0}^{\infty} x_i y_i, \quad x, y \in l_2.$$

Remark 2.1.7. A norm $\|\cdot\|$ on a normed space V is induced by a scalar product if and only if the parallelogram law is satisfied, e.g., it holds for $v, w \in V$

$$2\|v\|^2 + 2\|w\|^2 = \|v + w\|^2 + \|v - w\|^2.$$

Further, if $\|\cdot\|$ is induced by a scalar product $(\cdot, \cdot)$ then the Cauchy Schwarz inequality

$$|(v, w)| \leq \|v\|\|w\|$$

holds for any $v, w \in V$.

We conclude this section with an important definition.

Definition 2.1.8. A normed linear space V is called separable, if there exists a countable dense (w.r.t the norm on V) subset of V.

Example 2.1.9. A typical example of a separable Banach space is again $C[0, 1]$ with the norm $\|\cdot\|_\infty$. Since it is known from calculus lessons, that any continuous function can be approximated uniformly by polynomials (Theorem of Stone-Weierstrass).

2.2 Convexity

Definition 2.2.1. Let V be a vector space. A set $C \subset V$ is called convex, if for any $\lambda \in (0, 1)$ and $v, w \in C$ it holds

$$\lambda v + (1 - \lambda) w \in C.$$

The convex hull of a set $A \subset V$ is defined as

$$\bigcap_{\substack{C \supset A \\ C \text{ convex}}} C.$$

Definition 2.2.2. Let V be a vector space and $C \subset V$ be convex. A function $f : C \to \mathbb{R} \cup \{\infty\}$ is called convex, if (with the usual definitions for arithmetic with ∞) for any $\lambda \in (0,1)$ and $v, w \in C$ it holds

$$f(\lambda v + (1-\lambda)w) \leq \lambda f(v) + (1-\lambda) f(w).$$

Theorem 2.2.3. *Let V be a Hilbert space and $C \subset V$ nonempty, closed, and convex. Then there exists a unique mapping $\mathscr{P}_C : V \to C$ given by*

$$\|v - \mathscr{P}_C(v)\| = \operatorname{dist}(v, C) = \inf_{w \in C} \|v - w\| \quad \forall\, v \in V.$$

Proof. Let $v_k \in C$ be a minimizing sequence, e.g.,

$$\|v - v_k\| \to \inf_{w \in C} \|v - w\| =: d.$$

By the parallelogram law we get for any $k, l \in \mathbb{N}$

$$\|(v - v_k) - (v - v_l)\|^2 + \|(v - v_k) + (v - v_l)\|^2 = 2(\|v - v_l\|^2 + \|v - v_k\|^2).$$

This yields

$$\|v_k - v_l\|^2 = 2(\|v - v_l\|^2 + \|v - v_k\|^2 - 2\|v - \tfrac{1}{2}(v_k + v_l)\|^2).$$

Now, $\frac{1}{2}(v_k + v_l) \in C$ and thus $\|v - \frac{v_k + v_l}{2}\|^2 \geq d^2$. Hence, we get

$$\|v_k - v_l\|^2 \leq 2(\|v - v_l\|^2 + \|v - v_k\|^2 - 2d^2) \to 0 \quad (k, l \to \infty).$$

Now, V is complete, and C is closed, hence there exists $v_\infty \in C$ with $v_k \to v_\infty$. By continuity of the norm it follows

$$\|v - v_\infty\| = d.$$

To see uniqueness, assume that there is some $v'_\infty \in A$ satisfying

$$\|v - v'_\infty\| = d$$

an application of the parallelogram law yields as above

$$\|v_\infty - v'_\infty\| \leq 2(\|v - v_\infty\|^2 + \|v - v_\infty\|^2 - 2d^2) = 0$$

and hence $\mathscr{P}_C(v) = v_\infty$. □

Lemma 2.2.4. *Let V be a Hilbert space and $C \subset V$ nonempty, closed, and convex. The element $\mathscr{P}_C$ in Theorem 2.2.3 is given equivalently by the variational inequality*

$$(v - \mathscr{P}_C(v), c - \mathscr{P}_C(v)) \leq 0 \quad \forall c \in C.$$

Proof. Let $c \in C$ be given. Then for any $\lambda \in (0,1)$ it holds $(1-\lambda)\mathscr{P}_C(v) + \lambda c \in C$ and hence

$$\begin{aligned}\|v - \mathscr{P}_C(v)\|^2 &\leq \|v - ((1-\lambda)\mathscr{P}_C(v) + \lambda c)\|^2 \\ &= \|v - \mathscr{P}_C(v)\|^2 - 2\lambda(v - \mathscr{P}_C(v), c - \mathscr{P}_C(v)) + \lambda^2\|c - \mathscr{P}_C(v)\|^2.\end{aligned}$$

Dividing by λ and taking the limit $\lambda \to 0$ yields the desired variational inequality.

To see the converse, if the variational inequality holds we have

$$\begin{aligned}\|v - c\|^2 &= \|v - \mathscr{P}_C(v) + \mathscr{P}_C(v) - c\|^2 \\ &= \|v - \mathscr{P}_C(v)\|^2 + 2(v - \mathscr{P}_C(v), \mathscr{P}_C(v) - c) + \lambda^2\|c - \mathscr{P}_C(v)\|^2 \\ &\geq \|v - \mathscr{P}_C(v)\|^2.\end{aligned}$$

□

2.3 Linear Operators

In the following, let V, W be normed linear spaces.

Definition 2.3.1. A mapping $A\colon V \to W$ is called linear or linear operator if for any $\lambda \in \mathbb{R}$ and $v_1, v_2 \in V$ it holds

$$A(\lambda v_1 + v_2) = \lambda A v_1 + A v_2.$$

If $A\colon V \to \mathbb{R}$ is linear, then it is called a linear functional.

Definition 2.3.2. A linear operator $A\colon V \to W$ is bounded if there exists a constant c_A independent of v such that

$$\|Av\|_W \leq c_A \|v\|_V \quad \forall\, v \in V.$$

The set of all bounded linear operators from V to W is denoted by

$$\mathscr{L}(V, W).$$

If $V = W$ we write $\mathscr{L}(V) := \mathscr{L}(V, V)$.

Theorem 2.3.3. *$\mathscr{L}(V, W)$ is a normed linear space (exercise) with norm*

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

$\mathscr{L}(V, W)$ is a Banach space if W is a Banach space.

Theorem 2.3.4. *A linear operator $A\colon V \to W$ is bounded if and only if it is continuous.*

Example 2.3.5 (Multiplication operator)**.** Consider $V = C[0,1]$. Let $f \in V$ be an arbitrary function. Then we define a linear operator A by

$$(Ag)(x) = f(x)g(x) \quad \forall\, x \in [0,1], g \in C[0,1].$$

The operator A is an element in $\mathscr{L}(C[0,1], C[0,1])$ because

$$\|Ag\|_\infty = \max_{x\in[0,1]} |f(x)g(x)| \leq \max_{x\in[0,1]} |f(x)| \max_{x\in[0,1]} |g(x)| = \|f\|_\infty \|g\|_\infty.$$

Hence $\|A\| \leq \|f\|_\infty$. In fact $\|A\|_\infty = \|f\|_\infty$ (exercise).

Definition 2.3.6. The set $\mathscr{L}(V, \mathbb{R})$ is called (topological) dual space to V and is denoted by V^*.

Theorem 2.3.7 (Riesz representation)**.** *Let V be a Hilbert space with scalar product $(\cdot,\cdot)$. Then for any bounded linear functional $F \in V^*$ there exists a unique element $f \in V$ such that*

$$(f, v) = F(v) \quad \forall\, v \in V.$$

Proof. Assume w.l.o.g. that $F \neq 0$. The set $C = \mathscr{N}(F) = \{x \in H \,|\, F(x) = 0\}$ is clearly convex, closed and nonempty (Exercise). Hence by Theorem 2.2.3 the projection $\mathscr{P}_C$ is well defined.

We choose some $w \in V$ with $F(w) = 1$ and define $\tilde{f} = w - \mathscr{P}_C(w)$. We obtain $F(\tilde{f}) = 1$ and hence $\tilde{f} \neq 0$. From Lemma 2.2.4 (and the fact that C is a subspace) we obtain

$$(\tilde{f}, w') = 0 \quad \forall w' \in C = \mathscr{N}(F).$$

Now let $v \in V$ be arbitrary, then $v - F(v)\tilde{f} \in \mathscr{N}(F)$ and thus

$$(\tilde{f}, v) = (\tilde{f}, v - F(v)\tilde{f}) + (\tilde{f}, F(v)\tilde{f}) = F(v)\|\tilde{f}\|^2.$$

Hence $f = \frac{\tilde{f}}{\|\tilde{f}\|^2}$ yields the desired.

To see uniqueness we use the non degeneracy of the scalar product. Let $\hat{f} \in V$ be a second element with $(\hat{f}, v) = F(v)$ for all $v \in V$. Then it holds

$$0 = F(f - \hat{f}) - F(f - \hat{f}) = (f, f - \hat{f}) - (\hat{f}, f - \hat{f}) = \|f - \hat{f}\|^2.$$

□

For more general cases one may need the following generalization

Theorem 2.3.8 (Lax-Milgram). *Let V be a Hilbert space and $a: V \times V \to \mathbb{R}$ a bilinear form which satisfies for some constants $\alpha, \beta > 0$ that for any $u, v \in V$ it holds*

1. $|a(u,v)| \leq \beta \|u\| \|v\|$ *(Continuity)*
2. $\alpha \|v\|^2 \leq a(v,v)$ *(Coercivity).*

Then there exists a unique bijective $A \in \mathscr{L}(V)$ such that

$$(Au, v) = a(u,v) \quad \forall\, u, v \in V.$$

Further it holds

$$\|A\| \leq \beta, \quad \|A^{-1}\| \leq \frac{1}{\alpha}$$

Proof. Exercise. □

For reasons of notational simplicity it is sometimes convenient to define a duality pairing $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{B,B^*} : B \times B^* \to \mathbb{R}$ for arbitrary Banach spaces B by the following definition

$$\langle v, f \rangle = f(v) \quad \forall\, v \in B, f \in B^*.$$

Before we conclude this section we will define the bidual to a Banach space B as $B^{**} = (B^*)^* = \mathscr{L}(B^*, \mathbb{R})$. It is then clear, that we can define a mapping $i \in \mathscr{L}(B, B^{**})$ as follows:

$$\langle v', i(v) \rangle_{B^*,B^{**}} = \langle v, v' \rangle_{B,B^*} = v'(v) \quad \forall\, v \in B, v' \in B^*.$$

In fact the mapping i is an isometry, e.g., $\|i(v)\| = \|v\|$.

Definition 2.3.9. A Banach space B is called reflexive if i is surjective, e.g., B and B^{**} are isometrically isomorphic.

Example 2.3.10.
1. Every Hilbert space is reflexive.
2. The spaces l_p and $L^p(\Omega)$ ($\Omega \subset \mathbb{R}^n$) are reflexive if $1 < p < \infty$.
3. The space $C(\overline{\Omega})$, $L^1(\Omega)$, and $L^\infty(\Omega)$ are <u>not</u> reflexive if $\Omega \subset \mathbb{R}^n$ is open and nonempty.

2.4 Adjoints

Let U, V be Banach spaces. Let $A \in \mathscr{L}(U,V)$ be given. Then we can define for any $f \in V^*$ an element $g \in U^*$ by setting

$$g(u) = f(Au) \quad \forall u \in U$$

clearly g is linear and because of

$$|g(u)| \leq \|f\|_{V^*} \|A\|_{\mathscr{L}(U,V)} \|u\|_U$$

it holds $g \in U^*$ with

$$\|g\|_{U^*} \leq \|f\|_{V^*} \|A\|_{\mathscr{L}(U,V)}.$$

Definition 2.4.1. For any $A \in \mathscr{L}(U,V)$ the above mapping $f \mapsto g = f \circ A$ defines a linear operator $A^*\colon V^* \to U^*$, $f \mapsto f \circ A$ the adjoint or dual operator to A.

It is given by the relation

$$\langle Au, f\rangle_{V\times V^*} = \langle u, A^* f\rangle_{U\times U^*} \quad \forall\, u \in U, f \in V^*.$$

Moreover it holds $\|A^*\|_{\mathscr{L}(V^*,U^*)} = \|A\|_{\mathscr{L}(U,V)}$.

For Hilbert spaces there exists a natural choice for the duality pairing given by Theorem 2.3.7. This gives the following

Definition 2.4.2. Let U and V be Hilbert spaces and $A \in \mathscr{L}(U,V)$. Then we define the Hilbert space adjoint $A^* \in \mathscr{L}(V,U)$ to A by

$$(Au, v)_V = (u, A^* v)_U \quad \forall\, u \in U, v \in V.$$

Note that the notation is sloppy as there is a difference between the adjoint and the Hilbert space adjoint (Which one?). We will ignore this because it will be clear from the context which one is meant.

2.5 Weak Convergence

We have seen in some examples that the the usual finite dimensional proof for existence of minimizers fails, because in infinite dimensional spaces the closed unit ball is not compact. To circumvent this we will derive a new notion of convergence for these spaces.

Definition 2.5.1 (Weak convergence)**.** Let B be a Banach space. A sequence $v_k \in B$ converges weakly to some $v \in B$ if

$$\langle v_k, f\rangle \to \langle v, f\rangle \quad \forall f \in B^*.$$

This is written as $v_k \rightharpoonup v$.

Definition 2.5.2 (Weak* convergence)**.** Let B be a Banach space. A sequence $f_k \in B^*$ converges weakly* to some $f \in B^*$ if

$$\langle v, f_k\rangle \to \langle v, f\rangle \quad \forall v \in B.$$

This is written as $f_k \rightharpoonup^* f$.

Example 2.5.3. Consider the Hilbert space $L^2(0, 2\pi)$ with the sequence

$$v_k = \sin(kx)$$

Then we have for any $f \in L^2(0, 2\pi)$ that

$$(f, v_k) = \int_0^{2\pi} f(x) \sin(kx)\, dx \to 0$$

by Parseval's identity (The scalar product is up to a constant the k-th Fourier coefficient. This means $v_k \rightharpoonup 0$. Moreover we have that

$$\|v_k\|^2 = \pi.$$

And hence we immediately obtain that the norm is not weakly continuous (with the obvious definition) because

$$0 = \|0\| \neq \lim_{k\to\infty} \|v_k\| = \sqrt{\pi}.$$

The following important properties hold:

Lemma 2.5.4. *Let B be a Banach and H be a Hilbert space.*

1. *If B is reflexive then weak and weak* convergence coincide on B^*.*
2. *If $v_k \to v$ in B (strong convergence) then $v_k \rightharpoonup v$ (weak convergence).*
3. *If $v_k \rightharpoonup v$ in B then $\|v_k\| \leq C < \infty$.*
4. *If $u_k \to u$ and $v_k \rightharpoonup v$ in H then*

$$(u_k, v_k) \to (u, v).$$

   (The analog property holds for Banach spaces)

5. *If $v_k \rightharpoonup v$ in H and $\|v_k\| \to \|v\|$ then $v_k \to v$.*

Definition 2.5.5. A set $M \subset B$ of a Banach space B is weak sequentially closed if for any sequence $v_k \in M$ with weak limit $v \in B$ it holds $v \in M$.

A set $M \subset B$ of a Banach space B is weak sequentially compact if for any sequence $v_k \in M$ there exists a subsequence v_{k_l} and an element $v \in M$ with $v_{k_l} \rightharpoonup v$.

We have the following important properties:

Theorem 2.5.6. *Let B be a Banach space.*

1. *If $M \subset B$ is closed and convex then M is weak sequentially closed.*
2. *If B is reflexive and $M \subset B$ bounded, closed, and convex then M is weak sequential compact.*
3. *If B is separable then the ball $\overline{B_1(0)}$ in B^* is weakly* sequentially compact.*

Theorem 2.5.7. *Let B be a Banach space and $J : B \to \mathbb{R}$ be a continuous and convex functional. Then J is weakly lower semicontinuous, e.g., if $v_k \rightharpoonup v$ then*

$$\liminf_{k\to\infty} J(v_k) \geq J(v).$$

Remark 2.5.8. In particular the norm $\|\cdot\| : B \to \mathbb{R}$ is convex and continuous and thus weakly lower semicontinuous.

Definition 2.5.9. Let V, W be two Banach spaces. We call a map $A \in \mathscr{L}(V, W)$ compact, if for any bounded sequence $v_k \in V$ there exists a subsequence of Av_k which converges strongly.

In particular, if $A \in \mathscr{L}(V, W)$ is compact then if $v_k \rightharpoonup v$ in V it holds $Av_k \to Av$. (Exercise)

2.6 Lebesgue and Sobolev spaces

2.6.1 Domain regularity

Throughout this lecture, we consider $\Omega \subset \mathbb{R}^n$ to be a bounded domain, i.e., Ω is open, connected, and nonempty.

Definition 2.6.1. We say that a bounded domain Ω with boundary Γ is of class $C^{k,1}$ if the boundary of Ω is locally the graph of a $C^{k,1}$ function such that Ω is locally on one side of the boundary.

For the technical detail see, e.g., [1, 40]. For our purposes the most important regularity will be Lipschitz domains ($C^{0,1}$) domains.

An other important class is that of polygonally bounded domains, e.g., those domains whose boundary is a polygon, in 2d or polyhedral boundaries in 3d.

2.6.2 Lebesgue spaces

Definition 2.6.2. On a domain Ω we define the spaces $L^p(\Omega)$ as follows

$$L^p(\Omega) = \{f \mid f \text{ is Lebesgue-measurable, } \|f\|_p^p = \int_\Omega |f(x)|^p \, dx < \infty\}, \quad 1 \le p < \infty$$

$$L^\infty(\Omega) = \{f \mid f \text{ is Lebesgue-measurable, } \|f\|_\infty = \operatorname{ess\,sup}_{x\in\Omega} |f(x)| < \infty\}$$

The space $L^1_{\text{loc}}(\Omega)$ is defined as those functions that are in $L^1(C)$ for any compact set $C \subset \Omega$.

Remark 2.6.3. Note that from now on if we write down a norm or scalar product for functions without any indices this will be the $L^2(\Omega)$ norm or scalar product.

Theorem 2.6.4. *The dual space to $L^p(\Omega)$ for $1 \le p < \infty$ is isometrically isomorphic to $L^{p'}(\Omega)$ where $\frac{1}{p} + \frac{1}{p'} = 1$.*

Further they are reflexive for $1 < p < \infty$.

Theorem 2.6.5 (Lebesgue differentiation theorem)**.** *Let $f \in L^1(\Omega)$ for some bounded domain $\Omega \subset \mathbb{R}^n$. Then for almost all $x \in \Omega$ it holds*

$$f(x) = \lim_{r\to 0} \frac{1}{|B_r(x)|} \int_{B_r(x)} f(y)\, dy.$$

Proof. For a proof see [15]. □

2.6.3 Sobolev spaces

We recall that $C_0^\infty(\Omega)$ denotes the space of all arbitrarily times differentiable functions with compact support in Ω.

Motivated by the formula for partial integration of smooth functions we define the following.

Definition 2.6.6 (Weak derivatives)**.** Let $u \in L^1_{\text{loc}}(\Omega)$ and α a multi-index. If there exists a function $w \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega u(x)\partial^\alpha \varphi(x)\, dx = (-1)^{|\alpha|} \int_\Omega \varphi(x) w(x)\, dx \quad \forall\, \varphi \in C_0^\infty(\Omega)$$

then we call w the weak derivative of order α of u and write $w = \partial^\alpha u$.

As usual we denoted by ∂^α the term

$$\partial^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1} \cdots \frac{\partial^{\alpha_n}}{\partial x_n}.$$

Example 2.6.7. The function $u(x) = |x|$ on $\Omega = (-1, 1)$ has the weak derivative

$$\partial^1 u(x) = u'(x) = \begin{cases} -1 & x \in (-1, 0), \\ 1 & x \in (0, 1). \end{cases}$$

However, u has no higher weak derivatives (Exercise).

We can now define the Sobolev spaces

Definition 2.6.8. On a bounded domain Ω with given $k \in \mathbb{N}_0$ and $1 \leq p \leq \infty$, we define

$$W^{k,p}(\Omega) = \{u \,|\, \partial^\alpha u \in L^p(\Omega)\}.$$

Together with the norm

$$\|u\|_{k,p} = \Big(\sum_{|\alpha| \leq k} \|\partial^\alpha u\|_p^p \Big)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|u\|_{k,\infty} = \max_{|\alpha| \leq k} \|\partial^\alpha u\|_\infty$$

they are Banach spaces. For $1 < p < \infty$ they are reflexive.

Definition 2.6.9. On a bounded domain Ω, we define for $k \in \mathbb{N}_0$ and $1 \leq p < \infty$ we define $W_0^{k,p}(\Omega)$ as the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$.

In the special case that $p = 2$ these spaces are Hilbert spaces, we write $H^k(\Omega) = W^{k,2}(\Omega)$ and $H_0^k(\Omega) = W_0^{k,2}(\Omega)$.

Example 2.6.10. In the special case of $H^1(\Omega)$, which we will frequently use, the norm has the following form

$$\|u\|_{1,2}^2 = \int_\Omega u^2 + |\nabla u|^2 \, dx$$

The corresponding scalar product is given as

$$(u, v)_{1,2} = \int_\Omega uv + \nabla u \cdot \nabla v \, dx$$

where $\cdot$ denotes the euclidean scalar product in $\mathbb{R}^n$.

Theorem 2.6.11 (Boundary Traces). *Let Ω be a bounded Lipschitz domain and $1 \leq p \leq \infty$. Then there exists a unique operator $\tau \in \mathcal{L}(W^{1,p}(\Omega), L^p(\partial\Omega))$ such that*

$$\tau u = u\big|_{\partial\Omega} \qquad \forall u \in W^{1,p}(\Omega) \cap C(\overline{\Omega}).$$

In slightly sloppy notation, we will usually neglect to explicitly write down the operator τ when it is clear that the trace is meant.

With these preparations we can give an equivalent definition for the space $H_0^1(\Omega)$ by the following

$$H_0^1(\Omega) = \{u \in H^1(\Omega) \,|\, \tau u = 0\}. \tag{2.1}$$

Theorem 2.6.12 (Poincaré inequality). *Let Ω be a bounded domain. Then there exists a constant c_Ω depending on Ω only such that*

$$\|u\|^2 = \int_\Omega u^2 \, dx \leq c_\Omega \|\nabla u\|^2 = c_\Omega \int_\Omega |\nabla u|^2 \, dx \quad \forall\, u \in H_0^1(\Omega).$$

Corollary 2.6.13. *The norms $\|\cdot\|_{1,2}$ and $\|\nabla \cdot\|$ are equivalent on $H_0^1(\Omega)$.*

2.6.4 Embedding theorems

Now, in order to define constraints onto our solution variables it is nice to know which Sobolev space are contained in spaces of continuous functions. To see this we will recall several embedding theorems.

Theorem 2.6.14. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. Let $k_1, k_2 \in \mathbb{N}_0$ and $1 \leq p_1, p_2 < \infty$.*
a) If it holds

$$k_1 - \frac{n}{p_1} \geq k_2 - \frac{n}{p_2}, \quad k_1 \geq k_2$$

then the embedding $\mathrm{Id}\colon W^{k_1,p_1}(\Omega) \to W^{k_2,p_2}(\Omega)$ *exists and is continuous (but not compact). This means in particular that there exists a constant* c *depending on* $n, \Omega, k_1, k_2, p_1, p_2$ *such that*

$$\|u\|_{k_2,p_2} \leq c\|u\|_{k_1,p_1}.$$

b) If it holds

$$k_1 - \frac{n}{p_1} > k_2 - \frac{n}{p_2}, \quad k_1 > k_2$$

then the embedding $\mathrm{Id}\colon W^{k_1,p_1}(\Omega) \to W^{k_2,p_2}(\Omega)$ *is in addition compact.*

Theorem 2.6.15. *Let* $\Omega \subset \mathbb{R}^n$ *be a Lipschitz domain. Let* $k_1, k_2 \in \mathbb{N}_0$ *with* $k_1 \geq 1$ *and* $1 \leq p < \infty$ *and* $0 \leq \alpha \leq 1$. *If it holds*

$$k_1 - \frac{n}{p} > k_2 + \alpha,$$

then the embedding $\mathrm{Id}\colon W^{k_1,p}(\Omega) \to C^{k_2,\alpha}(\overline{\Omega})$ *exists and is continuous and compact.*

Theorem 2.6.16. *Let* $\Omega \subset \mathbb{R}^n$ *be a Lipschitz domain. Let* $k \in \mathbb{N}$, *then the embedding*

$$C^{k,1}(\overline{\Omega}) \to W^{k+1,\infty}(\Omega)$$

exists and is an isomorphism.

Example 2.6.17. We consider again the case of the space $H^1(\Omega)$. Now we consider the question whether or not these functions are continuous. To this end we note that the relevant inequality gets

$$1 - \frac{n}{2} > 0.$$

Hence if $n = 1$ we obtain that all $H^1(\Omega)$ functions are continuous. However the inequality fails for any dimension $n \geq 2$.

In fact for these dimensions functions in $H^1(\Omega)$ are not necessarily continuous. To see this let $\Omega = B_1(0) \subset \mathbb{R}^2$ and consider the function $u(x) = \ln(\ln(r^{-1}) + 1)$. A straightforward computation shows that $u \in H^1(\Omega)$ but clearly $u \notin C^0(\overline{\Omega})$.

3 Optimization in Banach Spaces I

We can now derive a result that closely reassembles the finite dimensional case given in Theorem 1.2.3.

Theorem 3.0.1 (An abstract existence result)**.** *Let Q be a Hilbert space, $j : Q \to \mathbb{R}$ be continuous, convex. Let $Q^{ad} \subset Q$ be closed and convex. Then if either of the following holds*

1. *Q^{ad} is bounded.*
2. *$j(q) \to \infty$ if $\|q\| \to \infty$.*

there exists a solution to

$$\min_{q \in Q^{ad}} j(q)$$

Proof. First we choose a minimizing sequence q_k, e.g., $j(q_k) \to \inf_{q \in Q^{\text{ad}}} j(q) := \overline{j}$.

Now we show that the set $\{q_k\}$ is bounded. If Q^{ad} is bounded there is nothing to show. Otherwise let $q_0 \in Q^{\text{ad}}$ be given, because $j(q) \to \infty$ as $\|q\| \to \infty$ there exists some $R > 0$ such that $j(q) \geq j(q_0)$ if $\|q\| \geq R$ and we can w.l.o.g take $q_k \in Q^{\text{ad}} \cap \overline{B_R(0)}$.

Now both Q^{ad} and $Q^{\text{ad}} \cap \overline{B_R(0)}$ are bounded, closed, and convex. By Theorem 2.5.6.2. there exists some $\overline{q} \in Q^{\text{ad}}$ and a subsequence q_k such that

$$q_k \rightharpoonup \overline{q}.$$

Finally, by Theorem 2.5.7 we note that

$$\overline{j} = \lim_{k \to \infty} j(q_k) \geq j(\overline{q}) \geq \overline{j}.$$

Hence, $\overline{q}$ is a solution. □

Remark 3.0.2. We remark, that Theorem 3.0.1 remains valid if Q is a reflexive Banach space and j is simply weakly lower semicontinuous. Further it is sufficient if Q^{ad} (or $Q^{\text{ad}} \cap \{q \,|\, j(q) \leq j(q_0)$ for some $q_0 \in Q^{\text{ad}}\}$) is weak sequentially compact.

4 Linear-quadratic elliptic optimization problems I

4.1 Poisson's equation

Although one could define an elliptic operator in a very general setting, e.g., a second order elliptic operator L in divergence form would look like

$$Lu = \nabla \cdot (A\nabla u + bu) + c \cdot \nabla u + du$$

with some $A\colon \Omega \to \mathbb{R}^{n\times n}$ symmetric and uniformly positive definite, $b, c\colon \Omega \to \mathbb{R}^n$ and $d \in \mathbb{R}$ and some additional boundary operators. However, for sake of notational simplicity we will throughout consider only a prototypical example namely Poisson's equation with homogeneous Dirichlet boundary values. We note however that everything said here can be transfered directly to other equations provided some conditions on the coefficients are satisfied.

Let Ω be a bounded Lipschitz domain, $f \in L^2(\Omega)$ a given function. Then we consider the following boundary value problem to determine u

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma = \partial\Omega \end{aligned} \tag{4.1}$$

where $\Delta = \partial_1^2 + \partial_2^2 = \operatorname{div}\operatorname{grad}$.

We will now consider the weak formulation of this equation. To obtain this we multiply with an arbitrary test function $\varphi \in H_0^1(\Omega)$ (or $\varphi \in C_0^\infty(\Omega)$) to get

$$-\int_\Omega \Delta u\varphi \, dx = \int_\Omega f\,\varphi \, dx.$$

Integration by parts gives

$$\int_\Omega \nabla u \cdot \nabla\varphi \, dx - \int_\Gamma \partial_\nu u\varphi \, ds = \int_\Omega f\,\varphi \, dx$$

with the outer (unit) normal ν to Ω. Now, we note that due to (2.1) we have $\varphi = 0$ on Γ and hence we obtain the weak formulation of the boundary value problem (4.1)

$$\int_\Omega \nabla u \cdot \nabla\varphi \, dx = \int_\Omega f\,\varphi \, dx \quad \forall\, \varphi \in H_0^1(\Omega) \tag{4.2}$$

or in short notation

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

Theorem 4.1.1. *For given $f \in L^2(\Omega)$ the problem (4.2) admits a unique weak solution $u \in H_0^1(\Omega)$. Further the stability estimate*

$$\|\nabla u\| \leq \sqrt{c_\Omega}\|f\|.$$

holds.

Proof. By Cauchy-Schwarz inequality and Poincaré inequality (Theorem 2.6.12) it holds

$$(f,\varphi) \leq \|f\|\|\varphi\| \leq \sqrt{c_\Omega}\|f\|\|\nabla\varphi\|.$$

Thus $F = (f,\cdot)$ defines an element in $(H_0^1(\Omega))^*$ with $\|F\|_{(H_0^1(\Omega))^*} \leq \sqrt{c_\Omega}\|f\|$. On the other hand we know from Corollary 2.6.13 that $(\nabla\cdot,\nabla\cdot)$ is the scalar product on $H_0^1(\Omega)$.

Hence by Riesz representation Theorem 2.3.7, we know that there exists a unique solution $u \in H_0^1(\Omega)$ to (4.2).

To see the stability estimate, we test (4.2) with the solution u and get

$$\|\nabla u\|^2 = (f,u) \leq \|f\|\|u\| \leq \sqrt{c_\Omega}\|f\|\|\nabla u\|$$

which proofs the assertion. □

We will consider again the space $(H_0^1(\Omega))^*$. We have seen in the above proof that a function $f \in L^2(\Omega)$ defines an element in $H_0^1(\Omega)^*$ as follows

$$(f,\cdot).$$

We choose $(\cdot,\cdot)$ as duality pairing on $H_0^1(\Omega)$. Then, we can represent the dual $H^{-1}(\Omega) = (H_0^1(\Omega))^*$ as follows

$$f \in H^{-1}(\Omega) \Leftrightarrow f = f_0 + \sum_{i=1}^{n} \partial_i f_i \quad \text{with } f_i \in L^2(\Omega).$$

The norm on $H^{-1}(\Omega)$ is then given as

$$\|f\|_{-1} = \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\nabla\varphi\|=1}} (f,\varphi)$$

By applying virtually the same arguments as in Theorem 4.1.1, we obtain the following

Theorem 4.1.2. *For any $f \in H^{-1}(\Omega)$ there exists a unique solution $u \in H_0^1(\Omega)$ to*

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

Further it holds

$$\|\nabla u\| \leq \|f\|_{-1}.$$

Proof. The proof is identical to the one for Theorem 4.1.1. □

4.2 Elliptic regularity

As we have seen in Theorem 4.1.1 the solution of the Poisson problem for given right hand side $f \in L^2(\Omega)$ is given naturally in the spaces $H_0^1(\Omega)$. In fact, the solution will exhibit some additional regularity. This is what we would suspect from the case of ODEs where the regularity of the right hand side transfers to regularity of the solution.

Theorem 4.2.1. *Assume that Ω is a $C^{1,1}$ domain, then for given $f \in L^2(\Omega)$ the solution u to (4.2) satisfies the regularity $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and there is a constant c (independent of u and f) such that*

$$\|u\|_{2,2} \leq c\|f\|.$$

More general, if Ω is a $C^{k+1,1}$ domain for some $k \geq 0$ then for $f \in H^k(\Omega)$ the solution u to (4.2) satisfies the regularity $u \in H^{k+2}(\Omega) \cap H_0^1(\Omega)$ and there is a constant c (independent of u and f) such that

$$\|u\|_{k+2,2} \leq c\|f\|_{k,2}.$$

Remark 4.2.2. Similar results hold for general elliptic operators provided the coefficients are sufficiently regular, see, e.g., [16].

Remark 4.2.3. However such a result does not hold on polygonal domains. To see this consider a cone $\Omega \subset \mathbb{R}^2$ with angle $\omega \in [0, 2\pi]$. Then the functions $u = \sin(\theta\pi/\omega)r^{\pi/\omega}$ (where r and θ are polar coordinates with respect to the angle) satisfy $-\Delta u = 0$ but are not in H^2 if $\omega > \pi$.

In fact the solutions u to (4.2) satisfy the H^2 bound in Theorem 4.2.1 if $\Omega \subset \mathbb{R}^2$ is convex. For more details see, e.g., [17].

4.3 Existence of Minimizers

We now come back to our model problem on a domain $\Omega \subset \mathbb{R}^2$.

$$\begin{aligned} \min J(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } u \in H_0^1(\Omega), \\ (\nabla u, \nabla\varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ q_{\min} \leq q \leq q_{\max} \quad \text{a.e. on } \Omega, \end{aligned} \tag{4.3}$$

with given $u^d \in L^2(\Omega)$, $\alpha > 0$, $q_{\min} < q_{\max} \in \mathbb{R}$. As we have seen in Theorem 4.1.1 (and Theorem 2.6.14) the equation $(\nabla u, \nabla\varphi) = (q,\varphi)$ defines a continuous linear operator

$L^2(\Omega) \to L^2(\Omega)$, where $q \mapsto u = u_q$. Hence the reduced cost functional $j(q) = J(q,u)$ is convex and continuous. We can apply Theorem 3.0.1 to obtain the existence of a solution $(\overline{q},\overline{u}) \in L^2(\Omega) \times H_0^1(\Omega)$ to (4.3). In fact, j is strictly convex if $\alpha > 0$, and hence the solution is unique (Exercise).

Before we conclude this section we consider a counter example to show that the convexity condition on j can not be removed in the infinite dimensional case.

Example 4.3.1. Consider the problem

$$\begin{aligned} \min J(q,u) = \frac{1}{2}\|u\|^2 + \int_\Omega (q(x)^2 - 1)^2 \, dx \\ \text{s.t. } u \in H_0^1(\Omega), \\ (\nabla u, \nabla \varphi) = (q, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ -1 \le q \le 1 \quad \text{a.e. on } \Omega, \end{aligned}$$

on $\Omega = (0,1)^2$. Let q_n be a sequence with $q_n \in \{\pm 1\}$ a.e. and $q_n \rightharpoonup^* 0$ in $L^\infty(\Omega) = L^1(\Omega)^*$ (Existence by checkerboards of vanishing fineness) and hence $q_n \rightharpoonup 0$ in $L^p(\Omega)$ for $1 < p < \infty$. Then, as the solution operator $S\colon L^2(\Omega) \to L^2(\Omega)$ defined by (4.2) is compact it holds that the corresponding states $u_n \to 0$ in $L^2(\Omega)$ and hence $J(q_n,u_n) \to 0$. However the only possibility to obtain $J(\overline{q},\overline{u}) = 0$ is that $\overline{u} = 0$ and because S is injective (Exercise!) this gives $\overline{q} = 0$ but $J(0,0) = 1$ in contradiction to the assumption.

5 Optimization in Banach Spaces II

5.1 Differentiability

Let U, V be two Banach spaces, and $F: U \to V$ be some mapping.

Definition 5.1.1. For given $u \in U$ and $\delta u \in U$ assume that the limit

$$\delta F(u;\delta u) = \lim_{t \downarrow 0} \frac{F(u+t\delta u) - F(u)}{t}$$

exists, then $\delta F(u;\delta u)$ is called the directional derivative of F at u in direction δu.

Remark 5.1.2. We have here some difference in the notation compared to the finite dimensional space which is mainly the case because in Banach spaces the directional derivative is used frequently in the calculus of variation, where it is denoted as δF.

Note that the directional derivative is not necessarily linear, e.g., $F: \mathbb{R} \to \mathbb{R}$, $F(u) = |u|$.

Definition 5.1.3. Assume that $\delta F(u;\delta u)$ exists for all $\delta u \in U$ and that there exists an operator $A \in \mathscr{L}(U,V)$ such that

$$A\delta u = \delta F(u;\delta u) \quad \forall\, \delta u \in U$$

then F is called Gâteaux differentiable at u. The mapping $A = F'(u)$ is called the Gâteaux derivative.

Definition 5.1.4. F is called Fréchet differentiable at u if there exists a function $\omega: U \to V$ with $\omega(\delta u) \in o(\|\delta u\|_U)$ such that

$$F(u+\delta u) = F(u) + F'(u)\delta u + \omega(\delta u).$$

In this case $F'(u)$ is called the Fréchet derivative.

The notion $\omega(\delta u) \in o(\|\delta u\|_U)$ means

$$\lim_{\delta u \to 0} \frac{\|\omega(\delta u)\|_V}{\|\delta u\|_U} = 0$$

Example 5.1.5.

1. Consider the mapping $A(u) = A_0 u + b$ with $A_0 \in \mathscr{L}(U,V)$ and $b \in V$. Then A is Fréchet differentiable with $F'(u) = A_0$ since

$$A(u+\delta u) = A_0 u + A_0 \delta u + b = A(u) + A_0 \delta u.$$

2. Let H be a Hilbert space with scalar product $(\cdot,\cdot)$. Then the natural norm $F(u) = \|u\|^2$ is Fréchet differentiable with

$$F'(u) = 2u.$$

   (Exercise)

3. Consider the space $U = C[0,1]$ together with the mapping $\sin\colon U \to U$ given by $u \mapsto \sin(u)$. (This is the Nemyzkii-operator generated by $\sin\colon \mathbb{R} \to \mathbb{R}$.) Then the directional derivative is given by

$$\begin{aligned}\delta F(u;\delta u)(x) &= \lim_{t \downarrow 0} \frac{1}{t}\Big(\sin(u(x) + t\delta u(x)) - \sin(u(x))\Big) \\ &= \cos(u(x))\delta u(x).\end{aligned}$$

   Now we know from Example 2.3.5 that the operator given by pointwise multiplication with $\cos(u(x))$ is in $\mathscr{L}(U,U)$ and hence sin is Gâteaux differentiable with $\sin' = \cos$. To obtain Fréchet differentiability we note that

$$|\sin(u(x)+\delta u(x)) - \sin(u(x)) - \cos(u(x))\delta u(x)| = \frac{1}{2}|\sin(\xi_x)|\delta u(x)^2 = \omega(\delta u)$$

   and hence

$$\frac{\sup_{x\in[0,1]} |\sin(u(x)+\delta u(x)) - \sin(u(x)) - \cos(u(x))\delta u(x)|}{\|\delta u\|_\infty} \leq \|\delta u\|_\infty$$

4. We now consider $\sin\colon L^p(0,1) \to L^p(0,1)$ for some $1 \leq p < \infty$. We note that the remainder in the Taylor expansion of the sin-function in $u = 0$ can be written pointwise as

$$\omega(\delta u) = \sin(0+\delta u(x)) - \sin(0) - \cos(0)\delta u(x) = \sin(\delta u(x)) - \delta u(x).$$

   Now we can define a sequence of functions

$$\delta u_\varepsilon(x) = \begin{cases} 1 & x \in [0,\varepsilon], \\ 0 & \text{otherwise,} \end{cases} \in L^p(0,1)$$

and obtain

$$\|\omega(\delta u_\varepsilon)\|_p^p = \int_0^1 |\omega(\delta u_\varepsilon(x))|^p \, dx = \int_0^\varepsilon (1-\sin(1))^p \, dx = (1-\sin(1))^p \varepsilon.$$

This gives

$$\frac{\|\omega(\delta u_\varepsilon)\|_p}{\|\delta u_\varepsilon\|_p} = \frac{(1-\sin(1))\varepsilon^{1/p}}{\varepsilon^{1/p}} = (1-\sin(1)) > 0$$

and hence sin is not differentiable between $L^p(\Omega)$ with $1 \leq p < \infty$.

Theorem 5.1.6. *Let U, V, W be Banach spaces and let $\mathcal{O}_U \subset U$ and $\mathcal{O}_V \subset V$ be open. Then the following is true:*

1. *If $f, g : \mathcal{O}_U \to V$ are Gâteaux or Fréchet differentiable then so is $\lambda f + g$ with $\lambda \in \mathbb{R}$ and it holds*
$$(\lambda f + g)'(u) = \lambda f'(u) + g'(u).$$

2. *(Mean-value theorem)*
If $f : \mathcal{O}_U \to V$ is Gâteaux differentiable and the interval $I = [u_0, u_1] = \{\lambda u_0 + (1-\lambda)u_1 \,|\, \lambda \in [0,1]\}$ is contained in $\mathcal{O}_U$, then
$$\|f(u_0) - f(u_1)\|_V \leq \sup_{\xi \in I} \|f'(\xi)\|_{\mathcal{L}(U,V)} \|u_0 - u_1\|_U$$

3. *If $f : \mathcal{O}_U \to V$ is Gâteaux differentiable and $f' : \mathcal{O}_U \to \mathcal{L}(U,V)$ is continuous, then f is Fréchet differentiable and we say that f is continuously differentiable and write $f \in C^1(\mathcal{O}_U, V)$.*

4. *(Chain rule)*
Let $f : \mathcal{O}_U \to V$ and $g : \mathcal{O}_V \to W$ with $f(\mathcal{O}_U) \subset \mathcal{O}_V$ be Fréchet differentiable in u_0 and $f(u_0)$, then $g \circ f$ is Fréchet differentiable in u_0 and it holds
$$(g \circ f)'(u_0) = g'(f(u_0)) \circ f'(u_0).$$

5. *(Implicit function theorem)*
Let $F : U \times V \supset \mathcal{O}_U \times \mathcal{O}_V \to W$ be continuously differentiable with $F(u_0, v_0) = 0$. Further, assume that the derivative of $v \mapsto F(u_0, v)$ is an isomorphism from V to W at $v = v_0$. Then there exists neighborhoods $N(u_0)$ of u_0 and $N(v_0)$ of v_0 such that for any $u \in N(u_0)$ the equation $F(u, v) = 0$ admits a unique solution $v = f(u) \in N(v_0)$. The mapping $f : N(u_0) \to V$ defined by this is continuously differentiable.

Example 5.1.7. Let Q, U be Hilbert spaces, $S \in \mathscr{L}(Q,U)$, and $u^d \in U$ fixed. Then $f(q) = \|Sq - u^d\|^2$ is differentiable by the chain rule. A direct application yields

$$f'(q) = 2(Sq - u^d) \circ S.$$

As this is sometimes misleading (in particular the meaning of $2(Sq-u^d)$ and $\circ$) we compute this in addition using directional derivatives to get

$$\delta f(q;\delta q) = (2(Sq - u^d), S\delta q).$$

We now assume there is an operator $S^* : U \to Q$ such that

$$(2(Sq - u^d), S\delta q) = (2S^*(Sq - u^d), \delta q) \quad \forall \delta q \in Q.$$

Then the Hilbert space representation of the derivative is

$$f'(q) = 2S^*(Sq - u^d).$$

5.2 Necessary Optimality Conditions

We will now derive some optimality conditions for solutions to optimization problems in Banach spaces. To this end, let Q be a Banach space, $Q^{\text{ad}} \subset Q$ a subset.

Definition 5.2.1. We say that $\overline{q} \in Q^{\text{ad}}$ is a local solution to the problem

$$\min_{q \in Q^{\text{ad}}} j(q) \tag{5.1}$$

if there exists $r > 0$ such that

$$j(\overline{q}) \leq j(q) \quad \forall\, q \in Q^{\text{ad}}, \|q - \overline{q}\|_Q \leq r.$$

A solution is global if this holds for all $r > 0$.

Definition 5.2.2. For a set $Q^{\text{ad}} \subset Q$ in a Banach space Q, we define the sequential tangent cone (or Bouligand cone) at $\overline{q}$ as follows

$$T(Q^{\text{ad}}, \overline{q}) := \{q \in Q \,|\, q = \lim_{k\to\infty} \tfrac{1}{t_k}(q_k - \overline{q}), t_k \downarrow 0, q_k \in Q^{\text{ad}}\}.$$

With this, we have the following

Theorem 5.2.3. *Let Q be a Banach space. Consider the problem*

$$\min_{q \in Q^{ad}} j(q)$$

with $Q^{ad} \subset Q$. Then for any (local) solution $\overline{q}$ at which j is Fréchet differentiable it holds

$$j'(\overline{q})q \geq 0 \quad \forall\, q \in T(Q^{ad}, \overline{q}).$$

Proof. Let $q \in T(Q^{\mathrm{ad}}, \overline{q})$ be given and t_k and $q_k \in Q^{\mathrm{ad}}$ be the approximating sequences. Then it holds

$$j'(\overline{q})q = \lim_{k\to\infty} \frac{1}{t_k} j'(\overline{q})(q_k - \overline{q}) = \lim_{k\to\infty} \frac{1}{t_k}(j(q_k) - j(\overline{q}) + \omega(q_k - \overline{q})).$$

Using $q_k \to \overline{q}$ and local optimality of $\overline{q}$ we get

$$j'(\overline{q})q \geq \lim_{k\to\infty} \frac{1}{t_k}\omega(q_k - \overline{q}) = \lim_{k\to\infty} \frac{\|q_k - \overline{q}\|_Q}{t_k} \frac{\omega(q_k - \overline{q})}{\|q_k - \overline{q}\|_Q} = \|q\|_Q \lim_{k\to\infty} \frac{\omega(q_k - \overline{q})}{\|q_k - \overline{q}\|_Q} = 0.$$

□

To come back to our generic setting of the reduced cost functional we restate this in a more convenient way

Theorem 5.2.4 (Necessary and sufficient optimiality conditions)**.** *Let Q be a Banach space, $Q^{ad} \subset Q$ be convex, and $j\colon Q^{ad} \to \mathbb{R}$ be directionally differentiable. Then if $\overline{q} \in Q^{ad}$ is a local solution to*

$$\min_{q \in Q^{ad}} j(q)$$

it holds the variational inequality (VI)

$$\delta j(\overline{q}; q - \overline{q}) \geq 0 \quad \forall\, q \in Q^{ad}. \tag{5.2}$$

Further, if j is convex on Q^{ad} then any $\overline{q} \in Q^{ad}$ which satisfies (5.2) is a global solution to (5.1).

Proof. The proof of the necessary condition is identical to the one for Theorem 1.2.5. (It can also be derived from Theorem 5.2.3 by noting that (Exercise)

$$\{q - \overline{q} \,|\, q \in Q^{\mathrm{ad}}\} \subset T(Q^{\mathrm{ad}}, \overline{q})$$

if j is Fréchet differentiable)

To see the sufficiency, we note that convexity of j implies that for any $q \in Q^{\mathrm{ad}}$ and any $\lambda \in (0,1)$ it holds

$$\frac{j(\overline{q} + \lambda(q - \overline{q})) - j(\overline{q})}{\lambda} \leq j(q) - j(\overline{q})$$

and thus

$$j(q) - j(\overline{q}) \geq \delta j(\overline{q}; q - \overline{q})$$

which shows the assertion. □

6 Linear-quadratic elliptic optimization problems II

We will now apply our results to the case of optimal control of Poisson's problem, e.g.,

$$\begin{aligned} \min J(q,u) &= \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } & u \in H_0^1(\Omega), \\ & (\nabla u, \nabla \varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ & q_{\min} \le q \le q_{\max} \quad \text{a.e. on } \Omega. \end{aligned} \tag{6.1}$$

We already know that the state equation defines a bounded linear operator $S\colon L^2(\Omega) \to L^2(\Omega)$, $q \mapsto Sq = u$. Hence, by Example 5.1.7, we know that

$$j'(\overline{q}) = S^*(S\overline{q} - u^d) + \alpha\overline{q}.$$

Then an application of Theorem 5.2.4 yields that for the unique solution of the reduced form of (6.1) it holds

$$(S^*(S\overline{q}-u^d) + \alpha\overline{q}, q-\overline{q}) \ge 0 \quad \forall\, q \in Q^{\text{ad}}$$

or, a little more concrete, for the solution $(\overline{q},\overline{u})$ of (6.1) it holds

$$\begin{aligned} (\nabla\overline{u}, \nabla\varphi) &= (\overline{q},\varphi) && \forall\, \varphi \in H_0^1(\Omega), \\ (S^*(\overline{u}-u^d) + \alpha\overline{q}, q-\overline{q}) &\ge 0 && \forall\, q \in Q^{\text{ad}}. \end{aligned}$$

In order to be able to evaluate the second line we need to understand what the operator S^* does.

Theorem 6.0.1. *The operator S given by the solution of (4.2) is self adjoint, e.g., $S^* = S$.*

Proof. To see this, we do the following computation where $q, p \in Q$ are arbitrary and $u_q = Sq$, $u_p = Sp$. Then we get

$$\begin{aligned} (S^*p, q) = (p, Sq) = (p, u_q) &= (\nabla u_p, \nabla u_q) \\ &= (\nabla u_q, \nabla u_p) = (q, u_p) \\ &= (q, Sp) = (Sp, q) \end{aligned}$$

for all $p, q \in Q$. □

In particular we obtain the following set of optimality conditions

Theorem 6.0.2. *Let $(\overline{q},\overline{u}) \in Q^{ad} \times H_0^1(\Omega)$ be a solution to (6.1). Then there exists some $\overline{z} \in H_0^1(\Omega)$ such that*

$$\begin{aligned} (\nabla\overline{u}, \nabla\varphi) &= (\overline{q},\varphi) && \forall\, \varphi \in H_0^1(\Omega), && (6.2) \\ (\nabla\overline{z}, \nabla\varphi) &= (\overline{u}-u^d,\varphi) && \forall\, \varphi \in H_0^1(\Omega), && (6.3) \\ (\overline{z} + \alpha\overline{q}, q-\overline{q}) &\ge 0 && \forall\, q \in Q^{ad}, && (6.4) \end{aligned}$$

Proof. The state equation (6.2) is clear. Further, we know from Theorem 6.0.1 that $S^* = S$ and hence if we define $\overline{z} \in H_0^1(\Omega)$ as a solution to (6.3) we get the desired relation $\overline{z} = S^*(\overline{u} - u^d)$ in $L^2(\Omega)$. This proofs the assertion. □

6.1 Formal Lagrange Calculus

The above computation of the adjoint equation is hardly intuitive. Hence we will give a more straight forward construction. To this end, we define the Lagrangian $\mathscr{L} : L^2(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega)$ by

$$\mathscr{L}(q,u,z) = J(q,u) + (q,z) - (\nabla u, \nabla z).$$

Then, without checking if this is allowed, we compute the (partial) derivatives of $\mathscr{L}$ and assume that

$$\mathscr{L}'_u(\overline{q},\overline{u},\overline{z}) = \mathscr{L}'_z(\overline{q},\overline{u},\overline{z}) = 0 \quad \text{in } H_0^1(\Omega)^*$$

as it would hold if we could apply Lagrange calculus (In fact we can and we will see this in the next chapter). This gives

$$\mathscr{L}'_u(\overline{q},\overline{u},\overline{z})\varphi = (\overline{u} - u^d, \varphi) - (\nabla\varphi, \nabla\overline{z}) = 0$$

for all $\varphi \in H_0^1(\Omega)$ which is exactly the form of the adjoint equation we derived earlier, but we note that it is much easier to obtain this equation this way.

6.2 Pointwise Discussion of the Variational Inequality

Now, we return to the (VI) (6.4). First, we remark that in fact the term $\overline{z} + \alpha\overline{q}$ is the $L^2(\Omega)$ representation of the gradient of the reduced cost functional j. To see this, we compute with $u = u(q), z = z(q) \in H_0^1(\Omega)$ given by (6.2) and (6.3)

$$\begin{aligned} j'(q)\delta q &= \mathscr{L}'_q(q,u,z)\delta q + \mathscr{L}'_u(q,u,z)S\delta q + \mathscr{L}'_z(q,u,z)S^*S\delta q \\ &= \mathscr{L}'_q(q,u,z)\delta q \\ &= (z + \alpha q, \delta q). \end{aligned}$$

We will now give some pointwise meaning to the (VI) (6.4).

Lemma 6.2.1. *Let $Q^{ad} \subset L^2(\Omega)$ be given by box constraints, e.g, $Q^{ad} = \{q \in L^2(\Omega) \,|\, q_{\min} \le q \le q_{\max}\}$. Then the variational inequality (6.4) holds if and only if it holds*

$$(\overline{z}(x) + \alpha\overline{q}(x))(q - \overline{q}(x)) \ge 0 \quad \forall\, q \in [q_{\min}, q_{\max}] \tag{6.5}$$

almost everywhere in Ω.

Proof. If the pointwise inequality holds so does the (VI) (6.4).

To see the converse, we note that $\overline{z}\,\overline{q}$, $\overline{q}^2$, $\overline{z}$, $\overline{q} \in L^1(\Omega)$. Hence by Lebesgue differentiation Theorem 2.6.5 almost all points in Ω are Lebesgue points of all these functions. Let now $x_0 \in \Omega$ be such a common Lebesgue point. For given $r > 0$ such that $B_r(x_0) \subset \Omega$ and $\lambda \in [0,1]$ we define a function $q_\lambda^r \in Q^{\mathrm{ad}}$ as follows

$$q_\lambda^r(x) = \begin{cases} \lambda q_{\min} + (1-\lambda) q_{\max} & x \in B_r(x_0), \\ \overline{q}(x) & \text{otherwise.} \end{cases}$$

Then we obtain from (6.4) that

$$\begin{aligned} 0 &\le \frac{1}{|B_r(x_0)|}(\overline{z} + \alpha\overline{q}, q_\lambda^r - \overline{q}) \\ &= \frac{1}{|B_r(x_0)|} \int_{B_r(x_0)} (\overline{z} + \alpha\overline{q})(q_\lambda^r - \overline{q})\, dx \\ &= \frac{1}{|B_r(x_0)|} \int_{B_r(x_0)} (\overline{z} + \alpha\overline{q})(\lambda q_{\min} + (1-\lambda) q_{\max} - \overline{q})\, dx. \end{aligned}$$

By taking the limit $r \to 0$ we get

$$0 \le (\overline{z}(x) + \alpha\overline{q}(x))(\lambda q_{\min} + (1-\lambda) q_{\max} - \overline{q}(x)).$$

The term $\lambda q_{\min} + (1-\lambda) q_{\max}$ can now take any value in $[q_{\min}, q_{\max}]$ by varying $\lambda \in [0,1]$. This yields the assertion. □

From this one immediately gets that the optimal control $\overline{q}$ satisfies

$$\begin{cases} \overline{q}(x) = q_{\min} & \text{if } \overline{z}(x) + \alpha\overline{q}(x) > 0, \\ \overline{q}(x) \in [q_{\min}, q_{\max}] & \text{if } \overline{z}(x) + \alpha\overline{q}(x) = 0, \\ \overline{q}(x) = q_{\max} & \text{if } \overline{z}(x) + \alpha\overline{q}(x) < 0. \end{cases} \tag{6.6}$$

This yields a pointwise representation of $\overline{q}$ if $\alpha > 0$, namely

$$\overline{q}(x) = \max(q_{\min}, \min(q_{\max}, -1/\alpha\overline{z}(x)) = \mathscr{P}_{[q_{\min}, q_{\max}]}(-1/\alpha\overline{z}(x)).$$

Remark 6.2.2. We note that the (VI) (6.4) can be rewritten for $\alpha > 0$ as

$$(1/\alpha\overline{z} + \overline{q}, \delta q - \overline{q}) \ge 0 \quad \forall \delta q \in Q^{\mathrm{ad}}.$$

By comparing this with Lemma 2.2.4 we see that in fact

$$\overline{q} = \mathscr{P}_{Q^{\mathrm{ad}}}(-1/\alpha\overline{z})$$

for which we have computed a pointwise representation above.

However, in the case $\alpha = 0$ the optimal control is not completely determined by the pointwise variational inequality (6.5). If in this case the set $\{x \in \Omega \,|\, \bar{z}(x) = 0\}$ has measure zero, then again (6.6) defines $\bar{q}$ which then has values in $\{q_{\min}, q_{\max}\}$ almost everywhere. This is then called a bang-bang control.

6.3 Lagrange Multipliers for Box-Control-Constraints

Let us assume that the conditions in Theorem 6.0.2 are fulfilled. This means $(\bar{q}, \bar{u}) \in Q^{\text{ad}} \times H_0^1(\Omega)$ solves the problem (6.1) if and only if there exists some $\bar{z} \in H_0^1(\Omega)$ such that:

$$\begin{aligned} (\nabla \bar{u}, \nabla \varphi) &= (\bar{q}, \varphi) && \forall\, \varphi \in H_0^1(\Omega), \\ (\nabla \bar{z}, \nabla \varphi) &= (\bar{u} - u^d, \varphi) && \forall\, \varphi \in H_0^1(\Omega), \\ (\bar{z} + \alpha \bar{q}, q - \bar{q}) &\geq 0 && \forall\, q \in Q^{\text{ad}}. \end{aligned}$$

We will now derive some Lagrange-multiplier for the control constraints.

Theorem 6.3.1. *Let $\bar{q} \in Q^{ad}$ be given, then it is equivalent*

1. *The variational inequality (6.4) (or (6.5), (6.6)) is satisfied.*
2. *There exist functions $\mu_{\min}, \mu_{\max} \in L^2(\Omega)$ such that*

$$\begin{aligned} \mu_{\min} &\geq 0, \\ \mu_{\max} &\geq 0, \\ \bar{z} + \alpha \bar{q} + \mu_{\max} - \mu_{\min} &= 0, \\ \mu_{\min}(x)(q_{\min} - \bar{q}(x)) &= 0 \quad \text{a.e. in } \Omega, \\ \mu_{\max}(x)(\bar{q}(x) - q_{\max}) &= 0 \quad \text{a.e. in } \Omega, \end{aligned}$$

(Note: the last two equations are called complementarity conditions.)

Proof. To see that 1. implies 2. we define pointwise almost everywhere

$$\begin{aligned} \mu_{\min} &= \max(0, \bar{z} + \alpha \bar{q}) = (\bar{z} + \alpha \bar{q})^+, \\ \mu_{\max} &= -\min(0, \bar{z} + \alpha \bar{q}) = (\bar{z} + \alpha \bar{q})^-. \end{aligned}$$

By this definition the first three conditions are clearly satisfied and $\mu_{\min}, \mu_{\max} \in L^2(\Omega)$.

To see the complementarity conditions we employ (6.6) to get the following cases (considered pointwise a.e.):

$\bar{z} + \alpha \bar{q} > 0$ Then $\bar{q} = q_{\min}$ and $\mu_{\max} = 0$ which asserts the complementarity.

$\bar{z} + \alpha \bar{q} < 0$ Then $\bar{q} = q_{\max}$ and $\mu_{\min} = 0$ which asserts the complementarity.

$\bar{z} + \alpha \bar{q} = 0$ Then $\mu_{\min} = \mu_{\max} = 0$ which asserts the complementarity.

This shows the first implication.

Now, to see the converse relation, let 2. be satisfied. We need to show that (6.5) is satisfied. To this end, let $q \in [q_{\min}, q_{\max}]$ be arbitrary. Then we conclude again using case-by-case analysis

$q_{\min} < \overline{q} < q_{\max}$ From complementarity we get that $\mu_{\min} = \mu_{\max} = 0$. Thus the third condition yields $\overline{z} + \alpha\overline{q} = 0$. Hence

$$(\overline{z} + \alpha\overline{q})(q - \overline{q}) = 0.$$

$q_{\min} = \overline{q}$ Then it holds $q - \overline{q} \geq 0$. As $q_{\min} < q_{\max}$ complementarity yields $\mu_{\max} = 0$. Finally, by the third relation we have

$$\overline{z} + \alpha\overline{q} = \mu_{\min} \geq 0$$

and thus

$$(\overline{z} + \alpha\overline{q})(q - \overline{q}) \geq 0.$$

$q_{\max} = \overline{q}$ Is analog to the case $q_{\min} = \overline{q}$.

□

With this we can summarize the necessary conditions in form of a Karush-Kuhn-Tucker system (KKT-system) as follows:

$$\begin{aligned}
(\nabla\overline{u}, \nabla\varphi) &= (\overline{q}, \varphi) && \forall\, \varphi \in H_0^1(\Omega) \\
(\nabla\varphi, \nabla\overline{z}) &= (\overline{u} - u^d, \varphi) && \forall\, \varphi \in H_0^1(\Omega) \\
\overline{z} + \alpha\overline{q} + \mu_{\max} - \mu_{\min} &= 0 && \text{a.e. in } \Omega \\
\mu_{\min} \geq 0, \quad q_{\min} - \overline{q} &\leq 0, && \mu_{\min}(q_{\min} - \overline{q}) = 0 \text{ a.e. in } \Omega \\
\mu_{\max} \geq 0, \quad \overline{q} - q_{\max} &\leq 0, && \mu_{\max}(\overline{q} - q_{\max}) = 0 \text{ a.e. in } \Omega
\end{aligned}$$

6.3.1 Regularity of Solutions

Finally we use the KKT-conditions to show that the optimal solutions indeed have higher regularity.

Theorem 6.3.2. *Let $u^d \in H^1(\Omega)$ and Ω of class $C^{2,1}$ and $\alpha > 0$. Then the solution $\overline{q}$ to (6.1) is in $W^{1,\infty}(\Omega)$ and $\overline{u} \in H^3(\Omega) \cap H_0^1(\Omega)$.*

Proof. By assumption we have $\overline{u} - u^d \in H^1(\Omega)$. Then by the elliptic regularity, see Section 4.2, we have that $\overline{z} \in H^3(\Omega)$ and thus the embedding Theorem 2.6.15 and Theorem 2.6.16 we have $\overline{z} \in H^3(\Omega) \subset C^{0,1}(\overline{\Omega}) = W^{1,\infty}(\Omega)$. Now, by Remark 6.2.2, we know

$$\overline{q} = \mathscr{P}_{Q^{\text{ad}}}(-1/\alpha\overline{z})$$

and because $\mathscr{P}_{Q^{\text{ad}}} : W^{1,\infty}(\Omega) \to W^{1,\infty}(\Omega)$ is continuous we get $\overline{q} \in W^{1,\infty}(\Omega)$.

Finally, because $W^{1,\infty}(\Omega) \subset H^1(\Omega)$ we get that $\overline{u} \in H^3(\Omega)$. □

7 Discretization

We are now concerned with the discretization of the infinite dimensional optimization problem (6.1). To do so, we will briefly review the finite element method which will be used to discretize the state equation. To this end, we consider the model problem to find $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \tag{7.1}$$

Then a generic way to discretize this equation is to choose some finite dimensional space V_h (for convenience let $V_h \subset H_0^1(\Omega)$ then V_h is called H_0^1-conform). Then one considers the problem to find $u_h \subset V_h$ such that

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \tag{7.2}$$

In particular, one searches for a solution u_h in the ansatz space V_h such that the variational equation is satisfied for all φ_h in the test space V_h (Although here they are the same.). The equation (7.2) has a unique solution by the same arguments that are used for (7.1).

Theorem 7.0.1. *Let u be the solution to (7.1) and u_h be the solution to (7.2). Then the following best-approximation property holds*

$$\|\nabla(u - u_h)\| \leq \inf_{\varphi_h \in V_h} \|\nabla(u - \varphi_h)\|.$$

Proof. The proof follows from the Galerkin orthogonality, e.g.,

$$(\nabla(u - u_h), \nabla \varphi_h) = 0 \quad \forall \varphi \in V_h.$$

□

By choosing a basis $\varphi_h^{(i)}$, $i = 1, \ldots, N$ of V_h the problem (7.2) an be rewritten as a linear algebraic problem as follows: Find $\mathbf{u_h} = (u_h^{(i)})_{i=1,\ldots,N}$ such that

$$\mathbf{A_h u_h} = \mathbf{f_h}$$

with the stiffness matrix

$$\mathbf{A_h} = (a_{ij})_{i,j=1,\ldots,N}, \quad a_{ij} = (\nabla \varphi_h^{(j)}, \nabla \varphi_h^{(i)})$$

and the right hand side

$$\mathbf{f_h} = ((f, \varphi_h^{(i)}))_{i=1,\ldots,N}.$$

Now, in order to be able to solve this problem efficiently one needs to choose the basis in a clever way, in particular such that it is „easy" to invert $\mathbf{A_h}$. The idea to choose an orthogonal basis is usually not computationally feasible, and may not be possible for general equations. One possible way to obtain a reasonably well behaved matrix $\mathbf{A_h}$ is given by the so called finite element method.

7.1 Linear Finite Elements for Elliptic Problems

For simplicity let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. Then we consider a sequence of decompositions (triangulations) $\mathcal{T}_h = T$ of $\overline{\Omega}$ into closed polygons T with $h := \max_{T\in\mathcal{T}_h} \operatorname{diam}(T) \to 0$. For simplicity let us assume that there exists some reference element $\widehat{T}$ such that any element $T \in \mathcal{T}_h$ is given by an affine-linear transformation of $\widehat{T}$. Typical elements T are triangles or quadrilaterals.

We require the following regularity properties for $\mathcal{T}_h$.

1. (Structural regularity or feasibility) Any two different elements T_1, $T_2 \in \mathcal{T}_h$ intersect at most in a vertex or along an entire edge.

2. (Shape regularity) For the radius ρ_T of the incircle and h_T of the circumscribed circle of an element $T \in \mathcal{T}_h$ it holds

$$\max_{T\in\mathcal{T}_h} \frac{h_T}{\rho_T} \leq c$$

   uniformly in $h \to 0$.

3. (Size regularity) It holds

$$\max_{T\in\mathcal{T}_h} h_T \leq c \min_{T\in\mathcal{T}_h} h_T$$

   uniformly in T.

We can now describe the construction of a finite dimensional space V_h. For simplicity let us assume that the decompositions $\mathcal{T}_h$ consist of triangles only. Then we can define

$$V_h^{(1)} := \{v_h \in C(\overline{\Omega}) \mid v_h\big|_T \in P_1(T),\ T \in \mathcal{T}_h,\ v_h\big|_\Gamma = 0\}.$$

where $P_1(T)$ denotes the space of polynomials of degree ≤ 1 on T. This defines the space of (piecewise) linear finite elements. It is straightforward to see that $V_h^{(1)} \subset H_0^1(\Omega)$.

Now to define a suitable basis of $V_h^{(1)}$ we denote the interior vertices of $\mathcal{T}_h$ by x_i for $i = 1, \ldots, N$. Then we define the nodal basis $\varphi_h^{(i)}$ (or Lagrange basis) as follows:

$$\varphi_h^{(i)}(x_j) = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Remark 7.1.1. Of course higher order polynomials are also possible, e.g., piecewise polynomials of degree two. Which then require additional basis functions, for example midpoints of the edges.

On quadrilateral elements the definition of piecewise polynomials is not satisfactory hence on considers usually functions that are piecewise bilinear (biquadratic, ...). However the definition on general quadrilaterals is more involved, as one needs to account for the transformation from the reference element (which may be non affine!).

As we have already seen, the error in the natural norm is given by the best approximation property. Hence to obtain error estimates it is natural to ask what is the best possible approximation. To answer this, we consider the Lagrange interpolation

Definition 7.1.2. For a function $u \in C(\overline{\Omega})$ we define the Lagrange interpolation $I_h u \colon C(\overline{\Omega}) \to V_h^{(1)}$ as follows

$$I_h u(x) = \sum_{i=1}^{N} u(x_i)\varphi_h^{(i)}(x).$$

Theorem 7.1.3. *Let $\mathcal{T}_h$ satisfy the regularity assumptions above. Then there exists a constant $c > 0$ independent of h such that for $u \in W^{2,p}(\Omega)$ it holds*

$$\|u - I_h u\|_p + h\|\nabla(u - I_h u)\|_p \leq ch^2\|u\|_{2,p}$$

for $1 < p \leq \infty$.

Theorem 7.1.4. *Let $\mathcal{T}_h$ satisfy the regularity assumptions above, and let Ω be convex. Then the solution u to (7.1) and u_h to (7.2) satisfy for $f \in L^2(\Omega)$*

$$\|u - u_h\| + h\|\nabla(u - u_h)\| \leq ch^2\|f\|.$$

Proof. From Theorem 7.0.1 and Theorem 7.1.3 we have that

$$\|\nabla(u - u_h)\| \leq ch\|u\|_{2,2}$$

and by the elliptic regularity results in Section 4.2 it follows

$$\|\nabla(u - u_h)\| \leq ch\|f\|.$$

For the error in the L^2-Norm we employ the Aubin-Nitsche trick. Define $e_h = u - u_h$ and $p \in H_0^1(\Omega)$ as solution to

$$(\nabla\varphi, \nabla p) = 1/\|e_h\|(e_h, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega).$$

Then, because $e_h \in L^2(\Omega)$ it is $p \in H^2(\Omega)$ and it holds

$$\|p\|_{2,2} \leq c\|e_h/\|e_h\|\| = c.$$

Then we obtain

$$\|e_h\| = (\nabla e_h, \nabla p) = (\nabla e_h, \nabla(p - I_h p)) \leq \|\nabla e_h\|\|\nabla(p - I_h p)\| \leq ch\|\nabla e_h\|$$

hence the assertion follows. □

Remark 7.1.5. The above theorem states that the operator $S\colon L^2(\Omega) \to L^2(\Omega)$ given by (7.1) and the discrete operator $S_h\colon L^2(\Omega) \to L^2(\Omega)$ given by (7.2) satisfy

$$\|(S - S_h)f\| \leq ch^2\|f\| \quad \forall\, f \in L^2(\Omega)$$

or equivalently

$$\|S - S_h\|_{\mathcal{L}(L^2(\Omega))} \leq ch^2.$$

7.2 Discretization of the Model Problem

We return to the model problem

$$\begin{aligned} \min J(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } u \in H_0^1(\Omega), \\ (\nabla u, \nabla\varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ q_{\min} \leq q \leq q_{\max} \quad \text{a.e. on } \Omega. \end{aligned} \tag{7.3}$$

Here we assume that $\Omega \subset \mathbb{R}^2$ is a convex, bounded polygonal domain. Further we assume that we have a sequence of feasible and (shape and size) regular triangulations $\mathcal{T}_h$ of Ω.

7.2.1 Variational Discretization

Now we can replace the state equation using the finite element method on a sequence of feasible and (shape and size) regular triangulations. This gives

$$\begin{aligned} \min J(q_h,u_h) = \frac{1}{2}\|u_h - u^d\|^2 + \frac{\alpha}{2}\|q_h\|^2 \\ \text{s.t. } u_h \in V_h, \\ (\nabla u_h, \nabla\varphi_h) = (q_h,\varphi_h) \quad \forall\, \varphi_h \in V_h, \\ q_{\min} \leq q_h \leq q_{\max} \quad \text{a.e. on } \Omega. \end{aligned} \tag{7.4}$$

We note that we did not discretize the control space, and hence this is (at least formally) an infinite dimensional problem.

Theorem 7.2.1. *For $\alpha > 0$ there exists a unique solution $(\overline{q}_h, \overline{u}_h) = (\overline{q}_h, S_h\overline{q}_h) \in Q^{ad} \times V_h$ to (7.4). This solution is (equivalently) characterized by the existence of $\overline{z}_h \in V_h$ such that the following holds:*

$$\begin{aligned} (\nabla\overline{u}_h, \nabla\varphi_h) &= (\overline{q}_h,\varphi) && \forall\, \varphi_h \in V_h, \\ (\nabla\varphi_h, \nabla\overline{z}_h) &= (\overline{u}_h - u^d, \varphi_h) && \forall\, \varphi_h \in V_h, \\ (\overline{z}_h + \alpha\overline{q}_h, q - \overline{q}_h) &\geq 0 && \forall\, q \in Q^{ad}. \end{aligned}$$

Proof. We note that the operator $S_h\colon L^2(\Omega) \to V_h \subset L^2(\Omega)$ is linear and continuous. Hence the reduced cost functional

$$j_h(q_h) = J(q_h, S_h q_h)$$

is convex and continuous and thus w.l.s.c. with $j_h(q_h) \to \infty$ as $\|q_h\| \to \infty$. Thus we can apply Theorem 3.0.1 to obtain existence of a solution $(\overline{q}_h, \overline{u}_h)$ to (7.4). Due to strict convexity of j_h this solution is unique.

Now as for j the functional j_h is continuously Fréchet differentiable and hence by Theorem 5.2.4 the solution $\overline{q}_h$ to

$$\min_{q_{\min} \le q_h \le q_{\max}} j_h(q_h)$$

is equivalently characterized by the variational inequality

$$(S_h^*(S_h\overline{q}_h - u^d) + \alpha\overline{q}_h, q - \overline{q}_h) \ge 0 \quad \forall q \in Q^{\mathrm{ad}}.$$

Now, to conclude we need to derive an equation that is defining S_h^*. To this end, we compute for $f, g \in L^2(\Omega)$

$$(S_h^* g, f) = (g, S_h f) = (\nabla S_h g, \nabla S_h f) = (\nabla S_h f, \nabla S_h g) = (f, S_h g)$$

Thus $S_h^* = S_h$ and the assertion follows. □

Remark 7.2.2. We remark that the optimality conditions obtained here are the discretized optimality conditions of the continuous case. This means that it does not matter whether we discretize the optimization problem and compute necessary optimality conditions or if we directly discretize the optimality conditions. This is often refered to as 'optimize-then-discretize = discretize-then-optimize'. This is not always the case, e.g., in finite difference schemes this need not happen.

Theorem 7.2.3. *Let $(\overline{q}, \overline{u})$ be the solution to (7.3) and $(\overline{q}_h, \overline{u}_h)$ be the solution to (7.4). Then there exists a constant (independent of h and α) such that*

$$\alpha\|\overline{q} - \overline{q}_h\|^2 + \|\overline{u} - \overline{u}_h\|^2 \le c\Big(1 + \frac{1}{\alpha}\Big)h^4.$$

Proof. We recall the necessary optimality condition for $\overline{q}$ and $\overline{q}_h$ namely

$$(S^*(S\overline{q} - u^d) + \alpha\overline{q}, q - \overline{q}) \ge 0 \quad \forall q \in Q^{\mathrm{ad}},$$
$$(S_h^*(S_h\overline{q}_h - u^d) + \alpha\overline{q}_h, q - \overline{q}_h) \ge 0 \quad \forall q \in Q^{\mathrm{ad}}.$$

Now we see that $\overline{q}_h$ is a feasible test function for the first inequality while $\overline{q}$ is a feasible test function for the second inequality. Together this gives

$$\begin{aligned}
0 \leq\ & (S^*(S\overline{q}-u^d)-S_h^*(S_h\overline{q}_h-u^d)+\alpha\overline{q}-\alpha\overline{q}_h,\overline{q}_h-\overline{q})\\
=\ & -\alpha\|\overline{q}_h-\overline{q}\|^2\\
& +(S\overline{q}-u^d,S(\overline{q}_h-\overline{q}))-(S_h\overline{q}_h-u^d,S_h(\overline{q}_h-\overline{q}))\\
=\ & -\alpha\|\overline{q}_h-\overline{q}\|^2\\
& +(S\overline{q}-S_h\overline{q}_h,S_h(\overline{q}_h-\overline{q}))+(S\overline{q}-u^d,(S-S_h)(\overline{q}_h-\overline{q}))\\
=\ & -\alpha\|\overline{q}_h-\overline{q}\|^2\\
& +(S\overline{q}-S_h\overline{q}_h,S_h\overline{q}_h-S\overline{q})+(S\overline{q}-S_h\overline{q}_h,S\overline{q}-S_h\overline{q})+(S\overline{q}-u^d,(S-S_h)(\overline{q}_h-\overline{q}))\\
=\ & -\alpha\|\overline{q}_h-\overline{q}\|^2-\|\overline{u}-\overline{u}_h\|^2\\
& +(S\overline{q}-S_h\overline{q}_h,(S-S_h)\overline{q})+((S-S_h)^*(S\overline{q}-u^d),\overline{q}_h-\overline{q})
\end{aligned}$$

We obtain, using $(S-S_h)^*=S^*-S_h^*$, Young's inequality, and a generic constant c (which may be different in each line of the inequality),

$$\begin{aligned}
\alpha\|\overline{q}_h-\overline{q}\|^2+\|\overline{u}-\overline{u}_h\|^2 &\leq \|S\overline{q}-S_h\overline{q}_h\|\|(S-S_h)\overline{q}\|+\|(S^*-S_h^*)(S\overline{q}-u^d)\|\|\overline{q}_h-\overline{q}\|\\
&\leq (\|(S-S_h)\overline{q}\|+\|S_h(\overline{q}-\overline{q}_h)\|)\|(S-S_h)\overline{q}\|\\
&\quad +\|(S^*-S_h^*)(S\overline{q}-u^d)\|\|\overline{q}_h-\overline{q}\|\\
&\leq ch^2\big(\|(S-S_h)\overline{q}\|+\|S_h(\overline{q}-\overline{q}_h)\|+\|\overline{q}_h-\overline{q}\|\big)\\
&\leq ch^4+ch^2\|\overline{q}_h-\overline{q}\|)\\
&\leq ch^4+\frac{c}{\alpha}h^4+\frac{\alpha}{2}\|\overline{q}_h-\overline{q}\|^2
\end{aligned}$$

By subtracting the last term, we get

$$\alpha\|\overline{q}_h-\overline{q}\|^2+\|\overline{u}-\overline{u}_h\|^2\leq c(1+\frac{1}{\alpha})h^4$$

which proofs the assertion. □

Remark 7.2.4. It is clear that if Q is a finite dimensional space, e.g., we have only finitely many control variables, then this discretization is sufficient to obtain a finite dimensional problem.

However, this technique yields a discrete optimal control problem even if Q is not finite dimensional. To see this one can again consider the case of pure control constraints, in this case the variational inequality

$$(\overline{z}_h+\alpha\overline{q}_h,q-\overline{q}_h)\geq 0\quad\forall q\in Q^{\mathrm{ad}}.$$

holds again pointwise almost everywhere, using the same arguments as in Lemma 6.2.1. yielding that

$$\overline{q}_h=\mathscr{P}_{Q^{\mathrm{ad}}}\Big(\frac{-1}{\alpha}\overline{z}_h\Big).$$

This can then be used to eliminate the control in the optimality conditions given by Theorem 7.2.1. This means that

$$\left(\mathscr{P}_{Q^{\mathrm{ad}}}\left(\frac{-1}{\alpha}\bar{z}_h\right), \bar{u}_h\right)$$

solves (7.4) if and only if $(\bar{u}_h, \bar{z}_h)$ solves the following nonlinear (and non smooth) problem

$$\begin{aligned}
(\nabla \bar{u}_h, \nabla \varphi_h) &= (\mathscr{P}_{Q^{\mathrm{ad}}}(-1/\alpha \bar{z}_h), \varphi) && \forall\, \varphi_h \in V_h,\\
(\nabla \varphi_h, \nabla \bar{z}_h) &= (\bar{u}_h - u^d, \varphi_h) && \forall\, \varphi_h \in V_h.
\end{aligned}$$

This technique is useful in particular if one uses simple constraints such as box-constraints when the projection can be evaluated easily.

This technique has been advertised in [21], see also [22]

7.2.2 Full Discretization

Finally, we want to discuss the full discretization of the optimization problem (7.3). To do so, we need to discretize the space Q. To this end, we consider the spaces

$$\begin{aligned}
Q_h^{(1)} &= \{q \in C(\overline{\Omega}) \,|\, q\big|_T \in P_1(T),\ T \in \mathscr{T}_h\},\\
Q_h^{(0)} &= \{q \in L^2(\Omega) \,|\, q\big|_T \in P_0(T),\ T \in \mathscr{T}_h\}.
\end{aligned}$$

Whether this gives a finite dimensional problem depends upon the set Q^{ad}, i.e., whether $Q^{\mathrm{ad},h} = Q^{\mathrm{ad}} \cap Q_h$ can be described by finitely many constraints. For simplicity, we set $Q^{\mathrm{ad,h}} = Q^{\mathrm{ad}} \cap Q_h$ – with $Q_h = Q_h^{(0)}$ or $Q_h = Q_h^{(1)}$ depending on the situation–. Now, if we assume to have box constraints, for $q_h \in Q_h^{(0)}$ it holds

$$q_h \in Q^{\mathrm{ad,h}} \text{ iff } q_h\big|_T \equiv \mathrm{const} \in [q_{\min}, q_{\max}] \quad \forall\, T \in \mathscr{T}_h.$$

To make this more clear, let $\mathscr{T}_h = \{T_i,\ i = 1, \ldots, M\}$. Then, we obtain a basis of $Q_h^{(0)}$ by defining

$$\hat{\varphi}_h^{(i)}\big|_{T_j} \equiv \begin{cases} 0 & i \neq j,\\ 1 & i = j. \end{cases}$$

In the case $q_h \in Q_h^{(1)}$, we need to argue a little more. Let $\{\hat{x}_i,\ i = 1, \ldots, M\}$ denote the set of vertices of $\mathscr{T}_h$, including those on the boundary, i.e., $M \geq N$. Then, the Lagrange basis $\hat{\varphi}_h^{(i)}$ of $Q_h^{(1)}$ is defined analog to the basis for $V_h^{(1)}$ by

$$\hat{\varphi}_h^{(i)}(\hat{x}_i) = \begin{cases} 0 & i \neq j,\\ 1 & i = j. \end{cases}$$

Now, $\hat{\varphi}_h^{(i)} \geq 0$ and $\sum_{i=1}^{M} \hat{\varphi}_h^{(i)} \equiv 1$. This gives that

$$q_{\min} \leq q_h \leq q_{\max} \text{ a.e. in } \Omega \text{ iff } q_h(\hat{x}_i) \in [q_{\min}, q_{\max}] \quad \forall\, i = 1, \ldots, M.$$

Note that this is no longer true for quadratic or higher order polynomials.

Now, in both cases, we can write $q_h = \sum_{i=1}^{M} q_h^{(i)} \hat{\varphi}_h^{(i)}$ and obtain the following fully discrete problem

$$\begin{aligned} \min J(q_h, u_h) = \frac{1}{2}\|u_h - u^d\|^2 + \frac{\alpha}{2}\|q_h\|^2 \\ \text{s.t. } u_h \in V_h, \\ (\nabla u_h, \nabla \varphi_h) = (q_h, \varphi_h) \quad \forall\, \varphi_h \in V_h, \\ q_{\min} \le q_h^{(i)} \le q_{\max} \quad \forall\, i = 1, \dots, M. \end{aligned} \tag{7.5}$$

With the identity $u_h = \sum_{i=1}^{N} u_h^{(i)} \varphi_h^{(i)}$ we get the equivalent quadratic program to determine $\mathbf{u_h} = (u_h^{(i)})_{i=1}^N$ and $\mathbf{q_h} = (q_h^{(i)})_{i=1}^M$

$$\begin{aligned} \min \frac{1}{2}\mathbf{u_h}^T \mathbf{M_1} \mathbf{u_h} - \mathbf{u_h}^T \mathbf{u^d} + \frac{\alpha}{2}\mathbf{q_h}^T \mathbf{M_2} \mathbf{q_h} \\ \text{s.t. } \mathbf{A_h u_h} = \mathbf{M_3 q_h}, \\ q_{\min} \le q_h^{(i)} \le q_{\max} \quad \forall\, i = 1, \dots, M. \end{aligned} \tag{7.6}$$

with the matrices

$$\mathbf{M_1} = \left((\varphi_h^{(i)}, \varphi_h^{(j)}\right)_{i,j=1}^N, \quad \mathbf{M_2} = \left((\hat{\varphi}_h^{(i)}, \hat{\varphi}_h^{(j)}\right)_{i,j=1}^M \quad \mathbf{M_3} = \left((\hat{\varphi}_h^{(j)}, \varphi_h^{(i)}\right)_{i,j=1}^{N,M}$$

and the data

$$\mathbf{u^d} = \left((\varphi_h^{(i)}, u^d\right)_{i=1}^N$$

This form can than readily be solved using standard optimization libraries.

Theorem 7.2.5. *For $\alpha > 0$ there exists a unique solution $(\bar{q}_h, \bar{u}_h) = (\bar{q}_h, S_h \bar{q}_h) \in Q^{ad,h} \times V_h$ to (7.5). This solution is (equivalently) characterized by the existence of $\bar{z}_h \in V_h$ such that the following holds:*

$$\begin{aligned} (\nabla \bar{u}_h, \nabla \varphi_h) &= (\bar{q}_h, \varphi) && \forall\, \varphi_h \in V_h, \\ (\nabla \varphi_h, \nabla \bar{z}_h) &= (\bar{u}_h - u^d, \varphi_h) && \forall\, \varphi_h \in V_h, \\ (\bar{z}_h + \alpha \bar{q}_h, q - \bar{q}_h) &\ge 0 && \forall\, q \in Q^{ad,h}. \end{aligned}$$

Proof. We note that the set $Q^{\mathrm{ad,h}} = Q^{\mathrm{ad}} \cap Q_h$ is closed and convex. Thus the assertion follows as in the proof of Theorem 7.2.1. □

Lemma 7.2.6. *For the L^2-projection, $P_h \colon L^2(\Omega) \to Q_h^{(0)}$ there exists a constant c such that for any element $T \in \mathcal{T}_h$ and $u \in H^1(T)$ it holds*

$$\|u - P_h u\|_T^2 \le c h^2 \|\nabla u\|_T^2$$

and $P_h u\big|_T = |T|^{-1} \int_T u \, dx$.

Although one can proof the a priori estimate by some lengthy calculations, as for the variational discretization with some additional tricks, we will split the proof into several sub steps. To begin with, we fix some notation. The discrete state and adjoint equation define a mapping $z_h : L^2(\Omega) \to V_h$ by the assignment

$$q \mapsto S_h q = u_h(q) \mapsto S_h^*(S_h q - u^d) = z_h(q).$$

Similarly, the continuous solution operators define a mapping $z : L^2(\Omega) \to V = H_0^1(\Omega)$ by the assignment

$$q \mapsto Sq = u(q) \mapsto S^*(Sq - u^d) = z(q).$$

Then, we have the reduced cost functionals $j, j_h : L^2(\Omega) \to \mathbb{R}$. defined as

$$j(q) = J(q, Sq), \quad j_h(q) = J(q, S_h q).$$

As we have seen they are Fréchet differentiable with

$$\begin{aligned} j'(q)\delta q &= (u(q) - u^d, u(\delta q)) + \alpha(q, \delta q) = (z(q) + \alpha q, \delta q), \\ j_h'(q)\delta q &= (u_h(q) - u^d, u_h(\delta q)) + \alpha(q, \delta q) = (z_h(q) + \alpha q, \delta q). \end{aligned}$$

We can show the following simple properties of these derivatives

Lemma 7.2.7. *There exists a constant $c > 0$ such that for any $\alpha \in (0, 1]$ and elements $p, q, r \in L^2(\Omega)$ the following holds:*

1. *(Discretization Error)*

$$|j'(q)r - j_h'(q)r| \le ch^2(1 + \|q\|)\|r\|$$

2. *(Lipschitz continuity)*

$$|j_h'(q)r - j_h'(p)r| \le c\|p - q\|\|r\|$$

Proof. We start to show 1. To this end, we use the representation of the derivatives to get

$$\begin{aligned} |j'(q)r - j_h'(q)r| &= |(z(q) + \alpha q - z_h(q) - \alpha q, r)| \\ &\le \|z(q) - z_h(q)\|\|r\| \\ &\le \|S^*(Sq - u^d) - S_h^*(S_h q - u^d)\|\|r\|. \end{aligned}$$

Now using the L^2-error estimate for $S - S_h$ in Theorem 7.1.4 we conclude

$$\begin{aligned} |j'(q)r - j_h'(q)r| &\le \|S^*(Sq - u^d) - S_h^*(S_h q - u^d)\|\|r\| \\ &\le \Big(\|(S^* - S_h^*)u^d\| + \|S^*Sq - S_h^*Sq\| + \|S_h^*Sq - S_h^*S_h q\|\Big)\|r\| \\ &\le \Big(ch^2\|u^d\| + ch^2\|Sq\| + c\|Sq - S_h q\|\Big)\|r\| \\ &\le \Big(ch^2\|u^d\| + ch^2\|Sq\| + c\|Sq - S_h q\|\Big)\|r\| \\ &\le ch^2(1 + \|q\|)\|r\| \end{aligned}$$

which shows the first assertion.

To see the second claim, we compute

$$\begin{aligned}|j_h'(q)r - j_h'(p)r| &= |(z_h(q) + \alpha q - z_h(p) - \alpha q, r)| \\ &\leq \Big(\|z_h(q) - z_h(p)\| + \alpha\|q-p\|\Big)\|r\| \\ &= \Big(\|S_h^*(S_h q - u^d - S_h p + u^d)\| + \alpha\|q-p\|\Big)\|r\| \\ &\leq \Big(c\|q-p\| + \alpha\|q-p\|\Big)\|r\|\end{aligned}$$

which shows the assertion. □

In the next step, we will now use these estimates to derive an upper bound for the error in the control.

Theorem 7.2.8. *Let $\overline{q}$ be the optimal control of problem (7.3) and $\overline{q}_h$ be the optimal control of problem (7.5). Then there exists a constant $c > 0$ depending on $\|\overline{q}\|$ such that for $\alpha \in (0,1]$*

$$\|\overline{q} - \overline{q}_h\| \leq c\Big(\frac{h^2}{\alpha} + (1 + \frac{1}{\alpha}) \inf_{p_h \in \widehat{Q}_h} \|\overline{q} - p_h\|\Big)$$

where the set $\widehat{Q}_h \subset Q^{ad,h}$ is given by

$$\widehat{Q}_h = \{p_h \in Q^{ad,h} \mid j'(\overline{q})(\overline{q}_h - p_h) \geq 0\}.$$

Proof. We start by taking $p_h \in \widehat{Q}_h$ arbitrary. Then, we obtain from the form of the derivative j_h' that

$$\begin{aligned}\alpha\|\overline{q}_h - p_h\|^2 &\leq (u_h(\overline{q}_h) - u_h(p_h) - u^d + u^d, u_h(\overline{q}_h) - u_h(p_h)) + \alpha(\overline{q}_h - p_h, \overline{q}_h - p_h) \\ &= j_h'(\overline{q}_h)(\overline{q}_h - p_h) - j_h'(p_h)(\overline{q}_h - p_h).\end{aligned}$$

Due to the discrete optimality condition the first term is non positive and we get, using the definition of $\widehat{Q}_h$,

$$\begin{aligned}\alpha\|\overline{q}_h - p_h\|^2 &\leq -j_h'(p_h)(\overline{q}_h - p_h) \\ &\leq j'(\overline{q})(\overline{q}_h - p_h) - j_h'(\overline{q})(\overline{q}_h - p_h) + j_h'(\overline{q})(\overline{q}_h - p_h) - j_h'(p_h)(\overline{q}_h - p_h).\end{aligned}$$

Applying Lemma 7.2.7, we conclude

$$\alpha\|\overline{q}_h - p_h\|^2 \leq ch^2(1 + \|\overline{q}\|)\|\overline{q}_h - p_h\| + c\|\overline{q} - p_h\|\|\overline{q}_h - p_h\|.$$

Thus, it holds

$$\|\overline{q}_h - p_h\| \leq c\frac{1}{\alpha}(h^2 + \|\overline{q} - p_h\|).$$

Together with the triangle inequality, we get

$$\begin{aligned}\|\overline{q} - \overline{q}_h\| &\leq \|\overline{q} - p_h\| + \|\overline{q}_h - p_h\| \\ &\leq \|\overline{q} - p_h\| + c\frac{1}{\alpha}(h^2 + \|\overline{q} - p_h\|)\end{aligned}$$

which shows the assertion. □

To finish the control error estimate, we need to estimate the best approximation error $\inf_{p_h \in \widehat{Q}_h} \|\overline{q} - p_h\|$. To this end, we define a reasonable approximation, see, e.g., [2] for a similar approach.

To proceed, we take $Q_h = Q_h^{(0)}$ to simplify the calculations.

Lemma 7.2.9. *The function p_h defined by*

$$p_h\big|_T = \begin{cases} q_{\min} & \text{if } \int_T (\overline{z} + \alpha\overline{q})\,dx > 0, \\ q_{\max} & \text{if } \int_T (\overline{z} + \alpha\overline{q})\,dx < 0, \\ |T|^{-1} \int_T \overline{q}\,dx & \text{if } \int_T (\overline{z} + \alpha\overline{q})\,dx = 0. \end{cases}$$

satisfies $p_h \in \widehat{Q}_h$, e.g., $p_h \in Q^{ad,h}$ and

$$j'(\overline{q})(\overline{q}_h - p_h) \geq 0.$$

Further, it holds

$$\|\overline{q} - p_h\| \leq ch\|\overline{q}\|_{1,\infty}.$$

Proof. It is clear, by definition, that $p_h \in Q^{\mathrm{ad,h}}$. To show that $p_h \in \widehat{Q}_h$, we show that the inequality holds element wise. Because

$$j'(\overline{q})(\overline{q}_h - p_h) = (\overline{z} + \alpha\overline{q}, \overline{q}_h - p_h) = \sum_{T \in \mathcal{T}_h} (\overline{z} + \alpha\overline{q}, \overline{q}_h - p_h)_T$$

this will show the assertion. This will be done by a case-by-case analysis:

a) $p_h = q_{\min}$ on T:
By definition, it holds $\int_T (\overline{z} + \alpha\overline{q})\,dx \geq 0$ and therefore

$$(\overline{z} + \alpha\overline{q}, \overline{q}_h - p_h)_T = (\overline{q}_h\big|_T - q_{\min}) \int_T (\overline{z} + \alpha\overline{q})\,dx \geq 0$$

b) $p_h = q_{\max}$ on T:
By definition, it holds $\int_T (\overline{z} + \alpha\overline{q})\,dx \leq 0$ and therefore

$$(\overline{z} + \alpha\overline{q}, \overline{q}_h - p_h)_T = (\overline{q}_h\big|_T - q_{\max}) \int_T (\overline{z} + \alpha\overline{q})\,dx \geq 0$$

c) otherwise:
We have that $\int_T (\overline{z} + \alpha\overline{q})\,dx = 0$ and thus

$$(\overline{z} + \alpha\overline{q}, \overline{q}_h - p_h)_T = (\overline{q}_h\big|_T - p_h\big|_T) \int_T (\overline{z} + \alpha\overline{q})\,dx = 0$$

Before we proceed to analyze the error $\|\overline{q} - p_h\|$ we note that by the results of Section 6.3.1 (Which still holds on a convex polygonal domain $\Omega \subset \mathbb{R}^2$) the regularity assumption $\overline{q} \in W^{1,\infty}(\Omega)$ is reasonable. We proceed by deriving bounds for the error on the elements by case differentiation

a) $p_h = q_{\min}$ on T:
By definition of p_h there exists $x_0 \in T$ with $\overline{z}(x_0) + \alpha\overline{q}(x_0) > 0$. In view of the projection formula (6.6), this means $\overline{q}(x_0) = q_{\min} = p_h(x_0)$. Using the fundamental theorem of calculus, we get

$$\begin{aligned}\|\overline{q} - p_h\|_T^2 &= \int_T |\overline{q}(x) - \overline{q}(x_0)|^2 \, dx \\ &\leq \int_T \Big(\int_0^1 |\nabla \overline{q}(x_0 + t(x - x_0))(x - x_0)| \, dt \Big)^2 dx \\ &\leq h^2 |T| \|\overline{q}\|_{1,\infty}^2.\end{aligned}$$

b) $p_h = q_{\max}$ on T:
Analog to a) it follows

$$\|\overline{q} - p_h\|_T^2 \leq h^2 |T| \|\overline{q}\|_{1,\infty}^2.$$

c) otherwise:
Then $p_h\big|_T = |T|^{-1} \int_T \overline{q} \, dx$ and thus $p_h\big|_T$ is the L^2-projection of $\overline{q}\big|_T$ on T and thus by standard estimates for $P_h \colon L^2(T) \to P_0(T)$ it follows

$$\|\overline{q} - p_h\|_T^2 \leq ch^2 \|\nabla u\|_T^2 \leq ch^2 |T| \|\overline{q}\|_{1,\infty}^2.$$

Together this yields

$$\|\overline{q} - p_h\|^2 \leq \sum_{T \in \mathcal{T}_h} \|\overline{q} - p_h\|_T^2 \leq \sum_{T \in \mathcal{T}_h} ch^2 |T| \|\overline{q}\|_{1,\infty}^2 \leq ch^2 \|\overline{q}\|_{1,\infty}^2 \sum_{T \in \mathcal{T}_h} |T| \leq ch^2 \|\overline{q}\|_{1,\infty}^2.$$

□

Now combining the results of Lemma 7.2.9 and Theorem 7.2.8 we have shown the following

Theorem 7.2.10. *Let $\overline{q}$ be the optimal control of problem (7.3) and $\overline{q}_h$ be the optimal control of problem (7.5). Then there exists a constant $c > 0$ depending on $\overline{q} \in W^{1,\infty}(\Omega)$ such that for $\alpha \in (0, 1]$*

$$\|\overline{q} - \overline{q}_h\| \leq c\Big(\frac{h^2}{\alpha} + (1 + \frac{1}{\alpha})h\Big).$$

From this one can immediately deduce the following sub-optimal convergence order for the state variable

Corollary 7.2.11. *Let $\overline{u}$ be the optimal state of problem (7.3) and $\overline{u}_h$ be the optimal state of problem (7.5). Then it holds*

$$\|\overline{u}-\overline{u}_h\| + \|\nabla(\overline{u}-\overline{u}_h)\| \leq ch$$

with a constant independent of h but depending on $\overline{q}$ and α.

Proof. The results follows immediately by noting that due to Poincaré's inequality (Theorem 2.6.12)

$$\|\overline{u}-\overline{u}_h\|+ \leq c\|\nabla(\overline{u}-\overline{u}_h)\|$$

and, by the finite element error estimate (Theorem 7.1.4), we have

$$\begin{aligned}\|\nabla(\overline{u}-\overline{u}_h)\| &\leq \|\nabla(S\overline{q}-S_h\overline{q})\| + \|\nabla(S_h\overline{q}-S_h\overline{q}_h)\| \\ &\leq ch\|\overline{q}\| + c\|\overline{q}-\overline{q}_h\| \\ &\leq ch.\end{aligned}$$

□

Remark 7.2.12. One can show that if $Q_h = Q_h^{(1)}$ then the best approximation error satisfies

$$\inf_{p_h\in\widehat{Q}_h} \|\overline{q}-p_h\| \leq ch^{3/2}$$

with a constant depending on $\overline{q}$.

An Optimal Order Estimate for the State Variable

In this section, we will show that, in fact, we can obtain the optimal convergence rate for the state variable in the L^2-norm, namely

Theorem 7.2.13. *Let $\overline{u}$ be the optimal state of problem (7.3) and $\overline{u}_h$ be the optimal state of problem (7.5). Then there exists a constant $c > 0$ depending on $\overline{q}$ such that*

$$\|\overline{u}-\overline{u}_h\| \leq ch^2\Big(1+\frac{1}{\alpha}\Big)$$

However, before we can finally obtain this result, we will need some preparation.

Definition 7.2.14. We define the mid-point interpolation $\mathscr{M}_h\colon C(\overline{\Omega}) \to Q_h^{(0)}$ as follows

$$\mathscr{M}_h(q)\big|_T = q(m_T) \quad \forall\, T \in \mathscr{T}_h$$

where m_T denotes the midpoint of the element T.

For this interpolation one can show the following estimates by standard scaling techniques (Bramble-Hilbert)

Lemma 7.2.15. *For the mid-point interpolation $\mathcal{M}_h$ the following estimates hold (for a sufficiently regular q):*

$$\left| \int_T q(x) - \mathcal{M}_h(q)(x)\, dx \right| \leq ch^2 |T|^{1/2} \|\nabla^2 q\|_T, \tag{7.7}$$

$$\|q - \mathcal{M}_h(q)\|_{\infty,T} \leq ch \|q\|_{1,\infty,T}, \tag{7.8}$$

$$\|q - \mathcal{M}_h(q)\|_T \leq ch \|\nabla^2 q\|_T, \tag{7.9}$$

$$\|\mathcal{M}_h(q)\| \leq \|q\|_\infty \tag{7.10}$$

To continue, we note that the optimal control $\overline{q}$ beeing only Lipschitz is limiting our possibility to increase the convergence order of the control variable in L^2 by increasing the polynomial degree. For the space $Q_h^{(1)}$ the rate would be $h^{3/2}$ instead of the desired h^2. Thus to obtain the desired rate of convergence in the L^2-norm we can not simply increase the polynomial degree in Q_h. Instead, we need a more refined technique. To this end, we introduce the following subsets $\mathcal{T}_h^r$ of elements T where $\overline{q}$ is regular and $\mathcal{T}_h^s$ where $\overline{q}$ has kinks as follows

$$\mathcal{T}_h^r = \{T \in \mathcal{T}_h \,|\, \overline{q}\big|_T \equiv q_{\min},\ \overline{q}\big|_T \equiv q_{\max},\ \text{or } q_{\min} < \overline{q}\big|_T < q_{\max}\},$$
$$\mathcal{T}_h^s = \mathcal{T}_h \setminus \mathcal{T}_h^1.$$

Corresponding to these sets, we can decompose Ω into

$$\Omega_h^r = \operatorname{int}\Big(\cup_{T\in\mathcal{T}_h^r} T\Big), \quad \Omega_h^s = \operatorname{int}\Big(\cup_{T\in\mathcal{T}_h^s} T\Big).$$

We now state the following

Assumption 7.2.16. We assume that there exists some constant $c > 0$, independent of h, such that

$$|\Omega_h^s| \leq \sum_{T\in\mathcal{T}_h^s} |T| \leq ch.$$

To proceed, we define the following norms for notational simplicity

$$\|q\|_h = \|q\|_{1,\infty,\Omega_h^s} + \|q\|_{2,2,\Omega_h^r}.$$

To continue, we derive some auxiliary results.

Lemma 7.2.17. *Let $\overline{q}$ be the solution to (7.3) and define $p_h = \mathcal{M}_h(\overline{q})$. Then for the corresponding states $\overline{u} = u(\overline{q}) = S\overline{q}$ and $u(p_h) = Sp_h$ it holds*

$$\|\overline{u} - u(p_h)\| \leq ch^2 \|\overline{q}\|_h$$

with a constant independent of h, α, and $\overline{q}$.

Proof. To begin with, we define the following dual variable $\lambda \in H_0^1(\Omega)$ as solution to

$$(\nabla\varphi, \nabla\lambda) = \frac{(\overline{u} - u(p_h), \varphi)}{\|\overline{u} - u(p_h)\|} \quad \forall \varphi \in H_0^1(\Omega).$$

In view of the elliptic regularity in Section 4.2, it is $\lambda \in H^2(\Omega)$ and it holds

$$\|\lambda\|_{2,2} \leq c.$$

Then, we obtain that

$$\begin{aligned} \|\overline{u} - u(p_h)\| &= (\nabla(\overline{u} - u(p_h)), \nabla\lambda) \\ &= (\overline{q} - p_h, \lambda) \\ &= (\overline{q} - p_h, \lambda)_{\Omega_h^r} + (\overline{q} - p_h, \lambda)_{\Omega_h^s}. \end{aligned} \tag{7.11}$$

Now, we estimate the two summands separately. On the regular part, we conclude

$$\begin{aligned} |(\overline{q} - p_h, \lambda)_{\Omega_h^r}| &\leq \sum_{T \in \mathscr{T}_h^r} \left| (\overline{q} - p_h, \lambda - \mathscr{M}_h(\lambda))_T + (\overline{q} - p_h, \mathscr{M}_h(\lambda))_T \right| \\ &\leq \sum_{T \in \mathscr{T}_h^r} \left(\|\overline{q} - p_h\|_T \|\lambda - \mathscr{M}_h(\lambda)\|_T + |\mathscr{M}_h(\lambda)|_T| \left| \int_T \overline{q} - p_h \, dx \right| \right). \end{aligned}$$

Thus using (7.7) and (7.9) we obtain (noting $p_h = \mathscr{M}_h(\overline{q})$)

$$\begin{aligned} |(\overline{q} - p_h, \lambda)_{\Omega_h^r}| &\leq ch^2 \sum_{T \in \mathscr{T}_h^r} \left(\|\overline{q}\|_{2,2,T} \|\lambda\|_{2,2,T} + |\mathscr{M}_h(\lambda)|_T| \, |T|^{1/2} \|\overline{q}\|_{2,2,T} \right) \\ &\leq ch^2 \sum_{T \in \mathscr{T}_h^r} \left(\|\overline{q}\|_{2,2,T} \|\lambda\|_{2,2,T} + \|\mathscr{M}_h(\lambda)\|_T \|\overline{q}\|_{2,2,T} \right). \end{aligned}$$

An application of the Cauchy-Schwarz inequality and (7.10) yield

$$|(\overline{q} - p_h, \lambda)_{\Omega_h^r}| \leq ch^2 \|\overline{q}\|_{2,2,\Omega_h^r} \|\lambda\|_{2,2,\Omega_h^r}. \tag{7.12}$$

The irregular part can be estimated as follows

$$\begin{aligned} |(\overline{q} - p_h, \lambda)_{\Omega_h^s}| &\leq \sum_{T \in \mathscr{T}_h^s} \left| (\overline{q} - p_h, \lambda)_T \right| \\ &\leq \|\overline{q} - p_h\|_{\infty,\Omega_h^s} \|\lambda\|_{\infty,\Omega_h^s} \sum_{T \in \mathscr{T}_h^s} |T|. \end{aligned}$$

Using Assumption 7.2.16, the continuous embedding $H^2(\Omega) \subset L^\infty(\Omega)$, and (7.8), we conclude

$$\begin{aligned} |(\overline{q} - p_h, \lambda)_{\Omega_h^s}| &\leq ch \|\overline{q}\|_{1,\infty,\Omega_h^s} \|\lambda\|_{2,2,\Omega_h^s} \sum_{T \in \mathscr{T}_h^s} |T| \\ &\leq ch^2 \|\overline{q}\|_{1,\infty,\Omega_h^s} \|\lambda\|_{2,2,\Omega_h^s}. \end{aligned} \tag{7.13}$$

Summing up, we get from (7.11), (7.12), and (7.13) that

$$\|\overline{u} - u(p_h)\| \leq ch^2 \|\overline{q}\|_{2,2,\Omega_h^r} \|\lambda\|_{2,2,\Omega_h^r} + ch^2 \|\overline{q}\|_{1,\infty,\Omega_h^s} \|\lambda\|_{2,2,\Omega_h^s}$$

and thus the assertion follows. □

Remark 7.2.18. The proof of Lemma 7.2.17 and, in particular, (7.12) and (7.13) show that, in fact, we have shown

$$\|\overline{q} - p_h\|_{-2} \leq ch^2$$

with a constant c depending on $\overline{q}$ noting that the estimates involved for the computation of an upper bound on

$$|(\overline{q} - p_h, \lambda)|$$

did not depend on the special choice of λ. Moreover, using $\lambda - P_h(\lambda)$ instead of $\lambda - \mathscr{M}_h(\lambda)$ would reveal that for any $\varepsilon > 0$

$$\|\overline{q} - p_h\|_{-1-\varepsilon} \leq ch^2.$$

In a next step, we will try to bound some of the terms occurring in the derivative of the reduced cost functional.

Lemma 7.2.19. *Let $\overline{q}$ be the solution to (7.3) and define $p_h = \mathscr{M}_h(\overline{q})$. Then for the corresponding adjoint states $\overline{z} = z(\overline{q}) = S^*(S\overline{q} - u^d)$ and $z_h(p_h) = S_h^*(S_h p_h - u^d)$ it holds*

$$\|\overline{z} - z_h(p_h)\| \leq ch^2(1 + \|\overline{q}\|_h)$$

with a constant $c > 0$, independent of h, α, and $\overline{q}$.

Proof. To show this, we introduce the intermediate quantity

$$\hat{z} = S^*(S_h p_h - u^d).$$

Then, we can split the error

$$\begin{aligned}
\|\overline{z} - z_h(p_h)\| &\leq \|\overline{z} - z(p_h)\| + \|z(p_h) - \hat{z}\| + \|\hat{z} - z_h(p_h)\| \\
&= \|S^*(S\overline{q} - u^d) - S^*(Sp_h - u^d)\| + \|S^*(Sp_h - u^d) - S^*(S_h p_h - u^d)\| \\
&\quad + \|S^*(S_h p_h - u^d) - S_h^*(S_h p_h - u^d)\| \\
&\leq c\|S\overline{q} - Sp_h\| + c\|(S - S_h)p_h\| + \|(S - S_h)(S_h p_h - u^d)\|.
\end{aligned}$$

Note that, we may not remove the operator S in the first term. Otherwise, we will only get convergence of order h. Now, we can estimate the first term using Lemma 7.2.17 and the rest by the a priori finite element estimate Theorem 7.1.4 to get

$$\begin{aligned}
\|\overline{z} - z_h(p_h)\| &\leq ch^2\|\overline{q}\|_h + ch^2\|p_h\| + ch^2(\|p_h\| + 1) \\
&\leq ch^2(1 + \|\overline{q}\|_h)
\end{aligned}$$

where the last inequality follows from (7.10). □

In the next step, we will show that the defined discrete value $p_h = \mathscr{M}_h(\overline{q})$ is a sufficiently good approximation to $\overline{q}_h$.

Lemma 7.2.20. *Let $\overline{q}$ be the solution to (7.3), $\overline{q}_h$ be the solution to (7.5) and define $p_h = \mathcal{M}_h(\overline{q})$. Then it holds*

$$\|\overline{q}_h - p_h\| \leq c\frac{h^2}{\alpha}(1 + \|\overline{q}\|_h)$$

with a constant $c > 0$, independent of h, α and $\overline{q}$.

Proof. From the pointwise optimality condition (6.5), we obtain for the midpoints m_T of an element T that

$$(\overline{z}(m_T) + \alpha\overline{q}(m_T))(\overline{q}_h(m_T) - \overline{q}(m_T)) \geq 0.$$

Now, we note that

$$p_h\big|_T = \mathcal{M}_h(\overline{q})\big|_T = \overline{q}(m_T)$$

and hence

$$(\overline{z}(m_T) + \alpha p_h(m_T))(\overline{q}_h(m_T) - p_h(m_T)) \geq 0.$$

From this, we obtain by integration over T and summation over all $T \in \mathcal{T}_h$ (noting that all quantities are constant on an element)

$$(\mathcal{M}_h(\overline{z}) + \alpha p_h, \overline{q}_h - p_h) \geq 0.$$

On the other hand, the discrete optimality condition, see Theorem 7.2.5, yields

$$(\overline{z}_h + \alpha\overline{q}_h, p_h - \overline{q}_h) \geq 0,$$

by noting that $p_h \in Q^{\mathrm{ad,h}}$.

Now, we add the two inequalities to get

$$(\mathcal{M}_h(\overline{z}) - \overline{z}_h + \alpha(p_h - \overline{q}_h), \overline{q}_h - p_h) \geq 0$$

or

$$\alpha\|\overline{q}_h - p_h\|^2 \leq (\mathcal{M}_h(\overline{z}) - \overline{z}_h, \overline{q}_h - p_h). \tag{7.14}$$

We proceed and insert various productive zeros

$$\begin{aligned}(\mathcal{M}_h(\overline{z}) - \overline{z}_h, \overline{q}_h - p_h) = (\mathcal{M}_h(\overline{z}) - \overline{z}, \overline{q}_h - p_h) + (\overline{z} - z_h(p_h), \overline{q}_h - p_h)\\ + (z_h(p_h) - \overline{z}_h, \overline{q}_h - p_h).\end{aligned} \tag{7.15}$$

For the first term on the right hand side, we note that $\overline{q}_h - p_h$ is constant on each element, and thus using (7.7)

$$\begin{aligned}(\mathcal{M}_h(\overline{z}) - \overline{z}, \overline{q}_h - p_h) &= \sum_{T\in\mathcal{T}_h} (\overline{q}_h(m_T) - p_h(m_T)) \int_T \overline{z}(m_T) - \overline{z}\, dx\\ &\leq ch^2 \sum_{T\in\mathcal{T}_h} |\overline{q}_h(m_T) - p_h(m_T)||T|^{1/2}\|\overline{z}\|_{2,2,T}\\ &= ch^2 \sum_{T\in\mathcal{T}_h} \|\overline{q}_h - p_h\|_T\|\overline{z}\|_{2,2,T}.\end{aligned}$$

Finally, using Cauchy Schwarz inequality and stability estimates, see Section 4.2, we get

$$(\mathscr{M}_h(\overline{z})-\overline{z},\overline{q}_h-p_h)\leq ch^2\|\overline{q}_h-p_h\|(1+\|\overline{q}\|_h). \tag{7.16}$$

For the second term, we obtain, using Lemma 7.2.19, that

$$\begin{aligned}(\overline{z}-z_h(p_h),\overline{q}_h-p_h)&\leq \|\overline{z}-z_h(p_h)\|\|\overline{q}_h-p_h\|\\&\leq ch^2(1+\|\overline{q}\|_h)\|\overline{q}_h-p_h\|.\end{aligned} \tag{7.17}$$

For the third term in (7.15), we use the definition of z_h by S_h^* to get

$$\begin{aligned}(z_h(p_h)-\overline{z}_h,\overline{q}_h-p_h)&=(u_h(p_h)-\overline{u}_h,\overline{u}_h-u_h(p_h))\\&=-\|u_h(p_h)-\overline{u}_h\|^2\leq 0.\end{aligned} \tag{7.18}$$

Thus, combining (7.14)–(7.18), we get

$$\alpha\|\overline{q}_h-p_h\|^2\leq ch^2\|\overline{q}_h-p_h\|\big((1+\|\overline{q}\|_h)+(1+\|\overline{q}\|_h)+0\big)$$

which proofs the assertion. □

Finally, we can proof the main result of this section:

Theorem 7.2.21. *Let $\overline{u}$ be the optimal state of problem (7.3) and $\overline{u}_h$ be the optimal state of problem (7.5). Then there exists a constant $c>0$ depending on $\|\overline{q}\|_h$ such that*

$$\|\overline{u}-\overline{u}_h\|\leq ch^2\big(1+\frac{1}{\alpha}\big)$$

Proof. To obtain the desired result, we split the error and get, with the notation of the previous lemmas,

$$\|\overline{u}-\overline{u}_h\|\leq\|\overline{u}-u(p_h)\|+\|u(p_h)-u_h(p_h)\|+\|u_h(p_h)-\overline{u}_h\|.$$

For the first term, we get from Lemma 7.2.17 that

$$\|\overline{u}-u(p_h)\|\leq ch^2\|\overline{q}\|_h.$$

For the second term, standard error estimates for the PDE, see Theorem 7.1.4 and (7.10), yield

$$\|u(p_h)-u_h(p_h)\|\leq ch^2\|p_h\|\leq ch^2\|\overline{q}\|_h.$$

For the third term, we note that $S_h\to S$ in $\mathscr{L}(L^2(\Omega))$ and thus $\|S_h\|\leq c$. This gives, together with Lemma 7.2.20,

$$\|u_h(p_h)-\overline{u}_h\|\leq c\|p_h-\overline{q}_h\|\leq c\frac{h^2}{\alpha}(1+\|\overline{q}\|_h).$$

Thus, the assertion follows. □

Remark 7.2.22. The combined proofs of Lemma 7.2.17, Lemma 7.2.19, and Lemma 7.2.20 show that in fact

$$\|\overline{q} - \overline{q}_h\|_{-2} \leq \|\overline{q} - p_h\|_{-2} + \|p_h - \overline{q}_h\| \leq ch^2$$

and thus Theorem 7.2.21 can directly be obtained by citing standard error estimates for the solution of the PDE with errors in the right hand side.

Remark 7.2.23. We note that using this convergence orders, one can readily deduce that, in fact, one can obtain the same order of convergence as in the variational discretization using the following post-processed control

$$\tilde{q}_h = \mathscr{P}_{Q^{\mathrm{ad}}}\Big(\frac{-1}{\alpha}\overline{z}_h\Big).$$

Namely, one gets

$$\|\tilde{q}_h - \overline{q}\| \leq ch^2$$

see, [27].

8 Algorithms

8.1 Unconstrained Case

We start our investigation with algorithms to tackle the unconstrained problem

$$\min_{q\in Q} j(q) = \frac{1}{2}\|Sq - u^d\|_U^2 + \frac{\alpha}{2}\|q\|_Q^2 \tag{8.1}$$

on a Hilbert spaces U, Q with $S \in \mathscr{L}(Q,U)$ and $\alpha > 0$.

To solve this problem numerically one could simply discretize the spaces and apply any standard optimization algorithm to the discrete optimization problem. However this approach has some drawbacks, for instance one usually needs to stop the algorithms at some time, typical stopping criteria would be to stop once $\|\nabla j\| \leq \text{TOL}$ in some norm. Clearly they are all equivalent in $\mathbb{R}^n$, but they may give trouble when refining the mesh. For instance using the space $Q_h^{(0)}$ the euclidean norm of the coordinate vector $\mathbf{q_h}$ would be

$$\|\mathbf{q_h}\|_{l^2}^2 = \sum_{i=1}^{N} (q_h^{(i)})^2$$

where as the $L^2(\Omega)$ norm of the corresponding vector $q_h = \sum_{i=1}^N q_h^{(i)} \hat{\varphi}_h^{(i)}$ is given by

$$\|q_h\|_{L^2(\Omega)}^2 \approx \frac{1}{N} \sum_{i=1}^{N} (q_h^{(i)})^2$$

this means that if $q_h \equiv 1$ then

$$\|\mathbf{q_h}\|_{l^2}^2 \to \infty \quad (N \to \infty), \qquad \|q_h\|_{L^2(\Omega)}^2 = \text{const} \quad (N \to \infty)$$

which implies that the use of the „wrong" discrete norm can make the problem harder to solve if $N \to \infty$. On the other hand, using a discrete norm like $\|q_h\|_X^2 = \approx \frac{1}{N^2} \sum_{i=1}^N (q_h^{(i)})^2$ would give

$$\|q_h\|_X^2 \to 0 \quad (N \to \infty)$$

and thus one would be able to satisfy the tolerance requirements simply by refining the discretization (This is clearly undesired!).

Thus we will present some algorithms in a function space setting, which implies the „right" norm to be used in the stopping criteria, and is useful to show mesh independent convergence of the algorithms.

Our aim is now to develop a method that generates a sequence q_k with $q_k \to \overline{q}$ where $\overline{q}$ solves (8.1).

We will start with a few definitions

Definition 8.1.1. We say that a method converges globally, if for any initial value q_0 the sequence q_k converges to a (local) solution $\overline{q}$ of (8.1).

Definition 8.1.2. We say that the convergence $q_k \to \overline{q}$ is linear with rate c if

$$\|q_{k+1} - \overline{q}\|_Q \leq c\|q_k - \overline{q}\|_Q.$$

It converges super-linear if

$$\|q_{k+1} - \overline{q}\|_Q \leq c_k\|q_k - \overline{q}\|_Q.$$

with $c_k \to 0$. It converges with rate α if for some constant c

$$\|q_{k+1} - \overline{q}\|_Q \leq c\|q_k - \overline{q}\|_Q^\alpha.$$

There is now a variety of algorithm available, we will consider here so called line-search algorithms, for the alternative trust-region algorithms we refer to the literature, e.g., [9].

Algorithm 8.1 Generic line-search method

Choose $q_0 \in Q$ and let $k = 0$.
while Stopping criterion not satisfied **do**
  Choose some direction d_k ▷ Should be descent, e.g., $\delta j(q_k; d_k) < 0$
  Choose some step length $t_k > 0$ ▷ Should give sufficient descent, e.g., ▷ $j(q_k + t_k d_k) \leq j(q_k) + \sigma \delta j(q_k, t_k d_k)$, $\sigma \in (0, 1)$)
  $q_{k+1} = q_k + t_k d_k$
  $k \leftarrow k + 1$
end while

We will now discuss a few possibilities to choose the direction.

8.1.1 Gradient descent

As a first (and most simple) choice for the direction one can consider the method of steepest descent. To this end we recall that we can define the gradient of j at q_k, using Riesz representation Theorem 2.3.7, as follows. Find $\nabla j(q_k) \in Q$ that satisfies

$$(\nabla j(q_k), \delta q)_Q = j'(q_k)\delta q \quad \forall\, \delta q \in Q.$$

It is then immediately clear that $d_k = -\nabla j(q_k)$ is a descent direction

$$j'(q_k)d_k = (\nabla j(q_k), d_k)_Q = -\|\nabla j(q_k)\|_Q^2 \leq 0.$$

Further if $j'(q_k)d_k = 0$ then $\nabla j(q_k) = 0$ and thus the necessary conditions for a minimizer are fulfilled. (For our model problem these are also sufficient.)

Now, to proceed, we notice that our model problem satisfies

$$j'(q_k)\delta q = (S^*Sq_k + \alpha q_k - S^*u^d, \delta q)_Q$$

and thus with the Hessian operator $H = S^*S + \alpha\,\mathrm{Id} \in \mathscr{L}(Q,Q)$ and $b = S^*u^d$ we have that

$$\nabla j(q_k) = Hq_k - b.$$

We notice that H is positive definite, because

$$(Hq,q)_Q = (Sq,Sq)_U + \alpha(q,q)_Q \geq \alpha\|q\|_Q^2.$$

For this problem, the step length can be computed analytically by finding t_k as solution to

$$\frac{d}{dt} j(q_k + td_k) = 0.$$

This means

$$0 = j'(q_k + td_k)d_k = (Hq_k + tHd_k - b, d_k)_Q = t(Hd_k, d_k)_Q + (Hq_k - b, d_k)_Q$$

and thus if $d_k \neq 0$ we have $(Hd_k, d_k)_Q \neq 0$ and hence

$$t_k = \frac{\|d_k\|_Q^2}{(Hd_k, d_k)_Q}.$$

With this choice it holds

$$j(q_k + t_k d_k) = j(q_k) + \frac{t_k}{2}(\nabla j(q_k), d_k)_Q,$$

compare Lemma 8.1.4, and thus we have the required descend condition.

In summary we obtain the practical gradient descent Algorithm 8.2.

Algorithm 8.2 Gradient descend

Choose $q_0 \in Q$ and let $k = 0$.
Choose $0 < \mathrm{TOL} < 1$ and $k_{\max} \in \mathbb{N}$
Calculate $d_0 = -\nabla j(q_0) = b - Hq_0$
Set $\delta_0 = \|d_k\|_Q$
while $\delta_k > \mathrm{TOL}\,\delta_0$ and $k < k_{\max}$ **do**
  $h_k = Hd_k$

$$t_k = \frac{\delta_k^2}{(h_k, d_k)_Q}$$

  $q_{k+1} = q_k + t_k d_k$
  $d_{k+1} = -\nabla j(q_{k+1}) = d_k - t_k h_k$
  $\delta_{k+1} = \|d_{k+1}\|_Q$
  $k \leftarrow k + 1$
end while

Remark 8.1.3.

1. The main effort in Algorithm 8.2 is hidden in the application of the operator H which consist in the evaluation of S and S^* (two PDE solves). This is the reason why the additional variable h_n is introduced, otherwise four PDE solves are necessary per iteration.
2. The algorithm can be applied to more general differentiable functionals j. Then the value t_k can no longer be determined by an exact line-search, but must be determined, e.g., by a back tracking line search procedure with some Armijo rule stopping criterion.
3. The calculation of ∇f can not be done exactly, but requires some discretization. This can be coped with using inexact variants of the gradient descent method (gradient like methods). It is still a field of research how to couple the fineness of the discretization with the required accuracy for the optimization routine.

Now, we would like to show convergence of the algorithm. To this end, we derive a relation between the step and the decrease in the function value

Lemma 8.1.4. *Let $q_k \in Q$ be arbitrary. Further define $d_k = b - Hq_k = -\nabla j(q_k)$ and $q_{k+1} = q_k + td_k$ for some $t \in \mathbb{R}$. Then it holds*

$$j(q_{k+1}) - j(q_k) = -t\|d_k\|_Q^2 + \frac{t^2}{2}(d_k, Hd_k)_Q \geq -t\|d_k\|^2$$

Proof. We have by definition of j and q_{k+1} that

$$\begin{aligned}
j(q_{k+1}) - j(q_k) &= \frac{1}{2}(S(q_k + td_k) - u^d, S(q_k + td_k) - u^d)_U + \frac{\alpha}{2}(q_k + td_k, q_k + td_k)_Q \\
&\quad - \frac{1}{2}(Sq_k - u^d, Sq_k - u^d)_U - \frac{\alpha}{2}(q_k, q_k)_Q \\
&= t(Sd_k, Sq_k - u^d)_U + \frac{t^2}{2}(Sd_k, Sd_k)_U + t(d_k, \alpha q_k)_Q + \frac{t^2}{2}(d_k \alpha d_k)_Q \\
&= t(d_k, S^*(Sq_k - u^d) + \alpha q_k)_Q + \frac{t^2}{2}(d_k, Hd_k)_Q \\
&= -t\|d_k\|_Q^2 + \frac{t^2}{2}(d_k, Hd_k)_Q.
\end{aligned}$$

□

We can now show the global convergence of the gradient descent Algorithm 8.2.

Theorem 8.1.5. *For any initial value q_0 Algorithm 8.2 converges to the solution of (8.1).*

Proof. To see the assertion, we start by showing that $\|d_k\|_Q \to 0$. To this end let us assume w.l.o.g. that $\|d_k\|_Q \neq 0$. Then due to positive definiteness of H we get by definition of t_k that

$$0 < t_k = \frac{\|d_k\|_Q^2}{(Hd_k, d_k)_Q} < \infty.$$

From Lemma 8.1.4 we obtain that

$$j(q_{k+1}) - j(q_k) = -t_k \|d_k\|_Q^2 + \frac{t_k^2}{2}(d_k, Hd_k)_Q = -\frac{t_k}{2}\|d_k\|_Q^2.$$

Now, because $j(q_k) \geq 0$ and monotone non-increasing it holds $j(q_k) \to \bar{j} \geq 0$. Thus, using telescope sums, we get

$$\sum_{k=0}^{n} \frac{t_k}{2}\|d_k\|_Q^2 = j(q_0) - j(q_{n+1}).$$

Taking the limit $n \to \infty$ yields

$$\sum_{k=0}^{\infty} \frac{t_k}{2}\|d_k\|^2 = j(q_0) - \bar{j}$$

and thus $\frac{t_k}{2}\|d_k\|^2 \to 0$.

To proceed, we use Lemma 8.1.4 to get

$$j(q_k + 2t_k d_k) - j(q_k) = -2t_k\|d_k\|_Q^2 + \frac{4t_k^2}{2}(d_k, Hd_k)_Q = 2t_k(-\|d_k\|_Q^2 + \|d_k\|_Q^2) = 0.$$

This implies (using continuity of H) that

$$\begin{aligned} 0 &= -2t_k\|d_k\|_Q^2 + \frac{4t_k^2}{2}(d_k, Hd_k)_Q \\ &\leq -2t_k\|d_k\|_Q^2 + 2t_k^2 c\|d_k\|_Q^2. \end{aligned}$$

Division by $2t_k$ yields

$$\|d_k\|^2 \leq c t_k \|d_k\|^2 \to 0.$$

This shows $\|d_k\| \to 0$.

To conclude the proof, we need to see that q_k converges to the solution $\bar{q}$ of (8.1). To see this we note that

$$\frac{\alpha}{2}\|q_k\|^2 \leq j(q_k) \to \bar{j}.$$

Thus, the sequence q_k is bounded and we get

$$\begin{aligned} \alpha\|q_k - \bar{q}\|_Q^2 &\leq \alpha\|q_k - \bar{q}\|_Q^2 + \|Sq_k - S\bar{q}\|_U^2 \\ &= (q_k - \bar{q}, \alpha(q_k - \bar{q}) + S^*(Sq_k - u^d - S\bar{q} + u^d))_Q \\ &= (q_k - \bar{q}, Hq_k - b - H\bar{q} + b)_Q \\ &= (q_k - \bar{q}, -d_k + 0) \to 0. \end{aligned}$$

□

Remark 8.1.6. The last step in the proof, namely to show that $q_k \to \overline{q}$ can also be seen by the fact that H is a compact perturbation of the identity, thus it is a Fredholm operator of index zero and hence surjective if it is injective. This means that because H is positive definite the inverse of H is a bounded linear operator which immediately implies the convergence of q_k given convergence of d_k.

Finally, we note that the following result (which is well known in finite dimensions) holds true

Lemma 8.1.7. *For a symmetric positive definite, and bounded operator H, assume that the spectrum $\sigma(H) \subset [\lambda, \Lambda]$ for some constants $0 < \lambda \leq \Lambda < \infty$. We define the condition number*

$$\kappa = \frac{\Lambda}{\lambda}.$$

Then the gradient descend method converges linearly in the norm induced by H with rate $\frac{(\kappa-1)}{(\kappa+1)}$, e.g.,

$$\|q_{k+1} - \overline{q}\|_H \leq \frac{(\kappa-1)}{(\kappa+1)} \|q_k - \overline{q}\|_H$$

where

$$\|q\|_H^2 = (Hq, q)_Q.$$

Proof. The proof in finite dimensions can be found in any introductory script to numerical analysis. For the infinite dimensional case, see, e.g., [12]. □

8.1.2 Newton Methods

A more advanced class of methods for the unconstrained problem is the Newton method. To this end we recall that a solution to (8.1) is equivalently given as the solution to the first order condition

$$0 = j'(\overline{q}) = H\overline{q} - b.$$

This can be solved by the following Algorithm 8.3.

The algorithm requires knowledge of the second derivatives of j. To this end we generally define the hessian of j in a point q to be given as $\nabla^2 j(q) \colon Q \to Q$ defined by

$$(\nabla^2 j(q)\delta q, \tau q) = j''(q; \delta q, \tau q) \quad \forall \delta q, \tau q \in Q.$$

In our case this means $\nabla^2 j(q) = H$.

Following the same line of arguments as in the finite dimensional case one can show that the method given in Algorithm 8.3 converges globally and locally quadratic for a twice continuously differentiable $j \colon Q \to \mathbb{R}$ given some appropriate conditions.

We already know that we can calculate the gradient $\nabla j(q)$ using the primal and adjoint (dual) problems

$$u = Sq, \quad z = S^*(u - u^d)$$

Algorithm 8.3 Newton method for the reduced formulation

Choose $q_0 \in Q$, $\rho \in (0,1)$, $\sigma \in (0,1/2)$, and let $k = 0$.
Choose $0 < \mathrm{TOL} < 1$ and $k_{\max} \in \mathbb{N}$
Calculate d_0 satisfying $\nabla^2 j(q_0) d_0 = -\nabla j(q_0)$
Set $\delta_0 = \|\nabla j(q_0)\|_Q$
while $\delta_k > \mathrm{TOL}\,\delta_0$ and $k < k_{\max}$ **do**
  Determine Step-Length: Set $l = 0$
  $q_{k+1} \leftarrow q_k + \rho^l d_k$
  $r = (\nabla j(q_k), d_k)_Q$.
  while $j(q_{k+1}) > j(q_k) + \sigma \rho^l r$ **do** ▷ Armijo type line-search
    $l \leftarrow l + 1$
    $q_{k+1} \leftarrow q_k - \rho^l d_k$
  end while
  Calculate d_{k+1} satisfying $\nabla^2 j(q_{k+1}) d_{k+1} = -\nabla j(q_{k+1})$
  Set $\delta_{k+1} = \|\nabla j(q_{k+1})\|_Q$
  $k \leftarrow k + 1$
end while

as

$$\nabla j(q) = z + \alpha q.$$

The main question remaining is how to solve the step-length problem

$$\nabla^2 j(q_{k+1}) d_{k+1} = -\nabla j(q_{k+1}).$$

To this end, we note that even after discretization the matrix $H_h = S_h^* S_h - \alpha I$ is dense, hence it is usually undesirable to build this matrix. Therefore we consider iterative methods for the solution of the linear equation

$$H d_{k+1} = b$$

with $H = \nabla^2 j(q_{k+1})$ and $b = -\nabla j(q_{k+1})$. One possibility would be the already discussed gradient descent method, or (more efficiently) a CG-method. To this end we note that one can compute the application H to a direction δq by the solution of two linear equations, namely one determines the $\delta u \in U$ by the tangent-equation

$$\delta u = S \delta q.$$

Then we can calculate $\delta z \in U$ by the dual-for-hessian equation

$$\delta z = S^* \delta u.$$

Then we can evaluate the element $H\delta q$ as

$$H\delta q = \delta z + \alpha \delta q$$

without the necessity to actually compute the operator H (or its discretization).

In the more general case where $u = u(q) = S(q)$ is given by

$$a(q, S(q))(\varphi) = 0 \quad \forall \varphi \in U$$

with a semilinear form $a : Q \times U \times U$, e.g., $a(q,u)(\cdot) \in \mathcal{U}^*$ for all $(q,u) \in Q \times U$ we can consider the problem

$$\min j(q) = J(q, S(q))$$

with some arbitrary function $J : Q \times U \to \mathbb{R}$. Then under additional assumptions one can obtain the gradient and the hessian using the Lagrange formalism. We have already seen this for the gradient in Section 6.1. Again we define

$$\mathcal{L}(q,u,z) = J(q,u) - a(q,u)(z).$$

Then it holds

$$j(q) = \mathcal{L}(q, u(q), z)$$

and hence (with $\delta_q \mathcal{L} = \mathcal{L}'_q, \ldots$) abbreviating $x = (q, u(q), z)$

$$j'(q)\delta q = \mathcal{L}'_q(x)\delta q + \mathcal{L}'_u(x)S'(q)\delta q + \mathcal{L}'_z(x)\frac{d}{dq}z(\delta q)$$

then noting that due to the choice $u = u(q)$ it holds $\mathcal{L}'_z(x) = 0$ we define $z = z(q)$ to solve

$$\mathcal{L}'_u(x)\varphi = 0 \quad \forall \varphi \in U$$

or equivalently

$$a_u(q, S(q))(\varphi, z(q)) = J'_u(q, S(q))\varphi \quad \forall \varphi \in U.$$

Then it holds with $x = (q, u(q), z(q))$

$$j'(q)\delta q = \mathcal{L}'_q(x)\delta q.$$

In a similar way we obtain for two directions δq, τq and the abbreviations $\delta u = u'(q)\delta q$, $\tau u = u'(q)\tau q$, $\delta\tau u = u''(q)(\tau q, \delta q)$ and the same for δz, τz, and $\delta\tau z$ the following representation for the hessian

$$\begin{aligned} j''(q)(\delta q, \tau q) = {} & \mathcal{L}''_{qq}(x)(\delta q, \tau q) + \mathcal{L}''_{qu}(x)(\delta q, \tau u) + \mathcal{L}''_{qz}(x)(\delta q, \tau z) \\ & + \mathcal{L}''_{uq}(x)(\delta u, \tau q) + \mathcal{L}''_{uu}(x)(\delta u, \tau u) + \mathcal{L}''_{uz}(x)(\delta u, \tau z) \\ & + \mathcal{L}''_{zq}(x)(\delta z, \tau q) + \mathcal{L}''_{zu}(x)(\delta z, \tau u) + \mathcal{L}''_{zz}(x)(\delta z, \tau z) \\ & + \mathcal{L}'_u(x)(\delta\tau u) + \mathcal{L}'_z(x)(\delta\tau z). \end{aligned}$$

By definition of $u(q)$ and $z(q)$ the last two terms vanish. Further the term $\mathcal{L}''_{zz} = 0$ since $\mathcal{L}$ is linear in z. Now, we can group these terms in order to determine a tangent equation to define $\delta u \in U$ such that

$$\mathcal{L}''_{qz}(x)(\delta q, \varphi) + \mathcal{L}''_{uz}(x)(\delta u, \varphi) = 0 \quad \forall \varphi \in U.$$

as well as a dual-for-hessian problem to determine $\delta z \in U$ such that

$$\mathscr{L}''_{qu}(x)(\delta q, \varphi) + \mathscr{L}''_{uu}(x)(\delta u, \varphi) + \mathscr{L}''_{zu}(x)(\delta z, \varphi) = 0 \quad \forall \varphi \in U.$$

Then the hessian has the following representation

$$j''(q)(\delta q, \tau q) = \mathscr{L}''_{qq}(x)(\delta q, \tau q) + \mathscr{L}''_{uq}(x)(\delta u, \tau q) + \mathscr{L}''_{zq}(x)(\delta z, \tau q)$$

which makes it easy to solve the representation

$$(\nabla j''(q)\delta q, \tau q)_Q = j''(q)(\delta q, \tau q)$$

to determine the object $\nabla j''(q)\delta q \in Q$.

In summary we can determine a more concrete formulation of Newtons method (assuming that $\nabla^2 j(q)$ remains positive definite. (Otherwise additional modifications in the step selections are required to guarantee decent.)

Algorithm 8.4 Newton method for the reduced formulation

Choose $q_0 \in Q$, $\rho \in (0,1)$, $\sigma \in (0, 1/2)$, and let $k = 0$.
Choose $0 < \mathrm{TOL} < 1$ and $k_{\max} \in \mathbb{N}$
Calculate $u(q_0)$ and $z(q_0)$
Calculate d_0 satisfying $\nabla^2 j(q_0)d_0 = -\nabla j(q_0)$ ▷ Use a CG-method.
▷ Needs δu and δz in each iteration.
Set $\delta_0 = \|\nabla j(q_0)\|_Q$
while $\delta_k > \mathrm{TOL}\,\delta_0$ and $k < k_{\max}$ **do**
  Determine Step-Length: Set $l = 0$
  $q_{k+1} \leftarrow q_k - \rho^l d_k$
  $r = (\nabla j(q_k), d_k)_Q$.
  while $j(q_{k+1}) > j(q_k) + \sigma\rho^l r$ **do** ▷ Armijo type line-search
    $l \leftarrow l + 1$
    $q_{k+1} \leftarrow q_k - \rho^l d_k$
  end while
  Calculate $u(q_0)$ and $z(q_0)$
  Calculate d_{k+1} satisfying $\nabla^2 j(q_{k+1})d_{k+1} = -\nabla j(q_{k+1})$ ▷ Use a CG-method.
  ▷ Needs δu and δz in each iteration.
  Set $\delta_{k+1} = \|\nabla j(q_{k+1})\|_Q$
  $k \leftarrow k + 1$
end while

8.1.3 Direct Solution of the KKT-System

Finally one could solve the corresponding KKT-system for the problem (8.1), namely

$$\begin{aligned} u &= Sq \\ z &= S^*(u - u^d) \\ q &= -z \end{aligned}$$

directly.

Remark 8.1.8. However if the operator S is nonlinear, then this will be a nonlinear system, and hence need to be solved again using some iterative method like gradient descent or newton.

We note that in contrast to the previous sections in the sequel of the iterations for the solution of the KKT-System one does not satisfy the state equation $u = S(q)$. Hence one can not use the cost-function j as a merit function to measure descend.

Possible ways are either to consider simply an appropriate norm of the residual, e.g., L^1 or squared L^2 norm. However one needs to be careful not to find a local minimizer of the merit function which is not a local minimizer of the function j, see, e.g., [28, Chapter 11].

Another possibility is to consider the problem as an equality constrained problem and apply appropriate merit-functions for this problem, see, e.g., [28, Chapter 15]. In any case the choice of the right norm is crucial.

Remark 8.1.9. Finally, one should note that depending on the situation at hand any of the possibilities may be a good choice. The newton method for the reduced formulation may need several solutions of linear PDEs per iteration but they are all of the same type as the state equation using as constraint, hence a good solver for this equation may be available. On the other hand the KKT-System is on one hand larger than the individual PDEs to be solved in the newton method and may be more difficult to solve (Exercise). On the other hand, especially if Q is large, one usually needs fewer solutions of the KKT-system than what is required in the application of the newton method for the reduced system.

8.2 Box Constraints for the Control

$$\begin{aligned} \min_{q \in Q^{\text{ad}}} j(q) &= \frac{1}{2}\|Sq - u^d\|_U^2 + \frac{\alpha}{2}\|q\|_Q^2 \\ Q^{\text{ad}} &= \{q_{\min} \le q \le q_{\max}\} \end{aligned} \tag{8.2}$$

8.2.1 Projected Gradient Method

In a first step we consider the natural extension of the gradient descend Algorithm 8.2 which is given in the following Algorithm 8.5

The Armijo-type line-search is justified since the projection is directional differentiable (Exercise). Its convergence properties are similar to those of the gradient descend method. Hence we will not analyze this method further.

Algorithm 8.5 Projected gradient method

Choose $q_0 \in Q^{\text{ad}}$, $\rho \in (0,1)$, $\alpha \in (0,1)$, TOL > 0 and let $k = 0$.
Calculate $d_0 = -\nabla j(q_0) = b - Hq_0$
while $\|q_k - \mathscr{P}_{Q^{\text{ad}}}(q_k + d_k)\|_Q >$ TOL **do**
  Determine Step-Length: Set $l = 0$
  $q_{k+1} \leftarrow \mathscr{P}_{Q^{\text{ad}}}(q_k + \rho^l d_k)$
  while $j(q_{k+1}) > j(q_k) + \alpha(d_k, q_k - q_{k+1})_Q$ **do** ▷ Armijo type line-search
    $l \leftarrow l + 1$
    $q_{k+1} \leftarrow \mathscr{P}_{Q^{\text{ad}}}(q_k + \rho^l d_k)$
  end while
  $d_{k+1} \leftarrow -\nabla j(q_{k+1}) = b - Hq_{k+1}$
  $k \leftarrow k + 1$
end while

8.2.2 Generalized Newton Methods

We know from Section 6.3 that the solution $\overline{q}$ of (8.2) is equivalently characterized by the existence of multipliers $\mu_{\min}$ and $\mu_{\max}$ such that the following holds:

$$\begin{aligned} S^*(S\overline{q} - u^d) + \alpha\overline{q} + \mu_{\max} - \mu_{\min} &= 0 && \text{a.e. in } \Omega \\ \mu_{\min} \geq 0, \quad q_{\min} - \overline{q} &\leq 0, && \mu_{\min}(q_{\min} - \overline{q}) = 0 \text{ a.e. in } \Omega \\ \mu_{\max} \geq 0, \quad \overline{q} - q_{\max} &\leq 0, && \mu_{\max}(\overline{q} - q_{\max}) = 0 \text{ a.e. in } \Omega. \end{aligned}$$

We would like to solve this system using some fast method, like newton. To do this we need to get rid of the inequality conditions. For this we define

Definition 8.2.1. We call a function $\Psi : \mathbb{R}^2 \to \mathbb{R}$ which satisfies

$$\Psi(x,y) = 0 \quad \Leftrightarrow \quad x \leq 0, y \leq 0, xy = 0.$$

a complementarity function.

As a simple example of such a function we consider

$$\psi(x,y) = x + \max(0, cy - x)$$

with some $c > 0$ (Exercise). Now, we can reformulate the system of optimality conditions as a system of equations

$$\begin{aligned} S^*(S\overline{q} - u^d) + \alpha\overline{q} + \mu_{\max} - \mu_{\min} &= 0, \\ -\mu_{\min} + \max(0, \mu_{\min} + c(q_{\min} - \overline{q})) &= 0, \\ -\mu_{\max} + \max(0, \mu_{\max} + c(\overline{q} - q_{\max})) &= 0. \end{aligned}$$

By defining $\mu = \mu_{max} - \mu_{\min}$ we can reduce the number of equations to two and get(Exercise)

$$\begin{aligned} S^*(S\overline{q} - u^d) + \alpha\overline{q} + \mu &= 0, \\ \mu - \min(0, \mu - c(q_{\min} - \overline{q})) - \max(0, \mu + c(\overline{q} - q_{\max})) &= 0. \end{aligned}$$

We now have obtained a system of equations which is equivalent to the solution of our problem, unfortunately it is not differentiable. However we will see that we can work with less regularity to design a newton like method.

Definition 8.2.2. Let $F\colon \operatorname{dom}(F) \subset X \to Y$ be a given mapping between two Banach spaces. We call F slant (or Newton) differentiable at $x \in X$ if there exists an open neighborhood $\mathcal{N}(x)$ and a mapping $G\colon \mathcal{N}(x) \to \mathcal{L}(X,Y)$ such that

$$\lim_{\|\delta x\|_X \downarrow 0} \frac{\|F(x+\delta x) - F(x) - G(x+\delta x)\delta x\|_Y}{\|\delta x\|_X} = 0.$$

Before we continue with our analysis we will state the first result which establishes the local fast convergence of the generalized Newton method.

Theorem 8.2.3. *Let x^* be a solution to $F(x) = 0$. Assume that F is slant differentiable with derivative G. If G is non singular and $\|G(x)^{-1}\|_{\mathcal{L}(Y,X)} \leq M$ on $\mathcal{N}(x^*)$ then the Newton iteration*

$$x_{k+1} = x_k - G(x_k)^{-1}F(x_k)$$

converges superlinearly to x^* if the initial value x_0 is close enough to x^*.

Proof. Let $r > 0$ be such that $B_r(x^*) \subset \mathcal{N}(x^*)$ then by definition of slant differentiability it holds

$$\|F(x_k) - F(x^*) - G(x_k)(x_k - x^*)\|_Y \leq c_r \|x_k - x^*\|_X$$

for all $x_k \in B_r(x^*)$ and a constant $c_r \to 0$ as $r \to 0$.

We initialize the iteration such that $x_0 \in B_{r_0}(x^*)$ such that $c_{r_0} < 1/M$. Now, by definition of the iteration it holds

$$\begin{aligned}\|x_{k+1} - x^*\|_X &= \|x_k - x^* - G(x_k)^{-1}F(x_k)\|_X \\ &\leq \|G(x_k)^{-1}\|_{\mathcal{L}(Y,X)} \|G(x_k)(x_k - x^*) - F(x_k) + F(x^*)\|_Y \\ &\leq c_{\|x_k - x^*\|_X} M \|x_k - x^*\|_X.\end{aligned}$$

By induction it follows that $c_{\|x_k - x^*\|_X} M \leq c_{r_0} M < 1$. Thus we have

$$\|x_{k+1} - x^*\|_X \leq M^{k+1} \prod_{l=0}^{k} c_{\|x_l - x^*\|_X} \leq (M c_{r_0})^{k+1} \to 0.$$

The super-linear convergence then follows as $c_{\|x_k - x^*\|_X} \to 0$ with $k \to \infty$. □

Before we continue to apply this to our model problem we will start with some examples

Example 8.2.4.

1. Any bounded linear operator A is slant differentiable with $G(x) = A$.
2. In a Hilbert space H the mapping $F(u) = \|u\|_H$ is slant differentiable (exercise).
3. The map $\max(0,\cdot)$ defined as

$$\max(0,\cdot)\colon L^r(\Omega) \to L^p(\Omega), \quad u \mapsto \max(0,u)$$

is slant differentiable for all $1 \leq p < r \leq \infty$. Its slant derivative $G(u) \in \mathscr{L}(L^r(\Omega), L^p(\Omega))$ is defined by pointwise multiplication with the function

$$g_u(x) = \begin{cases} 1 & u(x) > 0 \\ \delta & u(x) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\delta \in \mathbb{R}$ is arbitrary. This means

$$(G(u)\delta u)(x) = g_u(x)\delta u(x)$$

Then taking $\delta = 0$ we can write this as

$$G(u)\delta u = \chi_{\mathscr{A}(u)}\delta u$$

where $\mathscr{A}(u) = \{x \in \Omega | u(x) > 0\}$. χ_M is called characteristic function of the set M and is defined as

$$\chi_M = \begin{cases} 1 & x \in M \\ 0 & x \notin M \end{cases}$$

To see slant differentiability let $u, \delta u \in L^r(\Omega)$ be given. Then we denote the remainder by

$$\omega_{u,\delta u}(x) = \max(0, u(x) + \delta u(x)) - \max(0, u(x)) - g_{u+\delta u}(x)\delta u(x).$$

A simple calculation shows

$$|\omega_{u,\delta u}(x)| \begin{cases} \leq |u(x)| & (u(x) + \delta u(x))u(x) < 0, \\ \leq |u(x)| & (u(x) + \delta u(x)) = 0,\ u(x) \neq 0 \\ = 0 & \text{otherwise.} \end{cases} \tag{8.3}$$

From (8.3), we obtain that we only need to consider the set

$$\Omega_0^{\delta u} = \{x \in \Omega \,|\, u(x) \neq 0, u(x)(u(x) + \delta u(x)) \leq 0\}.$$

To do this, we define the following subsets $\Omega_\varepsilon \subset \Omega_0$ as follows

$$\Omega_\varepsilon^{\delta u} = \{x \in \Omega \,|\, |u| \geq \varepsilon, u(x)(u(x) + \delta u(x)) \leq 0\}.$$

Now, $|u(x)| \geq \varepsilon$ on $\Omega_\varepsilon^{\delta u}$ and thus to get the second defining inequality we have in addition $|\delta u(x)| \geq \varepsilon$ on $\Omega_\varepsilon^{\delta u}$. Hence it holds

$$\|\delta u\|_{L^r(\Omega)} \geq \|\delta u\|_{L^r(\Omega_\varepsilon^{\delta u})} \geq \varepsilon |\Omega_\varepsilon^{\delta u}|^{1/r}.$$

In particular, for any fixed $\varepsilon > 0$ it holds

$$\lim_{\|\delta u\|_{L^r(\Omega)} \to 0} |\Omega_\varepsilon^{\delta u}| = 0. \tag{8.4}$$

We continue by defining an other set

$$\widetilde{\Omega}_\varepsilon^u = \{x \in \Omega \,|\, 0 < |u(x)| \le \varepsilon\} \supset \Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}.$$

For this it holds

$$\widetilde{\Omega}_\varepsilon^u \subset \widetilde{\Omega}_{\varepsilon'}^u, \quad 0 < \varepsilon \le \varepsilon', \qquad \text{and} \qquad \bigcap_{\varepsilon>0} \widetilde{\Omega}_\varepsilon^u = \emptyset.$$

This implies

$$\lim_{\varepsilon\to 0} |\widetilde{\Omega}_\varepsilon^u| = 0. \tag{8.5}$$

By Hölder's inequality, we get for any function $f \in L^r(\Omega)$, and $1 \le p < r \le \infty$, it holds

$$\|f\|_p \le \|1\|_s \|f\|_r = |\Omega|^{1/s} \|f\|_r, \quad s = \begin{cases} \frac{pr}{r-p} & r < \infty \\ p & r = \infty. \end{cases}$$

Then, we use (8.3) to conclude

$$\begin{aligned}
\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}} &\le \frac{1}{\|\delta u\|_{L^r}} \Big(\int_{\Omega_0^{\delta u}} |u(x)^p| \, dx \Big)^{1/p} \\
&\le \frac{1}{\|\delta u\|_{L^r}} \Big\{ \Big(\int_{\Omega_\varepsilon^{\delta u}} |u(x)^p| \, dx \Big)^{1/p} + \Big(\int_{\Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}} |u(x)^p| \, dx \Big)^{1/p} \Big\} \\
&\le \frac{1}{\|\delta u\|_{L^r}} \Big\{ |\Omega_\varepsilon^{\delta u}|^{1/s} \Big(\int_{\Omega_\varepsilon^{\delta u}} |u(x)^r| \, dx \Big)^{1/r} + |\widetilde{\Omega}_\varepsilon^u|^{1/s} \Big(\int_{\Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}} |u(x)^r| \, dx \Big)^{1/r} \Big\} \\
&\le \frac{c}{\|\delta u\|_{L^r}} \Big(\int_{\Omega_0^{\delta u}} |u(x)|^r \, dx \Big)^{1/r} \Big(|\Omega_\varepsilon^{\delta u}|^{1/s} + |\widetilde{\Omega}_\varepsilon^u|^{1/s} \Big).
\end{aligned}$$

Now, we note that on $\Omega_0^{\delta u}$ it holds $|\delta u| \ge |u|$ and hence

$$\begin{aligned}
\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}} &\le \frac{c}{\|\delta u\|_{L^r}} \|\delta u\|_{L^r} \Big(|\Omega_\varepsilon^{\delta u}|^{1/s} + |\widetilde{\Omega}_\varepsilon^u|^{1/s} \Big) \\
&\le c \Big(|\Omega_\varepsilon^{\delta u}|^{1/s} + |\widetilde{\Omega}_\varepsilon^u|^{1/s} \Big).
\end{aligned}$$

Now, due to (8.5), the second summand can be made arbitrarily small independent of δu. Further, due to (8.4), the first summand goes to zero which shows the slant differentiability, e.g.

$$\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}} \to 0, \quad (\|\delta u\|_{L^r} \to 0).$$

Remark 8.2.5. Finally, we like to note that the restriction to consider $\max(0,u)$ as a mapping from L^r into L^p with $p < r$ is not just for convenience, but in fact the function

$$g_u(x)\begin{cases}1 & u(x)\geq 0\\ 0 & \text{otherwise}\end{cases}$$

is not a slant derivative if $r = p$ (exercise).

Now, we would like to apply our knowledge to the KKT-conditions given at the beginning of this section. But before we can do so we first need the following chain rule

Theorem 8.2.6. *Let $f:\operatorname{dom}(f)\subset X\to Y$ be continuously Fréchet differentiable and $g: Y\to Z$ be slant differentiable at $f(x)$ with derivative G. Assume that $\|G\|_{\mathscr{L}(Y,Z)}\leq c$ in $\mathscr{N}(f(x))$. Then $F = g\circ f$ is slant differentiable in x and the slant derivative is given by $G(f(x+\delta x))f'(x+\delta x)\in\mathscr{L}(X,Z)$ for δx sufficiently small.*

Proof. □

For further information regarding generalized Newton methods we refer to [23, 37].

Application to the model problem

We now return to the solution of our model problem (8.1). As we have seen already in the beginning of this section the problem is equivalent to the solution of the following problem:

$$\begin{aligned} S^*(S\overline{q}-u^d)+\alpha\overline{q}+\mu &= 0,\\ \mu-\min(0,\mu-c(q_{\min}-\overline{q}))-\max(0,\mu+c(\overline{q}-q_{\max})) &= 0.\end{aligned}$$

Now the first line gives us that $\mu = -S^*(S\overline{q}-u^d)-\alpha\overline{q}$. Thus we can eliminate any direct occurrence of $\overline{q}$ in the argument of the max or min function by choosing $c=\alpha$. This gives

$$S^*(S\overline{q}-u^d)+\alpha\overline{q}+\min(0,-S^*(S\overline{q}-u^d)-\alpha q_{\min})+\max(0,-S^*(S\overline{q}-u^d)-\alpha q_{\max})=0.$$

At present it is not so obvious what we have gained by this reinterpretation, as it is still $S^*(S\overline{q}-u^d)\in L^2(\Omega)$. However we remind that S^* is defined as follows, first we solve the PDE

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

to get $z\in H_0^1(\Omega)$ and then we use the embedding Theorem 2.6.14 of $H_0^1(\Omega)\subset L^p(\Omega)$ for some $p>2$. This shows, that for $q\in L^2(\Omega)$ the element $S^*(Sq-u^d)\in L^p(\Omega)$ for some $p>2$. Further the operators $\min,\max: L^p(\Omega)\to L^2(\Omega)$ are slant differentiable and hence we can apply the chain rule Theorem 8.2.6 to obtain the following

Theorem 8.2.7. *The mapping $F : L^2(\Omega) \to L^2(\Omega)$ defined by*

$$F(q) = S^*(Sq - u^d) + \alpha q + \min(0, -S^*(Sq - u^d) - \alpha q_{\min}) + \max(0, -S^*(Sq - u^d) - \alpha q_{\max})$$

is slant differentiable with derivative $G(q)$ given by

$$G(q) = S^*S + \alpha \,\mathrm{Id} - (\chi_{\mathscr{A}(q)^-} + \chi_{\mathscr{A}(q)^+})S^*S = \chi_{\mathscr{I}(q)}S^*S + \alpha \,\mathrm{Id}\,.$$

Here the active and inactive sets are defined as

$$\begin{aligned}\mathscr{A}(q)^- &= \{x \in \Omega \,|\, -S^*(Sq - u^d) - \alpha q_{\min} < 0\},\\ \mathscr{A}(q)^+ &= \{x \in \Omega \,|\, -S^*(Sq - u^d) - \alpha q_{\max} > 0\},\\ \mathscr{I}(q) &= \Omega \setminus (\mathscr{A}(q)^- \cup \mathscr{A}(q)^+).\end{aligned}$$

This now gives rise to the following algorithm for the model problem

Algorithm 8.6 Semi-smooth Newton method for the reduced formulation

Choose $q_0 \in Q$, and let $k = 0$.
Choose $0 < \mathrm{TOL} < 1$ and $k_{\max} \in \mathbb{N}$
Calculate d_0 satisfying $G(q_0)d_0 = -F(q_0)$
Set $\delta_0 = \|F(q_0)\|_Q$
while $\delta_k > \mathrm{TOL}\,\delta_0$ and $k < k_{\max}$ **do**
  $q_{k+1} \leftarrow q_k + d_k$
  Calculate d_{k+1} satisfying $G(q_{k+1})d_{k+1} = -F(q_{k+1})$
  Set $\delta_{k+1} = \|F(q_{k+1})\|_Q$
  $k \leftarrow k + 1$
end while

We note that during the computation of the direction d_k in Algorithm 8.6 we can again use some matrix free method and compute $G(q)\delta q$ by solving first for $\delta u = S(\delta q)$ and then $\delta z = S^*\delta u$, then

$$G(u)\delta q = \chi_{\mathscr{I}(q)}\delta z + \alpha\delta q.$$

For general problems one will need to do some kind of line search to ensure global convergence of the Algorithm 8.6.

Now to obtain the convergence behavior of this method we would like to apply Theorem 8.2.3. To do so we need to show that $G(q)^{-1}$ is bounded in a neighborhood of the solution.

Lemma 8.2.8. *The operator $G(q)$ defined by Theorem 8.2.7 is invertible and the inverse satisfies*

$$\|G(q)^{-1}\|_{\mathscr{L}(L^2(\Omega))} \leq \frac{2}{\alpha} + \frac{1}{\alpha^2}\|S^*S\|_{\mathscr{L}(L^2(\Omega))}.$$

Proof. Invertibility follows from the fact that $\chi_{\mathscr{I}(q)}S^*S + \alpha\,\mathrm{Id}$ is a Fredholm-operator on $L^2(\Omega)$. However, we will obtain invertibility later in Lemma 8.2.10.

Let $q, \delta q, F \in L^2(\Omega)$ be given. Further define $\mathscr{A}(q) = \mathscr{A}(q)^- \cup \mathscr{A}(q)^+$ and $\mathscr{I}(q)$ as in Theorem 8.2.7. Now the equation $G(q)\delta q = F$ is equivalent to

$$F = \chi_{\mathscr{I}(q)} S^* S \delta q + \alpha \delta q.$$

This gives

$$\chi_{\mathscr{A}(q)} \delta q = \chi_{\mathscr{A}(q)} \frac{1}{\alpha} F.$$

Now we use this to determine $\delta q\big|_{\mathscr{I}(q)}$. To this end, we consider the relation

$$S^* S(\chi_{\mathscr{I}(q)} \delta q) + \alpha \chi_{\mathscr{I}(q)} \delta q = F - S^* S(\chi_{\mathscr{A}(q)} \delta q)$$

on $\mathscr{I}(q)$.

Then it holds

$$\begin{aligned} \alpha \int_\Omega (\chi_{\mathscr{I}(q)} \delta q)^2 \, dx &\leq \int_\Omega (S^* S(\chi_{\mathscr{I}(q)} \delta q) \chi_{\mathscr{I}(q)} \delta q \, dx + \alpha \int_\Omega (\chi_{\mathscr{I}(q)} \delta q)^2 \, dx \\ &= \int_\Omega (F - S^* S(\chi_{\mathscr{A}(q)} \delta q)(\chi_{\mathscr{I}(q)} \delta q) \, dx. \end{aligned}$$

By Hölder's inequality, we get

$$\begin{aligned} \|G(q)^{-1} F\| = \|\delta q\| &\leq \|\chi_{\mathscr{I}(q)} \delta q\| + \|\chi_{\mathscr{A}(q)} \delta q\| \\ &\leq \frac{1}{\alpha}(\|F\| + \|S^* S(\chi_{\mathscr{A}(q)} \delta q)\|) + \|\chi_{\mathscr{A}(q)} \delta q\| \\ &\leq \frac{2}{\alpha}\|F\| + \frac{1}{\alpha^2}\|S^* S\|\|F\|. \end{aligned}$$

This shows the assertion. □

Corollary 8.2.9. *Algorithm 8.6 converges locally super-linear towards the unique solution $\overline{q}$ of (8.1).*

We now need to consider an elementary proof for the invertibility of the operator $G(q)$. Further this will give us some insight to the relation of the generalized newton method to active set methods. To this end we note that the Newton step

$$G(q_k) q_{k+1} = G(q_k) q_k - F(q_k)$$

with F and G as in Theorem 8.2.7 can equivalently be expressed by reintroducing μ_{k+1} as

$$(S^* S + \alpha \, \mathrm{Id}) q_{k+1} + \mu_{k+1} = S^* u^d.$$

Then the active sets are given as

$$\begin{aligned} \mathscr{A}(q, \mu)^- &= \{x \in \Omega \,|\, \mu - \alpha(q_{\min} - q) < 0\} = \mathscr{A}(q)^-, \\ \mathscr{A}(q, \mu)^+ &= \{x \in \Omega \,|\, \mu + \alpha(q - q_{\max}) > 0\} = \mathscr{A}(q)^+. \end{aligned}$$

Then the Newton step reads as follows (Exercise)

$$(S^* S + \alpha \, \mathrm{Id}) q_{k+1} + \mu_{k+1} = S^* u^d$$

$$\mu_{k+1} - (\chi_{\mathscr{A}(q_k, \mu_k)^-} + \chi_{\mathscr{A}(q_k, \mu_k)^+})(\mu_{k+1} + \alpha q_{k+1}) = -\chi_{\mathscr{A}(q_k, \mu_k)^-} \alpha q_{\min} - \chi_{\mathscr{A}(q_k, \mu_k)^+} \alpha q_{\max}$$

In particular

$$q_{k+1} = q_{\min} \text{ on } \mathscr{A}(q_k, \mu_k)^-, \quad q_{k+1} = q_{\max} \text{ on } \mathscr{A}(q_k, \mu_k)^+, \quad \mu_{k+1} = 0 \text{ on } \mathscr{I}(q_k).$$

With these preparations it is easy to see that

Lemma 8.2.10. *The Newton step q_{k+1} given by*

$$G(q_k)q_{k+1} = G(q_k)q_k - F(q_k)$$

with F, G as in Theorem 8.2.7 is equivalently given as the unique solution of the problem

$$\begin{aligned} &\min \frac{1}{2}\|Sq - u^d\|_U^2 + \frac{\alpha}{2}\|q\|_Q^2 \\ &s.t\, q = q_{\min} \quad on\ \mathcal{A}(q_k,\mu_k)^- \\ &s.t\, q = q_{\max} \quad on\ \mathcal{A}(q_k,\mu_k)^+ \end{aligned} \tag{8.6}$$

Proof. Compare the optimality conditions with the Newton equations. We note that on $\mathcal{A}(q_k,\mu_k)^\pm$ both clearly coincide. On the set $\mathcal{I}(q_k)$ we have that the solution q of the minimization problem satisfies $S^*(Sq-u^d)+\alpha q = 0$ which is the same as q_{k+1} noting that $\mu_{k+1} = 0$ on $\mathcal{I}(q_k)$. □

In particular, we know that $G(q_k)^{-1}F(q_k)$ exists.

Finally we can use this knowledge to derive an equivalent active set method. Where we recall that with $H = S^*S - \alpha\,\mathrm{Id}$ and $b = S^*u^d$ we have $\mu_{k+1} = b - Hq_{k+1}$.

Algorithm 8.7 primal-dual active set method (PDAS)

Choose $q_0 \in Q$, and let $k = 0$.
Set $\mu_0 = b - Hq_0$.
Choose $0 < \mathrm{TOL} < 1$ and $k_{\max} \in \mathbb{N}$
Compute $\mathcal{A}_0^- = \mathcal{A}(q_0,\mu_0)^-$, $\mathcal{A}_0^+ = \mathcal{A}(q_0,\mu_0)^+$, $\mathcal{I}_0 = \Omega \setminus (\mathcal{A}_0^- \cup \mathcal{A}_0^+)$.
Set $\delta_0 = \alpha^2\|q_0 - q_{\min}\|^2_{\mathcal{A}_0^-} + \alpha^2\|q_0 - q_{\max}\|^2_{\mathcal{A}_0^+} + \|\mu_0\|^2_{\mathcal{I}_0}$
while $\delta_k > \mathrm{TOL}\,\delta_0$ and $k < k_{\max}$ **do**
  Find q_{k+1} solving (8.6)
  Set $\mu_{k+1} = b - Hq_{k+1}$.
  Compute $\mathcal{A}_{k+1}^- = \mathcal{A}(q_{k+1},\mu_{k+1})^-$, $\mathcal{A}_{k+1}^+ = \mathcal{A}(q_{k+1},\mu_{k+1})^+$, $\mathcal{I}_{k+1} = \Omega \setminus (\mathcal{A}_{k+1}^- \cup \mathcal{A}_{k+1}^+)$.
  Set $\delta_{k+1} = \alpha^2\|q_{k+1} - q_{\min}\|^2_{\mathcal{A}_{k+1}^-} + \alpha^2\|q_{k+1} - q_{\max}\|^2_{\mathcal{A}_{k+1}^+} + \|\mu_{k+1}\|^2_{\mathcal{I}_{k+1}}$
  $k \leftarrow k+1$
end while

The term δ_k in the stopping criterion is chosen since due to the definition of $\mathcal{A}_{k+1}^\pm$ and $\mathcal{I}_{k+1}$ we have

$$\begin{aligned} \delta_{k+1} &= \alpha^2\|q_{k+1} - q_{\min}\|^2_{\mathcal{A}_{k+1}^-} + \alpha^2\|q_{k+1} - q_{\max}\|^2_{\mathcal{A}_{k+1}^+} + \|\mu_{k+1}\|^2_{\mathcal{I}_{k+1}} \\ &= \|\mu - \min(0, \mu - \alpha(q_{\min} - q_{k+1})) - \max(0, \mu + \alpha(q_{k+1} - q_{\max})\|^2_{\mathcal{A}_{k+1}^-} \\ &\quad + \|\mu - \min(0, \mu - \alpha(q_{\min} - q_{k+1})) - \max(0, \mu + \alpha(q_{k+1} - q_{\max})\|^2_{\mathcal{A}_{k+1}^+} \\ &\quad + \|\mu - \min(0, \mu - \alpha(q_{\min} - q_{k+1})) - \max(0, \mu + \alpha(q_{k+1} - q_{\max})\|^2_{\mathcal{I}_{k+1}} \\ &= \|F(q_{k+1})\|^2. \end{aligned}$$

Remark 8.2.11. Finally we note that in the finite dimensional case with $Q_h^{(0)}$ or $Q_h^{(1)}$ we can determine the active sets by the values of the coefficients for the Lagrange basis. This means if $q_k = \sum_{i=0}^{M} q_k^i \hat{\varphi}_h^{(i)}$ and $\mu_k = \sum_{i=0}^{M} \mu_k^i \hat{\varphi}_h^{(i)}$, then

$$\begin{aligned}
\mathscr{A}_k^- &= \{i \in \{1,\dots,M\} \,|\, \mu_k^{(i)} - \alpha(q_{\min} - q_k^{(i)}) < 0\},\\
\mathscr{A}_k^+ &= \{i \in \{1,\dots,M\} \,|\, \mu_k^{(i)} + \alpha(q_k^{(i)} q_{max}) > 0\},\\
\mathscr{I}_k &= \{1,\dots,M\} \setminus (\mathscr{A}_k^- \cup \mathscr{A}_k^-).
\end{aligned}$$

Remark 8.2.12. Note that the iterates q_k are infeasible. To obtain feasible iterates one could use specialized active set methods or barrier methods to assert feasibility of the iterates, we will discuss this in the context of state constraints.

Remark 8.2.13. In the case of nonlinear problems, one can again use both the generalized newton method, or some SQP method and solve the inner problems using primal-dual active set methods.

9 Optimization in Banach spaces III

Now, we want to consider additional constraints in the context of PDE constrained optimization as for instance pointwise constraints on the state variable. In contrast to the case of box constraints on the control, we will no longer be able to define Lagrange multipliers for our problems by hand. Instead, we will need some additional theory.

Hence we will now consider the general problem stated in the introduction

$$\begin{aligned} J(q,u) &\to \min \\ \text{s.t. } & q \in Q^{\text{ad}}, \\ & A(q,u) = 0, \\ & \hat{g}(q,u) \in -W^{\text{ad}}, \end{aligned} \tag{9.1}$$

with reflexive Banach spaces Q, U, Z, and a Banach space W, mappings $A\colon Q \times U \to Z^*$, $\hat{g}: Q \times U \to W$ and $J: Q \times U \to \mathbb{R}$, a closed and convex set $\emptyset \neq Q^{\text{ad}} \subset Q$ and a closed convex cone $\emptyset \neq W^{\text{ad}} \subset W$.

We already know from Section 5.2, in particular Theorem 5.2.3 that if J is continuously Fréchet-differentiable, then a local solution $(\overline{q}, \overline{u})$ of (9.1) satisfies

$$J'(\overline{q},\overline{u})(q,u) \geq 0 \quad \forall\, (q,u) \in T(X^{\text{ad}}, (\overline{q},\overline{u})),$$

with

$$X^{ad} = \{(q,u) \in Q \times U \,|\, q \in Q^{\text{ad}},\ A(q,u) = 0,\ \hat{g}(q,u) \in -W^{\text{ad}}\}$$

where $T(X^{\text{ad}}, (\overline{q},\overline{u}))$ is the sequential tangent cone given in Definition 5.2.2. Now, we would like to find corresponding Lagrange multipliers to get a better characterization of this necessary optimality condition.

To do so, we will again consider an abstract problem. Before we can state this, we start with some definitions

Definition 9.0.1. Let Y be a Banach space. A convex set $K \subset Y$ is called a convex cone if for any $y \in K$ and any $\lambda \geq 0$ it holds $\lambda y \in K$.

Definition 9.0.2. Let $K \subset Y$ be a convex cone. We write $y \geq_K 0$ if $y \in K$. (Analog $y \leq_K 0$ if $-y \in K$.)

This defines a partial ordering $\geq_K$ on Y.

Example 9.0.3.

- We can consider the set $K = \{x \in \mathbb{R}^2 \,|\, x_1 = 0\}$. This is a convex cone. Further it holds

$$x \geq_K 0 \quad \Leftrightarrow \quad x_1 = 0.$$

  In particular, there are elements in $\mathbb{R}^2$ that are neither $\leq_K 0$ nor $\geq_K 0$.

- Consider the set $\mathcal{S}^n$ of symmetric $n \times n$ matrices ($A \in \mathcal{S}^n$ iff $A = A^T$). Now, one can define the cone of positive semidefinite matrices (exercise)

$$\mathcal{S}^n_+ = \{A \in \mathcal{S}^n \,|\, x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n\}.$$

Now, our model problem can be stated as follows

$$\begin{aligned} f(x) &\to \min \\ \text{s.t. } x &\in C \\ g(x) &\leq_K 0. \end{aligned} \tag{9.2}$$

Where $f : X \to \mathbb{R}$, $g : X \to Y$ are both continuously Fréchet-differentiable, $C \subset X$ closed and convex, and $K \subset Y$ a closed and convex cone.

Example 9.0.4. Our model problem fits into this framework, in several ways. First, one can consider $f = J$, $g = (A, \hat{g})$, $X = Q \times U$, $Y = Z^* \times W$, $C = Q^{\text{ad}} \times \{0\}$, and $K = \{0\} \times W^{\text{ad}}$.

On the other hand, one can use the solution operator S (if it exists) to consider $X = Q$, $Y = W$, $f(q) = j(q) = J(q, S(q))$, $g(q) = \hat{g}(q, S(q))$, $C = Q^{\text{ad}}$, and $K = W^{\text{ad}}$.

Further, we can use a convex cone $K \subset Y$ to define non-negativity in the dual space Y^*.

Definition 9.0.5. Let $K \subset Y$ be a convex cone. Then we define the dual cone $K^+ \subset Y^*$ as follows

$$K^+ = \{y^* \in Y^* \,|\, \langle y, y^* \rangle_{Y,Y^*} \geq 0 \; \forall \, y \in K\} = \{y^* \in Y^* \,|\, \langle y, y^* \rangle_{Y,Y^*} \geq 0 \; \forall \, y \geq_K 0\}$$

Example 9.0.6.

- Consider $Y = L^2(\Omega)$ on a bounded domain Ω with the convex cone

$$K = \{y \in Y \,|\, y(x) \geq 0 \text{ for a.e. } x \in \Omega\}.$$

  Then by Theorem 2.3.7 it is $Y^* = Y$ and it holds $K^+ = K$ (exercise).

- Let Y be arbitrary and $K = \{0\}$. Then $y \geq_K 0$ iff $y = 0$ and hence $K^+ = Y^*$ since $\langle 0, y^* \rangle_{Y,Y^*} = 0$ for all $y^* \in Y^*$.

- Let Y be arbitrary and $K = Y$. Then $y \geq_K 0$ for all $y \in Y$ and hence $K^+ = \{0\}$ since $\langle y, y^* \rangle_{Y,Y^*} \geq 0$ for all $y \in Y$ only if $y^* = 0$.

Definition 9.0.7. Let $C \subset X$ be a convex subset of a Banach space X. Then we define the conical hull $C(x)$ to C in $x \in X$ as

$$C(x) = \{\lambda(c - x) \,|\, \lambda \geq 0,\ c \in C\}.$$

Remark 9.0.8. We have already seen in the exercises that we can now express the necessary optimality conditions (5.2)

$$(\nabla j(\overline{q}), q - \overline{q}) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}$$

in the abbreviated form

$$(\nabla j(\overline{q}), q) \geq 0 \quad \forall q \in Q^{\mathrm{ad}}(\overline{q})$$

or even shorter

$$j'(\overline{q}) \in Q^{\mathrm{ad}}(\overline{q})^+$$

Now, we can define Lagrange multipliers.

Definition 9.0.9. An element $\overline{\mu}$ is called a Lagrange multiplier for (9.2) at a solution $\overline{x}$ if $\overline{\mu} \in K^+$ and

$$f'(\overline{x}) + \overline{\mu} \circ g'(\overline{x}) \in C(\overline{x})^+,$$
$$\langle g(\overline{x}), \overline{\mu} \rangle_{Y,Y^*} = 0.$$

Here, the first inclusion means

$$f'(\overline{x})(\delta x) + \langle g'(\overline{x})\delta x, \overline{\mu} \rangle_{Y,Y^*} \geq 0 \quad \forall \delta x \in C(\overline{x})$$

Remark 9.0.10. The above definition for a Lagrange multiplier has the following reasoning: If one defines the Lagrange functional

$$\mathscr{L}(x, \mu) = f(x) + \langle g(x), \mu \rangle_{Y,Y^*}$$

then a pair $(\overline{x}, \overline{\mu}) \in X \times K^+$ is a saddle point of $\mathscr{L}$ if

$$\mathscr{L}(\overline{x}, \mu) \leq \mathscr{L}(\overline{x}, \overline{\mu}) \leq \mathscr{L}(x, \overline{\mu}) \quad \forall\, x \in C, \mu \in K^+.$$

In particular, if $\overline{x}$ is feasible for (9.2) and $(\overline{x}, \overline{\mu})$ is a saddle point, then $\overline{x}$ solves (9.2) (exercise).

Now, the conditions given in Definition 9.0.9 are necessary for $(\overline{x},\overline{\mu})$ to be a saddle point of the Lagrange functional. Further, if f and g are convex, then the conditions are also sufficient (exercise).

Now, we define some tangent cones, the first one was already given in Definition 5.2.2; where we defined the sequential tangent cone on $X^{\mathrm{ad}} = C \cap \{x \in X \mid g(x) \leq_K 0\}$

$$T(X^{\mathrm{ad}},\overline{x}) = \{x \in X \mid x = \lim_{k\to\infty} \frac{1}{t_k}(x_k - \overline{x}), t_k \downarrow 0, x_k \in X^{\mathrm{ad}}\}.$$

The second, we have encountered, too, but not yet given a name:

Definition 9.0.11. The cone

$$L(X^{\mathrm{ad}},\overline{x}) = \{x \in X \mid x \in C(\overline{x}),\ g'(\overline{x})x \in -K(-g(\overline{x}))\}$$

is called the linearizing cone.

As in, finite dimensional, nonlinear optimization; it is always

$$T(X^{\mathrm{ad}},\overline{x}) \subset L(X^{\mathrm{ad}},\overline{x})$$

but equality need not hold.

As we have already seen, in the the exercises, the linearizing cone is contained in the sequential tangent cone in the case of box constraints on the control, which means that g did not appear. To obtain the same statement in our case, we need to consider the following constraint qualification due to Zowe and Kurcyusz

Definition 9.0.12. We say that a point $\overline{x} \in X^{\mathrm{ad}}$ is regular if it holds

$$g'(\overline{x})C(\overline{x}) + K(-g(\overline{x})) = Y$$

or equivalently (see, [23, Sec 1.3]

$$0 \in \mathrm{int}(g'(\overline{x})(C - \overline{x}) + K + g(\overline{x}))$$

Theorem 9.0.13 (Lyusternik). *Let $g : X \to Y$ be continuously differentiable between two Banach spaces X, Y. Further let $X^{ad} = C \cap \{x \in X \mid g(x) \leq_K 0\} \subset X$ be given by a closed convex set C and a closed convex cone K. Assume that $\overline{x} \in X^{ad}$ is regular, then it holds*

$$\emptyset \neq L(X^{ad},\overline{x}) \subset T(X^{ad},\overline{x}).$$

Proof. See [39, Thm 5.2.5]. □

Before we come to the main result of this section we recall the well known separation theorem

Theorem 9.0.14 (Hahn-Banach). *Let Y be a Banach space, $V_1, V_2 \subset Y$ convex with V_1 open. Let $V_1 \cap V_2 = \emptyset$. Then there exists some $y^* \in Y^*$ such that*

$$\langle v_1, y^* \rangle_{Y,Y^*} < \langle v_2, y^* \rangle_{Y,Y^*} \quad \forall v_1 \in V_1, v_2 \in V_2.$$

Proof. See [38, Thm III.2.4]. □

Theorem 9.0.15. *Let a solution $\overline{x}$ to (9.2) be regular in the sense of Definition 9.0.12. Then there exists an associated Lagrange multiplier $\overline{\mu} \in Y^*$.*

Proof. We can define the set

$$A = \{(f'(\overline{x})x + r, g'(\overline{x})x + y) \,|\, x \in C(\overline{x}),\ r \geq 0,\ y \in K(-g(\overline{x}))\} \subset \mathbb{R} \times Y.$$

It is clear that A is a convex cone (exercise).

By Theorem 5.2.3 and Theorem 9.0.13, we have

$$f'(\overline{x})x \geq 0 \quad \forall x \in L(X^{\mathrm{ad}}, \overline{x}) \neq \emptyset.$$

This means that if $x \in L(X^{\mathrm{ad}}, \overline{x})$ then $f'(\overline{x})x + r \geq 0$. Otherwise, if $x \in C(\overline{x}) \setminus L(X^{\mathrm{ad}}, \overline{x})$ then $g'(\overline{x})x - y \neq 0$ and thus

$$A \subset (\mathbb{R}_{\geq 0} \times Y) \cup (\mathbb{R} \times Y \setminus \{0\})$$

and thus $(0,0)$ is a boundary point of A. Now, for any $s > 0$, we can take $x \in B_s(0)$, and see that $\|f'(\overline{x})x\| \leq c_s$ for a constant depending on s only. Hence, it follows that

$$A \supset M_s := \{c \geq c_s\} \times \{g'(\overline{x})x + y \,|\, x \in C(\overline{x}) \cap B_s(0),\ y \in K(-g(\overline{x}))\}.$$

By assumed regularity of $\overline{x}$, see Definition 9.0.12, it holds $\mathrm{int}(A) \supset \mathrm{int}(M_s) \neq \emptyset$.

Now, we employ Theorem 9.0.14 to the disjoint sets $\mathrm{int}(A)$ and $(0,0)$. Thus there exists $\alpha \in \mathbb{R}$ and $\overline{\mu} \in Y^*$ such that

$$\alpha a_1 + \langle a_2, \overline{\mu} \rangle_{Y,Y^*} > 0 \quad \forall (a_1, a_2) \in \mathrm{int}(A). \tag{9.3}$$

More concrete, it holds

$$\alpha(f'(\overline{x})x + r) + \langle g'(\overline{x})x + y, \overline{\mu} \rangle_{Y,Y^*} \geq 0 \quad \forall x \in C(\overline{x}),\ r \geq 0,\ y \in K(-g(\overline{x})).$$

First of all we pick $(x, r) = 0$. Then it follows by definition of $K(-g(\overline{x}))$

$$\langle k + g(\overline{x}), \overline{\mu} \rangle_{Y,Y^*} \geq 0 \quad \forall k \in K.$$

This means ($k \in K$ then also $\lambda k \in K$)

$$\lambda \langle k, \overline{\mu} \rangle_{Y,Y^*} \geq \langle -g(\overline{x}), \overline{\mu} \rangle_{Y,Y^*} > -\infty \quad \forall\ \lambda \geq 0.$$

In particular, taking $\lambda \to 0$, we get $\langle -g(\overline{x}), \overline{\mu} \rangle_{Y,Y^*} \leq 0$. Now, taking $\lambda \to \infty$, we get $\langle k, \overline{\mu} \rangle_{Y,Y^*} \geq 0$ for all $k \in K$ or $\overline{\mu} \in K^+$. Thus, by noting that $-g(\overline{x}) \in K$, we get $\langle g(\overline{x}), \overline{\mu} \rangle_{Y,Y^*} = 0$.

To see the inclusion $f'(\overline{x})+\overline{\mu}\circ g'(\overline{x}) \in C(\overline{x})^+$, we note that taking $(x,y)=(0,0)$ in (9.3) yields $\alpha \geq 0$. If one assumes for contradiction that $\alpha = 0$ one obtains

$$\langle g'(\overline{x})x + y, \overline{\mu}\rangle_{Y,Y^*} \geq 0 \quad \forall x \in C(\overline{x}),\ y \in K(-g(\overline{x})).$$

By assumption of regularity, see Definition 9.0.12 one has that $g'(\overline{x})C(\overline{x}) + K(-g(\overline{x})) = Y$ and hence $\overline{\mu} = 0$ in contradiction to the strict inequality sign in (9.3). Hence $\alpha > 0$ and we can assume w.l.o.g. that $\alpha = 1$. Now, taking $(r,y)=(0,0)$ yields

$$f'(\overline{x})x + \langle g'(\overline{x})x, \overline{\mu}\rangle_{Y,Y^*} = f'(\overline{x})x + \overline{\mu}\circ g'(\overline{x})x \geq 0 \quad \forall x \in C(\overline{x}).$$

This shows that $f'(\overline{x}) + \overline{\mu}\circ g'(\overline{x}) \in C(\overline{x})^+$. □

9.1 Constraint Qualifications

The problem with the above Theorem 9.0.15 is that for general (nonlinear) operators f, g it is difficult to check whether the solution –which we do not know– is a regular point. To this end there have been derived several conditions that ensure regularity.

Convex Problems

The most simple one is the following slater condition

Theorem 9.1.1. *Let f, g be convex, and let $\overline{x}$ be a solution to the problem (9.2). Let there be some element $\tilde{x} \in C$ such that*

$$-g(\tilde{x}) \in \operatorname{int}(K) \qquad \textit{in short: } g(\tilde{x}) <_K 0.$$

Then there exists a Lagrange-multiplier $\overline{\mu}$ at the solution $\overline{x}$ for the problem (9.2).

A noteworthy consequence of this can be seen if one considers the problem, with j as given in our model problem

$$\begin{aligned} &\min j(q) \\ &\text{s.t. } q \in L^2(\Omega),\ q \leq 0. \end{aligned}$$

Now if we rewrite the problem in the form of (9.2) then we can choose $C = \{v(x) \leq 0\}$ and $g = \mathrm{id}$, $K = L^2(\Omega)$. Then the conditions of the theorem are met, but the Lagrange multiplier is meaningless (it corresponds to the constraint $q \in L^2(\Omega)$). But we can directly compute the multipliers for the positivity constraint as we have seen.

On the other hand if we choose $C = L^2(\Omega)$ and $K = \{v(x) \geq 0\}$ then unfortunately $\operatorname{int}(K) = \emptyset$ (exercise) and hence we can no longer apply the Theorem 9.1.1.

Now we want to discuss how one can use the above result in different situations

Equality constraints

If we want to deal with equality constraints, i.e. $g(x) = 0$, then one takes $C = X$ and $K = \{0\}$. The constraint qualification then reads

$$g'(\overline{x})X = Y$$

which means that $g'(\overline{x})$ needs to be surjective. Then there exists a Lagrange multiplier $\overline{\mu}$ and since $C(\overline{x}) = X$ it holds

$$f'(\overline{x}) + \overline{\mu} \circ g'(\overline{x}) \in X^+ = \{0\}.$$

Inequality constraints

Before we start with the discussion of this case we note that the constraint qualification

$$g'(\overline{x})C(\overline{x}) + K(-g(\overline{x})) = Y$$

is equivalent to the fact that for any $y \in Y$ there exist $\lambda_1, \lambda_2 \geq 0$, $x \in C$ and $k \in K$ such that

$$\lambda_1 g'(\overline{x})(x - \overline{x}) + \lambda_2(k + g(\overline{x})) = y$$

Now if we have some inequality constrained problem the following linearized slater condition

$$\exists \tilde{x} \in C : g(\overline{x}) + g'(\overline{x})(\tilde{x} - \overline{x}) <_K 0$$

ensures that $\overline{x}$ is regular. To see this, we set $\lambda_1 = \lambda_2 = \lambda$ and then for any $y \in Y$ need to find λ, x, k such that

$$\lambda(g'(\overline{x})(x - \overline{x}) + k + g(\overline{x})) = y.$$

We take $x = \tilde{x}$ and define $\overline{y} = g(\overline{x}) + g'(\overline{x})(\tilde{x} - \overline{x})$ to get

$$\lambda k + \lambda \overline{y} = y.$$

Now, K is a cone and hence we can also find $k \in K$ such that

$$k + \lambda \overline{y} = y.$$

Since $\overline{y} <_K 0$, we can choose λ sufficiently large such that

$$y - \lambda \overline{y} \geq_K 0.$$

Then taking $k = y - \lambda \overline{y}$ solves the problem.

We note that again for this condition we need that K has interior points.

If both equality and inequality constraints appear, we can mix the corresponding conditions, see, e.g., [36].

10 Linear-quadratic elliptic optimization problems III

Now we want to use what we have learned on the following augmented optimization problem:

$$\begin{aligned} \min J(q,u) &= \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } & u \in H_0^1(\Omega), \\ & (\nabla u, \nabla \varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ & q_{\min} \le q \le q_{\max} \quad \text{a.e. on } \Omega, \\ & u \le u_{\max} \quad \text{on } \Omega. \end{aligned} \tag{10.1}$$

In this, we note that due to the regularity results in Section 4.2 we know that on a smooth or polygonally bounded domain Ω the solution u satisfies for any right hand side $q \in L^2(\Omega)$ that $u \in H_0^1(\Omega) \cap C(\overline{\Omega})$. In particular, we can define the operator $u \mapsto u - u_{\max}$. It is continuously Fréchet-differentiable as operator $C(\overline{\Omega}) \to C(\overline{\Omega})$. The cone

$$K = \{v \in C(\overline{\Omega}) \,|\, v(x) \ge 0,\ x \in \overline{\Omega}\}$$

has interior points. Hence, we can formulate this in the context of (9.2) by choosing S as solution operator of the PDE and setting $X = L^2(\Omega)$, $Y = C(\overline{\Omega})$, $f(q) = j(q) = J(q, Sq)$, $g(q) = Sq - u_{\max}$, and $C = Q^{\text{ad}}$. Now, let us assume that the linearized Slater condition is satisfied. To understand what this means, we note that

$$\operatorname{int}(K) = \{v \in C(\overline{\Omega}) \,|\, v(x) > 0,\ x \in \overline{\Omega}\}.$$

We see that the linearized Slater condition would be satisfied if there is a point $\tilde{q} \in Q^{\text{ad}}$ with $S\tilde{q} < u_{\max}$ as then clearly

$$g(\overline{q}) + g'(\overline{q})(\tilde{q} - \overline{q}) = S\overline{q} - u_{\max} + S(\tilde{q} - \overline{q}) = S\tilde{q} - u_{\max} <_K 0.$$

Under this assumption, using the results of the previous section, the existence of a Lagrange multiplier $\overline{\mu} \in K^+$ follows.

10.1 A side note on measures

To understand more clearly, what this means, we note that the space $C(\overline{\Omega})^*$ can be identified with the space $M(\overline{\Omega})$ of regular Borel measures (or Radon measures) by the following definition (Theorem of Riesz-Radon)

$$\langle u, \mu \rangle_{C,C^*} = \int_\Omega u(x)\, d\mu.$$

We would like to understand what such measures are. To this end, we let $\mathcal{B}$ be the smallest σ-algebra containing all closed and open subsets of $\overline{\Omega}$ (the so called Borel sets). This means

$$\begin{aligned} S \in \mathcal{B} &\Rightarrow \overline{\Omega} \setminus S \in \mathcal{B} \\ S_i \in \mathcal{B},\ i \in \mathbb{N} &\Rightarrow \bigcup_{i \in \mathbb{N}} S_i \in \mathcal{B} \end{aligned}$$

Definition 10.1.1. We say that a function $\mu\colon \mathscr{B} \to \mathbb{R}$ is a Borel measure if μ is σ-additive and it is of bounded variation, i.e., if S_i are pairwise disjoint sets for $i \in \mathbb{N}$ then

$$\mu\big(\cup_{i\in\mathbb{N}} S_i\big) = \sum_{i\in\mathbb{N}} \mu(S_i)$$

and $\|\mu\|_{\text{var}} < \infty$ where the variation of μ is given as

$$\|\mu\|_{\text{var}} = |\mu|(\overline{\Omega}) = \sup\Big\{\sum_{i\in\mathbb{N}} |\mu(S_i)| \,\Big|\, S_i \in \mathscr{B} \text{ pairwise disjoint, } S_i \subset \overline{\Omega}\Big\}.$$

For such a measure and a function $f : \overline{\Omega} \to \mathbb{R}$ one defines the integral $f \mapsto \int_\Omega f \, d\mu$ as usual by defining the integral for step-functions, then approximation of arbitrary functions by step functions.

Definition 10.1.2. A measure μ is called regular if for any $S \in \mathscr{B}$ it holds

$$\inf\Big\{|\mu|(O \setminus C) \,|\, C \subset S \subset O,\ C \text{ closed},\ O \text{ open}\Big\} = 0$$

Example 10.1.3. Typical examples for such measures are the Dirac-measures δ_y for given $y \in \overline{\Omega}$ which are defined as

$$\int_\Omega f(x)\, d\delta_y = f(y) \quad \forall f \in C(\overline{\Omega}).$$

Other typical measures are line measures, e.g., let $\gamma \subset \overline{\Omega}$ be a curve then we can define μ_γ as

$$\int_\Omega f(x)\, d\mu_\gamma = \int_\gamma f(s)\, ds \quad \forall f \in C(\overline{\Omega}).$$

Finally for a function $g \in L^1(\Omega)$ one can define μ_g as

$$\int_\Omega f(x)\, d\mu_g = \int_\Omega f(x) g(x), dx \quad \forall f \in C(\overline{\Omega}).$$

10.2 Recovering the complete KKT-conditions

We recall that Theorem 9.0.15 asserts; for a solution $\overline{q}$ of (10.1) there exists $\mu \in M(\overline{\Omega}) = C(\overline{\Omega})^*$ such that

$$\overline{\mu} \in K^+, \qquad \text{i.e.,} \int_\Omega f \, d\overline{\mu} \geq 0 \quad \forall f \geq 0, f \in C(\overline{\Omega}),$$

$$\langle \overline{u} - u_{\max}, \overline{\mu} \rangle_{C,C^*} = 0,$$
$$j'(\overline{q}) + \overline{\mu} \circ g'(\overline{u}) \circ S \in Q^{\text{ad}}(\overline{q})^+.$$

Again, the third inclusion can be rewritten as

$$\alpha(\overline{q}, q - \overline{q}) + (\overline{u} - u^d, S(q - \overline{q})) + \int_\Omega S(q - \overline{q}) \, d\overline{\mu} \geq 0 \quad \forall q \in Q^{\text{ad}}.$$

To get rid of the operator S applied to the direction $q - \overline{q}$, we can, once again, define our adjoint variable $\overline{z}$ as solution to the adjoint problem

$$\overline{z} = S^*(\overline{u} - u^d) + S^*(\overline{\mu}).$$

We have already seen the first term on the right hand side. However, the second term is new to us. To see that this is well defined, we note that $S \colon L^2(\Omega) \to C(\overline{\Omega})$ as already remarked at the beginning of this section. Hence, by definition, the adjoint operator $S^* \colon M(\overline{\Omega}) \to L^2(\Omega)^* = L^2(\Omega)$.

Remark 10.2.1. By a more thorough analysis one can obtain that in fact $S^* \colon M(\overline{\Omega}) \to W_0^{1,s'}(\Omega)$ is a bounded linear operator for any $s' < 2$ if $\Omega \subset \mathbb{R}^2$ is a $C^{1,1}$ domain, see [6].

We will use this and obtain that, in fact, $\overline{z}$ satisfies the equation

$$(\nabla \varphi, \nabla \overline{z}) = (\overline{u} - u^d, \varphi) + \int_\Omega \varphi \, d\overline{\mu} \quad \forall \varphi \in W_0^{1,s}(\Omega)$$

for any $s > 2$. Now, noting that by the embedding Theorem 2.6.14, we know that $S(q - \overline{q}) \in W^{1,s}(\Omega)$ for some $s > 2$. Thus, we get, using $(\nabla S q, \nabla z) = (q, z)$,

$$\alpha(\overline{q}, q - \overline{q}) + (q - \overline{q}, \overline{z}) = \alpha(\overline{q}, q - \overline{q}) + (\overline{u} - u^d, S(q - \overline{q})) + \int_\Omega S(q - \overline{q}) \, d\overline{\mu} \geq 0 \quad \forall q \in Q^{\text{ad}}.$$

Now, analog to Section 6.3 we can use the inequality to obtain existence of Lagrange multipliers $\mu_{\min}$ and $\mu_{\max}$. Finally, we obtain the following system of KKT-conditions

$$\begin{aligned}
(\nabla \overline{u}, \nabla \varphi) &= (\overline{q}, \varphi) && \forall\, \varphi \in H_0^1(\Omega), \\
(\nabla \varphi, \nabla \overline{z}) &= (\overline{u} - u^d, \varphi) + \int_\Omega \varphi \, d\overline{\mu} && \forall\, \varphi \in W_0^{1,s}(\Omega), \\
\overline{z} + \alpha \overline{q} + \mu_{\max} - \mu_{\min} &= 0 && \text{a.e. in } \Omega, \\
\mu_{\min} \geq 0, \quad q_{\min} - \overline{q} \leq 0, & \quad \mu_{\min}(q_{\min} - \overline{q}) = 0 && \text{a.e. in } \Omega, \\
\mu_{\max} \geq 0, \quad \overline{q} - q_{\max} \leq 0, & \quad \mu_{\max}(\overline{q} - q_{\max}) = 0 && \text{a.e. in } \Omega, \\
\overline{\mu} \in K^+, \quad \overline{u} - u_{\max} \leq 0, & \quad \langle \overline{u} - u_{\max}, \overline{\mu} \rangle_{C,C^*} = 0. &&
\end{aligned} \tag{10.2}$$

We will see in the exercises, that indeed the appearance of the space $M(\overline{\Omega})$ is not a flaw in the analysis, but indeed with the given assuptions on the problem data can not be avoided. More regularity of the multipliers can be shown, if, e.g., $u^d \in L^\infty$ as then $\mu \in H^{-1}(\Omega)$ is the worst possible case, see, [7].

11 Regularization

As we will see in the exercises, the direct solution of the discretized state constraint problem (10.1) will in general be difficult as $h \to \infty$. We will now derive some possibilities to reformulate the desired problem as limit of problems without state constraint. –Of course one can apply these techniques to remove control constraints as well–

To this end, we rewrite (10.1) as

$$\begin{aligned} \min J(q,u) + I_K(u_{\max} - u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2 + I_K(u_{\max} - u) \\ \text{s.t. } u \in H_0^1(\Omega), \\ (\nabla u, \nabla \varphi) = (q, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega) \end{aligned} \tag{11.1}$$

with $K = \{v \in C(\overline{\Omega}) \,|\, v(x) \geq 0,\ x \in \overline{\Omega}\}$ and the indicator function

$$I_K(v) = \begin{cases} 0 & v \in K, \\ \infty & \text{otherwise.} \end{cases}$$

In order to keep the presentation simple we have skiped the bounds on the constraints, although they could be incorporated.

We will assume throughout that the control and state constraints are compatible in the sense that there exists some Slater point $\tilde{q} \in Q^{\text{ad}}$ such that $\tilde{u} = S\tilde{q} < u_{\max}$. To simplify the presentation, we assume throughout that $\tilde{q} = 0$ and thus $0 < u_{\max}$ on $\overline{\Omega}$.

To obtain a regularized version of this problem one considers some sequence of sufficiently nice functions R_γ such that $R_\gamma \to I_K$ as $\gamma \to \infty$ in a sense that will become clear in the sequel. Then the regularized problems read as follows

$$\begin{aligned} \min J_\gamma(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2 + R_\gamma(u_{\max} - u) \\ \text{s.t. } u \in H_0^1(\Omega), \\ (\nabla u, \nabla \varphi) = (q, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega). \end{aligned} \tag{11.2}$$

There are two popular possibilities to choose R_γ.

One are so called barrier-methods which define some functional $R_\gamma = b_\gamma$ where b_γ is smooth in the interior of K and

$$b_\gamma(v) = \infty \text{ if } v \notin K \quad \text{and} \quad b_\gamma(v) \to \infty \text{ if } v \in K, v \to \partial K.$$

Clearly, we need for any fixed $v \in \text{int}(K)$ that $b_\gamma(v) \to 0$ as $\gamma \to \infty$. The motivation for such a functional is that one hopes to obtain iterates that are strictly feasible, i.e., $v \in \text{int}(K)$ so that the problem is differentiable. A typical choice for such a barrier functional on $K = \{v > 0\}$ would be

$$b_\gamma(v) = \frac{-1}{\gamma} \int_\Omega \ln(v(x))\, dx \quad v \in K$$

in analogy to the usual logarithmic barriers in finite dimensions. However, in contrast to the finite dimensional case the solutions to (11.2) will be feasible, but not strictly feasible in general. This is because we have

$$\int_\Omega -\log(v(x))\,dx < \infty$$

which gives $v(x) > 0$ almost everywhere. A way to circumvent this problem is the choice of higher order barrier functions like

$$b_\gamma(v) = \frac{1}{\gamma^p(p-1)} \int_\Omega \frac{1}{v(x)^{p-1}}\,dx \quad v \in K$$

for some $p > 1$. It can be shown that such a choice will yield strictly feasible states if p is sufficiently large, see [33, 34].

The second possibility are so called penalty-methods where one chooses $R_\gamma = p_\gamma$ with a function $p_\gamma = 0$ in K and

$$p_\gamma(v) \to \infty \text{ as } \gamma \to \infty \text{ if } v \notin K.$$

A typical choice would be the quadratic penalty or Moreau-Yosida regularization (Exercise) given by

$$p_\gamma(u_{\max} - u) = \frac{\gamma}{2}\|(u - u_{\max})^+\|^2.$$

This method will give infeasible iterates, in general. Moreover the problem is no longer smooth due to the nonsmooth term $(u - u_{\max})^+ = \max(0, u - u_{\max})$. It will however yield a slant differentiable system for the first order necessary conditions, see, e.g. [19] and the references therein.

There are several other possibilities like Lavrentiev type regularization where one replaces the constraint $u \le u_{\max}$ by a mixed constraint of the form $u + \varepsilon q \le u_{\max}$ with some $\varepsilon \to 0$. However due to there special requirements on the structure of the constraints we will not consider them any further.

Now, we note that the case of barrier-methods will require some extra work in the analysis, as we are then dealing with functions whose values are possibly ∞. Thus we refer to the given literature for the detailed analysis of this case. Instead we will have a look at the case of the quadratic penalty method.

11.1 Quadratic penalization

We will now consider the following model problem.

$$\begin{aligned} \min J_\gamma(q,u) &= J(q,u) + \frac{\gamma}{2}\|(u - u_{\max})^+\|^2 \\ \text{s.t. } u &\in H_0^1(\Omega), \\ (\nabla u, \nabla \varphi) &= (q, \varphi) \quad \forall\, \varphi \in H_0^1(\Omega) \end{aligned} \tag{11.3}$$

with J as in problem (11.1).

In a first step we note that for any $\gamma \ge 0$ there exists a unique solution $(\overline{q}, \overline{u}) \in Q \times U = L^2(\Omega) \times H_0^1(\Omega)$ to this problem. The proof of this is clear, as J is weakly lower semicontinuous and $\frac{\gamma}{2}\|(u-$

$u_{\max})^+\|^2 \geq 0$ and weakly lower semicontinuous and thus J_γ is weakly lower semicontinuous and we obtain existence of the optimal solution as for problem (4.3) using Theorem 3.0.1.

We begin our analysis by deriving convergence of the primal variables. The proof of this will be split into several steps.

Lemma 11.1.1. *There exists a constant $c > 0$, such that for any $\gamma \geq 0$ the solution $(\overline{q}_\gamma, \overline{u}_\gamma)$ to (11.3) satisfies*

$$\|\overline{q}_\gamma\| + \|\overline{u}_\gamma\|_{1,2} + \|\overline{u}_\gamma\|_\infty + \gamma\|(\overline{u}_\gamma - u_{\max})^+\|^2 \leq c.$$

Proof. To see this, we note first that the problem (11.3) has a solution for $\gamma = 0$. Now, we only consider the case $\gamma > 0$. Then we define $(\overline{q}, \overline{u})$ as the solution to (11.1).

We now show that

$$J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \leq J(\overline{q}, \overline{u})$$

for any $\gamma > 0$. This is clear, as $(\overline{q}, \overline{u})$ is feasible for (11.3) and $\|(\overline{u} - u_{\max})^+\| = 0$ and thus

$$J(\overline{q}, \overline{u}) = J_\gamma(\overline{q}, \overline{u})$$

and the assertion follows from the fact that $(\overline{q}_\gamma, \overline{u}_\gamma)$ solves (11.3).

Then it is clear, by definition of J_γ that

$$\alpha\|\overline{q}\|^2 + \gamma\|(\overline{u}_\gamma - u_{\max})^+\|^2 \leq \max(J(\overline{q}, \overline{u}), J(\overline{q}_0, \overline{u}_0))$$

for any $\gamma \geq 0$. Thus the assertion follows from $\|\overline{u}_\gamma\|_{1,2} + \|\overline{u}_\gamma\|_\infty \leq c\|\overline{q}_\gamma\|$. □

Now, we note that for any $\gamma \geq 0$ there exists a unique solution to (11.3) and, by the same arguments as in Theorem 6.0.2, we know that there exists some $\overline{z}_\gamma$ such that the triplet $(\overline{q}_\gamma, \overline{u}_\gamma, \overline{z}_\gamma)$ satisfies

$$\begin{aligned} (\nabla\overline{u}_\gamma, \nabla\varphi) &= (\overline{q}_\gamma, \varphi) && \forall\, \varphi \in H_0^1(\Omega), \\ (\nabla\overline{z}_\gamma, \nabla\varphi) &= (\overline{u}_\gamma - u^d, \varphi) + (\varphi, \overline{\mu}_\gamma) && \forall\, \varphi \in H_0^1(\Omega), \\ (\overline{z}_\gamma + \alpha\overline{q}_\gamma, q) &= 0 && \forall\, q \in Q, \end{aligned} \tag{11.4}$$

with $\overline{\mu}_\gamma = \gamma(\overline{u}_\gamma - u_{\max})^+$.

Theorem 11.1.2. *The solution $(\overline{q}_\gamma, \overline{u}_\gamma)$ to (11.3) converges to the solution $(\overline{q}, \overline{u})$ of (11.1) in $L^2(\Omega) \times H_0^1(\Omega) \cap C(\overline{\Omega})$ as $\gamma \to \infty$.*

Proof. To see this we note that by Lemma 11.1.1 we know that there exist a subsequence $(\overline{q}_\gamma, \overline{u}_\gamma)$ and some point $(q^*, u^*) \in Q^{\mathrm{ad}} \times H_0^1(\Omega) \cap C(\overline{\Omega})$ such that

$$\overline{q}_\gamma \rightharpoonup q^*, \quad \overline{u}_\gamma \to u^*$$

in $L^2(\Omega)$ and $H_0^1(\Omega) \cap C(\overline{\Omega})$ as $\gamma \to \infty$, note that Q^{ad} is weak sequentially closed and S is compact. By weak lower semicontinuity of J we know

$$J(q^*, u^*) \leq \liminf J(\overline{q}_\gamma, \overline{u}_\gamma) \leq \liminf J_\gamma(\overline{q}_\gamma, \overline{u}_\gamma) \leq J(\overline{q}, \overline{u}). \tag{11.5}$$

Next we want to see that in fact $u^* - u_{\max} \leq 0$. Assume for contradiction that there is some point $x_0 \in \Omega$ with $u^* - u_{\max} > 0$. Then there exists some $\delta > 0$ and some $\varepsilon > 0$ such that $u^* - u_{\max} > 2\varepsilon$ on $B_\delta(x_0)$. By definition of convergence in $C(\overline{\Omega})$ we then have that for all γ sufficiently large it holds $\overline{u}_\gamma - u_{\max} > \varepsilon$ on $B_\delta(x_0)$ this yields the contradiction to Lemma 11.1.1

$$c \geq \gamma \|(\overline{u}_\gamma - u_{\max})\|^2 \geq \gamma \varepsilon^2 |B_\delta(x_0)| \to \infty \ (\gamma \to \infty).$$

This shows that (q^*, u^*) are feasible for (11.1) and thus combining this with (11.5) we get

$$J(q^*, u^*) = J(\overline{q}, \overline{u})$$

and thus as the minimizers $(\overline{q}, \overline{u})$ are unique we get

$$(q^*, u^*) = (\overline{q}, \overline{u}).$$

As this argument is valid for all subsequences, we get that for the whole sequence $(\overline{q}_\gamma, \overline{u}_\gamma)$ it holds

$$\overline{q}_\gamma \rightharpoonup \overline{q}, \quad \overline{u}_\gamma \to \overline{u}.$$

Finally, to see strong convergence of $\overline{q}_\gamma$, we use the optimality conditions (11.4) and (10.2) to get

$$\begin{aligned}(\overline{z}_\gamma + \alpha \overline{q}_\gamma, \overline{q} - \overline{q}_\gamma) &= 0 \\ (-\overline{z} - \alpha \overline{q}, \overline{q} - \overline{q}_\gamma) &= 0\end{aligned}$$

and thus

$$\begin{aligned}\alpha \|\overline{q} - \overline{q}_\gamma\|^2 &= (\overline{z}_\gamma - \overline{z}, \overline{q} - \overline{q}_\gamma) \\ &= (\nabla(\overline{u} - \overline{u}_\gamma), \nabla(\overline{z}_\gamma - \overline{z})) \\ &= -\|\overline{u} - \overline{u}_\gamma\|^2 + \gamma(\overline{u} - \overline{u}_\gamma, (\overline{u}_\gamma - u_{\max})^+) + \int_\Omega \overline{u}_\gamma - \overline{u} \, d\overline{\mu}.\end{aligned}$$

Now noting that if $(\overline{u}_\gamma - u_{\max})^+ > 0$ we have $\overline{u}_\gamma > u_{\max} \geq \overline{u}$ the second term has negative sign, and thus we obtain

$$\alpha \|\overline{q} - \overline{q}_\gamma\|^2 \leq \int_\Omega \overline{u}_\gamma - \overline{u} \, d\overline{\mu} \to 0$$

which shows the assertion. □

In addition to the convergence stated in Theorem 11.1.2, one would like to know how large the error $\|\overline{q} - \overline{q}_\gamma\|$ in fact is for some given value $\gamma \geq 0$. To this end we need a little more information, a more detailed analysis can be found in the recent works [20, 42]. We start by the following result

Lemma 11.1.3. *Assume that $u_{\max} > 0$. Then the multiplier approximation $\overline{\mu}_\gamma = \gamma(\overline{u}_\gamma - u_{\max})^+$ is uniformly bounded in $L^1(\Omega)$.*

Proof. To see this we note that by definition we have

$$\overline{u}_\gamma \geq u_{\max} \geq \delta > 0$$

on the set $\{x \in \overline{\Omega} \,|\, \overline{u}_\gamma(x) - u_{\max} > 0\}$. Hence, it holds

$$\begin{aligned}
\|\overline{\mu}_\gamma\|_1 &= \gamma \int_\Omega 1(\overline{u}_\gamma - u_{\max})^+ \, dx \\
&= \frac{1}{\delta}\gamma \int_\Omega \delta(\overline{u}_\gamma - u_{\max})^+ \, dx \\
&\leq \frac{1}{\delta}\gamma \int_\Omega \overline{u}_\gamma(\overline{u}_\gamma - u_{\max})^+ \, dx.
\end{aligned}$$

Now, we can use (11.4) and test the adjoint equation with $\overline{u}_\gamma$ to get

$$\begin{aligned}
\gamma \int_\Omega \overline{u}_\gamma(\overline{u}_\gamma - u_{\max})^+ \, dx &= (\nabla \overline{u}_\gamma, \nabla \overline{z}_\gamma) + (u^d - \overline{u}_\gamma, \overline{u}_\gamma) \\
&= (\overline{q}_\gamma, \overline{z}_\gamma) + (u^d - \overline{u}_\gamma, \overline{u}_\gamma) \\
&= -\alpha \|\overline{q}_\gamma\|^2 + (u^d - \overline{u}_\gamma, \overline{u}_\gamma).
\end{aligned}$$

In view of Lemma 11.1.1, we obtain the desired result

$$\|\overline{\mu}_\gamma\|_1 \leq \frac{1}{\delta}(\alpha \|\overline{q}_\gamma\|^2 + (u^d - \overline{u}_\gamma, \overline{u}_\gamma)) \leq c.$$

□

Corollary 11.1.4. *Let $\Omega \subset \mathbb{R}^2$ be a $C^{1,1}$ domain. Then adjoint states $\overline{z}_\gamma$ are uniformly bounded in $W_0^{1,s'}(\Omega)$ for any $s' < 2$. Moreover, the states $\overline{u}_\gamma$ are uniformly bounded in $C^{0,1}(\Omega)$.*

Proof. We know from Remark 10.2.1 that $S^* \colon \mathcal{M}(\Omega) \to W_0^{1,s'}(\Omega)$ is bounded. Further, we have by Lemma 11.1.3 that

$$\|\overline{\mu}_\gamma\|_{C^*} = \sup_{\|v\|_\infty = 1} \int_\Omega v \, d\overline{\mu}_\gamma = \sup_{\|v\|_\infty = 1} \int_\Omega v \, \overline{\mu}_\gamma \, dx \leq \sup_{\|v\|_\infty = 1} \|v\|_\infty \|\overline{\mu}_\gamma\|_1 \leq c.$$

Hence, it is $\|\overline{z}_\gamma\|_{1,s'} \leq c$. By the relation $\overline{z}_\gamma = -\alpha \overline{q}_\gamma$ and the embedding Theorem 2.6.14 we have that

$$\|\overline{q}_\gamma\|_p \leq c$$

for some $p > 2$. Using elliptic regularity and the embedding theorems it follows $\overline{u}_\gamma \in W^{2,p}(\Omega) \subset C^{0,1}(\Omega)$ and

$$\|\overline{u}_\gamma\|_{C^{0,1}} \leq c \|\overline{q}_\gamma\|_p \leq c.$$

□

Lemma 11.1.5. *Let $\overline{q}_\gamma$ be the solution to (11.3) and $\overline{q}_\gamma$ be the solution to (11.1). Then it holds*

$$\|\overline{q} - \overline{q}_\gamma\|^2 \leq c\|(\overline{u}_\gamma - u_{\max})^+\|_\infty.$$

Proof. We know from the proof of Theorem 11.1.2 that

$$\alpha\|\overline{q} - \overline{q}_\gamma\|^2 \leq \int_\Omega \overline{u}_\gamma - \overline{u}\, d\overline{\mu}.$$

Now we can calculate

$$\begin{aligned}
\alpha\|\overline{q} - \overline{q}_\gamma\|^2 &\leq \int_\Omega \overline{u}_\gamma - u_{\max}\, d\overline{\mu} - \int_\Omega \overline{u} - u_{\max}\, d\overline{\mu} \\
&= \int_\Omega (\overline{u}_\gamma - u_{\max})^+\, d\overline{\mu} - \int_\Omega (\overline{u}_\gamma - u_{\max})^-\, d\overline{\mu} \\
&\leq \int_\Omega (\overline{u}_\gamma - u_{\max})^+\, d\overline{\mu} \\
&\leq \|(\overline{u}_\gamma - u_{\max})^+\|_\infty \|\overline{\mu}\|_{C^*}
\end{aligned}$$

which shows the assertion. □

Now all that remains is to estimate the maximal feasibility violation.

Lemma 11.1.6. *Let $\overline{u}_\gamma$ be the solution to (11.3) on a $C^{1,1}$ domain $\Omega \subset \mathbb{R}^2$ and let $u_{\max} \in C^{0,1})(\Omega)$. Then it holds*

$$\|(\overline{u}_\gamma - u_{\max})^+\|_\infty \leq c\gamma^{-1/3}$$

Proof. We begin by noting that due to Corollary 11.1.4 we have that $\overline{u}_\gamma$ are uniformly Lipschitz-continuous. Now define $g(x) = \overline{u}_\gamma(x) - u_{\max}$ and let $x^* \in \overline{\Omega}$ be such that

$$\varepsilon_\gamma = \max_{x \in \overline{\Omega}} g(x) = g(x^*).$$

Then we assume w.l.o.g. that $\varepsilon_\gamma > 0$. By L-continuity we get that

$$|g(x) - g(y)\| \leq L\|x - y\|.$$

and thus

$$g(y) \geq \varepsilon_\gamma/2 \quad \forall y \in B_{\varepsilon_\gamma/(2L)}(x^*).$$

Together we get from Lemma 11.1.3 that

$$\begin{aligned} c\gamma^{-1} &\geq \|(g)^+\|_1 \\ &\geq \int\limits_{\{g\geq \varepsilon_\gamma/2\}} |g(x)|\,dx \\ &\geq \frac{\varepsilon_\gamma}{2} \int\limits_{\{g\geq \varepsilon_\gamma/2\}} dx \\ &\geq c\frac{\varepsilon_\gamma}{2}\varepsilon_\gamma^2 = c\varepsilon_\gamma^3. \end{aligned}$$

This yields the assertion. □

Remark 11.1.7. We note that the combination of Lemma 11.1.5 and Lemma 11.1.6 gives a worst case convergence rate. In fact in our situation the states will be more regular than Lipschitz continuous. This will yield a larger convergence rate. Further the estimate in Lemma 11.1.6 assumes that we have only one point where the maximum is obtained and than the function decays in all directions. This is not always the case, e.g., for line measures this will be wrong, see [20] for an extensive discussion of this issue.

Remark 11.1.8. Further we note that the constants in all the theorems of this section did not depend upon the actual functions but rather on the norms of these functions which we can bound in terms of $J(\overline{q},\overline{u})$. In particular, the same estimates will hold if one discretizes the state equation as in Chapter 7 with constants independent of $h \to 0$. In particular, one will need uniform bounds on $\|\overline{q}_{h\gamma}\|_{1,s'}$ and $\|\overline{u}_{h\gamma}\|_{1,\infty}$

12 Discretization with State Constraints

As already remarked in Remark 11.1.8 the convergence of the penalty method will not depend upon the discretization of the state equation (or the control space). Hence it is sufficient to consider the discretization of the discrete limit problem to obtain an overall error estimate.

We will now consider the case of pure state constraints in order to avoid some technicalities. We use the same space V_h as in Chapter 7.

$$\begin{aligned} \min J(q_h, u_h) = \frac{1}{2}\|u_h - u^d\|^2 + \frac{\alpha}{2}\|q_h\|^2 \\ \text{s.t. } u_h \in V_h, \\ (\nabla u_h, \nabla \varphi_h) = (q_h, \varphi_h) \quad \forall\, \varphi_h \in V_h, \\ u_h \le u_{\max} \quad \text{on } \Omega. \end{aligned} \tag{12.1}$$

For simplicity, we will consider the variational discretization only. This means that we pick $Q_h = Q$. However, due to the necessary optimality conditions we would obtain exactly the same solution if $Q_h = V_h$. Moreover, we let $u_{\max} \ge c > 0$ be a continuous function so that, again, $\tilde{q} = 0$ is a Slater point.

In order for our problem to have good regularity properties for the solution as well as no technical difficulties in the discretization, we assume that $\Omega \subset \mathbb{R}^2$ is a polygonally bounded domain with all angles lower or equal to $\pi/2$.

Under this assumption we have the following result for the solution operators

Lemma 12.0.1. *Let $u = Sq$ and $u_h = S_h q$ be given by the solution operator of the PDE and its discrete counterpart. Assume that all angles of $\Omega \subset \mathbb{R}^2$ are lower or equal to $\pi/2$. Then for any $\varepsilon > 0$ there exists an $s' < 2$ and some constant $c > 0$ such that*

$$\|S - S_h\|_{\mathscr{L}(W_0^{1,s'}, L^\infty)} \le c h^{2-\varepsilon} = c h^\beta$$

or equivalently

$$\|u - u_h\|_\infty \le c h^{2-\varepsilon} \|q\|_{1,s'}.$$

Further if Ω is convex. Then

$$\|S - S_h\|_{\mathscr{L}(L^2, L^\infty)} \le ch$$

and

$$\|S^* - S_h^*\|_{\mathscr{L}(\mathscr{M}(\Omega), L^2)} \le ch.$$

Proof. See [31] for the finite element error on S in $\mathscr{L}(W_0^{1,s'}, L^\infty)$. The result for S^* is due to [5]. The estimate in $\mathscr{L}(L^2, L^\infty)$ is a standard finite element estimate. □

Further, we note that the following inverse estimate is true.

Theorem 12.0.2. *Let $u_h \in V_h$ be given, then for $1 \le s \le \infty$ it holds with a constant independent of h*

$$\|u_h\|_{1,s} \le ch^{-1}\|u_h\|_s$$

Proof. See any textbook on finite element analysis, e.g., [4]. □

The following arguments follow the works of [13, 26]. We already know that the solution $\overline{q}$ to (11.1) satisfies $\overline{q} \in W_0^{1,s'}(\Omega)$ for all $s < 2$. Now we want to see that the same holds true for the solutions $\overline{q}_h$ to (12.1), i.e., $\|\overline{q}_h\|_{1,s'} \le c$.

To this end, let us recall our usual constraint qualification

Assumption 12.0.3. We will assume throughout that the control and state constraints are compatible in the sense that there exists some Slater point $\widetilde{q} \in Q^{\mathrm{ad}} \cap H_0^1(\Omega)$ such that $\widetilde{u} = S\widetilde{q} < u_{\max}$. In particular, there is some $\tau > 0$ such that $\widetilde{u} \le u_{\max} - \tau$.

Indeed, here this assumption is satisfied with $\tilde{q} = 0$ due to the assumption $u_{\max} > 0$.

Lemma 12.0.4. *There exists some $h_0 > 0$ such that for any $h \in (0, h_0)$ the Slater point $\widetilde{q}$ defined in Assumption 12.0.3 is a Slater point for (12.1), i.e., there is some $\tau > 0$ such that $\widetilde{u}_h = S_h\widetilde{q} \le u_{\max} - \tau$.*

Remark 12.0.5. Indeed, in the present situation, $\tilde{q} = 0$ is clearly a Slater point, we will still show the lemma ignoring this fact to show how to proceed in general.

Proof. To see this, we note that $\overline{\Omega}$ is compact, and thus the continuous function $\widetilde{u}$ has a maximum

$$\max_{x \in \overline{\Omega}} \widetilde{u}(x) = \tau < u_{\max}$$

Now, in view of Lemma 12.0.1 there exists some value h_0 such that

$$ch_0^{2-\varepsilon}\|\widetilde{q}\|_{1,s'} < \tau/2.$$

Thus it holds

$$\begin{aligned}
\widetilde{u}_h(x) &= \widetilde{u}_h(x) - \widetilde{u}(x) + \widetilde{u}(x) \\
&\le \widetilde{u}(x) + \|\widetilde{u}_h - \widetilde{u}\|_\infty \\
&\le \max_{x \in \overline{\Omega}} \widetilde{u}(x) + ch^{2-\varepsilon}\|\widetilde{q}\|_{1,s'} \\
&\le u_{\max} - \tau + ch_0^{2-\varepsilon}\|\widetilde{q}\|_{1,s'} \\
&\le u_{\max} - \tau/2.
\end{aligned}$$

□

It is clear, that for each $h \in (0, h_0)$ the problem (12.1) has a unique solution and by the results of Chapter 9 there exists corresponding multipliers $\overline{z}_h \in V_h$ and $\overline{\mu}_h \in \mathcal{M}(\Omega)$ satisfying

$$\begin{aligned}
&(\nabla \overline{u}_h, \nabla \varphi_h) = (\overline{q}_h, \varphi_h) \quad \forall\, \varphi_h \in V_h,\\
&(\nabla \varphi_h, \nabla \overline{z}_h) = (\overline{u}_h - u^d, \varphi_h) + \int_\Omega \varphi_h \, d\overline{\mu}_h \quad \forall\, \varphi \in V_h,\\
&\overline{z}_h + \alpha \overline{q}_h = 0 \quad \text{a.e. in } \Omega,\\
&\qquad \overline{\mu}_h \in K^+, \quad \overline{u}_h - u_{\max} \le 0, \langle \overline{u}_h - u_{\max}, \overline{\mu}_h \rangle_{C,C^*} = 0.
\end{aligned} \tag{12.2}$$

Remark 12.0.6. Noting that V_h is finite dimensional, we can represent its dual as a finite dimensional space. Due to the definition of the nodal basis of V_h one such representation is that for any $\mu \in C^*$ there are $\mu_i \in \mathbb{R}$, $i = 1, \dots, N$ such that

$$\mu = \sum_{i=1}^{N} \mu_i \delta_{x_i} \in V_h^*$$

where the x_i denote the interior vertices of $\mathcal{T}_h$ and δ_{x_i} is the corresponding Dirac measure.

Lemma 12.0.7. *The solutions $(\overline{q}_h, \overline{u}_h)$ to (12.1) are uniformly bounded in $L^2(\Omega) \times H_0^1(\Omega) \cap C(\overline{\Omega})$*

Proof. The assertion follows from

$$\begin{aligned}
\alpha \|\overline{q}_h\|^2 &\le J(\overline{q}_h, \overline{u}_h)\\
&\le J(\widetilde{q}, \widetilde{u}_h)\\
&\le 2J(\widetilde{q}, \widetilde{u}) + \|\widetilde{u} - \widetilde{u}_h\|^2\\
&\le 2J(\widetilde{q}, \widetilde{u}) + ch^4 \|\widetilde{q}\|\\
&\le c,
\end{aligned}$$

together with the error estimate on $\|S - S_h\|_{\mathcal{L}(L^2, L^\infty)}$ given in Lemma 12.0.1. □

With these preparations, we can derive a uniform bound on the multipliers $\overline{\mu}_h$.

Lemma 12.0.8. *There exists a constant $c > 0$ such that $\|\overline{\mu}_h\|_{\mathcal{M}(\Omega)} \le c$ for all $h \in (0, h_0)$.*

Proof. To see this, let $f \in C(\overline{\Omega})$ be given with $\|f\|_\infty \le 1$. Then it follows

$$\begin{aligned}
\int_\Omega f \, d\overline{\mu}_h &\le \int_\Omega |f| \, d\overline{\mu}_h\\
&\le \int_\Omega 1 \, d\overline{\mu}_h.
\end{aligned}$$

Further, by assumption we have $\overline{u}_h = u_{\max} > 0$ on the support of $\overline{\mu}_h$. Thus, we get from the conditions (12.2)

$$\begin{aligned} u_{\max} \int_\Omega f \, d\overline{\mu}_h &\le u_{\max} \int_\Omega 1 \, d\overline{\mu}_h \\ &= \int_\Omega \overline{u}_h \, d\overline{\mu}_h \\ &= (\nabla \overline{z}_h, \nabla \overline{u}_h) + (u^d - \overline{u}_h, \overline{u}_h) \\ &= (\overline{q}_h, \overline{z}_h) + (u^d - \overline{u}_h, \overline{u}_h) \\ &= -\alpha \|\overline{q}_h\|^2 + (u^d - \overline{u}_h, \overline{u}_h). \end{aligned}$$

Boundedness of $\|\overline{\mu}_h\|_{\mathcal{M}(\Omega)} \le c$ follows from Lemma 12.0.7. □

Before we proceed we need some additional results for finite elements.

Theorem 12.0.9. *Let the triangulation be regular in the sense given in Section 7.1. Then for the L^2-projection $P_h : L^2(\Omega) \to V_h$ it holds*

$$\|P_h u\|_{1,s} \le c \|u\|_{1,s} \quad \forall u \in W_0^{1,s}(\Omega)$$

and

$$\|u - P_h u\|_s \le ch \|u\|_{1,s} \quad \forall u \in W_0^{1,s}(\Omega).$$

Proof. See [10] for the stability estimate. For the error estimate use [14] together with standard interpolation estimates, see, e.g., [4]. □

With these preparations the following is easy to see.

Theorem 12.0.10. *The solutions $\overline{q}_h$ to (12.1) are uniformly bounded in $W_0^{1,s'}(\Omega)$ for any $s' < 2$.*

Proof. Due to the relation $\alpha \overline{q}_h + \overline{z}_h = 0$ it is sufficient to show the boundedness of

$$\overline{z}_h = S_h^*(\overline{u}_h - u^d + \overline{\mu}_h).$$

To do so, we define

$$z^h = S^*(\overline{u}_h - u^d + \overline{\mu}_h).$$

Then, we denote by P_h the L^2-projection $L^2(\Omega) \to V_h$ and get

$$\|\overline{z}_h\|_{1,s'} \le \|\overline{z}_h - P_h z^h\|_{1,s'} + \|z^h - P_h z^h\|_{1,s'} + \|z^h\|_{1,s'}.$$

By stability of the L^2-projection in $W^{1,s'}$, see Theorem 12.0.9, and boundedness of $S^* : \mathcal{M}(\Omega) \to W_0^{1,s'}(\Omega)$, we get for the last two summands

$$\|z^h - P_h z^h\|_{1,s'} + \|z^h\|_{1,s'} \le c(\|\overline{u}_h - u^d\| + \|\overline{\mu}_h\|_{\mathcal{M}(\Omega)})$$

which is bounded due to Lemma 12.0.7 and Lemma 12.0.8. To estimate the first term, we use an inverse estimate Theorem 12.0.2 to get

$$\begin{aligned}\|\overline{z}_h - P_h z^h\|_{1,s'} &\leq ch^{-1}\|\overline{z}_h - P_h z^h\|_{s'} \\ &\leq ch^{-1}(\|\overline{z}_h - z^h\|_{s'} + \|z^h - P_h z^h\|_{s'}).\end{aligned}$$

The second term is bounded by error estimates for the L^2-projection in Theorem 12.0.9. For the first, we use Lemma 12.0.1 and the embedding of L^2 into $L^{s'}$ to get

$$\|\overline{z}_h - z^h\|_{s'} \leq c\|\overline{z}_h - z^h\| \leq ch(\|\overline{u}_h - u^d\| + \|\overline{\mu}_h\|_{\mathcal{M}(\Omega)})$$

and together with Lemma 12.0.8 the assertion follows. □

Now, we are prepared to show the convergence of the control variables. To do so, we will need to construct some feasible points for the problems.

Lemma 12.0.11. *Let $\widetilde{q}$ be given by Assumption 12.0.3 and $h < h_0$ and $\tau > 0$ given by Lemma 12.0.4. Further, let $(\overline{q}, \overline{u})$ be the solution to (11.1). Then the control*

$$\widehat{q}_h = (1 - c_1 h^\beta)\overline{q} + c_1 h^\beta \widetilde{q}$$

and the corresponding discrete state $\widehat{u}_h = S_h \widehat{q}_h$ are feasible for (12.1) if $\beta = 2 - \varepsilon$ and $c_1 > c\|\overline{q}\|_{1,s'}/\tau$ with ε, c, and s' given by Lemma 12.0.1 for all h small enough such that $(1 - c_1 h^\beta) > 0$.

Conversely, let $(\overline{q}_h, \overline{u}_h)$ be the solution to (12.1). Then the control

$$\widehat{q}^h = (1 - c_2 h^\beta)\overline{q}_h + c_2 h^\beta \widetilde{q}$$

and the corresponding state $\widehat{u}^h = S\widehat{q}^h$ are feasible for (11.1) if $c_2 > c\|\overline{q}_h\|_{1,s'}/\tau$.

Proof. We know by assumption that $\max_{x\in\overline{\Omega}} S\widetilde{q} \leq u_{\max} - \tau$ and $\max_{x\in\overline{\Omega}} S_h\widetilde{q} \leq u_{\max} - \tau$. Thus it follows

$$\begin{aligned}S_h\widehat{q}_h &= (1 - c_1 h^\beta)S_h\overline{q} + c_1 h^\beta S_h\widetilde{q} \\ &= (1 - c_1 h^\beta)(S\overline{q} - S\overline{q} + S_h\overline{q}) + c_1 h^\beta S_h\widetilde{q} \\ &\leq (1 - c_1 h^\beta)u_{\max} + (1 - c_1 h^\beta)\|(S - S_h)\overline{q}\|_\infty + c_1 h^\beta S_h\widetilde{q} \\ &\leq (1 - c_1 h^\beta)u_{\max} + (1 - c_1 h^\beta)\|S - S_h\|_{\mathcal{L}(W_0^{1,s'},L^\infty)}\|\overline{q}\|_{1,s'} + c_1 h^\beta(u_{\max} - \tau).\end{aligned}$$

By Lemma 12.0.1, we get

$$S_h\widehat{q}_h \leq u_{\max} + (1 - c_1 h^\beta)ch^\beta\|\overline{q}\|_{1,s'} - c_1 h^\beta \tau.$$

Due to the choice of c_1, we have

$$(1 - c_1 h^\beta)c\|\overline{q}\|_{1,s'} \leq c\|\overline{q}\|_{1,s'} \leq c_1\tau$$

and thus

$$S_h\widehat{q}_h \leq u_{\max}$$

which means that $(\widehat{q}_h, \widehat{u}_h)$ is feasible for (12.1).

By an analog computation, we get that $(\widehat{q}^h, \widehat{u}^h)$ is feasible for (11.1). □

We will now give a slightly different proof for the main convergence theorem than those given in [13, 26] which follows the lines of [29].

Theorem 12.0.12. *Under the assumptions of Lemma 12.0.11, it holds*

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

Proof. To see the assertion, we note that by construction it holds

$$\|\overline{q}-\widehat{q}_h\| = \|\overline{q}-(1-c_1h^\beta)\overline{q}+c_1h^\beta\widetilde{q}\| = ch^\beta\|\overline{q}+\widetilde{q}\| \le ch^\beta$$

and further (Exercise)

$$\|\overline{u}-\widehat{u}_h\| \le c\|\overline{q}-\widehat{q}_h\|+ch^2 \le ch^\beta.$$

The functional $J(q,u)$ is continuously Fréchet differentiable and thus locally Lipschitz continuous (Mean-value Theorem 5.1.6). In particular, this means that there is a constant L_δ depending monotonically on $\|\overline{q}\|$, $\|\overline{u}\|$, and possibly δ (although not here) such that

$$|J(\overline{q},\overline{u})-J(q,u)| \le L_\delta(\|\overline{q}-q\|+\|\overline{u}-u\|) \quad \forall\,(q,u) \text{ with } \|\overline{q}-q\|+\|\overline{u}-u\| \le \delta$$

with $L_{\delta_1} \le L_{\delta_2}$ if $\delta_1 \le \delta_2$.

This means that

$$|J(\overline{q},\overline{u})-J(\widehat{q}_h,\widehat{u}_h)| \le L_{ch_0^\beta}ch^\beta = ch^\beta.$$

Further, by Lemma 12.0.11, we know that $(\widehat{q}_h,\widehat{u}_h)$ are feasible for (12.1) and thus

$$J(\overline{q}_h,\overline{u}_h) \le J(\widehat{q}_h,\widehat{u}_h) \le J(\overline{q},\overline{u})+ch^\beta.$$

Repeating the same argument for $(\overline{q}_h,\overline{u}_h)$ and $(\widehat{q}^h,\widehat{u}^h)$ gives that

$$|J(\overline{q}_h,\overline{u}_h)-J(\widehat{q}^h,\widehat{u}^h)| \le ch^\beta$$

and

$$J(\overline{q},\overline{u}) \le J(\widehat{q}^h,\widehat{u}^h) \le J(\overline{q}_h,\overline{u}_h)+ch^\beta$$

where we note that due to the bound $J(\overline{q}_h,\overline{u}_h) \le J(\overline{q},\overline{u})+ch^\beta$ the constant does not depend on h.

Combining the estimates gives

$$J(\overline{q}_h,\overline{u}_h)-ch^\beta \le J(\overline{q},\overline{u}) \le J(\overline{q}_h,\overline{u}_h)+ch^\beta$$

and thus the assertion. □

Now, we can show convergence of the primal variables. To this end we note that the cost functional is uniformly convex. This follows from uniform convexity of the involved norms. This follows from the parallelogram law

Lemma 12.0.13. *For any $v, w \in L^2(\Omega)$ it holds*

$$\|\tfrac{1}{2}(v-w)\|^2 + \|\tfrac{1}{2}(v+w)\|^2 = \tfrac{1}{2}\|v\|^2 + \tfrac{1}{2}\|w\|^2$$

Proof. The proof follows by simple calculation

$$\begin{aligned}\|\tfrac{1}{2}(v-w)\|^2 + \|\tfrac{1}{2}(v+w)\|^2 &= \tfrac{1}{4}\|v\|^2 - \tfrac{1}{2}(v,w) + \tfrac{1}{4}\|w\|^2 + \tfrac{1}{4}\|v\|^2 + \tfrac{1}{2}(v,w) + \tfrac{1}{4}\|w\|^2 \\ &= \tfrac{1}{2}\|v\|^2 + \tfrac{1}{2}\|w\|^2.\end{aligned}$$

□

Corollary 12.0.14. *Under the assumptions of Theorem 12.0.12 it holds*

$$\|\overline{q} - \overline{q}_h\| \leq ch^{\beta/2}.$$

Proof. Let $\widehat{q}_h$ be defined by Lemma 12.0.11. Then it follows

$$\|\overline{q} - \overline{q}_h\| \leq \|\overline{q} - \widehat{q}_h\| + \|\widehat{q}_h - \overline{q}_h\| \leq ch^{\beta} + \|\widehat{q}_h - \overline{q}_h\|.$$

We can apply Lemma 12.0.13 to the functions $\widehat{q}_h, \overline{q}_h$ and $\widehat{u}_h - u^d, \overline{u}_h - u^d$ to get

$$\begin{aligned}\frac{\alpha}{4}\|\widehat{q}_h - \overline{q}_h\|^2 \leq \frac{\alpha}{4}\|\widehat{q}_h - \overline{q}_h\|^2 &+ \frac{1}{4}\|\widehat{u}_h - \overline{u}_h\|^2 \\ = \frac{\alpha}{2}\|\widehat{q}_h\|^2 + \frac{\alpha}{2}\|\overline{q}_h\|^2 &- \frac{\alpha}{4}\|\widehat{q}_h + \overline{q}_h\|^2 \\ &+ \frac{1}{2}\|\widehat{u}_h - u^d\|^2 + \frac{1}{2}\|\overline{u}_h - u^d\|^2 - \frac{1}{4}\|\widehat{u}_h + \overline{u}_h - 2u^d\|^2 \\ = J(\widehat{q}_h, \widehat{u}_h) + J(\overline{q}_h, \overline{u}_h) &- 2J\big(\tfrac{1}{2}(\widehat{q}_h + \overline{q}_h), \tfrac{1}{2}(\widehat{u}_h + \overline{u}_h)\big).\end{aligned}$$

Now, $(\frac{1}{2}(\widehat{q}_h + \overline{q}_h), \frac{1}{2}(\widehat{u}_h + \overline{u}_h))$ is feasible for (12.1) and thus

$$\begin{aligned}\frac{\alpha}{4}\|\widehat{q}_h - \overline{q}_h\|^2 &\leq J(\widehat{q}_h, \widehat{u}_h) + J(\overline{q}_h, \overline{u}_h) - 2J(\overline{q}_h, \overline{u}_h) \\ &\leq J(\widehat{q}_h, \widehat{u}_h) - J(\overline{q}, \overline{u}) + J(\overline{q}, \overline{u}) - J(\overline{q}_h, \overline{u}_h) \\ &\leq ch^{\beta}.\end{aligned}$$

□

Now we have seen, that the control variable converges in L^2 and that $\|\overline{q}_h\|_{1,s'}$ is bounded. The next natural step would be the analysis of the discretized and regularized problem

$$\begin{aligned}\min J_\gamma(q_{h\gamma}, u_{h\gamma}) = \frac{1}{2}\|u_{h\gamma} - u^d\|^2 + \frac{\alpha}{2}\|q_{h\gamma}\|^2 + \frac{\gamma}{2}\|(u_{h\gamma} - u_{\max})^+\|^2 \\ \text{s.t. } u_{h\gamma} \in V_h(\Omega), \\ (\nabla u_{h\gamma}, \nabla \varphi_h) = (q_{h\gamma}, \varphi_h) \quad \forall\, \varphi_h \in V_h.\end{aligned} \tag{12.3}$$

In view of Remark 11.1.8, the only thing left to see is that $\|\overline{u}_{h\gamma}\|_{1,\infty}$ is bounded independent of h. That this is in fact the case follows directly from the following result in finite element approximation which is valid under the assumptions of Lemma 12.0.1:

$$\|S - S_h\|_{\mathscr{L}(W_0^{1,s'},W^{1,\infty})} \leq ch^{1-\varepsilon}.$$

In particular the main result of Chapter 11 namely Lemma 11.1.5 and Lemma 11.1.6 carry over directly to to the solutions $(\overline{q}_{h\gamma}, \overline{u}_{h\gamma})$ of (12.3) and $(\overline{q}_h, \overline{u}_h)$ of (12.1). This means that

$$\|\overline{q}_h - \overline{q}_{h\gamma}\| \leq c\gamma^{-1/6}$$

with a constant $c > 0$ independent of h and γ.

Remark 12.0.15. The above convergence rate would suggest to choose γ and h such that

$$\gamma^{-1/6} \approx h^{\beta/2} = h^{1-\varepsilon}$$

in order to get best possible rate of convergence for minimal effort. Unfortunately, this is not so simple as the convergence rate with respect to γ is usually better, see Remark 11.1.7 and the exercises. This would then in turn give over-regularized systems, i.e., γ being too small. However, an a priori determination of the correct convergence rate is difficult if not impossible as the rate depends upon the structure of the active set of the limit solution which is unknown.

Further, up to now we do not know the precise size of the constants involved in the error estimates making a balancing of the error contributions based upon the asymptotic convergence rate doubtful at best.

Outlook

We remark that this lecture did not cover all recent topics in optimization with partial differential equations. Thus for a broader view on the topic the interested reader may wish to continue her/his studies with some of the topics we neglected. These are for instance

- Nonlinear elliptic PDEs, and in particular second order necessary and sufficient optimality conditions.
- Parabolic and hyperbolic PDEs.
- Non differentiable cost functionals like L^1 norms.
- Non differentiable constraints (i.e., non differentiable equations).

Bibliography

[1] R. A. Adams and J. J. F. Fournier. Sobolev Spaces, volume 140 of Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for a semilinear elliptic control problem. Comput. Optim. Appl., 23:201–229, 2002.

[3] D. Braess. Finite elements. Cambridge University Press, 2001. Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker.

[4] S. C. Brenner and L. R. Scott. The Mathematical Theory of Finite Element Methods. Springer Verlag, New York, 3. edition, 2008.

[5] E. Casas. L^2 estimates for the finite element method for the dirichlet problem with singular data. Numer. Math., 47:627–632, 1985.

[6] E. Casas. Control of an elliptic problem with pointwise state constraints. SIAM J. Control Optim., 24(6):1309–1318, 1986.

[7] E. Casas, M. Mateos, and B. Vexler. New regularity results and improved error estimates for optimal control problems with state constraints. ESAIM Control Optim. Calc. Var., 20 (3):803–822, 2014.

[8] P. G. Ciarlet. The Finite Element Method for Elliptic Problems, volume 40 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 2002.

[9] A. R. Conn, N. I. M. Gould, and P. L. Toint. Trust-Region Methods, volume 1 of MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[10] M. Crouzeix and V. Thomeé. The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces. Math. Comp., 48(178):521–532, 1987.

[11] B. Dacorogna. Direct Methods in the Calculus of Variations, volume 78 of Applied Mathematical Sciences. Springer, second edition, 2008.

[12] J. W. Daniel. The conjugate gradient method for linear and nonlinear operator equations. SIAM J. Numer. Anal., 4:10–26, 1967.

[13] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. SIAM J. Numer. Anal., 45:1937–1953, 2007.

[14] J. Douglas, Jr., T. Dupont, and L. Wahlbin. The stability in L^q of the L^2-projection into finite element function spaces. Numer. Math., 23:193–197, 1974/75.

[15] M. Giaquinta, G. Modica, and J. Souček. Cartesian Currents in the Calculus of Variations I. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, Berlin – Heidelberg – New York, 1. edition, 1998.

[16] D. Gilbarg and N. S. Trudinger. Elliptic Partial Differential Equations of Second Order, volume 224 of Grundlehren der mathematischen Wissenschaften. Springer, revised 3. edition, 2001.

[17] P. Grisvard. Elliptic Problems in Nonsmooth Domains. Monographs and studies in Mathematics. Pitman, Boston, 1. edition, 1985.

[18] R. Herzog. Vorlesungsskript: Optimale Steuerung partieller Differentialgleichungen. gehalten im SS 2010 Technische Universität Chemnitz.

[19] M. Hintermüller and K. Kunisch. PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. SIAM J. Optim., 20(3):1133–1156, 2009.

[20] M. Hintermüller, A. Schiela, and W. Wollner. The length of the primal-dual path in Moreau-Yosida-based path-following for state constrained optimal control. SIAM J. Optim., 24(1): 108–126, 2014.

[21] M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. Comp. Optim. Appl., 30(1):45–61, 2005.

[22] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. Optimization with PDE Constraints, volume 23 of Mathematical Modelling: Theory and Applications. Springer, 2009.

[23] K. Ito and K. Kunisch. Lagrange Multiplier Approach to Variational Problems and Applications, volume 15 of Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), 2008.

[24] M. Lavrentieff. Sur quelques problèmes du calcul des variations. Ann. Mat. Pura Appl., 4 (1):7–28, 1927.

[25] C. Meyer. Vorlesungsskript: Optimale Steuerung partieller Differentialgleichungen. gehalten im WS 2010/2011 TU Dortmund.

[26] C. 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.

[27] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. SIAM J. Control Optim., 43(3):970–985, 2004.

[28] J. Nocedal and S. J. Wright. Numerical Optimization. Springer, 1999.

[29] C. Ortner and W. Wollner. A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state. Numer. Math., 118(3):587–600, 2011.

[30] W. Rudin. Functional analysis. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., 1991.

[31] A. H. Schatz and L. B. Wahlbin. Maximum norm estimates in the finite element method on plane polygonal domains. Part 1. Math. Comp., 32(141):73–109, 1978.

[32] A. Schiela. Lecture notes: Optimization of complex systems. gehalten im WS 2010/2011 and der Uni Hamburg.

[33] A. Schiela. Barrier methods for optimal control problems with state constraints. SIAM J. Optim., 20(2):1002–1031, 2009.

[34] A. Schiela and W. Wollner. Barrier methods for optimal control problems with convex nonlinear gradient state constraints. SIAM J. Optim., 21(1):269–286, 2011.

[35] G. Strang and G. J. Fix. An Analysis of the Finite Element Method. Prentice-Hall Inc., 1973.

[36] F. Tröltzsch. Optimale Steuerung partieller Differentialgleichungen. Vieweg, 1. edition, 2005.

[37] M. Ulbrich. Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2011.

[38] D. Werner. Funktionalanalysis. Springer, Berlin – Heidelberg – New York, 2005.

[39] J. Werner. Optimization Theory and Applications. Advanced Lectures in Mathematics. Vieweg, 1984.

[40] J. Wloka. Partielle Differentialgleichungen. Sobolewräume und Randwertaufgabe. Teubner, Leipzig, 1. edition, 1982.

[41] J. Wloka. Partial Differential Equations. Cambridge University Press, 1987.

[42] W. Wollner. A priori error estimates for optimal control problems with constraints on the gradient of the state on nonsmooth polygonal domains. In K. Bredies, C. Clason, K. Kunisch, and G. von Winckel, editors, Control and Optimization with PDE Constraints, volume 164 of International Series of Numerical Mathematics, pages 193–215. Birkhäuser, 2013.