Lecture Notes of the Autumn School

Modelling and Optimization with Partial Differential Equations

Hamburg, September 26-30, 2005

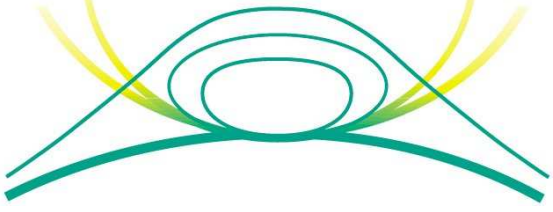
Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich

supported by

Universität Hamburg

and

SFB 609



Technische Universität Dresden

Ackknowledgements

The autumn school was supported by the Collaborative Research Center 609, located at the Technische Universität Dresden, and sponsored by the Deutsche Forschungsgemeinschaft, and by the Schwerpunkt Optimierung und Approximation at the Department Mathematik of the Universität Hamburg. All support is greatfully acknowledged.

Contents

Ackknowledgements 1

Chapter 1. Analytical Background and Optimality Theory 5
1. Introduction and examples 5
2. Linear functional analysis and Sobolev spaces 12
3. Existence of optimal controls 32
4. Reduced problem, sensitivities and adjoints 37
5. Optimality conditions 44

Chapter 2. Optimization Methods in Banach Spaces 63
1. Synopsis 63
2. Globally convergent methods in Banach spaces 64
3. Newton-based methods – A preview 73
4. Generalized Newton methods 79
5. Semismooth Newton methods in function spaces 88
6. Sequential Quadratic Programming 96
7. Further aspects 106

Chapter 3. Discrete concepts in pde constrained optimization 109
1. Introduction 109
2. Stationary model problem 110
3. Time dependent problems with control constraints 133
4. State constraints (joint with Klaus Deckelnick, Magdeburg) 135

Chapter 4. Applications 153
1. Introduction 153
2. Optimal Semiconductor Design 153
3. Optimal Control of Glass Cooling 167
4. Optimal Control of Traffic Networks 177

Bibliography 189

CHAPTER 1

Analytical Background and Optimality Theory

Stefan Ulbrich
Fachbereich Mathematik
TU Darmstadt

1. Introduction and examples

1.1. Introduction. The modelling and numerical simulation of complex systems plays an important role in physics, engineering, mechanics, chemistry, medicine, finance, and in other disciplines. Very often, mathematical models of complex systems result in partial differential equations (PDEs). For example heat flow, diffusion, wave propagation, fluid flow, elastic deformation, option prices and many other phenomena can be modelled by using PDEs. Therefore, these notes could just as well be entitled Optimization with partial differential equations. However, many of the techniques that we will develop carry over to systems that are not necessarily described by PDEs.

In most applications, the ultimate goal is not only the mathematical modelling and numerical simulation of the complex system, but rather the optimization or optimal control of the considered process. Typical examples are the optimal control of a thermal treatment in cancer therapy and the optimal shape design of an aircraft. The resulting optimization problems are very complex and a thorough mathematical analysis is necessary to design efficient solution methods.

There exist many different types of partial differential equations. We will focus on linear and semilinear elliptic and parabolic PDEs. For these PDEs the existence and regularity of solutions is well understood and we will be able to develop a fairly complete theory.

Abstractly speaking, we will consider problems of the following form

$$\min_{w\in W} f(w) \quad \text{subject to} \quad E(w)=0, \quad C(w)\in\mathcal{K}, \tag{1.1}$$

where $f : W \to \mathbb{R}$ is the objective function, $E : W \to Z$ and $C : W \to V$ are operators between Banach spaces, and $\mathcal{K} \subset V$ is a closed convex cone.

In most cases, the spaces W, Z and V are (generalized) function spaces and the operator equation $E(w) = 0$ represents a PDE or a system of coupled PDEs. The constraint

$$C(w) \in \mathcal{K}$$

is considered as an abstract inequality constraint. Sometimes (e.g., in the case of bound constraints), it will be convenient to replace the inequality constraint by a constraint of the form $w \in S$, where $S \subset W$ is a closed convex set:

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad E(w) = 0, \quad w \in \mathcal{S}. \tag{1.2}$$

Here "s.t." abbreviates "subject to".

To get the connection to finite dimensional optimization, consider the case

$$W = \mathbb{R}^n, \quad Z = \mathbb{R}^p, \quad V = \mathbb{R}^m, \quad \mathcal{K} = (-\infty, 0]^m.$$

Then the problem (1.1) becomes a nonlinear optimization problem

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad E(w) = 0, \quad C(w) \leq 0. \tag{1.3}$$

Very often, we will have additional structure: The optimization variable w admits a natural splitting into two parts, a state $y \in Y$ and a control (or design) $u \in U$, where Y and U are Banach spaces. Then $W = Y \times U$, $w = (y, u)$, and the problem reads

$$\min_{y \in Y, u \in U} f(y, u) \quad \text{s.t.} \quad E(y, u) = 0, \quad C(y, u) \in \mathcal{K}. \tag{1.4}$$

Here, $y \in Y$ describes the state (e.g., the velocity field of a fluid) of the considered system, which is described by the equation $E(y, u) = 0$ (in our context usually a PDE). The control (or design, depending on the application) $u \in U$ is a parameter that shall be adapted in an optimal way.

The splitting of the optimization variable $w = (y, u)$ into a state and a control is typical in the optimization of complex systems. Problems with this structure are called *optimal control problems*. In most cases we will consider, the state equation $E(y, u) = 0$ admits, for every $u \in U$, a unique corresponding solution $y(u)$, because the state equation is a well posed PDE for y in which u appears as a parameter. Several examples will follow below.

We use the finite-dimensional problem (1.3) to give a teaser about important questions we will be concerned with.

1. Existence of solutions.

Denote by f^* the optimal objective function value. First, we show, using the properties of the problem at hand, that f^* is achievable and finite. Then, we consider a minimizing sequence (w^k), i.e., $E(w^k) = 0$, $C(w^k) \leq 0$, $f(w^k) \to f^*$. Next, we prove that (w^k) is bounded (which has to be verified for the problem at hand). Now we do something that *only works in finite dimensions*: We conclude that, due to boundedness, (w^k) contains a convergent subsequence $(w_k)_K \to w^*$. Assuming the continuity of f, E and C we see that

$$f(w^*) = \lim_{K \ni k \to \infty} f(w^k) = f^*, \; E(w^*) = \lim_{K \ni k \to \infty} E(w^k) = 0, \; C(w^*) = \lim_{K \ni k \to \infty} C(w^k) \leq 0.$$

Therefore, w^* solves the problem.

We note that for doing the same in Banach space, we need a replacement for the compactness argument, which will lead us to weak convergence and weak compactness. Furthermore, we need the continuity of the function f and of the operators E and C with respect to the norm topology and/or the weak topology.

2. Uniqueness

Uniqueness usually relies on strict convexity of the problem, i.e., f strictly convex, E linear and C_i convex. This approach can be easily transfered to the infinite-dimensional case.

3. Optimality conditions

Assuming continuous differentiability of the functions f, C, and E, and that the constraints satisfy a regularity condition on the constraints, called *constraint qualification* (CQ) at the solution, the following first-order optimality conditions hold true at a solution w^*:

Karush-Kuhn-Tucker conditions:

There exist Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that (w^*, λ^*, μ^*) solves the following KKT-system:

$$\begin{aligned} \nabla f(w) + C'(w)^T\lambda + E'(w)^T\mu &= 0, \\ E(w) &= 0, \\ C(w) \leq 0, \quad \lambda \geq 0, \quad C(w)^T\lambda &= 0. \end{aligned}$$

Here, the column vector $\nabla f(w) = f'(w)^T \in \mathbb{R}^n$ is the gradient of f and $C'(w) \in \mathbb{R}^{m\times n}$, $E'(w) \in \mathbb{R}^{p\times n}$ are the Jacobian matrices of C and E.

All really efficient optimization algorithms for (1.3) build upon these KKT-conditions. Therefore, it will be very important to derive first order optimality conditions for the infinite-dimensional problem (1.1). Since the KKT-conditions involve derivatives, we have to extend the notion of differentiability to operators between Banach spaces. This will lead us to the concept of Fréchet-differentiability. For concrete problems, the appropriate choice of the underlying function spaces is not always obvious, but it is crucial for being able to prove the Fréchet-differentiability of the function f and the operators C, E and for verifying constraint qualifications.

4. Optimization algorithms

As already said, modern optimization algorithms are based on solving the KKT system. For instance, for problems without inequality constraints, the KKT system reduces to the following $(n+p)\times(n+p)$ system of equations:

$$G(w,\mu) \stackrel{\text{def}}{=} \begin{pmatrix} \nabla f(w) + E'(w)^T\mu \\ E(w) \end{pmatrix} = 0. \tag{1.5}$$

One of the most powerful algorithms for equality constrained optimization, the Lagrange-Newton method, consists in applying Newton's method to the equation (1.5):

Lagrange-Newton method:

For $k = 0, 1, 2, \ldots$:

1. STOP if $G(w^k, \mu^k) = 0$.
2. Compute $s^k = (s_w^k, s_\mu^k)^T$ by solving

$$G'(w^k, \mu^k)s^k = -G(w^k, \mu^k)$$

and set $w^{k+1} := w^k + s_w^k$, $\mu^{k+1} := \mu^k + s_\mu^k$.

Since G involves first derivatives, the matrix $G'(w, \mu)$ involves second derivatives. For the development of Lagrange-Newton methods for the problem class (1.1) we thus need second derivatives of f and E.

There are many more aspects that will be covered, but for the time being we have given sufficient motivation for the material to follow.

1.2. Examples for optimization problems with PDEs. We give several simple, but illustrative examples for optimization problems with PDEs.

1.3. Optimization of a stationary heating process. Consider a solid body occupying the domain $\Omega \subset \mathbb{R}^3$. Let $y(x)$, $x \in \Omega$ denote the temperature of the body at the point x.

We want to heat or cool the body in such a way that the temperature distribution y coincides as good as possible with a desired temperature distribution $y_d : \Omega \to \mathbb{R}$.

Boundary control. If we apply a temperature distribution $u : \partial\Omega \to \mathbb{R}$ to the boundary of Ω then the temperature distribution y in the body is given by the *Laplace equation*

$$-\Delta y(x) = 0, \quad x \in \Omega \tag{1.6}$$

together with the boundary condition of *Robin type*

$$\kappa \frac{\partial y}{\partial \nu}(x) = \beta(x)\,(u(x) - y(x)), \quad x \in \partial\Omega,$$

where $\kappa > 0$ is the heat conduction coefficient of the material of the body and $\beta : \partial\Omega \to (0, \infty)$ is a positive function modelling the heat transfer coefficient to the exterior.

Here, Δy is the Laplace operator defined by

$$\Delta y(x) = \sum_{i=1}^{n} y_{x_i x_i}(x)$$

with the abbreviation

$$y_{x_i x_i}(x) = \frac{\partial^2 y}{\partial x_i^2}(x)$$

and $\frac{\partial y}{\partial \nu}(x)$ is the derivative in the direction of the outer unit normal $\nu(x)$ of $\partial\Omega$ at x, i.e.,

$$\frac{\partial y}{\partial \nu}(x) = \nabla y(x) \cdot \nu(x), \quad x \in \partial\Omega.$$

As we will see, the Laplace equation (1.6) is an *elliptic* partial differential equation of second order.

In practice, the control u is restricted by additional constraints, for example by upper and lower bounds

$$a(x) \le u(x) \le b(x), \quad x \in \partial\Omega.$$

To minimize the distance of the actual and desired temperature y and y_d, we consider the following optimization problem.

$$\begin{aligned} \min \qquad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\int_\Omega (y(x)-y_d(x))^2\,dx + \frac{\alpha}{2}\int_{\partial\Omega} u(x)^2\,dS(x) \\ \text{subject to} \quad & -\Delta y = 0 \quad \text{on } \Omega, \\ & \frac{\partial y}{\partial \nu} = \frac{\beta}{\kappa}(u-y) \quad \text{on } \partial\Omega, \qquad \text{(State equation)} \\ & a \le u \le b \quad \text{on } \partial\Omega \qquad \text{(Control constraints).} \end{aligned}$$

The first term in the objective functional $f(y,u)$ measures the distance of y and y_d, the second term is a regularization term with parameter $\alpha \ge 0$ (typically $\alpha \in [10^{-5}, 10^{-3}]$), which leads to improved smoothness properties of the optimal control for $\alpha > 0$.

If we set

$$E(y,u) \stackrel{\text{def}}{=} \begin{pmatrix} -\Delta y \\ \frac{\partial y}{\partial \nu} - \frac{\beta}{\kappa}(u-y) \end{pmatrix}, C(y,u) \stackrel{\text{def}}{=} \begin{pmatrix} a-u \\ u-b \end{pmatrix},$$

where Y and U are appropriately chosen Banach spaces of functions

$$y : \Omega \to \mathbb{R}, \quad u : \partial\Omega \to \mathbb{R},$$

$Z = Z_1 \times Z_2$ with appropriately chosen Banach spaces Z_1, Z_2 of functions

$$z_1 : \Omega \to \mathbb{R}, \quad z_2 : \partial\Omega \to \mathbb{R},$$

$V = U \times U$, and

$$\mathcal{K} = \{(v_1, v_2) \in U \times U \,:\, v_i(x) \le 0,\ x \in \partial\Omega\},$$

then the above optimal control problem is of the form (1.1).

One of the crucial points will be to choose the above function spaces in such a way that f, E, and C are continuous and sufficiently often differentiable, to ensure existence of solutions, the availability of optimality conditions, etc.

Boundary control with radiation boundary. If we take heat radiation at the boundary of the body into account, we obtain a nonlinear Stefan-Boltzmann boundary condition. This leads to the semilinear state equation (i.e., the highest order term is still linear)

$$\begin{aligned} -\Delta y &= 0 \quad \text{on } \Omega, \\ \frac{\partial y}{\partial \nu} &= \frac{\beta}{\kappa}(u^4 - y^4) \quad \text{on } \partial\Omega. \end{aligned}$$

This is a problem of the form (1.1) with

$$E(y,u) \stackrel{\text{def}}{=} \begin{pmatrix} -\Delta y \\ \frac{\partial y}{\partial \nu} - \frac{\beta}{\kappa}\,(u^4 - y^4) \end{pmatrix}$$

and the rest as before.

Distributed control. Instead of heating at the boundary it is in some applications also possible to apply a distributed heat source as control. This can for example be achieved by using electro-magnetic induction.

If the boundary temperature is zero then, similar as above, we obtain the problem

$$\begin{aligned} \min \qquad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\int_\Omega (y(x) - y_d(x))^2\,dx + \frac{\alpha}{2}\int_\Omega u(x)^2\,dx \\ \text{subject to} \quad & -\Delta y = \gamma\, u \quad \text{on } \Omega, \\ & y = 0 \ \text{ on } \partial\Omega, \\ & a \le u \le b \text{ on } \Omega. \end{aligned}$$

Here, the coefficient $\gamma : \Omega \to [0,\infty)$ weights the control. The choice $\gamma = 1_{\Omega_c}$ for some control region $\Omega_c \subset \Omega$ restricts the action of the control to the control region Ω_c.

If we assume a surrounding temperature y_a then the state equation changes to

$$\begin{aligned} -\Delta y &= \gamma\, u \quad \text{on } \Omega, \\ \frac{\partial y}{\partial \nu} &= \frac{\beta}{\kappa}\,(y_a - y) \quad \text{on } \partial\Omega. \end{aligned}$$

Problems with state constraints. In addition to control constraint also *state constraints*

$$l \le y \le r$$

with functions $l < r$ are of practical interest. They are much harder to handle than control constraints.

1.4. Optimization of an unsteady heating processes. In most applications, heating processes are time-dependent. Then the temperature $y : \Omega \times [0,T] \to \mathbb{R}$ depends on space and time. We set

$$Q \stackrel{\text{def}}{=} \Omega \times (0,T), \quad \Sigma = \partial\Omega \times (0,T).$$

Boundary control. Let y_d be a desired temperature distribution at the end time T and y_0 be the initial temperature of the body. To find a control $u : \Sigma \to \mathbb{R}$ that minimizes the distance of the actual temperature $y(\cdot,T)$ at the end time and the desired temperature y_d, we consider similar as above the

following optimization problem.

$$\begin{aligned} \min \qquad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\int_\Omega (y(T,x)-y_d(x))^2\,dx + \frac{\alpha}{2}\int_0^T\int_{\partial\Omega} u(x,t)^2\,dS(x)\,dt \\ \text{subject to} \quad & y_t - \Delta y = 0 \quad \text{on } Q, \\ & \frac{\partial y}{\partial \nu} = \frac{\beta}{\kappa}(u-y), \\ & y(x,0) = y_0(x) \quad \text{on } \Omega \\ & a \le u \le b \quad \text{on } \Sigma. \end{aligned}$$

Here, y_t denotes the partial derivative with respect to time and Δy is the Laplace operator in space. The PDE

$$y_t - \Delta y = 0$$

is called *heat equation* and is the prototype of a *parabolic* partial differential equation.

Similarly, unsteady boundary control with radiation and unsteady distributed control can be derived from the steady counterparts.

Optimal control problems with linear state equation and quadratic objective function are called *linear-quadratic*. If the PDE is nonlinear in lower order terms then the PDE is called *semilinear*.

1.5. Optimal design. A very important dscipline is optimal design. Here, the objective is to optimize the shape of some object. A typical example is the optimal design of a wing or a whole airplane with respect to certain objective, e.g., minimal drag, maximum lift or a combination of both.

Depending on the quality of the mathematical model employed, the flow around a wing is described by the Euler equations or (better) by the compressible Navier-Stokes equations. Both are systems of PDEs. A change of the wing shape would then result in a change of the spatial flow domain Ω and thus, the design parameter is the domain Ω itself or a description of it (e.g. a surface describing the shape of the wing). Optimization problems of this type are very challenging.

Therefore, we look here at a much simpler example:

Consider a very thin elastic membrane spanned over the domain $\Omega \subset \mathbb{R}^2$. Its thickness $u(x) > 0$, $x \in \Omega$, varies (but is very small). At the boundary of Ω, the membrane is clamped at the level $x_3 = 0$.

Given a vertical force distribution $g : \Omega \to \mathbb{R}$ acting from below, the membrane takes the equilibrium position described by the graph of the function $y : \Omega \to \mathbb{R}$ (we assume that the thickness is negligibly compared to the displacement). For small displacement, the mathematical model for this membrane then is given by the following elliptic PDE:

$$\begin{aligned} -\text{div}(u\nabla y) &= g \quad \text{on } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

Here, $\text{div}\, v = \sum_i (v_i)_{x_i}$ denotes the divergence of $v : \Omega \to \mathbb{R}^2$.

The design goal consists in finding an optimal thickness u subject to the thickness constraints

$$a(x) \leq u(x) \leq b(x) \quad x \in \Omega$$

and the volume constraint

$$\int_\Omega u(x)\,dx \leq V$$

such that the compliance

$$f(y) = \int_\Omega g(x)y(x)\,dx$$

of the membrane is as small as possible. The smaller the compliance, the stiffer the membrane with respect to the load g. We obtain the following optimal design problem

$$\begin{aligned} \min \qquad & f(y) \stackrel{\text{def}}{=} \int_\Omega g(x)y(x)dx \\ \text{subject to} \quad & -\operatorname{div}(u\nabla y) = g \quad \text{on } \Omega, \\ & y = 0 \quad \text{on } \partial\Omega, \\ & a \leq u \leq b \quad \text{on } \Omega, \\ & \int_\Omega u(x)\,dx \leq V. \end{aligned}$$

2. Linear functional analysis and Sobolev spaces

We have already seen that PDEs do in practical relevant situations not necessarily have classical solutions. A satisfactory solution theory can be developed by using Sobolev spaces and functional analysis.

We recall first several basics on Banach and Hilbert spaces. Details can be found in any book on linear functional analysis, e.g., [**3**], [**46**], [**65**], [**80**], [**81**].

2.1. Banach and Hilbert spaces.

2.2. Basic definitions.

DEFINITION 2.1. (Norm, Banach space)
Let X be a real vector space.

i) *A mapping $\|\cdot\| : X \mapsto [0,\infty)$ is a* norm *on X, if*
   1) $\|u\| = 0 \iff u = 0$,
   2) $\|\lambda u\| = |\lambda| u \;\forall\, u \in X,\ \lambda \in \mathbb{R}$,
   3) $\|u+v\| \leq \|u\| + \|v\| \;\forall\, u, v \in X$.

ii) *A normed real vector space X is called (real)* Banach space *if it is complete, i.e., if any Cauchy sequence (u_n) has a limit $u \in X$, more precisely, if $\lim_{m,n\to\infty} \|u_m - u_n\| = 0$ then there is $u \in X$ with $\lim_{n\to\infty} \|u_n - u\| = 0$.*

EXAMPLE 2.2.

(1) *The function space*

$$C(\bar{\Omega}) = \big\{u : \bar{\Omega} \to \mathbb{R} \ : \ u \textit{ continuous}\big\}$$

is a Banach space with the sup-norm

$$\|u\|_{C(\bar{\Omega})} = \sup_{x\in\bar{\Omega}} |u(x)|.$$

(2) *For a multiindex* $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ *we define its order by* $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^n \alpha_i$ *and associate the* $|\alpha|$*-th order partial derivative at* x

$$D^\alpha u(x) \stackrel{\text{def}}{=} \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}(x).$$

The spaces

$$C^k(\bar{\Omega}) = \big\{u \in C(\bar{\Omega}) \ : \ D^\alpha u \in C(\bar{\Omega}) \textit{ for } |\alpha| \le k\big\}$$

are Banach spaces with the norm

$$\|u\|_{C^k(\bar{\Omega})} \stackrel{\text{def}}{=} \sum_{|\alpha|\le k} \|D^\alpha u\|_{C(\bar{\Omega})}.$$

DEFINITION 2.3. (Inner product, Hilbert space)
Let H *be a real vector space.*

i) *A mapping* $(\cdot,\cdot) : H \times H \mapsto \mathbb{R}$ *is an* inner product *on* X*, if*
   1) $(u, v) = (v, u)\ \forall\, u, v \in H$*,*
   2) *For every* $v \in H$ *the mapping* $u \in H \mapsto (u, v)$ *is linear,*
   3) $(u, u) \ge 0\ \forall\, u \in H$ *and* $(u, u) = 0 \iff u = 0$*.*

ii) *A vector space* H *with inner product* $(\cdot,\cdot)$ *and associated norm*

$$\|u\| \stackrel{\text{def}}{=} \sqrt{(u, u)}$$

is called Pre-Hilbert space*.*

iii) *A Pre-Hilbert space* $(H, (\cdot,\cdot))$ *is called* Hilbert space *if it is complete under its norm* $\|u\| \stackrel{\text{def}}{=} \sqrt{(u, u)}$*.*

EXAMPLE 2.4. *Let* $\emptyset \ne \Omega \subset \mathbb{R}^n$ *be open and bounded. Then* $(C(\bar{\Omega}), (\cdot,\cdot)_{L^2})$ *is a Pre-Hilbert space with the* L^2*-inner product*

$$(u, v)_{L^2} = \int_\Omega u(x)\, v(x)\, dx.$$

Note that $(C(\bar{\Omega}), (\cdot,\cdot)_{L^2})$ *is not complete (why?).*

THEOREM 2.5. *Let* H *be a Pre-Hilbert space. Then the* Cauchy-Schwarz inequality *holds*

$$|(u, v)| \le \|u\|\,\|v\| \quad \forall\, u, v \in H.$$

Many spaces arising in applications have the important property that they contain a countable dense subset.

DEFINITION 2.6. *A Banach space X is called* separable *if it contains a countable dense subset. I.e., there exists $Y = \{x_i \in X \,:\, i \in \mathbb{N}\} \subset X$ such that*

$$\forall\, x \in X,\ \forall\, \varepsilon > 0 :\ \exists\, y \in Y :\ \|x - y\|_X < \varepsilon.$$

EXAMPLE 2.7. *For bounded Ω the space $C(\bar{\Omega})$ is separable (the polynomials with rational coefficients are dense by Weierstraß's approximation theorem).*

2.3. Linear operators and dual space. Obviously, linear partial differential operators define linear mappings between function spaces. We recall the following definition.

DEFINITION 2.8. (Linear operator)
Let X, Y be normed vector spaces with norms $\|\cdot\|_X$, $\|\cdot\|_Y$.

i) *A mapping $A : X \to Y$ is called* linear operator *if it satisfies*

$$A(\lambda u + \mu v) = \lambda Au + \mu Av \quad \forall\, u, v \in X,\ \lambda, \mu \in \mathbb{R}.$$

The range *of A is defined by*

$$R(A) \stackrel{\text{def}}{=} \{y \in Y \,:\, \exists\, x \in X :\ y = Ax\}$$

and the null space *of A by*

$$N(A) \stackrel{\text{def}}{=} \{x \in X \,:\, Ax = 0\}.$$

ii) *By $\mathcal{L}(X, Y)$ we denote the space of all linear operators $A : X \to Y$ that are bounded in the sense that*

$$\|A\|_{X,Y} \stackrel{\text{def}}{=} \sup_{\|u\|_X = 1} \|Au\|_Y < \infty.$$

$\mathcal{L}(X, Y)$ is a normed space with the operator norm $\|\cdot\|_{X,Y}$.

THEOREM 2.9. *If Y is a Banach space then $\mathcal{L}(X, Y)$ is a Banach space.*

The following theorem tells us, as a corollary, that if Y is a Banach space, any operator $A \in \mathcal{L}(X, Y)$ is determined uniquely by its action on a dense subspace.

THEOREM 2.10. *Let X be a normed space, Y be a Banach space and let $U \subset X$ be a dense subspace (carrying the same norm as X). Then for all $A \in \mathcal{L}(U, Y)$, there exists a unique extension $\tilde{A} \in \mathcal{L}(X, Y)$ with $\tilde{A}|_U = A$. For this extension, there holds $\|\tilde{A}\|_{X,Y} = \|A\|_{U,Y}$.*

DEFINITION 2.11. (Linear functionals, dual space)

i) *Let X be a Banach space. A bounded linear operator $u^* : X \to \mathbb{R}$, i.e., $u^* \in \mathcal{L}(X, \mathbb{R})$ is called a* bounded linear functional *on X.*
ii) *The space $X^* \stackrel{\text{def}}{=} \mathcal{L}(X, \mathbb{R})$ of linear functionals on X is called* dual space *of X and is (by Theorem 2.9) a Banach space with the operator norm*

$$\|u^*\| \stackrel{\text{def}}{=} \sup_{\|u\|_X = 1} |u^*(u)|.$$

iii) *We use the notation*

$$\langle u^*, u\rangle_{X^*,X} \stackrel{\text{def}}{=} u^*(u).$$

$\langle\cdot,\cdot\rangle_{X^*,X}$ *is called the* dual pairing *of* X^* *and* X.

Of essential importance is the following

THEOREM 2.12. (Riesz representation theorem)
The dual space H^* *of a Hilbert space* H *is isometric to* H *itself. More precisely, for every* $v \in H$ *the linear functional* u^* *defined by*

$$\langle u^*, u\rangle_{H^*,H} \stackrel{\text{def}}{=} (v,u)_H \;\forall\, u \in H$$

is in H^* *with norm* $\|u^*\|_{H^*} = \|v\|_H$. *Vice versa, for any* $u^* \in H^*$ *there exists a unique* $v \in H$ *such that*

$$\langle u^*, u\rangle_{H^*,H} = (v,u)_H \;\forall\, u \in H$$

and $\|u^*\|_{H^*} = \|v\|_H$.

In particular, a Hilbert space is reflexive.

DEFINITION 2.13. *Let* X, Y *be Banach spaces. Then for an operator* $A \in \mathcal{L}(X,Y)$ *the dual operator* $A^* \in \mathcal{L}(Y^*, X^*)$ *is defined by*

$$\langle A^*u, v\rangle_{X^*,X} = \langle u, Av\rangle_{Y^*,Y} \quad \forall\, u \in Y^*,\ v \in X.$$

It is easy to check that $\|A^*\|_{Y^*,X^*} = \|A\|_{X,Y}$.

2.4. Sobolev spaces. To develop a satisfactory theory for PDEs, it is necessary to replace the classical function spaces $C^k(\bar\Omega)$ by *Sobolev spaces* $W^{k,p}(\Omega)$. Roughly speaking, the space $W^{k,p}(\Omega)$ consists of all functions $u \in L^p(\Omega)$ that possess (weak) partial derivatives $D^\alpha u \in L^p(\Omega)$ for $|\alpha| \le k$.

We recall

2.5. Lebesgue spaces. Our aim is to characterize the function space $L^p(\Omega)$ that is complete under the L^p-norm, where

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u(x)|^p\,dx\right)^{1/p}, \quad p \in [1,\infty),$$
$$\|u\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| (= \sup_{x\in\Omega} |u(x)| \text{ for } u \in C(\bar\Omega)).$$

2.6. Lebesgue measurable functions and Lebesgue integral.

DEFINITION 2.14. *A collection* $\mathcal{S} \subset \mathcal{P}(\mathbb{R}^n)$ *of subsets of* $\mathbb{R}^n$ *is called* σ*-algebra on* $\mathbb{R}^n$ *if*

i) $\emptyset, \mathbb{R}^n \in \mathcal{S}$,
ii) $A \in \mathcal{S}$ *implies* $\mathbb{R}^n \setminus A \in \mathcal{S}$,
iii) *if* $(A_k)_{k\in\mathbb{N}} \subset \mathcal{S}$ *then* $\bigcup_{k=1}^\infty A_k \in \mathcal{S}$.

A measure $\mu : \mathcal{S} \to [0,\infty]$ *is a mapping with the following properties:*

i) $\mu(\emptyset) = 0$.
ii) *If* $(A_k)_{k\in\mathbb{N}} \subset \mathcal{S}$ *is a sequence of pairwise disjoint sets then*

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) \qquad (\sigma\text{-additivity}).$$

Of essential importance is the σ-algebra of Lebesgue measurable sets with corresponding Lebesgue measure.

THEOREM 2.15. *There exists the* σ*-algebra* $\mathcal{B}_n$ *of Lebesgue measurable sets on* $\mathbb{R}^n$ *and the Lebesgue measure* $\mu : \mathcal{B}_n \to [0, \infty]$ *with the properties:*

i) $\mathcal{B}_n$ *contains all open sets (and thus all closed sets).*
ii) μ *is a measure on* $\mathcal{B}_n$.
iii) *If* B *is any ball in* $\mathbb{R}^n$ *then* $\mu(B) = |B|$.
iv) *If* $A \subset B$ *with* $B \in \mathcal{B}_n$ *and* $\mu(B) = 0$ *then* $A \in \mathcal{B}_n$ *and* $\mu(A) = 0$ *(*$(\mathbb{R}^n, \mathcal{B}_n, \mu)$ *is a complete measure space).*

The sets $A \in \mathcal{B}_n$ *are called* Lebesgue measurable.

Notation: If some property holds for all $x \in \mathbb{R} \setminus N$ with $N \subset \mathcal{B}_n$, $\mu(N) = 0$, then we say that it holds almost everywhere (a.e.). □

DEFINITION 2.16. *We say that* $f : \mathbb{R}^n \to [-\infty, \infty]$ *is* Lebesgue measurable *if*

$$\{x \in \mathbb{R}^n \,:\, f(x) > \alpha\} \in \mathcal{B}_n \quad \forall\, \alpha \in \mathbb{R}.$$

If $A \in \mathcal{B}_n$ *and* $f : A \to [-\infty, \infty]$ *then we call* f *Lebesgue measurable on* A *if* $f1_A$ *is Lebesgue measurable. Here, we use the convention* $f1_A = f$ *on* A *and* $f1_A = 0$ *otherwise.*

Remark For open $\Omega \subset \mathbb{R}^n$ any function $f \in C(\Omega)$ is measurable, since $\{f > \alpha\}$ is relatively open in Ω (and thus open). □

We now extend the classical integral to Lebesgue measurable functions.

DEFINITION 2.17. *The set of nonnegative elementary functions is defined by*

$$E_+(\mathbb{R}^n) \stackrel{\text{def}}{=} \left\{ f = \sum_{k=1}^{m} \alpha_k 1_{A_k} \,:\, (A_k)_{1\le k\le m} \subset \mathcal{B}_n \text{ pairwise disjoint, } \alpha_k \ge 0,\, m \in \mathbb{N} \right\}.$$

The Lebesgue integral of $f = \sum_{k=1}^{m} \alpha_k 1_{A_k} \in E_+(\mathbb{R}^n)$ *is defined by*

$$\int_{\mathbb{R}^n} f(x)\, d\mu(x) \stackrel{\text{def}}{=} \sum_{k=1}^{m} \alpha_k \mu(A_k).$$

An extension to general Lebesgue measurable functions is obtained by the following fact.

LEMMA 2.18. *For any sequence (f_k) of Lebesgue measurable functions also*

$$\sup_k f_k, \quad \inf_k f_k, \quad \limsup_{k\to\infty} f_k, \quad \liminf_{k\to\infty} f_k$$

are Lebesgue measurable.

For any Lebesgue measurable function $f \geq 0$ there exists a monotone increasing sequence $(f_k)_{k\in\mathbb{N}} \subset E_+(\mathbb{R}^n)$ with $f = \sup_k f_k$.

This motivates the following definition of the Lebesgue integral.

DEFINITION 2.19. (Lebesgue integral)

i) *For a nonnegative Lebesgue measurable function $f : \mathbb{R}^n \to [0,\infty]$ we define the Lebesgue integral of f by*
$$\int_{\mathbb{R}^n} f(x)\, d\mu(x) \stackrel{\text{def}}{=} \sup_k \int_{\mathbb{R}^n} f_k(x)\, d\mu(x),$$
where $(f_k)_{k\in\mathbb{N}} \subset E_+(\mathbb{R}^n)$ is a monotone increasing sequence with $f = \sup_k f_k$.

ii) *For a Lebesgue measurable function $f : \mathbb{R}^n \to [-\infty,\infty]$ we define the Lebesgue integral by*
$$\int_{\mathbb{R}^n} f(x)\, d\mu(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} f^+(x)\, d\mu(x) - \int_{\mathbb{R}^n} f^-(x)\, d\mu(x)$$
with $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$ if one of the terms on the right hand side is finite. In this case f is called integrable.

iii) *If $A \in \mathcal{B}_n$ and $f : A \to [-\infty,\infty]$ is a function such that $f1_A$ is integrable then we define*
$$\int_A f(x)\, d\mu(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} f(x)1_A(x)\, d\mu(x).$$

Notation: In the sequel we will write dx instead of $d\mu(x)$. □

2.7. Definition of Lebesgue spaces. Clearly, we can extend the L^p-norm to Lebesgue measurable functions.

DEFINITION 2.20. *Let $\Omega \in \mathcal{B}_n$. We define for $p \in [1,\infty)$ the seminorm*

$$\|u\|_{L^p(\Omega)} \stackrel{\text{def}}{=} \left(\int_{\mathbb{R}^n} |u(x)|^p\right)^{1/p}.$$

and

$$\|u\|_{L^\infty(\Omega)} \stackrel{\text{def}}{=} \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| \stackrel{\text{def}}{=} \inf\{\alpha \geq 0 \,:\, \mu(\{|u| > \alpha\}) = 0\}.$$

Now, for $1 \leq p \leq \infty$ we define the spaces

$$\mathcal{L}^p(\Omega) \stackrel{\text{def}}{=} \left\{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable} \,:\, \|u\|_{L^p(\Omega)} < \infty\right\}.$$

These are not normed space since there exist mesurable functions $u : \Omega \to \mathbb{R}$, $u \neq 0$, with $\|u\|_{L^p} = 0$.

We use the equivalence relation

$$u \sim v \text{ in } L^p(\Omega) \ :\Longleftrightarrow\ \|u-v\|_{L^p(\Omega)} = 0 \overset{\text{by Lemma 2.21}}{\Longleftrightarrow} u = v \text{ a.e.}$$

to define $L^p(\Omega) = \mathcal{L}^p(\Omega)/\sim$ *as the space of equivalence classes of a.e. identical functions, equipped with the norm* $\|\cdot\|_{L^p}$.

Finally we define

$$\mathcal{L}^p_{loc}(\Omega) \overset{\text{def}}{=} \{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable} : \ u \in \mathcal{L}^p(K) \text{ for all } K \subset \Omega \text{ compact}\}$$

and set $L^p_{loc}(\Omega) \overset{\text{def}}{=} \mathcal{L}^p_{loc}(\Omega)/\sim$.

In the following we will consider elements of L^p *and* L^p_{loc} *as functions that are known up to a set of measure zero.*

Remark It is easy to see that $L^p(\Omega) \subset L^1_{loc}(\Omega)$ for all $p \in [1,\infty]$. □

We collect several important facts of Lebesgue spaces.

LEMMA 2.21. *For all* $u, v \in \mathcal{L}^p(\Omega)$, $p \in [1,\infty]$ *we have*

$$\|u-v\|_{L^p} = 0 \iff u = v \text{ a.e.}.$$

Proof. The assertion is obvious for $p = \infty$. For $p \in [1,\infty)$ let $w = u - v$.

"$\Longrightarrow$:" We have for all $k \in \mathbb{N}$

$$0 = \|w\|_{L^p} \geq \frac{1}{k}\mu(\{|w| \geq 1/k\})^{1/p}.$$

Hence $\mu(\{w \geq 1/k\}) = 0$ and consequently

$$\mu(w \neq 0) = \mu\left(\bigcup_{k=1}^{\infty} \{|w| \geq 1/k\}\right) \leq \sum_{k=1}^{\infty} \mu\{|w| \geq 1/k\}) = 0.$$

"$\Longleftarrow$:" If $w = 0$ a.e. then $|w|^p = 0$ on $\mathbb{R}^n \setminus N$ for some N with $\mu(N) = 0$. Hence, $|w|^p = \sup_k w_k$ with $(w_k) \subset E_+(\mathbb{R}^n)$, where without restriction $w_k = 0$ on $\mathbb{R}^n \setminus N$. Hence $\int_{\mathbb{R}^n} w_k\, dx = 0$ and consequently $\int_{\mathbb{R}^n} |w|^p dx = 0$. □

THEOREM 2.22. (Fischer-Riesz) *The spaces* $L^p(\Omega)$, $p \in [1,\infty]$, *are Banach spaces. The space* $L^2(\Omega)$ *is a Hilbert space with inner product*

$$(u,v) \overset{\text{def}}{=} \int_\Omega uv\, dx.$$

LEMMA 2.23. (Hölder inequality)
Let $\Omega \in \mathcal{B}_n$. *Then for all* $p \in [1,\infty]$ *we have with the dual exponent* $q \in [1,\infty]$ *satisfying* $\frac{1}{p} + \frac{1}{q} = 1$ *for all* $u \in L^p(\Omega)$ *and* $v \in L^q(\Omega)$ *the* Hölder inequality

$$uv \in L^1(\Omega) \text{ and } \|uv\|_{L^1} \leq \|u\|_{L^p}\|v\|_{L^q}.$$

Now we can characterize the dual space of L^p-spaces.

THEOREM 2.24. *Let $\Omega \in \mathcal{B}_n$, $p \in [1,\infty)$ and $q \in (1,\infty]$ the dual exponent satisfying $\frac{1}{p}+\frac{1}{q}=1$. Then the dual space $(L^p(\Omega))^*$ can be identified with $L^q(\Omega)$ by means of the isometric isomorphism*

$$v \in L^q(\Omega) \mapsto u^* \in (L^p(\Omega))^*, \quad \text{where } \langle u^*, u\rangle_{(L^p)^*,L^p} \stackrel{\text{def}}{=} \int_\Omega u(x)v(x)\,dx.$$

Remark Note however that L^1 is only a subspace of $(L^\infty)^*$. □

2.8. Density results and convergence theorems. A fundamental result is the following:

THEOREM 2.25 (Dominated convergence theorem). *Let $\Omega \in \mathcal{B}_n$. Assume that $f_k : \Omega \to \mathbb{R}$ are measurable with*

$$f_k \to f \text{ a.e.} \quad \text{and} \quad |f_k| \le g \text{ a.e.}$$

with a function $g \in \mathcal{L}^1(\Omega)$. Then $f_k, f \in \mathcal{L}^1(\Omega)$ and

$$\int_\Omega f_k\,dx \to \int_\Omega f\,dx, \quad f_k \to f \quad \text{in } L^1(\Omega).$$

Next we state the important fact that the set of "nice" functions

$$C_c^\infty(\Omega) \stackrel{\text{def}}{=} \big\{u \in C^\infty(\bar\Omega) \,:\, \operatorname{supp}(u) \subset \Omega \text{ compact}\big\}$$

is actually dense in $L^p(\Omega)$ for all $p \in [1,\infty)$.

LEMMA 2.26. *Let $\Omega \subset \mathbb{R}^n$ be open. Then $C_c^\infty(\Omega)$ is dense in $L^p(\Omega)$ for all $p \in [1,\infty)$.*

A quite immediate consequence is the following useful result.

LEMMA 2.27. *Let $\Omega \subset \mathbb{R}^n$ be open and $f \in L^1_{loc}(\Omega)$ with*

$$\int_\Omega f(x)\varphi(x)\,dx = 0 \quad \forall\, \varphi \in C_c^\infty(\Omega).$$

Then $f = 0$ a.e.

2.9. Weak derivatives. The definition of weak derivatives is motivated by the fact that for any function $u \in C^k(\bar\Omega)$ and any multiindex $\alpha \in \mathbb{N}_0^n$, $|\alpha \le k$, the identity holds (integrate $|\alpha|$-times by parts)

$$\int_\Omega D^\alpha u\varphi\,dx = (-1)^{|\alpha|}\int_\Omega uD^\alpha\varphi\,dx, \quad \forall\,\varphi \in C_c^\infty(\Omega). \tag{2.1}$$

This motivates the definition

DEFINITION 2.28. *Let $\Omega \subset \mathbb{R}^n$ be open and let $u \in L^1_{loc}(\Omega)$. If there exists a function $w \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega w\varphi\,dx = (-1)^{|\alpha|}\int_\Omega uD^\alpha\varphi\,dx, \quad \forall\,\varphi \in C_c^\infty(\Omega) \tag{2.2}$$

then $D^\alpha u := w$ is called the α-th weak partial derivative of u.

Remark

(1) By Lemma 2.27, (2.2) determines the weak derivative $D^\alpha u \in L^1_{loc}(\Omega)$ uniquely.
(2) For $u \in C^k(\bar{\Omega})$ and $\alpha \in \mathbb{N}_0^n$, $|\alpha| \le k$, the classical derivative $w = D^\alpha u$ satisfies (2.1) and thus (2.2). Hence, the weak derivative is consistent with the classical derivative. □

2.10. Regular domains and integration by parts. For $k \in \mathbb{N}_0$ and $\beta \in (0, 1]$ let

$$C^{k,\beta}(\mathbb{R}^n) = \left\{ u \in C^k(\mathbb{R}^n) \,:\, D^\alpha u \text{ } \beta\text{-Hölder continuous for } |\alpha| = k \right\}.$$

Here, f is β-Hölder continuous if there exists a constant $C > 0$ such that

$$|f(x) - f(y)| \le C|x - y|^\beta \quad \forall\, x, y.$$

Of course, 1-Hölder continuity is Lipschitz continuity.

We set $C^{k,0}(\mathbb{R}^n) = C^k(\mathbb{R}^n) =$.

DEFINITION 2.29. ($C^{k,\beta}$-boundary, unit normal field)
Let $\Omega \subset \mathbb{R}^n$ be open and bounded.

a) We say that Ω has a *$C^{k,\beta}$-boundary*, $k \in \mathbb{N}_0 \cup \{\infty\}$, $0 \le \beta \le 1$, if for any $x \in \partial U$ there exists $r > 0$, $k \in \{1, \dots, n\}$, and a function $\gamma \in C^k(\mathbb{R}^{n-1})$ such that

$$\Omega \cap B(x; r) = \{ y \in B(x; r) \,:\, y_k < \gamma(y_1, \dots y_{k-1}, y_{k+1}, \dots, y_n) \}.$$

Instead of $C^{0,1}$-boundary we say also *Lipschitz-boundary*.
b) If $\partial\Omega$ is $C^{0,1}$ then we can define a.e. the *unit outer normal field* $\nu : \partial\Omega \to \mathbb{R}^n$, where $\nu(x)$, $\|\nu(x)\|_2 = 1$, is the outward pointing unit normal of $\partial\Omega$ at x.
c) Let $\partial\Omega$ be $C^{0,1}$. We call the directional derivative

$$\frac{\partial u}{\partial \nu}(x) \stackrel{\text{def}}{=} \nu(x) \cdot \nabla u(x), \quad x \in \partial\Omega,$$

the *normal derivative* of u.

We recall the Gauß-Green theorem (integration by parts formula).

THEOREM 2.30. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with C^1-boundary. Then for all $u, v \in C^1(\bar{\Omega})$*

$$\int_\Omega u_{x_i}(x) v(x)\, dx = -\int_\Omega u(x) v_{x_i}(x)\, dx + \int_{\partial\Omega} u(x) v(x) \nu_i(x)\, dS(x).$$

2.11. Sobolev spaces. We will now introduce subspaces $W^{k,p}(\Omega)$ of functions $u \in L^p(\Omega)$, for which the weak derivatives $D^\alpha u$, $|\alpha| \le k$, are in $L^p(\Omega)$.

DEFINITION 2.31. *Let $\Omega \subset \mathbb{R}^n$ be open. For $k \in \mathbb{N}_0$, $p \in [1, \infty]$, we define the* Sobolev space *$W^{k,p}(\Omega)$ by*

$$W^{k,p}(\Omega) = \{ u \in L^p(\Omega) \,:\, u \text{ has weak derivatives } D^\alpha u \in L^p(\Omega) \text{ for all } |\alpha| \le k \} \tag{2.3}$$

equipped with the norm

$$\|u\|_{W^{k,p}(\Omega)} \stackrel{\text{def}}{=} \left(\sum_{|\alpha|\le k} \|D^\alpha u\|_{L^p}^p \right)^{1/p}, \quad p \in [1,\infty),$$
$$\|u\|_{W^{k,\infty}(\Omega)} \stackrel{\text{def}}{=} \sum_{|\alpha|\le k} \|D^\alpha u\|_{L^\infty(\Omega)}.$$

REMARK 2.32. • *The set $C^\infty(\Omega) \cap W^{k,p}(\Omega)$, $k \in \mathbb{N}_0$, $1 \le p < \infty$, is dense in $W^{k,p}(\Omega)$. Hence, $W^{k,p}(\Omega)$ is the completion of $\{u \in C^\infty(\Omega) : \|u\|_{W^{k,p}} < \infty\}$ with respect to the norm $\|\cdot\|_{W^{k,p}}$.*

• *If Ω is a bounded Lipschitz-domain then $C^\infty(\bar\Omega)$ is dense in $W^{k,p}(\Omega)$, $k \in \mathbb{N}_0$, $1 \le p < \infty$.*

Notations:

(1) In the case $p = 2$ one writes $H^k(\Omega) \stackrel{\text{def}}{=} W^{k,2}(\Omega)$. We note that $W^{0,p}(\Omega) = L^p(\Omega)$ for $p \in [1,\infty]$.
(2) For weak partial derivatives we use also the notation u_{x_i}, $u_{x_i x_j}$, $u_{x_i x_j x_k}, \ldots$
(3) For $u \in H^1(\Omega)$ we set

$$\nabla u(x) = \begin{pmatrix} u_{x_1}(x) \\ \vdots \\ u_{x_n}(x) \end{pmatrix}.$$

□

Remark Simple examples show that weak differentiability does not necessarily ensures continuity. We have for example with $\Omega \stackrel{\text{def}}{=} B(0;1)$ and $u(x) \stackrel{\text{def}}{=} \|x\|^{-\beta}$ that

$$u \in W^{1,p}(\Omega) \iff \beta < \frac{n-p}{p}.$$

□

THEOREM 2.33. *Let $\Omega \subset \mathbb{R}^n$ be open, $k \in \mathbb{N}_0$, and $p \in [1,\infty]$. Then $W^{k,p}(\Omega)$ is a Banach space. Moreover, the space $H^k(\Omega) = W^{k,2}(\Omega)$ is a Hilbert space with inner product*

$$(u,v)_{H^k(\Omega)} = \sum_{|\alpha|\le k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}.$$

To incorporate homogeneous boundary conditions already in the function space we define the following subspace.

DEFINITION 2.34. *Let $\Omega \subset \mathbb{R}^n$ be open. For $k \in \mathbb{N}_0$, $p \in [1,\infty]$, we denote by*

$$W_0^{k,p}(\Omega)$$

the closure of $C_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$ (i.e., for any $u \in W_0^{k,p}(\Omega)$ there exists a sequence $(\varphi_i) \subset C_c^\infty(\Omega)$ with $\lim_{i\to\infty} \|u - \varphi_i\|_{W^{k,p}(\Omega)} = 0$). The space is equipped with the same norm as $W^{k,p}(\Omega)$ and is a Banach space. The space $H_0^k(\Omega) = W_0^{k,2}(\Omega)$ is a Hilbert space.

REMARK 2.35. $W_0^{k,p}(\Omega)$ contains exactly all $u \in W^{1,p}(\Omega)$ such that $D^\alpha u = 0$ for $|\alpha| \le k-1$ on $\partial\Omega$ with an appropriate interpretation of the *traces* $D^\alpha u|_{\partial\Omega}$. □

We consider next the appropriate assignment of boundary values (so called *boundary traces*) for functions $u \in W^{k,p}(\Omega)$ if Ω has Lipschitz-boundary.

If $u \in W^{k,p}(\Omega) \cap C(\bar\Omega)$ then the boundary values can be defined in the classical sense by using the continuous extension. However, since $\partial\Omega$ is a set of measure zero and functions $u \in W^{k,p}(\Omega)$ are only determinded up to a set of measure zero, the definition of boundary values requires care. We resolve the problem by defining a *trace operator*.

THEOREM 2.36. *Assume that $\Omega \subset \mathbb{R}^n$ is open and bounded with Lipschitz-boundary. Then for all $p \in [1,\infty]$ there exists a unique bounded linear operator*

$$T : W^{1,p}(\Omega) \to L^p(\partial\Omega)$$

such that

$$Tu = u|_{\partial\Omega} \quad \forall\, u \in W^{1,p}(\Omega) \cap C(\bar\Omega).$$

Here, $\|T\|_{W^{1,p}(\Omega),L^p(\partial\Omega)}$ depends only on Ω and p. Tu is called the trace *of u on $\partial\Omega$.*

2.12. Poincaré's inequality. We have seen that the trace of functions in $H_0^k(\Omega)$, $k \ge 0$, vanishes. For the treatment of boundary value problems it will be useful that the semi-norm

$$|u|_{H^k(\Omega)} \stackrel{\text{def}}{=} \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{L^2}^2 \right)^{1/2} \tag{2.4}$$

defines an equivalent norm on the Hilbert space $H_0^k(\Omega)$. It is obvious that

$$|u|_{H^k(\Omega)} \le \|u\|_{H^k(\Omega)}.$$

We will now show that also

$$\|u\|_{H^k(\Omega)} \le C\, |u|_{H^k(\Omega)} \quad \forall\, u \in H_0^k(\Omega). \tag{2.5}$$

THEOREM 2.37. (Poincaré's inequality)
Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then there exists a constant $C > 0$ with

$$|u|_{H^k(\Omega)} \le \|u\|_{H^k(\Omega)} \le C\, |u|_{H^k(\Omega)} \quad \forall\, u \in H_0^k(\Omega). \tag{2.5}$$

2.13. Sobolev imbedding theorem. Sobolev spaces are embedded in classical spaces:

THEOREM 2.38. *Let $\Omega \subset \mathbb{R}^n$ be open, bounded with Lipschitz-boundary. Let $m \in \mathbb{N}$, $1 \leq p < \infty$.*

i) *For all $k \in \mathbb{N}_0$, $0 < \beta < 1$ with*

$$m - \frac{n}{p} \geq k + \beta$$

one has the continuous embedding

$$W^{m,p}(\Omega) \subset C^{k,\beta}(\bar{\Omega}).$$

More precisely, there exists a constant $C > 0$ such that for all $u \in W^{m,p}(\Omega)$ possibly after modification on a set of measure zero $u \in C^{k,\beta}(\bar{\Omega})$ and

$$\|u\|_{C^{k,\beta}(\bar{\Omega})} \leq C\|u\|_{W^{m,p}(\Omega)}.$$

ii) *For all $k \in \mathbb{N}_0$, $0 \leq \beta \leq 1$ with*

$$m - \frac{n}{p} > k + \beta$$

one has the compact embedding

$$W^{m,p}(\Omega) \subset\subset C^{k,\beta}(\bar{\Omega}),$$

i.e., closed balls in $W^{m,p}(\Omega)$ are relatively compact in $C^{k,\beta}(\bar{\Omega})$.

iii) *For $q \geq 1$ and $l \in \mathbb{N}_0$ with $m - n/p \geq l - n/q$ one has the continuous embedding*

$$W^{m,p}(\Omega) \subset W^{l,q}(\Omega).$$

The embedding is compact if $m - n/p > l - n/q$ and for $l = 0$ we have $W^{0,q}(\Omega) = L^q(\Omega)$.

For arbitrary open bounded $\Omega \subset \mathbb{R}^n$ i), ii), iii) hold for $W_0^{m,p}(\Omega)$ instead of $W^{m,p}(\Omega)$.

Proof. See for example **[3]**, **[1]**, **[28]**. □

EXAMPLE 2.39. *For $n \leq 3$ we have the continuous imbedding $H^1(\Omega) \subset L^6(\Omega)$ and the compact imbedding $H^2(\Omega) \subset\subset C(\bar{\Omega})$ for $n \leq 3$.*

2.14. The dual space H^{-1} of H_0^1. The dual space of the Hilbert space $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega)$. This space can be characterized as follows:

THEOREM 2.40. *For the space $H^{-1}(\Omega)$, $\Omega \subset \mathbb{R}^n$ open, the following holds:*

$$H^{-1}(\Omega) = \left\{ v \in H_0^1(\Omega) \mapsto (f^0, v)_{L^2} + \sum_{j=1}^{n} (f^j, v_{x_j})_{L^2} \ : \ f_j \in L^2(\Omega) \right\}.$$

Furthermore,

$$\|f\|_{H^{-1}} = \min \left\{ \left(\sum_{j=0}^{n} \|f^j\|_{L^2}^2 \right)^{1/2} \ : \ \langle f, v \rangle_{H^{-1}, H_0^1} = (f^0, v)_{L^2} + \sum_{j=1}^{n} (f^j, v_{x_j})_{L^2}, \ f^j \in L^2(\Omega) \right\}.$$

Proof. "$\subset$": Let $f \in H^{-1}(\Omega)$. By the Riesz representation theorem, there exists a unique $u \in H_0^1(\Omega)$ with

$$(u,v)_{H^1} = \langle f, v\rangle_{H^{-1},H_0^1} \quad \forall\, v \in H_0^1(\Omega).$$

Set $f^0 = u$, $f^j = u_{x_j}$, $j \geq 1$.

Then

$$(f^0,v)_{L^2} + \sum_{j=1}^n (f^j, v_{x_j})_{L^2} = (u,v)_{L^2} + \sum_{j=1}^n (u_{x_j}, v_{x_j})_{L^2} = (u,v)_{H^1} = \langle f, v\rangle_{H^{-1},H_0^1} \quad \forall\, v \in H_0^1(\Omega).$$

"$\supset$": For $g_0, \ldots, g_n \in L^2(\Omega)$, consider

$$g : v \in H_0^1(\Omega) \mapsto (g^0, v)_{L^2} + \sum_{j=1}^n (g^j, v_{x_j})_{L^2}.$$

Obviously, g is linear. Furthermore, for all $v \in H_0^1(\Omega)$, there holds

$$\begin{aligned}
|(g^0,v)_{L^2} + \sum_{j=1}^n (g^j, v_{x_j})_{L^2}| &\leq \|g^0\|_{L^2}\|v\|_{L^2} + \sum_{j=1}^n \|g^j\|_{L^2}\|v_{x_j}\|_{L^2} \\
&\leq \left(\sum_{j=0}^n \|g^j\|_{L^2}^2\right)^{1/2} \left(\|v\|_{L^2}^2 + \sum_{j=1}^n \|v_{x_j}\|_{L^2}\right)^{1/2} \\
&= \left(\sum_{j=0}^n \|g^j\|_{L^2}^2\right)^{1/2} \|v\|_{H^1}.
\end{aligned}$$

This shows $g \in H^{-1}(\Omega)$ and

$$\|g\|_{H^{-1}} \leq \left(\sum_{j=0}^n \|g^j\|_{L^2}^2\right)^{1/2}.$$

Now let $f = g$, let u be the Riesz representation, and choose

$$(f^0, \ldots, f^n) = (u, u_{x_1}, \ldots, u_{x_n})$$

as above. Then by the Riesz representation theorem

$$\|g\|_{H^{-1}}^2 = \|f\|_{H^{-1}}^2 = \|u\|_{H^1}^2 = \|u\|_{L^2}^2 + \sum_{j=1}^n \|u_{x^j}\|_{L^2}^2 = \sum_{j=0}^n \|f^j\|_{L^2}^2.$$

□

2.15. Weak solutions of elliptic PDEs. In this section we sketch the theory of weak solutions for elliptic second order partial differential equations. For more details we refer, e.g., to [**3**], [**28**], [**65**], [**74**], [**80**].

2.16. Weak solutions of the Poisson equation.

2.16.1. *Dirichlet boundary conditions.* We start with the elliptic boundary value problem

$$-\Delta y = f \quad \text{on } \Omega, \tag{2.6}$$

$$y = 0 \quad \text{on } \partial\Omega, \quad \text{(Dirichlet condition)} \tag{2.7}$$

where $\Omega \subset \mathbb{R}^n$ is an open, bounded set and $f \in L^2(\Omega)$. This admits discontinuous right hand sides f, e.g. source terms f that act only on a subset of Ω. Since a classical solution $y \in C^2(\Omega) \cap C^1(\bar{\Omega})$ exists at best for continuous right hand sides, we need a generalized solution concept. It is based on a *variational formulation* of (2.6)–(2.7).

To this end let us assume that $y \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is a classical solution of (2.6)–(2.7). Then we have $y \in H_0^1(\Omega)$ by Remark 2.35. Multiplying by $v \in C_c^\infty(\Omega)$ and integrating over Ω yields

$$-\int_\Omega \Delta y \, v \, dx = \int_\Omega f v \, dx \quad \forall\, v \in C_c^\infty(\Omega). \tag{2.8}$$

It is easy to see that (2.6) and (2.8) are equivalent for classical solutions. Now integration by parts gives

$$-\int_\Omega y_{x_i x_i} \, v \, dx = \int_\Omega y_{x_i} v_{x_i} \, dx - \int_{\partial\Omega} y_{x_i} \, v \, \nu_i \, dS(x) = \int_\Omega y_{x_i} v_{x_i} \, dx. \tag{2.9}$$

Note that the boundary integral vanishes, since $v|_{\partial\Omega} = 0$. Thus, (2.8) is equivalent to

$$\int_\Omega \nabla y \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall\, v \in C_c^\infty(\Omega). \tag{2.10}$$

We note that this variational equation makes already perfect sense in a larger space:

LEMMA 2.41. *The mapping*

$$(y, v) \in H_0^1(\Omega)^2 \mapsto a(u, v) \stackrel{\text{def}}{=} \int_\Omega \nabla y \cdot \nabla v \, dx \in \mathbb{R}$$

is bilinear and bounded:

$$|a(y, v)| \le \|y\|_{H^1} \|v\|_{H^1}. \tag{2.11}$$

For $f \in L^2(\Omega)$, the mapping

$$v \in H_0^1(\Omega) \mapsto \int_\Omega f v \, dx \in \mathbb{R}$$

is linear and bounded:

$$\left| \int_\Omega f v \, dx \right| = (f, v)_{L^2} \le \|f\|_{L^2} \|v\|_{L^2} \le \|f\|_{L^2} \|v\|_{H_0^1}. \tag{2.12}$$

Proof. Clearly, $a(y,v)$ is bilinear. The boundedness follows from

$$\begin{aligned}|a(y,v)| &\le \int_\Omega |\nabla y(x)^T \nabla v(x)|\,dx \le \int_\Omega \|\nabla y(x)\|_2 \|\nabla v(x)\|_2\,dx \\ &\le \|\|\nabla y\|_2\|_{L^2}\|\|\nabla v\|_2\|_{L^2} = |y|_{H^1}|v|_{H^1} \le \|y\|_{H^1}\|v\|_{H^1} = \|y\|_V\|v\|_V,\end{aligned}$$

where we have applied the Cauchy-Schwarz inequality.

The second assertion is trivial. □

By density and continuity, we can extend (2.10) to $y \in H_0^1(\Omega)$ and $v \in H_0^1(\Omega)$. We arrive at the *variational formulation*

$$\int_\Omega \nabla y \cdot \nabla v\,dx = \int_\Omega fv\,dx \quad \forall\, v \in H_0^1(\Omega). \tag{2.13}$$

We summarize: (2.6) and (2.13) are equivalent for a classical solution $y \in C^2(\Omega) \cap C^1(\bar\Omega)$. But the variational formulation (2.13) makes already perfectly sense for $y \in H_0^1(\Omega)$ and $f \in L^2(\Omega)$. This motivates the following definition.

DEFINITION 2.42. *A function $y \in H_0^1(\Omega)$ is called* weak solution *of the boundary value problem* (2.6)–(2.7) *if it satisfies the* variational formulation *or* weak formulation

$$\int_\Omega \nabla y \cdot \nabla v\,dx = \int_\Omega fv\,dx \quad \forall\, v \in H_0^1(\Omega). \tag{2.13}$$

In order to allow a uniform treatment of more general equations than (2.6)–(2.7), we introduce the following abstract notation. Let

$$V = H_0^1(\Omega),$$

$$a(y,v) = \int_\Omega \nabla y \cdot \nabla v\,dx, \quad y,v \in V, \tag{2.14}$$

$$F(v) = (f,v)_{L^2(\Omega)}, \quad v \in V. \tag{2.15}$$

Then $a : V \times V \to \mathbb{R}$ is a bilinear form, $F \in V^*$ is a linear functional on V and (2.13) can be written as

$$\text{Find } y \in V\text{:} \quad a(y,v) = F(v) \quad \forall\, v \in V. \tag{2.16}$$

Remark Since $a(y,\cdot) \in V^*$ for all $y \in V$ and $y \in V \mapsto a(y,\cdot) \in V^*$ is conitnuous and linear, there exists a bounded linear operator $A : V \to V^*$ with

$$a(y,v) = \langle Ay, v\rangle_{V^*,V} \quad \forall\, y,v \in V. \tag{2.17}$$

Then (2.16) can be written in the form

$$\text{Find } y \in V\text{:} \quad Ay = F. \tag{2.18}$$

□

We have the following important existence and uniqueness result for solutions of (2.16).

LEMMA 2.43. (Lax-Milgram lemma)
Let V be a real Hilbert space with inner product $(\cdot,\cdot)_V$ and let $a : V \times V \to \mathbb{R}$ be a bilinear form that satisfies with constants $\alpha_0, \beta_0 > 0$

$$|a(y,v)| \le \alpha_0 \|y\|_V \|v\|_V \quad \forall\, y, v \in V, \qquad \textit{(boundedness)} \tag{2.19}$$

$$a(y,y) \ge \beta_0 \|y\|_V^2 \quad \forall\, y \in V \qquad \textit{(V-coercivity).} \tag{2.20}$$

Then for any bounded linear functional $F \in V^*$ the variational equation (2.16) *has a unique solution $y \in V$. Moreover, u satisfies*

$$\|y\|_V \le \frac{1}{\beta_0} \|F\|_{V^*}. \tag{2.21}$$

In particular the operator A defined in (2.17) *satisfies*

$$A \in \mathcal{L}(V, V^*), \quad A^{-1} \in \mathcal{L}(V^*, V), \quad \|A^{-1}\|_{V^*,V} \le \frac{1}{\beta_0}.$$

Remark If $a(\cdot,\cdot)$ is symmetric, i.e., if $a(y,v) = a(v,y)$ for all $y, v \in V$, then the Lax-Milgram lemma is an immediate consequence of the Riesz representation theorem. In fact, in this case $(u,v) := a(u,v)$ defines a new inner product on V and the existence of a unique solution of (2.16) follows directly from the Riesz representation theorem. □

Application of the Lax-Milgram lemma to (2.13) yields

THEOREM 2.44. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz-boundary.*

Then the bilinear form a in (2.14) *is bounded and V-coercive for $V = H_0^1(\Omega)$ and the associated operator $A \in \mathcal{L}(V, V^*)$ in* (2.17) *has a bounded inverse. In particular,* (2.6)–(2.7) *has for all $f \in L^2(\Omega)$ a unique weak solution $y \in H_0^1(\Omega)$ given by* (2.13) *and satisfies*

$$\|y\|_{H^1(\Omega)} \le C_P \|f\|_{L^2(\Omega)},$$

where C_P depends on Ω but not on f.

Proof. We verify the hypotheses of Lemma 2.43. Clearly, $a(y,u)$ in (2.14) is bilinear. The boundedness 2.19 follows from (2.11) Using the Poincaré's inequality (2.5) we obtain

$$a(y,y) = \int_\Omega \nabla y \cdot \nabla y \, dx = |y|^2_{H_0^1(\Omega)} \ge \frac{1}{C^2} \|y\|^2_{H_0^1(\Omega)} = \frac{1}{C^2} \|y\|_V^2$$

which shows the V-coercivity (2.20).

Finally, the definition of F in (2.15) yields

$$\|F\|_{V^*} = \sup_{\|v\|_V = 1} F(v) = \sup_{\|v\|_V = 1} (f,v)_{L^2(\Omega)} \le \sup_{\|v\|_V = 1} \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \le \|f\|_{L^2(\Omega)}.$$

Thus, the assertion holds with $C_P = C^2$ by the Lax-Milgram lemma. □

2.16.2. *Boundary conditions of Robin type.* We have seen that in heating applications the boundary condition is sometimes of Robin type. We consider now problems of the form

$$-\Delta y + c_0 y = f \quad \text{on } \Omega, \tag{2.22}$$

$$\frac{\partial y}{\partial \nu} + \alpha y = g \quad \text{on } \partial\Omega, \quad \text{(Robin condition)} \tag{2.23}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ are given and $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\partial\Omega)$ are nonnegative coefficients.

Weak solutions can be defined similarly as above. If y is a classical solution of (2.22)–(2.23) then for any test function $v \in C^1(\bar\Omega)$ integration by parts, see (2.9), yields as above

$$\begin{aligned}&\int_\Omega (-\Delta y + c_0 y)\, v\, dx = \\ &= \int_\Omega \nabla y \cdot \nabla v\, dx + (c_0 y, v)_{L^2(\Omega)} - \int_{\partial\Omega} \frac{\partial y}{\partial \nu}\, v\, dS(x) = \int_\Omega f v\, dx \quad \forall\, v \in C^1(\bar\Omega).\end{aligned}$$

Inserting the boundary condition $\frac{\partial y}{\partial \nu} = -\alpha y + g$ we arrive at

$$\int_\Omega \nabla y \cdot \nabla v\, dx + (c_0 y, v)_{L^2(\Omega)} + (\alpha y, v)_{L^2(\partial\Omega)} = (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\partial\Omega)} \quad \forall\, v \in H^1(\Omega). \tag{2.24}$$

The extension to $v \in H^1(\Omega)$ is possible, since for $y \in H^1(\Omega)$ both sides are continuous with respect to $v \in H^1(\Omega)$ and since $C^1(\bar\Omega)$ is dense in $H^1(\Omega)$.

DEFINITION 2.45. *A function $y \in H^1(\Omega)$ is called* weak solution *of the boundary value problem* (2.22)–(2.23) *if it satisfies the* variational formulation *or* weak formulation (2.24).

To apply the general theory, we set

$$\begin{aligned} V &= H^1(\Omega), \\ a(y, v) &= \int_\Omega \nabla y \cdot \nabla v\, dx + (c_0 y, v)_{L^2(\Omega)} + (\alpha y, v)_{L^2(\partial\Omega)}, \quad y, v \in V, \\ F(v) &= (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\partial\Omega)}, \quad v \in V. \end{aligned} \tag{2.25}$$

The Lax-Milgram lemma yields similarly as above

THEOREM 2.46. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz-boundary and let $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\partial\Omega)$ be nonnegative with $\|c_0\|_{L^2(\Omega)} + \|\alpha\|_{L^2(\partial\Omega)} > 0$.*

Then the bilinear form a in (2.25) *is bounded and V-coercive for $V = H^1(\Omega)$ and the associated operator $A \in \mathcal{L}(V, V^*)$ in* (2.17) *has a bounded inverse. In particular,* (2.6)–(2.7) *has for all $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ a unique weak solution $y \in H^1(\Omega)$ given by* (2.24) *and satisfies*

$$\|y\|_{H^1(\Omega)} \le C_R(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}),$$

where C_R depends on Ω, α, c_0 but not on f, g.

Proof. The proof is an application of the Lax-Milgram lemma. The boundedness of $a(y,v)$ and of $F(v)$ follows by the trace theorem. The V-coercivity is a consequence of a generalized Poincaré inequality. □

A refined analysis yields the following result [**74**].

THEOREM 2.47. *Let the assumptions of the previous theorem hold and let $r > n/2$, $s > n-1$, $n \geq 2$. Then for any $f \in L^r(\Omega)$ and $g \in L^s(\partial\Omega)$ there exists a unique weak solution $y \in H^1(\Omega) \cap C(\bar{\Omega})$ of* (2.6)–(2.7). *There exists a constant $C_\infty > 0$ with*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq C_R(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\partial\Omega)}),$$

where C_∞ depends on Ω, α, c_0 but not on f, g.

An analogous result holds for homogeneous Dirichlet boundary conditions instead of Robin boundary conditions [**49**].

2.17. Weak solutions of uniformly elliptic equations. More generally, we can consider general second order elliptic PDEs of the form

$$Lu = f \quad \text{on } \Omega \tag{2.26}$$

with

$$Ly \stackrel{\text{def}}{=} -\sum_{i,j=1}^{n} (a_{ij} y_{x_i})_{x_j} + c_0 y, \quad a_{ij}, c_0 \in L^\infty, \quad c_0 \geq 0, \quad a_{ij} = a_{ji} \tag{2.27}$$

and L is assumed to be *uniformly elliptic* in the sense that there is a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \theta\|\xi\|^2 \quad \text{for almost all } x \in \Omega \text{ and all } \xi \in \mathbb{R}^n. \tag{2.28}$$

For example in the case of Dirichlet boundary conditions

$$y|_{\partial\Omega} = 0$$

the weak formulation of (2.26) is given by

$$\text{Find } y \in V := H_0^1(\Omega)\text{: } a(y,v) = (f,v)_{L^2(\Omega)} \quad \forall\, v \in V$$

with the bilinear form

$$a(y,v) = \int_\Omega \sum_{i,j=1}^{n} (a_{ij}\, y_{x_i} v_{x_j} + c\, y\, v)\, dx.$$

Our previous results remain true, if in the case of the Robin boundary condition the normal derivative is replaced by the conormal derivative

$$\frac{\partial y}{\partial \nu_A}(x) \stackrel{\text{def}}{=} \nabla y(x) \cdot A(x)\nu(x), \quad A(x) = (a_{ij}(x)), \tag{2.29}$$

For continuous solutions, we have to assume

2.18. An existence and uniqueness result for semilinear elliptic equations. We finally state an existence and uniqueness result for a uniformly elliptic semilinear equation

$$\begin{aligned} Ly + d(x,y) &= f \quad \text{on } \Omega \\ \frac{\partial y}{\partial \nu} + \alpha y + b(x,y) &= g \quad \text{on } \partial\Omega \end{aligned} \tag{2.30}$$

where the operator L is given by

$$Ly := -\sum_{i,j=1}^{n} (a_{ij} y_{x_i})_{x_j} + c_0 y, \quad a_{ij}, c_0 \in L^\infty, \quad c_0 \geq 0, \quad a_{ij} = a_{ji} \tag{2.27}$$

and L is assumed to be uniformly elliptic in the sense that there is a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \theta \|\xi\|^2 \quad \text{for almost all } x \in \Omega \text{ and all } \xi \in \mathbb{R}^n. \tag{2.28}$$

Moreover, we assume that $0 \leq \alpha \in L^\infty(\partial\Omega)$ and that the functions $d : \Omega \times \mathbb{R} \to \mathbb{R}$, and $b : \partial\Omega \times \mathbb{R} \to \mathbb{R}$ satisfy

$$\begin{aligned} &d(x,\cdot) \text{ is continuous and monotone increasing for a.a. } x \in \Omega, \\ &b(x,\cdot) \text{ is continuous and monotone increasing for a.a. } x \in \partial\Omega, \\ &d(\cdot,y), b(\cdot,y) \text{ measurable for all } y \in \mathbb{R}. \end{aligned} \tag{2.31}$$

Under these assumptions the theory of maximal monotone operators and a technique of Stampacchia can be applied to extend Theorem 2.47 to the semilinear elliptic equation (2.30), see for example [**74**].

THEOREM 2.48. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz-boundary, let $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\partial\Omega)$ be nonnegative with $\|c_0\|_{L^2(\Omega)} + \|\alpha\|_{L^2(\partial\Omega)} > 0$ and let* (2.28)*,* (2.31) *be satisfied. Moreover, let $r > n/2$, $s > n-1$, $2 \leq n \leq 3$. Then* (2.30)*,* (2.27) *has for any $f \in L^r(\Omega)$ and $g \in L^s(\partial\Omega)$ a unique weak solution $y \in H^1(\Omega) \cap C(\bar{\Omega})$. There exists a constant $C_\infty > 0$ with*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq C_R(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\partial\Omega)} + 1),$$

where C_∞ depends on Ω, α, c_0 but not on f, g, b, d.

2.19. Gâteaux- and Fréchet Differentiability. We extend the notion of differentiability to operators between Banach spaces.

DEFINITION 2.49. *Let $F : U \subset X \to Y$ be an operator with X, Y Banach spaces and $U \neq \emptyset$ open.*

a) F is called directionally differentiable *at $x \in U$ if the limit*

$$dF(x,h) = \lim_{t \to 0^+} \frac{F(x+th) - F(x)}{t} \in Y$$

exists for all $h \in X$. In this case, $dF(x,h)$ is called directional derivative of F in the direction h.

b) F is called Gâteaux differentiable *at $x \in U$ if F is directionally differentiable at x and the directional derivative $F'(x) : X \ni h \mapsto dF(x,h) \in Y$ is bounded and linear, i.e., $F'(x) \in \mathcal{L}(X,Y)$.*

c) F is called Fréchet differentiable *at $x \in U$ if F is Gâteaux differentiable at x and if the following approximation condition holds:*

$$\|F(x+h) - F(x) - F'(x)h\|_Y = o(\|h\|_X) \quad \textit{for } \|h\|_X \to 0.$$

d) If F is directionally-/G-/F-differentiable at every $x \in V$, $V \subset U$ open, then F is called directionally-/G-/F-differentiable on V.

Higher derivatives can be defined as follows:

If F is G-differentiable in a neighborhood V of x, and $F' : V \to \mathcal{L}(X,Y)$ is itself G-differentiable at x, then F is called twice G-differentiable at x. We write $F''(x) \in \mathcal{L}(X, \mathcal{L}(X,Y))$ for the second G-derivative of F at x. It should be clear now how the kth derivative is defined.

In the same way, we define F-differentiability of order k.

It is easy to see that F-differentiablity of F at x implies continuity of F at x. We say that F is k-times continuously F-differentiable if F is k-times F-differentiable and $F^{(k)}$ is continuous.

We collect a couple of facts:

a) The chain rule holds for F-differentiable operators:

$$\begin{aligned} &H(x) = G(F(x)),\ F, G \text{ F-differentiable at } x \text{ and } F(x), \text{ respectively} \\ \implies\ &H \text{ F-differentiable at } x \text{ with } H'(x) = G'(F(x))F'(x). \end{aligned}$$

Moreover, if F is G-differentiable at x and G is F-differentiable at $F(x)$, then H is G-differentiable and the chain rule holds. As a consequence, also the sum rule holds for F- and G-differentials.

b) If F is G-differentiable on a neighborhood of x and F' is continuous at x then F is F-differentiable at x.

c) If $F : X \times Y \to Z$ is F-differentiable at (x,y) then $F(\cdot, y)$ and $F(x, \cdot)$ are F-differentiable at x and y, respectively. These derivatives are called partial derivatives and denoted by $F'_x(x,y)$ and $F'_y(x,y)$, respectively. There holds (since F is F-differentiable)

$$F'(x,y)(h_x, h_y) = F'_x(x,y)h_x + F'_y(x,y)h_y.$$

d) If F is G-differentiable in a neighborhood V of x, then for all $h \in X$ with $\{x + th \,:\, 0 \le t \le 1\} \subset V$, the following holds:

$$\|F(x+h) - F(x)\|_Y \le \sup_{0<t<1} \|F'(x+th)h\|_Y$$

If $t \in [0,1] \mapsto F'(x+th)h \in Y$ is continuous, then

$$F(x+h) - F(x) = \int_0^1 F'(x+th)h \, dx,$$

where the Y-valued integral is defined as a Riemann integral.

We only prove the last assertion: As a corollary of the Hahn-Banach theorem, we obtain that for all $y \in Y$ there exists a $y^* \in Y^*$ with $\|y^*\|_{Y^*} = 1$ and

$$\|y\|_Y = \langle y^*, y \rangle_{Y^*,Y}.$$

Hence,

$$\|F(x+h) - F(x)\|_Y = \max_{\|y^*\|_{Y^*}=1} d(1) \quad \text{with} \quad d_{y^*}(t) = \langle y^*, F(x+th) - F(x) \rangle_{Y^*,Y}.$$

By the chain rule for G-derivatives, we obtain that d is G-differentiable in a neighborhood of $[0,1]$ with

$$d'_{y^*}(t) = \langle y^*, F'(x+th)h \rangle_{Y^*,Y}.$$

G-differentiability of $d : (-\varepsilon, 1+\varepsilon) \to \mathbb{R}$ means that d is differentiable in the classical sense. The mean value theorem yields

$$\langle y^*, F(x+h) - F(x) \rangle_{Y^*,Y} = d(1) = d_{y^*}(1) - d_{y^*}(0) = d'_{y^*}(\tau) \leq \sup_{0<t<1} d'_{y^*}(t)$$

for appropriate $\tau \in (0,1)$. Therefore,

$$\begin{aligned} \|F(x+h) - F(x)\|_Y &= \max_{\|y^*\|_{Y^*}=1} d_{y^*}(1) \leq \sup_{\|y^*\|_{Y^*}=1} \sup_{0<t<1} \langle y^*, F'(x+th)h \rangle_{Y^*,Y} \\ &= \sup_{0<t<1} \sup_{\|y^*\|_{Y^*}=1} \langle y^*, F'(x+th)h \rangle_{Y^*,Y} = \sup_{0<t<1} \|F'(x+th)h\|_Y. \end{aligned}$$

3. Existence of optimal controls

In the introduction we have discussed several examples of optimal control problems. We will now consider the question whether there exists an optimal solution. To this purpose, we need a further ingredient from functional analysis, the concept of weak convergence.

3.1. Weak convergence. In infinite dimensional spaces bounded, closed sets are no longer compact. In order to obtain compactness results, one has to use the concept of weak convergence.

DEFINITION 3.1. *Let X be a Banach space. We say that a sequence $(x_k) \subset X$* converges weakly *to $x \in X$, written*

$$x_k \rightharpoonup x,$$

if

$$\langle x^*, x_k \rangle_{X^*,X} \to \langle x^*, x \rangle_{X^*,X} \quad \text{as } k \to \infty \quad \forall\, x^* \in X^*.$$

It is easy to check that strong convergence $x_k \to x$ implies weak convergence $x_k \rightharpoonup x$. Moreover, one can show:

THEOREM 3.2. *i) Let X be a normed space and let $(x_k) \subset X$ be weakly convergent to $x \in X$. Then (x_k) is bounded.*

ii) Let $C \subset X$ be a closed convex subset of the normed space X. Then C is weakly closed.

DEFINITION 3.3. *A Banach space X is called* reflexive *if the mapping $x \in X \mapsto \langle \cdot, x \rangle_{X^*,X} \in (X^*)^*$ is surjective, i.e., if for any $x^{**} \in (X^*)^*$ there exists $x \in X$ with*

$$\langle x^{**}, x^* \rangle_{(X^*)^*,X^*} = \langle x^*, x \rangle_{X^*,X} \quad \forall\, x^* \in X^*.$$

Remark: Note that for any $x \in X$ the mapping $x^{**} := \langle \cdot, x \rangle_{X^*,X}$ is in $(X^*)^*$ with $\|x^{**}\|_{(X^*)^*} \leq \|x\|_X$, since

$$|\langle x^*, x \rangle_{X^*,X}| \leq \|x^*\|_{X^*} \|x\|_X.$$

One can show that actually $\|x^{**}\|_{(X^*)^*} = \|x\|_X$. □

Remark: L^p is for $1 < p < \infty$ reflexive, since we have the isometric isomorphisms $(L^p)^* = L^q$, $1/p + 1/q = 1$, and thus $((L^p)^*)^* = (L^q)^* = L^p$. Moreover, any Hilbert space is reflexive by the Riesz representation theorem. □

The following result is important.

THEOREM 3.4. (Weak sequential compactness) *Let X be a reflexive Banach space. Then the following holds*

i) *Every bounded sequence $(x_k) \subset X$ contains a weakly convergent subsequence, i.e., there are $(x_{k_i}) \subset (x_k)$ and $x \in X$ with $x_{k_i} \rightharpoonup x$.*
ii) *Every bounded, closed and convex subset $C \subset X$ is weakly sequentially compact, i.e., every sequence $(x_k) \subset C$ contains a weakly convergent subsequence $(x_{k_i}) \subset (x_k)$ with $x_{k_i} \rightharpoonup x$, where $x \in C$.*

For a proof see for example [**3**], [**81**].

THEOREM 3.5. (Lower semicontinuity) *Let X be a Banach space. Then any continuous, convex functional $F : X \to \mathbb{R}$ is weakly lower semicontinuous, i.e.*

$$x_k \rightharpoonup x \quad \Longrightarrow \quad \liminf_{k\to\infty} F(x_k) \geq F(x).$$

Finally, it is valuable to have mappings that map weakly convergent sequences to strongly convergent ones.

DEFINITION 3.6. *A linear operator $A : X \to Y$ between normed spaces is called* compact *if it maps bounded sets to relatively compact sets, i.e.,*

$$M \subset X \text{ bounded} \implies \overline{AM} \subset Y \text{ compact.}$$

Since compact sets are bounded (why?), compact operators are automatically bounded and thus continuous. The connection to weak/strong convergence is as follows.

LEMMA 3.7. *Let $A : X \to Y$ be a compact operator between normed spaces. Then, for all $(x_k) \subset X$, $x_k \rightharpoonup x$, there holds*

$$Ax_k \to Ax.$$

Proof. From $x_k \longrightarrow x$ and $A \in \mathcal{L}(X,Y)$ we see that $Ax_k \longrightarrow Ax$. Since (x_k) is bounded (Theorem 3.2), there exists a bounded set $M \subset X$ with $x \in M$ and $(x_k) \subset M$. Now assume $Ax_k \not\to Ax$. Then there exist $\varepsilon > 0$ and a subsequence $(Ax_k)_K$ with $\|Ax_k - Ax\|_Y \geq \varepsilon$ for all $k \in K$. Since $\overline{AM}$ is compact, the sequence $(Ax_k)_K$ possesses a convergent subsequence $(Ax_k)_{K'} \to y$. The continuity of the norm implies

$$\|y - Ax\|_Y \geq \varepsilon.$$

But since $(Ax_k)_{K'} \longrightarrow Ax$ and $(Ax_k)_{K'} \to y$ we must have $y = Ax$, which is a contradiction. □

3.2. Existence result for a general problem. All linear-quadratic optimization problems in the introduction can be converted to a linear-quadratic optimization problem of the form

$$\begin{aligned} \min_{(y,u)\in Y\times U} \quad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\|Qy - q_d\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 \\ \text{subject to} \quad & Ay + Bu = g, \quad u \in U_{ad},\ y \in Y_{ad} \end{aligned} \tag{3.1}$$

where H, U are Hilbert spaces, Y, Z are Banach spaces and $q_d \in H$, $g \in Z$, Y is reflexive, $A \in \mathcal{L}(Y,Z)$, $B \in \mathcal{L}(U,Z)$, $Q \in \mathcal{L}(Y,H)$ and the the following assumption holds.

ASSUMPTION 1.
(1) $\alpha \geq 0$, $U_{ad} \subset U$ *is convex, closed and in the case* $\alpha = 0$ *bounded.*
(2) $Y_{ad} \subset Y$ *is convex and closed, such that* (3.1) *has a feasible point.*
(3) $A \in \mathcal{L}(Y,Z)$ *has a bounded inverse.*

DEFINITION 3.8. *A state-control pair* $(\bar{y}, \bar{u}) \in Y_{ad} \times U_{ad}$ *is called* optimal *for* (3.1)*, if* $A\bar{y} + B\bar{u} = g$ *and*

$$f(\bar{y}, \bar{u}) \leq f(y,u) \quad \forall\, (y,u) \in Y_{ad} \times U_{ad},\ Ay + Bu = g.$$

We prove first the following existence result for (3.1).

THEOREM 3.9. *Let assumption 1 hold. Then problem* (3.1) *has an optimal solution* $(\bar{y}, \bar{u})$*. If* $\alpha > 0$ *then the solution is unique.*

Proof. Denote the feasible set by

$$W_{ad} := \{(y,u) \in Y \times U \,:\, (y,u) \in Y_{ad} \times U_{ad}, Ay + Bu = g\}.$$

Since $f \geq 0$ and W_{ad} is nonempty, the infimum

$$f^* := \inf_{(y,u)\in W_{ad}} f(y,u)$$

exists and hence we find a minimizing sequence $(y_k, u_k) \subset W_{ad}$ with

$$\lim_{k\to\infty} f(y_k, u_k) = f^*.$$

The sequence (u_k) is bounded, since by assumption either U_{ad} is bounded or $\alpha > 0$. In the latter case the boundedness follows from

$$f(y_k, u_k) \geq \frac{\alpha}{2}\|u_k\|_U^2.$$

Since, $A \in \mathcal{L}(Y,Z)$, $B \in \mathcal{L}(U,Z)$, and $A^{-1} \in \mathcal{L}(Z,Y)$, this implies that also the state sequence (y_k) given by $y_k = A^{-1}(g - Bu_k)$ is bounded. Hence,

$$(y_k, u_k) \subset W_{ad} \cap (\bar{B}_Y(r) \times \bar{B}_U(r)) =: M$$

for $r > 0$ large enough, where $\bar{B}_Y(r), \bar{B}_U(r)$ denote the closed balls of radius r in Y, U. By assumption $Y_{ad} \times U_{ad}$ is closed, convex and thus also W_{ad} is closed and convex. Thus, the set M is bounded, closed and convex and consequently by Theorem 3.4 weakly sequentially compact. Therefore, there exists a weakly convergent subsequence $(y_{k_i}, u_{k_i}) \subset (y_k, u_k)$ and some $(\bar{y}, \bar{u}) \in W_{ad}$ with $(y_{k_i}, u_{k_i}) \rightharpoonup (\bar{y}, \bar{u})$ as $i \to \infty$. Finally, $(y,u) \in Y \times U \to f(y,u)$ is obviously continuous and convex. We conclude by Theorem 3.5 that

$$f^* = \lim_{i \to \infty} f(y_{k_i}, u_{k_i}) \geq f(\bar{y}, \bar{u}) \geq f^*,$$

where the last inequality follows from $(\bar{y}, \bar{u}) \in W_{ad}$. Therefore, $(\bar{y}, \bar{u})$ is the optimal solution of (3.1). If $\alpha > 0$ then $u \mapsto f(A^{-1}(g - Bu), u)$ is strictly convex, which contradicts the existence of more than one minimizer. □

Remark Actually, the reflexivity of Y is not needed. In fact, We can use that $Ay + Bu = g$ implies $y = A^{-1}(g - Bu)$ and thus the problem (3.1) is equivalent to

$$\min_{u \in U} \ \hat{f}(u) \quad \text{s.t.} \quad u \in \hat{U}_{ad}$$

with

$$\hat{f}(u) = f(A^{-1}(g - Bu), u), \quad \hat{U}_{ad} = \left\{ u \in U \ : \ u \in U_{ad}, \ A^{-1}(g - Bu) \in Y_{ad} \right\}.$$

It is easy to see that $\hat{f}$ is continuous and convex and $\hat{U}_{ad}$ is closed and convex. An argumentation as before shows that a minimizing sequence is bounded and thus contains a weakly convergent subsequence convergent to some $\bar{u} \in \hat{U}_{ad}$. Lower semicontinuity implies the optimality of $\bar{u}$. Setting $\bar{y} = A^{-1}(g - B\bar{u})$, we obtain a solution $(\bar{y}, \bar{u})$ of (3.1).

3.3. Existence results for nonlinear problems. The existence result can be extended to nonlinear problems

$$\min_{(y,u) \in Y \times U} \ f(y,u) \text{ subject to } \quad E(y,u) = 0, \quad u \in U_{ad}, \ y \in Y_{ad}, \tag{3.2}$$

$f : Y \times U \to \mathbb{R}$, $E : Y \times U \to Z$ continuous, U and Y reflexive Banach spaces.

Similarly as above, existence can be shown under the following assumptions.

ASSUMPTION 2.

(1) *$U_{ad} \subset U$ is convex, bounded and closed.*
(2) *$Y_{ad} \subset Y$ is convex and closed, such that* (3.2) *has a feasible point.*
(3) *The state equation $E(y,u) = 0$ has a continuous, bounded solution operator $u \in U_{ad} \mapsto y(u) \in Y$.*
(4) *$(y,u) \in Y \times U \mapsto E(y,u) \in Z$ is continuous under weak convergence.*
(5) *f is sequentially weakly lower semicontinuous.*

To show 4., one uses usually compact embeddings $Y \subset\subset \tilde{Y}$ to convert weak convergence in Y to strong convergence in $\tilde{Y}$.

EXAMPLE 3.10. *To show 4. for the semilinear state equation*

$$y \in Y := H^1(\Omega) \mapsto E(y,u) := -\Delta y + y^3 - u \in Y^*,$$

one can proceed as follows. Let $\Omega \subset \mathbb{R}^n$ open and bounded with Lipschitz boundary. Then the imbedding $Y := H^1(\Omega) \subset\subset L^5(\Omega)$ is compact for $n = 2, 3$. Therefore, $y_k \longrightarrow y$ weakly in Y implies $y_k \to y$ strongly in $L^5(\Omega)$ and thus $y_k^3 \to y^3$ strongly in $L^{5/3}(\Omega) = L^{5/2}(\Omega)^* \subset Y^*$ (see below), and thus strongly in Y^*.

To prove $y_k^3 \to y^3$ in $L^{5/3}(\Omega)$, we first observe that $y_k^3, y^3 \in L^{5/3}(\Omega)$ obviously holds. Next, we prove

$$|b^3 - a^3| \le 3(|a|^2 + |b|^2)|b - a|.$$

In fact, $b^3 - a^3 = \phi(1) - \phi(0)$ with $\phi(t) = (a + t(b-a))^3$. Hence,

$$|b^3 - a^3| = \left| \int_0^1 \phi'(t)\, dt \right| \le \int_0^1 |\phi'(t)|\, dt.$$

Now

$$|\phi'(t)| = 3|(a + t(b-a))^2 (b-a)| \le 3 \max(a^2, b^2)|b-a| \le 3(a^2 + b^2)|b-a|.$$

Therefore,

$$\|y_k^3 - y^3\|_{L^{5/3}} \le 3\|(y_k^2 + y^2)|y_k - y|\|_{L^{5/3}} \le \|y_k^2 |y_k - y|\|_{L^{5/3}} + \|y^2 |y_k - y|\|_{L^{5/3}}.$$

We estimate, using the Hölder inequality with $p = 3/2$ and $q = 3$,

$$\|v^2 w\|_{L^{5/3}} = \||v|^{10/3} |w|^{5/3}\|_{L^1}^{3/5} \le \||v|^{10/3}\|_{L^{3/2}}^{3/5} \||w|^{5/3}\|_{L^3}^{3/5} = \||v|^5\|_{L^1}^{2/5} \||w|^5\|_{L^1}^{1/5} = \|v\|_{L^5}^2 \|w\|_{L^5}.$$

This shows

$$\|y_k^3 - y^3\|_{L^{5/3}} \le \|y_k^2 |y_k - y|\|_{L^{5/3}} + \|y^2 |y_k - y|\|_{L^{5/3}}. \le (\|y_k\|_{L^5}^2 + \|y\|_{L^5}^2)\|y_k - y\|_{L^5} \to 2\|y\|_{L^5}^2 \cdot 0 = 0.$$

3.4. Applications.

3.5. Distributed control of elliptic equations. We apply the result first to the distributed optimal control of a steady temperature distribution with boundary temperature zero.

$$\begin{aligned} \min \quad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 \\ \text{subject to} \quad & -\Delta y = \gamma\, u \quad \text{on } \Omega, \\ & y = 0 \quad \text{on } \partial\Omega, \\ & a \le u \le b \quad \text{on } \Omega, \end{aligned} \tag{3.3}$$

where

$$\gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \ge 0, a, b \in L^2(\Omega), \quad a \le b.$$

The form of f and the assumptions on a, b suggest the choice $U = L^2(\Omega)$ and

$$U_{ad} = \{u \in U \,:\, a \le u \le b\}.$$

Then $U_{ad} \subset U$ is bounded, closed and convex.

We know from Theorem 2.44 that the weak formulation of the boundary value problem

$$\begin{aligned} -\Delta y &= \gamma\, u \quad \text{on } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

can be written in the form

$$\text{Find } y \in Y := H_0^1(\Omega): \quad a(y,v) = (\gamma u, v)_{L^2(\Omega)} \quad \forall\, v \in Y.$$

with $a(y,v) = \int_\Omega \nabla y \cdot \nabla v \, dx$, or short

$$Ay + Bu = 0,$$

where $A \in \mathcal{L}(Y, Y^*)$, is the operator representing a, see (2.17), and $B \in \mathcal{L}(U, Y^*)$ is defined through $Bu = -(\gamma u, \cdot)_{L^2(\Omega)}$. By Theorem 2.44, $A \in \mathcal{L}(Y, Y^*)$ has a bounded inverse. Therefore, Assumption 1 is satisfied with the choice $Z = Y^*$. Finally, setting $g = 0$ and $Q = I_{Y,U}$ with the trivial, continuous imbedding $I_{Y,U} : y \in Y \to y \in U$, (3.3) is equivalent to (3.1).

4. Reduced problem, sensitivities and adjoints

We consider again optimal control problems of the form

$$\min_{y \in Y, u \in U} f(y,u) \quad \text{subject to} \quad E(y,u) = 0, \quad (y,u) \in W_{ad}, \tag{4.1}$$

where $f : Y \times U \to \mathbb{R}$ is the objective function, $E : Y \times U \to Z$ is an operator between Banach spaces, and $W_{ad} \subset W := Y \times Z$ is a nonempty closed set.

We assume that f and E are continuously F-differentiable and that the state equation

$$E(y,u) = 0$$

possesses for each ("reasonable") $u \in U$ a unique corresponding solution $y(u) \in Y$. Thus, we have a solution operator $u \in U \mapsto y(u) \in Y$. Furthermore, we assume that $E'_y(y(u),u) \in \mathcal{L}(Y,Z)$ is continuously invertible. Then the implicit function theorem ensures that $y(u)$ is continuously differentiable. An equation for the derivative $y'(u)$ is obtained by differentiating the equation $E(y(u),u) = 0$ with respect to u:

$$E'_y(y(u),u)y'(u) + E'_u(y(u),u) = 0.$$

Inserting $y(u)$ in (4.1), we obtain the reduced problem

$$\min_{u \in U} \hat{f}(u) \stackrel{\text{def}}{=} f(y(u),u) \quad \text{subject to} \quad u \in \hat{U}_{ad} \stackrel{\text{def}}{=} \{u \in U : (y(u),u) \in W_{ad}\}. \tag{4.2}$$

It will be important to investigate the possibilities of computing the derivative of the reduced objective function $\hat{f}$.

Essentially, there are two methods to do this:

- The sensitivity approach,
- The adjoint approach.

4.1. Sensitivity approach. Sensitivities are directional derivatives. For $u \in U$ and a direction $s \in U$, the chain rule yields for the sensitivity of $\hat{f}$:

$$d\hat{f}(u,s) = \langle \hat{f}(u), s\rangle_{U^*,U} = \langle f'_y(y(u),u), y'(u)s\rangle_{Y^*,Y} + \langle f'_u(y(u),u), s\rangle_{U^*,U}.$$

In this expression, the sensitivity $dy(u,s) = y'(u)s$ appears. Differentiating $E(y(u),u) = 0$ in the direction s yields

$$E'_y(y(u),u)y'(u)s + E'_u(y(u),u)s = 0.$$

Hence, the sensitivity $\delta_s y = dy(u,s)$ is given as the solution of the linearized state equation

$$E'_y(y(u),u)\delta_s y = -E'_u(y(u),u)s.$$

Therefore, to compute the directional derivative $d\hat{f}(u,s) = \langle \hat{f}(u), s\rangle_{U^*,U}$ via the sensitivity approach, the following steps are required:

1. Compute the sensitivity $\delta_s y = dy(u,s)$ by solving

$$E'_y(y(u),u)\delta_s y = -E'_u(y(u),u)s. \tag{4.3}$$

2. Compute $d\hat{f}(u,s) = \langle \hat{f}'(u), s\rangle_{U^*,U}$ via

$$d\hat{f}(u,s) = \langle f'_y(y(u),u), \delta_s y\rangle_{Y^*,Y} + \langle f'_u(y(u),u), s\rangle_{U^*,U}.$$

This procedure is expensive if the whole derivative $\hat{f}'(u)$ is required, since this means that for a basis B of U, all the directional derivatives

$$d\hat{f}(u,b), \quad b \in B,$$

have to be computed. Each of them requires the solution of one linearized state equation (4.3) with $s = b$.

This is a effort that grows linearly in the dimension of U.

Actually, computing all sensitivities of $\delta_b y = y'(u)b$, $b \in B$, is equivalent to computing the whole operator $y'(u)$. As we will see now, much less is needed for the derivative of $\hat{f}$.

4.2. Adjoint approach. We now derive a more efficient way of representing the derivative of $\hat{f}$. From

$$\begin{aligned}\langle \hat{f}'(u), s\rangle_{U^*,U} &= \langle f'_y(y(u),u), y'(u)s\rangle_{Y^*,Y} + \langle f'_u(y(u),u), s\rangle_{U^*,U} \\ &= \langle y'(u)^* f'_y(y(u),u), s\rangle_{U^*,U} + \langle f'_u(y(u),u), s\rangle_{U^*,U}\end{aligned}$$

we see that

$$\hat{f}'(u) = y'(u)^* f'_y(y(u),u) + f'_u(y(u),u).$$

Therefore, not the operator $y'(u) \in \mathcal{L}(U,Y)$, but only the vector $y'(u)^* f'_y(y(u),u) \in U^*$ is really required.

Since

$$y'(u)^* f'_y(y(u),u) = -E'_u(y(u),u)^* E'_y(y(u),u)^{-*} f'_y(y(u),u),$$

it follows that

$$y'(u)^* f'_y(y(u),u) = E'_u(y(u),u)^* p(u),$$

where the *adjoint state* $p = p(u) \in Z^*$ solves the

Adjoint Equation:

$$E'_y(y(u),u)^* p = -f'_y(y(u),u). \tag{4.4}$$

We thus have

$$\hat{f}'(u) = E'_u(y(u),u)^* p(u) + f'_u(y(u),u).$$

The derivative $\hat{f}'(u)$ can thus be computed via the adjoint approach as follows:

1. Compute the adjoint state by solving the adjoint equation

$$E'_y(y(u),u)^* p = -f'_y(y(u),u).$$

2 Compute $\hat{f}'(u)$ via

$$\hat{f}'(u) = E'_u(y(u),u)^* p(u) + f'_u(y(u),u).$$

4.3. Application to a linear-quadratic optimal control problem. We consider the linear-quadratic optimal control problem

$$\begin{aligned} \min_{(y,u)\in Y\times U} \quad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\|Qy - q_d\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 \\ \text{subject to} \quad & Ay + Bu = g, \quad u \in U_{ad},\ y \in Y_{ad} \end{aligned} \tag{4.5}$$

where H, U are Hilbert spaces, Y, Z are Banach spaces and $q_d \in H$, $g \in Z$, $A \in \mathcal{L}(Y,Z)$, $B \in \mathcal{L}(U,Z)$, $Q \in \mathcal{L}(Y,H)$ andlet Assumption 1 hold.

$$E(y,u) = Ay + Bu - g, W_{ad} = Y_{ad} \times U_{ad}.$$

By assumption, there exists a continuous affine linear solution operator

$$U \ni u \mapsto y(u) = A^{-1}(g - Bu) \in Y.$$

For the derivatives we have

$$\begin{aligned} \langle f'_y(y,u), s_y\rangle_{Y^*,Y} &= (Qy - q_d, Qs_y)_H, = \langle Q^*(Qy - q_d), s_y\rangle_{Y^*,Y} \\ \langle f'_u(y,u), s_u\rangle_{U^*,U} &= \alpha(u,s_u)_U, \\ E'_y(y,u)s_y &= As_y, \\ E'_u(y,u)s_y &= Bs_u, \end{aligned}$$

Therefore,

$$\begin{aligned} f'_y(y,u) &= (Qy - q_d, Q\cdot)_H \\ f'_u(y,u) &= \alpha(u,\cdot)_U, \\ E'_y(y,u) &= A, \\ E'_u(y,u) &= B. \end{aligned}$$

If we choose the Riesz representations $U^* = U$, $H^* = H$, then

$$f'_y(y,u) = (Qy - q_d, Q\cdot)_H = \langle Qy - q_d, Q\cdot\rangle_{H^*,H} = \langle Q^*(Qy - q_d), \cdot\rangle_{Y^*,Y} = Q^*(Qy - q_d),$$
$$f'_u(y,u) = \alpha(u,\cdot)_U = \alpha u.$$

The reduced objective function is

$$\hat{f}(u) = f(y(u),u) = \frac{1}{2}\|Q(A^{-1}(g - Bu)) - q_d\|_H^2 + \frac{\alpha}{2}\|u\|_U^2.$$

For evaluation of $\hat{f}$, we first solve the state equation

$$Ay + Bu = g$$

to obtain $y = y(u)$ and then we evaluate $f(y,u)$. In the following, let $y = y(u)$.

Sensitivity Approach:

For $s \in U$, we obtain $d\hat{f}(u,s) = \langle \hat{f}'(u), s\rangle_{U^*,U}$ by first solving the linearized state equation

$$A\delta_s y = -Bs$$

for $\delta_s y$ and then setting

$$d\hat{f}(u,s) = ((Qy - q_d), Q\delta_s y)_H + \alpha(u,s)_U.$$

Adjoint Approach:

We obtain $\hat{f}'(u)$ by first solving the adjoint equation

$$A^*p = -((Qy - q_d); Q\cdot)_H \quad (= -Q^*(Qy - q_d) \text{ if } H^* = H)$$

for the adjoint state $p = p(u) \in Z^*$ and then setting

$$\hat{f}'(u) = B^*p + \alpha(u,\cdot)_U \quad (= B^*p + \alpha u \text{ if } U^* = U).$$

Next, let us consider the concrete example of the elliptic control problem

$$\begin{aligned} \min \qquad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\int_\Omega (y(x) - y_d(x))^2\,dx + \frac{\alpha}{2}\int_\Omega u(x)^2\,dx \\ \text{subject to} \quad & -\Delta y = \gamma\, u \quad \text{on } \Omega, \\ & \frac{\partial y}{\partial \nu} = \frac{\beta}{\kappa}\,(y_a - y) \ \text{ on } \partial\Omega, \\ & a \le u \le b \text{ on } \Omega. \end{aligned}$$

The appropriate spaces are

$$U = L^2(\Omega), \quad Y = H^1(\Omega)$$

and we assume

$$a, b \in U, \quad y_d \in L^2(\Omega), \quad \alpha > 0, \quad y_a \in L^2(\partial\Omega), \quad \gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \geq 0.$$

The coefficient γ weights the control and y_a can be interpreted as the surrounding temperature in the case of the heat equation. $\beta > 0$ and $\kappa > 0$ are coefficients.

The weak formulation of the state equation is

$$y \in Y, \quad a(y, v) = (\gamma u, v)_{L^2(\Omega)} + ((\beta/\kappa) y_a, v)_{L^2(\partial\Omega)} \quad \forall\, v \in Y = H^1(\Omega)$$

with

$$a(y, v) = \int_\Omega \nabla y^T \nabla v \, dx + ((\beta/\kappa) y, v)_{L^2(\partial\Omega)}.$$

Now let $Z = Y^*$, $H = L^2(\Omega)$ and

- $A \in \mathcal{L}(Y, Y^*)$ the operator induced by a, i.e., $Ay = a(y, \cdot)$,
- $B \in \mathcal{L}(U, Y^*)$, $Bu = -(\gamma u, \cdot)_{L^2(\Omega)}$,
- $g \in Y^*$, $g = ((\beta/\kappa) y_a, \cdot)_{L^2(\partial\Omega)}$,
- $U_{ad} = \{u \in U \,:\, a \leq u \leq b \text{ on } \Omega\}$,
- $Q \in \mathcal{L}(Y, H)$, $Qy = y$.

Then, we arrive at a linear quadratic problem of the form (4.5).

We compute the adjoints. Note that all spaces are Hilbert spaces and thus reflexive. In particular, we identify the dual of $U = L^2$ with U by working with $\langle \cdot, \cdot \rangle_{U^*,U} = (\cdot, \cdot)_{L^2(\Omega)}$. We do the same with $H = L^2$ We thus have

$$\begin{aligned} A^* &\in \mathcal{L}(Z^*, Y^*) = \mathcal{L}(Y^{**}, Y^*) = \mathcal{L}(Y, Y^*), \\ B^* &\in \mathcal{L}(Z^*, U^*) = \mathcal{L}(Y^{**}, U) = \mathcal{L}(Y, U), \\ Q^* &\in \mathcal{L}(H^*, Y^*) = \mathcal{L}(H, Y^*). \end{aligned}$$

For A^* we obtain

$$\langle A^* v, w \rangle_{Y^*,Y} = \langle v, Aw \rangle_{Z^*,Z} = \langle Aw, v \rangle_{Y^*,Y} = a(w, v) = a(v, w) = \langle Av, w \rangle_{Y^*,Y} \quad \forall\, v, w \in Y.$$

Here, we have used that obviuously a is a symmetric bilinear form. Therefore, $A^* = A$.

For B^* we have

$$\begin{aligned} (B^* v, w)_U &= \langle B^* v, w \rangle_{U^*,U} = \langle v, Bw \rangle_{Z^*,Z} = \langle v, Bw \rangle_{Y,Y^*} = (v, -\gamma w)_{L^2} \\ &= -(\gamma v, w)_U \quad \forall\, v \in Y,\ w \in U. \end{aligned}$$

Hence $B^* v = -\gamma v$. Finally, for Q^* we obtain

$$\langle Q^* v, w \rangle_{Y^*,Y} = \langle v, Qw \rangle_{H^*,H} = (v, w)_{L^2(\Omega)}.$$

Therefore, $Q^* v = (v, \cdot)_{L^2(\Omega)}$.

This means that

$$f'_y(y,u) = (Q^*(Qy - y_d), \cdot)_{L^2(\Omega)} = (y - y_d, \cdot)_{L^2(\Omega)}.$$

Taking all together, the adjoint equation thus reads

$$Ap = -(y - y_d, \cdot)_{L^2(\Omega)},$$

which is the weak form of

$$\begin{aligned} -\Delta p &= -(y - y_d) \quad \text{on } \Omega, \\ \frac{\partial p}{\partial \nu} + \frac{\beta}{\kappa} p &= 0 \ \text{ on } \partial\Omega, \end{aligned}$$

The adjoint gradient representation then is

$$\hat{f}'(u) = B^* p(u) + f'_u(y(u), u) = -\gamma p + \alpha u.$$

4.4. A different view of the adjoint approach. The adjoint gradient representation can also be derived in a different way. Consider (4.1) and define the Lagrange function $L : Y \times U \times Z^* \to \mathbb{R}$,

$$L(y,u,p) = f(y,u) + \langle p, E(y,u)\rangle_{Z^*,Z}.$$

Inserting $y = y(u)$ gives, for arbitrary $p \in Z^*$,

$$\hat{f}(u) = f(y(u),u) = f(y(u),u) + \langle p, E(y(u),u)\rangle_{Z^*,Z} = L(y(u),u,p).$$

Differentaiting this, we obtain

$$\langle \hat{f}'(u), s\rangle_{U^*,U} = \langle L'_y(y(u),u,p), y'(u)s\rangle_{Y^*,Y} + \langle L'_u(y(u),u,p), s\rangle_{U^*,U}. \tag{4.6}$$

Now we choose a special $p = p(u)$, namely such that

$$L'_y(y(u),u,p) = 0. \tag{4.7}$$

This is nothing else but the adjoint equation. In fact,

$$\langle L'_y(y,u,p), d\rangle_{Y^*,Y} = \langle f'_y(y,u), d\rangle_{Y^*,Y} + \langle p, E'_y(y,u)d\rangle_{Z^*,Z} = \langle f'_y(y,u) + E'_y(y,u)^* p, d\rangle_{Y^*,Y}.$$

Therefore,

$$L'_y(y(u),u,p) = f'_y(y(u),u) + E'_y(y(u),u)^* p.$$

Now, choosing $p = p(u)$ according to (4.7), we obtain from (4.6) that

$$\hat{f}'(u) = L'_u(y(u),u,p(u)) = f'_u(y(u),u) + E'_u(y(u),u)^* p(u). \tag{4.8}$$

This is exactly the adjoint gradient representation.

4.5. Second derivatives. We can use the Lagrange function based approach to derive the second derivative of $\hat{f}$.

To this end, assume that f and E are twice continuously differentiable. As already noted, for all $p \in Z^*$ we have the identity

$$\hat{f}(u) = f(y(u), u) = L(y(u), u, p).$$

Differentiating this in the direction $s_1 \in U$ yields (see above)

$$\langle \hat{f}'(u), s_1 \rangle_{U^*,U} = \langle L'_y(y(u), u, p), y'(u)s_1 \rangle_{Y^*,Y} + \langle L'_u(y(u), u, p), s_1 \rangle_{U^*,U}.$$

Differentiating this once again in the direction $s_2 \in U$ gives

$$\begin{aligned}\langle \hat{f}''(u)s_2, s_1 \rangle_{U^*,U} = {} & \langle L'_y(y(u), u, p), y''(u)(s_1, s_2) \rangle_{Y^*,Y} \\ & + \langle L''_{yy}(y(u), u, p)y'(u)s_2, y'(u)s_1 \rangle_{Y^*,Y} \\ & + \langle L''_{yu}(y(u), u, p)s_2, y'(u)s_1 \rangle_{Y^*,Y} \\ & + \langle L''_{uy}(y(u), u, p)y'(u)s_2, s_1 \rangle_{U^*,U} \\ & + \langle L'_{uu}(y(u), u, p)s_2, s_1 \rangle_{U^*,U}.\end{aligned}$$

Now we choose $p = p(u)$, i.e., $L'_y(y(u), u, p(u)) = 0$. Then the term containing $y''(u)$ drops out and we arrive at

$$\begin{aligned}\langle \hat{f}''(u)s_2, s_1 \rangle_{U^*,U} = {} & \langle L''_{yy}(y(u), u, p(u))y'(u)s_2, y'(u)s_1 \rangle_{Y^*,Y} \\ & + \langle L''_{yu}(y(u), u, p(u))s_2, y'(u)s_1 \rangle_{Y^*,Y} \\ & + \langle L''_{uy}(y(u), u, p(u))y'(u)s_2, s_1 \rangle_{U^*,U} \\ & + \langle L'_{uu}(y(u), u, p(u))s_2, s_1 \rangle_{U^*,U}.\end{aligned}$$

This shows

$$\begin{aligned}\hat{f}''(u) &= y'(u)^* L''_{yy}(y(u), u, p(u))y'(u) + y'(u)^* L''_{yu}(y(u), u, p(u)) \\ &\quad + L''_{uy}(y(u), u, p(u))y'(u) + L''_{uu}(y(u), u, p(u)) \\ &= T(u)^* L''_{ww}(y(u), u, p(u))T(u)\end{aligned} \tag{4.9}$$

with

$$T(u) = \begin{pmatrix} y'(u) \\ I_U \end{pmatrix} \in \mathcal{L}(U, Y \times U), \quad L''_{ww} = \begin{pmatrix} L''_{yy} & L''_{yu} \\ L''_{uy} & L''_{uu} \end{pmatrix}.$$

Here $I_U \in \mathcal{L}(U, U)$ is the identity.

Note that $y'(u) = -E'_y(y(u), u)^{-1} E'_u(y(u), u)$ and thus

$$T(u) = \begin{pmatrix} y'(u) \\ I_U \end{pmatrix} = \begin{pmatrix} -E'_y(y(u), u)^{-1} E'_u(y(u), u) \\ I_U \end{pmatrix}. \tag{4.10}$$

Usually, the Hessian representation (4.9) is not used to compute the whole operator $\hat{f}''(u)$. Rather, it is used to compute operator-vector-products $\hat{f}''(u)s$ as follows:

1. Compute the sensitivity
$$\delta_s y = y'(u)s = -E'_y(y(u),u)^{-1}E'_u(y(u),u)s.$$
This requires one linearized state equation solve.

2. Compute
$$\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} L''_{yy}(y(u),u,p(u))\delta_s y + L''_{yu}(y(u),u,p(u))s \\ L''_{uy}(y(u),u,p(u))\delta_s y + L''_{uu}(y(u),u,p(u))s \end{pmatrix}.$$

3. Compute
$$h_3 = y'(u)^* h_1 = -E'_u(y(u),u)^* E'_y(y(u),u)^{-*} h_1.$$
This requires and adjoint equation solve.

4. Set $\hat{f}''(u)s = h_2 + h_3$.

This procedure can be used to apply iterative solvers to the Newton equation
$$\hat{f}''(u^k)s^k = -\hat{f}'(u^k).$$

Example:

For the linear-quadratic optimal control problem (4.5) with $U^* = U$ and $H^* = H$ we have
$$\begin{aligned} L(y,u,p) &= f(y,u) + \langle p, Ay + Bu\rangle_{Z^*,Z}, \\ L'_y(y,u,p) &= Q^*(Qy - q_d) + A^*p, \\ L'_u(y,u,p) &= \alpha u + B^*p, \\ L''_{yy}(y,u,p) &= Q^*Q, \\ L''_{yu}(y,u,p) &= 0, \\ L''_{yu}(u,y,p) &= 0, \\ L''_{uu}(y,u,p) &= \alpha I_U. \end{aligned}$$
From this, all the steps in the above algorithm can be derived easily.

5. Optimality conditions

5.1. Optimality conditions for simply constrained problems. We consider the problem
$$\min_{w\in W} f(w) \quad \text{s.t.} \quad w \in \mathcal{S}, \tag{5.1}$$
where W is a Banach space, $f : W \to \mathbb{R}$ is Gâteaux-differentiable and $\mathcal{S} \subset W$ is nonempty, closed, and convex.

THEOREM 5.1. *Let W be a Banach space and $\mathcal{S} \subset W$ be nonempty and convex. Furthermore, let $f : V \to \mathbb{R}$ be defined on an open neighborhood of $\mathcal{S}$. Let $\bar{w}$ be a local solution of* (5.1) *at which f is Gâteaux-differentiable. Then the following optimality condition holds:*
$$\bar{w} \in \mathcal{S}, \quad \langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W} \geq 0 \quad \forall\, w \in \mathcal{S}. \tag{5.2}$$

If f is convex on $\mathcal{S}$, the condition (5.2) *is necessary and sufficient for global optimality.*

If, in addition, f is strictly convex on $\mathcal{S}$, then there exists at most one solution of (5.1)*, or, equivalently, of* (5.2)*.*

If W is reflexive, $\mathcal{S}$ is closed and convex, and f is convex and continuous with

$$\lim_{w\in\mathcal{S},\|w\|_W\to\infty} f(w) = \infty,$$

then there exists a (global = local) solution of (5.1)*.*

Remark: A condition of the form (5.2) is called variational inequality.

Proof. Let $w \in \mathcal{S}$ be arbitrary. By the convexity of $\mathcal{S}$ we have $w(t) = \bar{w} + t(w - \bar{w}) \in \mathcal{S}$ for all $t \in [0, 1]$. Now the optimality of $\bar{w}$ yields

$$f(\bar{w} + t(w - \bar{w})) - f(\bar{w}) \geq 0 \quad \forall\, t \in [0, 1]$$

and thus

$$\langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W} = \lim_{t\to 0^+} \frac{f(\bar{w} + t(w - \bar{w})) - f(\bar{w})}{t} \geq 0.$$

Since $w \in \mathcal{S}$ was arbitrary, the proof is complete.

Now let f be convex. Then

$$f(w) - f(\bar{w}) \geq \langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W} \quad \forall\, w \in \mathcal{S}. \tag{5.3}$$

In fact, for all $t \in (0, 1]$,

$$f(\bar{w} + t(w - \bar{w})) \leq (1 - t)f(\bar{w}) + tf(w).$$

Hence,

$$f(w) - f(\bar{w}) \qquad = \frac{f(\bar{w} + t(w - \bar{w})) - f(\bar{w})}{t} \overset{t\to 0^+}{\longrightarrow} \langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W}.$$

Now from (5.2) and (5.3) it follows that

$$f(w) - f(\bar{w}) \geq \langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W} \geq 0 \quad \forall\, w \in \mathcal{S}.$$

Thus, $\bar{w}$ is optimal.

If f is strictly convex and $\bar{w}_1, \bar{w}_2$ are two global solutions, the point $(\bar{w}_1 + \bar{w}_2)/2 \in \mathcal{S}$ would be a better solution, unless $\bar{w}_1 = \bar{w}_2$.

Now let the assumptions of the last assertion hold and let $(w_k) \in \mathcal{S}$ be a minimizing sequence. Then (w_k) is bounded (otherwise $f(w_k) \to \infty$) and thus (w_k) contains a weakly convergent subsequence $(w_k)_K \longrightarrow \bar{w}$. Since $\mathcal{S}$ is convex and closed, it is weakly closed and thus $\bar{w} \in \mathcal{S}$. From the continuity and convexity of f we conclude that f is weakly sequentially lower semicontinuous and thus

$$f(\bar{w}) \leq \lim_{K\ni k\to\infty} f(w_k) = \inf_{w\in\mathcal{S}} f(w).$$

Thus, $\bar{w}$ solves the minimization problem. □

In the case of a closed convex set $\mathcal{S}$ in a *Hilbert space* W, we can rewrite the variational inequality in the form

$$\bar{w} - P(\bar{w} - \gamma \nabla f(w)) = 0$$

where $\gamma > 0$ is a fixed parameter and $\nabla f(w) \in W$ is the Riesz representation of $f'(w) \in W^*$.

To prove this, we need some knowledge about the projection onto closed convex sets.

LEMMA 5.2. *Let $\mathcal{S} \subset W$ be a nonempty closed convex subset of the Hilbert space W and denote by $P : W \to \mathcal{S}$ the projection onto $\mathcal{S}$, i.e.,*

$$P(w) \in \mathcal{S}, \quad \|P(w) - w\|_W = \min_{v \in \mathcal{S}} \|v - w\|_W \qquad \forall\, w \in W.$$

Then:

a) *P is well-defined.*

b) *For all $w, z \in W$ there holds:*

$$\begin{aligned} &z = P(w) \quad \Longleftrightarrow \\ &z \in \mathcal{S}, \quad (w - z, v - z)_W \le 0 \quad \forall\, v \in \mathcal{S}. \end{aligned}$$

c) *P is nonexpansive, i.e.,*

$$\|P(v) - P(w)\|_W \le \|v - w\|_W \quad \forall\, v, w \in W.$$

d) *P is monotone, i.e.,*

$$(P(v) - P(w), v - w)_W \ge 0 \quad \forall\, v, w \in W.$$

Furthermore, equality holds if and only if $P(v) = P(w)$.

e) *For all $w \in \mathcal{S}$ and $d \in W$, the function*

$$\phi(t) \stackrel{\text{def}}{=} \frac{1}{t} \|P(w + td) - w\|_W, \quad t > 0,$$

is nonincreasing.

Proof. a):

The function $W \ni w \mapsto \|w\|_W^2$ is strictly convex: For all $w_1, w_2 \in W$, $w_1 \neq w_2$, and all $t \in (0, 1)$;

$$\|w_1 + t(w_2 - w_1)\|_W^2 = \|w_1\|_W^2 + 2t(w_1, w_2 - w_1)_W + t^2 \|w_2 - w_1\|_W^2 =: p(t).$$

The function on the right is a strictly convex parabola. Hence,

$$\|w_1 + t(w_2 - w_1)\|_W^2 = p(t) < (1 - t)p(0) + tp(1) = (1 - t)\|w_1\|_2^2 + t\|w_2\|_2^2.$$

Therefore, for all $w \in W$, the function

$$f(v) = \frac{1}{2} \|v - w\|_W^2$$

is strictly convex. Furthermore, it tends to ∞ as $\|v\|_W \to \infty$. Hence, by Theorem 5.1, the problem

$$\min_{v\in\mathcal{S}} f(v)$$

possesses a unique solution $\bar{v}$, and thus $P(w) = \bar{v}$ is uniquely defined.

b):

The function f defined above is obviously F-differentiable with

$$\langle f'(v), s\rangle_{W^*,W} = (v - w, s)_W \quad \forall\, s \in W.$$

Since $P(w) = \bar{v}$ minimizes f on $\mathcal{S}$, we have by Theorem 5.1 that $z = P(w)$ if and only if $z \in \mathcal{S}$ and

$$z \in \mathcal{S}, \quad \langle f'(z), v - z\rangle_{W^*,W} = (z - w, v - z)_W \geq 0 \quad \forall\, v \in \mathcal{S}.$$

c):

We use b):

$$(v - P(v), P(w) - P(v))_W \leq 0,$$
$$(w - P(w), P(v) - P(w))_W \leq 0.$$

Adding these two inequalities gives

$$(w - v + P(v) - P(w), P(v) - P(w)) = (w - v, P(v) - P(w))_W + \|P(v) - P(w)\|_W^2 \leq 0.$$

Hence, by the Cauchy-Schwarz inequality

$$\|P(v) - P(w)\|_W^2 \leq (v - w, P(v) - P(w))_W \leq \|v - w\|_W \|P(v) - P(w)\|_W. \tag{5.4}$$

d):

The assertion follows immediately from the first inequality in (5.4).

e):

We follow [CM87]. Let $t > s > 0$. If $\|P(w + td) - w\|_W \leq \|P(w + sd) - w\|_W$ then obviously $\phi(s) > \phi(t)$.

Now let $\|P(w + td) - w\|_W > \|P(w + sd) - w\|_W$.

Using the Cauchy-Schwarz inequality, for any $u, v \in W$ we have

$$\begin{aligned}
&\|v\|_W(u, u - v)_W - \|u\|_W(v, u - v)_W \\
&\quad = \|v\|_W\|u\|_W^2 - \|v\|_W(u, v)_W - \|u\|_W(v, u)_W + \|u\|_W\|v\|_W^2 \\
&\quad \geq \|v\|_W\|u\|_W^2 - \|v\|_W\|u\|_W\|v\|_W - \|u\|_W\|v\|_W\|u\|_W + \|u\|_W\|v\|_W^2 = 0.
\end{aligned}$$

Now, set $u := P(w+td) - w$, $v := P(w+sd) - w$, and $w_\tau = w + \tau d$. Then

$$\begin{aligned}(u, u-v)_W - (td, P(w_t) - P(w_s))_W &= (P(w_t) - w - td, P(w_t) - P(w_s))_W \\ &= (P(w_t) - w_t, P(w_t) - P(w_s))_W \leq 0, \\ (v, u-v)_W - (sd, P(w_t) - P(w_s))_W &= (P(w_s) - w - sd, P(w_t) - P(w_s))_W \\ &= (P(w_s) - w_s, P(w_t) - P(w_s))_W \geq 0.\end{aligned}$$

Thus,

$$\begin{aligned}0 &\leq \|v\|_W (u, u-v)_W - \|u\|_W (v, u-v)_W \\ &\leq \|v\|_W (td, P(w_t) - P(w_s))_W - \|u\|_W (sd, P(w_t) - P(w_s))_W \\ &= (t\|v\|_W - s\|u\|_W)(d, P(w_t) - P(w_s))_W.\end{aligned}$$

Now, due to the monotonicity of P,

$$(d, P(w_t) - P(w_s))_W = \frac{1}{t-s}(w_t - w_s, P(w_t) - P(w_s))_W > 0,$$

since $P(w_t) \neq P(w_s)$. Therefore,

$$0 \leq t\|v\|_W - s\|u\|_W = ts(\phi(s) - \phi(t)).$$

□

LEMMA 5.3. *Let W be a Hilbert space, $\mathcal{S} \subset W$ be nonempty, closed, and convex. Furthermore, let P denote the projection onto $\mathcal{S}$. Then, for all $y \in W$ and all $\gamma > 0$, the following conditions are equivalent:*

$$w \in \mathcal{S}, \quad (y, v-w)_W \geq 0 \quad \forall\, v \in \mathcal{S}. \tag{5.5}$$

$$w - P(w - \gamma y) = 0. \tag{5.6}$$

Proof. Let (5.5) hold. Then with $w_\gamma = w - \gamma y$ we have

$$(w_\gamma - w, v - w)_W = -\gamma (y, v-w)_W \leq 0 \quad \forall\, v \in \mathcal{S}.$$

By Lemma 5.2 b), this implies $w = P(w_\gamma)$ as asserted in (5.6).

Conversely, let (5.6) hold. Then with the same notation as above we obtain $w = P(w_\gamma) \in \mathcal{S}$. Furthermore, Lemma 5.2 b) yields

$$(y, v-w)_W = -\frac{1}{\gamma}(w_\gamma - w, v - w) \geq 0 \quad \forall\, v \in \mathcal{S}.$$

□

COROLLARY 5.4. *Let W be a Hilbert space and $\mathcal{S} \subset W$ be nonempty, closed, and convex. Furthermore, let $f : V \to \mathbb{R}$ be defined on an open neighborhood of $\mathcal{S}$. Let $\bar{w}$ be a local solution of* (5.1) *at which f is Gâteaux-differentiable. Then the following optimality condition holds:*

$$\bar{w} = P(\bar{w} - \gamma \nabla f(\bar{w})) \tag{5.7}$$

Here, $\gamma > 0$ is arbitrary but fixed and $\nabla f(w) \in W$ denotes the Riesz-representation of $f'(w) \in W^*$.

5.2. Optimality conditions for control-constrained problems. We consider a general possibly nonlinear problem of the form

$$\min_{(y,u)\in Y\times U} \quad f(y,u) \quad \text{subject to} \quad E(y,u)=0, \quad u\in U_{ad}. \tag{5.8}$$

We make the

ASSUMPTION 3.
(1) *$U_{ad}\subset U$ is nonempty and convex.*
(2) *$f: Y\times U\to\mathbb{R}$and $E: Y\times U\to Z$ are continuously Fréchet differentiable and U, Y, Z are Banach spaces.*
(3) *For all $u\in V$ in a neighborhood $V\subset U$ of U_{ad}, the state equation $E(y,u)=0$ has a unique solution $y=y(u)\in Y$.*
(4) *$E'_y(y(u),u)\in\mathcal{L}(Y,Z)$ has a bounded inverse for all $u\in U_{ad}$.*

Obviously, the general linear-quadratic optimization problem

$$\begin{aligned}\min_{(y,u)\in Y\times U} \quad & f(y,u)\stackrel{\text{def}}{=}\frac{1}{2}\|Qy-q_d\|_H^2+\frac{\alpha}{2}\|u\|_U^2\\ \text{subject to} \quad & Ay+Bu=g, \quad u\in U_{ad},\end{aligned} \tag{5.9}$$

is a special case of (5.8), where H, U are Hilbert spaces, Y, Z are Banach spaces and $q_d\in H$, $g\in Z$, $A\in\mathcal{L}(Y,Z)$, $B\in\mathcal{L}(U,Z)$, $Q\in\mathcal{L}(Y,H)$. Moreover, Assumption 1 ensures Assumption 3, since $E'_y(y,u)=A$.

5.3. A general first order optimality condition. Now consider problem (5.8) and let Assumption 3 hold. Then we can formulate the reduced problem

$$\min_{u\in U}\hat{f}(u) \quad \text{s.t.} \quad u\in U_{ad} \tag{5.10}$$

with the reduced objective functional

$$\hat{f}(u):=f(y(u),u),$$

where $V\ni u\mapsto y(u)\in Y$ is the solution operator of the state equation. We have the following general result.

THEOREM 5.5. *Let Assumption 3 hold. If $\bar{u}$ is a local solution of the reduced problem* (5.10) *then $\bar{u}\in U_{ad}$ and $\bar{u}$ satisfies the variational inequality*

$$\langle\hat{f}'(\bar{u}),u-\bar{u}\rangle_{U^*,U}\geq 0 \quad \forall\, u\in U_{ad}. \tag{5.11}$$

Proof. We can directly apply Theorem 5.1. □

Depending on the structure of U_{ad} the variational inequality (5.11) can be expressed in a more convenient form. We show this for the case of box constraints.

LEMMA 5.6. *Let $U = L^2(\Omega)$, $a, b \in L^2(\Omega)$, $a \le b$, and U_{ad} be given by*

$$U_{ad} = \left\{ u \in L^2(\Omega) \,:\, a \le u \le b \right\}$$

We work with $U^* = U$ write $\nabla \hat{f}(u)$ for the derivative to emphasize that this is the Riesz representation. Then the following conditions are equivalent:

i) $\bar{u} \in U_{ad}$, $(\nabla \hat{f}(\bar{u}), u - \bar{u})_U \ge 0 \quad \forall\, u \in U_{ad}$.

ii) $\bar{u} \in U_{ad}$, $\nabla \hat{f}(\bar{u})(x) \begin{cases} = 0, & \text{if } a(x) < \bar{u}(x) < b(x), \\ \ge 0, & \text{if } a(x) = \bar{u}(x) < b(x), \\ \le 0, & \text{if } a(x) < \bar{u}(x) = b(x), \end{cases}$ *for a.a.* $x \in \Omega$.

iii) *There are* $\bar{z}_a, \bar{z}_b \in U^* = L^2(\Omega)$ *with*

$$\begin{aligned} &\nabla \hat{f}(\bar{u}) + \bar{z}_b - \bar{z}_a = 0, \\ &\bar{u} \ge a, \quad \bar{z}_a \ge 0, \quad \bar{z}_a\,(\bar{u} - a) = 0, \\ &\bar{u} \le b, \quad \bar{z}_b \ge 0, \quad \bar{z}_b\,(b - \bar{u}) = 0. \end{aligned}$$

iv) *For any $\gamma > 0$: $\bar{u} = P_{U_{ad}}(\bar{u} - \gamma \nabla \hat{f}(\bar{u}))$, with $P_{U_{ad}}(u) = \min(\max(a, u), b)$.*

Proof. ii) $\Longrightarrow$ i): If $\nabla \hat{f}(\bar{u})$ satisfies ii) then it is obvious that $\nabla \hat{f}(\bar{u})\,(u - \bar{u}) \ge 0$ a.e. for all $u \in U_{ad}$ and thus

$$(\nabla \hat{f}(\bar{u}), u - \bar{u})_U = \int_\Omega \nabla \hat{f}(\bar{u})(u - \bar{u})\, dx \ge 0 \quad \forall\, u \in U_{ad}.$$

i) $\Longrightarrow$ ii): Clearly, ii) is the same as

$$\nabla \hat{f}(\bar{u})(x) \begin{cases} \ge 0 & \text{a.e. on } I_a = \{x \,:\, a(x) \le \bar{u}(x) < b(x)\} \\ \le 0 & \text{a.e. on } I_b = \{x \,:\, a(x) < \bar{u}(x) \le b(x)\} \end{cases}$$

Assume this is not true. Then, without loss of generality, there exists a set $M \subset I_a$ of positive measure with $\nabla \hat{f}(\bar{u})(x) < 0$ on M. Now choose $u = \bar{u} + 1_M\,(b - \bar{u})$. Then $u \in U_{ad}$, $u - \bar{u} > 0$ on M and $u - \bar{u} = 0$ elsewhere. Hence, we get the contradiction

$$(\nabla \hat{f}(\bar{u}), u - \bar{u})_U = \int_M \underbrace{\nabla \hat{f}(\bar{u})}_{<0}\, \underbrace{(b - \bar{u})}_{>0}\, dx < 0.$$

ii) $\Longrightarrow$ iii): Let $\bar{z}_a = \max(\nabla \hat{f}(\bar{u}), 0)$, $\bar{z}_b = \max(-\nabla \hat{f}(\bar{u}), 0)$. Then $a \le \bar{u} \le b$ and $\bar{z}_a, \bar{z}_b \ge 0$ hold trivially. Furthermore,

$$\begin{aligned} \bar{u}(x) > a(x) &\Longrightarrow \nabla \hat{f}(\bar{u})(x) \le 0 \Longrightarrow \bar{z}_a(x) = 0, \\ \bar{u}(x) < b(x) &\Longrightarrow \nabla \hat{f}(\bar{u})(x) \ge 0 \Longrightarrow \bar{z}_b(x) = 0. \end{aligned}$$

iii) $\Longrightarrow$ ii):

$$a(x) < \bar{u}(x) < b(x) \Longrightarrow \bar{z}_a = \bar{z}_b = 0 \Longrightarrow \nabla \hat{f}(\bar{u}) = 0,$$
$$a(x) = \bar{u}(x) < b(x) \Longrightarrow \bar{z}_b = 0 \Longrightarrow \nabla \hat{f}(\bar{u}) = \bar{z}_a \geq 0,$$
$$a(x) < \bar{u}(x) = b(x) \Longrightarrow \bar{z}_a = 0 \Longrightarrow \nabla \hat{f}(\bar{u}) = -\bar{z}_b \leq 0.$$

ii) $\Longleftrightarrow$ iv): This is easily verified.

Alternatively, we can use Lemma 5.3 to prove the equivalence of i) and iv). □

5.4. Necessary first order optimality conditions. Next, we use the adjoint representation of the derivative

$$\hat{f}'(u) = E'_u(y(u), u)^* p(u) + f'_u(y(u), u), \tag{5.12}$$

where the adjoint state $p(u) \in Z^*$ solves the adjoint equation

$$E'_y(y(u), u)^* p = -f'_y(y(u), u). \tag{5.13}$$

For compact notation, we recall the definition of the Lagrange function associated with (5.8)

$$L : Y \times U \times Z^* \to \mathbb{R}, \quad L(y, u, p) = f(y, u) + \langle p, E(y, u) \rangle_{Z^*, Z}.$$

The representation (5.12) of $\hat{f}'(\bar{u})$ yields the following corollary of Theorem 5.5.

COROLLARY 5.7. *Let $(\bar{y}, \bar{u})$ an optimal solution of the problem* (5.8) *and let Assumption 3 hold. Then there exists an adjoint state (or Lagrange multiplier) $\bar{p} \in Z^*$ such that the following optimality conditions hold*

$$E(\bar{y}, \bar{u}) = 0, \tag{5.14}$$
$$E'_y(\bar{y}, \bar{u})^* \bar{p} = -f'_y(\bar{y}, \bar{u}), \tag{5.15}$$
$$\bar{u} \in U_{ad}, \quad \langle f'_u(\bar{y}, \bar{u}) + E'_u(\bar{y}, \bar{u})^* \bar{p}, u - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall\, u \in U_{ad}, \tag{5.16}$$
(5.17)

Using the Lagrange function we can write (5.14)–(5.16) *in the compact form*

$$L'_p(\bar{y}, \bar{u}, \bar{p}) = E(\bar{y}, \bar{u}) = 0, \tag{5.14}$$
$$L'_y(\bar{y}, \bar{u}, \bar{p}) = 0, \tag{5.15}$$
$$\bar{u} \in U_{ad}, \quad \langle L'_u(\bar{y}, \bar{u}, \bar{p}), u - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall\, u \in U_{ad}. \tag{5.16}$$

Proof. We have only to combine (5.11), (5.13), and (5.12). □

To avoid dual operators, one can also use the equivalent form

$$E(\bar{y}, \bar{u}) = 0, \tag{5.18}$$
$$\langle L'_y(\bar{y}, \bar{u}, \bar{p}), v \rangle_{Y^*, Y} = 0 \quad \forall\, v \in Y \tag{5.19}$$
$$\bar{u} \in U_{ad}, \quad \langle L'_u(\bar{y}, \bar{u}, \bar{p}), u - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall\, u \in U_{ad}. \tag{5.20}$$

5.5. Applications.

5.5.1. *General linear-quadratic problem.* We apply the result to the linear-quadratic problem

$$\begin{aligned} &\min_{(y,u)\in Y\times U} && f(y,u) := \frac{1}{2}\|Qy - q_d\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 \\ &\text{subject to} && Ay + Bu = g, \quad u \in U_{ad} \end{aligned} \tag{5.21}$$

under Assumption 1. Then

$$E(y,u) = Ay + Bu - g, \quad E'_y(y,u) = A, \quad E'_u(y,u) = B$$

and Corollary 5.7 is applicable. We only have to compute L'_y and L'_u for the Lagrange function

$$\begin{aligned} L(y,u,p) &= f(y,u) + \langle p, Ay + Bu - g\rangle_{Z^*,Z} \\ &= \frac{1}{2}(Qy - q_d, Qy - q_d)_H + \frac{\alpha}{2}(u,u)_U + \langle p, Ay + Bu - q\rangle_{Z^*,Z}. \end{aligned}$$

We have with the identification $H^* = H$ and $U^* = U$

$$\begin{aligned} \langle L'_y(\bar y, \bar u, \bar p), v\rangle_{Y^*,Y} &= (Q\bar y - q_d, Qv)_H + \langle \bar p, Av\rangle_{Z^*,Z} \\ &= \langle Q^*(Q\bar y - q_d) + A^*\bar p, v\rangle_{Y^*,Y} \quad \forall\, v \in Y \end{aligned} \tag{5.22}$$

and

$$\begin{aligned} (L'_u(\bar y, \bar u, \bar p), w)_U &= \alpha(\bar u, w)_U + \langle \bar p, Bw\rangle_{Z^*,Z} \\ &= (\alpha\bar u + B^*\bar p, w)_U \quad \forall\, w \in U. \end{aligned} \tag{5.23}$$

Thus (5.14)–(5.16) take the form

$$A\bar y + B\bar u = g, \tag{5.24}$$

$$A^*\bar p = -Q^*(Q\bar y - q_d), \tag{5.25}$$

$$\bar u \in U_{ad}, \quad (\alpha\bar u + B^*\bar p, u - \bar u)_U \geq 0 \quad \forall\, u \in U_{ad}. \tag{5.26}$$

5.5.2. *Distributed control of elliptic equations.* We consider next the distributed optimal control of a steady temperature distribution with boundary temperature zero

$$\begin{aligned} &\min && f(y,u) := \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 \\ &\text{subject to} && -\Delta y = \gamma\, u \quad \text{on } \Omega, \\ &&& y = 0 \quad \text{on } \partial\Omega, \\ &&& a \leq u \leq b \quad \text{on } \Omega, \end{aligned} \tag{5.27}$$

where

$$\gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \geq 0, a, b \in L^2(\Omega), \quad a \leq b.$$

We have already observed that (5.27) has the form (5.21) with

$$U = H = L^2(\Omega), \quad Y = H_0^1(\Omega), \quad Z = Y^*, \quad g = 0, \quad Q = I_{Y,H},$$

and

$$A \in \mathcal{L}(Y, Y^*), \quad \langle Ay, v\rangle_{Y^*,Y} = a(y,v) = \int_\Omega \nabla y \cdot \nabla v \, dx,$$
$$B \in \mathcal{L}(U, Y^*), \quad \langle Bu, v\rangle_{Y^*,Y} = -(\gamma u, v)_{L^2(\Omega)}.$$

As a Hilbert space, Y is reflexive and $Z^* = Y^{**}$ can be identified with Y through

$$\langle p, y^*\rangle_{Y^{**},Y^*} = \langle y^*, p\rangle_{Y^*,Y} \quad \forall\, y^* \in Y^*,\ p \in Y = Y^{**}.$$

This yields

$$\langle p, Ay\rangle_{Z^*,Z} = \langle Ay, p\rangle_{Y^*,Y} = a(y,p) = a(p,y).$$

Let $(\bar y, \bar u) \in Y \times U$ be an optimal solution. Then by Corollary 5.7 and (5.22), (5.23) the optimality system in the form (5.18)–(5.20) reads

$$a(\bar y, v) - (\gamma \bar u, v)_{L^2(\Omega)} = 0 \quad \forall\, v \in Y, \tag{5.28}$$

$$(\bar y - y_d, v)_{L^2\Omega} + a(\bar p, v) = 0 \quad \forall\, v \in Y, \tag{5.29}$$

$$a \le \bar u \le b, \quad (\alpha \bar u - \gamma \bar p, u - \bar u)^2_L(\Omega) \ge 0, \quad \forall\, u \in U,\ a \le u \le b. \tag{5.30}$$

Now the adjoint equation (5.28) is just the weak formulation of

$$-\Delta \bar p = -(\bar y - y_d), \quad \bar p|_{\partial\Omega} = 0.$$

Applying Lemma 5.6 we can summarize

THEOREM 5.8. *If $(\bar y, \bar u)$ is an optimal solution of* (5.27) *then there exist $\bar p \in H^1_0(\Omega)$, $\bar z_a, \bar z_b \in L^2(\Omega)$ such that the following optimality conditions hold in the weak sense.*

$$\begin{aligned}
&-\Delta \bar y = \gamma \bar u, \quad \bar y|_{\partial\Omega} = 0,\\
&-\Delta \bar p = -(\bar y - y_d), \quad \bar p|_{\partial\Omega} = 0,\\
&\alpha \bar u - \gamma \bar p + \bar z_b - \bar z_a = 0,\\
&\bar u \ge a, \quad \bar z_a \ge 0, \quad \bar z_a\,(\bar u - a) = 0,\\
&\bar u \le b, \quad \bar z_b \ge 0, \quad \bar z_b\,(b - \bar u) = 0.
\end{aligned}$$

5.5.3. *Distributed control of semilinear elliptic equations.* We consider next the distributed optimal control of a semilinear elliptic PDE:

$$\begin{aligned}
\min \qquad & f(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}\\
\text{subject to} \quad & -\Delta y + y^3 = \gamma\, u \quad \text{on } \Omega,\\
& y = 0 \quad \text{on } \partial\Omega,\\
& a \le u \le b \quad \text{on } \Omega,
\end{aligned} \tag{5.31}$$

where

$$\gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \ge 0, a, b \in L^\infty(\Omega), \quad a \le b.$$

Let $n \leq 3$. By the theory of monotone operators one can show that there exists a continuous solution operator of the state equation

$$u \in U := L^2(\Omega) \to y \in Y := H_0^1(\Omega).$$

Let $A : H_0^1(\Omega) \to H_0^1(\Omega)^*$ be the operator associated with the bilinear form $a(y, v) = \int_\Omega \nabla y \cdot \nabla v \, dx$ for the Laplace operator $-\Delta y$ and let

$$N : y \to y^3.$$

Then the weak formulation of the state equation can be written in the form

$$E(y, u) := Ay + N(y) - \gamma u = 0.$$

By the Sobolev imbedding theorem [] one has for $n \leq 3$ the continuous imbedding

$$H_0^1(\Omega) \subset L^6(\Omega).$$

Moreover, the mapping $N : y \in L^6(\Omega) \to y^3 \in L^2(\Omega)$ is continuously Fréchet differentiable with

$$N'(y)v = 2y^2 v.$$

At this point, it is convenient to prove first the following extension of H$\ddot{\text{o}}$lder's inequality:

LEMMA 5.9. *Let $\omega \subset \mathbb{R}^n$ be measurable. Then, for all $p_i, p \in [1, \infty]$ with $1/p_1 + \cdots + 1/p_k = 1/p$ and all $u_i \in L^{p_i}(\Omega)$, there holds $u_1 \cdots u_k \in L^p(\Omega)$ and*

$$\|u_1 \cdots u_k\|_{L^p} \leq \|u_1\|_{L^{p_1}} \cdots \|u_k\|_{L^{p_k}}.$$

Proof. We use induction. For $k = 1$ the assertion is trivial and for $k = 2$ we obtain it from Hölder's inequality: From $1/p_1 + 1/p_2 = 1/p$ we see that $1/q_1 + 1/q_2 = 1$ holds for $q_i = p_i/p$ and thus

$$\begin{aligned} \|u_1 u_2\|_{L^p} &= \||u_1|^p |u_2|^p\|_{L^1}^{1/p} \leq \||u_1|^p\|_{L^{q_1}}^{1/p} \||u_2|^p\|_{L^{q_2}}^{1/p} \\ &= \||u_1|^{pq_1}\|_{L^1}^{1/p_1} \||u_2|^{pq_2}\|_{L^1}^{1/p_2} = \|u_1\|_{L^{p_1}} \|u_2\|_{L^{p_2}}. \end{aligned}$$

As a consequence, $u_1 u_2 \in L^p(\Omega)$ and the assertion is shown for $k = 2$.

For $1, \ldots, k-1 \to k$, let $q \in [1, \infty]$ be such that

$$\frac{1}{q} + \frac{1}{p_k} = \frac{1}{p}.$$

Then we have $1/p_1 + \cdots + 1/p_{k-1} = 1/q$ and thus (using the assertion for $k-1$), we obtain $u_1 \cdots u_{k-1} \in L^q(\Omega)$ and

$$\|u_1 \cdots u_{k-1}\|_{L^q} \leq \|u_1\|_{L^{p_1}} \cdots \|u_{k-1}\|_{L^{p_{k-1}}}.$$

Therefore, using the assertion for $k = 2$,

$$\|u_1 \cdots u_k\|_{L^p} \leq \|u_1 \cdots u_{k-1}\|_{L^q} \|u_k\|_{L^{p_k}} = \|u_1\|_{L^{p_1}} \cdots \|u_k\|_{L^{p_k}}.$$

□

We now return to the proof of the F-differentiabilty of N: We just have to apply the Lemma with $p_1 = p_2 = p_3 = 6$ and $p = 2$:

$$\begin{aligned}\|(y+h)^3 - y^3 - 3y^2h\|_{L^2} &= \|3yh^2 + h^3\|_{L^2} = 3\|y\|_{L^6}\|h\|_{L^6}^2 + \|h\|_{L^6}^3 \\ &= O(\|h\|_{L^6}^2) = o(\|h\|_{L^6}).\end{aligned}$$

This shows the F-differentiability of N with derivative N'. Furthermore, to prove the continuity of N', we estimate

$$\begin{aligned}\|(N'(y+h) - N'(y))v\|_{L^2} &= 3\|((y+h)^2 - y^2)v\|_{L^2} = 3\|(y+h)hv\|_{L^2} \\ &= 3\|y+h\|_{L^6}\|h\|_{L^6}\|v\|_{L^6}.\end{aligned}$$

Hence,

$$\|N'(y+h) - N'(y)\|_{L^2,L^6} \le 3\|y+h\|_{L^6}\|h\|_{L^6} \overset{\|h\|_{L^6}\to 0}{\longrightarrow} 0.$$

Therefore, $E : Y \times U \to Y^* =: Z$ is continuously Fréchet differentiable with

$$E_y'(y,u)v = Av + 3y^2v, \quad E_u'(y,u)w = -\gamma w.$$

Finally, $E_y'(y,u) \in \mathcal{L}(Y,Z)$ has a bounded inverse, since for any $y \in Y$ the equation

$$Av + 3y^2v = f$$

has a bounded solution operator $f \in Z \to v \in Y$. Hence, Assumption (OPT) is satisfied. The optimality conditions are now very similar to the linear-quadratic problem (5.27) with the only difference that now $E_y'(y,u)v = Av + 2y^2v$: Let $(\bar y, \bar u) \in Y \times U$ be an optimal solution. Then by Corollary 5.7 the optimality system in the form (5.18)–(5.20) reads

$$A\bar y + \bar y^3 - \gamma \bar u = 0, \tag{5.32}$$

$$(\bar y - y_d, v)_L^2\Omega + a(\bar p, v) + (3\bar y^2\bar p, v)_L^2(\Omega) = 0 \quad \forall\, v \in Y, \tag{5.33}$$

$$a \le \bar u \le b, \quad (\alpha\bar u - \gamma\bar p, u - \bar u)_L^2(\Omega) \ge 0, \quad \forall\, a \le u \le b. \tag{5.34}$$

Now the adjoint equation (5.33) is just the weak formulation of

$$-\Delta\bar p + 3\bar y^2\bar p = -(\bar y - y_d), \quad \bar p|_{\partial\Omega} = 0.$$

Applying Lemma 5.6 we can summarize

THEOREM 5.10. *If $(\bar y, \bar u)$ is an optimal solution of* (5.31) *then there exist $\bar p \in H_0^1(\Omega)$, $\bar z_a, \bar z_b \in L^2(\Omega)$ such that the following optimality system holds in the weak sense.*

$$\begin{aligned}&-\Delta\bar y = \gamma\bar u, \quad \bar y|_{\partial\Omega} = 0,\\ &-\Delta\bar p + 3\bar y^2\bar p = -(\bar y - y_d), \quad \bar p|_{\partial\Omega} = 0,\\ &\alpha\bar u - \gamma\bar p + \bar z_b - \bar z_a = 0,\\ &\bar u \ge a, \quad \bar z_a \ge 0, \quad \bar z_a\,(\bar u - a) = 0,\\ &\bar u \le b, \quad \bar z_b \ge 0, \quad \bar z_b\,(b - \bar u) = 0.\end{aligned}$$

5.6. Optimality conditions for problems with general constraints. We sketch now the theory of optimality conditions for general problems of the form

$$\min_{w\in W} f(w) \quad \text{subject to} \quad G(w)\in\mathcal{K}, \quad w\in\mathcal{C}. \tag{5.35}$$

Here, $f : W \to \mathbb{R}$, $G : W \to V$ are continuously Fréchet differentiable with Banach spaces W, V, $\mathcal{C} \subset V$ is non-empty, closed and convex, and $\mathcal{K} \subset V$ is a closed convex cone. Here, $\mathcal{K}$ is a cone if

$$\forall\, \lambda > 0 : v \in \mathcal{K} \Longrightarrow \lambda v \in \mathcal{K}.$$

We denote the feasible set by

$$W_{ad} := \{w \in W \,:\, G(w)\in\mathcal{K}, \quad w\in\mathcal{C}\}\,.$$

Remark It is no restriction not to include equality constraints. In fact

$$E(w) = 0, \quad C(w)\in\mathcal{K}_C$$

is equivalent to

$$G(w) := \begin{pmatrix} E(w)\\ C(w)\end{pmatrix} \in \{0\}\times\mathcal{K}_C =: \mathcal{K}.$$

5.7. A basic first order optimality condition. Let $\bar w$ be a local solution of (5.35). To develop an extension of Theorem 5.5, we define the cone of feasible directions as follows.

DEFINITION 5.11. *Let $W_{ad} \subset W$ be nonempty. The* tangent cone *of W_{ad} at $w \in W_{ad}$ is defined by*

$$T(W_{ad}; w) = \Big\{ s\in W \,:\, \exists\, \eta_k > 0, w_k \in W_{ad} :\ \lim_{k\to\infty} w_k = w,\ \lim_{k\to\infty} \eta_k(w_k - w) = s \Big\}\,.$$

Then we have the following optimality condition.

THEOREM 5.12. *Let $f : W \to \mathbb{R}$ be continuously Fréchet differentiable. Then for any local solution $\bar w$ of* (5.35) *the following optimality condition holds.*

$$\bar w \in W_{ad} \quad \textit{and} \quad \langle f'(\bar w), s\rangle_{W^*,W} \geq 0 \quad \forall\, s\in T(W_{ad};\bar w). \tag{5.36}$$

Proof. $\bar w \in W_{ad}$ is obvious. Let $s \in T(W_{ad}; \bar w)$ be arbitrary. Then there exist $(w_k) \subset W_{ad}$ and $\eta_k > 0$ with $w_k \to \bar w$ und $\eta_k(w_k - \bar w) \to s$. This yields for all sufficiently large k

$$0 \leq \eta_k(f(w_k) - f(\bar w)) = \langle f'(\bar w), \eta_k(w_k - \bar w)\rangle_{W^*,W} + \eta_k o(\|w_k - \bar w\|_W) \to \langle f'(\bar w), s\rangle_{W^*,W}$$

since $\eta_k o(\|w_k - \bar w\|_W) \to 0$, which follows from $\eta_k(w_k - \bar w) \to s$. □

5.8. Constraint qualification and Robinsons's regularity condition. We want to replace the tangent cone by a cone with a less complicated representation. Linearization of the constraints (assuming G is continuously differentiable) leads us to the *linearization cone* at a point $\bar{w} \in W_{ad}$ defined by

$$L(W_{ad}, G, \mathcal{K}, \mathcal{C}; \bar{w}) = \{\eta\, d \,:\, \eta > 0,\ d \in W,\ G(\bar{w}) + G'(\bar{w})d \in \mathcal{K},\ \bar{w} + d \in \mathcal{C}\}\,.$$

Assume now that the a local solution $\bar{w}$ of (5.35) satisfies the

Constraint Qualification:

$$L(W_{ad}, G, \mathcal{C}, \mathcal{K}; \bar{w}) \subset T(W_{ad}; \bar{w}) \tag{5.37}$$

Then the following result is obvious.

THEOREM 5.13. *Let $f : W \to \mathbb{R}$, $G : W \to V$ be continuously Fréchet differentiable, with Banach-spaces W, V. Further let $\mathcal{C} \subset V$ be non-empty, closed and convex, and let $\mathcal{K} \subset V$ be a closed convex cone. Then at every local solution $\bar{w}$ of* (5.35) *satisfying* (5.37) *the following optimality condition holds.*

$$\bar{w} \in W_{ad} \quad \textit{and} \quad \langle f'(\bar{w}), s\rangle_{W^*,W} \geq 0 \quad \forall\, s \in L(W_{ad}, G, \mathcal{C}, \mathcal{K}; \bar{w}). \tag{5.38}$$

Remark If G is affine linear, then (5.37) is satisfied. In fact, let $s \in L(W_{ad}, G, \mathcal{C}, \mathcal{K}; \bar{w})$. Then $s = \eta d$ with $\eta > 0$ and $d \in W$,

$$G(\bar{w} + d) = G(\bar{w}) + G'(\bar{w})d \in \mathcal{K}, \quad \bar{w} + d \in \mathcal{C}.$$

Since $G(\bar{w}) \in \mathcal{K}$ and $\bar{w} \in \mathcal{C}$, the convexity of $\mathcal{K}$ and $\mathcal{C}$ yields $w_k := \bar{w} + \frac{\eta}{k} d \in W_{ad}$. Choosing $\eta_k = 1/k$ shows that $s \in T(W_{ad}; \bar{w})$. □

In general, (5.37) can be ensured if $\bar{w}$ satisfies the

Regularity Condition of Robinson:

$$0 \in \operatorname{int}\left(G(\bar{w}) + G'(\bar{w})\,(\mathcal{C} - \bar{w}) - \mathcal{K}\right). \tag{5.39}$$

We have the following important and deep result by Robinson [**66**].

THEOREM 5.14. *Robinson's regularity condition* (5.39) *implies the constraint qualification* (5.37).

Proof. See [**66**, Thm. 1, Cor. 2]. □

5.9. Karush-Kuhn-Tucker conditions. Using Robinson's regularity condition, we can write the optimality condition (5.38) in a more explicit form.

THEOREM 5.15. *(Zowe and Kurcyusz* [**82**]*)*
Let $f : W \to \mathbb{R}$, $G : W \to V$ be continuously Fréchet differentiable, with Banach-spaces W, V. Further let $\mathcal{C} \subset V$ be non-empty, closed and convex, and let $\mathcal{K} \subset V$ be a closed convex cone. Then for any local solution $\bar{w}$ of (5.35) *at which Robinson's regularity condition* (5.39) *is satisfied, the following optimality condition holds:*

There exists a Lagrange multiplier $\bar{q} \in V^*$ with

$$G(\bar{w}) \in \mathcal{K}, \tag{5.40}$$

$$\bar{q} \in \mathcal{K}^\circ := \{q \in V^* \,:\, \langle q, v\rangle_{V^*,V} \le 0 \quad \forall\, v \in \mathcal{K}\}\,, \tag{5.41}$$

$$\langle \bar{q}, G(\bar{w})\rangle_{V^*,V} = 0, \tag{5.42}$$

$$\bar{w} \in \mathcal{C}, \quad \langle f'(\bar{w}) + G'(\bar{w})^*\bar{q}, w - \bar{w}\rangle_{W^*,W} \ge 0 \quad \forall\, w \in \mathcal{C}. \tag{5.43}$$

Using the Lagrangian function

$$L(w,q) := f(w) + \langle q, G(w)\rangle_{V^*,V}$$

we can write (5.43) *in the compact form*

$$\bar{w} \in \mathcal{C}, \quad \langle L'_w(\bar{w},\bar{q}), w - \bar{w}\rangle_{W^*,W} \ge 0 \quad \forall\, w \in \mathcal{C}. \tag{5.43}$$

Proof. Under Robinson's regularity condition (5.39), a separation argument can be used to derive (5.41)–(5.43), see [**82**]. □

A similar result can be shown if $\mathcal{K}$ is a closed convex set instead of a closed convex cone, see [**11**], but then (5.41), (5.42) have a more complicated structure.

5.10. Application to PDE-constrained optimization. In PDE-constrained optimization, we have usually a state equation and constraints on control and/or state. Therefore, we consider as a special case the problem

$$\min_{(y,u)\in Y\times U} f(y,u) \quad \text{subject to } E(y,u) = 0, \quad C(y) \in \mathcal{K}_C, \quad u \in U_{ad}, \tag{5.44}$$

where $E : Y \times U \to Z$ and $C : Y \to V$ are continuously Fréchet differentiable, $\mathcal{K}_C \subset V$ is a closed convex cone in a Banach space $\tilde{Y} \supset Y$ and $U_{ad} \subset U$ is a closed convex set. We set

$$G : \begin{pmatrix} y \\ u \end{pmatrix} \in W := Y \times U \mapsto \begin{pmatrix} E(y,u) \\ C(y) \end{pmatrix} \in Z \times V, \quad \mathcal{K} = \{0\} \times \mathcal{K}_C, \quad \mathcal{C} = Y \times U_{ad}.$$

Then (5.44) has the form (5.35) and Robinson's regularity condition at a feasible point $\bar{w} = (\bar{y}, \bar{u})$ reads

$$0 \in \operatorname{int}\left(\begin{pmatrix} 0 \\ C(\bar{y}) \end{pmatrix} + \begin{pmatrix} E'_y(\bar{w}) & E'_u(\bar{w}) \\ C'(\bar{y}) & 0 \end{pmatrix}\begin{pmatrix} Y \\ U_{ad} - \bar{u} \end{pmatrix} - \begin{pmatrix} 0 \\ \mathcal{K}_C \end{pmatrix}\right). \tag{5.45}$$

We rewrite now (5.40)–(5.43) for our problem. The multiplier has the form $q = (p, \lambda) \in Z^* \times V^*$ and the Lagrangian function is given by

$$\mathcal{L}(y,u,q,\lambda) = f(y,u) + \langle p, E(y,u)\rangle_{Z^*,Z} + \langle \lambda, C(y)\rangle_{V^*,V} = L(y,u,p) + \langle \lambda, C(y)\rangle_{V^*,V}$$

with the Lagrangian

$$L(y,u,p) = f(y,u) + \langle p, E(y,u)\rangle_{Z^*,Z}$$

for the equality constraints.

Since $\mathcal{K} = \{0\} \times \mathcal{K}_C$, we have

$$\mathcal{K}^\circ = V^* \times \mathcal{K}_C^\circ$$

and thus (5.40)–(5.43) read

$$\begin{aligned}
&E(\bar y,\bar u)=0,\quad C(\bar y)\in\mathcal{K}_C,\\
&\bar\lambda\in\mathcal{K}_C^\circ,\quad \langle\bar\lambda,C(\bar y)\rangle_{V^*,V}=0,\\
&\langle L_y'(\bar y,\bar u,\bar p)+C'(\bar y)^*\bar\lambda,y-\bar y\rangle_{Y^*,Y}\geq 0\quad\forall\, y\in Y,\\
&\bar u\in U_{ad},\quad \langle L_u'(\bar y,\bar u,\bar p),u-\bar u\rangle_{U^*,U}\geq 0\quad\forall\, u\in U_{ad}.
\end{aligned}$$

This yields finally

$$E(\bar y,\bar u)=0,\quad C(\bar y)\in\mathcal{K}_C, \tag{5.46}$$

$$\bar\lambda\in\mathcal{K}_C^\circ,\quad \langle\bar\lambda,C(\bar y)\rangle_{V^*,V}=0, \tag{5.47}$$

$$L_y(\bar y,\bar u,\bar p)+C'(\bar y)^*\bar\lambda=0, \tag{5.48}$$

$$\bar u\in U_{ad},\quad \langle L_u(\bar y,\bar u,\bar p),u-\bar u\rangle_{U^*,U}\geq 0\quad\forall\, u\in U_{ad}. \tag{5.49}$$

Remark Without the state constraint $C(y)\in\mathcal{K}_C$ (which can formally be removed by omitting everything involving C or by making the constraint trivial, e.g, $C(y)=y$, $V=Y$, $\mathcal{K}_C=Y$), we recover exactly the optimality conditions (5.14)–(5.16) of Corollary 5.7. □

We show next that the following Slater-type condition implies Robinson's regularity condition (5.45).

LEMMA 5.16. *Let $\bar w\in W_{ad}$. If $E_y'(\bar w)\in\mathcal{L}(Y,Z)$ is surjective and if there exist $\tilde u\in U_{ad}$ and $\tilde y\in Y$ with*

$$\begin{aligned}
&E_y'(\bar w)(\tilde y-\bar y)+E_u'(\bar w)(\tilde u-\bar u)=0,\\
&C(\bar y)+C'(\bar y)(\tilde y-\bar y)\in\operatorname{int}(\mathcal{K}_C)
\end{aligned}$$

then Robinson's regularity condition (5.45) *is satisfied.*

Proof. Let

$$\tilde v:=C(\bar y)+C'(\bar y)(\tilde y-\bar y).$$

Then there exists $\varepsilon>0$ with

$$\tilde v+B_V(2\varepsilon)\subset\mathcal{K}_C.$$

Here $B_V(\varepsilon)$ is the open ε-ball in V. Furthermore, there exists $\delta>0$ with

$$C'(\bar y)B_Y(\delta)\subset B_V(\varepsilon).$$

Using that $\tilde u\in U_{ad}$ and $\tilde y-\bar y+B_Y(\delta)\subset Y$ we have

$$\begin{aligned}
&\begin{pmatrix}0\\ C(\bar y)\end{pmatrix}+\begin{pmatrix}E_y'(\bar w) & E_u'(\bar w)\\ C'(\bar y) & 0\end{pmatrix}\begin{pmatrix}Y\\ U_{ad}-\bar u\end{pmatrix}-\begin{pmatrix}0\\ \mathcal{K}_C\end{pmatrix}\\
\supset&\begin{pmatrix}0\\ C(\bar y)\end{pmatrix}+\begin{pmatrix}E_y'(\bar w) & E_u'(\bar w)\\ C'(\bar y) & 0\end{pmatrix}\begin{pmatrix}\tilde y-\bar y+B_Y(\delta)\\ \tilde u-\bar u\end{pmatrix}-\begin{pmatrix}0\\ \tilde v+B_V(2\varepsilon)\end{pmatrix}\\
=&\begin{pmatrix}E_y'(\bar w)\\ C'(\bar y)\end{pmatrix}B_Y(\delta)+\begin{pmatrix}0\\ B_V(2\varepsilon)\end{pmatrix}\supset\begin{pmatrix}E_y'(\bar w)B_Y(\delta)\\ B_V(\varepsilon)\end{pmatrix}.
\end{aligned}$$

In the last step we have used $C'(\bar y)B_Y(\delta) \subset B_V(\varepsilon)$ and that, for all $v \in B_V(\varepsilon)$, therte holds $v + B_V(2\varepsilon) \supset B_V(\varepsilon)$. By the open mapping theorem $E_y'(\bar w)B_Y(\varepsilon)$ is open in Z and contains 0. Thererefore, the set on the left hand side is an open neighborhood of 0 in $Z \times V$. □

5.11. Applications.

5.11.1. *Elliptic problem with state constraints.* We consider the problem

$$\begin{aligned} \min \qquad & f(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \\ \text{subject to} \quad & -\Delta y + y = \gamma\, u \quad \text{on } \Omega, \\ & \frac{\partial y}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \\ & y \geq 0 \quad \text{on } \Omega. \end{aligned} \tag{5.50}$$

Let $n \leq 3$. We know from Theorem 2.47 that for $u \in U := L^2(\Omega)$ there exists a unique weak solution $y \in H^1(\Omega) \cap C(\bar\Omega)$ of the state equation. We can write the problem in the form

$$\min f(y,u) \quad \text{subject to} \quad Ay + Bu = 0, \quad y \geq 0.$$

where $Bu = -\gamma u$, and A is induced by the bilinear form $a(y,v) = \int_\Omega \nabla y \cdot \nabla v\, dx + (y,v)_{L^2(\Omega)}$.

With appropriate spaces $Y \subset H^1(\Omega)$, $Z \subset H^1(\Omega)^*$ and $V \supset Y$ we set

$$E : \begin{pmatrix} y \\ u \end{pmatrix} \in Y \times U \mapsto Ay + Bu \in Z, \quad C(y) = y, \quad \mathcal{K}_C = \{v \in V \,:\, v \geq 0\}, \quad U_{ad} = U$$

and arrive at a problem of the form (5.44). For the naive choice $V = Y = H^1(\Omega)$, $Z = Y^*$, the cone $\mathcal{K}_C$ has no interior point. But since $Bu = -\gamma u \in L^2(\Omega)$, we know that all solutions y of the state equation live in the space

$$Y = \left\{ y \in H^1(\Omega) \cap C(\bar\Omega) \,:\, Ay \in U^* = L^2(\Omega) \right\}$$

and Y is a Banach space with the norm $\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar\Omega)} + \|Ay\|_{L^2(\Omega)}$ (why?). Then $A : Y \mapsto L^2(\Omega) =: Z$ is bounded and by Theorem 2.47 also surjective. Finally, we choose $V = C(\bar\Omega)$, then $V \supset Y$ and $\mathcal{K}_C \subset V$ has an interior point.

Now assume that there exists $\tilde y \in Y$, $\tilde y > 0$ and $\tilde u \in U$ with (note that $E_y' = A, E_u' = B$)

$$A(\tilde y - \bar y) + B(\tilde u - \bar u) = 0.$$

For example in the case $\gamma \equiv 1$ the choice $\tilde y = \bar y + 1$, $\tilde u = \bar u + 1$ works. Then by Lemma 5.16 Robinson's regularity assumption is satisfied. Therefore, at a solution $(\bar y, \bar u)$ the necessary conditions (5.46)–(5.49) are satisfied: Using that

$$L(y,u,p) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + (p, Ay + Bu)_{L^2(\Omega)}$$

we obtain

$$
\begin{aligned}
&A\bar{y} + B\bar{u} = 0, \quad \bar{y} \geq 0,\\
&\bar{\lambda} \in \mathcal{K}_C^\circ, \quad \langle \bar{\lambda}, \bar{y} \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = 0,\\
&(\bar{y} - y_d, v)_{L^2(\Omega)} + (\bar{p}, Av)_{L^2(\Omega)} + \langle \bar{\lambda}, v \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = 0,\\
&(\alpha\bar{u} - \gamma\bar{p}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall\, u \in U.
\end{aligned}
$$

One can show that the set $\mathcal{K}_C^\circ \subset C(\bar{\Omega})^*$ of nonpositive functionals on $C(\bar{\Omega})$ can be identified with nonpositive regular Borel measures, i.e.

$$
\begin{aligned}
&\lambda \in \mathcal{K}_C^\circ \quad \Longleftrightarrow\\
&\langle \lambda, v \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = -\int_\Omega v(x)\, d\mu_\Omega(x) - \int_{\partial\Omega} v(x)\, d\mu_{\partial\Omega}(x) \text{ with nonneg. measures } \mu_\Omega, \mu_{\partial\Omega}.
\end{aligned}
$$

Therefore, the optimality system is formally a weak formulation of the following system.

$$
\begin{aligned}
&-\Delta\bar{y} + \bar{y} = \gamma\bar{u} \text{ on } \Omega, \quad \frac{\partial y}{\partial \nu} = 0 \quad \text{on } \partial\Omega,\\
&\bar{y} \geq 0, \quad \bar{\mu}_\Omega, \quad \bar{\mu}_{\partial\Omega} \text{ nonnegative regular Borel measures,}\\
&\int_\Omega \bar{y}(x)\, d\mu_\Omega(x) + \int_{\partial\Omega} \bar{y}(x)\, d\mu_{\partial\Omega}(x) = 0,\\
&-\Delta\bar{p} + \bar{p} = -(\bar{y} - y_d) + \bar{\mu}_\Omega \text{ on } \Omega, \quad \frac{\partial p}{\partial \nu} = \bar{\mu}_{\partial\Omega} \quad \text{on } \partial\Omega,\\
&\alpha\bar{u} + \gamma\bar{p} = 0.
\end{aligned}
$$

CHAPTER 2

Optimization Methods in Banach Spaces

Michael Ulbrich
Lehrstuhl Optimierung
TU München

1. Synopsis

The aim of this chapter is to give an introduction to selected optimization algorithms that are well-suited for PDE-constrained optimization. For the development and analysis of such algorithms, a functional analytic setting is the framework of choice. Therefore, we will develop optimization methods in this abstract setting and then return to concrete problems later.

Optimization methods are iterative algorithms for finding (global or local) solutions of minimization problems. Usually, we are already satisified if the method can be proved to converge to *stationary* points. These are points that satisfy the first-order optimality conditions. Besides global convergence, which will not be the main focus of this chapter, fast local convergence is desired. All fast converging optimization methods use the idea of Newton's method in some sense. Therefore, our main focus will be on Newton-type methods for optimization problems in Banach spaces.

Optimization methods generate a sequence (w^k) of iterates. Essentially, as already indicated, there are two desirable properties an optimization algorithm should have:

1. Global convergence:

   There are different flavors to formulate global convergence; here is a selection:

   a) Every accumluation point of w^k is a stationary point.

   b) For some continuous stationarity measure $\Sigma(w)$, e.g., $\Sigma(w) := \|f'(w)\|_{W^*}$ in the unconstrained case, there holds

   $$\lim_{k\to\infty} \Sigma(w^k) = 0.$$

   c) There exists an accumluation point of (w^k) that is stationary.

d) For the continuous stationarity measure $\Sigma(w)$ there holds

$$\liminf_{k\to\infty} \Sigma(w^k) = 0.$$

Note that b) implies a) and c) implies d).

2. Fast local convergence.

These are local results in a neighborhood of a stationary point w^*:
There exists $\delta > 0$ such that, for all $w^0 \in W$ with $\|w^0 - w^*\|_W < \delta$, we have $w^k \to w^*$ and

$$\|w^{k+1} - w^*\|_W = o(\|w^k - w^*\|_W), \quad \text{(q-superlinear convergence)}$$

or even, for $\alpha > 0$,

$$\|w^{k+1} - w^*\|_W = O(\|w^k - w^*\|_W^{1+\alpha}). \quad \text{(q-superlinear convergence with order } 1+\alpha)$$

The case $1 + \alpha = 2$ is called q-quadratic convergence.

We begin with a discussion of globalization concepts. Then, in the rest of this chapter, we present locally fast convergent methods that all can be viewed as Newton-type methods.

Notation. If W is a Banach space, we denote by W^* its dual space. The Frechet-derivative (F-derivative) of an operator $G : X \to Y$ between Banach spaces is denoted by $G' : X \to \mathcal{L}(X, Y)$, where $\mathcal{L}(X, Y)$ are the bounded linear operators $A : X \to Y$. In particular, the derivative of a real-valued function $f : W \to \mathbb{R}$ is denoted by $f' : W \to W^*$. In case of a Hilbert space W, the gradient $\nabla f : W \to W$ is the Riesz representation of f', i.e.,

$$(\nabla f(w), v)_W = \langle f'(w), v\rangle_{W^*,W} \quad \forall\, v \in W.$$

Here $\langle f'(w), v\rangle_{W^*,W}$ denotes the dual pairing between the dual space $W^* = \mathcal{L}(W, \mathbb{R})$ and W and $(\cdot,\cdot)_W$ is the inner product. Note that in Hilbert space we can do the identification $W^* = W$ via $\langle\cdot,\cdot\rangle_{W^*,W} = (\cdot,\cdot)_W$, but this is not always done.

2. Globally convergent methods in Banach spaces

2.1. Unconstrained optimization. For understanding how global convergence can be achieved, it is important to look at unconstrained optimization first:

$$\min_{w\in W} f(w)$$

with W a real Banach space and $f : W \to \mathbb{R}$ continuously F-differentiable.

The first-order optimality conditions for a local minimum $w^* \in W$ are well-known:

$w^* \in W$ satisfies

$$f'(w) = 0.$$

We develop a general class of methods that is globally convergent: *Descent methods*.

The idea of descent methods is to find, at the current (kth) iterate $w^k \in W$, a direction $s^k \in W$ such that $\phi_k(t) \stackrel{\text{def}}{=} f(w^k + ts^k)$ is decreasing at $t = 0$:

$$\phi_k'(0) = \langle f'(w^k), s^k \rangle_{W^*,W} < 0.$$

Of course, this descent can be very small. However, from the (sharp) estimate

$$\phi_k'(0) = \langle f'(w^k), s^k \rangle_{W^*,W} \geq -\|f'(w^k)\|_{W^*} \|s^k\|_W$$

it is natural to derive the following quality requirement ("angle" condition)

$$\langle f'(w^k), s^k \rangle_{W^*,W} \leq -\eta \|f'(w^k)\|_{W^*} \|s^k\|_W \tag{2.1}$$

for the descent direction. Here $\eta \in (0, 1)$ is fixed.

A second ingredient of a descent method is a step size rule to obtain a step size $\sigma_k > 0$ such that

$$\phi_k(\sigma_k) < \phi_k(0).$$

Then, the new iterate is computed as $w^{k+1} := w^k + \sigma_k s^k$. Overall, we obtain:

ALGORITHM 2.1 (General descent method).

0. Choose an initial point $w^0 \in W$.

For $k = 0, 1, 2, \ldots$:

1. If $f'(w^k) = 0$, STOP.

2. Choose a descent direction $s^k \in W$: $\langle f'(w^k), s^k \rangle_{W^*,W} < 0$.

3. Choose a step size $\sigma_k > 0$.

4. Set $w^{k+1} := w^k + \sigma_k s^k$.

In this generality, it is not possible to prove global convergence. We need additional requirements on the quality of the descent direction and the step sizes:

1. Admissibility of the search directions:

$$\frac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W} \stackrel{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad \|f'(w^k)\|_{W^*} \stackrel{k\to\infty}{\longrightarrow} 0.$$

2. Admissibility of the step sizes:

$$f(w^k + \sigma_k s^k) - f(w^k) \stackrel{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad \frac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W} \stackrel{k\to\infty}{\longrightarrow} 0.$$

These conditions become more intuitive by realizing that the expression $\dfrac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W}$ is the slope of f at w^k in the direction s^k:

$$\frac{d}{dt} f\left(w^k + \frac{s^k}{\|s^k\|_W}\right)\bigg|_{t=0} = \frac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W}.$$

Therefore, admissible step sizes means that if the f-decrease become smaller and smaller then the slopes along the s^k have to become smaller and smaller. And admissible search directions means that if the slopes along the s^k become smaller and smaller then the steepest possible slope has to become smaller and smaller.

With these two conditions at hand, we can prove global convergence.

THEOREM 2.2. *Let f be continuously F-differentiable and (w^k), (s^k), (σ_k) be generated by Algorithm* 2.1. *Assume that (σ_k) and (s_k) are admissible and that $(f(w^k))$ is bounded below. Then*

$$\lim_{k\to\infty} f'(w^k) = 0.$$

In particular, every accumulation point of (w^k) is a stationary point.

Proof. Let $(f(w^k))$ be bounded below by $f^* \in \mathbb{R}$. Then

$$f(w^0) - f^* = \sum_{k=0}^{\infty} (f(w^k) - f(w^{k+1})) = \sum_{k=0}^{\infty} |f(w^k + \sigma_k s^k) - f(w^k)|$$

and thus $f(w^k + \sigma_k s^k) - f(w^k) \to 0$. By the admissibility of (σ_k), this implies

$$\frac{\langle f'(w^k), s^k\rangle_{W^*,W}}{\|s^k\|_W} \overset{k\to\infty}{\longrightarrow} 0.$$

Now the admissibility of (s^k) yields

$$\|f'(w^k)\|_{W^*} \overset{k\to\infty}{\longrightarrow} 0.$$

If w^* is an accumulation point of (w^k), then there exists a subsequence $(w^k)_K \to w^*$ and due to monotonicity of $f(w^k)$ we conclude $f(w^k) \geq f(w^*)$ for all k. Hence, by continuity,

$$f'(w^*) = \lim_{k\to\infty} f'(w^k) = 0$$

□

There are two questions open:

a) How can we check in practice if a search direction is admissible or not?

b) How can we compute admissible step sizes?

An answer to question a) is provided by the following Lemma:

LEMMA 2.3. *If the search directions (s^k) satisfy the angle condition* (2.1) *then they are admissible.*

Proof. The angle condition yields

$$\|f'(w^k)\|_{W^*} \leq -\frac{1}{\eta} \frac{\langle f'(w^k), s^k\rangle_{W^*,W}}{\|s^k\|_W}.$$

□

The mother of all step size rules is the

2.1.1. *Armijo rule:* Choose the maximum $\sigma_k \in \{1, 1/2, 1/4, \ldots\}$ for which

$$f(w^k + \sigma_k s^k) - f(w^k) \leq \gamma \sigma_k \langle f'(w^k), s^k \rangle_{W^*,W}.$$

Here $\gamma \in (0,1)$ is a constant. The next result shows that Armijo step sizes exist.

LEMMA 2.4. *Let f' be uniformly continuous on $N_0^\rho = \{w + s : f(w) \leq f(w^0),\ \|s\|_W \leq \rho\}$ for some $\rho > 0$. Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $w^k \in W$ with $f(w^k) \leq f(w^0)$ and all $s^k \in W$ that satisfy*

$$\frac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W} \leq -\varepsilon,$$

there holds

$$f(w^k + \sigma s^k) - f(w^k) \leq \gamma \sigma \langle f'(w^k), s^k \rangle_{W^*,W} \quad \forall\, \sigma \in [0, \delta/\|s^k\|_W].$$

Proof. We have, with appropriate $\tau_\sigma \in [0, \sigma]$,

$$\begin{aligned} f(w^k + \sigma s^k) - f(w^k) &= \sigma \langle f'(w^k + \tau_\sigma s^k), s^k \rangle_{W^*,W} \\ &\leq \sigma \langle f'(w^k), s^k \rangle_{W^*,W} + \sigma \|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} \|s^k\|_W \\ &= \gamma \sigma \langle f'(w^k), s^k \rangle_{W^*,W} + \rho_k(\sigma), \end{aligned}$$

where

$$\rho_k(\sigma) := (1-\gamma)\sigma \langle f'(w^k), s^k \rangle_{W^*,W} + \sigma \|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} \|s^k\|_W.$$

Now we use the uniform continuity of f' to choose $\delta \in (0, \rho)$ so small that

$$\|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} < (1-\gamma)\varepsilon \quad \forall\, \sigma \in [0, \delta/\|s^k\|_W].$$

This is possible since

$$\|\tau_\sigma s^k\|_W \leq \sigma \|s^k\|_W \leq \delta.$$

Then

$$\begin{aligned} \rho_k(\sigma) &= (1-\gamma)\sigma \langle f'(w^k), s^k \rangle_{W^*,W} + \sigma \|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} \|s^k\|_W \\ &\leq -(1-\gamma)\varepsilon\sigma \|s^k\|_{W^*,W} + (1-\gamma)\varepsilon\sigma \|s^k\|_W = 0. \end{aligned}$$

□

Next, we prove the admissibility of Armijo step sizes under mild conditions.

LEMMA 2.5. *Let f' be uniformly continuous on $N_0^\rho = \{w + s : f(w) \leq f(w^0),\ \|s\|_W \leq \rho\}$ for some ρ. We consider Algorithm* 2.1*, where (σ_k) is generated by the Armijo rule and the s^k are chosen such that they are not too short in the following sense:*

$$\|s^k\|_W \geq \phi\left(-\frac{\langle f'(w^k), s^k \rangle_{W^*,W}}{\|s^k\|_W} \right),$$

where $\phi : [0, \infty) \to [0, \infty)$ is monotonically increasing and satisfies $\phi(t) > 0$ for all $t > 0$. Then the step sizes (σ_k) are admissible.

Proof. Assume that there exist an infinite set K and $\varepsilon > 0$ such that

$$\frac{\langle f'(w^k), s^k\rangle_{W^*,W}}{\|s^k\|_W} \leq -\varepsilon \quad \forall\, k \in K.$$

Then

$$\|s^k\|_W \geq \phi\left(-\frac{\langle f'(w^k), s^k\rangle_{W^*,W}}{\|s^k\|_W}\right) \geq \phi(\varepsilon) =: \eta > 0 \quad \forall\, k \in K.$$

By Lemma 2.4, for $k \in K$ we have either $\sigma_k = 1$ or $\sigma_k \geq \delta/(2\|s^k\|)$. Hence,

$$\sigma_k \|s^k\|_W \geq \min\{\delta/2, \eta\} \quad \forall\, k \in K.$$

This shows

$$\begin{aligned} f(w^k + \sigma_k s^k) - f(w^k) &\leq \gamma\sigma_k \langle f'(w^k), s^k\rangle_{W^*,W} = \gamma\sigma_k \|s^k\|_W \frac{\langle f'(w^k), s^k\rangle_{W^*,W}}{\|s^k\|_W} \\ &\leq -\gamma \min\{\delta/2, \eta\}\varepsilon \quad \forall\, k \in K. \end{aligned}$$

Therefore

$$f(w^k + \sigma_k s^k) - f(w^k) \not\to 0.$$

□

In the Banach space setting, the computation of descent directions is not straightforward. Note that the negative derivative of f is *not* suitable, since $W^* \ni f'(w^k) \notin W$.

In the Hilbert space setting, however, we *can* choose $W^* = W$ and $\langle\cdot,\cdot\rangle_{W^*,W} = (\cdot,\cdot)_W$ by the Riesz representation theorem. Then we have $f'(w^k) = \nabla f(w^k) \in W$ and $-\nabla f(w^k)$ is the direction of steepest descent, as we will show below.

Certainly the most well-known descent method is the steepest descent method. In Banach space, the steepest descent directions of f at w are defined by $s = td_{sd}$, $t > 0$, where d_{sd} solves

$$\min_{\|d\|_W=1} \langle f'(w), d\rangle_{W^*,W}.$$

Now consider the case where $W = W^*$ is a Hilbert space. Then

$$d_{sd} = -\frac{\nabla f(w)}{\|\nabla f(w)\|_W}$$

In fact, by the Cauchy-Schwarz inequality,

$$\begin{aligned} \min_{\|d\|_W=1} \langle f'(w), d\rangle_{W^*,W} &= \min_{\|d\|_W=1} (\nabla f(w), d)_W \leq -\|\nabla f(w)\|_W \\ &= \left(\nabla f(w), -\frac{\nabla f(w)}{\|\nabla f(w)\|_W}\right)_W \end{aligned}$$

Therefore, $-\nabla f(w)$ is a steepest descent direction. This is the reason why the steepest descent method is also called gradient method.

It should be mentioned that the steepest descent method is usually very inefficient. Therefore, the design of efficient globally convergent methods works as follows: A locally fast convergent method

(e.g., Newton's method) is used to generate trial steps. If the generated step satisfies a (generalized) angle test ensuring admissibility of the step, the step is selected. Otherwise, another search direction is chosen, e.g., the steepest descent direction.

2.2. Optimization with simple constraints. We now develop descent methods for simply constrained problems of the form

$$\min f(w) \quad \text{s.t.} \quad w \in S \tag{2.2}$$

with W a Hilbert space, $f : W \to \mathbb{R}$ continuously F-differentiable, and $S \subset W$ closed and convex.

EXAMPLE 2.6. *A scenarion frequently found in practice is*

$$W = L^2(\Omega), \quad S = \left\{u \in L^2(\Omega) \,:\, a(x) \le u(x) \le b(x) \text{ a.e. on } \Omega\right\}$$

with L^∞-functions a, b. It is then very easy to compute the projection P_S onto S, which will be needed in the following:

$$P_S(w)(x) = P_{[a(x),b(x)]}(w(x)) = \max(a(x), \min(w(x), b(x))).$$

The presence of the constraint set S requires to take care that we stay feasible with respect to S, or, (if we think of an infeasible method) that we converge to feasibility. In the following, we consider a feasible algorithm, i.e., $w^k \in S$ for all k.

If w^k is feasible and we try to apply the unconstrained descent method, we have the difficulty that already very small step sizes $\sigma > 0$ can result in points $w^k + \sigma s^k$ that are infeasible. The backtracking idea of considering only those $\sigma \ge 0$ for which $w^k + \sigma s^k$ is feasible is not viable, since very small step sizes or even $\sigma_k = 0$ might be the result.

Therefore, instead of performing a line search along the ray $\left\{w^k + \sigma s^k \,:\, \sigma \ge 0\right\}$, we perform a line search along the projected path

$$\left\{P_S(w^k + \sigma s^k) \,:\, \sigma \ge 0\right\},$$

where P_S is the projection onto S. Of course, we have to ensure that along this path we achieve sufficient descent as long as w^k is not a stationary point. Unfortunately, not any descent direction is suitable here.

EXAMPLE 2.7. *Consider*

$$S = \left\{w \in \mathbb{R}^2 \,:\, w_1 \ge 0,\ w_1 + w_2 \ge 3\right\}, \quad f(w) = 5w_1^2 + w_2^2.$$

Then, at $w^k = (1,2)^T$, we have $\nabla f(w^k) = (10,4)^T$. Since f is convex quadratic with minimum $w^* = 0$, the Newton step is

$$d^k = -w^k = -(1,2)^T.$$

This is a descent direction, since

$$\nabla f(w^k)^T d^k = -18.$$

But, for $\sigma \ge 0$, there holds

$$P_S(w^k - \sigma d^k) = P_S((1-\sigma)(1,2)^T) = (1-\sigma)\begin{pmatrix}1\\2\end{pmatrix} + \sigma\begin{pmatrix}3/2\\3/2\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix} + \frac{\sigma}{2}\begin{pmatrix}1\\-1\end{pmatrix}.$$

From

$$\nabla f(w^k)^T \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 6$$

we see that we are getting ascent, not descent, along the projected path, although d^k is a descent direction.

The example shows that care must be taken in choosing appropriate search directions for projected methods. Since the projected descent properties of a search direction are more complicated to judge than in the unconstrained case, it is out of the scope of this chapter to give a general presentation of this topic. In the finite dimensional setting, we refer to [**47**] for a detailed discussion. Here, we only consider the projected gradient method.

ALGORITHM 2.8 (Projected gradient method).

0. Choose $w^0 \in S$.

For $k = 0, 1, 2, 3, \ldots$:

1. Set $s^k = -\nabla f(w^k)$.

2. Choose σ_k by a projected step size rule.

3. Set $w^{k+1} := P_S(w^k + \sigma_k s^k)$.

For abbreviation, let

$$w_\sigma^k = w^k - \sigma \nabla f(w^k).$$

We will prove global convergence of this method. To do this, we need to collect some facts about the projection operator P_S.

The following result shows that along the projected steepest descent path we achieve a certain amount of descent:

LEMMA 2.9. *Let W be a Hilbert space and let $f : W \to \mathbb{R}$ be continuously F-differentiable on a neighborhood of the closed convex set S. Let $w^k \in S$ and assume that ∇f is α-order Hölder-continuous with modulus $L > 0$ on*

$$\left\{(1-t)w^k + tP_S(w_\sigma^k) \,:\, 0 \le t \le 1\right\}.$$

for some $\alpha \in (0, 1]$. Then there holds

$$f(P_S(w_\sigma^k)) - f(w^k) \le -\frac{1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 + L\|P_S(w_\sigma^k) - w^k\|_W^{1+\alpha}$$

Proof.

$$\begin{aligned} f(P_S(w_\sigma^k)) - f(w^k) &= (\nabla f(v_\sigma^k), P_S(w_\sigma^k) - w^k)_W \\ &= (\nabla f(w^k), P_S(w_\sigma^k) - w^k)_W + (\nabla f(v_\sigma^k) - \nabla f(w^k), P_S(w_\sigma^k) - w^k)_W \end{aligned}$$

with appropriate $v_\sigma^k \in \left\{(1-t)w^k + tP_S(w_\sigma^k) \,:\, 0 \le t \le 1\right\}$.

Now, since $w_\sigma^k - w^k = \sigma s^k = -\sigma \nabla f(w^k)$ and $w^k = P_S(w^k)$, we obtain

$$\begin{aligned}
-\sigma(\nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &= (w_\sigma^k - w^k, P_S(w_\sigma^k) - w^k)_W \\
&= (w_\sigma^k - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\
&= (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\
&\quad + \underbrace{(w_\sigma^k - P_S(w_\sigma^k), P_S(w_\sigma^k) - P_S(w^k))_W}_{\geq 0} \\
&\geq (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\
&= \|P_S(w_\sigma^k) - w^k\|_W^2.
\end{aligned}$$

Next, we use

$$\|v_\sigma^k - w^k\|_W \leq \|P_S(w_\sigma^k) - w^k\|_W.$$

Hence,

$$\begin{aligned}
(\nabla f(v_\sigma^k) - \nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &\leq \|\nabla f(v_\sigma^k) - \nabla f(w^k)\|_W \|P_S(w_\sigma^k) - w^k\|_W \\
&\leq L\|v_\sigma^k - w^k\|_W^\alpha \|P_S(w_\sigma^k) - w^k\|_W \\
&\leq L\|P_S(w_\sigma^k) - w^k\|_W^{1+\alpha}.
\end{aligned}$$

□

We now consider the following

2.2.1. *Projected Armijo rule:* Choose the maximum $\sigma_k \in \{1, 1/2, 1/4, \ldots\}$ for which

$$f(P_S(w^k + \sigma_k s^k)) - f(w^k) \leq -\frac{\gamma}{\sigma_k}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^2.$$

Here $\gamma \in (0, 1)$ is a constant.

In the unconstrained case, we recover the ordinary Armijo rule:

$$f(P_S(w^k + \sigma_k s^k)) - f(w^k) = f(w^k + \sigma_k s^k) - f(w^k),$$

$$-\frac{\gamma}{\sigma_k}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^2 = -\frac{\gamma}{\sigma_k}\|\sigma_k s^k\|_W^2 = -\gamma\sigma_k\|s^k\|_W^2 = \gamma\sigma_k(\nabla f(w^k), s^k)_W.$$

As a stationarity measure $\Sigma(w) = \|p(w)\|_W$ we use the norm of the *projected gradient*

$$p(w) \stackrel{\text{def}}{=} w - P_S(w - \nabla f(w)).$$

In fact, the first-order optimality conditions for (2.2) are

$$w \in S, \quad (\nabla f(w), v - w)_W \geq 0 \quad \forall\, v \in S.$$

By Lemma 5.2, this is equivalent to

$$w - P_S(w - \nabla f(w)) = 0.$$

As a next result we show that projected Armijo step sizes exist.

LEMMA 2.10. *Let W be a Hilbert space and let $f : W \to \mathbb{R}$ be continuously F-differentiable on a neighborhood of the closed convex set S. Then, for all $w^k \in S$ with $p(w^k) \neq 0$, the projected Armijo rule terminates sucessfully.*

Proof. We proceed as in the proof of Lemma 2.9 and obtain (we have not assumed Hölder continuity of ∇f here)

$$f(P_S(w_\sigma^k)) - f(w^k) \leq \frac{-1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W).$$

It remains to show that, for all small $\sigma > 0$,

$$\frac{\gamma - 1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W) \leq 0$$

But this follows easily from (Lemma 5.2 e)):

$$\frac{\gamma - 1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 \leq \underbrace{(\gamma - 1)\|p(w^k)\|_W}_{<0}\|P_S(w_\sigma^k) - w^k\|_W.$$

□

THEOREM 2.11. *Let W be a Hilbert space, $f : W \to \mathbb{R}$ be continuously F-differentiable, and $S \subset W$ be nonempty, closed, and convex. Consider Algorithm* 2.1 *and assume that $f(w^k)$ is bounded below. Furthermore, let ∇f be α-order Hölder continuous on*

$$N_0^\rho = \{w + s : f(w) \leq f(w^0),\ \|s\|_W \leq \rho\}$$

for some $\alpha > 0$ and some $\rho > 0$. Then

$$\lim_{k\to\infty} \|p(w^k)\|_W = 0.$$

Proof. Set $p^k = p(w^k)$ and assume $p^k \not\to 0$. Then there exist $\varepsilon > 0$ and an infinite set K with $\|p^k\|_W \geq \varepsilon$ for all $k \in K$.

By construction we have that $f(w^k)$ is monotonically decreasing and by assumption the sequence is bounded below. Since ∇f is α-order Hölder continuous on N_0, we have for all $k \in K$

$$f(w^k) - f(w^{k+1}) \geq \frac{\gamma}{\sigma_k}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^2 \geq \gamma\sigma_k\|p^k\|_W^2 \geq \gamma\sigma_k\varepsilon^2,$$

where we have used the Armijo condition and Lemma 5.2 e). This shows $(\sigma_k)_K \to 0$ and $(\|P_S(w^k + \sigma_k s^k) - w^k\|_W$ 0.

For large $k \in K$ we have $\sigma_k \leq 1/2$ and therefore, the Armijo condition did not hold for the step size $\sigma = 2\sigma_k$. Hence,

$$\begin{aligned} -\frac{\gamma}{2\sigma_k}\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 &\leq f(P_S(w^k + 2\sigma_k s_k)) - f(w^k) \\ &\leq -\frac{1}{2\sigma_k}\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 + L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}. \end{aligned}$$

Here, we have applied Lemma 2.9 and the fact that by Lemma 5.2 e)

$$\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W \leq 2\|P_S(w^k + \sigma_k s^k) - w^k\|_W \overset{K\ni k\to\infty}{\longrightarrow} 0.$$

Hence,

$$\frac{1-\gamma}{2\sigma_k}\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 \leq L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}.$$

From this we derive

$$(1-\gamma)\|p^k\|_W\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W \leq L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}.$$

Hence,

$$(1-\gamma)\varepsilon \leq L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{\alpha} \leq L2^{\alpha}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^{\alpha} \overset{K\ni k\to\infty}{\longrightarrow} 0.$$

This is a contradiction. □

A careful choice of search directions will allow to extend the convergence theory to more general classes of projected descent algorithms. For instance, in finite dimensions, q-superlinearly convergent projected Newton methods and their globalization are investigated in [**47, 10**]. In an L^2 setting, the superlinear convergence of projected Newton methods was investigated by Kelley and Sachs in [**48**].

2.3. General optimization problems. For more general optimization problems than we discussed so far, one usually globalizes by choosing step sizes based on an Armijo-type rule that is applied to a suitable merit function. For instance, if we consider problems of the form

$$\min_w \ f(w) \quad \text{s.t.} \quad E(w), \quad C(w) \in K,$$

with functions $f : W \to \mathbb{R}$, $E : W \to Z$, and $C : W \to V$, where W, Z, and V are Banach spaces and $K \subset V$ is a closed convex cone, a possible choice for a merit function is

$$m_\rho(w) = f(w) + \rho\|E(w)\|_W + \rho\,\mathrm{dist}(C(w), K)$$

with penalty parameter $\rho > 0$. In the case of equality constraints, a global convergence result for reduced SQP methods based on this merit function is presented in [**45**]. Other merit functions can be constructed by taking the norm of the residual of the KKT system, the latter being reformulated as a nonsmooth operator equation, see section 5. This residual-based type of globalization, however, does not take into account second-order information.

3. Newton-based methods – A preview

To give an impression of modern Newton-based for optimization problems approaches, we first consider all these methods in the finite dimensional setting: $W = \mathbb{R}^n$.

3.1. Unconstrained problems – Newton's method. Consider

$$\min_{w\in\mathbb{R}^n} f(w) \tag{3.1}$$

with $f:\mathbb{R}^n\to\mathbb{R}$ twice continuously differentiable.

From analysis we know that the first-order optimality conditions are:

$$\nabla f(w)=0. \tag{3.2}$$

Newton's method for (3.1) is obtained by applying Newton's method to the equation (3.2).

This, again, is done by linearizing $G=\nabla f$ about the current iterate w^k:

$$G(w^k)+G'(w^k)s^k=0,\quad w^{k+1}=w^k+s^k.$$

It is well-known – and will be proved later in a much more general context – that Newton's method converges q-superlinearly if G is C^1 and $G'(w^*)$ is invertible.

3.2. Simple constraints. Let $f:\mathbb{R}^n\to\mathbb{R}$ be C^2 and let $S\subset\mathbb{R}^n$ be a nonempty closed convex set.

We consider the problem

$$\min_{w\in\mathbb{R}^n} f(w)\quad\text{s.t.}\quad w\in S.$$

The optimality conditions, written in a form that directly generalizes to a Banach space setting, are: w^* satisfies

$$w\in S,\quad \nabla f(w)^T(v-w)\geq 0\ \forall\, v\in S. \tag{3.3}$$

This is a *Variational Inequality*, which we abbreviate VI($\nabla f, S$).

Note that the necessity of VI($\nabla f, S$) can be derived very easily: For any $v\in S$, the line segment $\{w^*+t(v-w^*)\,:\,0\leq t\leq 1\}$ connecting w^* and v is contained in S (convexity) and therefore, the function

$$\phi(t):=f(w^*+t(w-w^*))$$

is nondecreasing at $t=0$:

$$0\leq\phi'(0)=\nabla f(w^*)^T(v-w^*).$$

Similarly, in the Banach space setting, we will have that w^* solves

$$w\in S,\quad \langle f'(w),v-w\rangle_{W^*,W}\geq 0\ \forall\, v\in S$$

with $S\subset W$ closed, convex and $f':W\to W^*$.

Note that if $S=\mathbb{R}^n$, then (3.3) is equivalent to (3.2).

3.2.1. *Nonsmooth reformulation approach and generalized Newton methods.* In the development of projected descent methods we already used the important fact that the VI (3.3) is equivalent to

$$w - P_S(w - \theta \nabla f(w)) = 0, \tag{3.4}$$

where $\theta > 0$ is fixed.

EXAMPLE 3.1. *If S is a box, i.e.,*

$$S = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

then $P_S(w)$ can be computed very easily as follows:

$$P_S(w)_i = \max(a_i, \min(w_i, b_i)).$$

It is instructive (and not difficult) to check the equivalence of (3.3) *and* (3.4) *by hand.*

The function

$$\Phi(w) := w - P_S(w - \theta \nabla f(w))$$

is Lipschitz continuous (P_S is non-exapnsive and ∇f is C^1), but cannot be expected to be differentiable. Therefore, *at a first sight*, Newton's method is *not* applicable.

However, a second look shows that Φ has nice properties if S is sufficiently nice. To be concrete, let

$$S = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

be a box in the following. Then Φ is *piecewise* continuously differentiable, i.e., it consists of finitely many C^1-pieces $\Phi^j : \mathbb{R}^n \to \mathbb{R}^n$, $j = 1, \ldots, m$. More precisely, every component Φ_i of Φ consists of three pieces:

$$w_i - a_i, \quad w_i - b_i, \quad w_i - (w_i - \theta \nabla f(w)_i) = \theta \nabla f(w)_i,$$

hence Φ consists of (at most) 3^n pieces Φ^j.

Denote by

$$A(w) = \left\{ j \ : \ \Phi^j(w) = \Phi(w) \right\}$$

the active indices at w and by

$$I(w) = \left\{ j \ : \ \Phi^j(w) \neq \Phi(w) \right\}$$

the set of inactive indices at w.

By continuity, $I(w) = I(w^*)$ in a neighborhood U of w^*. Now consider the following

ALGORITHM 3.2 (Generalized Newton's method for piecewise C^1 equations).

0. Chose w^0 (sufficiently close to w^*).

For $k = 0, 1, 2, \ldots$:

1. Choose $M_k \in \left\{ (\Phi^j)'(w^k) \ : \ j \in A(w^k) \right\}$ and solve

$$M_k s^k = -\Phi(w^k).$$

2. Set $w^{k+1} = w^k + s^k$.

For w^k close to w^*, we have $A(w^k) \subset A(w^*)$ and thus s^k is the Newton step for the C^1 equation

$$\Phi^{j_k}(w) = 0,$$

where $j_k \in A(w^k) \subset A(w^*)$ is the active index with $M_k = (\Phi^{j_k})'(w^k)$.

Therefore, if all the finitely many Newton processes for

$$\Phi^j(w) = 0, \quad j \in A(w^*)$$

converge locally fast, our generalized Newton's method converges locally fast, too. In particular, this is the case if f is C^2 and all $(\Phi^j)'(w^*)$, $j \in A(w^*)$, are invertible.

3.2.2. *SQP methods.* A further appealing idea is to obtain an iterative method by linearizing ∇f in VI($\nabla f, S$) about the current iterate $w^k \in S$:

$$w \in S, \quad (\nabla f(w^k) + \nabla^2 f(w^k)(w - w^k))^T (v - w) \geq 0 \quad \forall\, v \in S.$$

The solution w^{k+1} of this VI is then the new iterate. The resulting method, of course, can just as well be formulated for general variational inequalities VI(F, S) with C^1-function $F : \mathbb{R}^n \to \mathbb{R}^n$. We obtain the following method:

ALGORITHM 3.3 (Josephy-Newton method for VI(F, S)).

0. Choose $w^0 \in S$ (sufficiently close to the solution w^* of VI(F, S)).

For $k = 0, 1, 2 \ldots$

1. STOP if w^k solves VI(F, S) (holds if $w^k = w^{k-1}$).

2. Compute the solution w^{k+1} of

$$\begin{aligned} &\text{VI}(F(w^k) + F'(w^k)(\cdot - w^k), S) : \\ &w \in S, \quad (F(w^k) + F'(w^k)(w - w^k))^T (v - w) \geq 0 \quad \forall\, v \in S \end{aligned}$$

that is closest to w^k.

It is easily seen that VI($F(w^k) + F'(w^k)(\cdot - w^k), S$) are the first-order optimality conditions of the problem

$$\min_{w \in \mathbb{R}^n} \nabla f(w^k)^T (w - w^k) + \frac{1}{2}(w - w^k)^T \nabla^2 f(w^k)(w - w^k) \quad \text{s.t.} \quad w \in S.$$

The objective function is quadratic, and in the case of box constraints, we have a box-constrained quadratic program.

This is why this approach is called sequential quadratic programming.

ALGORITHM 3.4 (Sequential Quadratic Programming for simple constraints).

0. Chose $w^0 \in \mathbb{R}^n$ (sufficiently close to w^*).

For $k = 0, 1, 2, \ldots$:

1. *Compute the first-order optimal point s^k of the QP*

$$\min_{s\in\mathbb{R}^n} \nabla f(w^k)^T s + \frac{1}{2} s^T \nabla^2 f(w^k) s \quad \textit{s.t.} \quad w^k + s \in S$$

that is closest to 0.

2. *Set* $w^{k+1} = w^k + s^k$.

The local convergence analysis of the Josephy-Newton method is intimately connected with the locally unique and Lipschitz-stable solvability of the parametrized VI

$$\begin{aligned} &\mathrm{VI}(F(w^*) + F'(w^*)(\cdot - w^*) - p, S): \\ &w \in S, \quad (F(w^*) + F'(w^*)(w - w^*) - p)^T (v - w) \geq 0 \quad \forall\, v \in S. \end{aligned}$$

In fact, if there exist open neighborhoods $U_p \subset \mathbb{R}^n$ of 0, $U_w \subset \mathbb{R}^n$ of w^*, and a Lipschitz continuous function $U_p \ni p \mapsto w(p) \in U_w$ such that $w(p)$ is the unique solution of $\mathrm{VI}(F(w^*)+F'(w^*)(\cdot - w^*) - p, S)$ in U_w, then $\mathrm{VI}(F, S)$ is called *strongly regular* at w^*.

As we will see, strong regularity implies local q-superlinear convergence of the above SQP method if f is C^2.

In the case $S = \mathbb{R}^n$ we have

$$\mathrm{VI}(F, \mathbb{R}^n): \quad F(w) = 0$$

Hence, the Josephy-Newton method for $\mathrm{VI}(F, \mathbb{R}^n)$ is Newton's method for $F(w) = 0$. Furthermore, from

$$\mathrm{VI}(F(w^*) + F'(w^*)(\cdot - w^*) + p, \mathbb{R}^n): \quad F(w^*) + F'(w^*)(w - w^*) + p = 0$$

we see that in this case strong regularity is the same as the invertibility of $F'(w^*)$.

3.3. General inequality constraints. We now consider general nonlinear optimization in $\mathbb{R}^n$:

$$\min_{w\in\mathbb{R}^n} f(w) \quad \text{s.t.} \quad E(w) = 0, \quad C(w) \leq 0, \tag{3.5}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $E : \mathbb{R}^n \to \mathbb{R}^p$, and $C : \mathbb{R}^n \to \mathbb{R}^m$ are C^2 and $\leq$ is meant component-wise.

Denote by

$$L(w, \lambda, \mu) = f(w) + \lambda^T C(w) + \mu^T E(w)$$

the Lagrange function of problem (3.5).

Under a CQ, the first-order optimality conditions (KKT conditions) hold at (w^*, λ^*, μ^*):

$$\begin{aligned} &\nabla_w L(w, \lambda, \mu) = \nabla f(w) + C'(w)^T \lambda + E'(w)^T \mu = 0, \\ &\lambda \geq 0, \quad \nabla_\lambda L(w, \lambda, \mu)^T (z - \lambda) = C(w)^T (z - \lambda) \leq 0 \quad \forall\, z \geq 0, \\ &\nabla_\mu L(w, \lambda, \mu) = E(w) = 0. \end{aligned} \tag{3.6}$$

REMARK 3.5.

a) An easy way to remember these conditions is the following: (w^*, λ^*, μ^*) is a first-order saddle point of L on $\mathbb{R}^n \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$.

b) The second equation can be equivalently written in the following way:

$$\lambda \geq 0, \quad C(w) \leq 0, \quad C(w)^T\lambda = 0.$$

The KKT system consists of two equations and the variational inequality VI($-C(w), \mathbb{R}^m_+$). This is a VI w.r.t. λ that is parametrized by w. Also, since equations are special cases of variational inequalities, we have that (3.6) is in fact the same as VI($-\nabla L, \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$).

We now can use the same techniques as for simple constraints.

3.3.1. *Nonsmooth reformulation approach and generalized Newton methods.* Using the projection, we rewrite the VI in (3.6) as a nonsmooth equation:

$$\Phi(w, \lambda) := \lambda - P_{\mathbb{R}^m_+}(\lambda + \theta C(w)) = 0.$$

The reformulated KKT system

$$G(w, \lambda, \mu) := \begin{pmatrix} \nabla f(w) + C'(w)^T\lambda + E'(w)^T\mu \\ \Phi(w, \lambda) \\ E(w) \end{pmatrix} = 0$$

is a system of $n + m + p$ equations in $n + m + p$ unknowns.

The function on the left is C^1, except for the second row which is piecewise C^1. Therefore, the generalized Newton's method for piecewise smooth equations (Alg. 3.2) can be applied. It is q-superlinearly convergent if $(G^j)'(w^*, \lambda^*, \mu^*)$ is invertible for all active indices $j \in A(w^*, \lambda^*, \mu^*)$.

3.3.2. *SQP methods.* As already observed, the KKT system is identical to VI($-\nabla L, \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$).

The SQP method for (3.5) can now be derived as in the simply constrained case by linearizing $-\nabla L$ about the current iterate $x^k := (w^k, \lambda^k, \mu^k)$: The resulting subproblem is VI($-\nabla L(x^k) - \nabla L(x^k)(\cdot - x^k), \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$), or, in detail:

$$\begin{aligned} \nabla_w L(x^k) + \nabla^2_{wx} L(x^k)(x - x^k) &= 0 \\ \lambda \geq 0, \quad (C(w_k) + C'(w^k)(w - w^k))^T(z - \lambda) &\leq 0 \quad \forall\, z \geq 0, \\ E(w^k) + E'(w^k)(w - w^k) &= 0. \end{aligned} \tag{3.7}$$

As in the simply constrained case, it is straightforward to verify that (3.7) is equivalent to the KKT conditions of the following quadratic program:

$$\begin{aligned} &\min_w \nabla f(w^k)^T(w - w^k) + \frac{1}{2}(w - w^k)^T\nabla^2_{ww} L(x^k)(w - w^k) \\ &\text{s.t. } E(w^k) + E'(w^k)(w - w^k) = 0, \quad C(w_k) + C'(w^k)(w - w^k) \leq 0. \end{aligned}$$

4. Generalized Newton methods

We have seen in the previous section that we can reformulate KKT systems of finite dimensional optimization problems as nonsmooth equations. This also holds true for PDE-constrained optimization with inequality constraints, as we will sketch below. In finite dimensions, we observed that a projection-based reformulation results in a piecewise C^1-function to which a Newton-type method can be applied. In order to develop similar approaches in a function space framework, it is important to find minimum requirements on the operator $G : X \to Y$ that allow us to develop and analyze a Newton-type method for the (possibly nonsmooth) operator equation

$$G(x) = 0. \tag{4.1}$$

4.1. Motivation: Application to optimal control. We will show now that the optimality conditions of constrained optimal control problems can be converted to nonsmooth operator equations.

Consider the following elliptic optimal control problem:

$$\min_{y\in H_0^1, u\in L^2} J(y,u) \stackrel{\text{def}}{=} \|y - y_d\|_{L^2}^2 + \frac{\gamma}{2}\|u\|_{L^2}^2 \quad \text{s.t.} \quad Ay = u,\ \alpha \le u \le \beta.$$

Here, $y \in H_0^1(\Omega)$ is the state, which is defined on the open bounded domain $\Omega \subset \mathbb{R}^n$, and $u \in L^2(\Omega)$ is the control. Furthermore, $A : H_0^1(\Omega) \to H^{-1}(\Omega) = H_0^1(\Omega)^*$ is a (for simplicity) linear elliptic partial differential operator, e.g., $A = -\Delta$.

The control is subject to pointwise bounds $\alpha < \beta$. The objective is to drive the state as close to $y_d \in L^2(\Omega)$ as possible. The second part penalizes excessive control costs; the parameter $\gamma > 0$ is typically small.

We eliminate the state y via the state equation, i.e., $y = y(u) = A^{-1}u$, and obtain the reduced problem

$$\min_{u\in L^2} f(u) \stackrel{\text{def}}{=} J(y(u),u) \stackrel{\text{def}}{=} \frac{1}{2}\|A^{-1}u - y_d\|_{L^2}^2 + \frac{\gamma}{2}\|u\|_{L^2}^2 \quad \text{s.t.} \quad \alpha \le u \le \beta.$$

The feasible set is

$$S = \big\{u \in L^2(\Omega) \,:\, \alpha \le u \le \beta\big\}$$

and the optimality conditions are given by

$$\text{VI}(\nabla f, S): \quad u \in S, \quad (\nabla f(u), v-u)_{L^2} \ge 0 \quad \forall\, v \in S.$$

Using the projection $P_S(u) = P_{[\alpha,\beta]}(u(\cdot))$ onto S, this can be rewritten as

$$\Phi(u) \stackrel{\text{def}}{=} u - P_{[\alpha,\beta]}(u - \theta\nabla f(u)) = 0,$$

where $\theta > 0$ is fixed. This is a nonsmooth operator equation in the space $L^2(\Omega)$. Hence, we were able to convert the optimality system into a nonsmooth operator equation.

4.2. A general superlinear convergence result. Consider the operator equation (4.1) with $G : X \to Y$, X, Y Banach spaces.

A general Newton-type method for (4.1) has the form

ALGORITHM 4.1 (Generalized Newton's method).

0. Choose $x^0 \in X$ (sufficiently close to the solution x^*.)

For $k = 0, 1, 2, \ldots$:

1. Choose an invertible operator $M_k \in \mathcal{L}(X, Y)$.

2. Obtain s^k by solving

$$M_k s = -G(x^k), \tag{4.2}$$

and set $x^{k+1} = x^k + s^k$.

We now investigate the generated sequence (x^k) in a neighborhood of a solution $x^* \in X$, i.e., $G(x^*) = 0$.

For the distance $d^k := x^k - x^*$ to the solution we have

$$M_k d^{k+1} = M_k(x^{k+1} - x^*) = M_k(x^k + s^k - x^*) = M_k d^k - G(x^k) = G(x^*) + M_k d^k - G(x^k).$$

Hence, we obtain:

1. (x^k) converges q-linearly to x^* with rate $\gamma \in (0, 1)$ iff

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \le \gamma \|d^k\|_X \quad \forall\, k \text{ with } \|d_k\|_X \text{ sufficiently small.} \tag{4.3}$$

2. (x^k) converges q-superlinearly to x^* iff

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = o(\|d^k\|_X) \quad \text{for } \|d_k\|_X \to 0. \tag{4.4}$$

3. (x^k) convergences with q-order $1 + \alpha > 1$ iff

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d_k\|_X \to 0. \tag{4.5}$$

In 1., the esimate is meant uniformly in k, i.e., there exists $\delta_\gamma > 0$ such that

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \le \gamma \|d^k\|_X \quad \forall\, k \text{ with } \|d_k\|_X < \delta_\gamma.$$

In 2., $o(\|d^k\|_X)$ is meant uniformly in k, i.e., for all $\eta \in (0, 1)$, there exists $\delta_\eta > 0$ such that

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \le \eta \|d^k\|_X \quad \forall\, k \text{ with } \|d_k\|_X < \delta_\eta.$$

The condition in 3. and those stated below are meant similarly.

It is convenient, and often done, to split the smallness assumption on

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X$$

in two parts:

1. **Regularity condition**:

$$\|M_k^{-1}\|_{Y,X} \leq C \quad \forall\, k \geq 0. \tag{4.6}$$

2. **Approximation condition**:

$$\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X = o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \to 0. \tag{4.7}$$

or

$$\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d^k\|_X \to 0. \tag{4.8}$$

We obtain

THEOREM 4.2. *Consider the operator equation* (4.1) *with $G : X \to Y$, where X and Y are Banach spaces. Let (x^k) be generated by the generalized Newton method (Alg.* 4.1*). Then:*

1. *If x^0 is sufficiently close to x^* and* (4.3) *holds then $x^k \to x^*$ q-linearly with rate γ.*
2. *If x^0 is sufficiently close to x^* and* (4.4) *(or* (4.6) *and* (4.7)*) holds then $x^k \to x^*$ q-superlinearly.*
3. *If x^0 is sufficiently close to x^* and* (4.5) *holds (or* (4.6) *and* (4.8)*) then $x^k \to x^*$ q-superlinearly with order $1 + \alpha$.*

Proof. 1. Let $\delta > 0$ be so small that (4.3) holds for all x^k with $\|d^k\|_X < \delta$. Then, for x^0 satisfying $\|x^0 - x^*\|_X < \delta$, we have

$$\begin{aligned}\|x^1 - x^*\|_X = \|d^1\|_X &= \|M_0^{-1}(G(x^* + d^0) - G(x^*) - M_0 d^0)\|_X \leq \gamma \|d^0\|_X \\ &= \gamma \|x^0 - x^*\|_X < \delta.\end{aligned}$$

Inductively, let $\|x^k - x^*\|_X < \delta$. Then

$$\begin{aligned}\|x^{k+1} - x^*\|_X = \|d^{k+1}\|_X &= \|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \\ &\leq \gamma \|d^k\|_X = \gamma \|x^k - x^*\|_X < \delta.\end{aligned}$$

Hence, we have

$$\|x^{k+1} - x^*\|_X \leq \gamma \|x^k - x^*\|_X \quad \forall\, k \geq 0.$$

2. Fix $\gamma \in (0, 1)$ and let $\delta > 0$ be so small that (4.3) holds for all x^k with $\|d^k\|_X < \delta$. Then, for x^0 satisfying $\|x^0 - x^*\|_X < \delta$, we can apply 1. to conclude $x^k \to x^*$ with rate γ.

Now, (4.4) immediately yields

$$\begin{aligned}\|x^{k+1} - x^*\|_X = \|d^{k+1}\|_X &= \|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = o(\|d^k\|_X) \\ &= o(\|x^k - x^*\|_X) \quad (k \to \infty).\end{aligned}$$

3. As in 2, but now

$$\begin{aligned}\|x^{k+1} - x^*\|_X = \|d^{k+1}\|_X &= \|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = O(\|d^k\|_X^{1+\alpha}) \\ &= O(\|x^k - x^*\|_X^{1+\alpha}) \quad (k \to \infty).\end{aligned}$$

□

We emphasize that an inexact solution of the Newton system (4.2) can be interpreted as a solution of the same system, but with M_k replaced by a perturbed operator $\tilde{M}_k$. Since the condition (4.4) (or the conditions (4.6) and (4.7)) remain valid if M_k is replaced by a perturbed operator $\tilde{M}_k$ and the perturbation is sufficiently small, we see that the fast convergence of the generalized Newton's method is not affected if the system is solved inexactly and the accuracy of the solution is controlled suitably. The Dennis-Moré condition [**25**] characterizes perturbations that are possible without destroying q-superlinear convergence.

We will now specialize on particular instances of generalized Newton methods. The first one, of course, is Newton's method itself.

4.3. The classical Newton's method. In the classical Newton's method, we assume that G is continuously F-differentiable and choose $M_k = G'(x^k)$.

The regularity condition then reads

$$\|G'(x^k)^{-1}\|_{Y,X} \leq C \quad \forall\, k \geq 0.$$

By Banach's Lemma (asserting continuity of $M \mapsto M^{-1}$), this holds true if G' is continuous at x^* and

$$G'(x^*) \in \mathcal{L}(X,Y) \text{ is continuously invertible.}$$

This condition is the textbook regularity requirement in the analysis of Newton's method.

Fréchet differentiability at x^* means

$$\|G(x^* + d^k) - G(x^*) - G'(x^*)d^k\|_X = o(\|d^k\|_X).$$

Now, due to the continuity of G',

$$\begin{aligned}\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X &= \|G(x^* + d^k) - G(x^*) - G'(x^* + d^k)d^k\|_X \\ &\leq \|G(x^* + d^k) - G(x^*) - G'(x^*)d^k\|_X + \|(G'(x^*) - G'(x^* + d^k))d^k\|_X \\ &= o(\|d^k\|_X) + \|G'(x^*) - G'(x^* + d^k)\|_{X,Y}\|d^k\|_X = o(\|d^k\|_X).\end{aligned}$$

Therefore, we have proved the superlinear approximation condition.

If G' is α-order Hölder continuous near x^*, we even obtain the approximation condition of order $1+\alpha$. In fact, let $L > 0$ be the modulus of Hölder continuity. Then

$$\begin{aligned}
\|G(x^* + d^k) - G(x^*) - M_k d^k\|_Y &= \|G(x^* + d^k) - G(x^*) - G'(x^* + d^k)d^k\|_Y \\
&= \left\| \int_0^1 (G'(x^* + td^k) - G'(x^* + d^k))d_k \, dt \right\|_Y \\
&\le \int_0^1 \|G'(x^* + td^k) - G'(x^* + d^k)\|_{X,Y} \, dt \, \|d^k\|_X \\
&\le L \int_0^1 (1-t)^\alpha \|d^k\|_X^\alpha \, dt \, \|d^k\|_X \\
&= \frac{L}{1+\alpha} \|d^k\|_X^{1+\alpha} = O(\|d^k\|_X^{1+\alpha}).
\end{aligned}$$

Summarizing, we have proved the following

COROLLARY 4.3. *Let $G : X \to Y$ be a continuously F-differentiable operator between Banach spaces and assume that $G'(x^*)$ is continuously invertible at the solution x^*. Then Newton's method (i.e., Alg.* 4.1 *with $M_k = G'(x^k)$ for all k) converges locally q-superlinearly. If, in addition, G' is α-order Hölder continuous near x^*, the order of convergence is $1 + \alpha$.*

REMARK 4.4. *The choice of M_k in the ordinary Newton's method, $M_k = G'(x^k)$, is* point-based*, since it depends on the point x^k.*

4.4. Generalized differential and semismoothness. If G is nonsmooth, the question arises if a suitable substitute for G' can be found. We follow [**75, 77**] here; a related approach can be found in [**50**]. Thinking at subgradients of convex functions, which are set-valued, we consider set-valued generalized differentials $\partial G : X \rightrightarrows Y$. Then we will choose M_k point-based, i.e.,

$$M_k \in \partial G(x^k).$$

If we want every such choice M_k to satisfy the superlinear approximation condition, then we have to require

$$\sup_{M \in \partial G(x^* + d)} \|G(x^* + d) - G(x^*) - Md\|_X = o(\|d\|_X) \quad \text{for } \|d\|_X \to 0.$$

This approximation property is called semismoothness [**75, 77**]:

DEFINITION 4.5 (Semismoothness). *Let $G : X \to Y$ be a continuous operator between Banach spaces. Furthermore, let be given the set-valued mapping $\partial G : X \rightrightarrows Y$ with nonempty images (which we will call generalized differential in the sequel). Then*

a) G is called ∂G-semismooth at $x \in X$ if

$$\sup_{M \in \partial G(x+d)} \|G(x + d) - G(x) - Md\|_X = o(\|d\|_X) \quad \textit{for } \|d\|_X \to 0.$$

b) G is called ∂G-semismooth of order $\alpha > 0$ at $x \in X$ if

$$\sup_{M \in \partial G(x+d)} \|G(x+d) - G(x) - Md\|_X = O(\|d\|_X^{1+\alpha}) \quad \textit{for } \|d\|_X \to 0.$$

LEMMA 4.6. *If $G : X \to Y$ is continuously F-differentiable near x, then G is $\{G'\}$-semismooth at x. Furthermore, if G' is α-order Hölder continuous near x, then G is $\{G'\}$-semismooth at x of order α.*

Proof.

$$\begin{aligned}\|G(x+d) - G(x) - G'(x+d)d\|_Y \leq \\ \leq \|G(x+d) - G(x) - G'(x)d\|_Y + \|G'(x)d - G'(x+d)d\|_Y \\ \leq o(\|d\|_X) + \|G'(x) - G'(x+d)\|_{X,Y}\|d\|_X = o(\|d\|_X).\end{aligned}$$

Here, we have used the definition of F-differentiablity and the continuity of G'.

In the case of α-order Hölder continuity we have to work a little bit more:

$$\begin{aligned}\|G(x+d) - G(x) - G'(x+d)d\|_Y &= \left\| \int_0^1 (G'(x+td) - G'(x+d))d\,dt \right\|_Y \\ \leq \int_0^1 \|G'(x+td) - G'(x+d)\|_{X,Y}\,dt\,\|d\|_X &\leq \int_0^1 L(1-t)^\alpha \|d\|_X^\alpha\,dt\,\|d\|_X \\ = \frac{L}{1+\alpha}\|d\|_X^{1+\alpha} &= O(\|d\|_X^{1+\alpha}).\end{aligned}$$

□

EXAMPLE 4.7. *For locally Lipschitz-continuous functions $G : \mathbb{R}^n \to \mathbb{R}^m$, the standard choice for ∂G is Clarke's generalized Jacobian:*

$$\partial^{cl} G(x) = \textit{conv}\left\{M \,:\, x^k \to x,\ G'(x^k) \to M,\ G \textit{ differentiable at } x^k\right\}. \tag{4.9}$$

This definition is justified since G' exists almost everywhere on $\mathbb{R}^n$ by Rademacher's theorem (which is a deep result).

REMARK 4.8. *The classical definition of semismoothness for functions $G : \mathbb{R}^n \to \mathbb{R}^m$* **[59, 64]** *is equivalent to $\partial^{cl}G$-semismoothness, where $\partial^{cl}G$ is Clarke's generalized Jacobian defined in* (4.9)*, in connection with directional differentiability of G.*

Next, we give a concrete example of a semismooth function:

EXAMPLE 4.9. *Consider $\psi : \mathbb{R} \to \mathbb{R}$, $\psi(x) = P_{[a,b]}(x)$, then Clarke's generalized derivative is*

$$\partial^{cl}\psi(x) = \begin{cases} 0 & x < a \textit{ or } x > b, \\ 1 & a < x < b, \\ \textit{conv}\{0,1\} = [0,1] & x = a \textit{ or } x = b. \end{cases}$$

The $\partial^{cl}\psi$-semismoothness of ψ can be shown easily:

For all $x \notin \{a, b\}$ we have that ψ is continuously differentiable in a neighborhood of x with $\partial^{cl}\psi \equiv \{\psi'\}$. Hence, by Lemma 4.6*, ψ is $\partial^{cl}\psi$-semismooth at x.*

For $x = a$, we estimate explicitly: For small $d > 0$, we have $\partial^{cl}\psi(x) = \{\psi'(a + d)\} = \{1\}$ and thus

$$\sup_{M \in \partial^{cl}\psi(x+d)} |\psi(x + d) - \psi(x) - Md| = a + d - a - 1 \cdot d = 0.$$

For small $d < 0$, we have $\partial^{cl}\psi(x) = \{\psi'(a + d)\} = \{0\}$ and thus

$$\sup_{M \in \partial^{cl}\psi(x+d)} |\psi(x + d) - \psi(x) - Md| = a - a - 0 \cdot d = 0.$$

Hence, the semismoothness of ψ at $x = a$ is proved.

For $x = b$ we can do exactly the same.

The class of semismooth operators is closed with respect to a wide class of operations, see [**75**]:

THEOREM 4.10. *Let X, Y, Z, X_i, Y_i be Banach spaces.*

a) If the operators $G_i : X \to Y_i$ are ∂G_i-semismooth at x then (G_1, G_2) is $(\partial G_1, \partial G_2)$-semismooth at x.

b) If $G_i : X \to Y$, $i = 1, 2$, are ∂G_i-semismooth at x then $G_1 + G_2$ is $(\partial G_1 + \partial G_2)$-semismooth at x.

c) Let $G_1 : Y \to Z$ and $G_2 : X \to Y$ be ∂G_i-semismooth at $G_2(x)$ and x, respectively. Assume that ∂G_1 is bounded near $y = G_2(x)$ and that G_2 is Lipschitz continuous near x. Then $G = G_1 \circ G_2$ is ∂G-semismooth with

$$\partial G(x) = \{M_1 M_2 \,:\, M_1 \in \partial G_1(G_2(x)),\ M_2 \in \partial G_2(x)\}.$$

Proof. Parts a) and b) are straightforward to prove.

Part c):

Let $y = G_2(x)$ and consider $d \in X$. Let $h(d) = G_2(x + d) - y$. Then

$$\|h(d)\|_Y = \|G_2(x + d) - G_2(x)\|_Y \le L_2 \|d\|_Y.$$

Hence, for $M_1 \in \partial G_1(G_2(x + d))$ and $M_2 \in \partial G_2(x + d)$, we obtain

$$\begin{aligned}
&\|G_1(G_2(x + d)) - G_1(G_2(x)) - M_1 M_2 d\|_Z = \\
&\quad = \|G_1(y + h(d)) - G_1(y) - M_1 h(d) + M_1(G_2(x + d) - G_2(x) - M_2 d)\|_Z \\
&\quad \le \|G_1(y + h(d)) - G_1(y) - M_1 h(d)\|_Y + \|M_1\|_{Y,Z} \|G_2(x + d) - G_2(x) - M_2 d\|_Y
\end{aligned}$$

By assumption, there exists C with $\|M_1\|_{Y,Z} \leq C$. Taking the supremum with respect to M_1, M_2 and using the semismoothness gives

$$\begin{aligned}
&\sup_{M \in \partial G(x+d)} \|G(x+d) - G(x) - Md\|_Z \\
&\quad \leq \sup_{M_1 \in \partial G_1(y+h(d))} \|G_1(y+h(d)) - G_1(y) - M_1 h(d)\|_Y \\
&\quad + C \sup_{M_2 \in \partial G_2(x+d)} \|G_2(x+d) - G_2(x) - M_2 d\|_Y \\
&\quad = o(\|h(d)\|_Y) + o(\|d\|_X) = o(\|d\|_X).
\end{aligned}$$

□

4.5. Semismooth Newton methods. The semismoothness concept ensures the approximation property required for generalized Newton methods. In addition, we need a regularity condition, which can be formulated as follows:

There exist constants $C > 0$ and $\delta > 0$ such that

$$\|M^{-1}\|_{Y,X} \leq C \quad \forall\, M \in \partial G(x)\ \forall\, x \in X,\ \|x - x^*\|_X < \delta. \tag{4.10}$$

Under these two assumptions, the following generalized Newton method for semismooth operator equations is q-superlinearly convergent:

ALGORITHM 4.11 (Semismooth Newton's method).

0. Choose $x^0 \in X$ (sufficiently close to the solution x^*.)

For $k = 0, 1, 2, \ldots$:

1. Choose $M_k \in \partial G(x^k)$.

2. Obtain s^k by solving

$$M_k s = -G(x^k),$$

and set $x^{k+1} = x^k + s^k$.

The local convergence result is a simple corollary of Theorem 4.2:

THEOREM 4.12. *Let $G : X \to Y$ be continuous and ∂G-semismooth at a solution x^* of* (4.1)*. Furthermore, assume that the regularity condition* (4.10) *holds. Then there exists $\delta > 0$ such that for all $x^0 \in X$, $\|x^0 - x^*\|_X < \delta$, the semismooth Newton method (Alg.* 4.11*) converges q-superlinearly to x^*.*

If G is ∂G-semismooth of order $\alpha > 0$ at x^*, then the convergence is of order $1 + \alpha$.

Proof. The regularity condition (4.10) implies (4.6) as long as x^k is close enough to x^*. Furthermore, the semismoothness of G at x^* ensures the q-superlinear approximation property (4.7).

In the case of α-order semismoothness, the approximation property with order $1 + \alpha$ holds.

Therefore, Theorem 4.2 yields the assertions. □

4.5.1. *Semismooth Newton method for finite dimensional KKT systems.* At the beginning of this chapter we have seen that we can rewrite the KKT conditions of the NLP

$$\min f(w) \quad \text{s.t.} \quad E(w) = 0, \quad C(w) \leq 0$$

in the following form:

$$G(x) \stackrel{\text{def}}{=} \begin{pmatrix} \nabla_w L(w, \lambda, \mu) \\ \lambda - P_{\mathbb{R}^p_+}(\lambda + C(w)) \\ E(w) \end{pmatrix} = 0,$$

where we have set $x = (w, \lambda, \mu)$. With the developed results, we now can show that the function G on the left is semismooth. In fact, $\nabla_w L$ is $\{\nabla^2_{wx} L\}$-semismooth and E is E'-semismooth.

Furthermore, as shown above, $\psi(t) = P_{\mathbb{R}_+}(t)$ is $\partial^{cl}\psi$-semismooth with

$$\partial^{cl}\psi(t) = 0 \ (t < 0), \quad \partial^{cl}\psi(t) = 1 \ (t > 0), \quad \partial^{cl}\psi(0) = [0, 1].$$

Hence, by the sum and chain rules from Theorem 4.10

$$\phi_i(w, \lambda_i) \stackrel{\text{def}}{=} \lambda_i - P_{\mathbb{R}_+}(\lambda_i + C_i(w)),$$

is semismooth with respect to

$$\partial\phi_i(w, \lambda_i) := \left\{(-g_i C_i'(w), 1 - g_i) \ : \ g_i \in \partial^{cl}\psi(\lambda_i + C_i(w))\right\}.$$

Therefore, the function $\Phi(w, \lambda) = \lambda - P_{\mathbb{R}^p_+}(\lambda + C(w))$ is semismooth with respect to

$$\partial\Psi(w, \lambda) := \left\{(-D_g C_i'(w), I - D_g) \ : \ D_g = \text{diag}(g_i), \ g_i \in \partial^{cl}\psi(\lambda_i + C_i(w))\right\}.$$

This shows that G is semismooth with respect to

$$\partial G(x) \stackrel{\text{def}}{=} \left\{ \begin{pmatrix} \nabla^2_{ww} L(x) & C'(w)^T & E'(w)^T \\ -D_g C(w)' & I - D_g & 0 \\ E'(w) & 0 & 0 \end{pmatrix} \ ; \ D_g = \text{diag}(g_i), \ g_i \in \partial^{cl}\psi(\lambda_i + C_i(w)) \right\}.$$

Under the regularity condition

$$\|M^{-1}\| \leq C \quad \forall\, M \in \partial G(x) \quad \forall\, x, \ \|x - x^*\| < \delta,$$

where $x^* = (w^*, \lambda^*, \mu^*)$ is a KKT triple, Theorem 4.12 is applicable and yields the q-superlinear convergence of the semismooth Newton method.

REMARK 4.13. *The compact-valuedness and the upper semicontinuity of Clarke's generalized differential* [**21**] *even allows to reduce the regularity condition to*

$$\|M^{-1}\| \leq C \quad \forall\, M \in \partial G(x^*).$$

REMARK 4.14. *We also can view G as a piecewise smooth equation and apply Algorithm* 3.2*. In fact, it can be shown that Clarke's generalized Jacobian is the convex hull of the Jacobians of all essentially active pieces* [**70, 75**]*. We are not going into details here.*

4.5.2. *Discussion.* So far, we have looked at semismooth Newton methods from an abstract point of view. The main point, however, is to prove semismoothness for concrete instances of nonsmooth operators. In particular, we aim at reformulating KKT systems arising in PDE-constrained optimization in the same way as we did this in finite dimensions in the above section. We will investigate this in detail in the next section 5.

It should be mentioned that the class of semismooth Newton method includes as a special case the *primal dual active set strategy*, a modern approach found in the literature, see [**9, 39**].

5. Semismooth Newton methods in function spaces

In the finite dimensional setting we have shown that variational inequalities and complementarity conditions can be reformulated as nonsmooth equations. We also described how generalized Newton methods can be developed that solve these nonsmooth equations.

In section 4.5 we introduced the concept of semismoothness for nonsmooth operators and developed superlinearly convergent generalized Newton methods for semismooth operator equations. We now will show that, similar to the finite dimensional case, it is possible to reformulate variational inequalities and complementarity conditions in function space.

5.1. Pointwise bound constraints in L^2. Let $\Omega \subset \mathbb{R}^n$ be measurable with $0 < |\Omega| < \infty$. We consider the problem

$$\min_{u \in L^2(\Omega)} f(u) \quad a \le u \le b \quad \text{a.e. on } \Omega$$

with $f : L^2(\Omega) \to \mathbb{R}$ twice continuously F-differentiable. We can admit unilateral constraints ($a \le u$ or $u \le b$) just as well. To avoid distinguishing cases, we will focus on the bilateral case $a, b \in L^\infty(\Omega)$, $b - a \ge \nu > 0$ on Ω. We also could consider problems in L^p, $p \ne 2$. However, for the sake of compact presentation, we focus on the case $p = 2$, which is the most important situation.

It is convenient to transform the bounds to constant bounds, e.g., via

$$u \mapsto \frac{u - a}{b - a}.$$

Hence, we will consider the problem

$$\min_{u \in L^2(\Omega)} f(u) \quad \alpha \le u \le \beta \quad \text{a.e. on } \Omega \tag{5.1}$$

with constants $\alpha < \beta$. Let $U = L^2(\Omega)$ and $S = \{u \in L^2(\Omega) \,:\, \alpha \le u \le \beta\}$. We choose the standard dual pairing $\langle \cdot, \cdot \rangle_{U^*,U} = (\cdot, \cdot)_{L^2}$ and then have $U^* = U = L^2(\Omega)$. The optimality conditions are

$$u \in S, \quad (\nabla f(u), v - u)_{L^2} \ge 0 \quad \forall\, v \in S.$$

We now use the projection P_S onto S, which is given by

$$P_S(v)(x) = P_{[\alpha,\beta]}(v(x)), \quad x \in \Omega.$$

Then the optimality conditions can be written as

$$\Phi(u) := u - P_S(u - \theta \nabla f(u)) = 0, \tag{5.2}$$

where $\theta > 0$ is arbitrary, but fixed. Note that, since P_S conincides with the pointwise projection onto $[\alpha, \beta]$, we have

$$\Phi(u)(x) = u(x) - P_{[\alpha,\beta]}(u(x) - \theta \nabla f(u)(x)).$$

Our aim now is to define a generalized differential $\partial\Phi$ for Φ in such a way that Φ is semismooth.

By the chain rule and sum rule that we developed, this reduces to the question how a suitable differential for the superposition $P_{[\alpha,\beta]}(v(\cdot))$ can be defined.

5.2. Semismoothness of superposition operators. More generally than the superposition operator in the previous subsection, we look at the superposition operator

$$\Psi : L^p(\Omega)^m \to L^q(\Omega), \quad \Psi(w)(x) = \psi(w_1(x), \dots, w_m(x)).$$

with $1 \le q \le p \le \infty$.

Here, $\psi : \mathbb{R}^m \to \mathbb{R}$ is assumed to be Lipschitz continuous. Since we aim at semismoothness of Ψ, it is more than natural to assume semismoothness of ψ. As differential we choose Clarke's generalized differential $\partial^{cl}\psi$. Now it is reasonable to define $\partial\Psi$ in such a way that, for all $M \in \partial\Psi(w + d)$, the remainder

$$|(\Psi(u + d) - \Psi - Md)(x)| = |\psi(w(x) + d(x)) - \psi(w(x)) - (Md)(x)|$$

becomes pointwise small if $|d(x)|$ is small. By semismoothness of ψ, this, again, holds true if $(Md)(x) \in \partial^{cl}(w(x) + d(x))$ is satisified.

Hence, we define:

DEFINITION 5.1. *Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous and ($\partial^{cl}\psi$-) semismooth. For $1 \le q \le p \le \infty$, consider*

$$\Psi : L^p(\Omega)^m \to L^q(\Omega), \quad \Psi(w)(x) = \psi(w_1(x), \dots, w_m(x)).$$

We define the differential

$$\begin{aligned}
&\partial\Psi : L^p(\Omega)^m \rightrightarrows \mathcal{L}(L^p(\Omega), L^q(\Omega)),\\
&\partial\Psi(w) = \left\{M \,:\, Mw = g^T w,\ g \in L^\infty(\Omega)^m,\ g(x) \in \partial^{cl}\psi(w(x)) \text{ for a.a. } x \in \Omega\right\}.
\end{aligned}$$

The operator Φ in (5.2) is naturally defined as a mapping from L^2 to L^2. Therefore, since ∇f maps to L^2, we would like the superposition $v \mapsto P_{[\alpha,\beta]}(v(\cdot))$ to be semismooth from L^2 to L^2. But this is not true, as the following Lemma shows in great generality.

LEMMA 5.2. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be any locally Lipschitz continuous function that is* not *affine linear. Furthermore, let $\Omega \subset \mathbb{R}^n$ be nonempty, open and bounded. Then, for all $q \in [1, \infty)$, the operator*

$$\Psi : L^q(\Omega) \ni u \mapsto \psi(u(\cdot)) \in L^q(\Omega)$$

is not *$\partial\Psi$-semismooth.*

Proof. Fix $b \in \mathbb{R}$ and choose $g_b \in \partial\psi(b)$ Since ψ is not affine linear, there exists $a \in \mathbb{R}$ with

$$\psi(a) \neq \psi(b) + g_b(a-b).$$

Hence,

$$\rho := |\psi(b) - \psi(a) - g_b(b-a)| > 0.$$

Let $x_0 \in \Omega$ and $U_\varepsilon = (x_0 - h_\varepsilon, x_0 + h_\varepsilon)^n$, $h_\varepsilon = \varepsilon^{1/n}/2$. Define

$$u(x) = a, \quad x \in \Omega, \qquad d_\varepsilon(x) = \begin{cases} b-a & x \in U_\varepsilon, \\ 0 & x \notin U_\varepsilon. \end{cases}$$

Then

$$\|d_\varepsilon\|_{L^q} = \left(\int_\Omega |d_\varepsilon(x)|^q \, dx \right)^{1/q} = \left(\int_{U_\varepsilon} |b-a|^q dx \right)^{1/q} = \varepsilon^{1/q} |b-a|.$$

Choose some $g_a \in \partial\psi(a)$ and define

$$g_\varepsilon(x) = \begin{cases} g_b & x \in U_\varepsilon, \\ g_a & x \notin U_\varepsilon. \end{cases}$$

Then $M : L^q(\Omega) \ni v \mapsto g_\varepsilon \cdot v \in L^q(\Omega)$ is an element of $\partial\Psi(u + d_\varepsilon)$. Now, for all $x \in \Omega$,

$$|\psi(u(x) + d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x) d_\varepsilon(x)| = \begin{cases} |\psi(b) - \psi(a) - g_b(b-a)| = \rho > 0, & x \in U_\varepsilon, \\ |\psi(a) - \psi(a) - g_a(a-a)| = 0, & x \notin U_\varepsilon. \end{cases}$$

Therefore,

$$\begin{aligned} \|\Psi(u + d_\varepsilon) - \Psi(u) - M d_\varepsilon\|_{L^q} &= \left(\int_\Omega |\psi(u(x) + d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x) d_\varepsilon(x)|^q \, dx \right)^{1/q} \\ &= \left(\int_{U_\varepsilon} \rho^q \, dx \right)^{1/q} = \varepsilon^{1/q} \rho = \frac{\rho}{|b-a|} \|d_\varepsilon\|_{L^q}. \end{aligned}$$

□

Note that the trouble is not caused by the nonsmoothness of ψ, but by the nonlinearity of ψ.

Fortunately, Ulbrich [**75, 77**] proved a result that helps us. To formulate the result in its full generality, we extend our definition of generalized differentials to superposition operators of the form $\psi(G(\cdot))$, where G is a continuously F-differentiable operator.

DEFINITION 5.3. *Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz continuous and ($\partial^{cl}\psi$-) semismooth. Furthermore, let $1 \leq q \leq p \leq \infty$ be given, consider*

$$\Psi_G : L^p(\Omega)^m \to L^q(\Omega), \quad \Psi_G(y)(x) = \psi(G(y)(x)),$$

where $G : Y \to L^p(\Omega)^m$ is continuously F-differentiable and Y is a Banach space. We define the differential

(5.3)

$$\partial\Psi_G : Y \rightrightarrows \mathcal{L}(Y, L^q(\Omega)),$$

$$\partial\Psi_G(y) = \left\{ M \,:\, Mv = g^T(G'(y)v),\ g \in L^\infty(\Omega)^m,\ g(x) \in \partial^{cl}\psi(G(y)(x)) \text{ for a.a. } x \in \Omega \right\}.$$

Note that this is just the differential that we would obtain by the construction in part c) of Theorem 4.10.

Now we can state the following semismoothness result.

THEOREM 5.4. *Let $\Omega \subset \mathbb{R}^n$ be measurable with $0 < |\Omega| < \infty$. Furthermore, let $\psi : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous and semismooth. Let Y be a Banach space, $1 \le q < p \le \infty$, and assume that the operator $G : Y \to L^q(\Omega)^m$ is F-differentiable and that G maps Y locally Lipschitz continuously to $L^p(\Omega)$. Then, the operator*

$$\Psi_G : Y \to L^q(\Omega), \quad \Psi_G(y)(x) = \psi(G(y)(x)),$$

is $\partial\Psi_G$-semismooth, where $\partial\Psi_G$ is defined in (5.3)*.*

Addition: Under additional assumptions, the operator Ψ_G is $\partial\Psi_G$-semismooth of order $\alpha > 0$.

5.3. Pointwise bound constraints in L^2 revisited. We return to the operator Φ defined in (5.2). To be able to prove the semismoothness of $\Phi : L^2 \to L^2$ definied in (5.2), we thus need some kind of smoothing property of the mapping

$$u \mapsto u - \theta\nabla f(u).$$

Therefore, we assume that ∇f has the following structure:

$$\begin{aligned} &\text{There exist } \gamma > 0 \text{ and } p > 2 \text{ such that} \\ &\nabla f(u) = \gamma u + B(u), \\ &B : L^2(\Omega) \to L^2(\Omega) \text{ continuously F-differentiable,} \\ &B : L^2(\Omega) \to L^p(\Omega) \text{ locally Lipschitz continuous.} \end{aligned} \tag{5.4}$$

This structure is met by many optimal control problems, as illustrated in section 5.4.

If we now choose $\theta = 1/\gamma$, then we have

$$\Phi(u) = u - P_{[\alpha,\beta]}(u - (1/\gamma)(\gamma u + B(u))) = u - P_{[\alpha,\beta]}(-(1/\gamma)B(u)).$$

Therefore, we have achieved that the operator inside the projection satisfies the requirements of Theorem 5.4. We obtain:

THEOREM 5.5. *Consider the problem* (5.1) *with $\alpha < \beta$ and let $f : L^2(\Omega) \to L^2(\Omega)$ satisfy condition* (5.4)*. Then, for $\theta = 1/\gamma$, the operator Φ in the reformulated optimality conditions* (5.2) *is $\partial\Phi$-semismooth with*

$$\begin{aligned} &\partial\Phi : L^2(\Omega) \rightrightarrows \mathcal{L}(L^2(\Omega), L^2(\Omega)), \\ &\partial\Phi(u) = \Big\{ M \,;\; M = I + \frac{g}{\gamma} \cdot B'(u),\ g \in L^\infty(\Omega), \\ &\qquad\qquad g(x) \in \partial^{cl} P_{[\alpha,\beta]}(-1/\gamma B(u)(x)) \text{ for a.a. } x \in \Omega \Big\}. \end{aligned}$$

Here,

$$\partial P_{[\alpha,\beta]}(t) \begin{cases} 0 & t < \alpha \text{ or } t > \beta, \\ 1 & \alpha < t < \beta, \\ [0,1] & t = \alpha \text{ or } t = \beta. \end{cases}$$

Proof. Setting $q = 2$, $\psi = P_{[\alpha,\beta]}$ and $G = -(1/\gamma)B$, we can apply Theorem 5.4 and obtain that the operator $\Psi_G : L^2(\Omega) \to L^2(\Omega)$ is $\partial\Psi_G$-semismooth. Therefore, $\Phi = \gamma I + \Psi_G$ is $(\gamma I + \partial\Psi_G)$-semismooth by Theorem 4.10. Since $\partial\Phi = \gamma I + \partial\Psi_G$, the proof is complete. □

For the applicability of the semismooth Newton method (Alg. 4.11) we need, in addition, the following regularity condition:

$$\|M^{-1}\|_{L^2,L^2} \le C \quad \forall\, M \in \partial\Phi(u)\ \forall\, u \in L^2(\Omega),\ \|u - u^*\|_{L^2} < \delta.$$

Sufficient conditions for this regularity assumption in the flavor of second order sufficient optimality conditions can be found in [**76, 75**].

5.4. Application to optimal control. Consider the following elliptic optimal control problem:

$$\min_{y\in H_0^1, u\in L^2} J(y,u) \stackrel{\text{def}}{=} \|y - y_d\|_{L^2}^2 + \frac{\gamma}{2}\|u\|_{L^2}^2 \quad \text{s.t.} \quad Ay = r + Ru,\ \alpha \le u \le \beta. \tag{5.5}$$

Here, $y \in H_0^1(\Omega)$ is the state, which is defined on the open bounded domain $\Omega \subset \mathbb{R}^n$, and $u \in L^2(\Omega_c)$ is the control, which is defined on the open bounded domain $\Omega_c \subset \mathbb{R}^m$. Furthermore, $A : H_0^1(\Omega) \to H^{-1}(\Omega) = H_0^1(\Omega)^*$ is a (for simplicity) linear elliptic partial differential operator, e.g., $A = -\Delta$, and $r \in H^{-1}(\Omega)$ is given.

The control operator $R : L^{p'}(\Omega_c) \to H^{-1}(\Omega)$ is continuous and linear, with $p' \in [1, 2)$ (the reason why we do not choose $p' = 2$ here will become clear later; note however, that $L^2(\Omega_c)$ is continuously embedded in $L^{p'}(\Omega_c)$). For instance, distributed control on the whole domain Ω would correspond to the choice $\Omega_c = \Omega$ and $R : u \in L^{p'}(\Omega) \mapsto u \in H^{-1}(\Omega)$, where p' is chosen in such a way that $H_0^1(\Omega)$ is continuously embedded in the dual space $L^p(\Omega)$, $p = p'/(p'-1)$, of $L^{p'}(\Omega)$.

The control is subject to pointwise bounds $\alpha < \beta$. The objective is to drive the state as close to $y_d \in L^2(\Omega)$ as possible. The second part penalizes excessive control costs; the parameter $\gamma > 0$ is typically small.

We eliminate the state y via the state equation, i.e., $y = y(u) = A^{-1}(r + Ru)$, and obtain the reduced problem

$$\min_{u\in L^2} f(u) \stackrel{\text{def}}{=} J(y(u), u) \stackrel{\text{def}}{=} \frac{1}{2}\|y(u) - y_d\|_{L^2}^2 + \frac{\gamma}{2}\|u\|_{L^2}^2 \quad \text{s.t.} \quad \alpha \le u \le \beta.$$

This problem is of the form (5.1).

For the gradient we obtain

$$(\nabla f(u), d)_{L^2} = (y(u) - y_d, y'(u)d)_{L^2(\Omega)} + \gamma u, d)_{L^2(\Omega_c)} = (y'(u)^*(y(u) - y_d) + \gamma(u, d)_{L^2(\Omega_c)}$$

Therefore,

$$\begin{aligned}\nabla f(u) &= y'(u)^*(y(u) - y_d) + \gamma u = R^*(A^{-1})^*(A^{-1}(r + Ru) - y_d) + \gamma u \\ &= \gamma u + R^*(A^{-1})^*(A^{-1}(r + Ru) - y_d) \stackrel{\text{def}}{=} \gamma u + B(u).\end{aligned}$$

Since $R \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$, we have $R^* \in \mathcal{L}(H_0^1(\Omega), L^p(\Omega_c))$ with $p = p'/(p'-1) > 2$. Hence, the affine linear operator

$$B(u) = R^*(A^{-1})^*(A^{-1}(r + Ru) - y_d)$$

is a continuous affine linear mapping $L^2(\Omega_c) \to L^p(\Omega)$.

Hence, we can apply Theorem 5.4 to rewrite the optimality conditions as a semismooth operator equation

$$\Phi(u) \stackrel{\text{def}}{=} u - P_{[\alpha,\beta]}(-(1/\gamma)B(u)) = 0.$$

The Newton system reads

$$(I + \frac{1}{\gamma} g^k \cdot B'(u^k))s^k = -\Phi(u^k), \tag{5.6}$$

where $g \cdot B'(u)$ stands for $v \mapsto g \cdot (B'(u)v)$ and $g^k \in L^\infty(\Omega_c)$ is chosen such that

$$g^k(x) \begin{cases} = 0 & -(1/\gamma)B(u^k)(x) \notin [\alpha,\beta], \\ = 1 & -(1/\gamma)B(u^k)(x) \in (\alpha,\beta), \\ \in [0,1] & -(1/\gamma)B(u^k)(x) \in \{\alpha,\beta\}. \end{cases}$$

The linear operator on the left has the form

$$M_k \stackrel{\text{def}}{=} I + \frac{1}{\gamma} g^k \cdot B'(u^k) = I + \frac{1}{\gamma} g^k \cdot R^*(A^{-1})^* A^{-1} R.$$

For solving (5.6), it can be advantagous to note that s^k solves (5.6) if and only if $s^k = d_u^k$ and $(d_y^k, d_u^k, d_\mu^k)^T$ solves

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -\frac{1}{\gamma} g^k \cdot R^* \\ A & -R & 0 \end{pmatrix} \begin{pmatrix} d_y^k \\ d_u^k \\ d_\mu^k \end{pmatrix} = \begin{pmatrix} 0 \\ -\Phi(u^k) \\ 0 \end{pmatrix} \tag{5.7}$$

As we will see later in section 7.2, this is system is amenable to multigrid methods.

5.5. General optimization problems with inequality constraints in L^2. We now consider problems of the form

$$\min_{w \in W} f(w) \quad E(w) = 0, \quad C_j(w) \le 0 \text{ a.e. on } \Omega_j,\ j = 1, \dots, m.$$

Here W and Z are Banach spaces, $f : W \to \mathbb{R}$, $E : W \to Z$, and $C_j : W \to L^2(\Omega_j)$ are twice continuously F-differentiable. The sets $\Omega_j \subset \mathbb{R}^{n_j}$ are assumed to be open and bounded.

The main application we have in mind are control-constrained optimal control problems with L^2-control u and state $y \in Y$:

$$\min_{y \in Y, u \in L^2(\Omega)} J(y,u) \quad E(y,u) = 0, \quad a_i \le u_i \le b_i,\ i = 1, \dots, r,$$

with $y \in Y$ denoting the state, $u \in L^2(\Omega_1) \times \cdots \times L^2(\Omega_r)$ denoting the controls, and $a_i, b_i \in L^\infty(\Omega_i)$.

In this case, we have

$$w = (y,u), \quad m = 2r, \quad C_{2i-1}(y,u) = a_i - u_i,\ C_{2i}(y,u) = u_i - b_i,\ i = 1, \dots, r.$$

To simplify the presentation, consider the case $m = 1$, i.e.,

$$\min_{w \in W} f(w) \quad E(w) = 0, \quad C(w) \le 0 \text{ a.e. on } \Omega. \tag{5.8}$$

The Lagrange function is given by

$$L : W \times L^2(\Omega) \times Z^*, \quad L(w, \lambda, \mu) = f(w) + (\lambda, C(w))_{L^2} + \langle \mu, E(w) \rangle_{Z^*.Z}.$$

Assuming that a CQ holds at the solution $w^* \in W$, the KKT conditions hold:

There exist $\lambda^* \in L^2(\Omega)$ and $\mu^* \in Z^*$ such that (w^*, λ^*, μ^*) satisfies

$$L'_w(w, \lambda, \mu) = 0, \tag{5.9}$$

$$E(w) = 0, \tag{5.10}$$

$$C(w) \leq 0, \quad \lambda \geq 0, \quad (\lambda, C(w))_{L^2} = 0. \tag{5.11}$$

The last line can equivalently be written as VI($-C, K$) with $K = \{u \in L^2(\Omega) \,:\, u \geq 0\}$ and this VI can again be rewritten using the projection onto K:

$$\lambda - P_K(\lambda + \theta C(w)) = 0.$$

Since $P_K(u) = P_{[0,\infty)}(u(\cdot))$, we again have to deal with a superposition operator.

To make the whole KKT system a semismooth equation, we need to get a smoothing operator inside of the projection.

We need additional structure to achieve this. Since it is not very enlightening to define this structure in full generality, we look at an example.

5.6. Application to an elliptic control problem. Very similar as in section 5.4, we consider the following control-constrained elliptic optimal control problem

$$\begin{aligned} \min_{y \in H_0^1(\Omega), u \in L^2(\Omega)} \; & J(y, u) \stackrel{\text{def}}{=} \|y - y_d\|_{L_2}^2 + \frac{\gamma}{2} \|u\|_{L^2}^2 \\ \text{s.t. } & Ay = r + Ru, \quad u \leq b. \end{aligned} \tag{5.12}$$

Here $\Omega \subset \mathbb{R}^n$ is an open bounde domain and $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a second order linear elliptic operator, e.g., $A = -\Delta$. Furthermore, $b \in L^\infty(\Omega)$ is an upper bound on the control, $r \in H^{-1}(\Omega)$ is a source term, and $C \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$, $p' \in [1, 2)$ is the control operator. For a more detailed explanation of the problem setting, see section 5.4.

We convert this control problem into the form (5.8) by setting

$$\begin{aligned} & w = (y, u), \quad W = Y \times U, \quad Y = H_0^1(\Omega), \quad U = L^2(\Omega), \quad Z = H^{-1}(\Omega), \\ & E(y, u) = Ay - Ru - r, \quad C(y, u) = u - b. \end{aligned}$$

Note that E is a continuous linear operator and C is a continuous affine linear operator. Hence,

$$E'_y(y, u) = A, \quad E'_u(y, u) = -R, \quad C'_y(y, u) = 0, \quad C'_u(y, u) = I.$$

The Lagrange function is

$$L(y, u, \lambda, \mu) = J(y, u) + (\lambda, C(y, u)_{L^2}) + \langle \mu, E(y, u) \rangle_{H_0^1, H^{-1}}.$$

We write down the optimality conditions:

$$\begin{aligned} L_y'(y,u,\lambda,\mu) = J_y(y,u) + C_y'(y,u)^*\lambda + E_y'(y,u)^*\mu = y - y_d + A^*\mu = 0,\\ L_u'(y,u,\lambda,\mu) = J_u(y,u) + C_u'(y,u)^*\lambda + E_u'(y,u)^*\mu = \gamma u + \lambda - R^*\mu = 0,\\ \lambda \geq 0, \quad C(y,u) = u - b \leq 0, \quad (\lambda, C(y,u))_{L^2} = (\lambda, u-b)_{L^2} = 0,\\ E(y,u) = Ay - Ru - r = 0. \end{aligned}$$

The second equation yields $\lambda = R^*\mu - \gamma u$ and inserting this, we arrive at

$$\begin{aligned} y - y_d + A^*\mu &= 0, && \text{(adjoint equation)}\\ R^*\mu - \gamma u \geq 0, \quad u \leq b, \quad (R^*\mu - \gamma u, u-b)_{L^2} &= 0, &&\\ Ay - Ru - f &= 0, && \text{(state equation)} \end{aligned}$$

We can reformulate the complementarity condition by using the projection $P_{[0,\infty)}$ as follows:

$$b - u - P_{[0,\infty)}(b - u - \theta(R^*\mu - \gamma u)) = 0.$$

If we choose $\theta = 1/\gamma$, this simplifies to

$$\Phi(u,\mu) := u - b + P_{[0,\infty)}(b - (1/\gamma)R^*\mu) = 0.$$

Since $R^* \in \mathcal{L}(H_0^1(\Omega), L^p(\Omega))$ with $p = p'/(p'-1) > 2$, we see that

$$(u,\mu) \in L^2(\Omega) \times H_0^1(\Omega) \mapsto b - (1/\gamma)R^*\mu \in L^p(\Omega)$$

is continuous and affine linear, and thus Φ is $\partial\Phi$-semismooth w.r.t.

$$\begin{aligned} &\partial\Phi : L^2(\Omega) \times H_0^1(\Omega) \rightrightarrows \mathcal{L}(L^2(\Omega) \times H_0^1(\Omega), L^2(\Omega)),\\ &\partial\Phi(u,\mu) = \big\{ M \,;\, M = (I, -(g/\gamma)\cdot R^*),\ g \in L^\infty(\Omega),\\ &\qquad\qquad g(x) \in \partial^{cl} P_{[0,\infty)}(b(x) - (1/\gamma)R^*\mu(x)) \text{ for a.a. } x \in \Omega \big\}. \end{aligned}$$

Here,

$$\partial P_{[0,\infty)}(t) \begin{cases} 0 & t < 0,\\ 1 & t > 0,\\ [0,1] & t = 0. \end{cases}$$

The semismooth Newton system looks as follows

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -(g^k/\gamma)\cdot R^* \\ A & -R & 0 \end{pmatrix} \begin{pmatrix} s_y \\ s_u \\ s_\mu \end{pmatrix} = - \begin{pmatrix} y^k - y_d + A^*\mu^k \\ u^k - b + P_{[0,\infty)}(b - (1/\gamma)R^*\mu^k) \\ Ay^k - Ru^k - f \end{pmatrix} \tag{5.13}$$

It is important to note that this equation has exactly the same linear operator on the left as the extended system in (5.7). In particular, the regularity condition for the Newton system (5.13) is closely connected to the regularity condition for (5.6).

6. Sequential Quadratic Programming

6.1. Lagrange-Newton methods for equality constrained problems. We consider

$$\min_{w\in W} f(w) \quad \text{s.t.} \quad E(w) = 0 \tag{6.1}$$

with $f : W \to \mathbb{R}$ and $E : W \to Z$ twice continuously F-differentiable.

If w^* is a local solution and a CQ holds (i.e., $E'(w^*)$ is surjective), then the KKT conditions hold:

There exists a Lagrange multiplier $\mu^* \in Z^*$ such that (w^*, μ^*) satisfies

$$L'_w(w,\mu) = f'(w) + E'(w)^*\mu = 0,$$
$$L'_\mu(w,\mu) = E(w) = 0.$$

Setting

$$x = (w,\mu), \quad G(w,\mu) = \begin{pmatrix} L'_w(w,\mu) \\ E(w) \end{pmatrix},$$

the KKT conditions form a nonlinear equation

$$G(x) = 0.$$

To this equation we can apply Newton's method:

$$G'(x^k)s^k = -G(x^k).$$

Written in detail,

$$\begin{pmatrix} L''_{ww}(w^k,\mu^k) & E'(w^k)^* \\ E'(w^k) & 0 \end{pmatrix} \begin{pmatrix} s^k_w \\ s^k_\mu \end{pmatrix} = -\begin{pmatrix} L'_w(w^k,\mu^k) \\ E(w^k) \end{pmatrix}. \tag{6.2}$$

We need a regularity condition:

$$\begin{pmatrix} L''_{ww}(w^*,\mu^*) & E'(w^*)^* \\ E'(w^*) & 0 \end{pmatrix} \quad \text{is boundedly invertible.} \tag{6.3}$$

THEOREM 6.1. *Let f and E be twice continuously F-differentiable. Let (w^*,μ^*) be a KKT pair of* (6.1) *at which the regularity condition* (6.3) *holds. Then there exists $\delta > 0$ such that, for all $(w^0,\mu^0) \in W \times Z^*$ with $\|(w^0,\mu^0) - (w^*,\mu^*)\|_{W\times Z^*} < \delta$, the Lagrange-Newton iteration converges q-superlinearly to (w^*,μ^*).*

If the second derivatives of f and E are locally Lipschitz continuous, then the rate of convergence is q-quadratic.

Proof. We just have to apply the convergence theory of Newton's method.

If the second derivatives of f and E are locally Lipschitz continuous, then G' is locally Lipschitz continuous, and thus we have q-quadratic convergence. □

So far, it is not clear what the connection is between the Lagrange-Newton method and sequential quadratic programming.

However, the connection is very close. Consider the following quadratic program:

SQP subproblem:

$$\begin{aligned} &\min_{d\in W} \langle f'(w^k), d\rangle_{W^*,W} + \frac{1}{2}\langle L''_{ww}(w^k,\mu^k)d, d\rangle_{W^*,W} \\ &\text{s.t. } E(w^k) + E'(w^k)d = 0. \end{aligned} \tag{6.4}$$

The constraint is linear with derivative $E'(w^k)$. As we will show below, $E'(w^k)$ is surjective for w^k close to w^*.

Therefore, at a solution d^k of (6.4), the KKT conditions hold:

There exists $\mu^k_{qp} \in Z^*$ such that (d^k, μ^k_{qp}) solves

$$\begin{aligned} f'(w^k) + L''_{ww}(w^k,\mu^k)d + E'(w^k)^*\mu_{qp} &= 0 \\ E(w^k) + E'(w^k)d &= 0. \end{aligned} \tag{6.5}$$

It is now easily seen that (d^k, μ^k_{qp}) solves (6.5) if and only if $(s^k_w, s^k_\mu) = (d^k, \mu^k_{qp} - \mu^k)$ solves (6.2).

Hence, locally, the Lagrange-Newton method is equivalent to the following method:

ALGORITHM 6.2 (SQP method for equality constrained problems).

0. Choose (w^0, μ^0) (sufficiently close to (w^*, μ^*)).

For $k = 0, 1, 2, \ldots$:

1. If (w^k, μ^k) is a KKT pair of (6.1)*, STOP.*

2. Compute the KKT pair (d^k, μ^{k+1}) of

$$\begin{aligned} &\min_{d\in W} \langle f'(w^k), d\rangle_{W^*,W} + \frac{1}{2}\langle L''_{ww}(w^k,\mu^k)d, d\rangle_{W^*,W} \\ &\text{s.t. } E(w^k) + E'(w^k)d = 0. \end{aligned}$$

that is closest to $(0, \mu^k)$.

3. Set $w^{k+1} = w^k + s^k$.

For solving the SQP subproblems in step 2, it is important to know if for w^k close to w^*, the operator $E'(w^k)$ is indeed surjective and if there exists a unique solution to the QP.

LEMMA 6.3. *Let W be a Hilbert space and Z be a Banach space. Furthermore, let $E : W \to Z$ be continuously F-differentiable and let $E'(w^*)$ be surjective. Then $E'(w)$ is surjective for all w close to w^*.*

Proof. We set $B = E'(w^*)$, and $B(w) = E'(w)$, and do the splitting $W = W_0 \perp W_1$ with $W_0 = \text{Kern}(B)$. We then see that $B|_{W_1} \in \mathcal{L}(W_1, Z)$ is bijective and thus continuously invertible (open mapping theorem). Now, by continuity, for $w \to w^*$ we have $B(w) \to B$ in $\mathcal{L}(W, Z)$ and thus

also $B(w)|_{W_1} \to B|_{W_1}$ in $\mathcal{L}(W_1, Z)$. Therefore, by the Lemma of Banach, $B(w)|_{W_1}$ is continuously invertible for w close to w^* and thus $B(w)$ is onto. □

Next, we show a second-order sufficient condition for the QP.

LEMMA 6.4. *Let W be a Hilbert space and Z be a Banach space. Furthermore, let $f : W \to \mathbb{R}$ and $E : W \to Z$ be twice continuously F-differentiable. Let $E(w^*) = 0$ assume that $E'(w^*)$ is surjective. In addition, let the following second-order sufficient condition hold at (w^*, μ^*):*

$$\langle d, L''_{ww}(w^*, \mu^*)d\rangle_{W,W^*} \geq \alpha \|d\|_W^2 \quad \forall\, d \in W \text{ with } E'(w^*)d = 0,$$

where $\alpha > 0$ is a constant. Then, there exists $\delta > 0$ such that for all $(w, \mu) \in W \times Z^*$ with $\|(w, \mu) - (w^*, \mu^*)\|_{W\times Z^*} < \delta$ the following holds:

$$\langle d, L''_{ww}(w, \mu)d\rangle_{W,W^*} \geq \frac{\alpha}{2} \|d\|_W^2 \quad \forall\, d \in W \text{ with } E'(w)d = 0,$$

Proof. Set $B = E'(w^*)$, $B(w) = E'(w)$, $W_0 = \text{Kern}(W)$ and split $W = W_0 \perp W_1$. Remember that $B|_{W_1} \in \mathcal{L}(W_1, Z)$ is continuously invertible.

For any $d \in \text{Kern}(B(x))$ there exist unique $d_0 \in W_0$ and $d_1 \in W_1$ with $d = d_0 + d_1$. Our first aim is to show that d_1 is small. In fact,

$$\|Bd_1\|_Z = \|Bd\|_Z = \|(B - B(w))d\| \leq \|B - B(w)\|_{W,Z} \|d\|_W.$$

Hence,

$$\|d_1\|_W \leq \|(B|_{W_1})^{-1}\|_{Z,W_1} \|B - B(w)\|_{W,Z} \|d\|_W \stackrel{\text{def}}{=} \xi(w)\|d\|_W.$$

Therefore, setting $x = (w, \mu)$,

$$\begin{aligned}
\langle L''_{ww}(x)d, d\rangle_{W^*,W} &= \langle L''_{ww}(x^*)d, d\rangle_{W^*,W} + \langle (L''_{ww}(x) - L''_{ww}(x^*))d, d\rangle_{W^*,W} \\
&= \langle L''_{ww}(x^*)d_0, d_0\rangle_{W^*,W} + \langle L''_{ww}(x^*)(d + d_0), d_1\rangle_{W^*,W} + \langle (L''_{ww}(x) - L''_{ww}(x^*))d, d\rangle_{W^*,W} \\
&\geq \alpha\|d_0\|_W^2 - \|L''_{ww}(x^*)\|_{W,W^*}(\|d\|_W + \|d_0\|_W)\|d_1\|_W - \|L''_{ww}(x) - L''_{ww}(x^*)\|_{W,W^*}\|d\|_W^2 \\
&\geq (\alpha(1 - \xi^2(w)) - 2\|L''_{ww}(x^*)\|_{W,W^*}\xi(w) - \|L''_{ww}(x) - L''_{ww}(x^*)\|_{W,W^*})\|d\|_W^2 \\
&=: \alpha(x)\|d\|_W^2.
\end{aligned}$$

By continuity, $\alpha(x) \to \alpha$ for $x \to x^*$. □

A sufficient condition for the regularity condition (6.3) is the following:

LEMMA 6.5. *Let W be a Hilbert space, let $E'(w^*)$ be surjective (this is a CQ), and assume that the following second order sufficient condition holds:*

$$\langle d, L''_{ww}(w^*, \mu^*)d\rangle_{W,W^*} \geq \alpha \|d\|_W^2 \quad \forall\, d \in W \text{ with } E'(w^*)d = 0.$$

Then the regularity condition (6.3) *holds.*

Proof. For brevity, set $A = L''_{ww}(w^*, \mu^*)$ and $B = E'(w^*)$. We consider the unique solvability of

$$\begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix} \begin{pmatrix} w \\ \mu \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}.$$

Denote by W_0 the null space of B and by W_1 its orthogonal complement. Then $W = W_0 \perp W_1$ and W_0, W_1 are Hilbert spaces.

Since B is surjective, the equation $Bw = r_2$ is solvable and the set of all solutions is $w_1(r_2) + W_0$, where $w_1(r_2) \in W_1$ is uniquely determined.

We have

$$\langle d, Ad\rangle_{W,W^*} \geq \alpha \|d\|_W^2 \quad \forall\, d \in W_0.$$

Hence, by the Lax-Milgram Lemma, there exists a unique solution $w_0(r_1, r_2) \in W_0$ to the problem

$$w_0 \in W_0,\ \langle d, Aw_0\rangle_{W,W^*} = \langle d, r_1 - Aw_1(r_2)\rangle_{W,W^*} \quad \forall\, d \in W_0.$$

Since B is surjective, the closed range theorem yields

$$\text{Kern}(B^*) = (BW)^\perp = Z^\perp = \{0\}.$$

Hence, B^* is injective. Also, since $BW = Z$ is closed, the closed range theorem yields

$$B^* Z^* = \text{Kern}(B)^\perp = W_0^\perp$$

Here, for $S \subset X$

$$S^\perp = \{x' \in X^* \,:\, \langle x', s\rangle_{X^*,X} = 0 \,\forall\, s \in S\}\,.$$

By construction, $r_1 - Aw_0(r_1, r_2) - Aw_1(r_2) \in W_0^\perp$. Hence, there exists a unique $\mu(r_1, r_2) \in Z^*$ such that

$$\mu(r_1, r_2) = r_1 - Aw_0(r_1, r_2) - Aw_1(r_2)$$

Therefore, we have found the unique solution

$$\begin{pmatrix} w \\ \mu \end{pmatrix} = \begin{pmatrix} w_0(r_1, r_2) + w_1(r_2) \\ \mu(r_1, r_2) \end{pmatrix}.$$

□

6.2. The Josephy-Newton method. In the previous section, we were able to derive the SQP method for equality-constrained problems by applying Newton's method to the KKT system.

For inequality constrained problems this is not directly possible since the KKT system consists of operator equations and a variational inequality. As we will see, such a combination can be most elegantly written as a

6.2.1. *Generalized Equation:*

$$\text{GE}(G, N): \qquad 0 \in G(x) + N(x).$$

Here, $G : X \to Y$ is assumed to be continuously F-differentiable and $N : X \rightrightarrows Y$ is a set-valued mapping with closed graph.

For instance, the variational inequality VI(F, S), with $F : W \to W^*$ and $S \subset W$ closed and convex, can be written as

$$0 \in F(w) + N_S(w),$$

where N_S is the normal cone mapping of S:

DEFINITION 6.6. *Let $S \subset W$ be a nonempty closed convex subset of the Banach space W. The* normal cone $N_S(w)$ *of S at $w \in W$ is defined by*

$$N_S(w) = \begin{cases} \{y \in W^* \,:\, \langle y, z-w\rangle_{W^*,W} \le 0 \;\forall\, z \in S\}\,, & w \in S, \\ \emptyset & w \notin S. \end{cases}$$

This defines a set-valued mapping $N_S : W \rightrightarrows W^*$.

The Josephy-Newton method for generalized equations looks as follows:

ALGORITHM 6.7 (Josephy-Newton method for GE(G, N)).

0. Choose $x^0 \in X$ (sufficiently close to the solution x^* of GE(G, N)).

For $k = 0, 1, 2 \ldots$

1. STOP if x^k solves GE(G, N) (holds if $x^k = x^{k-1}$).

2. Compute the solution x^{k+1} of

$$\begin{aligned} &\text{GE}(G(x^k) + G'(x^k)(\cdot - x^k), N) : \\ &0 \in G(x^k) + G'(x^k)(x - x^k) + N(x) \end{aligned}$$

that is closest to x^k.

In the ordinary Newton's method, which corresponds to $N(x) = \{0\}$ for all x, an essential ingredient is the regularity condition that $G'(x^*)$ is continuously invertible.

This means that the linearized equation

$$p = G(x^*) + G'(x^*)(x - x^*)$$

possesses the unique solution $x(p) = G'(x^*)^{-1}p$, which of course depends linearly and thus Lipschitz continuously on $p \in Y$.

The appropriate generalization of this regularity condition is the following:

DEFINITION 6.8 (Strong regularity). *The generalized equation GE(G, N) is called* strongly regular *at a solution x^* if there exist $\delta > 0$, $\varepsilon > 0$ and $L > 0$ such that, for all $p \in Y$, $\|p\|_Y < \delta$, there exists a unique $x = x(p) \in X$ with $\|x(p) - x^*\|_X < \varepsilon$ such that*

$$p \in G(x^*) + G'(x^*)(x - x^*) + N(x)$$

and $x(p)$ is Lipschitz continuous:

$$\|x(p_1) - x(p_2)\|_X \le L\|p_1 - p_2\|_Y \quad \forall\, p_1, p_2 \in Y,\ \|p_i\|_X < \delta,\ i = 1, 2.$$

It is a milestone result of Robinson ([**67**]) that then the following holds:

THEOREM 6.9. *Let X, Y, and Z be Banach spaces. Furthermore, let $z^* \in Z$ be fixed and assume that x^* is a solution of*

$$\text{GE}(G(z^*, \cdot), N) : \quad 0 \in G(z^*, x) + N(x)$$

at which the GE is strongly regluar with Lipschitz modulus L. Assume that G is F-differentiable with repect to x near (z^*, x^*) and that G and G'_x are continuous at (z^*, x^*).

Then, for every $\varepsilon > 0$, there exist neighborhoods $Z_\varepsilon(z^*)$ of z^*, $X_\varepsilon(x^*)$ of x^*, and a mapping $x : Z_\varepsilon(z^*) \to X_\varepsilon(x^*)$ such that, for all $z \in Z_\varepsilon(z^*)$, $x(z)$ is the (locally) unique solution of the generalized equation

$$0 \in G(z, x) + N(x), \quad x \in X_\varepsilon(x^*).$$

In addition,

$$\|x(z_1) - x(z_2)\|_X \leq (L + \varepsilon)\|G(z_1, x(z_2)) - G(z_2, x(z_2))\|_Y \quad \forall\, z_1, z_2 \in Z_\varepsilon(z^*).$$

From this, it is not difficult to derive fast local convergence of the Josephy-Newton method:

THEOREM 6.10. *Let X, Y be Banach spaces, $G : X \to Y$ continuously F-differentiable, and let $N : X \rightrightarrows Y$ be set-valued with closed graph. If x^* is a strongly regular solution of* GE(G, N)*, then the Josephy-Newton method (Alg.* 6.7*) is locally q-superlinearly convergent in a neighborhood of x^*. If, in addition, G' is α-Hölder continuous near x^*, then the order of convergence is $1 + \alpha$.*

Proof. For compact notation, we set $B_\delta(x) = \{y \in X : \|y - x\|_X < \delta\}$.

Let L be the Lipschitz modulus of strong regularity. We set $Z = X$, $z^* = x^*$ and consider

$$\bar{G}(z, x) \stackrel{\text{def}}{=} G(z) + G'(z)(x - z).$$

Since $\bar{G}(z^*, \cdot)$ is affine linear, we have

$$\bar{G}(z^*, x^*) + \bar{G}'_x(z^*, x^*)(x - x^*) = G(z^*, x) = G(z^*) + G'(z^*)(x - z^*) = G(x^*) + G'(x^*)(x - x^*).$$

Therefore, GE$(\bar{G}(z^*, \cdot), N)$ is strongly regular at x^* with Lipschitz constant L. Theorem 6.9 is applicable and thus, for $\varepsilon > 0$, there exist neighborhoods $Z_\varepsilon(x^*)$ of $z^* = x^*$ and $X_\varepsilon(x^*)$ of x^* such that, for all $z \in Z_\varepsilon(x^*)$,

$$0 \in \bar{G}(z, x) + N(x) = G(z) + G'(z)(x - z) + N(x), \quad x \in X_\varepsilon(x^*)$$

has a unique solution $x(z)$ that satisfies

$$\begin{aligned}
&\forall\, z_1, z_2 \in Z_\varepsilon(z^*) = Z_\varepsilon(x^*) : \\
&\|x(z_1) - x(z_2)\|_X \leq (L + \varepsilon)\|\bar{G}(z_1, x(z_2)) - \bar{G}(z_2, x(z_2))\|_Y \\
&\quad = (L + \varepsilon)\|G(z_1) - G(z_2) + G'(z_1)(x(z_2) - z_1) - G'(z_2)(x(z_2) - z_2)\|_Y
\end{aligned}$$

If we choose $z_1 = z \in Z_\varepsilon(x^*)$ and $z_2 = x^*$, we obtain $x(z^2) = x^*$ and thus forall $z \in Z_\varepsilon(x^*)$:

$$\begin{aligned}
\|x(z) - x^*\|_X &\le (L+\varepsilon)\|G(z) - G(x^*) + G'(z)(x^* - z) - G'(x^*)(x^* - x^*)\|_Y \\
&= (L+\varepsilon)\|G(z) - G(x^*) - G'(z)(z - x^*)\|_Y \\
&\le (L+\varepsilon)\|G(z) - G(x^*) - G'(x^*)(z - x^*)\|_Y \\
&+ (L+\varepsilon)\|(G'(x^*) - G'(z))(z - x^*)\|_Y \qquad (6.6)\\
&\le (L+\varepsilon)\|G(z) - G(x^*) - G'(x^*)(z - x^*)\|_Y \\
&+ (L+\varepsilon)\|G'(x^*) - G'(z)\|_{X,Y}\|z - x^*\|_X \\
&= o(\|z - x^*\|_X) \quad (z \to x^*).
\end{aligned}$$

In the last estimate, we have used the F-differentiability of G and the continuity of G'.

Now choose $\delta > 0$ such that $B_{3\delta}(x^*) \subset Z_\varepsilon(x^*)$. By possibly reducing δ, we achieve

$$x(z) \in B_{\delta/2}(x^*) \subset B_\delta(x^*) \quad \forall\, z \in B_\delta(x^*).$$

Now observe that, for $x^k \in B_\delta(x^*)$, the unique solution of GE($G(x^k) + G'(x^k)(\cdot - x^k), N$) in $Z_\varepsilon(x^*)$ is given by $x(x^k) \in B_{\delta/2}(x^*)$.

From

$$\|x(x^k) - x^k\| \le \|x(x^k) - x^*\|_X + \|x^* - x^k\|_X < \frac{\delta}{2} + \delta = \frac{3}{2}\delta$$

and $B_{3\delta}(x^*) \subset Z_\varepsilon(x^*)$ we conclude that $x(x^k)$ is the solution of GE($G(x^k) + G'(x^k)(\cdot - x^k), N$) that is closest to x^k. Hence, for $x^k \in B_\delta(x^*)$, we have

$$x^{k+1} = x(x^k) \in B_\delta(x^*), \quad \|x^{k+1} - x^*\|_X \le \frac{1}{2}\|x^k - x^*\|_X$$

Thus, if we choose $x^0 \in B_\delta(x^*)$, we obtain by induction $x^k \to x^*$.

Furthermore, from (6.6) it follows that

$$\|x^{k+1} - x^*\|_X = \|x(x^k) - x^*\|_X = o(\|x^k - x^*\|_X) \quad (k \to \infty).$$

This proves the q-superlinear convergence.

If G' is α-order Hölder continuous at x^*, then we can improve the estimate (6.6):

$$\begin{aligned}
\|x(z) - x^*\|_X &\le (L+\varepsilon)\|G(z) - G(x^*) - G'(z)(z - x^*)\|_Y \\
&= (L+\varepsilon)\left\|\int_0^1 G'(x^* + t(z - x^*)) - G'(z))(z - x^*)\, dt\right\|_Y \\
&\le (L+\varepsilon)\int_0^1 \|G'(x^* + t(z - x^*)) - G'(z)\|_{X,Y}\, dt \|z - x^*\|_X \\
&\le (L+\varepsilon)\int_0^1 t\|z - x^*\|_X^\alpha\, dt\, \|z - x^*\|_X = \frac{L+\varepsilon}{1+\alpha}\|z - x^*\|_X^{1+\alpha} \\
&= O(\|z - x^*\|_X^{1+\alpha}) \quad (z \to x^*).
\end{aligned}$$

Hence,

$$\|x^{k+1} - x^*\|_X = \|x(x^k) - x^*\|_X = O(\|x^k - x^*\|_X^{1+\alpha}) \quad (k \to \infty).$$

□

6.3. SQP methods for inequality constrained problems. We consider the problem

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad E(w) = 0,\ C(w) \in K. \tag{6.7}$$

with $f : W \to \mathbb{R}$, $E : W \to Z$, and $C : W \to V$ twice continuously F-differentiable. Furthermore, W ,Z, V are Banach spaces, and V is reflexive (i.e., $V^{**} = V$), and $K \subset V$ is a nonempty closed convex cone.

For this problem, we define the Lagrange function

$$L(w, \lambda, \mu) = f(w) + \langle \lambda, C(w) \rangle_{V^*,V} + \langle \mu, E(w) \rangle_{Z^*,Z}.$$

We will need the notion of the polar cone.

DEFINITION 6.11. *Let X be a Banach space and let $K \subset X$ be a nonempty closed convex cone. Then the* polar cone *of K is defined by*

$$K^\circ = \{y \in X^* \,:\, \langle y, x \rangle_{X^*,X} \leq 0 \,\forall\, x \in K\},$$

Obviously, K° is a closed convex cone.

Recall also the definition of the normal cone mapping (Def. 6.6).

Under a constraint qualification, the following KKT conditions hold:

There exist Lagrange multipliers $\lambda^* \in K^\circ$ and $\mu^* \in Z^*$ such that (w^*, λ^*, μ^*) satisifes

$$\begin{aligned} &L'_w(w, \lambda, \mu) = 0, \\ &C(w) \in K, \quad \lambda \in K^\circ, \quad \langle \lambda, C(w) \rangle_{V^*,V} = 0, \\ &E(w) = 0, \end{aligned}$$

Note that, since $V^{**} = V$, we have $K^\circ \subset V$.

The second condition can be shown to be equivalent to VI($-C(w), K^\circ$). This is a VI w.r.t. λ with a constant operator parametrized by w.

Now comes the trick, see, e.g., [**4**]:

By means of the normal cone N_{K°, it is easily seen that VI($-C(w), K^\circ$) is equivalent to the generalized equation

$$0 \in -C(w) + N_{K^\circ}(\lambda).$$

Therefore, we can write the KKT system as a generalized equation:

$$0 \in \begin{pmatrix} L_w(w, \lambda, \mu) \\ -C(w) \\ E(w) \end{pmatrix} + \begin{pmatrix} \{0\} \\ N_{K^\circ}(\lambda) \\ \{0\} \end{pmatrix}. \tag{6.8}$$

Setting

$$N(w,\lambda,\mu) = \begin{pmatrix} \{0\} \\ N_{K^\circ}(\lambda) \\ \{0\} \end{pmatrix},$$

and noting $L'_\lambda(w,\lambda,\mu) = C(w)$, $L'_\mu(w,\lambda,\mu) = E(w)$, we can write (6.8) very compactly as GE$(-L', N)$.

The closed graph of the normal cone mapping is proved in the next lemma.

LEMMA 6.12. *Let X be a Banach spaces and $S \subset X$ be nonempty, closed, and convex. Then the normal cone mapping N_S has closed graph.*

Proof. Let graph$(N_S) \ni (x^k, y^k) \to (x^*, y^*)$. Then $y^k \in N_S(x^k)$ and thus $x^k \in S$, since otherwise $N_S(x^k)$ would be empty. Since S is closed, $x^* \in S$ follows. Now, for all $z \in S$, by continuity

$$\langle y^*, z - x^* \rangle_{X^*,X} = \lim_{k\to\infty} \underbrace{\langle y^k, z - x^k \rangle_{X^*,X}}_{\leq 0} \leq 0,$$

hence $y^* \in N_S(x^*)$. Therefore, $(x^*, y^*) \in$ graph(N_S). □

If we now apply the Josephy-Newton method to (6.8), we obtain the following subproblem (we set $x^k = (w^k, \lambda^k, \mu^k)$):

$$0 \in \begin{pmatrix} L_w(x^k) \\ -C(w^k) \\ E(w^k) \end{pmatrix} + \begin{pmatrix} L''_{ww}(x^k) & C'(w^k)^* & E'(w^k)^* \\ -C'(w^k) & 0 & 0 \\ E'(w^k) & 0 & 0 \end{pmatrix} \begin{pmatrix} w - w^k \\ \lambda - \lambda^k \\ \mu - \mu^k \end{pmatrix} + \begin{pmatrix} \{0\} \\ N_{K^\circ}(\lambda) \\ \{0\} \end{pmatrix}. \tag{6.9}$$

It is not difficult to see that (6.9) are exactly the KKT conditions of the following quadratic optimization problem:

6.3.1. *SQP subproblem:*

$$\begin{aligned} \min_{w\in W}\ & \langle f'(w^k), w - w^k \rangle_{W^*,W} + \frac{1}{2}\langle L''_{ww}(x^k)(w - w^k), w - w^k \rangle_{W^*,W} \\ \text{s.t. } & E(w^k) + E'(w^k)(w - w^k) = 0, \quad C(w^k) + C'(w^k)(w - w^k) \in K. \end{aligned}$$

In fact, the Lagrange function of the QP is

$$\begin{aligned} L^{qp}(x) = {} & \langle f'(w^k), w - w^k \rangle_{W^*,W} + \frac{1}{2}\langle L''_{ww}(x^k)(w - w^k), w - w^k \rangle_{W^*,W} \\ & + \langle \lambda, C(w^k) + C'(w^k)(w - w^k) \rangle_{W^*,W} + \langle \mu, E(w^k) + E'(w^k)(w - w^k) \rangle_{Z^*,Z}. \end{aligned}$$

Since

$$\begin{aligned} L^{qp\prime}_w(x) &= f'(w^k) + L''_{ww}(x^k)(w - w^k) + C'(w^k)^*\lambda + E'(w^k)^*\mu \\ &= L'_w(x^k) + L''_{ww}(x^k)(w - w^k) + C'(w^k)^*(\lambda - \lambda^k) + E'(w^k)^*(\mu - \mu^k), \end{aligned}$$

we see that writing down the KKT conditions for the QP in the form (6.8) gives exactly the generalized equation (6.9).

We obtain:

ALGORITHM 6.13 (SQP method for inequality constrained problems).

0. Choose (w^0, λ^0, μ^0) *(sufficiently close to* (w^*, λ^*, μ^*)*).*

For $k = 0, 1, 2, \ldots$*:*

1. If (w^k, λ^k, μ^k) *is a KKT triple of* (6.7)*, STOP.*

2. Compute the KKT triple $(d^k, \lambda^{k+1}, \mu^{k+1})$ *of*

$$\min_{d \in W} \langle f'(w^k), d\rangle_{W^*,W} + \frac{1}{2}\langle L''_{ww}(w^k, \lambda^k, \mu^k)d, d\rangle_{W^*,W}$$
$$\text{s.t. } E(w^k) + E'(w^k)d = 0, \quad C(w^k) + C'(w^k)d \in K.$$

that is closest to $(0, \lambda^k, \mu^k)$.

3. Set $w^{k+1} = w^k + d^k$.

Since this method is the Josephy-Newton algorithm applied to (6.8), we can derive local convergence results immediately if a Robinson's strong regularity condition is satisfied. This condition has to be verified from case to case and is connected to second order sufficient optimality conditions. As an example where strong regularity is verified for an optimal control problem, we refer to [**32**].

6.3.2. *Application to optimal control.* For illustration, we consider the nonlinear elliptic optimal control problem

$$\min_{y \in H^1_0, u \in L^2} J(y,u) \stackrel{\text{def}}{=} \|y - y_d\|^2_{L^2} + \frac{\gamma}{2}\|u\|^2_{L^2} \quad \text{s.t.} \quad Ay + y^3 + y = u,\ u \le b. \tag{6.10}$$

Here, $y \in H^1_0(\Omega)$ is the state, which is defined on the open bounded domain $\Omega \subset \mathbb{R}^n$, $n \le 3$, and $u \in L^2(\Omega)$ is the control. Furthermore, $A : H^1_0(\Omega) \to H^{-1}(\Omega) = H^1_0(\Omega)^*$ is a linear elliptic partial differential operator, e.g., $A = -\Delta$. Finally $b \in L^\infty(\Omega)$ is an upper bound on the control. We convert this control problem into the form (6.7) by setting

$$Y = H^1_0(\Omega), \quad U = L^2(\Omega), \quad Z = H^{-1}(\Omega),$$
$$E(y,u) = Ay + y^3 + y - u, \quad C(y,u) = u - b,$$
$$K = \left\{u \in L^2(\Omega) \,:\, u \le 0 \text{ a.e. on } \Omega\right\}.$$

One can show (note $n \le 3$) that the operator E is twice continuously F-differentiable with

$$E'_y(y,u) = A + 3y^2 \cdot I + I, E''_{yy}(y,u)(h_1,h_2) = 6yh_1h_2$$

(the other derivatives are obvious due to linearity). Therefore, given $x^k = (y^k, u^k, \lambda^k, \mu^k)$, the SQP subproblem reads

$$\min_{d_y, d_u} (y^k - y_d, d_y)_{L^2} + A\gamma(u^k, d_u)_{L^2} + \frac{1}{2}\|d_y\|^2_{L^2} + \frac{1}{2}\langle \mu^k, 6y^k d_y^2\rangle_{H^1_0, H^{-1}} + \frac{\gamma}{2}\|d_u\|^2_{L^2}$$
$$\text{s.t. } Ay^k + (y^k)^3 + y^k - u^k + Ad_y + 3(y^k)^2 d_y + d_y - d_u = 0,$$
$$u_k + d_u \le b.$$

7. Further aspects

7.1. Mesh independence. For numerical computations, we have to discretize the problem (Finite elements, finite differences,...) and to apply the developed optimization methods to the discretized, finite dimensional problem. One such situation would be, for instance, to apply an SQP method to the discretization (P_h) of the infinite dimensional problem (P). If this is properly done, we can interpret the discrete SQP method as an inexact (i.e. perturbed) version of the SQP method applied to (P).

Abstractly speaking, we have an infinite dimensional problem (P) and an algorithm A for its solution. Furthermore, we have a family of finite dimensional approximations (P_h) of (P), and discrete versions A_h of algorithm A. Here $h > 0$ denotes the accuracy of discretization (with increasing accuracy as $h \to 0$). Starting from x^0 and the corresponding discrete point x_h^0, respectively, the algorithms A and A_h will generate sequences (x^k) and (x_h^k), respectively. Mesh independence means that the convergence behavior of (x^k) and (x_h^k) become more and more alike as the discretization becomes more and more accurate, i.e., as $h \to 0$. This means, for instance, that q-superlinear convergence of Alg. A on a δ-neighborhood of the solution implies the same rate of convergence for Alg. A_h on a δ-neighborhood of the corresponding discrete solution as soon as h is sufficiently small.

Mesh independence results for Newton's method were established in, e.g., [**2, 26**]. The mesh independence of SQP methods and Josephy-Newton methods was shown, e.g., in [**5, 27**]. Furthermore, the mesh independence of semismooth Newton methods was established in [**40**].

7.2. Application of fast solvers. An important ingredient in PDE constrained optimization is the combination of optimization methods with efficient solvers (sparse linear solvers, multigrid, preconditioned Krylov subspace methods, etc.). It is by far out of the scope of these notes to give details. Instead, we focus on just two simple examples.

For both semismooth reformulations of the elliptic control problems (5.5) and (5.12), we showed that the semismooth Newton system is equivalent to

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -\frac{1}{\gamma} g^k \cdot R^* \\ A & -R & 0 \end{pmatrix} \begin{pmatrix} d_y^k \\ d_u^k \\ d_\mu^k \end{pmatrix} = \begin{pmatrix} r_1^k \\ r_2^k \\ r_3^k \end{pmatrix} \tag{7.1}$$

with appropriate right hand side. Here $A \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ is an elliptic operator, $R \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$ with $p' \in [1, 2)$, and $g^k \in L^\infty(\Omega_c)$ with $g^k \in [0, 1]$ almost everywhere. We can do block elimination to obtain

$$\begin{pmatrix} I & A^* & 0 \\ A & -\frac{1}{\gamma} R(g^k \cdot R^*) & 0 \\ 0 & -\frac{g^k}{\gamma} \cdot R^* & I \end{pmatrix} \begin{pmatrix} s_y \\ s_\mu \\ s_u \end{pmatrix} = - \begin{pmatrix} r_1^k \\ R r_2^k + r_3^k \\ r_2^k \end{pmatrix}$$

The first two rows form a 2×2 elliptic system for which very efficient fast solvers (e.g., multigrid [**34**]) exist.

Similar techniques can successfully be used, e.g., for elastic contact problems [**79**].

7.3. Other methods. Our treatment of Newton-type methods is not at all complete. There exist, for instance, interior point methods that are very well suited for optimization problems in function spaces, see [**37, 78**].

CHAPTER 3

Discrete concepts in pde constrained optimization

M. Hinze
Universität Hamburg
Department Mathematik
Schwerpunkt Approximation und Optimierung
Bundestraße 55
D-20146 Hamburg, Germany,
email:hinze@math.uni-hamburg.de

1. Introduction

This chapter presents the state of the art of discrete concepts in pde constrained optimization including control and state constraints. So far, concepts without constraints are fairly well understood, and theory and praxis for control constraints are strongly emerging. However, the development of reliable numerical approaches for state constraints is still an open issue and requires further intensive research.

We illustrate all concepts at hand of model pdes which are well understood w.r.t. analysis and discretization concepts. This allows to focus the presentation on structural aspects inherent in optimal control problems with pde constraints.

2. Stationary model problem

We consider the *Mother Problem*

$$(\mathbb{P}) \quad \begin{cases} \min_{(y,u)\in Y\times U} J(y,u) := \frac{1}{2}\|y-z\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U \\ \text{s.t.} \\ \quad -\Delta y = Bu \quad \text{in } \Omega, \\ \quad\quad y = 0 \quad\quad \text{on } \partial\Omega, \\ \text{and} \\ u \in U_{\text{ad}} \subseteq U. \end{cases} \tag{2.1}$$

Here, $\Omega \subset \mathbb{R}^n$ denotes an open, bounded sufficiently smooth (polyhedral) domain, $Y := H^1_0(\Omega)$, the operator $B : U \to H^{-1}(\Omega) \equiv Y^*$ denotes the (linear, continuous) control operator, and U_{ad} is assumed to be a closed and convex subset of the Hilbert space U.

EXAMPLE 2.1.

(1) $U := L^2(\Omega)$, $B : L^2(\Omega) \to H^{-1}(\Omega)$ *Injection,* , $U_{ad} := \{v \in L^2(\Omega); a \le v(x) \le b$ *a.e. in* $\Omega\}, a, b \in L^\infty(\Omega)$.

(2) $U := \mathbb{R}^m$, $B : \mathbb{R}^m \to H^{-1}(\Omega)$, $Bu := \sum_{j=1}^m u_j F_j$, $F_j \in H^{-1}(\Omega)$ *given* , $U_{ad} := \{v \in \mathbb{R}^m; a_j \le v_j \le b_j\}, a < b$.

We already know that problem $\mathbb{P}$ admits a unique solution $(y,u) \in H^1(\Omega) \times U$, and that $(\mathbb{P})$ equivalently can be rewritten as the optimization problem

$$\min_{u\in U_{\text{ad}}} \hat{J}(u) \tag{2.2}$$

for the reduced functional $\hat{J}(u) := J(y(u),u) \equiv J(SBu,u)$ over the set U_{ad}, where $S : Y^* \to Y$ denotes the solution operator associated with $-\Delta$. We further know that the first order necessary (and here also sufficient) optimality conditions take the form

$$(\hat{J}'(u), v-u)_U \ge 0 \text{ for all } v \in U_{\text{ad}} \tag{2.3}$$

where $\hat{J}'(u) = \alpha u + B^*S^*(SBu - z) \equiv \alpha u + B^*p$, with $p := S^*(SBu - z)$ denoting the adjoint variable. The function p in our setting satisfies

$$\begin{aligned} -\Delta p &= y - z \quad \text{in } \Omega, \\ p &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

To discretize $(\mathbb{P})$ we concentrate on Finite Element approaches and make the following assumptions.

ASSUMPTION 4.
$\Omega \subset \mathbb{R}^n$ *denotes a polyhedral domain,* $\bar{\Omega} = \cup_{j=1}^{nt} \bar{T}_j$ *with admissible quasi-uniform sequences of partitions* $\{T_j\}_{j=1}^{nt}$ *of* Ω*, i.e. with* $h_{nt} := \max_j$ *diam* T_j *and* $\sigma_{nt} := \min_j\{\sup$ *diam* $K; K \subseteq T_j\}$ *there holds* $c \le \frac{h_{nt}}{\sigma_{nt}} \le C$ *uniformly in* nt *with positive constants* $0 < c \le C < \infty$ *independent of* nt*. We abbreviate* $\tau_h := \{T_j\}_{j=1}^{nt}$.

In order to tackle $(\mathbb{P})$ numerically we shall distinguish two different approaches. The first is called *First discretize, then optimize*, the second *First optimize, then discretize*. It will turn out that both approaches under certain circumstances lead to the same numerical results. However, from a structural point of view they are completely different.

2.1. First discretize, then optimize. The *First discretize, then optimize* approach works as follows. All quantities in $(\mathbb{P})$ are discretized a-priori, which results in a finite dimensional optimization problem. To discretize we replace the spaces Y and U by finite dimensional subspaces Y_h and U_d, the set U_{ad} by some discrete counterpart U_{ad}^d, and the functionals, integrals and dualities by appropriate discrete surrogates. Having in mind Assumption 4 we set for $k \in \mathbb{N}$

$$W_h := \{v \in C^0(\bar{\Omega}); v_{|_{T_j}} \in \mathbb{P}_k(T_j) \text{ for all } 1 \leq j \leq nt\} =: \langle \phi_1, \dots, \phi_{ng} \rangle, \text{ and}$$
$$Y_h := \{v \in W_h, v_{|_{\partial\Omega}} = 0\} =: \langle \phi_1, \dots, \phi_n \rangle \subseteq Y,$$

with some $0 < n < ng$. The resulting Ansatz for y_h then is of the form $y_h(x) = \sum_{i=1}^{n} y_i \phi_i$. Further, with $u^1, \dots, u^m \in U$, we set $U_d := \langle u^1, \dots, u^m \rangle$ and $U_{\text{ad}}^d := P_{U_{\text{ad}}}^d(U_d)$, where $P_{U_{\text{ad}}}^d : U \to U_{\text{ad}}$ is a sufficiently smooth (nonlinear) mapping. It is convenient to assume that U_{ad}^d may be represented in the form

$$U_{\text{ad}}^d = \left\{ u \in U; u = \sum_{j=1}^{m} s_j u^j, s \in C \right\}$$

with $C \subset \mathbb{R}^m$ denoting a convex closed set. Finally let $z_h := I_h z = \sum_{i=1}^{ng} z_i \phi_i$, where $I_h : L^2(\Omega) \to W_h$ denotes a continuous interpolation operator. Now we replace problem $(\mathbb{P})$ by

$$(\mathbb{P}_{(h,d)}) \quad \begin{cases} \min_{(y_h,u_d)\in Y_h \times U_d} J_{(h,d)}(y,u) := \frac{1}{2}\|y_h - z_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_d\|^2_U \\ \text{s.t.} \\ \quad a(y_h, v_h) \;=\; \langle Bu_d, v_h \rangle_{Y^*,Y} \quad \text{for all } v_h \in Y_h, \\ \text{and} \\ u_d \in U_{\text{ad}}^d. \end{cases} \tag{2.4}$$

Here, we have set $a(y,v) := \int_\Omega \nabla y \nabla v dx$. Introducing the Finite Element stiffness matrix $A := (a_{ij})_{i,j=1}^n$, $a_{ij} := a(\phi_i, \phi_j)$, the Finite Element Mass matrix $M := (m_{ij})_{i,j=1}^{ng}$, $m_{ij} := \int_\Omega \phi_i \phi_j dx$, the matrix $E := (e_{ij})_{i=1,\dots,n;j=1,\dots,m}$, $e_{ij} = \langle Bu^j, \phi_i \rangle_{Y^*,Y}$, and the control mass matrix $F := (f_{ij})_{i,j=1}^m$, $f_{ij} := (u^i, u^j)_U$, allows us to rewrite $(\mathbb{P}_{(h,d)})$ in the form

$$(\mathbb{P}_{(n,m)}) \quad \begin{cases} \min_{(y,s)\in \mathbb{R}^n \times \mathbb{R}^m} Q(y,s) := \frac{1}{2}(y-z)^t M (y-z) + \frac{\alpha}{2} s^t F s \\ \text{s.t.} \\ \quad Ay \;=\; Es \\ \text{and} \\ s \in C. \end{cases} \tag{2.5}$$

This is now a finite dimensional optimization problem with quadratic objective, linear equality constraints, and admissibility characterized by the closed, convex set $C \subset \mathbb{R}^m$. Since the matrix A is spd, problem $(\mathbb{P}_{(n,m)})$ is equivalent to minimizing the reduced functional $\hat{Q}(s) := Q(A^{-1}Es, s)$ over the set C. And of course does $(\mathbb{P}_{(n,m)})$ admit a unique solution $(y, s) \in \mathbb{R}^n \times C$ which is characterized by the finite dimensional variational inequality

$$(\hat{Q}'(s), t - s)_{\mathbb{R}^m} \geq 0 \text{ for all } t \in C, \tag{2.6}$$

with $\hat{Q}'(s) = \alpha Fs + E^t A^{-t} M(A^{-1}Es - z) \equiv \alpha Fs + E^t p$, where $p := A^{-t}M(A^{-1}Es - z)$ denotes the discrete adjoint vector to whom we associate the discrete adjoint variable $p_h := \sum_{i=1}^{n} p_i \phi_i$. Comparing this with the expression for $\hat{J}'(u)$ from above, we note that the operator E takes the role the control operator B, and the inverse of the stiffness matrix A that of the solution operator S.

Problem $(\mathbb{P}_{(n,m)})$ now can be solved numerically with the help of appropriate solution algorithms, which should exploit the structure of the problem. We fix the following

NOTE 2.2. *In the First discretize, then optimize approach the discretization of the adjoint variable p is determined by the Ansatz for the discrete state y_h.*

In the *First optimize, then discretize* approach discussed next, this is different.

2.2. First optimize, the discretize. The starting point for the present approach is the system of first order necessary optimality conditions for problem $(\mathbb{P})$ stated next;

$$(\mathbb{OS}) \quad \begin{cases} -\Delta y = Bu & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \\ -\Delta p = y - z & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega, \\ (\alpha u + B^* p, v - u)_U \geq 0 & \text{for all } v \in U_{\text{ad}}. \end{cases} \tag{2.7}$$

Now we discretize everything related to the state y, the control u, and to functionals, integrals, and dualities as in Section 2.1. Further, we have the freedom to also select an appropriate discretization of the adjoint variable p. Here we choose continuous Finite Elements of order l on τ, which leads to the Ansatz $p_h(x) = \sum_{i=1}^{q} p_i \chi_i(x)$, where $\langle \chi_1, \ldots, \chi_q \rangle \subset Y$ denotes the Ansatz space for the adjoint variable. Forming the adjoint stiffness matrix $\tilde{A} := (\tilde{a}_{ij})_{i,j=1}^{q}$, $\tilde{a}_{ij} := a(\chi_i, \chi_j)$, the matrix $\tilde{E} := (\tilde{e}_{ij})_{i=1,\ldots,q;j=1,\ldots,m}$, $\tilde{e}_{ij} = \langle Bu^j, \chi_i \rangle_{Y^*,Y}$, and $T := (t_{ij})_{i=1,\ldots,n;j=1,\ldots,q}$, $t_{ij} := \int_\Omega \phi_i \chi_j dx$, the discrete analogon to $(\mathbb{OS})$ reads

$$(\mathbb{OS})_{(n,q,m)} \quad \begin{cases} Ay = Es, \\ \tilde{A}p = T(y - z), \\ (\alpha Fs + \tilde{E}^t p, t - s)_{\mathbb{R}^m} \geq 0 \text{ for all } t \in C. \end{cases} \tag{2.8}$$

Since the matrices A and $\tilde{A}$ are spd, this system is equivalent to the variational inequality

$$(\alpha Fs + \tilde{E}^t \tilde{A}^{-1} T(A^{-1}Es - z), t - s)_{\mathbb{R}^m} \geq 0 \text{ for all } t \in C. \tag{2.9}$$

Before we relate the approaches of Section 2.1 and Section 2.2 let us give some examples, compare Example 2.1.

EXAMPLE 2.3.

(1) $U := L^2(\Omega)$, $B : L^2(\Omega) \to H^{-1}(\Omega)$ *Injection,* , $U_{ad} := \{v \in L^2(\Omega); a \leq v(x) \leq b$ *a.e. in* $\Omega\}, a, b \in L^\infty(\Omega)$. *Further let* $k = l = 1$ *(linear Finite Elements for* y *and* p*),* $U_d := \langle u^1, \ldots, u^{nt} \rangle$*, where* $u^k_{|_{T_i}} = \delta_{ki}$ $(k, i = 1, \ldots, nt)$ *are piecewise constant functions (i.e.* $m = nt$*),* $C := \prod_{i=1}^{nt} [a_i, b_i]$*, where* $a_i := a(barycenter(T_i))$, $b_i := b(barycenter(T_i))$.

(2) *As in (1), but* $U_d := \langle u^1, \ldots, u^{ng} \rangle$*, where* $u^k_{|_{D_i}} = \delta_{ki}$ $(k, i = 1, \ldots, ng)$ *are piecewise constant functions (i.e.* $m = ng$*), with* D_i *denoting the patch associated to the vertex* P_i $(i = 1, \ldots, ng)$ *of the barycentric dual triangulation of* τ*,* $C := \prod_{i=1}^{ng} [a_i, b_i]$*, where* $a_i := a(P_i)$, $b_i := b(P_i)$.

(3) *As in (1), but* $U_d := \langle \phi_1, \ldots, \phi_{ng} \rangle$ *(i.e.* $m = ng$*),* $C := \prod_{i=1}^{ng} [a_i, b_i]$*, where* $a_i := a(P_i)$, $b_i := b(P_i)$*, with* P_i $(i = 1, \ldots, ng)$ *denote the vertices of the triangulation* τ.

(4) *(Compare Example 2.1): As in (1), but* $U := \mathbb{R}^m$, $B : \mathbb{R}^m \to H^{-1}(\Omega)$, $Bu := \sum_{j=1}^{m} u_j F_j$, $F_j \in H^{-1}(\Omega)$ *given* , $U_{ad} := \{v \in \mathbb{R}^m; a_j \leq v_j \leq b_j\}, a < b$*,* $U_d := \langle e_1, \ldots, e_m \rangle$ *with* $e_i \in \mathbb{R}^m$ $(i = 1, \ldots, m)$ *denoting the* $i-$*th unit vector,* $C := \prod_{i=1}^{ng} [a_i, b_i] \equiv U_d$.

2.3. Discussion and implications. Now let us compare the approaches of the two previous sections. It is clear that choosing the same Ansatz spaces for the state y and the adjoint variable p in the *First optimize, then discretize* approach leads to an optimality condition which is identical to that of the *First discretize, then optimize* approach in (2.6), since then $T \equiv M$. However, choosing a different approach for p in general leads to (2.9) with a non-symmetric matrix T, with the consequence that the matrix $\alpha F + \tilde{E}^t \tilde{A}^{-1} T A^{-1} E$ not longer represents a symmetric matrix. This is in contrast to the matrix $\hat{Q}''(s) = \alpha F + E^t A^{-1} M A^{-1} E$ of the *First discretize, then optimize* approach. Moreover, the expression $\alpha F s + \tilde{E}^t \tilde{A}^{-1} T (A^{-1} E s - z)$ in general does not represent a gradient, which is in contrast to $\hat{Q}'(s) = \alpha F s + E^t A^{-t} M (A^{-1} E s - z)$ which in fact is the gradient of the reduced finite dimensional functional $\hat{Q}(s)$.

In many situations the adjoint variable p is much more regular than the state y. For example, if z is a smooth function, the domain Ω has smooth boundary and B denotes the injection as in Example 2.1(1), the adjoint variable p admits two more weak derivatives than the state y, whose regularity in the control constrained case is restricted through the regularity of the control u, which in the present example is not better than $H^{1,r}$ for some $r \leq \infty$, no matter how smooth the boundary of Ω is. So it could be meaningful to use Ansatz functions with higher polynomial degree for p than for y.

There is up to now no general recipe which approach has to be preferred, and it should depend on the application and computational resources which approach to take for tackling the numerical solution of

the optimization problem. However, the numerical approach taken should to some extent reflect and preserve the structure which is inherent in the infinite dimensional optimization problem $(\mathbb{P})$. This can be best explained in the case without control constraints, i.e. $U_{\text{ad}} \equiv U$. Then the first order necessary optimality conditions for $(\mathbb{P})$ read

$$\hat{J}'(u) = \alpha u + B^* S^*(SBu - z) \equiv \alpha u + B^* p = 0 \text{ in } U.$$

Now let us for the moment consider Example 2.1(1), in which this equation becomes

$$\hat{J}'(u) = \alpha u + p = 0 \text{ in } L^2(\Omega).$$

For proceeding on the numerical level this identity clearly gives us the advice to relate to each other the discrete Ansätze for the control u and the adjoint variable p. This remains true also in the presence of control constraints, for which this smooth operator equation has to be replaced by the nonsmooth operator equation

$$u = P_{U_{\text{ad}}}\left(-\frac{1}{\alpha}p\right) \text{ in } L^2(\Omega), \tag{2.10}$$

where $P_{U_{\text{ad}}}$ denotes the orthogonal projection in U (here $= L^2(\Omega)$) onto the admissible set of controls. In any case, optimal control and corresponding adjoint state are related to each other, and this should be reflected by every numerical approach to be taken for the solution of problem $(\mathbb{P})$.

NOTE 2.4. *Controls should be discretized conservative, i.e. according to the relation between the adjoint state and the control given by the first order optimality condition. This rule should be obeyed in both, the First discretize, then optimize, and in the First optimize, then discretize approach.*

2.4. A structure exploiting discretization concept. The concepts presented in the subsequent subsections are introduced in [**41**]. Let us closer investigate (2.10) (in the general setting now) in terms of the simple fix-point iteration given next.

ALGORITHM 2.5.

- u *given*
- *do until convergence*
  $u^+ = P_{U_{ad}}\left(-\frac{1}{\alpha}B^* p(u)\right)$, $u = u^+$.

In this algorithm $p(u)$ is obtained by first solving $y = SBu$, and then $p = S^*(SBu - z)$. To obtain a discrete algorithm we now replace the solution operators S, S^* by their discrete counterparts S_h, S_h^* obtained by a Finite Element discretization, say. The discrete algorithm then reads

ALGORITHM 2.6.

- u *given*
- *do until convergence*
  $u^+ = P_{U_{ad}}\left(-\frac{1}{\alpha}B^* p_h(u)\right)$, $u = u^+$,

where $p_h(u)$ is obtained by first solving $y = S_h B u$, and then solving $p_h = S_h^*(S_h B u - z)$. We note that in this algorithm the control is not discretized. Only state and co-state are discretized. Two questions immediately arise.

(1) Is Algorithm 2.6 numerically implementable?
(2) Do Algorithms 2.5, 2.6 converge?

Let us first discuss question (2). Since both algorithms are fix-point algorithms, sufficient conditions for convergence are given by the relations $\alpha > \|B^* S^* S B\|_{\mathcal{L}(U)}$ for Algorithm 2.5, and by $\alpha > \|B^* S_h^* S_h B\|_{\mathcal{L}(U)}$ for Algorithm 2.6, since $P_{U_{\text{ad}}} : U \to U_{\text{ad}}$ denotes the orthogonal projection which is Lipschitz continuous with Lipschitz constant $L = 1$. However, we already know that (2.10) for every $\sigma > 0$ is equivalent (in the general setting) to the equation

$$G(u) = u - P_{U_{\text{ad}}}\left(u - \sigma \hat{J}'(u)\right) \equiv u - P_{U_{\text{ad}}}\left(u - \sigma(\alpha u + B^* p)\right) = 0 \text{ in } U, \tag{2.11}$$

so that we may apply a semi-smooth Newton algorithm, or a primal-dual active set strategy to its numerical solution. Since local convergence for these algorithms applied to 2.11 can be guaranteed for every choice of $\sigma > 0$ we in particular may set $\sigma := \frac{1}{\alpha}$. To anticipate discussion, for this choice of parameter we will obtain that the semi-smooth Newton method, and the primal-dual active set strategy are both numerically implementable in the discrete case.

Question (1) admits the answer *Yes*, whenever for given u it is possible to numerically evaluate the expression

$$P_{U_{\text{ad}}}\left(-\frac{1}{\alpha} B^* p_h(u)\right)$$

in the $i - th$ iteration of Algorithm 2.6 with an numerical overhead which is *independent* of the iteration counter of the algorithm. To illustrate this fact let us turn back to Example 2.1(1), i.e. $U = L^2(\Omega)$ and B denoting the injection, with $a \equiv const1$, $b \equiv const2$. In this case it is easy to verify that

$$P_{U_{\text{ad}}}(v)(x) = P_{[a,b]}(v(x)) = \max\{a, \min\{v(x), b\}\},$$

so that in every iteration of Algorithm 2.6 we have to form the control

$$u^+(x) = P_{[a,b]}\left(-\frac{1}{\alpha} p_h(x)\right), \tag{2.12}$$

which for in the one-dimensional setting is illustrated in Figure 2.4.

To construct the function u^+ it is sufficient to characterize the intersection of the bounds a, b (understood as constant functions) and the function $-\frac{1}{\alpha} p_h$ on every simplex T of the triangulation $\tau = \tau_h$. For piecewise linear finite element approximations of p we have the following theorem.

THEOREM 2.7. *Let u^+ denote the function of* (2.12)*, with p_h denoting a piecewise linear, continuous finite element function, and constant bounds $a < b$. Then there exists a partition $\kappa_h = \{K_1, \ldots K_{l(h)}\}$*

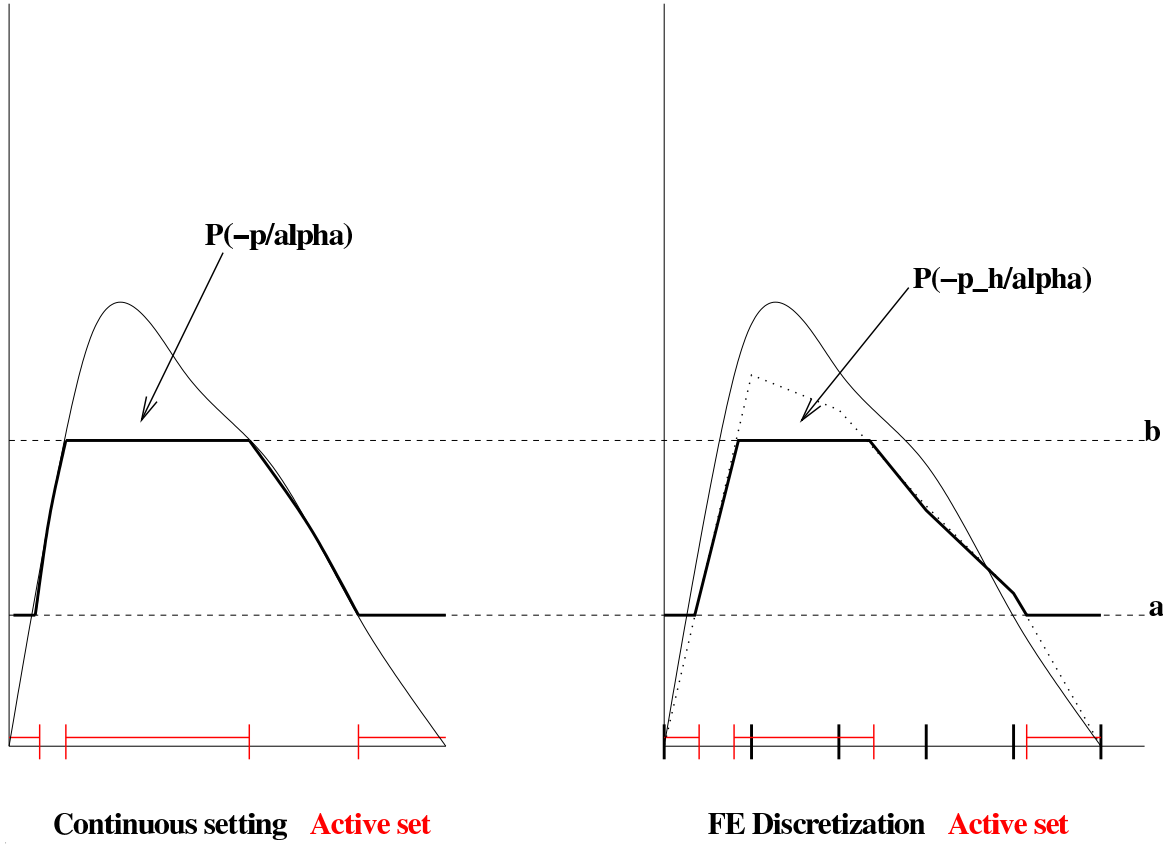


FIGURE 2.1. Projection of $-\frac{1}{\alpha}p$ (left) and $-\frac{1}{\alpha}p_h$ (right) in one space dimension, p_h discretized with linear finite elements. Finite element grid given by black bars.

of Ω such that u^+ restricted to K_j $(j = 1, \ldots, l(h))$ is a polynomial either of degree zero or one. For $l(h)$ there holds

$$l(h) \leq Cnt(h),$$

with a positive constant $C \leq 3$ and $nt(h)$ denoting the number of simplexes in τ_h. In particular, the vertices of the discrete active set associated to u^+ need not coincide with finite element nodes.

Proof: Abbreviate $\xi_h^a := -\frac{1}{\alpha}p_h^* - a$, $\xi_h^b := b - \frac{1}{\alpha}p_h^*$ and investigate the zero level sets 0_h^a and 0_h^b of ξ_h^a and ξ_h^b, respectively.

Case $n = 1$: $0_h^a \cap T_i$ is either empty or a point $S_i^a \in T_i$. Every point S_i^a subdivides T_i into two sub-intervals. Analogously $0_h^b \cap T_i$ is either empty or a point $S_i^b \in T_i$. Further $S_i^a \neq S_i^b$ since $a < b$. The maximum number of sub-intervals of T_i induced by 0_h^a and 0_h^b therefore is equal to three. Therefore, $l(h) \leq 3nt(h)$, i.e. $C = 3$.

Case $n = 2$: $0_h^a \cap T_i$ is either empty or a vertex of τ_h or a line $L_i^a \subset T_i$, analogously $0_h^b \cap T_i$ is either empty or a vertex of τ_h or a line $L_i^b \subset T_i$. Since $a < b$ the lines L_i^a and L_i^b do not intersect. Therefore, similar considerations as in the case $n = 1$ yield $C = 3$.

Case $n \in \mathbb{N}$: $0_h^a \cap T_i$ is either empty or a part of a $k-$dimensional hyperplane $(k < n)$ $L_i^a \subset T_i$, analogously $0_h^b \cap T_i$ is either empty or a part of $k-$dimensional hyperplane $(k < n)$ $L_i^b \subset T_i$. Since $a < b$ the surfaces L_i^a and L_i^b do not intersect. Therefore, similar considerations as in the case $n = 2$ yield $C = 3$. This completes the proof.

It is now clear that the proof of the previous theorem easily extends to functions p_h which are piecewise polynomials of degree $k \in \mathbb{N}$, and bounds a, b which are piecewise polynomials of degree $l \in \mathbb{N}$ and $m \in \mathbb{N}$, respectively, since the difference of a, b and p_h in this case also represents a piecewise polynomial function whose projection on every element can be easily characterized.

We now have that Algorithm 2.6 is numerically implementable, but only converges for a certain parameter range of α. A locally (super-linearly) convergent algorithm for the numerical solution of equation (2.11) is the semi-smooth Newton method, since the function G is semi-smooth in the sense of [**77**, Example 5.6].

Before we proceed let us define

$$\hat{J}_h(u) := J(S_h B u, u), \quad u \in U$$

and consider the following infinite dimensional optimization problem

$$\min_{u \in U_{\text{ad}}} \hat{J}_h(u). \tag{2.13}$$

According to (2.2) this problem admits a unique solution $u_h \in U_{\text{ad}}$ which is characterized by the variational inequality

$$(\hat{J}_h'(u_h), v - u_h)_U \geq 0 \text{ for all } v \in U_{\text{ad}}, \tag{2.14}$$

which in turn is equivalent to the non-smooth operator equation (compare (2.11))

$$G_h(u) = u - P_{U_{\text{ad}}}\left(u - \sigma \hat{J}_h'(u)\right) \equiv u - P_{U_{\text{ad}}}\left(u - \sigma(\alpha u + B^* p_h)\right) = 0 \text{ in } U,$$

where similar as above

$$J_h'(u) = \alpha u + B^* S_h^*(S_h B u - z) \equiv \alpha u + B^* p_h(u).$$

The considerations made above now imply that the unique solution u_h of the infinite dimensional optimization problem (2.13) can be numerically computed either by Algorithm 2.6 (for α large enough), or by a semi-smooth Newton method (since the function G_h also is semi-smooth), however in both cases *without* a further discretization step.

2.5. Error estimates. Next let us investigate the error $\|u - u_h\|_U$ between the solutions u of (2.3) and u_h of (2.13).

THEOREM 2.8. *Let u denote the unique solution of* (2.2)*, and u_h the unique solution of* (2.13)*. Then there holds*

$$\|u-u_h\|_U^2 \le \frac{1}{\alpha}\left\{(B^*(p(u)-\tilde{p}_h(u)),u_h-u)_U + \right. \\ \left. + \int_\Omega (y_h(u_h)-y_h(u))(y(u)-y_h(u))dx \right\}, \tag{2.15}$$

where $\tilde{p}_h(u) := S_h^*(SBu-z)$, $y_h(u) := S_hBu$, and $y(u) := SBu$.

Proof: Since (2.13) is an optimization problem defined on all of U_{ad}, the unique solution u of (2.2) is an admissible test function in (2.14). Let us emphasize, that this is different for approaches, where the control space is discretized explictly. In this case we may only expect that u_h is an admissible test function for the continuous problem (if ever). So let us test (2.3) with u_h, and (2.14) with u, and then add the resulting variational inequalities. This leads to

$$(\alpha(u-u_h)+B^*S^*(SBu-z)-B^*S_h^*(S_hBu_h-z),u_h-u)_U \ge 0.$$

This inequality is equivalent to

$$\alpha\|u-u_h\|_U^2 \le \\ \le (B^*(p(u)-\tilde{p}_h(u))+B^*(\tilde{p}_h(u)-p_h(u))+B^*(p_h(u)-p_h(u_h)),u_h-u)_U .$$

Let us investigate the third addend on the right hand side of this inequality. By definition of the adjoint variables there holds

$$(B^*(p_h(u)-p_h(u_h),u_h-u)_U = \langle p_h(u)-p_h(u_h),B(u_h-u)\rangle_{Y^*,Y} = \\ = a(y_h(u_h)-y_h(u),p_h(u)-p_h(u_h)) = \int_\Omega (y_h(u_h)-y_h(u))(y_h(u)-y_h(u_h))dx = \\ = -\|y_h(u)-y_h(u_h)\|_{L^2(\Omega)}^2 \le 0.$$

Furthermore, for the second addend we have

$$(B^*(\tilde{p}_h(u)-p_h(u),u_h-u)_U = \langle \tilde{p}_h(u)-p_h(u),B(u_h-u)\rangle_{Y^*,Y} = \\ = a(y_h(u_h)-y_h(u),\tilde{p}_h(u)-p_h(u)) = \int_\Omega (y_h(u_h)-y_h(u))(y(u)-y_h(u))dx,$$

so that the claim of the theorem follows.

What are the consequences of Theorem 2.15? From the structure of this estimate we immediately infer that an error estimate for $\|u-u_h\|_U$ is at hand, if

- an error estimate for $\|B^*(p(u)-\tilde{p}_h(u)\|_U$ is available, and
- the mapping $u \mapsto y_h(u)$ from U to $L^2(\Omega)$ is Lipschitz continuous, and
- an error estimate for $\|y(u)-y_h(u)\|_{L^2(\Omega)}$ is available.

This means, that the error of $\|u - u_h\|_U$ is completely determined by the approximation properties of the discrete solution operators S_h and S_h^*.

NOTE 2.9. *The error $\|u-u_h\|_U$ between the solution u of problem* (2.2) *and u_h of* (2.13) *is completely determined by the approximation properties of the discrete solution operators S_h and S_h^*.*

Let us revisit Example 2.1. Then $U = L^2(\Omega)$ and B denotes the injection. Then $y = SBu \in H^2(\Omega) \cap H_0^1(\Omega)$ (if for example $\Omega \in C^{1,1}$ or Ω convex). Let us estimate the two addenda on the right side of the inequality sign in (2.15). There holds

$$(B^*(p(u) - \tilde{p}_h(u)), u - u_h)_U = \int_\Omega (p(u) - \tilde{p}_h(u))(u - u_h)dx \leq$$
$$\leq \|p(u) - \tilde{p}_h(u)\|_{L^2(\Omega)}\|u - u_h\|_{L^2(\Omega)} \leq ch^2\|y(u)\|_{L^2(\Omega)}\|u - u_h\|_{L^2(\Omega)},$$

and

$$\int_\Omega (y_h(u) - y_h(u_h))(y(u) - y(u_h))dx \leq$$
$$\leq \|y(u) - y_h(u)\|_{L^2(\Omega)}\|y_h(u) - y_h(u_h)\|_{L^2(\Omega)} \leq$$
$$\leq ch^2\|u\|_{L^2(\Omega)}\|y_h(u) - y_h(u_h)\|_{L^2(\Omega)}.$$

To obtain an error estimate of the form $\|u-u_h\|_{L^2(\Omega)} \leq ch^2$ it remains to show the Lipschitz continuity of y_h w.r.t. the control u which is easy to verify in the following way.

$$\int_\Omega |y_h(u) - y_h(u_h)|^2 dx \leq c_p^2 a(y_h(u) - y_h(u_h), y_h(u) - y_h(u_h)) =$$
$$= c_p^2 \int_\Omega (y_h(u) - y_h(u_h))(u - u_h)dx \leq$$
$$\leq c_p^2\|y_h(u) - y_h(u_h)\|_{L^2(\Omega)}\|u - u_h\|_{L^2(\Omega)}.$$

Here c_p denotes the constant of the Poincaré inequality. Combining these estimates we immediately obtain

THEOREM 2.10. *Let u and u_h denote the solutions of problem* (2.2) *and* (2.13)*, respectively in the setting of Example 2.1(1). Then there holds*

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2 \left\{\|y(u)\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}\right\}.$$

And this theorem is also valid for the setting of Example 2.1(2) if we require $F_j \in L^2(\Omega)$ $(j = 1, \ldots, m)$. This is an easy consequence of the fact that for a function $z \in H^{-1}(\Omega)$ there holds $B^*z \in \mathbb{R}^m$ with $(B^*z)_i = \langle F_i, z\rangle_{Y^*,Y}$ for $i = 1, \ldots, m$.

THEOREM 2.11. *Let u and u_h denote the solutions of problem* (2.2) *and* (2.13)*, respectively in the setting of Example 2.1(2). Then there holds*

$$\|u - u_h\|_{\mathbb{R}^m} \leq ch^2 \left\{\|y(u)\|_{L^2(\Omega)} + \|u\|_{\mathbb{R}^m}\right\},$$

where the positive constant now depends on the functions F_j $(j = 1, \dots, m)$.

Proof: It suffices to estimate

$$(B^*(p(u) - \tilde{p}_h(u)), u - u_h)_{\mathbb{R}^m} =$$

$$= \sum_{j=1}^{m} \left\{ \int_\Omega F_j(p(u) - \tilde{p}_h(u))dx(u - u_h)_j \right\} \leq$$

$$\leq \|p(u) - \tilde{p}_h(u)\|_{L^2(\Omega)} \left(\sum_{j=1}^{m} \int_\Omega |F_j|^2 dx \right)^{\frac{1}{2}} \|u - u_h\|_{\mathbb{R}^m} \leq$$

$$\leq ch^2 \|y(u)\|_{L^2(\Omega)} \|u - u_h\|_{\mathbb{R}^m}.$$

The reminder terms can be estimated as above.

2.6. Boundary control. The structure of all considerations of the previous subsections remain valid also for inhomogeneous Neumann and Dirichlet boundary control problems. Let us consider the model problems

$$(\mathbb{NC}) \quad \begin{cases} \min_{(y,u)\in Y\times U} J(y,u) := \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U \\ \text{s.t.} \\ \quad -\Delta y = 0 \quad \text{in } \Omega, \\ \quad \partial_\eta y = Bu - \gamma y \quad \text{on } \partial\Omega, \\ \text{and} \\ u \in U_{\text{ad}} \subseteq U, \end{cases} \tag{2.16}$$

and

$$(\mathbb{DC}) \quad \begin{cases} \min_{(y,u)\in Y\times U} J(y,u) := \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U \\ \text{s.t.} \\ \quad -\Delta y = 0 \quad \text{in } \Omega, \\ \quad y = Bu \quad \text{on } \partial\Omega, \\ \text{and} \\ u \in U_{\text{ad}} \subseteq U, \end{cases} \tag{2.17}$$

where in both cases $B : U \to L^2(\Gamma)$ with $\Gamma := \partial\Omega$. To anticipate the discussion we note that the Dirichlet problem for y in $(\mathbb{DC})$ for $Bu \in L^2(\Gamma)$ is understood in the very weak sense.

2.6.1. *Neumann and Robin-type boundary control.* Let us first consider problem $(\mathbb{NC})$ which equivalently can be rewritten in the form

$$\min_{u\in U_{\text{ad}}} \hat{J}(u) \tag{2.18}$$

for the reduced functional $\hat{J}(u) := J(y(u), u) \equiv J(SBu, u)$ over the set U_{ad}, where $S : Y^* \to Y$ for $Y := H^1(\Omega)$ denotes the weak solution operator of the Neumann boundary value problem for $-\Delta$,

i.e. $y = Sf$ iff

$$a(y,v) := \int_\Omega \nabla y \nabla v dx + \int_\Gamma \gamma y v d\Gamma = \langle f, v\rangle_{Y^*,Y} \text{ for all } v \in Y,$$

and the action of $Bu \in L^2(\Gamma)$ as an element $EBu \in Y^*$ is defined by

$$\langle EBu, v\rangle_{Y^*,Y} := \int_\Gamma Buv d\Gamma \text{ for all } v \in Y.$$

We further know that the first order necessary (and here also sufficient) optimality conditions here take the form

$$(\hat{J}'(u), v-u)_U \geq 0 \text{ for all } v \in U_{\text{ad}} \tag{2.19}$$

where $\hat{J}'(u) = \alpha u + B^*E^*S^*(SEBu - z) \equiv \alpha u + B^*E^*p$, with $p := S^*(SEBu - z)$ denoting the adjoint variable. Here $E^* : Y \to L^2(\Gamma)$ denotes the trace operator. From here onwards let us not longer distinguish between B and EB. The function p in our setting satisfies the following Poisson problem with Neumann (Robin-type) boundary conditions;

$$\begin{aligned} -\Delta p &= y - z \quad \text{in } \Omega, \\ \partial_\eta p + \gamma p &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

We now define the discrete analogon to problem (2.18) as in the previous subsection;

$$\min_{u\in U_{\text{ad}}} \hat{J}_h(u), \tag{2.20}$$

where for $u \in U$ we set $\hat{J}_h(u) := J(S_hBu, u)$ with S_h denoting the discrete analogon of S. According to (2.18) this problem admits a unique solution $u_h \in U_{\text{ad}}$ which is characterized by the variational inequality

$$(\hat{J}_h'(u_h), v-u_h)_U \geq 0 \text{ for all } v \in U_{\text{ad}}, \tag{2.21}$$

where similar as above

$$J_h'(u) = \alpha u + B^*S_h^*(S_hBu - z) \equiv \alpha u + B^*p_h(u).$$

We notice that the whole exposition can be done by *copy and paste* from Section 2.4, the structure of the optimization problem its discretization does not depend on where control is applied. It is completely characterized by the operators S, S_h, and B (as well as by E). For Neumann boundary control the analogon to Theorem 2.8 reads

THEOREM 2.12. *Let u denote the unique solution of* (2.18)*, and u_h the unique solution of* (2.20)*. Then there holds*

$$\|u-u_h\|_U^2 \leq \frac{1}{\alpha}\left\{(B^*(p(u) - \tilde{p}_h(u)), u_h - u)_U + \right. \\ \left. + \int_\Omega (y_h(u_h) - y_h(u))(y(u) - y_h(u))dx \right\}, \tag{2.22}$$

where $\tilde{p}_h(u) := S_h^*(SBu - z)$, $y_h(u) := S_hBu$, and $y(u) := SBu$.

The proof of this theorem is analogous to that of Theorem 2.8 and is left as an exercise.

2.6.2. *Dirichlet boundary control.* Now we switch to problem $(\mathbb{DC})$ which equivalently can be rewritten in the form

$$\min_{u\in U_{\text{ad}}} \hat{J}(u) \tag{2.23}$$

for the reduced functional $\hat{J}(u) := J(y(u),u) \equiv J(SBu,u)$ over the set U_{ad}, where $S : Y^* \to L^2(\Omega)$ for $Y := H^2(\Omega)\cap H_0^1(\Omega)$ denotes the the very-weak solution operator of the Dirichlet boundary value problem for $-\Delta$, i.e. for $f \in Y^*$ and $u \in U$ there holds $y = S(f + EBu)$ iff

$$a(y,v) := \int_\Omega y(-\Delta v)dx = \langle f, v\rangle_{Y^*,Y} - \int_\Gamma Bu\partial_\eta v d\Gamma \text{ for all } v \in Y.$$

Here, the action of $Bu \in L^2(\Gamma)$ as an element $EBu \in Y^*$ is defined by

$$\langle EBu, v\rangle_{Y^*,Y} := \int_\Gamma Bu\partial_\eta v d\Gamma \text{ for all } v \in Y.$$

The first order necessary (and here also sufficient) optimality conditions here again take the form

$$(\hat{J}'(u), v-u)_U \geq 0 \text{ for all } v \in U_{\text{ad}} \tag{2.24}$$

where $\hat{J}'(u) = \alpha u - B^*E^*S^*(SEBu - z) \equiv \alpha u - B^*E^*p$, with $p := S^*(SEBu - z)$ denoting the adjoint variable. Here $E^* : Y \to L^2(\Gamma)$ denotes the trace operator of first order, i.e. for $v \in Y$ there holds $E^*v = (\partial_\eta v)_{|\Gamma}$. From here onwards let us not longer distinguish between B and EB, so that $\hat{J}'(u) = \alpha u - B^*\partial_\eta p$. The function p in our setting satisfies the following Poisson problem with homogeneous Dirichlet boundary conditions;

$$\begin{aligned} -\Delta p &= y - z \quad \text{in } \Omega, \\ p &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

To define an appropriate discrete approach for (2.23) in the present situation is a little bit more involved due to the following fact.

NOTE 2.13. *We intend to approximate the solution y of the Dirichlet boundary value problem in* (2.23) *and the adjoint variable p by piecewise polynomials y_h and p_h of order k greater or equal to one, say. Then it is clear that it makes no sense to prescribe boundary values for y_h represented by (restrictions of) piecewise polynomials of order $k-1$. However, the discrete analogon of the variational inequality* (2.24) *exactly proposes this, since $\partial_\eta p_h$ is a piecewise polynomial of order $k-1$ on Γ.*

We introduce the L^2 projection Π_h onto boundary functions which are piecewise polynomial of degree $k \geq 1$ and continuous on the boundary grid induced by triangulation of Ω on the boundary Γ. For $v \in L^2(\Gamma)$ we define $\Pi_h v$ to be the continuous, piecewise polynomial of degree k defined by the relation

$$\int_\Gamma \Pi_h v w_h d\Gamma = \int_\Gamma v w_h d\Gamma \text{ for all } w_h \in \text{trace}(W_h),$$

where W_h is defined in Section 2.1. The numerical approximation $S_hBu := y_h \in W_h$ of the very weak solution y of the state equation with boundary values Bu is defined by the relation

$$\int_\Omega \nabla y_h \nabla v_h dx = 0 \text{ for all } v_h \in Y_h, \text{ and } y_h = \Pi_h(Bu) \text{ on } \Gamma,$$

and the numerical approximation p_h of the adjoint variable p as the usual finite element approximation $p_h := S_h^*(S_hE(Bu) - z)$, i.e.

$$\int_\Omega \nabla p_h \nabla v_h dx = \int_\Omega (y_h - z)v_h dx \text{ for all } v_h \in Y_h.$$

The discrete analogon of the optimization problem (2.23) reads

$$\min_{u\in U_{\text{ad}}} \hat{J}_h(u), \tag{2.25}$$

where for $u \in U$ we set $\hat{J}_h(u) := J(S_hBu, u)$ with S_h denoting the discrete analogon to S. It admits a unique solution $u_h \in U_{\text{ad}}$. After a short calculation we obtain for $u \in U_{\text{ad}}$

$$J_h'(u) = \alpha u - B^*\partial_\eta p_h(u),$$

where the discrete flux $\partial_\eta p_h(u)$ in the latter equation is a continuous, piecewise polynomial function of degree k on the boundary grid defined through the relation

$$\int_\Gamma \partial_\eta p_h(u) w_h d\Gamma := \int_\Omega \nabla p_h \nabla w_h dx - \int_\Omega (y_h(u) - z)w_h dx \text{ for all } w_h \in W_h. \tag{2.26}$$

The unique solution $u_h \in U_{\text{ad}}$ of problem (2.25) satisfies the variational inequality

$$(J_h'(u_h), v - u_h)_U \geq 0 \text{ for all } v \in U_{\text{ad}}, \tag{2.27}$$

which also represents a sufficient condition for u_h to solve problem (2.25). For Dirichlet boundary control the analogon to Theorem 2.8 reads

THEOREM 2.14. *Let u denote the unique solution of* (2.23)*, and u_h the unique solution of* (2.27)*. Then there holds*

$$\|u - u_h\|_U^2 \leq \frac{1}{\alpha}\Big\{ -\Big(B^*(\partial_\eta p(u) - \widetilde{\partial_\eta p_h(u)}), u_h - u\Big)_U + \\ - \int_\Omega (y_h(u_h) - y_h(u))(y(u) - y_h(u))dx \Big\}, \tag{2.28}$$

where $\widetilde{\partial_\eta p_h(u)}$ denotes the discrete flux associated to $y(u) = SBu$, and $y_h(u) := S_hBu$.

Proof: We test equation (2.24) with u_h, equation (2.27) with the solution u of problem (2.23), and add the variational inequalities (2.24) and (2.27). This leads to

$$\alpha\|u - u_h\|_U \leq \\ \leq -\left(B^*(\partial_\eta p(u) - \partial_\eta p_h(u)), u_h - u\right)_U - \left(B^*(\partial_\eta p_h(u) - \partial_\eta p_h(u_h)), u_h - u\right)_U.$$

From the definition of B, Π_h and of S_h it follows that

$$-\left(B^*(\partial_\eta p_h(u) - \partial_\eta p_h(u_h)), u_h - u\right)_U = -\|y_h(u_h) - y_h(u)\|^2_{L^2(\Omega)} \leq 0.$$

Further,

$$\left(B^*(\partial_\eta p(u) - \partial_\eta p_h(u)), u_h - u\right)_U =$$
$$= \left(B^*(\partial_\eta p(u) - \widetilde{\partial_\eta p_h(u)}), u_h - u\right)_U + \left(B^*(\widetilde{\partial_\eta p_h(u)} - \partial_\eta p_h(u)), u_h - u\right)_U,$$

and, by the definition of Π_h and the discrete fluxes,

$$\left(B^*(\widetilde{\partial_\eta p_h(u)} - \partial_\eta p_h(u)), u_h - u\right)_U = \int_\Gamma (y_h(u_h) - y_h(u))(\widetilde{\partial_\eta p_h(u)} - \partial_\eta p_h(u))d\Gamma =$$
$$= \int_\Omega \nabla(y_h(u_h) - y_h(u))\nabla(\tilde{p}_h(u) - p_h(u))dx - \int_\Omega (y(u) - y_h(u))(y_h(u_h) - y_h(u))dx.$$

Since $\tilde{p}_h(u) = p_h(u) = 0$ on Γ, there holds

$$\int_\Omega \nabla(y_h(u_h) - y_h(u))\nabla(\tilde{p}_h(u) - p_h(u))dx = 0,$$

so that the claim of the theorem follows.

Let us finally emphasize the similarities in Theorem 2.8, Theorem 2.12 and Theorem 2.14. All three estimates contain the term

$$\int_\Omega (y_h(u_h) - y_h(u))(y(u) - y_h(u))dx,$$

which stems from the first part of the cost functional in problem (2.1). But also the first addend of the right-hand sides in (2.15), (2.22), and (2.28) have a very similar structure. They may be rewritten as

$$\left(B^\backprime p(u) - \tilde{p}_h(u)), u_h - u\right)_U$$

in estimates (2.15) and (2.22), where $E : Y \to Y$ denotes the identity in (2.15), and as

$$-\left(B^*(\partial_\eta p(u) - B^*\widetilde{\partial_\eta p_h(u)}), u_h - u\right)_U$$

in estimate (2.28).

2.7. Numerical examples. In the present section we present numerical results for the discrete approach presented in the previous subsections, and also numerical comparisons to other commonly used discrete approaches. Let us begin with the following distributed control problem.

EXAMPLE 2.15. (Distributed control)
We consider problem (2.1) with Ω denoting the unit circle, $U_{\text{ad}} := \{v \in L^2(\Omega); -0.2 \leq u \leq 0.2\} \subset L^2(\Omega)$ and $B : L^2(\Omega) \to Y^*(\equiv H^{-1}(\Omega))$ the injection. Further we set $z(x) := (1 - |x|^2)x_1$ and $\alpha = 0.1$. The numerical discretization of state and adjoint state is performed with linear, continuous finite elements.

Here we consider the scenario that the exact solution of the problem is not known in advance (although it is easy to construct example problems where exact state, adjoint state and control are known, see [**74**]). Instead we use the numerical solutions computed on a grid with $h = \frac{1}{256}$ as references. To present numerical results it is convenient to introduce the *Experimental Order of Convergence*, brief EOC, which for some positive error functional E is defined by

$$\text{EOC} := \frac{\ln E(h_1) - \ln E(h_2)}{\ln h_1 - \ln h_2}. \tag{2.29}$$

Fig. 2.15 presents the numerical results for $h = \frac{1}{8}$. Fig. 2.15 presents a numerical comparison for active sets obtained by the numerical approach discussed so far, and obtained by a conventional approach which uses piecewise linear, continuous finite elements also for the a-priori discretization of controls. We observe a significant better resolution of active sets by the approach presented in the previous subsections. In Tables 2.1 - 2.3 the experimental order of convergence for different error functionals is presented for the state, adjoint state, and control. We use the abbreviations $E_{y_{L2}}$ for the error in the L^2-norm, $E_{y_{sup}}$ for the error in the L^∞-norm, $E_{y_{sem}}$ for the error in the H^1-seminorm, and $E_{y_{H_1}}$ for the error in the H^1-norm. Table 2.4 presents the results for the controls of the conventional approach which should be compared to the numbers of Table 2.3. Table 2.5 presents the order of convergence of the active sets for the approach presented here and for the conventional approach. As error functional we use in this case the area

$$E_a := |(A \setminus A_h) \cup (A_h \setminus A)|$$

of the symmetric difference of discrete and continuous active sets. EOC with the corresponding subscripts denotes the associated experimental order of convergence.

As a result we obtain, that the approach presented here provides a much better approximation of the controls and active sets than the conventional approach. In particular the errors in the L^2- and L^∞-norm are much smaller than the corresponding ones in the conventional approach. Let us also note that the results in the conventional approach would become even more worse if we would use piecewise constants as Ansatz for the controls. For theoretical and numerical results of conventional approaches let us refer to [**6**].

Let us note that similar numerical results can be obtained by an approach of Meyer and Rösch presented in [**56**]. The authors in a preliminary step compute a piecewise constant optimal control $\bar{u}$ and with its help compute in a post-processing step a projected control u through

$$u = P_{U_{\text{ad}}}(-\frac{1}{\alpha} B^* p_h(\bar{u})).$$

However, the numerical analysis of their approach underlies much more restrictions than the approach presented here, and requires assumptions on the $d-1$-dimensional Hausdorff measure of the discrete active set induced by the optimal control.

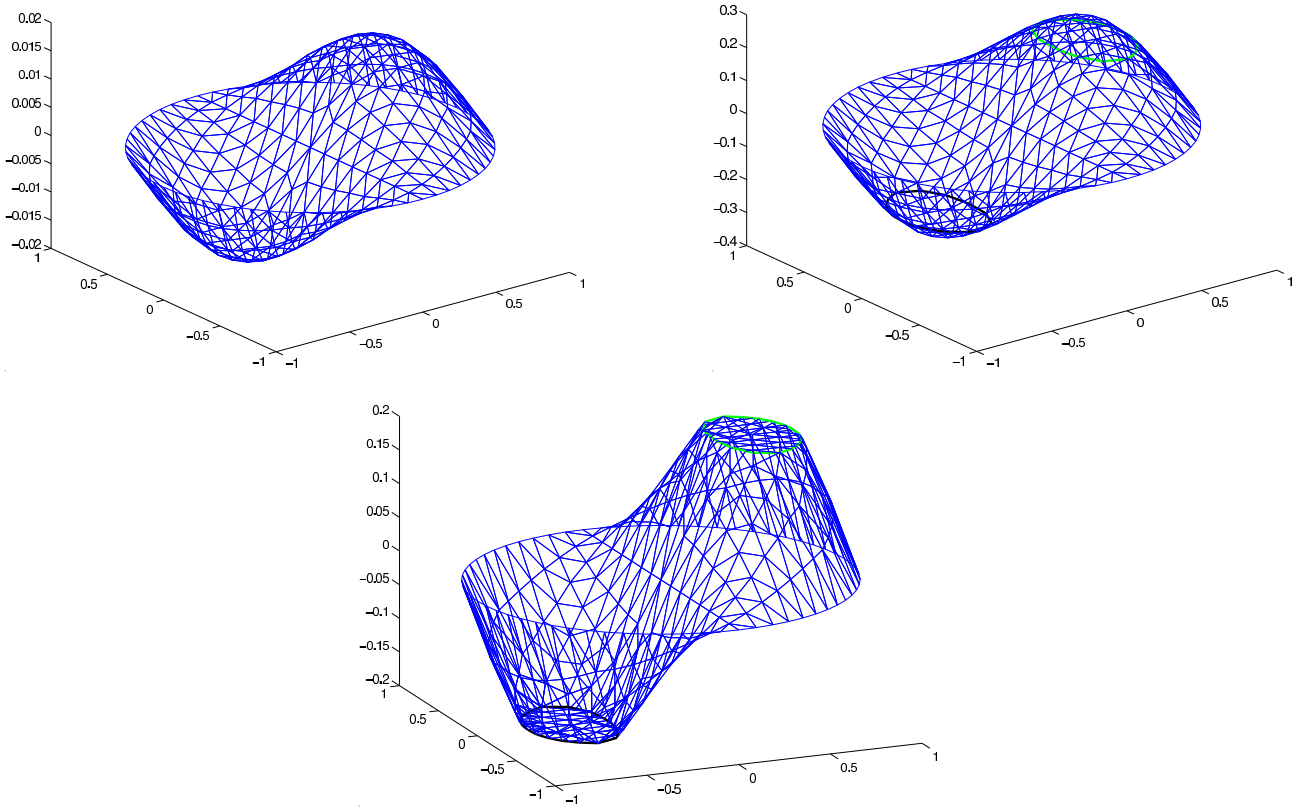


FIGURE 2.2. Numerical results of distributed control: Optimal state (left), corresponding adjoint state (middle) and associated optimal control (right). The black and green lines, respectively depict the boarders of the active set.

h	$E_{y_{L2}}$	$E_{y_{sup}}$	$E_{y_{sem}}$	$E_{y_{H_1}}$	$EOC_{y_{L2}}$	$EOC_{y_{sup}}$	$EOC_{y_{sem}}$	$EOC_{y_{H_1}}$
1/1	1.47e-2	1.63e-2	5.66e-2	5.85e-2	-	-	-	-
1/2	5.61e-3	6.02e-3	2.86e-2	2.92e-2	1.39	1.44	0.98	1.00
1/4	1.47e-3	1.93e-3	1.38e-2	1.39e-2	1.93	1.64	1.06	1.08
1/8	3.83e-4	5.02e-4	6.89e-3	6.90e-3	1.94	1.95	1.00	1.01
1/16	9.65e-5	1.26e-4	3.44e-3	3.45e-3	1.99	2.00	1.00	1.00
1/32	2.40e-5	3.14e-5	1.71e-3	1.71e-3	2.01	2.00	1.01	1.01
1/64	5.73e-6	7.78e-6	8.37e-4	8.37e-4	2.06	2.01	1.03	1.03
1/128	1.16e-6	1.85e-6	3.74e-4	3.74e-4	2.30	2.07	1.16	1.16

TABLE 2.1. Errors (columns left) and EOC (columns right) of state for different error functionals. As reference solution y_h for $h = \frac{1}{256}$ is taken.

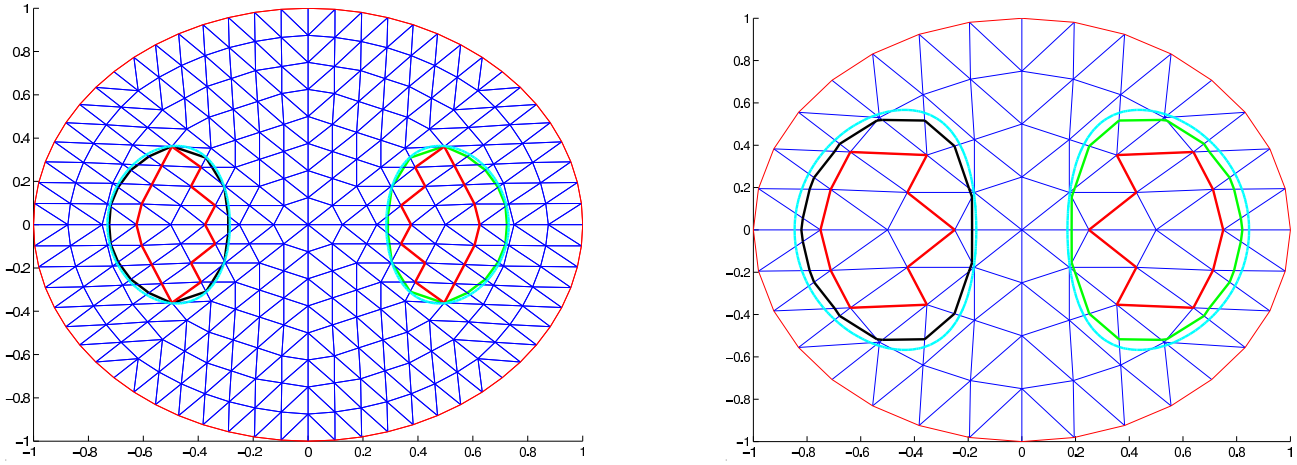


FIGURE 2.3. Numerical comparison of active sets obtained by the approach presented here, and those obtained by a conventional approach with piecewise linear, continuous controls: $h = \frac{1}{8}$ and $\alpha = 0.1$ (left), $h = \frac{1}{4}$ and $\alpha = 0.01$ (right). The red line depicts the boarder of the active set in the conventional approach, the cyan line the exact boarder, the black and green lines, respectively the boarders of the active set in the approach presented here.

h	$E_{p_{L2}}$	$E_{p_{sup}}$	$E_{p_{sem}}$	$E_{p_{H_1}}$	$EOC_{p_{L2}}$	$EOC_{p_{sup}}$	$EOC_{p_{sem}}$	$EOC_{p_{H_1}}$
1/1	2.33e-2	2.62e-2	8.96e-2	9.26e-2	-	-	-	-
1/2	6.14e-3	7.75e-3	4.36e-2	4.40e-2	1.92	1.76	1.04	1.07
1/4	1.59e-3	2.50e-3	2.17e-2	2.18e-2	1.95	1.64	1.00	1.02
1/8	4.08e-4	6.52e-4	1.09e-2	1.09e-2	1.97	1.94	0.99	0.99
1/16	1.03e-4	1.64e-4	5.48e-3	5.48e-3	1.99	1.99	1.00	1.00
1/32	2.54e-5	4.14e-5	2.73e-3	2.73e-3	2.01	1.99	1.01	1.01
1/64	6.11e-6	1.04e-5	1.33e-3	1.33e-3	2.06	1.99	1.03	1.03
1/128	1.27e-6	2.61e-6	5.96e-4	5.96e-4	2.27	1.99	1.16	1.16

TABLE 2.2. Errors (columns left) and EOC (columns right) of adjoint state for different error functionals. As reference solution p_h for $h = \frac{1}{256}$ is taken.

EXAMPLE 2.16. (Robin-type boundary control)
Now we consider Robin-type boundary control and in particular compare the approach presented here with that taken by Casas et al. in [**19**]. In order to compare our numerical results to exact solutions we consider an optimal control problem which slightly differs from that formulated in (2.16). The following example is taken from [**19**]. The computational domain is the unit square $\Omega := [0, 1]^2 \subset \mathbb{R}^2$.

h	$E_{u_{L2}}$	$E_{u_{sup}}$	$E_{u_{sem}}$	$E_{u_{H_1}}$	$EOC_{u_{L2}}$	$EOC_{u_{sup}}$	$EOC_{u_{sem}}$	$EOC_{u_{H_1}}$
1/1	2.18e-1	2.00e-1	8.66e-1	8.93e-1	-	-	-	-
1/2	5.54e-2	7.75e-2	4.78e-1	4.81e-1	1.97	1.37	0.86	0.89
1/4	1.16e-2	2.30e-2	2.21e-1	2.22e-1	2.25	1.75	1.11	1.12
1/8	3.02e-3	5.79e-3	1.15e-1	1.15e-1	1.94	1.99	0.94	0.95
1/16	7.66e-4	1.47e-3	6.09e-2	6.09e-2	1.98	1.98	0.92	0.92
1/32	1.93e-4	3.67e-4	2.97e-2	2.97e-2	1.99	2.00	1.03	1.03
1/64	4.82e-5	9.38e-5	1.41e-2	1.41e-2	2.00	1.97	1.07	1.07
1/128	1.17e-5	2.37e-5	6.40e-3	6.40e-3	2.04	1.98	1.14	1.14

TABLE 2.3. Errors (columns left) and EOC (columns right) of control for different error functionals. As reference solution u_h for $h = \frac{1}{256}$ is taken.

h	$E_{u_{L2}}$	$E_{u_{sup}}$	$E_{u_{sem}}$	$E_{u_{H_1}}$	$EOC_{u_{L2}}$	$EOC_{u_{sup}}$	$EOC_{u_{sem}}$	$EOC_{u_{H_1}}$
1/1	2.18e-1	2.00e-1	8.66e-1	8.93e-1	-	-	-	-
1/2	6.97e-2	9.57e-2	5.10e-1	5.15e-1	1.64	1.06	0.76	0.79
1/4	1.46e-2	3.44e-2	2.39e-1	2.40e-1	2.26	1.48	1.09	1.10
1/8	4.66e-3	1.65e-2	1.53e-1	1.54e-1	1.65	1.06	0.64	0.64
1/16	1.57e-3	8.47e-3	9.94e-2	9.94e-2	1.57	0.96	0.63	0.63
1/32	5.51e-4	4.33e-3	6.70e-2	6.70e-2	1.51	0.97	0.57	0.57
1/64	1.58e-4	2.09e-3	4.05e-2	4.05e-2	1.80	1.05	0.73	0.73
1/128	4.91e-5	1.07e-3	2.50e-2	2.50e-2	1.68	0.96	0.69	0.69

TABLE 2.4. Conventional approach: Errors (columns left) and EOC (columns right) of control for different error functionals. As reference solution u_h for $h = \frac{1}{256}$ is taken.

The optimization problem reads

$$\begin{aligned}\min J(y,u) &= \frac{1}{2}\int_\Omega (y(x)-y_\Omega)^2 dx + \frac{\alpha}{2}\int_\Gamma u(x)^2 d\sigma(x) + \int_\Gamma e_u(x)u(x)d\sigma(x) \\ &\quad + \int_\Gamma e_y(x)y(x)d\sigma(x)\end{aligned}$$

s.t. $(y,u) \in H^1(\Omega)\times L^\infty(\Gamma)$, $u \in U_{\text{ad}} = \{u \in L^\infty(\Gamma) : 0 \le u(x) \le 1\}$, and (y,u) satisfying the linear state equation

$$\begin{aligned}-\Delta y(x) + c(x)y(x) &= e_1(x) \text{ in } \Omega \\ \partial_\nu y(x) + y(x) &= e_2(x) + u(x) \text{ on } \Gamma,\end{aligned}$$

h	conventional approach E_a	EOC_a	our approach E_a	EOC_a
1/1	5.05e-1	-	5.11e-1	-
1/2	5.05e-1	0.00	3.38e-1	0.60
1/4	5.05e-1	0.00	1.25e-1	1.43
1/8	2.60e-1	0.96	2.92e-2	2.10
1/16	1.16e-1	1.16	7.30e-3	2.00
1/32	4.98e-2	1.22	1.81e-3	2.01
1/64	1.88e-2	1.41	4.08e-4	2.15
1/128	6.98e-3	1.43	8.51e-5	2.26

TABLE 2.5. Errors (columns left) and EOC (columns right) of active sets. As reference set that corresponding to the control u_h for $h = \frac{1}{256}$ is taken. The order of convergence seems to tend to 1.5 in the classical approach, if we are optimistic. The order of convergence of the approach presented here is clearly 2, and its errors are two orders of magnitude smaller than those produced by the conventional approach.

where $\alpha = 1$, $c(x_1, x_2) = 1 + x_1^2 - x_2^2$, $e_y(x_1, x_2) = 1$, $y_\Omega(x_1, x_2) = x_1^2 + x_1 x_2$, $e_1(x_1, x_2) = -2 + (1 + x_1^2 - x_2^2)(1 + 2x_1^2 + x_1 x_2 - x_2^2)$,

$$e_u(x_1, x_2) = \begin{cases} -1 - x_1^3 & \text{on } \Gamma_1 \\ -1 - \min(8(x_2 - 0.5)^2 + 0.5, 1 - 15x_2(x_2 - 0.25) \\ \qquad (x_2 - 0.75)(x_2 - 1)) & \text{on } \Gamma_2 \\ -1 - x_1^2 & \text{on } \Gamma_3 \\ -1 - x_2(1 - x_2) & \text{on } \Gamma_4, \end{cases}$$

and

$$e_2(x_1, x_2) = \begin{cases} 1 - x_1 + 2x_1^2 - x_1^3 & \text{on } \Gamma_1 \\ 7 + 2x_2 - x_2^2 - \min(8(x_2 - 0.5)^2 + 0.5, 1) & \text{on } \Gamma_2 \\ -2 + 2x_1 + x_1^2 & \text{on } \Gamma_3 \\ 1 - x_2 - x_2^2 & \text{on } \Gamma_4. \end{cases}$$

Here $\Gamma_1, \ldots, \Gamma_4$ denote the boundary parts of the unit square numbered counterclockwise beginning at bottom. The adjoint equation for this example is given by

$$\begin{aligned} -\Delta p + c(x)p &= y(x) - y_\Omega(x) \text{ in } \Omega \\ \partial_\nu p + p &= e_y(x) \text{ on } \Gamma, \end{aligned}$$

and the optimal control is given by

$$u = \text{Proj}_{U_{ad}}(-\frac{1}{\alpha}(p + e_u)) \text{ on } \Gamma. \tag{2.30}$$

To solve for u we iterate (2.30), i.e. we apply the fix-point iteration of Algorithm 2.6. The corresponding numerical results can be found in Tables 2.6-2.7 and Figures 2.16-2.5.

h	δy_{L^2}	δy_{L^∞}	δp_{L^2}	δp_{L^∞}	δu_{L^2}	δu_{L^∞}
2^{-0}	0.21922165	0.16660113	0.00981870	0.01171528	0.01293312	0.00975880
2^{-1}	0.05490636	0.05592789	0.00283817	0.00375928	0.00412034	0.00375928
2^{-2}	0.01379774	0.01802888	0.00077525	0.00108642	0.00111801	0.00099280
2^{-3}	0.00345809	0.00554111	0.00019969	0.00028092	0.00028729	0.00025594
2^{-4}	0.00086531	0.00165357	0.00005038	0.00007065	0.00007250	0.00006447
2^{-5}	0.00021639	0.00048246	0.00001263	0.00001769	0.00001819	0.00001615
2^{-6}	0.00005410	0.00013819	0.00000316	0.00000443	0.00000455	0.00000404
2^{-7}	0.00001353	0.00003899	0.00000079	0.00000111	0.00000114	0.00000101
2^{-8}	0.00000338	0.00001086	0.00000020	0.00000028	0.00000028	0.00000025
2^{-4}	0.00056188				0.04330776	0.11460900
2^{-5}	0.00014240				0.02170775	0.05990258
2^{-6}	0.00003500				0.01086060	0.03060061
2^{-7}	0.00000897				0.00543114	0.01546116

TABLE 2.6. Errors in the approach presented here (top part) and in the approach of [**19**] (bottom part). We observe that the error in the controls in the approach presented here on the initial grid already is smaller than the error produced by the approach of [**19**] on a grid with mesh size $h = 2^{-7}$.

h	y_{L^2}	y_{L^∞}	p_{L^2}	p_{L^∞}	u_{L^2}	u_{L^∞}
2^{-1}	1.997345	1.574758	1.790572	1.639862	1.650235	1.376247
2^{-2}	1.992541	1.633258	1.872222	1.790877	1.881837	1.920876
2^{-3}	1.996386	1.702064	1.956905	1.951362	1.960359	1.955685
2^{-4}	1.998688	1.744588	1.986941	1.991434	1.986431	1.989070
2^{-5}	1.999575	1.777112	1.996193	1.997494	1.995161	1.997047
2^{-6}	1.999873	1.803728	1.998912	1.999222	1.998106	1.999024
2^{-7}	1.999964	1.825616	1.999700	1.999725	1.999174	1.999834
2^{-8}	1.999991	1.843640	1.999932	1.999950	1.999609	1.999918

TABLE 2.7. EOC for the approach presented here in the case of Robin-type boundary control. For a comparison to the approach of [**19**] see also Fig. 2.5.

EXAMPLE 2.17. (Dirichlet boundary control)
Here we consider problem (2.17) with $U = L^2(\Gamma)$, $\alpha = 1$ and $U_{\text{ad}} = \{u \in U; 0 \leq u \leq 0.9\}$, i.e. $B \equiv Id$. Again we choose $\Omega = (0,1)^2$. The desired state is given by $z = -\text{sign}(x - 0.5 - \frac{0.1}{\pi})$. State and adjoint state are discretized with piecewise linear, continuous Ansatz functions as described in

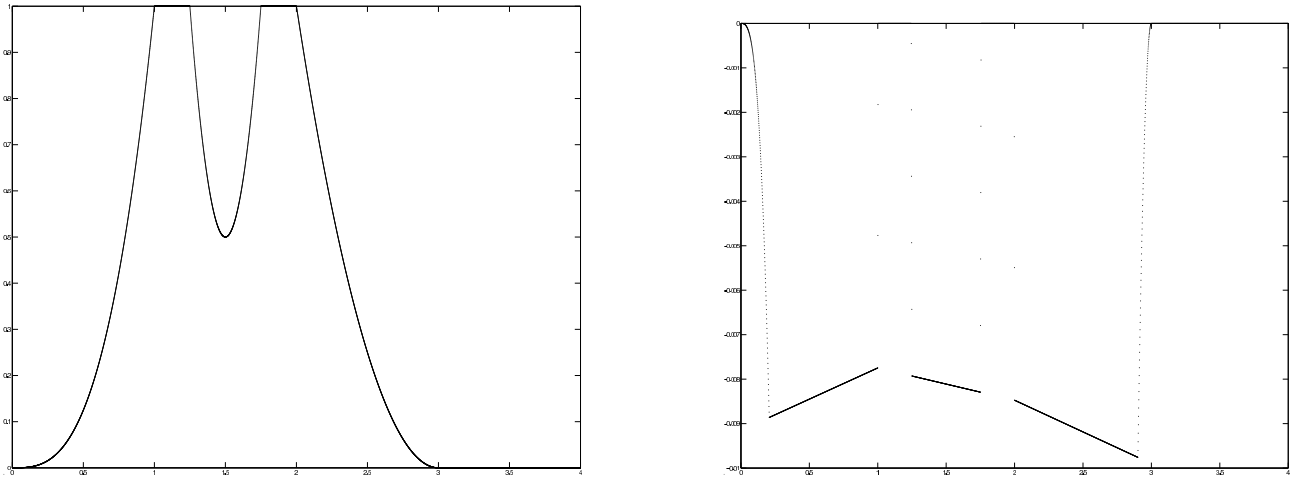


FIGURE 2.4. Exact control (left) and error of exact and numerically computed control on the initial grid containing 4 triangles, i.e. $h = \frac{1}{2}$ (right).

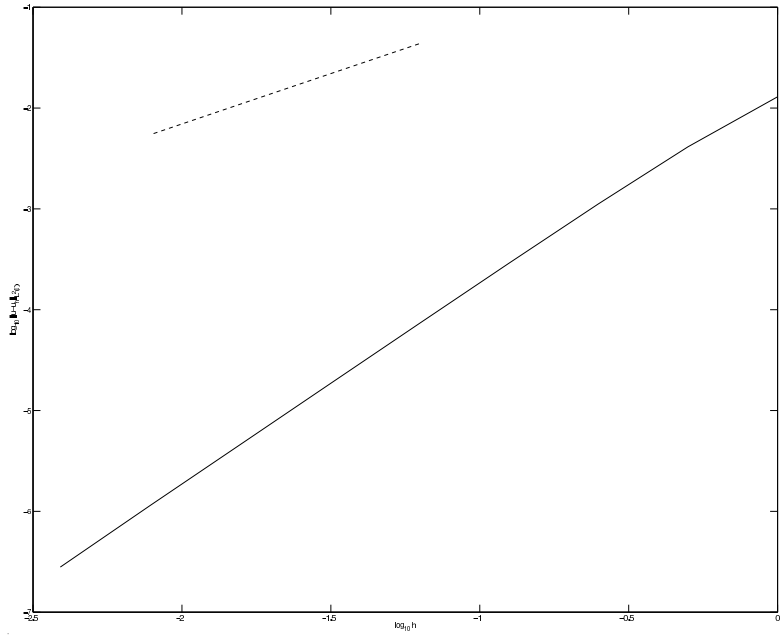


FIGURE 2.5. Numerical comparison of EOC of controls for $E(h) := \|u - u_h\|_{L^2(\Gamma)}$: Approach of [**19**] (dashed) and the approach presented here (solid). The latter yields quadratic convergence, whereas the approach of [**19**] only shows linear convergence.

Subsection 2.6.2. The variational inequality (2.27) motivates the solution algorithm

$$u_h^+ = P_{U_{ad}} \left(\frac{1}{\alpha} \partial_\eta p_h(u_h) \right).$$

We investigate two different approaches; approach 1 in this algorithm uses $\partial_\eta p_h(u_h)$, which represents a piecewise constant (on the boundary grid) L^2 function. Let us emphasize that we not yet have available theory for this approach (which in fact seems to be the natural one if we would replace the continuous quantities in (2.24) by their discrete counterparts). The second approach in this algorithm uses the piecewise linear, continuous discrete flux $\partial_\eta p_h(u_h)$ defined by (2.26). For $h = 2^{-6}$ the value of the cost functional in the optimal solution for the second approach is $J = 0.47473792124624$. The numerical results are summarized in Table 2.8 and are better than those which one would expect from the theoretical investigations in [**8**] (for the state equation) and [**20**] (for the control problem). However, in the case of Dirichlet boundary control many questions are still open and a lot of research has to be done.

h	y_{L^2}	y_{L^∞}	p_{L^2}	p_{L^∞}	u_{L^2}	u_{L^∞}
1-2	-44.315839	-45.874172	2.252319	1.449921	-Inf	-Inf
2-3	-2.658752	-2.692762	0.890090	0.631871	-2.710238	-2.947286
3-4	0.513148	0.230017	1.605929	1.322948	0.559113	0.709528
4-5	0.864432	0.633565	1.641025	1.616581	0.867286	0.687088
5-6	0.955413	0.898523	1.474113	1.599350	0.937568	0.794933
6-7	0.969762	0.711332	1.239616	1.497993	0.936822	0.878459
7-8	0.992879	0.987835	1.106146	1.342300	0.986749	0.960009
8-9	0.990927	0.858741	1.035620	1.177092	0.982189	0.976724
1-2	-0.015094	-0.950093	2.273887	1.599015	-0.464738	-0.950093
2-3	1.479164	1.040787	0.909048	0.498459	1.194508	1.040787
3-4	1.484622	0.855688	1.720355	1.540523	0.979140	0.855688
4-5	1.647971	0.701102	1.873278	1.835947	1.360098	0.701102
5-6	1.545075	0.764482	1.910160	1.895133	1.253975	0.764482
6-7	1.424251	0.798198	1.955067	1.875618	1.227700	0.798198
7-8	1.163258	0.825129	1.915486	1.819988	1.173902	0.825129
8-9	1.020300	0.845442	1.742227	1.722124	1.099603	0.845442

TABLE 2.8. EOC for Dirichlet boundary control: Approach 1 (top part), for which theory is not yet available, Approach 2 (bottom part), for which the theory of Subsection 2.6.2 applies. In both cases we observe linear convergence of the states and controls. The adjoint state also converges linear for approach 1, but seems to converge quadratically in approach 2.

NOTE 2.18. *We note that in all numerical examples presented in the previous subsections, (variants) of the fix-point iteration of Algorithm 2.6 are used. Let us recall that convergence of this algorithm can only be guaranteed for parameter values $\alpha > 0$ large enough. For small parameters $\alpha > 0$ semi-smooth Newton methods* [**77**] *or primal-dual active set strategies* [**39**] *should be applied to the numerical solution of the discrete problems, compare the discussion associated to* (2.11)*. Finally we note that our solution algorithms performs independent of the finite element mesh, i.e. is mesh-independent. This may easily explained by the fact that the iteration of Algorithm 2.6 is defined on*

the infinite dimensional space U of controls, which we have not discretized. Thus, the finite element discretization from the viewpoint of the control problem has more of the flavor of a parametrization than of a discretization.

3. Time dependent problems with control constraints

For the time–dependent case we present the analysis of Discontinuous Galerkin approximations w.r.t. time for an abstract linear–quadratic model problem. The underlying analysis turns out to be very similar to that of the previous section for the stationary model problem.

3.1. Mathematical model, state equation. Let V, H denote separable Hilbert spaces, so that $(V, H = H^*, V^*)$ forms a Gelfand triple. We denote by $a : V \times V \to \mathbb{R}$ a bounded, coercive (and symmetric) bilinear form, and again by U the Hilbert space of controls, and by $B : U \to \mathcal{L}^2(U, L^2(V^*))$ the linear control operator. Here, and from here onwards we write $L^p(S) \equiv L^p((0,T);S)$ where S denotes a Banach space and $T > 0$. For $y_0 \in H$ we consider the state equation

$$\left.\begin{array}{rcll} \int_0^T \langle y_t, v\rangle_{V^*,V} + a(y,v)dt &=& \int_0^T \langle (Bu)(t), v\rangle_{V^*,V} dt & \forall\, v \in L^2(V), \\ (y(0), v)_H &=& (y_0, v)_H & \forall\, v \in V, \end{array}\right\} :\Longleftrightarrow y = \mathcal{T}Bu,$$

which for every $u \in U$ admits a unique solution $y = y(u) \in W := \{w \in L^2(V), w_t \in L^2(V^*)\}$, see e.g. [**80**].

3.2. Optimization problem. We consider the optimization problem

$$(TP) \quad \left\{ \begin{array}{l} \min_{(y,u)\in W\times U_{\text{ad}}} J(y,u) := \frac{1}{2}\|y - z\|^2_{L^2(H)} + \frac{\alpha}{2}\|u\|^2_U \\ \text{s.t. } y = \mathcal{T}Bu, \end{array}\right. \tag{3.1}$$

where $U_{\text{ad}} \subseteq U$ denotes a closed, convex subset. Introducing the reduced cost functional

$$\hat{J}(u) := J(y(u), u),$$

the necessary (and in the present case also sufficient) optimality conditions take the form

$$\Big(\hat{J}'(u), v - u\Big) \geq 0 \text{ for all } v \in U_{\text{ad}}.$$

Here

$$\hat{J}'(u) = \alpha u + B^* p(y(u)),$$

where the adjoint state p solves the adjoint equation

$$\left.\begin{array}{rcll} \int_0^T \langle -p_t, w\rangle_{V^*,V} + a(w,p)dt &=& \int_0^T (y - z, w)_H & \forall\, w \in W, \\ (p(T), v)_H &=& 0, & v \in V. \end{array}\right.$$

This variational inequality is equivalent to the semi–smooth operator equation

$$u = P_{U_{\text{ad}}}\left(-\frac{1}{\alpha}B^* p(y(u))\right)$$

with $P_{U_{\text{ad}}}$ denoting the orthogonal projection in U onto U_{ad}.

3.3. Discretization. Let $V_h \subset V$ denote a finite dimensional subspace, and let $0 = t_0 < t_1 < \cdots < t_m = T$ denote a time grid with grid width δt. We set $I_n := (t_{n-1}, t_n]$ for $n = 1, \ldots, m$ and seek discrete states in the space

$$V_{h,\delta t} := \{\phi : [0,T] \times \Omega \to \mathbb{R}, \phi(t,\cdot)_{|\bar{\Omega}} \in V_h, \phi(\cdot,x)_{|I_n} \in \mathbb{P}_r \text{ for } n = 1, \ldots, m\}.$$

i.e. $y_{h,\delta t}$ is a polynomial of degree $r \in \mathbb{N}$ w.r.t. time. Possible choices of V_h in applications include polynomial finite element spaces, and also wavelet spaces, say. We define the discontinuous Galerkin w.r.t. time approximation (dG(r)-approximation) $\tilde{y} = y_{h,\delta t}(u) \equiv \mathcal{T}_{h,\delta t} Bu \in V_{h,\delta t}$ of the state y as unique solution of

$$A(\tilde{y}, v) := \sum_{n=1}^{m} \int_{I_n} (\tilde{y}_t, v)_H + a(\tilde{y}, v) dt + \sum_{n=1}^{m} ([\tilde{y}]^{n-1}, v^{n-1+})_H + (\tilde{y}^{0+}, v^{0+})_H =$$

$$= (y_0, v^{0+})_H + \int_0^T \langle (Bu)(t), v \rangle_{V^*,V} dt \text{ for all } v \in V_{h,\delta t}.$$

Here,

$$v^{n+} := \lim_{t \searrow t_n} v(t,\cdot),\ v^{n-} := \lim_{t \nearrow t_n} v(t,\cdot), \text{ and } [v]^n := v^{n+} - v^{n-}.$$

The discrete counterpart of the optimal control problem reads

$$(P_{h,\delta t}) \quad \min_{u \in U_{\text{ad}}} \hat{J}_{h,\delta t}(u) := J(y_{h,\delta t}(u), u)$$

and it admits a unique solution $u_{h,\delta t} \in U_{\text{ad}}$. We further have

$$\hat{J}'_{h,\delta t}(v) = v + B^* p_{h,\delta t}(y_{h,\delta t}(v)),$$

where $p_{h,\delta t}(y_{h,\delta t}(v)) \in V_{h,\delta t}$ denotes the unique solution of

$$A(v, p_{h,\delta t}) = \int_0^T (y_{h,\delta t} - z, v)_H dt \text{ for all } v \in V_{h,\delta t}.$$

Further, the unique discrete solution $u_{h,\delta t}$ satisfies

$$(u_{h,\delta t} + B^* p_{h,\delta t}, v - u_{h,\delta t})_H \geq 0 \text{ for all } v \in U_{\text{ad}}.$$

As in the continuous case this variational inequality is equivalent to a semi–smooth operator equation, namely

$$u_{h,\delta t} = P_{U_{\text{ad}}} \left(-\frac{1}{\alpha} B^* p_{h,\delta t}(y_{h,\delta t}(u_{h,\delta t})) \right).$$

For this discrete approach the proof of the following theorem follows the lines of the proof of Theorem 2.8.

THEOREM 3.1. *Let $u, u_{h,\delta t}$ denote the unique solutions of (P) and $(P_{h,\delta t})$, respectively. Then*

$$\|u-u_{h,\delta t}\|_U^2 \le \frac{1}{\alpha}\Big\{(B^*(p(u)-\tilde p_{h,\delta t}(u)),u_{h,\delta t}-u)_U + \int_0^T (y_{h,\delta t}(u_{h,\delta t})-y_{h,\delta t}(u))(y(u)-y_{h,\delta t}(u))_H\,dt\Big\}, \tag{3.2}$$

where $\tilde p_{h,\delta t}(u) := \mathcal{T}_{h,\delta t}^*(\mathcal{T}Bu-z)$, $y_{h,\delta t}(u) := \mathcal{T}_{h,\delta t}Bu$, and $y(u) := \mathcal{T}Bu$.

4. State constraints (joint with Klaus Deckelnick, Magdeburg)

Let us now sketch the actually most promising discrete approach in the presence of state constraints. As model problem we take

$$(\mathbb{S})\quad \begin{cases} \min_{(y,u)\in Y\times U} J(y,u) := \frac12\int_\Omega |y-y_0|^2 + \frac{\alpha}{2}\|u-u_0\|_U^2 \\ \text{s.t.} \\ \left.\begin{aligned} -\Delta y + y &= u && \text{in } \Omega,\\ \partial_\eta y &= 0 && \text{on } \Gamma,\end{aligned}\right\} :\Longleftrightarrow y=\mathcal{G}(u) \\ \text{and} \\ y\in Y_{\text{ad}} := \{y\in L^\infty(\Omega), y(x)\le b(x) \text{ a.e. in } \Omega\}. \end{cases} \tag{4.1}$$

Here, $\Omega\subset\mathbb{R}^n$ denotes an open, bounded sufficiently smooth (polyhedral) domain, $U\equiv L^2(\Omega)$, $\alpha>0$, and $y_0,u_0\in H^1(\Omega)$ as well as $b\in W^{2,\infty}(\Omega)$ are given functions. We denote by $\mathcal{M}(\bar\Omega)$ the space of Radon measures which is defined as the dual space of $C^0(\bar\Omega)$ and endowed with the norm

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

The difficulty of problem $(\mathbb{S})$ stems from the pointwise *state constraint* $y(x)\le b(x)$ a.e. in Ω. However, the analysis of (4.1) is well understood and sketched in [**74**, Section 6.2.1] for the problem under consideration. Since the state constraints form a convex set and the cost functional is quadratic it is not difficult to establish the existence of a unique solution $u\in L^2(\Omega)$ to this problem. Moreover, from [**16**, Theorem 5.2] we infer (compare also [**15**, Theorem 2])

THEOREM 4.1. *A function $u\in L^2(\Omega)$ is a solution of* (4.1) *if and only if there exist $\mu\in\mathcal{M}(\bar\Omega)$ and $p\in L^2(\Omega)$ such that with $y=\mathcal{G}(u)$ there holds*

$$\int_\Omega p\big(-\Delta v+v\big) = \int_\Omega (y-z)v + \int_{\bar\Omega} v\,d\mu \qquad \forall v\in H^2(\Omega) \text{ with } \partial_\nu v=0 \text{ on } \partial\Omega \tag{4.2}$$

$$p+\alpha(u-u_0)=0 \qquad \text{a.e. in } \Omega \tag{4.3}$$

$$\mu\ge 0,\ y(x)\le b(x) \text{ a.e. in } \Omega \quad \text{and} \quad \int_{\bar\Omega}(b-y)\,d\mu = 0. \tag{4.4}$$

The study of (4.1) is complicated by the presence of the measure μ on the right hand side of (4.2). As a consequence, the solution p of this problem is no longer in $H^1(\Omega)$ but only in $W^{1,s}(\Omega)$ for all $1 \leq s < \frac{d}{d-1}$. This fact also accounts for the form of the weak formulation (4.2). In its classical form the first order optimality system of problem (4.1) reads

$$
\left\{
\begin{array}{rcll}
-\Delta y + y &=& u & \text{in } \Omega, \\
\partial_\eta y &=& 0 & \text{on } \Gamma, \\
-\Delta p + p &=& y - z - \mu & \text{in } \Omega, \\
\partial_\eta y &=& 0 & \text{on } \Gamma, \\
p + \alpha(u - u_0) &=& 0 & \text{a.e. in } \Omega, \\
\int\limits_{\bar{\Omega}} y - b d\mu &=& 0 & \\
\multicolumn{4}{l}{\text{and}} \\
\multicolumn{4}{l}{\mu \geq 0,\ y(x) \leq b(x) \text{ a.e. in } \Omega.}
\end{array}
\right. \tag{4.5}
$$

Let us first provide Example 6.2 from [**58**] (in slightly modified form) which illustrates that Lagrange multipliers in fact occur as measures.

EXAMPLE 4.2. (see [**58**, Example 6.2])
We set $\Omega := B_1(0) \subset \mathbb{R}^2$, $r = r(x) := |x|$ and denote by $\Phi(r) := -\frac{1}{2\pi} \log r$ the fundamental solution of Poisson's equation on the unit disc in $\mathbb{R}^2$. Then

$$-\Delta \Phi = \delta_0,$$

where δ_0 denotes the Dirac measure concentrated in $0 \in \mathbb{R}^2$. The solution of (4.5) with $u_0 = 4 + \frac{1}{4\alpha\pi} r^2 - \frac{1}{2\alpha\pi} \log r$, and $b(x) := r^4 + 4$ and $z(x) := 4 + \frac{1}{\pi} - \frac{1}{4\pi} r^2 + \frac{1}{2\pi} \log r$ then is given by

$$y(x) \equiv 4,\ p(x) = \frac{1}{4\pi} r^2 - \frac{1}{2\pi} \log r,\ u(x) \equiv 4, \text{ and } \mu = \delta_0,$$

and $\mu \in C(\Omega)^*$, but $\mu \notin H^1(\Omega)^*$.

The development of numerical approaches to tackle (4.1) and/or (4.5) is ongoing. In the following we sketch the approach presented in [**23**] and present finite element approximations of problem (4.1). The underlying idea consists in approximating the cost functional J by a sequence of functionals J_h where h is a mesh parameter related to a sequence of triangulations. The definition of J_h involves the approximation of the state equation by linear finite elements and enforces constraints on the state in the nodes of the triangulation. We shall prove that the minima of J_h converge in L^2 to the minimum of J as $h \to 0$ and that the states convergence strongly in H^1 as well as uniformly and derive corresponding error bounds.
To the authors knowledge only few earlier attempts have been made to develop a finite element analysis for state constrained elliptic control problems. In [**17**] Casas proves convergence of finite element approximations to optimal control problems for semi-linear elliptic equations with finitely many state constraints. Casas and Mateos extend these results in [**18**] to a less regular setting for the states and prove convergence of finite element approximations to semi-linear distributed and boundary control problems.

Let us comment on further approaches that tackle optimization problems for pdes with state constraints. A *Lavrentiev-type regularization* of problem (4.1) is investigated in [**57**]. In this approach the state constraint $y \le b$ in (4.1) is replaced by the mixed constraint $\epsilon u + y \le b$, with $\epsilon > 0$ denoting a regularization parameter. It turns out that the associated Lagrange multiplier μ_ϵ belongs to $L^2(\Omega)$. The resulting optimization problems are solved either by interior-point methods or primal-dual active set strategies, compare [**58**]. The development of numerical approaches to tackle (4.1) is ongoing. An excellent overview can be found in [**37, 38**], where among other things penalty methods are discussed, and also further references are given.

4.1. Finite element discretization. Let $\mathcal{T}_h$ be a triangulation of Ω with maximum mesh size $h := \max_{T\in\mathcal{T}_h} \mathrm{diam}(T)$ and vertices $x_1, \dots, x_m$. We suppose that $\bar\Omega$ is the union of the elements of $\mathcal{T}_h$ so that element edges lying on the boundary are curved. In addition, we assume that the triangulation is quasi-uniform in the sense that there exists a constant $\kappa > 0$ (independent of h) such that each $T \in \mathcal{T}_h$ is contained in a ball of radius $\kappa^{-1}h$ and contains a ball of radius κh. Let us define the space of linear finite elements,

$$X_h := \{v_h \in C^0(\bar\Omega) \mid v_h \text{ is a linear polynomial on each } T \in \mathcal{T}_h\}.$$

In what follows it is convenient to introduce a discrete approximation of the solution operator $\mathcal{G}$. For a given function $v \in L^2(\Omega)$ we denote by $z_h = \mathcal{G}_h(v) \in X_h$ the solution of the discrete Neumann problem

$$\int_\Omega \big(\nabla z_h \cdot \nabla v_h + z_h v_h\big) \;=\; \int_\Omega v v_h \quad \text{for all } v_h \in X_h.$$

It is well–known that for all $v \in L^2(\Omega)$

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\| \;\le\; Ch^2\|v\|, \tag{4.6}$$

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\|_{L^\infty} \;\le\; Ch^{2-\frac{d}{2}}\|v\|. \tag{4.7}$$

Here, $\|\cdot\|$ denotes the L^2–norm. We propose the following approximation of the control problem (4.1):

$$\begin{aligned} &\min_{u\in L^2(\Omega)} J_h(u) := \frac{1}{2}\int_\Omega |y_h - P_h y_0|^2 + \frac{\alpha}{2}\int_\Omega |u - P_h u_0|^2 \\ &\text{subject to } \; y_h = \mathcal{G}_h(u) \text{ and } y_h(x_j) \le b(x_j) \text{ for } j = 1, \dots, m. \end{aligned} \tag{4.8}$$

Here, P_h denotes the L^2–projection, i.e.

$$\int_\Omega P_h z\, v_h = \int_\Omega z\, v_h \qquad \forall v_h \in X_h. \tag{4.9}$$

It is well–known that

$$\|z - P_h z\| \le Ch\|z\|_{H^1} \qquad \forall z \in H^1(\Omega). \tag{4.10}$$

Problem (4.8) represents a convex infinite-dimensional optimization problem of similar structure as problem (4.1), but with only finitely many equality and inequality constraints which form a convex

admissible set. Again we can apply [**16**, Theorem 5.2] which together with [**15**, Corollary 1] yields (compare also the analysis of problem (P) in [**17**])

LEMMA 4.3. *Problem (4.8) has a unique solution $u_h \in L^2(\Omega)$. There exist $\mu_1, \dots, \mu_m \in \mathbb{R}$ and $p_h \in X_h$ such that with $y_h = \mathcal{G}_h(u_h)$ and $\mu_h = \sum_{j=1}^m \mu_j \delta_{x_j}$ we have*

$$\int_\Omega \big(\nabla p_h \cdot \nabla v_h + p_h v_h\big) = \int_\Omega (y_h - P_h y_0) v_h + \int_{\bar\Omega} v_h d\mu_h \qquad \text{for all } v_h \in X_h, \tag{4.11}$$

$$p_h + \alpha(u_h - P_h u_0) = 0 \text{ in } \Omega, \tag{4.12}$$

$$\mu_j \geq 0,\ y_h(x_j) \leq b(x_j), j = 1, \dots, m \text{ and } \int_{\bar\Omega} \big(I_h b - y_h\big) d\mu_h = 0. \tag{4.13}$$

Here, δ_x denotes the Dirac measure concentrated at x and I_h is the usual Lagrange interpolation operator.

REMARK 4.4. *From* (4.12) *we deduce that in problem* (4.8) *it is sufficient to minimize over controls $u \in X_h$ instead of $u \in L^2(\Omega)$ in order to obtain the same unique solution u_h. For the resulting finite dimensional optimization problem the result of Lemma 4.3 then follows from e.g.* [**62**, Theorem 12.1].

Now let us introduce the finite element matrices

$$A = (a_{ij}), a_{ij} := \int_\Omega \nabla\phi_i \nabla\phi_j + \phi_i\phi_j dx, \text{ and } M = (m_{ij}), m_{ij} := \int_\Omega \phi_i\phi_j dx,\ (i, j = 1, \dots, nv).$$

Then we may rewrite the system of Lemma 4.3 in the form

$$\left\{ \begin{array}{rcl} Ay &=& Mu, \\ Ap &=& M(y - z) - \mu, \\ p + \alpha(u - u_0) &=& 0, \\ \mu &=& \max(0, \mu + (y - b)), \end{array} \right. \tag{4.14}$$

where now $y, p, u, \mu, b \in \mathbb{R}^{nv}$ denote the nodal vectors associated to the corresponding finite element Ansatz functions.

We have the following convergence result.

THEOREM 4.5. *Let $u_h \in L^2(\Omega)$ be the optimal solution of (4.8) with corresponding state $y_h \in X_h$ and adjoint variables $p_h \in X_h$ and $\mu_h \in \mathcal{M}(\bar\Omega)$. Then, as $h \to 0$ we have*

$$u_h \to u \text{ in } L^2(\Omega), \quad y_h \to y \text{ in } H^1(\Omega) \text{ and in } C^0(\bar\Omega),$$

where u is the solution of (4.1) with corresponding state y.

Proof. Let $\underline{b} := \min_{x \in \bar\Omega} b(x)$. Since $\underline{b} = \mathcal{G}_h(\underline{b})$ and $\underline{b} \leq b(x_j)$ for $j = 1, \dots, m$ we have

$$\frac{1}{2} \int_\Omega |y_h - P_h y_0|^2 + \frac{\alpha}{2} \int_\Omega |u_h - P_h u_0|^2 = J_h(u_h) \leq J_h(\underline{b}) \leq C(y_0, u_0, \underline{b}).$$

This implies that there exists a constant C which is independent of h such that

$$\|y_h\|,\ \|u_h\|,\ \|p_h\| \leq C \qquad \text{for all } 0 < h \leq 1. \tag{4.15}$$

Note that the bound on p_h follows from (4.12). In order to estimate μ_h we use $v_h \equiv 1$ in (4.11) and obtain for every $f \in C^0(\bar\Omega), |f| \le 1$

$$\int_{\bar\Omega} f d\mu_h \le \sum_{j=1}^m \mu_j |f(x_j)| \le \sum_{j=1}^m \mu_j = \int_{\bar\Omega} 1 d\mu_h = \int_\Omega \big(p_h + P_h y_0 - y_h\big) \le C$$

by (4.15). This yields

$$\|\mu_h\|_{\mathcal{M}(\bar\Omega)} \le C \qquad \text{for all } 0 < h \le 1. \tag{4.16}$$

In view of (4.15), (4.16) there exists a sequence $h \to 0$ and $\hat u, \hat p \in L^2(\Omega)$ as well as $\hat\mu \in \mathcal{M}(\bar\Omega)$ such that

$$u_h \rightharpoonup \hat u, \quad p_h \rightharpoonup \hat p \text{ in } L^2(\Omega), \quad \text{ and } \mu_h \rightharpoonup \hat\mu \text{ in } \mathcal{M}(\bar\Omega). \tag{4.17}$$

Since $\mathcal{G}$ is compact as an operator from $L^2(\Omega)$ into $C^0(\bar\Omega)$ we have, after passing to a further subsequence if necessary,

$$\mathcal{G}(u_h) \to \mathcal{G}(\hat u) \qquad \text{in } C^0(\bar\Omega) \tag{4.18}$$

and hence

$$\|y_h - \mathcal{G}(\hat u)\|_{L^\infty} \le \|\mathcal{G}_h(u_h) - \mathcal{G}(u_h)\|_{L^\infty} + \|\mathcal{G}(u_h) - \mathcal{G}(\hat u)\|_{L^\infty} \le Ch^{2-\frac{d}{2}}\|u_h\| + \|\mathcal{G}(u_h) - \mathcal{G}(\hat u)\|_{L^\infty}$$

so that $y_h \to \mathcal{G}(\hat u) =: \hat y$ in $C^0(\bar\Omega)$ as $h \to 0$ by (4.15) and (4.18). A similar argument shows that $y_h \to \hat y$ in $H^1(\Omega)$.

Let us now pass to the limit in (4.11)–(4.13). To begin, let $v \in H^2(\Omega)$ with $\partial_\nu v = 0$ on $\partial\Omega$ and denote by $R_h v$ the Ritz projection of v. Recalling (4.17), (4.11) and the fact that $R_h v \to v$ in $C^0(\bar\Omega)$ we obtain

$$\begin{aligned}
\int_\Omega \hat p\big(-\Delta v + v\big) \;&\leftarrow\; \int_\Omega p_h\big(-\Delta v + v\big) = \int_\Omega \big(\nabla p_h \cdot \nabla v + p_h v\big) \\
&= \int_\Omega \big(\nabla p_h \cdot \nabla R_h v + p_h R_h v\big) = \int_\Omega (y_h - P_h y_0) R_h v + \int_{\bar\Omega} R_h v d\mu_h \\
&\to \int_\Omega (\hat y - y_0) v + \int_{\bar\Omega} v d\hat\mu.
\end{aligned}$$

Using (4.17) we may pass to the limit in (4.12) and deduce $\hat p + \alpha(\hat u - u_0) = 0$ a.e. in Ω. Clearly, $\hat\mu \ge 0$; since $y_h \le I_h b$ in $\bar\Omega$ and $y_h \to \hat y$ in $C^0(\bar\Omega)$ we have $\hat y \le b$ in $\bar\Omega$. Furthermore, recalling that $\int_{\bar\Omega}(I_h b - y_h) d\mu_h = 0$ we obtain in the limit

$$\int_{\bar\Omega} (b - \hat y) d\hat\mu = 0.$$

Lemma 4.1 now implies that $\hat u$ is a solution of (4.1); as the solution of this problem is unique we must have $u = \hat u$ and hence $y = \hat y$ and the whole sequence is convergent.
Let us finally prove that $u_h \to u$ in $L^2(\Omega)$. To begin, note that by (4.7)

$$\mathcal{G}_h\big(u - \gamma h^{2-\frac{d}{2}}\big) = \mathcal{G}_h(u) - \mathcal{G}(u) + \mathcal{G}(u) - \gamma h^{2-\frac{d}{2}} \le Ch^{2-\frac{d}{2}}\|u\| + b - \gamma h^{2-\frac{d}{2}} \le b$$

in $\bar\Omega$, provided that γ is large enough. Evaluating the above inequality at the nodes $x_1,\dots,x_m$ we see that $\mathcal{G}_h(u-\gamma h^{2-\frac{d}{2}})$ is admissible for the discrete problem and hence $J_h(u_h)\le J_h(u-\gamma h^{2-\frac{d}{2}})$ or

$$\frac{\alpha}{2}\|u_h-P_hu_0\|^2\le\frac{\alpha}{2}\|u-\gamma h^{2-\frac{d}{2}}-P_hu_0\|^2+\frac{1}{2}\|\mathcal{G}_h(u)-\gamma h^{2-\frac{d}{2}}-P_hy_0\|^2-\frac{1}{2}\|y_h-P_hy_0\|^2.$$

Since $y_h\to y,\ \mathcal{G}_h(u)\to\mathcal{G}(u)=y$ in $L^2(\Omega)$ we infer that

$$\limsup_{h\to0}\|u_h-P_hu_0\|^2\le\|u-u_0\|^2\le\liminf_{h\to0}\|u_h-P_hu_0\|^2,$$

where the second inequality is a consequence of the weak convergence $u_h-P_hu_0\rightharpoonup u-u_0$. Thus, $\|u_h-P_hu_0\|^2\to\|u-u_0\|^2$ which implies $u_h-P_hu_0\to u-u_0$ in L^2 and hence $u_h\to u_0$ in L^2. □

4.2. Error analysis. Let us now turn to the error analysis and start with a couple of auxiliary results.

LEMMA 4.6. *Suppose that $u,u_h\in L^2(\Omega)$ are the optimal solutions of (4.1) and (4.8) respectively with corresponding states $y\in H^2(\Omega), y_h\in X_h$. Let $v\in L^2(\Omega)$ and $z=\mathcal{G}(v), z_h=\mathcal{G}_h(v)$. Then*

$$J(u)+\frac{1}{2}\int_\Omega|z-y|^2+\frac{\alpha}{2}\int_\Omega|v-u|^2+\int_{\bar\Omega}(b-z)d\mu = J(v) \tag{4.19}$$

$$J_h(u_h)+\frac{1}{2}\int_\Omega|z_h-y_h|^2+\frac{\alpha}{2}\int_\Omega|v-u_h|^2+\int_{\bar\Omega}\big(I_hb-z_h\big)d\mu_h = J_h(v) \tag{4.20}$$

Proof. An elementary calculation using (4.2) shows

$$\begin{aligned}J(v)-J(u) &= \frac{1}{2}\int_\Omega|z-y|^2+\frac{\alpha}{2}\int_\Omega|v-u|^2+\int_\Omega(z-y)(y-y_0)+\alpha\int_\Omega(u-u_0)(v-u)\\ &= \frac{1}{2}\int_\Omega|z-y|^2+\frac{\alpha}{2}\int_\Omega|v-u|^2+\int_\Omega p\big(-\Delta(z-y)+(z-y)\big)\\ &\quad-\int_{\bar\Omega}(z-y)d\mu+\alpha\int_\Omega(u-u_0)(v-u).\end{aligned}$$

Since $z=\mathcal{G}(v), y=\mathcal{G}(u)$ we have

$$\int_\Omega p\big(-\Delta(z-y)+(z-y)\big)=\int_\Omega p(v-u),$$

so that (4.3) and (4.4) finally imply

$$J(v)-J(u)=\frac{1}{2}\int_\Omega|z-y|^2+\frac{\alpha}{2}\int_\Omega|v-u|^2+\int_{\bar\Omega}(b-z)d\mu.$$

The second claim follows in a similar way. □

REMARK 4.7. *Note that in the above $z=\mathcal{G}(v), z_h=\mathcal{G}_h(v)$ do not necessarily have to be admissible for the minimization problems.*

The next lemma examines in more detail the approximation of J by J_h.

LEMMA 4.8. *Suppose that* $v \in W^{1,s}(\Omega)$ *for some* $\frac{2d}{d+2} \le s \le 2$. *Then*

$$|J(v) - J_h(v)| \le Ch^{2+\frac{d}{2}-\frac{d}{s}}\big(\|u_0\|_{H^1}\|v\|_{W^{1,s}} + \|v\|^2 + \|y_0\|_{H^1}^2 + \|u_0\|_{H^1}^2\big).$$

Proof. Let $z = \mathcal{G}(v), z_h = \mathcal{G}_h(v)$. Then

$$J(v) - J_h(v) = \frac{1}{2}\int_\Omega \big(|z-y_0|^2 - |z_h - P_h y_0|^2\big) + \frac{\alpha}{2}\int_\Omega \big(|v-u_0|^2 - |v-P_h u_0|^2\big).$$

Using (4.9), (4.10), (4.6) and

$$\|y\|_{H^2} \le C\|u\|_{L^2}.$$

we obtain

$$\begin{aligned}
|\int_\Omega \big(|z-y_0|^2 - |z_h - P_h y_0|^2\big)| &= |\int_\Omega (z - y_0 - z_h + P_h y_0)(z - y_0 + z_h - P_h y_0)| \\
&= |\int_\Omega \big((z-z_h)(z-y_0+z_h-P_h y_0) - (y_0 - P_h y_0)(z - y_0 - P_h(z-y_0)\big)| \\
&\le C\|z-z_h\|\big(\|z\| + \|z_h\| + \|y_0\|\big) + Ch^2\|y_0\|_{H^1}\big(\|z\|_{H^1} + \|y_0\|_{H^1}\big) \\
&\le Ch^2\big(\|v\|^2 + \|y_0\|_{H^1}^2\big).
\end{aligned}$$

For the second term we obtain in a similar way

$$\int_\Omega \big(|v-u_0|^2 - |v-P_h u_0|^2\big) = \int_\Omega (u_0 - P_h u_0)w = \int_\Omega (u_0 - P_h u_0)(w - P_h w),$$

where $w = u_0 + P_h u_0 - 2v$ and where we have used (4.9). Applying Lemma 4.16 from the Appendix we infer

$$\begin{aligned}
|\int_\Omega \big(|v-u_0|^2 - |v-P_h u_0|^2\big)| &\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\|u_0\|_{H^1}\|w\|_{W^{1,s}} \\
&\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\|u_0\|_{H^1}\big(\|u_0\|_{H^1} + \|v\|_{W^{1,s}}\big).
\end{aligned}$$

This proves the lemma. □

LEMMA 4.9. *Suppose that* $v \in W^{1,s}(\Omega)$ *for some* $1 < s < \frac{d}{d-1}$. *Then*

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\|_{L^\infty} \le Ch^{3-\frac{d}{s}}|\log h|\ \|v\|_{W^{1,s}}.$$

Proof. Let $z = \mathcal{G}(v),\ z_h = \mathcal{G}_h(v)$. Elliptic regularity theory implies that $z \in W^{3,s}(\Omega)$ from which we infer that $z \in W^{2,q}(\Omega)$ with $q = \frac{ds}{d-s}$ using a well–known embedding theorem. Furthermore, we have

$$\|z\|_{W^{2,q}} \le c\|z\|_{W^{3,s}} \le c\|v\|_{W^{1,s}}. \tag{4.21}$$

Using Theorem 2.2 and the following Remark in [**69**] we have

$$\|z - z_h\|_{L^\infty} \le c|\log h| \inf_{\chi \in X_h} \|z - \chi\|_{L^\infty}, \tag{4.22}$$

which, combined with a well–known interpolation estimate, yields

$$\|z - z_h\|_{L^\infty} \le ch^{2-\frac{d}{q}}|\log h|\|z\|_{W^{2,q}} \le ch^{3-\frac{d}{s}}|\log h|\|v\|_{W^{1,s}}$$

in view (4.21) and the relation between s and q. □

Our next aim is to derive a uniform bound on $\|u_h\|_{W^{1,s}}$ for $s < \frac{d}{d-1}$.

LEMMA 4.10. *Let $1 < s < \frac{d}{d-1}$. Then there exists a constant c, which is independent of h, such that*

$$\|u_h\|_{W^{1,s}} \le c \qquad \textit{for all } 0 < h \le 1.$$

Proof. In view of (4.12) we have

$$\|u_h\|_{W^{1,s}} \le \frac{1}{\alpha}\|p_h\|_{W^{1,s}} + \|P_h u_0\|_{H^1} \le \frac{1}{\alpha}\|p_h\|_{W^{1,s}} + c,$$

so that it is sufficient to bound $\|p_h\|_{W^{1,s}}$.
Let s' be such that $\frac{1}{s} + \frac{1}{s'} = 1$ and suppose that $\phi \in L^{s'}(\Omega)$. Let us denote by $\psi \in W^{2,s'}(\Omega)$ the unique solution of the Neumann problem

$$\begin{aligned} -\Delta\psi + \psi &= \phi \quad \text{in } \Omega \\ \partial_\nu \psi &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Integration by parts and (4.11) yield

$$\begin{aligned} \int_\Omega p_h\,\phi &= \int_\Omega \big(\nabla p_h \cdot \nabla\psi + p_h\psi\big) = \int_\Omega \big(\nabla p_h \cdot \nabla R_h\psi + p_h R_h\psi\big) \\ &= \int_\Omega (y_h - P_h y_0) R_h\psi + \int_{\bar\Omega} R_h\psi\, d\mu_h, \end{aligned} \tag{4.23}$$

where $R_h\psi$ is the Ritz projection of ψ. Arguing similarly as in Theorem 1 of [**14**] one shows that there exists a unique solution $p^h \in W^{1,s}(\Omega)$ of the problem

$$\int_\Omega p^h\big(-\Delta v + v\big) = \int_\Omega (y_h - P_h y_0)v + \int_{\bar\Omega} v\, d\mu_h \quad \forall v \in H^2(\Omega) \text{ with } \partial_\nu v = 0 \text{ on } \partial\Omega. \tag{4.24}$$

Furthermore, there exists a constant $c = c(s) > 0$ such that

$$\|p^h\|_{W^{1,s}} \le c\big(\|y_h - P_h y_0\| + \|\mu_h\|_{\mathcal{M}(\bar\Omega)}\big) \le c \tag{4.25}$$

uniformly in h in view of (4.15) and (4.16). If we use $v = \psi$ in (4.24) and combine it with (4.23) we obtain

$$\begin{aligned} \int_\Omega (p^h - p_h)\phi &= \int_\Omega (y_h - P_h y_0)(\psi - R_h\psi) + \int_{\bar\Omega} (\psi - R_h\psi) d\mu_h \\ &\le ch^2\|\psi\|_{H^2}\big(\|y_h\| + \|P_h y_0\|\big) + \|\psi - R_h\psi\|_{L^\infty}\|\mu_h\|_{\mathcal{M}(\bar\Omega)} \\ &\le ch^2\|\psi\|_{H^2} + ch^{2-\frac{d}{s'}}|\log h|\|\psi\|_{W^{2,s'}} \\ &\le ch^{2-\frac{d}{s'}}|\log h|\|\phi\|_{L^{s'}}. \end{aligned}$$

Note that we have again applied (4.22) in order to control $\|\psi - R_h\psi\|_{L^\infty}$. Since $\phi \in L^{s'}(\Omega)$ is arbitrary we infer

$$\|p^h - p_h\|_{L^s} \le ch^{2-\frac{d}{s'}}|\log h|.$$

Interpolation and inverse estimates then give

$$\|\nabla p_h\|_{L^s} \le c\|\nabla p^h\|_{L^s} + ch^{1-\frac{d}{s'}}|\log h| \le c$$

by (4.25) and since $1-\frac{d}{s'} = \frac{d-1}{s}\big(\frac{d}{d-1}-s\big) > 0$. □

Let us finally turn to an error estimate for the optimal controls and the optimal states.

THEOREM 4.11. *Let u and u_h be the solutions of (4.1) and (4.8) respectively. For every $\epsilon > 0$ there exists $C_\epsilon > 0$ such that*

$$\|u-u_h\| + \|y-y_h\|_{H^1} \le C_\epsilon h^{2-\frac{d}{2}-\epsilon}.$$

Proof. Let us define $\tilde{y}^h := \mathcal{G}(u_h) \in H^2(\Omega)$ and $\tilde{y}_h := \mathcal{G}_h(u) \in X_h$. Then Lemma 4.6 implies

$$\begin{aligned} J(u) + \frac{1}{2}\int_\Omega |\tilde{y}^h - y|^2 + \frac{\alpha}{2}\int_\Omega |u_h-u|^2 + \int_{\bar\Omega}(b-\tilde{y}^h)d\mu &= J(u_h) \\ J_h(u_h) + \frac{1}{2}\int_\Omega |\tilde{y}_h - y_h|^2 + \frac{\alpha}{2}\int_\Omega |u-u_h|^2 + \int_{\bar\Omega}\big(I_h b - \tilde{y}_h\big)d\mu_h &= J_h(u). \end{aligned}$$

Since $u = u_0 - \frac{1}{\alpha}p \in W^{1,s}(\Omega)$ for all $\frac{2d}{d+2} \le s < \frac{d}{d-1}$ we obtain with the help of Lemma 4.8

$$\begin{aligned} &\frac{1}{2}\int_\Omega |\tilde{y}^h - y|^2 + \frac{1}{2}\int_\Omega |\tilde{y}_h - y_h|^2 + \alpha\int_\Omega |u_h-u|^2 \\ &= J(u_h) - J(u) + J_h(u) - J_h(u_h) - \int_{\bar\Omega}(b-\tilde{y}^h)d\mu - \int_{\bar\Omega}\big(I_h b - \tilde{y}_h\big)d\mu_h \qquad (4.26) \\ &\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\Big(\|u_0\|_{H^1}\big(\|u\|_{W^{1,s}} + \|u_h\|_{W^{1,s}}\big) + \|u\|^2 + \|u_h\|^2 + \|y_0\|_{H^1}^2 + \|u_0\|_{H^1}^2\Big) \\ &\quad + \int_{\bar\Omega}(\tilde{y}^h - b)d\mu + \int_{\bar\Omega}(\tilde{y}_h - I_h b)d\mu_h. \end{aligned}$$

Let us first consider the last two integrals. We have for $x \in \bar\Omega$

$$\begin{aligned} \tilde{y}^h(x) - b(x) &= (\tilde{y}^h(x) - y_h(x)) + (y_h(x) - (I_h b)(x)) + ((I_h b)(x) - b(x)) \\ &\le \|\mathcal{G}(u_h) - \mathcal{G}_h(u_h)\|_{L^\infty} + \|I_h b - b\|_{L^\infty}, \end{aligned}$$

since $y_h(x_j) \le b(x_j), j = 1,\dots,m$ implies that $y_h \le I_h b$ in $\bar\Omega$. If we combine Lemma 4.9 with Lemma 4.10 we infer

$$\int_{\bar\Omega}(\tilde{y}^h - b)d\mu \le ch^{3-\frac{d}{s}}|\log h|\,\|u_h\|_{W^{1,s}} + Ch^2|b|_{W^{2,\infty}} \le ch^{3-\frac{d}{s}}|\log h|.$$

Similarly we have from (4.4)

$$\begin{aligned} \tilde{y}_h(x) - (I_h b)(x) &= (\tilde{y}_h(x) - y(x)) + (y(x) - b(x)) + (b(x) - (I_h b)(x)) \\ &\le \|\mathcal{G}_h(u) - \mathcal{G}(u)\|_{L^\infty} + \|b - I_h b\|_{L^\infty}, \end{aligned}$$

so that (4.16) and Lemma 4.9 give

$$\int_{\bar\Omega}\big(y_h - I_h b\big)d\mu_h \le ch^{3-\frac{d}{s}}|\log h|\,\|u\|_{W^{1,s}} + Ch^2|b|_{W^{2,\infty}} \le ch^{3-\frac{d}{s}}|\log h|.$$

Inserting these estimates into (4.26) and applying again Lemma 4.10 we derive

$$\|u-u_h\|^2+\|y-y_h\|^2\le ch^{3-\frac{d}{s}}|\log h|.$$

If we now choose s sufficiently close to $\frac{d}{d-1}$ we obtain

$$\|u-u_h\|^2+\|y-y_h\|^2\le C_\epsilon h^{4-d-2\epsilon}.$$

Finally, in order to obtain the error bound for y in H^1 we note that

$$\int_\Omega\big(\nabla(y-y_h)\cdot\nabla v_h+(y-y_h)v_h\big)=\int_\Omega(u-u_h)v_h$$

for all $v_h\in X_h$, from which one derives the desired estimate using standard finite element techniques and the bound on $\|u-u_h\|$. □

In general we only expect weak convergence of μ_h to μ. Nevertheless we have the following partial result.

COROLLARY 4.12. *Let $K\subset\bar\Omega$ be compact with $K\cap supp\mu=\emptyset$. For every $\epsilon>0$ there exists a constant C_ϵ such that*

$$\mu_h(K)\le C_\epsilon h^{2-\frac{d}{2}-\epsilon}.$$

Proof. By Lemma 4.17 in the Appendix there exists a nonnegative function $\phi\in C^2(\bar\Omega)$ which satisfies

$$\phi\ge 1 \text{ on } K,\ \phi=0 \text{ on supp}\mu,\ \partial_\nu\phi=0 \text{ on } \partial\Omega.$$

Since $\mu_h\ge 0$ we obtain from (4.11)

$$\begin{aligned}
\mu_h(K) &\le \int_{\bar\Omega}\phi\,d\mu_h=\int_{\bar\Omega}(\phi-R_h\phi)d\mu_h+\int_{\bar\Omega}R_h\phi\,d\mu_h\\
&= \int_{\bar\Omega}(\phi-R_h\phi)d\mu_h+\int_\Omega\big(\nabla p_h\cdot\nabla R_h\phi+p_hR_h\phi\big)-\int_\Omega(y_h-P_hy_0)R_h\phi\\
&= \int_{\bar\Omega}(\phi-R_h\phi)d\mu_h+\int_\Omega\big(\nabla p_h\cdot\nabla\phi+p_h\phi\big)-\int_\Omega(y_h-P_hy_0)R_h\phi\\
&= \int_{\bar\Omega}(\phi-R_h\phi)d\mu_h+\int_\Omega p_h(-\Delta\phi+\phi)-\int_\Omega(y_h-P_hy_0)R_h\phi,
\end{aligned}$$

where R_h is again the Ritz projection. On the other hand, (4.2) and the fact that $\phi=0$ on suppμ imply

$$\int_\Omega(y-y_0)\phi-\int_\Omega p(-\Delta\phi+\phi)=0.$$

Combining this relation with the first estimate we derive

$$\begin{aligned}\mu_h(K) &\leq \int_{\bar\Omega}(\phi - R_h\phi)d\mu_h + \int_\Omega (p_h - p)(-\Delta\phi + \phi) + \int_\Omega (y_h - P_h y_0)(\phi - R_h\phi)\\ &\quad + \int_\Omega (y - y_h - y_0 + P_h y_0)\phi\\ &\leq \|\phi - R_h\phi\|_{L^\infty}\|\mu_h\|_{\mathcal{M}(\bar\Omega)} + \|p - p_h\|\|\phi\|_{H^2} + \big(\|y_h\| + \|P_h y_0\|\big)\|\phi - R_h\phi\|\\ &\quad + \big(\|y - y_h\| + \|y_0 - P_h y_0\|\big)\|\phi\|\\ &\leq C\|\phi - R_h\phi\|_{L^\infty} + C_\epsilon h^{2-\frac{d}{2}-\epsilon} \leq C_\epsilon h^{2-\frac{d}{2}-\epsilon}\end{aligned}$$

in view of (4.3), (4.12) and Theorem 4.11. □

REMARK 4.13. *We mention here a second approach that differs from the one discussed above in the way in which the inequality constraints are realized. Denote by $D_1, \dots, D_m$ the cells of the dual mesh. Each cell D_i is associated with a vertex x_i of $\mathcal{T}_h$ and we have*

$$\bar\Omega = \cup_{i=1}^m D_i, \quad \textit{int}(D_i) \cap \textit{int}(D_j) = \emptyset, i \neq j.$$

In (4.8), we now impose the constraints

$$\fint_{D_j} (y_h - I_h b) \leq 0 \textit{ for } j = 1, \dots, m \tag{4.27}$$

on the discrete solution $y_h = \mathcal{G}_h(u)$. Here, we have abbreviated $\fint_{D_j} f = \frac{1}{|D_j|}\int_{D_j} f$. The measure μ_h that appears in Lemma 4.3 now has the form $\mu_h = \sum_{j=1}^m \mu_j \fint_{D_j} \cdot\, dx$, and the pointwise constraints in (4.13) *are replaced by those of* (4.27)*. Introducing the matrix*

$$C = (c_{ij}), c_{ij} := \fint_{D_i} \phi_j, \ (i, j = 1, \dots, nv),$$

the corresponding optimality system in matrix form then reads

$$\left\{\begin{array}{rcl} Ay &=& Mu,\\ Ap &=& M(y - z) - C\mu,\\ p + \alpha(u - u_0) &=& 0,\\ \mu &=& \max\left(0, \mu + C^t(y - b)\right),\end{array}\right. \tag{4.28}$$

where now $y, p, u, \mu, b \in \mathbb{R}^{nv}$ again denote the nodal vectors associated to the corresponding finite element Ansatz functions. This system admits a unique solution, since it represents the first-order necessary (and also sufficient) optimality system of the following quadratic optimization with convex constraints;

$$(\mathbb{S}) \quad \left\{\begin{array}{l} \min_{(y_h,u)\in Y_h\times U} J(y_h, u) := \frac{1}{2}\|y_h - P_h z\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - u_0\|^2_U\\ \textit{s.t.} y_h = \mathcal{G}_h(u)\\ \textit{and}\\ y_h \in Y_{ad}^h := \left\{y_h \in L^\infty(\Omega), \fint_{D_j}(y_h - I_h b) \leq 0 \textit{ for all } j = 1, \dots, m\right\}.\end{array}\right. \tag{4.29}$$

The error analysis for the resulting numerical method can be carried out in the same way as shown above with the exception of Theorem 4.11, where the bounds on $\tilde{y} - b$ and $\tilde{y}_h - I_h b$ require a different argument. In this case, additional terms of the form

$$\|f - \fint_{D_j} f\|_{L^\infty(D_j)}$$

have to be estimated. Since these will in general only be of order $O(h)$, this analysis would only give $\|u - u_h\|, \|y - y_h\|_{H^1} = O(\sqrt{h})$. The numerical test example in §4 suggests that at least $\|u - u_h\| = O(h)$, but we are presently unable to prove such an estimate.

4.3. Numerical examples.

EXAMPLE 4.14. The following test problem is taken - in a slightly modified form - from [**58**], Example 6.2. Let $\Omega := B_1(0)$, $\alpha > 0$,

$$y_0(x) := 4 + \frac{1}{\pi} - \frac{1}{4\pi}|x|^2 + \frac{1}{2\pi}\log|x|, \quad u_0(x) := 4 + \frac{1}{4\alpha\pi}|x|^2 - \frac{1}{2\alpha\pi}\log|x|$$

and $b(x) := |x|^2 + 4$. We consider the cost functional

$$J(u) := \frac{1}{2}\int_\Omega |y - y_0|^2 + \frac{\alpha}{2}\int_\Omega |u - u_0|^2,$$

where $y = \mathcal{G}(u)$. By checking the optimality conditions of first order one verifies that $u \equiv 4$ is the unique solution of (4.1) with corresponding state $y \equiv 4$ and adjoint states

$$p(x) = \frac{1}{4\pi}|x|^2 - \frac{1}{2\pi}\log|x| \text{ and } \mu = \delta_0.$$

The finite element counterparts of y, u, p and μ are denoted by y_h, u_h, p_h and μ_h.

To investigate the experimental order of convergence (see (2.29) for its definition) for our model problem we choose a sequence of uniform partitions of Ω containing five refinement levels, starting with eight triangles forming a uniform octagon as initial triangulation of the unit disc. The corresponding grid sizes are $h_i = 2^{-i}$ for $i = 1, \ldots, 5$. As error functionals we take $E(h) = \|(u, y) - (u_h, y_h)\|$ and $E(h) = \|(u, y) - (u_h, y_h)\|_{H^1}$ and note, that the error $p - p_h$ is related to $u - u_h$ via (4.12). We solve problems (4.8) using the QUADPROG routine of the MATLAB OPTIMIZATION TOOLBOX. The required finite element matrices for the discrete state and adjoint systems are generated with the help of the MATLAB PDE TOOLBOX. Furthermore, for discontinuous functions f we use the quadrature rule

$$\int_\Omega f(x)dx \approx \sum_{T\in\mathcal{T}_h} f\left(x_{s(T)}\right)|T|,$$

where $x_{s(T)}$ denotes the barycenter of T. In all computations we set $\alpha = 1$.

In Table 4.9, we present EOCs for problem (4.8) (case $S = D$) and the approach sketched in Remark 4.13 (case $S = M$). As one can see, the error $\|u - u_h\|$ behaves in the case $S = D$ as predicted by Theorem 4.11, whereas the errors $\|y - y_h\|$ and $\|y - y_h\|_{H^1}$ show a better convergence behaviour. On the finest level we have $\|u - u_h\| = 0.003117033$, $\|y - y_h\| = 0.000123186$ and $|y - y_h|_{H^1} =$

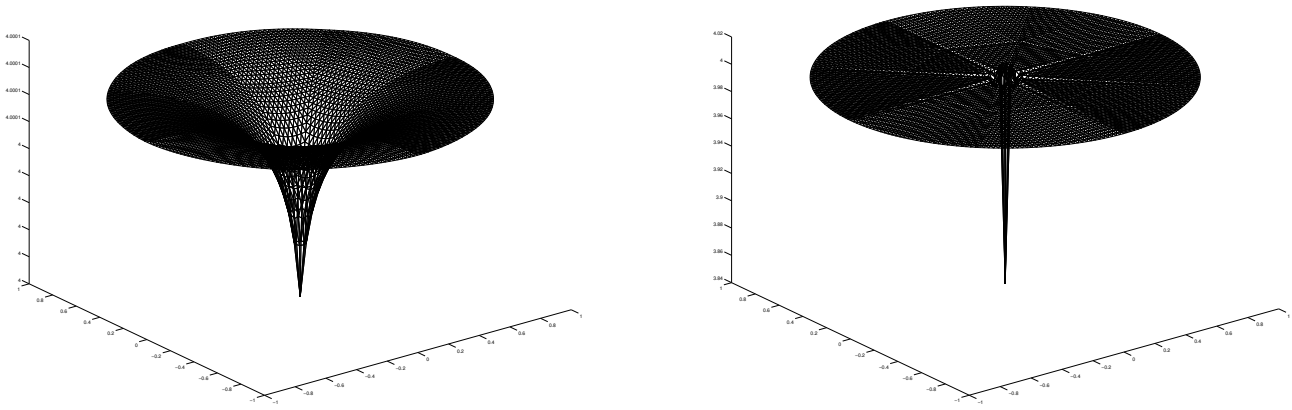


FIGURE 4.1. Numerically computed state y_h (left) and control u_h (right) for $h = 2^{-5}$ in the case $S = D$.

0.000083757. Furthermore, all coefficients of μ_h are equal to zero, except the one in front of δ_0 whose value is 0.99946494 The errors $\|u - u_h\|$, $\|y - y_h\|$ and $\|y - y_h\|_{H^1}$ in the case $S = M$ show a better EOC than in the case $S = D$. This can be explained by the fact that the exact solutions y and u are very smooth, and that the relaxed form of the state constraints introduce a smearing effect on the numerical solutions at the origin. On the finest level we have $\|u - u_h\| = 0.001020918$, $\|y - y_h\| = 0.000652006$ and $|y - y_h|_{H^1} = 0.000037656$. Furthermore, the coefficient of μ_h corresponding to the patch containing the origin has the value 1.0640946.

Figures 4.1 and 4.2 present the numerical solutions y_h and u_h for $h = 2^{-5}$ in the case $S = D$ and $S = M$, respectively. We note that using equal scales on all axes would give completely flat graphs in all four figures.

	(S=D)	(S=M)	(S=D)	(S=M)	(S=D)	(S=M)
Level	$\|u - u_h\|$	$\|u - u_h\|$	$\|y - y_h\|$	$\|y - y_h\|$	$\|y - y_h\|_{H^1}$	$\|y - y_h\|_{H^1}$
1	0.788985	0.654037	0.536461	0.690302	0.860516	0.688531
2	0.759556	1.972784	1.147861	2.017836	1.272400	2.015602
3	0.919917	1.962191	1.389378	2.004383	1.457095	2.004286
4	0.966078	1.856687	1.518381	1.989727	1.564204	1.990566
5	0.986686	1.588722	1.598421	1.979082	1.632772	1.979945

TABLE 4.9. Experimental order of convergence

EXAMPLE 4.15. The second test problem is taken from [**57**], Example 2. It reads

$$\min_{u \in L^2(\Omega)} J(u) = \frac{1}{2} \int_\Omega |y - y_0|^2 + \frac{1}{2} \int_\Omega |u - u_0|^2$$
$$\text{subject to } y = \mathcal{G}(u) \text{ and } y(x) \geq b(x) \text{ in } \Omega.$$

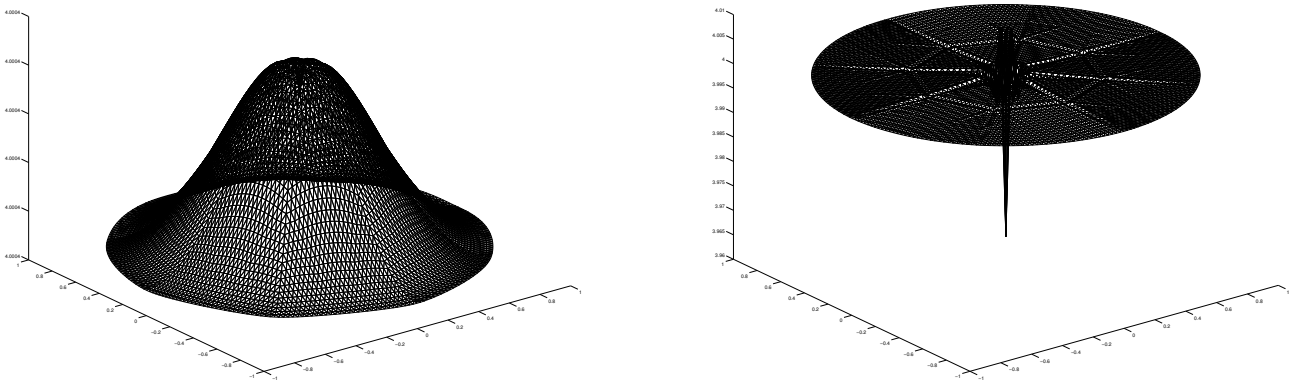


FIGURE 4.2. Numerically computed state y_h (left) and control u_h (right) for $h = 2^{-5}$ in the case $S = M$.

Here, Ω denotes the unit square,

$$b(x) = \begin{cases} 2x_1 + 1, & x_1 < \frac{1}{2}, \\ 2, & x_1 \geq \frac{1}{2}, \end{cases} \quad y_0(x) = \begin{cases} x_1^2 - \frac{1}{2}, & x_1 < \frac{1}{2}, \\ \frac{1}{4}, & x_1 = \frac{1}{2}, \\ \frac{3}{4}, & x_1 > \frac{1}{2}, \end{cases}$$

and

$$u_0(x) = \begin{cases} \frac{5}{2} - x_1^2, & x_1 < \frac{1}{2}, \\ \frac{9}{4}, & x_1 \geq \frac{1}{2}. \end{cases}$$

The exact solution is given by $y \equiv 2$ and $u \equiv 2$ in Ω. The corresponding Lagrange multiplier $p \in H^1(\Omega)$ is given by

$$p(x) = \begin{cases} \frac{1}{2} - x_1^2, & x_1 < \frac{1}{2}, \\ \frac{1}{4}, & x_1 \geq \frac{1}{2}. \end{cases}$$

The multiplier μ has the form

$$\int_{\bar{\Omega}} f d\mu = \int_{\{x_1 = \frac{1}{2}\}} f ds + \int_{\{x_1 > \frac{1}{2}\}} f dx, \qquad f \in C^0(\bar{\Omega}). \tag{4.30}$$

In our numerical computations we use uniform grids generated with the POIMESH function of the MATLAB PDE TOOLBOX. Integrals containing y_0, u_0 are numerically evaluated by substituting y_0, u_0 by their piecewise linear, continuous finite element interpolations $I_h y_0, I_h u_0$. The grid size of a grid containing l horizontal and l vertical lines is given by $h_l = \frac{\sqrt{2}}{l+1}$. Fig. 4.3 presents the numerical results for a grid with $h = \frac{\sqrt{2}}{36}$ in the case (S=D). The corresponding values of μ_h on the same grid are presented in Fig. 4.4. They reflect the fact that the measure consists of a lower dimensional part which is concentrated on the line $\{x \in \Omega \,|\, x_1 = \frac{1}{2}\}$ and a regular part with a density $\chi_{|\{x_1 > \frac{1}{2}\}}$. We again note that using equal scales on all axes would give completely flat graphs for y_h as well as for u_h.

We compute EOCs for the two different sequences of grid-sizes $s_o = \{h_1, h_3, \ldots, h_{19}\}$ and $s_e = \{h_0, h_2, \ldots, h_{18}\}$. We note that the grids corresponding to s_o contain the line $x_1 = \frac{1}{2}$. Table 4.10

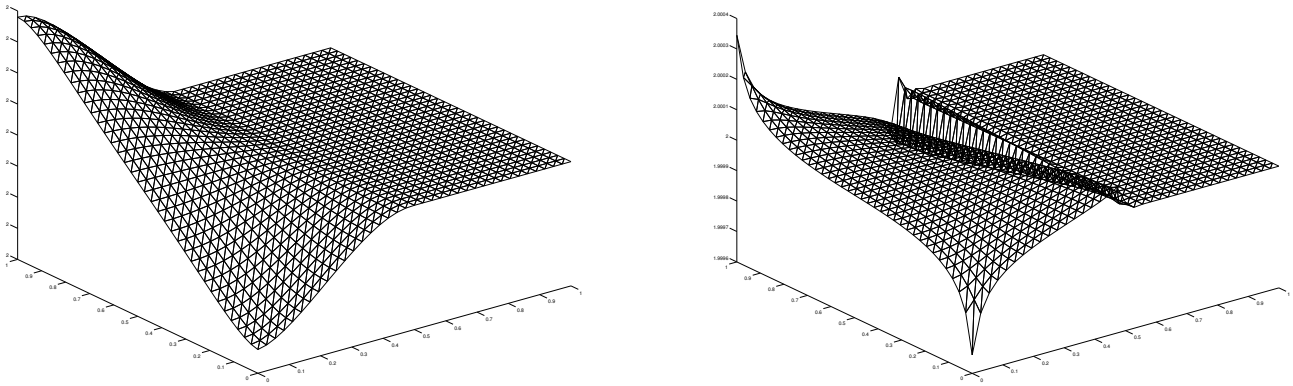


FIGURE 4.3. Numerically computed state y_h (left) and control u_h (right) for $h = \frac{\sqrt{2}}{36}$ in the case $S = D$.

presents EOCs for s_o, and Table 4.11 presents EOCs for s_e. For the sequence s_o we observe superconvergence in the case (S=D), although the discontinuous function y_0 for the quadrature is replaced by its piecewise linear, continuous finite element interpolant $I_h y_0$. Let us note that further numerical experiments show that the use of the quadrature rule (4.14) for integrals containing the function y_0 decreases the EOC for $\|u-u_h\|$ to $\frac{3}{2}$, whereas EOCs remain close to 2 for the other two errors $\|y-y_h\|$ and $\|y - y_h\|_{H^1}$. For this sequence also the case (S=M) behaves twice as good as expected by our arguments in Remark 4.13. For the sequence s_e the error $\|u - u_h\|$ in the case (S=D) approximately behaves as predicted by our theory, in the case (S=M) it behaves as for the sequence s_o. The errors $\|y - y_h\|$ and $\|y - y_h\|_{H^1}$ behave that well, since the the exact solutions y and u are very smooth. For h_{19} we have in the case (S=D) $\|u - u_h\| = 0.000103428$, $\|y - y_h\| = 0.000003233$ and $|y - y_h|_{H^1} = 0.000015155$, and in the case (S=M) $\|u - u_h\| = 0.011177577$, $\|y - y_h\| = 0.000504815$ and $|y - y_h|_{H^1} = 0.001547907$. We observe that the errors in the case $S = M$ are two magnitudes larger than in the case (S=D). This can be explained by the fact that an Ansatz for the multiplier μ with a linear combination of Dirac measures is better suited to approximate measures concentrated on singular sets than a piecewise constant Ansatz as in the case (S=M). Finally, Table 4.12 presents $\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$ and $\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$ for s_o in the case (S=D). As one can see $\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$ tends to 1, the length of $\{x_1 = 1/2\}$, and $\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$ tends to 1/2, the area of $\{x_1 > 1/2\}$. These numerical findings indicate that $\mu_h = \sum_{i=1}^{m} \mu_i \delta_{x_i}$ well approximates μ, since $\int_{\bar{\Omega}} d\mu_h = \sum_{i=1}^{m} \mu_i$, and that μ_h also well resolves the structure of μ, see (4.30). For all numerical computations of this example we have $\mu_i = 0$ for $x_i \in \{x_1 < 1/2\}$.

4.4. Appendix.

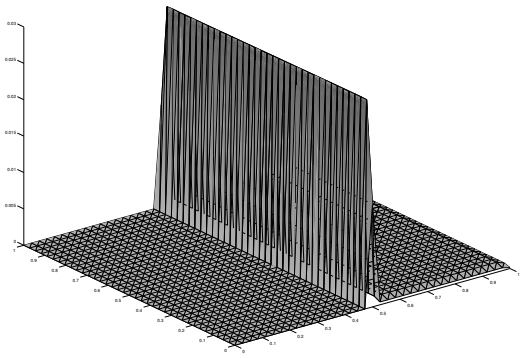


FIGURE 4.4. Numerically computed multiplier μ_h for $h = \frac{\sqrt{2}}{36}$ in the case $S = D$.

	(S=D)	(S=M)	(S=D)	(S=M)	(S=D)	(S=M)
Level	$\|u - u_h\|$	$\|u - u_h\|$	$\|y - y_h\|$	$\|y - y_h\|$	$\|y - y_h\|_{H^1}$	$\|y - y_h\|_{H^1}$
1	1.669586	0.448124	1.417368	0.544284	1.594104	0.384950
2	1.922925	1.184104	1.990906	1.473143	1.992097	1.239771
3	2.000250	1.456908	2.101633	1.871948	2.080739	1.745422
4	2.029556	1.530303	2.125168	2.427634	2.108241	2.348036
5	2.041913	1.260744	2.124773	2.743918	2.116684	2.563363
6	2.047106	1.142668	2.117184	1.430239	2.117739	1.318617
7	2.048926	1.177724	2.107828	1.503463	2.115633	1.409563
8	2.049055	1.194893	2.098597	1.578342	2.112152	1.497715
9	2.048312	1.194802	2.090123	1.622459	2.108124	1.549495

TABLE 4.10. Experimental order of convergence, $x_1 = \frac{1}{2}$ grid line

LEMMA 4.16. *Let* $\frac{2d}{d+2} \leq s \leq 2$ *and* $v \in W^{1,s}(\Omega)$. *Then*

$$\|v - P_h v\| \leq C h^{1+\frac{d}{2}-\frac{d}{s}} \|v\|_{W^{1,s}}.$$

Proof. The assertion is clear if $s = \frac{2d}{d+2}$ or if $s = 2$ so that we may assume $\frac{2d}{d+2} < s < 2$. Let us write

$$\int_\Omega |v - P_h v|^2 = \int_\Omega |v - P_h v|^{\frac{sd-2d+2s}{s}} |v - P_h v|^{\frac{d(2-s)}{s}}$$

	(S=D)	(S=M)	(S=D)	(S=M)	(S=D)	(S=M)
Level	$\|u-u_h\|$	$\|u-u_h\|$	$\|y-y_h\|$	$\|y-y_h\|$	$\|y-y_h\|_{H^1}$	$\|y-y_h\|_{H^1}$
1	0.812598	0.460528	1.160789	2.154570	0.885731	1.473561
2	1.361946	0.406917	2.042731	0.597846	1.918942	0.405390
3	1.228268	1.031763	1.832573	1.392796	1.700124	1.088595
4	1.245030	1.262257	1.678233	1.621110	1.570580	1.392408
5	1.252221	1.416990	1.646124	1.844165	1.554434	1.686808
6	1.256861	1.505759	1.696309	2.128776	1.620231	2.021210
7	1.264456	1.489061	1.627539	2.507863	1.559065	2.415552
8	1.260157	1.316627	1.640964	2.989867	1.580113	2.818148
9	1.265599	1.169109	1.686579	1.601263	1.635084	1.460153

TABLE 4.11. Experimental order of convergence, $x_1 = \frac{1}{2}$ not a grid line

Level	$\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$	$\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$
1	1.13331662624081	0.36552954225441
2	1.06315278164899	0.43644163287114
3	1.03989323182608	0.45990635060758
4	1.02893022155910	0.47095098878247
5	1.02265064139378	0.47727091447291
6	1.01855129775903	0.48139306499280
7	1.01569011772403	0.48426838085822
8	1.01359012331610	0.48637773715316
9	1.01198410389649	0.48799027450619

TABLE 4.12. Approximation of the multiplier in the case (S=D), $x_1 = \frac{1}{2}$ grid line

and apply Hölder's inequality with $p = \frac{s^2}{sd-2d+2s}, q = \frac{s^2}{(d-s)(2-s)}$ which implies

$$\begin{aligned} \|v - P_h v\|^2 &\leq \|v - P_h v\|_{L^s}^{\frac{sd-2d+2s}{s}} \|v - P_h v\|_{L^{\frac{ds}{d-s}}}^{\frac{d(2-s)}{s}} \\ &\leq \|v - P_h v\|_{L^s}^{\frac{sd-2d+2s}{s}} \big(\|v\|_{L^{\frac{ds}{d-s}}} + \|P_h v\|_{L^{\frac{ds}{d-s}}} \big)^{\frac{d(2-s)}{s}}. \end{aligned}$$

We infer from [**24**] that

$$\|v - P_h v\|_{L^s} \leq Ch \|v\|_{W^{1,s}}, \quad \|P_h v\|_{L^{\frac{ds}{d-s}}} \leq C \|v\|_{L^{\frac{ds}{d-s}}}$$

which, together with the continuous embedding $W^{1,s}(\Omega) \hookrightarrow L^{\frac{ds}{d-s}}(\Omega)$, gives

$$\|v - P_h v\|^2 \leq c h^{\frac{sd-2d+2s}{s}} \|v\|_{W^{1,s}}^2$$

so that the assertion follows. □

LEMMA 4.17. *Suppose that K and $\tilde{K}$ are two disjoint compact subsets of $\bar{\Omega}$. Then there exists a nonnegative function $\phi \in C^2(\bar{\Omega})$ which satisfies*

$$\partial_\nu \phi = 0 \text{ on } \partial\Omega, \quad \phi \geq 1 \text{ on } K, \quad \phi = 0 \text{ on } \tilde{K}.$$

Proof. For $r > 0$ let us define $\Omega_r := \{x \in \bar{\Omega} \,|\, \text{dist}(x, \partial\Omega) < r\}$. In view of the smoothness of $\partial\Omega$ there exists $\delta > 0$ such that for each $x \in \Omega_\delta$ there exists a unique point $y = y(x) \in \partial\Omega$ with

$$x = y - \text{dist}(x, \partial\Omega)\nu(y)$$

(see [**31**], 14.6). Since $K \cap \tilde{K} = \emptyset$ we may assume that $\text{dist}(K, \tilde{K}) > \delta$. Let us define

$$\Gamma_K := \{y(x) \,|\, x \in K \cap \overline{\Omega_{\frac{\delta}{2}}}\}, \quad \Gamma_{\tilde{K}} := \{y(x) \,|\, x \in \tilde{K} \cap \overline{\Omega_{\frac{\delta}{2}}}\}.$$

Γ_K and $\Gamma_{\tilde{K}}$ are disjoint, compact subsets of $\partial\Omega$, since $\text{dist}(K, \tilde{K}) > \delta$ and $x \mapsto y(x)$ is continuous. Let $\phi_1 \in C^2(\partial\Omega)$ be a nonnegative function satisfying $\phi_1 \geq 1$ on Γ_K, $\phi_1 = 0$ on $\Gamma_{\tilde{K}}$. By setting $\phi_1(x) = \phi_1(y(x))$ we extend ϕ_1 as a C^2 function to Ω_δ. Clearly, $\partial_\nu \phi_1 = 0$ on $\partial\Omega$. Let $\psi \in C^2(\bar{\Omega})$ be a nonnegative cut–off function with $\psi = 1$ in $\Omega_{\frac{\delta}{4}}$ and $\psi = 0$ in $\bar{\Omega} \setminus \Omega_{\frac{\delta}{2}}$. Then $\phi_2 := \psi\phi_1$ satisfies

$$\partial_\nu \phi_2 = 0 \text{ on } \partial\Omega, \quad \phi_2 \geq 1 \text{ on } K \cap \Omega_{\frac{\delta}{4}}, \quad \phi_2 = 0 \text{ on } \tilde{K}.$$

Finally, choose a nonnegative function $\phi_3 \in C^2(\bar{\Omega})$ with

$$\phi_3 \geq 1 \text{ on } K \cap (\bar{\Omega} \setminus \Omega_{\frac{\delta}{4}}), \quad \phi_3(x) = 0 \text{ if } \text{dist}(x, K \cap (\bar{\Omega} \setminus \Omega_{\frac{\delta}{4}})) \geq \frac{\delta}{8}.$$

Then, $\partial_\nu \phi_3 = 0$ on $\partial\Omega$, $\phi_3 = 0$ on $\tilde{K}$ and $\phi := \phi_2 + \phi_3$ has the required properties. □

CHAPTER 4

Applications

René Pinnau
Fachbereich Mathematik
Universität Kaiserslautern

1. Introduction

The following three sections are devoted to the study of three industrial applications, in which optimization with partial differential equations plays a crucial role. To give you an overview on the the different mathematical settings which can be handled with the general optimal control calculus, we will focus on large scale optimal control problems involving the three well–known types of partial differential equations, namely elliptic, parabolic and hyperbolic equations. And since real world applications lead generally to mathematically involved problems, we study especially nonlinear systems of equations. The examples are chosen in such a way that they are up to date and represent the present status of the mathematical tools, which are employed for their solution. The industrial fields we will cover are ranging from semiconductor design over glass production to automated traffic control. Since most people will not be familiar with the underlying physics and mathematical models, we will start each section with a modelling part for the derivation of the equations, which is followed by the analytical and numerical study of the related optimal control problems.

2. Optimal Semiconductor Design

In the first lecture on numerical mathematics each student learns that the enormous speed-up of numerical simulations during the last 30 years is stemming from two facts, namely the significant improvement of algorithms and the ongoing miniaturization in electronics which allows for faster computing times. During the last lectures we already learned how one can develop fast numerical algorithms, so we will now shortly study the impact mathematical of optimization on advanced semiconductor design. In this industry there are several stages at which optimization and control is necessary. Think e.g. of circuit design, thermal control of the circuit board or, on a smaller level, the design of each semiconductor device, or even the control of the production process itself. Presently, the most popular semiconductor device is the so-called MOSFET (metal oxide silicium field effect transistor), which is employed in many applications (see Figure 2.1) [**73**]. In the design cycle one changes the geometry

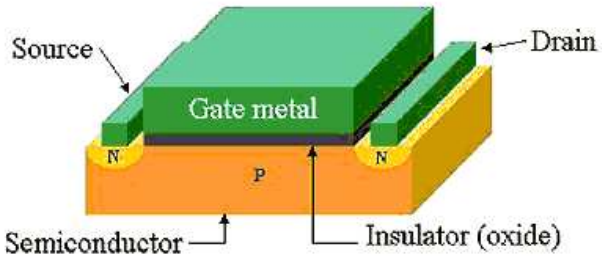


FIGURE 2.1. MOSFET Device

of the device (miniaturization!) and the so-called doping profile, which describes the the density of charged background ions describing the specific type of the device. In the conventional design cycle simulation tools are employed to compute the so called current-voltage characteristics of the device, from which the engineer can deduce the many performance characteristica of the device. This is done for a certain set of design parameters and then, the parameters are adjusted empirically. Thus, the total design time depends crucially on the knowledge and experience of the electrical engineer.

In standard applications a working point, i.e. a certain voltage–current pair, for the device is fixed. Especially for MOSFET devices in portable systems it is most important to have on the one hand a low leakage current (in the off–state), which maximizes the battery lifetime, and on the other hand to maximize the drive current (in the on–state) [**72**]. Now, we want to study how one can apply the previously introduced techniques to optimize such a device, especially we want to find a solution to the following design question:

> Is it possible to gain an amplified current at the working point only by a slight change of the doping profile?

We will proceed in several steps. First, we motivate the system of nonlinear equations, which is describing the electronic behavior of the semiconductor device. There are many semiconductor models at hand, but we will concentrate in the next section on the so-called drift diffusion model. Then, we state the optimization problem in mathematical terms and study its solution.

2.1. Modeling. In this section we give a brief introduction into the physics and mathematics of semiconductor devices, which is far from being comprehensive. If the reader wants to go into the details we suggest to have look into the books by *Sze* [**73**] or *Selberherr* [**71**]. Clearly, the most important features of semiconductor devices are due to electromagnetic effects, i.e such a device reacts on applied voltages. Here, we will only consider electrostatic effects ignoring electrodynamics and magnetic phenomena. Further, we will ignore quantum effects, which are getting increasingly important due to the shrinking device size.

In general one might say, the semiconductor is a specifically modified crystal. The modification of the underlying crystal (consisting e.g. of Silicium atoms) are due a preparation of the surface (to build

metallic or insulating contacts) and due to the implantation of impurities (e.g. Aluminum atoms). This has to be done since the electronic behavior of a homogeneous semiconductor is rather boring. But due to the replacement of atoms in the crystal, which is the so-called doping process, we get an inhomogeneous semiconductor which exhibits the desired electronic performance. There exist several sophisticated technologies to achieve the desired doping. And since these processes can be controlled on the nanometer scale, it is possible to fabricate nowadays devices, which have a gate length of 60 nanometers. Nevertheless, there is still a strong need for the (automated) design of the semiconductor device, i.e. how the doping has to be adjusted such that the device shows the desired behavior.

Normally, one implants atoms which have more (donator atoms) or less (acceptor atoms) electrons participating at binding interactions. While Silicium atoms have four binding electrons, Phosphor atoms have five and Aluminum atoms have three. If a Silicium atom is replaced by a Phosphor atom we have one additional electron, which is not necessary for the binding and which can therefor move freely in the crystal. Hence, the Phosphor atom donates one electron to the conductivity band. But if the Silicium atom is replaced by an Aluminum atom, then the additional electron which is needed for the binding is taken from the surrounding atoms and a hole is generated.

Note, that also these holes contribute to the charge transport, since the Silicium atom which is then positively charged will attract an electron from one surrounding atom. This process repeats and charge transport takes place by the missing electrons, i.e. the holes. Experiments suggest that the charge transport by holes can be considered as charge transport by real particles which have a positive charge q.

Now that we have a feeling how charge transport takes place in the semiconductor, let us assume that the semiconductor occupies a bounded domain $\Omega \subset \mathbb{R}^3$. So far, our assumptions imply that there is an instantenous *electric field* $\mathbf{E}(x)$, $\mathbf{x} \in \Omega$ which is only determined by the position of the charged particles.

EXAMPLE 2.1. *In school one learns that the force* $\mathbf{F}$ *acting between two charges* q_1 *and* q_2 *in the points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *is given by Coulomb's law*

$$\mathbf{F} = q_1 \cdot q_2 \cdot \frac{\mathbf{x}_2 - \mathbf{x}_1}{\left|\mathbf{x}_2 - \mathbf{x}_1\right|^3},$$

where $\left|\mathbf{z}\right|^2 = \left|(z_1, z_2, z_3)^T\right|^2 = z_1^2 + z_2^2 + z_3^2$.

Hence, we could describe the overall charge transport the semiconductor just by considering an ensemble of charged particles interacting via the electric field: Put an electron with velocity $\mathbf{v}_0$ *in the point* $\mathbf{x}_0 \in \Omega$. *Then there will be an interaction of the electron with the electric field, which can be describes by Coulomb's law and Newton's second law:*

$$m_e \frac{d^2\,\mathbf{x}(t)}{dt^2} = -q\mathbf{E}(\mathbf{x}(t)),$$

where m_e *is the electron mass and* q *is the elementary charge. Further, we would have the initial conditions*

$$\mathbf{x}(t=0) = \mathbf{x}_0 \quad \textit{and} \quad \ldots \mathbf{x}t(t=0) = \mathbf{v}_0.$$

The above example shows us that this ensemble of electrons will act as an electronic device, since the presence of an electric field leads to an energy transport performed by the charged particles. Clearly, this description is computationally not efficient since there will be billions of particles even in a very tiny piece of the semiconductor. For this reason we introduce the *electron density* $n(\mathbf{x})$ with unit m^{-3} (number of particles per cubic meter). This function can be interpreted as follows: Consider again that the semiconductor occupies the domain Ω and assume that this domain contains a large number of electrons. Now assume that there is a subdomain $\omega \subset \Omega$ which is large compared to the size of one electron. Then the total number of electrons in this subdomain is given by

$$\int_\omega n(\mathbf{x})\ d\mathbf{x}.$$

Since the number of particles in a domain is always nonnegative, we directly have $n \geq 0$. The density of holes $p(\mathbf{x})$ is defined respectively.

Further, we introduce the *mean electron velocity* v_n, which has the following meaning: Assume that there is a subdomain $\omega \subset \Omega$ which is large compared to the size of one electron. Then the avergage velocity of electrons in this subdomain is given by

$$\int_\omega \mathbf{v}_n(\mathbf{x})\ d\mathbf{x}.$$

In analogy, we define the mean hole velocity v_p. Finally, we introduce the *electron and hole current densities* by

$$\mathbf{J}_n = q\, n\, \mathbf{v}_n, \qquad \mathbf{J}_p = -q\, p\, \mathbf{v}_p.$$

Now, we will motivate the set of partial differential equations connecting those quantities.

2.1.1. *The Potential Equation.* First, we give a mathematically tractable relation between the charge densities and the electric field. This can be done by the introduction of the *electrostatic potential* V which is defined as a solution of *Poisson's equation*

$$-\epsilon \Delta V = q(n - p + N_A - N_D),$$

where ϵ is the dielectric constant of the semiconductor material and N_A, N_D are the densities of acceptor and donator atoms, respectively. Here, we assumed that each donator atom contributes just one electron as well as each acceptor atom contributes just one hole. One can show that then the electric field can be expressed as

$$\mathbf{E} = -\nabla V.$$

Note that the potential is not uniquely defined by this equations, since one might add an arbitrary constant and will still get the same electric field. Especially, if the equation is posed on a bounded domain the prescription of boundary data will be essential. Introducing the *doping profile*

$$C(\mathbf{x}) := N_D(\mathbf{x}) - N_A(\mathbf{x})$$

we get the equation

$$-\epsilon \Delta V = q(n - p - C), \tag{2.1}$$

where the function $q(n - p - C)$ is called the *space charge*.

REMARK 2.1. *For the sake of simplicity and notational convenience we assume in the following that all physical parameters in our model are constant.*

2.1.2. *The Continuity Equations.* The current density $\mathbf{J}(x)$ in the semiconductor consists of the electron and the hole current density, i.e.

$$\mathbf{J} = \mathbf{J}_n + \mathbf{J}_p.$$

Note that that only the full current can be measured. If we assume that we have conservation of charged particles and no generation and recombination processes are present, then it holds for each subdomain $\omega \subset \Omega$ with smooth boundary Σ that

$$I_\Sigma = \int_\Sigma \mathbf{J} \cdot \nu \, ds = 0.$$

Hence, Gauß' theorem implies directly

$$\int_\omega \operatorname{div} \mathbf{J} \, dx = 0$$

and since this holds for any subdomain ω the variational lemma yields the differential form of the continuity equation

$$\operatorname{div} \mathbf{J} = 0.$$

Taking into account $\mathbf{J} = \mathbf{J}_n + \mathbf{J}_p$ we get

$$\operatorname{div} \mathbf{J}_n = \operatorname{div} \mathbf{J}_p = 0.$$

2.1.3. *The Current Densities.* These equations are by far not sufficient to prescribe the charge transport in the semiconductor. Especially, we need additional relations for the current densities. In many applications one can successfully assume that the current densities are entirely determined by the particle densities and by the electrostatic potential. Here, we will consider two contributions, namely the convective current density and the diffusion current density.

The *convective current density* encounters for the acceleration of charged particles in an electric field and it is assumed to be essentially proportional to the electric field, i.e.

$$\mathbf{J}_n^{conv} = q \, \mu_n \, n \, \nabla V, \qquad \mathbf{J}_p^{conv} = -q \, \mu_p \, n \, \nabla V,$$

where μ_n and μ_p are the mobilities of electrons and holes, respectively.

The *diffusion current density* accounts for the ensemble of many charged particles which tends to compensate density fluctuations. Hence, this causes an additional movement of the particles, the so–called diffusion. We assume that these diffusion current densities are given by

$$\mathbf{J}_n^{diff} = q \, D_n \, \nabla n, \qquad \mathbf{J}_p^{diff} = q \, D_p \, \nabla p,$$

where the diffusion coefficients D_n and D_p are assumed to be positive constants.

Finally, we get the current density relations

$$\mathbf{J}_n = \mathbf{J}_n^{diff} + \mathbf{J}_n^{conv} = q\, D_n\, \nabla n + q\, \mu_n\, n\, \nabla V,$$
$$\mathbf{J}_p = \mathbf{J}_p^{diff} + \mathbf{J}_p^{conv} = q\, D_p\, \nabla p - q\, \mu_p\, p\, \nabla V.$$

These can be further simplified by assuming the *Einstein relations*

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{k_B\, T}{q} =: U_T,$$

where T is the (constant) temperature of electrons and holes and k_B is the Boltzmann constant. Here, U_T is called the thermal voltage.

Summarizing we get the so-called *drift diffusion model* which was first introduced by Van Rosbroeck (cf. [**73, 54**] and the references therein):

$$\mathbf{J}_n = q\,(D_n \nabla n + \mu_n\, n\, \nabla V)\,, \tag{2.2a}$$
$$\mathbf{J}_p = -q\,(D_p \nabla p - \mu_p\, p\, \nabla V)\,, \tag{2.2b}$$
$$\operatorname{div} \mathbf{J}_n = 0, \tag{2.2c}$$
$$\operatorname{div} \mathbf{J}_p = 0, \tag{2.2d}$$
$$-\epsilon \Delta V = q(n - p - C). \tag{2.2e}$$

Henc, the drift diffusion model consists of a coupled system of nonlinear elliptic partial differential equations, which makes its mathematical analysis quite involved. To get a well posed problem we have further to prescribe additional boundary data. We assume that the boundary $\partial\Omega$ of the domain Ω splits into two disjoint parts Γ_D and Γ_N, where Γ_D models the Ohmic contacts of the device and Γ_N represents the insulating parts of the boundary. Let ν denote the unit outward normal vector along the boundary. First, assuming charge neutrality ($n - p - C = 0$) and thermal equilibrium ($np = n_i^2$) at the Ohmic contacts Γ_D and, secondly, zero current flow and vanishing electric field at the insulating part Γ_N yields the following set of boundary data

$$n = n_D, \quad p = p_D, \quad V = V_D \quad \text{on } \Gamma_D, \tag{2.2f}$$
$$\mathbf{J}_n \cdot \nu = \mathbf{J}_p \cdot \nu = \nabla V \cdot \nu = 0 \quad \text{on } \Gamma_N, \tag{2.2g}$$

where n_D, p_D, V_D are given by

$$n_D = \frac{C + \sqrt{C^2 + 4\, n_i^2}}{2}, \quad p_D = \frac{-C + \sqrt{C^2 + 4\, n_i^2}}{2}, \quad V_D = -U_T\, \log\left(\frac{n_D}{n_i}\right) + U, \quad \text{on } \Gamma_D.$$

Here, U denotes the applied biasing voltage, which is e.g. applied between the source and the drain contact of the MOSFET device, and n_i the intrinsic carrier density of the semiconductor. Note, that the main unknowns are the densities n and p as well as the potential V.

2.1.4. *Scaling.* These model is not only challenging from the analytical point view, but also due to the severe numerical problems it is causing. To understand this it is most convenient to rewrite the

equations in nondimensional form using following diffusion scaling

$$n \to C_m\,\tilde{n}, \qquad p \to C_m\,\tilde{p}, \qquad \mathbf{x} \to L\,\tilde{\mathbf{x}},$$
$$C \to C_m\,\tilde{C}, \qquad V \to U_T\,\tilde{V}, \qquad \mathbf{J}_{n,p} \to \frac{q\,U_T\,C_m\,\mu_{n,p}}{L}\tilde{\mathbf{J}}_{n,p}$$

where L denotes a characteristic device length, C_m the maximal absolute value of the background doping profile and $\mu_{n,p}$ a characteristic value for the respective mobilities. Defining the dimensionless *Debye length*

$$\lambda^2 = \frac{\epsilon\,U_T}{q\,C_m\,L^2}$$

the scaled equations read

$$\operatorname{div}\mathbf{J}_n = 0, \qquad \mathbf{J}_n = \nabla n + n\,\nabla V, \tag{2.3a}$$
$$\operatorname{div}\mathbf{J}_p = 0, \qquad \mathbf{J}_p = -(\nabla p - p\,\nabla V), \tag{2.3b}$$
$$-\lambda^2\Delta V = n - p - C, \tag{2.3c}$$

where we omitted the tilde for notational convenience. The Dirichlet boundary conditions transform to

$$n_D = \frac{C + \sqrt{C^2 + 4\,\delta^4}}{2}, \quad p_D = \frac{-C + \sqrt{C^2 + 4\,\delta^4}}{2}, \quad V_D = -\log\left(\frac{n_D}{\delta^2}\right) + U, \quad \text{on } \Gamma_D, \tag{2.3d}$$

where $\delta^2 = n_i/C_m$ denotes the scaled intrinsic density.

For typical device parameters we get for the Debye length $\lambda^2 = 10^{-3}$ and $\delta^2 = 10^{-4}$. Hence, the DD model is singular perturbed which has to be encountered in the numerical treatment. One realizes that there will be large gradients in the potential and thus also in the particle densities near to rapid changes in the doping profile, the so called junctions. In general, one employs the Scharfetter–Gummel discretization [**68**] for the discretization, which can be interpreted as an exponentially fitted scheme.

2.2. Optimization. After setting up the underlying model equations we turn our attention again to the design question. Remember that the main objective in optimal semiconductor design is to get an improved current flow at a specific contact of the device, e.g. focusing on the reduction of the leakage current (in the off-state) in MOSFET devices or maximizing the drive current (in the on–state) [**72**]. In both cases a certain working point is fixed and one tries to achieve the objective by a change of the doping profile C. Hence, the objective of the optimization, the current flow over a contact Γ, is given by

$$I = \int_\Gamma \mathbf{J}\cdot\nu\;ds = \int_\Gamma (\mathbf{J}_n + \mathbf{J}_p)\cdot\nu\;ds, \tag{2.4}$$

where the current density $\mathbf{J}$ for a specific doping profile C is given by the solution of the DD model (2.3).

Now, we want to embed the design question into the optimal control context. We want to minimize a cost functional of tracking type

$$Q(n,p,V,C) := \frac{1}{2}\left|\int_\Gamma \mathbf{J}\cdot\nu\, ds - I^*\right|^2 + \frac{\gamma}{2}\int_\Omega \left|\nabla(C-\bar{C})\right|^2\, dx, \tag{2.5}$$

where $\bar{C}$ is a given reference doping profile, I^* is a desired current flow, and the parameter $\gamma > 0$ allows to adjust the deviations from $\bar{C}$. Clearly, C is acting here as the control parameter. The introduction of $\bar{C}$ is necessary to ensure that we change not the type of the semiconductor device during the optimization.

Since the current density $\mathbf{J}$ is given by a solution of the DD model this yields altogether a constrained optimization problem. This problem can be clearly tackled by an optimization approach, but only recently efforts were made to solve the design problem using mathematical sound optimization techniques [**13, 12, 29, 30, 43, 42, 44**]. In [**52**] *Lee et al.* present a finite–dimensional least-squares approach for adjusting the parameters of a semiconductor to fit a given, ideal IVC. Their work is purely numerical and has its focus on testing different approaches to solve the least-squares problem.

We describe in the following how one can apply the idea of the adjoints to this problem. For this purpose we introduce the state $x \stackrel{\text{def}}{=} (n,p,V)$ and an admissible set of controls $\mathcal{C} \subset H^1(\Omega)$ and rewrite the state equations (2.3) shortly as $e(x,C) = 0$. Due to the nonlinear structure of the equations we define the state space $X \stackrel{\text{def}}{=} x_D + X_0$, where $x_D \stackrel{\text{def}}{=} (n_D,p_D,V_D)$ denotes the boundary data introduced in (2.3) and $X_0 \stackrel{\text{def}}{=} \left(H^1_{0,\Gamma_D}(\Omega)\cap L^\infty(\Omega)\right)^3$, where we define

$$H^1_{0,\Gamma_D}(\Omega) \stackrel{\text{def}}{=} \left\{\phi \in H^1(\Omega) : \phi|_{\Gamma_D} = 0\right\},$$

as well as $Z \stackrel{\text{def}}{=} [H^1(\Omega)]^3$. Then, one can show that $e : X\times H^1(\Omega) \to Z^*$ is well–defined and infinitely often differentiable [**43**]. Now, the mathematically precise optimization problem reads

$$\min_{X\times\mathcal{C}} Q(n,p,V,C) \quad \text{such that} \quad e(n,p,V,C) = 0. \tag{2.6}$$

We restrict the set of admissible controls to

$$\mathcal{C} \stackrel{\text{def}}{=} \{C \in H^1(\Omega) : C = \bar{C} \text{ on } \Gamma_D\}. \tag{2.7}$$

This is necessary for the solvability of the state system and for the continuous dependence of the state on the control C, since the boundary data in (2.3) does depend on C. In fact, there are various results on the solvability of the state system (c.f. [**60, 54, 55**] and the references therein). For completeness we state the the following existence results, for which the proof can be found in [**61**].

PROPOSITION 2.1. *Assume sufficient regularity of the boundary and the data. Then for each* $C \in H^1(\Omega)$ *and all boundary data* (n_D,p_D,V_D) *with*

$$\frac{1}{K} \le n_D(\mathbf{x}), p_D(\mathbf{x}) \le K, \quad \mathbf{x}\in\Omega, \quad \textit{and } \|V_D\|_{L^\infty(\Omega)} \le K$$

for some $K \geq 1$, there exists a solution $(\mathbf{J}_n, \mathbf{J}_p, n, p, V) \in [L^2(\Omega)]^2 \times (H^1(\Omega) \cap L^\infty(\Omega))^3$ of system (2.3) *fulfilling*

$$\frac{1}{L} \leq n(\mathbf{x}), p(\mathbf{x}) \leq L, \quad \mathbf{x} \in \Omega, \quad \textit{and } \|V\|_{L^\infty(\Omega)} \leq L$$

for some constant $L = L(\Omega, K, \|C\|_{L^p(\Omega)}) \geq 1$, where the embedding $H^1(\Omega) \hookrightarrow L^p(\Omega)$ holds.

The idea of the proof to write down a fixed point mapping decoupling the equations and to use Schauder's fixed point theorem to get the existence of a fixed point. The compactness of the mapping is derived by energy estimates and Stampacchia's truncation method, which ensures the uniform bounds on the solution.

What makes this system special is that there exists in general no unique solution and this is even physically reasonable, since there are devices, like the thyristor, whose performance relies on the multiplicity of solutions. Nevertheless, one can ensure uniqueness near to the thermal equilibrium state, i.e. for small applied biasing voltages U. Clearly, this has also impact on the optimization problem. Especially, we cannot consider the reduced cost functional in each regime and also the linearized operator e_x is in general not boundedly invertible. But still one can proof the existence of a minimizer [**43**].

THEOREM 2.2. *The constrained minimization problem* (2.6) *admits at least one solution $(n^*, p^*, V^*, C^*) \in X \times \mathcal{C}$.*

The proof uses the standard techniques presented during this week. I.e. one extracts a convergent minimizing sequence using the coercivity of the cost functional, employs the bounds for the state system given in Proposition 2.1 to get convergent subsequences of the state variables and uses the weak lower semicontinuity of the cost functional.

Since the set given by the constraint is not convex, we can in general not expect the uniqueness of the minimizer. Here, one can in fact show analytically that for special choices of the reference doping $\bar{C}$ there exist at least two solutions and for other choices there is numerical evidence. This is due to the fact that the minimizer has the possibility to interchange the roles of the electron and the hole current densities (see Figure 2.2 and Figure 2.3). Clearly, this has also some impact on the construction and convergence of numerical schemes. Espescially, the choice of an appropriate starting point for iterative algorithms is then crucial.

2.2.1. *The First–order Optimality System.* In this section we want to discuss the first–order optimality system which is, as we already know, the basis for all optimization methods seeking at least a stationary point. Since we have a constrained optimization problem, we write the first–oder optimality system using the Lagrangian $\mathcal{L} : X \times \mathcal{C} \times Z \to \mathbb{R}$ associated to problem (2.6) defined by

$$\mathcal{L}(x, C, \xi) \stackrel{\text{def}}{=} Q(x, C) + \langle e(x, C), \xi \rangle_{Z^*, Z},$$

where $\xi \stackrel{\text{def}}{=} (\xi^n, \xi^p, \xi^V)$ denotes the adjoint variable. For the existence of a Lagrange multiplier associated to an optimal solution (x^*, C^*) of (2.6) it is sufficient that the operator $e'(x^*, C^*)$ is surjective.

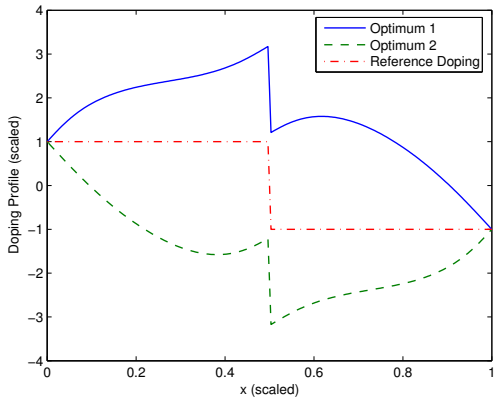


FIGURE 2.2. Optimized Doping Profiles for a Symmetric n–p–diode

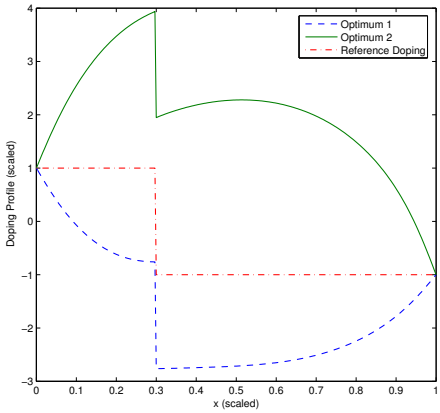


FIGURE 2.3. Optimized Doping Profiles for an Unsymmetric n–p–diode

Note the equivalence

$$e'(x,C)[(v,\tilde{C})] = g \quad \text{in } Z^* \quad \Leftrightarrow \quad e_x(x,C)[v] = g - e_C(x,C)[\tilde{C}] \quad \text{in } Z^*.$$

For the DD model this does in general not hold, but one can ensure the bounded invertibility of $e'(x^*,C^*)$ for small current densities [**55**]. This idea can be used to prove the unique existence of adjoint states [**43**].

THEOREM 2.3. *There exists a constant $j = j(\Omega,\lambda,U) > 0$ such that for each state $x \in X$ with*

$$\left\| \frac{\mathbf{J}_n^2}{n} \right\|_{L^\infty(\Omega)} + \left\| \frac{\mathbf{J}_p^2}{p} \right\|_{L^\infty(\Omega)} \leq j$$

there exists an adjoint state $\xi \in Z$ fulfilling $e_x^*(x,C)\xi = -Q_x(x,C)$.

Hence, at least for small current densities there exists a unique Lagrange multiplier ξ^* such that together with an optimal solution (x^*,C^*) it fulfills the first–order optimality system

$$\mathcal{L}'(x^*,C^*,\xi^*) = 0. \tag{2.8}$$

We can rewrite this equations in a more concise form:

$$\begin{aligned} e(x^*,C^*) &= 0 \quad \text{in } Z^*, \\ e_x^*(x^*,C^*)\xi^* + Q_x(x^*,C^*) &= 0 \quad \text{in } X^*, \\ e_C(x^*,C^*)\xi^* + Q_C(x^*,C^*) &= 0 \quad \text{in } \mathcal{C}^*. \end{aligned}$$

I.e., a critical point of the Lagrangian has to satisfy the state system (2.3), as well as the adjoint system. The derivation of this system is an easy exercise just using the techniques presented in the previous

lectures, yielding

$$\Delta\xi^n - \nabla V\,\nabla\xi^n = \xi^V, \tag{2.9a}$$

$$\Delta\xi^p + \nabla V\,\nabla\xi^p = -\xi^V, \tag{2.9b}$$

$$-\lambda^2\Delta\xi^V + \operatorname{div}(n\,\nabla\xi^n) - \operatorname{div}(p\,\nabla\xi^p) = 0, \tag{2.9c}$$

supplemented with the boundary data

$$\xi^{J_n} = \begin{cases} \int_\Gamma J_n\cdot\nu\;ds - I_n^*, & \text{on } \Gamma,\\ 0, & \text{on } \Gamma_D\setminus\Gamma, \end{cases} \tag{2.9d}$$

$$\xi^{J_p} = \begin{cases} \int_\Gamma J_p\cdot\nu\;ds - I_p^*, & \text{on } \Gamma,\\ 0, & \text{on } \Gamma_D\setminus\Gamma, \end{cases} \tag{2.9e}$$

$$\xi^V = 0, \text{ on } \Gamma_D, \tag{2.9f}$$

as well as

$$\xi^n\cdot\nu = \xi^p\cdot\nu = \nabla\xi^V\cdot\nu = 0 \quad \text{ on } \Gamma_N. \tag{2.9g}$$

Further we have the optimality condition

$$\gamma\Delta\left(C-\bar C\right) = \xi^V \quad \text{ in } \Omega, \tag{2.10a}$$

$$C = \bar C \quad \text{ on } \Gamma_D, \quad \nabla C\cdot\nu = \nabla\bar C\cdot\nu \quad \text{ on } \Gamma_N. \tag{2.10b}$$

2.3. Numerical Results. Finally, we want to discuss the behavior of two numerical methods applied to this optimization problem. The first adequate and easy to implement numerical method for the solution of (2.6) is the following gradient algorithm.

ALGORITHM 2.1.

(1) *Choose* $C_0\in\mathcal{C}$.
(2) *For* $k=1,2,\ldots$ *compute* $C_k = C_{k-1} - \alpha_k\hat Q'(C_{k-1})$

Here, $\hat Q(C)\overset{\text{def}}{=} Q(x(C),C)$ denotes the reduced cost functional, which can be introduced near to the thermal equilibrium state, and $\hat Q'(C)$ is the Riesz representative of its first variation. The evaluation of

$$\hat Q'(C) = Q_C(x,C) + e_C^*\xi$$

requires the solution of the nonlinear state system (2.3) for x as well as a solution of the linear adjoint system (2.9) for ξ and finally a linear solve of a Poisson problem to get the correct Riesz representative.

REMARK 2.2. *There exist various choices for the parameters* α_k *ensuring the convergence of this algorithm to a critical point, like the Armijo or the Goldstein rule. The overall numerical performance of this algorithm relies on an appropriate choice of the step–size rule for* α_k*, since these methods require in general consecutive evaluations of the cost functional requiring additional solves of the nonlinear state system* [**53**].

We apply Algorithm 2.1 for the optimal design of an unsymmetric n–p–diode (for the reference doping profile see Figure 2.4). We already learned that the cost functional employed so far might admit multiple minimizers. For this reason we study here a slightly different functional of the form

$$Q(n,p,V,C) = \frac{1}{2}\left|\int_\Gamma \mathbf{J}_n \cdot \nu \, ds - I_n^*\right|^2 + \frac{1}{2}\left|\int_\Gamma \mathbf{J}_p \cdot \nu \, ds - I_p^*\right|^2 + \frac{\gamma}{2}\int_\Omega \left|\nabla(C-\bar{C})\right|^2 \, dx.$$

This allows to adjust the electron and hole current separately. The computations were performed on a uniform grid with 1000 points and the scaled parameters were set to $\lambda^2 = 10^{-3}$, $\delta^2 = 10^{-2}$ and $U = 10$. For the parameter γ we chose $2 \cdot 10^{-2}$. The step–size α_k is computed by an exact one dimensional linesearch

$$\alpha_k = \operatorname{argmin}_\alpha \hat{Q}\left(C_{k-1} - \alpha \hat{Q}'(C_{k-1})\right).$$

The iteration terminates when the relative error $\left\|\hat{Q}'(C_k)\right\|_{H^1} / \left\|\hat{Q}'(C_0)\right\|_{H^1}$ is less than $5 \cdot 10^{-4}$.

In Figure 2.4 we present the optimized doping profiles for different choices of I_n^*, I_p^*, i.e. we are seeking an amplification of either the hole current ($I_n^* = \mathbf{J}_n^*, I_p^* = 1.5 \cdot \mathbf{J}_p^*$) or of the electron current ($I_n^* = 1.5 \cdot \mathbf{J}_n^*, I_p^* = \mathbf{J}_p^*$) or of both of them $I_n^* = 1.5 \cdot \mathbf{J}_n^*, I_p^* = 1.5 \cdot \mathbf{J}_p^*$) by 50%.

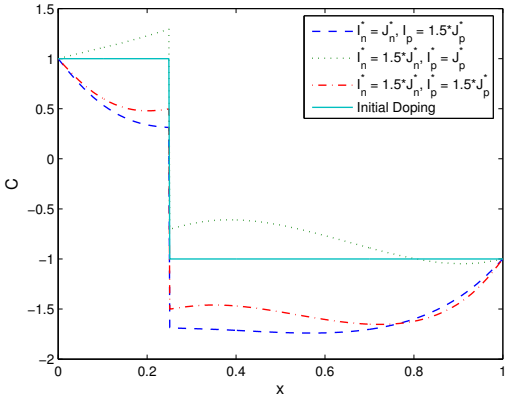


FIGURE 2.4. Optimized Doping Profiles

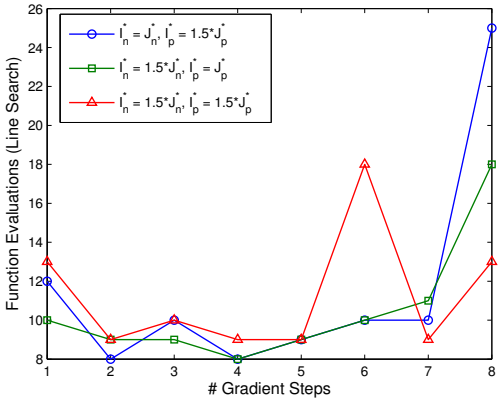


FIGURE 2.5. Function Evaluations for the Line Search

To get an impression of the overall performance of the method we also have to consider the nonlinear solves needed for the exact one dimensional linesearch. These are presented in Figure 2.5 and one realizes that this is indeed the numerically most expensive part.

Finally, we want to discuss the performance of the Newton algorithm presented in **???**. Again, we tried to achieve an increase of the electron and hole current by 50 % and studied the influence of the regularization parameter γ. The different resulting optimal doping profiles can be found in Figure 2.6. As expected we get larger deviations from $\bar{C}$ for decreasing γ, which on the other hand also allows for a better reduction of the observation as can be seen in Figure 2.7. For all three cases we already get a significant reduction after two steps and the algorithm terminates rather quickly. Only for the smallest

value of γ we need two more iterations to meet the stopping criterion, which can be explained by a loss of convexity or, equivalently, a weaker definiteness of the Hessean.

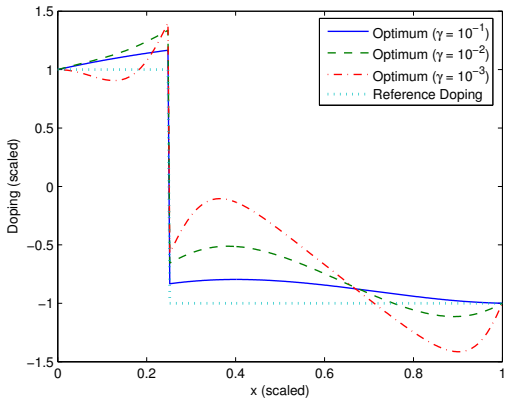


FIGURE 2.6. Dependence of the optimum on γ

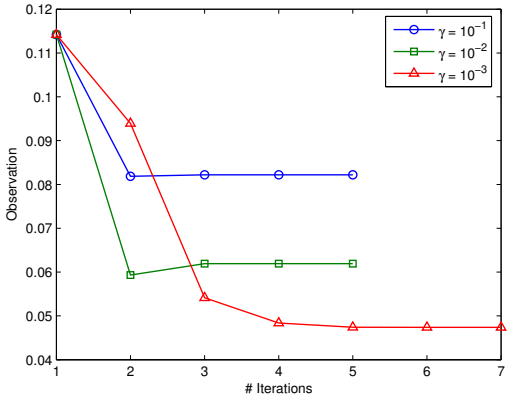


FIGURE 2.7. Dependence of the observation on γ

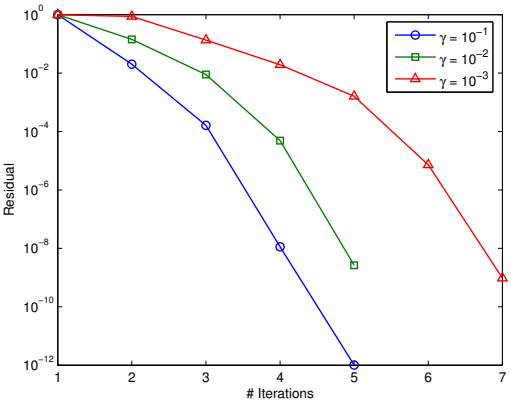


FIGURE 2.8. Dependence of the residual on γ

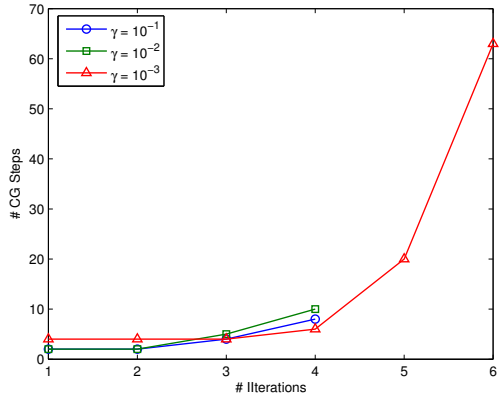


FIGURE 2.9. Dependence of the CG iteration on γ

The conjugate gradient algorithm in the inner loop was terminated when the norm of the gradient became sufficiently small; to be more precise, in the j-th conjugate gradient step for the computation of the update in Newton step k we stop if the residual r_k^j satisfies

$$\frac{\|r_j^k\|}{\left\|\hat{Q}'(C^0)\right\|} \leq \min\left\{\left(\frac{\|\hat{Q}'(C^k)\|}{\|\hat{Q}'(C^0)\|}\right)^q, p\,\frac{\|\hat{Q}'(C^k)\|}{\|\hat{Q}'(C^0)\|}\right\} \quad \text{or} \quad j \geq 100. \tag{2.11}$$

Note, that q determines the order of the outer Newton algorithm, such that p should be chosen in the open interval $(1, 2)$. The value of p is important for the first step of Newton's method, as for $k = 0$ the norm quotients are all 1; for later steps, the influence of q becomes increasingly dominant.

To get deeper insight into the convergence behavior of the algorithm, we present in Figure 2.8 the norm of the residual during the iteration for different values of γ. Here, we used $q = 2$ to get the desired quadratic convergence behavior. Again, one realizes that the convergence deteriorates with decreasing γ. Since the overall numerical effort is spent in the inner loop, we show the number of conjugate gradient steps in Figure 2.9. Here, one realizes the drastic influence of the regularization parameter.

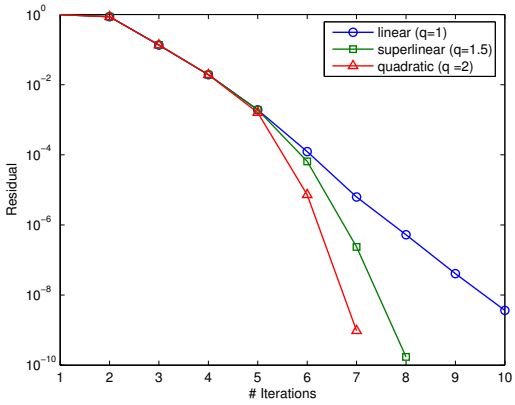


FIGURE 2.10. Dependence of the residual on q

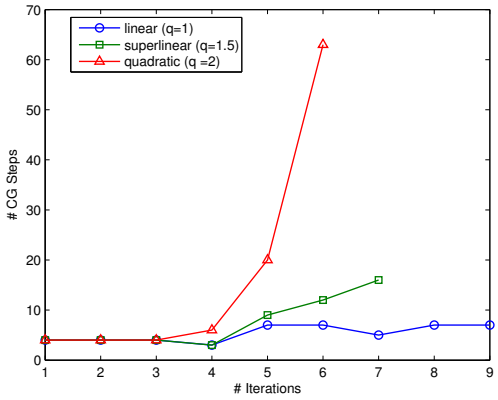


FIGURE 2.11. Dependence of the CG iteration on q

The next numerical test was devoted to the stopping criterion of the inner iteration and the influence of the exponent q. In Figure 2.10 the decrease of the residual is depicted for different values of $q = 1, 1.5,$ or 2. As predicted by the general theory one gets linear, superlinear and quadratic convergence. Note, that for all three cases we have a linear convergence behavior at the beginning of the iteration due to the globalization of the Newton algorithm. Clearly, the parameter q strongly influences the number of conjugate gradient steps, which can be seen from Figure 2.11. While in the linear case ($q = 1$) we have an almost constant amount of CG steps in each each iteration, we get, as expected, a drastic increase towards the end of the iteration for the quadratic case ($q = 2$). Hence, the overall numerical effort in terms of CG steps is despite of the quadratic convergence much larger compared to the relaxed stopping criterion, which only yields linear convergence!

3. Optimal Control of Glass Cooling

Although glass manufacturing is a very old industry, one has to be aware that it is nowadays technically rather advanced. This is stemming from the strong need for high quality glass products, like lenses for laser optics or mirrors for space telescopes. But clearly one also wants to influence the production process for lower quality fabriques, like monitors or car windows. There are many stages in the production process where optimal control techniques can be used. We will focus here on the stage where a hot melt of glass is cooled in a controlled environment, e.g. a furnace. During cooling, large temperature differences i.e. large gradients have to be avoided since they lead to thermal stress in the material. This may cause cracks or, in the case of high quality glass, affect the quality of the resulting product or device. Hence, the process has to be managed in such a way that temperature gradients are sufficiently small. Another related question concerns chemical reactions during the cooling process, which have to be activated and triggered. Here, one again wants to avoid spatial gradients since these reactions have to take place homogeneously in the glass. We will see that these two different question can be mathematically embedded into the same optimal control problem. The presentation is again

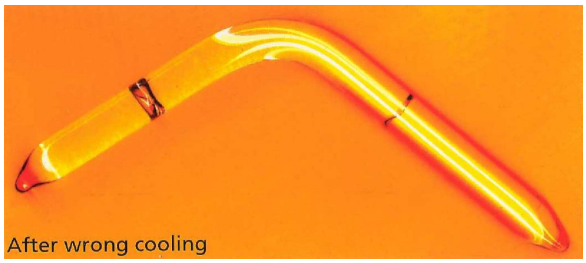


FIGURE 3.1. This happens after wrong cooling!

done in three steps. First, we discuss the equations which can be used for the simulation of the cooling process. Then, we state and discuss the optimal control problem and, finally, we present numerical results.

3.1. Modeling. The modeling of glass cooling has to take into account that this process involves very high temperatures up to 1500 K. In this temperature range heat transfer will be dominated by radiation and not by diffusion anymore. Hence, we have first to understand how radiation can be modelled.

3.1.1. *Radiation.* Thermal radiation can be viewed as electro–magnetic waves or, alternatively, as photons. It is characterized by its speed of propagation c, wavelength λ and frequency ν, which are related by $c = \lambda \cdot \nu$. The most important difference to heat conduction and convection is that it is a long–range, non–local phenomenon in contrast to the local, microscopic diffusion effect.

REMARK 3.1. *Note, that the magnitude of conduction and convection is linear in the temperature T, whereas radiation depends essentially on the fourth power of T, which shows that this effect gets increasingly important for higher temperatures.*

In general, engineers are only interested in the energy of the radiative field and they describe it using the radiative intensity $I = I(x, t, \omega, \nu)$, which depends on the position x, the time t, the angular direction ω and on the frequency ν. To derive an equation for I, we consider a small portion Δx of a ray in direction ω.

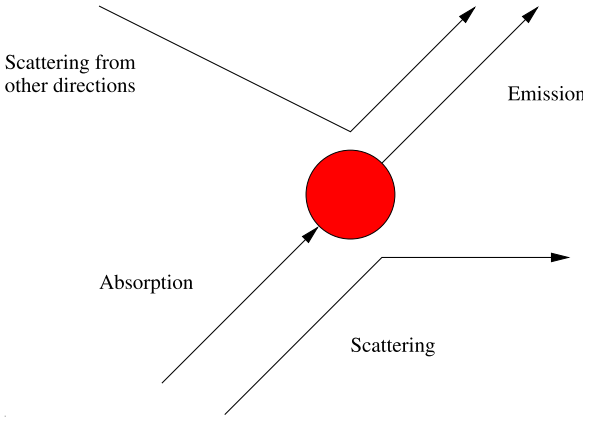


FIGURE 3.2. Radiative Effects

There, one looses energy due to absorption $-\kappa I \Delta x$, where κ is the absorption coefficient of the material, which might also depend on T and ν. Further, one gains energy due to emission $+\kappa B \Delta x$, where

$$B(T, \nu) = n_G^2 \frac{2h\nu^3}{c^2} \Big(e^{\frac{h\nu}{kT}} - 1 \Big)^{-1}$$

is Planck's function for black body radiation. Another source for energy loss is scattering $-\sigma I \Delta x$, where σ is the scattering constant of the material. But one can also gain energy due to back scattering, i.e. one has to collect the distributions from all incoming directions $+\frac{\sigma}{4\pi} \int_\omega I(\omega')\ d\omega'\ \Delta x$. Now, we can write down the balance equation for the radiative intensity

$$I(x + c\omega\Delta t, \omega, t + \Delta t) - I(x, \omega, t) = \Big(-\kappa I + \kappa B - \sigma I + \frac{\sigma}{4\pi} \int_\omega I(\omega')\ d\omega' \Big) \Delta x$$

Going to the limit $\Delta t \to 0, \Delta x = c\Delta t$ yields

$$\frac{1}{c} \partial_t I + \omega \cdot \nabla I + (\kappa + \sigma) I = \frac{\sigma}{4\pi} \int_\omega I(\omega')\ d\omega' + \kappa B. \tag{3.1}$$

This equation holds for all times $t \in \mathbb{R}^+$, all spatial points $x \in \Omega$, all angles $\omega \in S^2$ and all frequencies $\nu \in [\nu_0, \infty)$! To get an impression of the computational effort let us assume that we use a discretization with

$$60 \text{ angles} \times 10 \text{ frequency bands} \times 8000 \text{ spatial points} \times 100 \text{ time steps}.$$

This yields 500 millions discrete variables! Indeed, this leads to a large scale optimization problem. To be honest, we will not even dare to use this equation directly, but instead we will use techniques from asymptotic analysis to derive a numerically tractable model. Finally, we pose some physically

reasonable assumptions which will significantly simplify the presentation. Note that c is large and hence we will drop the time derivative. Further, we assume that no scattering occurs in the glass, i.e. $\sigma = 0$.

3.1.2. *SP_N–approximations.* Using a diffusion scaling we introduce the optical thickness of the material as a small parameter

$$\varepsilon = \frac{1}{\kappa_{\text{ref}} x_{\text{ref}}} \approx \frac{\text{mean free path}}{\text{reference length}}.$$

Then, the remaining scaled equation reads

$$\varepsilon \omega \cdot \nabla I = \kappa(B - I)$$

Now, the idea is to invert the transport operator

$$\Big(1 + \frac{\varepsilon}{\kappa}\omega \cdot \nabla\Big) I = B$$

formally using the Neumann series. Then it holds for $\rho := \int_\omega I \, d\omega$ in the limit $\varepsilon \to 0$ the asymptotic expansion

$$4\pi B = \Big[1 - \frac{\varepsilon^2}{3\kappa^2}\Delta - \frac{4\varepsilon^4}{45\kappa^4}\Delta^2 - \frac{44^\varepsilon 6}{945\kappa^6}\Delta^3\Big]\rho + O(\varepsilon^8)$$

This yields the SP_N-approximations [**51**] of order $\mathcal{O}(\varepsilon^{2N})$.

In the following we will only employ the SP_1–approximation. Since the radiative intensity depends crucially on the temperature, we need to couple our approximate equation with the heat equation which yields the overall system

$$\partial_t T = k\,\Delta T + \frac{1}{3\kappa}\Delta\rho, \tag{3.2a}$$

$$-\varepsilon^2 \frac{1}{3\kappa}\Delta\rho + \kappa\rho = 4\pi\kappa\, a\, T^4 \tag{3.2b}$$

This has to be supplmented with appropriate initial conditions $T(x, 0) = T_0(x)$ and boundary data

$$\frac{h}{\varepsilon k} T + n \cdot \nabla T = \frac{h}{\varepsilon k} u, \tag{3.2c}$$

$$\frac{3\kappa}{2\varepsilon}\rho + n \cdot \nabla\rho = \frac{3\kappa}{2\varepsilon} 4\pi\, a\, u^4. \tag{3.2d}$$

Here, we assume that we have heat loss over the boundary only due to Newton's cooling law, where h is the heat transfer coefficient, and that we have semi–transparent boundary data for the mean radiative intensity ρ. Here, u denotes the ambient temperature which will act in the following as the control variable.

This leads alltogether to an optimal boundary control problem for an parabolic/elliptic system, which can be treated numerically with standard finite element techniques.

REMARK 3.2. *Reasonable regularity assumptions on the data ensure the existence of a unique solution to system* (3.2).

3.2. Optimization. We intend to minimize cost functionals of tracking type having the form

$$J(T,u) = \frac{1}{2}\left\|T - T_d\right\|^2_{L^2(0,1;L^2(\Omega))} + \frac{\delta}{2}\left\|u - u_d\right\|^2_{H^1(0,1;\mathbb{R})}, \tag{3.3}$$

Here, $T_d = T_d(t,x)$ is a specified temperature profile, which is typically given by engineers. In glass manufacturing processes, T_d is used to control chemical reactions in the glass, especially their activation energy and the reaction time. For the quality of the glass it is essential that these reactions happen spatially homogeneously, such that we will later on require that T_d is independent of x. The control variable u, which is considered to be space–independent, enters the cost functional as a penalizing and regularizingterm, where additionally a known cooling curve u_d can be prescribed. The parameter δ allows to adjust the effective heating costs of the cooling process. The main subject is now the study of the following boundary control problem

$$\begin{aligned} &\min\ J(T,\rho,u) \text{ w.r.t.}(T,\rho,u),\\ &\text{subject to the } SP_1\text{–system (3.2).} \end{aligned} \tag{3.4}$$

For notational convenience we define

$$\begin{gathered} Q \overset{\text{def}}{=} (0,1)\times\Omega, \quad \Sigma \overset{\text{def}}{=} (0,1)\times\partial\Omega,\\ V \overset{\text{def}}{=} L^2(0,1;H^1(\Omega)), \quad U \overset{\text{def}}{=} H^1(0,1;\mathbb{R}),\\ W \overset{\text{def}}{=} \left\{\phi\in V : \phi_t \in V^*\right\}, \quad X \overset{\text{def}}{=} W\times V, \quad Z \overset{\text{def}}{=} V\times V\times L^2(\Omega). \end{gathered}$$

Then, we define $X_\infty \overset{\text{def}}{=} X\cap[L^\infty(Q)]^2$ as the space of states $x \overset{\text{def}}{=} (T,\rho)$ and U is the space of controls. Finally, we set $\alpha = \frac{h}{\varepsilon k}$, $\gamma = \frac{3\kappa}{2\varepsilon}$.

We define the state/control pair $(x,u)\in X_\infty\times U$ and the nonlinear operator $e \overset{\text{def}}{=} (e_1,e_2,e_3) : X_\infty\times U\to Z^*$ via

$$\begin{aligned} \langle e_1(x,u),\phi\rangle_{V^*,V} \overset{\text{def}}{=}\ & \langle\partial_t T,\phi\rangle_{V^*,V} + k\,(\nabla T,\nabla\phi)_{L^2(Q)} + \frac{1}{3\kappa}\,(\nabla\rho,\nabla\phi)_{L^2(Q)}\\ &+ k\alpha(T-u,\phi)_{L^2(\Sigma)} + \frac{1}{3\kappa}\gamma(\rho - 4\pi a u^4,\phi)_{L^2(\Sigma)} \end{aligned} \tag{3.5a}$$

and

$$\langle e_2(x,u),\phi\rangle_{V^*,V} \overset{\text{def}}{=} \frac{\varepsilon^2}{3\kappa}\,(\nabla\rho,\nabla\phi)_{L^2(Q)} + \kappa(\rho - 4\pi\kappa a\,T^4,\phi)_{L^2(Q)} + \frac{\varepsilon^2}{3\kappa}\gamma(\rho - 4\pi a u^4,\phi)_{L^2(\Sigma)} \tag{3.5b}$$

for all $\phi\in V$. Further, we define $e_3(x,u) \overset{\text{def}}{=} T(0) - T_0$.

REMARK 3.3. *Note, that for $d\le 2$ it is in fact possible to use X itself as the state space, but for $d=3$ we cannot guarantee that e_2 is well defined due to the fourth–order nonlinearity in T.*

Then the minimization problem (3.4) can be shortly written as

$$\begin{aligned} &\min J(x,u) \text{ over } (x,u)\in X_\infty\times U,\\ &\text{subject to } e(x,u) = 0 \text{ in } Z^*. \end{aligned} \tag{3.6}$$

In fact, one can show the existence of a minimizer.

THEOREM 3.1. *There exists a minimizer* $(x^*, u^*) \in X_\infty \times U$ *of the constrained minimization problem* (3.7).

The proof uses the standard techniques presented during this week. I.e. one extracts a convergent minimizing sequence using the coercivity of the cost functional, uses the bounds for the state system to get convergent subsequences of the state variables and uses the weak lower semicontinuity of the cost functional.

3.2.1. *Derivatives.* In the following we provide the derivative information, which is necessary for the application of the Newton algorithm.

Owing to the fact that the system (3.2) is uniquely solvable [**63**], we may reformulate the minimization problem (3.6) introducing the *reduced cost functional* $\hat{J}$ as

$$\begin{aligned} &\text{minimize} \quad \hat{J}(u) \stackrel{\text{def}}{=} J(x(u), u) \quad \text{over} \quad u \in U \\ &\text{where} \quad x(u) \in X \quad \text{satisfies} \quad e(x(u), u) = 0. \end{aligned} \tag{3.7}$$

The numerical realization of Newton's method relies on derivative information on J and e, or $\hat{J}$ respectively. Formally, these can derived as follows: First, the implicit function theorem leads to the following derivative of x at u in a direction δu:

$$x'(u)\delta u = -e_x^{-1}(x(u), u)e_u(x(u), u)\delta u.$$

Using the chain rule one obtains

$$\left\langle \hat{J}'(u), \delta u \right\rangle = \left\langle J_u(x(u), u) - e_u^*(x(u), u)e_x^{-*}(x(u), u)J_x(x(u), u), \delta u \right\rangle.$$

Here, $e_x^*(x, u)\xi$ denotes the adjoint of the linearization of e at (x, u) in the direction ξ. We define the adjoint variable $\xi = (\xi_T, \xi_\rho, \xi_{T0})$ by

$$\xi = -e_x^{-*}(x(u), u)J_x(x(u), u) \in Z.$$

Assuming enough regularity of the solution one gets the Riesz representative of the derivative

$$\hat{J}'(u) = J_u(x(u), u) + e_u^*(x(u), u)\xi. \tag{3.8}$$

EXAMPLE 3.2. *In case of the cost functional* (3.3)*, the adjoint variable can be characterized as the variational solution of*

$$-\partial_t \xi_T = k\,\Delta\xi_T + 16\pi a\kappa\, T^3\,\xi_\rho - (T - T_d), \tag{3.9a}$$

$$-\frac{\varepsilon^2}{3\kappa}\,\Delta\xi_\rho + \kappa\,\xi_\rho = \frac{1}{3\kappa}\,\Delta\xi_T, \quad \textit{in } Q \tag{3.9b}$$

with boundary conditions

$$k(n \cdot \nabla\xi_T + \alpha\,\xi_T) = 0, \tag{3.9c}$$

$$n \cdot \nabla\xi_T + \gamma\,\xi_T + \varepsilon^2(n \cdot \nabla\xi_\rho + \gamma\,\xi_\rho) = 0, \quad \textit{on } \Sigma \tag{3.9d}$$

and terminal condition

$$\xi_T(1) = 0 \quad \textit{in } \Omega. \tag{3.9e}$$

Introducing the Lagrangian $\mathcal{L} : X \times U \times Z \to \mathbb{R}$ associated to (3.6) defined by

$$\mathcal{L}(x, u, \xi) \stackrel{\text{def}}{=} J(x, u) + \langle e(x, u), \xi \rangle .$$

we know that there exists a critical point of the Lagrangian. In fact, for an optimal solution there exists a unique Lagrange multiplier [**63**].

THEOREM 3.3. *Let $(x^*, u^*) \in X \times U$ denote an optimal solution. Then there exists a unique Lagrange multiplier $\xi^* \in Z^*$ such that the triple (x^*, u^*, ξ^*) satisfies*

$$\mathcal{L}'(x^*, u^*, \xi^*) = 0 \quad \textit{in } X^* \times U^* \times Z^*.$$

Let $(x^*, u^*) \in X \times U$ denote an optimal solution. Then the second derivative of the Lagrangian is formally given by

$$\mathcal{L}''(x^*, u^*, \xi^*) = \begin{pmatrix} J_{xx}(x^*, u^*) + \langle e_{xx}(x^*, u^*)(\cdot, \cdot), \xi^* \rangle & 0 & e_x^*(x^*, u^*) \\ 0 & J_{uu}(x^*, u^*) + \langle e_{uu}(x^*, u^*)(\cdot, \cdot), \xi^* \rangle & e_u^*(x^*, u^*) \\ e_x(x^*, u^*) & e_u(x^*, u^*) & 0 \end{pmatrix}.$$

Defining the operator

$$\mathcal{T}(x, u) \stackrel{\text{def}}{=} \begin{pmatrix} -e_x^{-1}(x, u)\, e_u(x, u) \\ id_U \end{pmatrix}$$

we can write the reduced Hessean as

$$\hat{J}''(u) \stackrel{\text{def}}{=} \mathcal{T}^*(x, u)\mathcal{L}_{yy}(x, u, \xi)\mathcal{T}(x, u), \tag{3.10}$$

where $y \stackrel{\text{def}}{=} (x, u)$, i.e. $\mathcal{L}_{yy}$ is the upper left 2×2–block of $\mathcal{L}''$.

3.2.2. *Newton's Method.* In this section we describe the second order optimization algorithm, i.e. we apply Newton's method for the computation of an optimal control for the reduced cost functional. The algorithm reads formally

ALGORITHM 3.1.

(1) *Choose u_0 in a neighborhood of u^*.*
(2) *For $k = 0, 1, 2, \ldots$*
   (a) *Solve $\hat{J}''(u_k)\delta u_k = -\hat{J}'(u_k)$,*
   (b) *Set $u_{k+1} = u_k + \delta u_k$.*

REMARK 3.4. *The solution of the system in step (ii.a) is done iteratively by using a conjugate gradient algorithm embedded inside the Newton algorithm, as the computation of a discretization of the Hessean would require a significant numerical effort, while a conjugate gradient based approach leads to the same result with a fraction of the demands on memory and computation time. The conjugate*

gradient algorithm only requires the applications of the Hessean on a sequence of direction vectors δu to be computed, so that no (direct) solution of the large system in (ii.a) is required.

ALGORITHM 3.2.

(1) *Choose u_0 in a neighborhood of u^*.*
(2) *For $k = 0, 1, 2, \ldots$*
   (a) *Evaluate $\hat{J}'(u_k)$ and set $\delta u_k^j = 0$*
   (b) *For $j = 0, 1, 2, \ldots$ do until convergence*
      (i) *Evaluate $q_k^j = \hat{J}''(u_k)\delta u_k^j$*
      (ii) *Compute an approximation δu_k^{j+1} for δu_k, e.g. by a cg–step*
   (c) *Set $u_{k+1} = u_k + \delta u_k$*

Each application of the reduced Hessian $\hat{J}''(u_k)$ during the j-th cg–step requires two linear solves, in detail

$$v_k^j = e_x^{-1}(x_k, u_k)e_u(x_k, u_k)\delta u_k^j$$

and

$$w_k^j = e_x^{-*}(x_k, u_k)\left\{J_{xx}(x_k, u_k)(v_k^j, \cdot) + \left\langle e_{xx}(x_k, u_k)(v_k^j, \cdot), \xi_k\right\rangle_{Z^*,Z}\right\}.$$

EXAMPLE 3.4. *Especially, for the cost functional* (3.3) *one has to apply successively the following steps*

(1) *Solve the linearized state system (see system 3.2)*

$$\partial_t v_T = k\Delta v_T + \frac{1}{3\kappa}\Delta v_\rho \tag{3.11a}$$

$$-\frac{\varepsilon^2}{3\kappa}\Delta v_\rho + \kappa v_\rho = 16\pi\kappa a T_k^3\, v_T \tag{3.11b}$$

with boundary conditions

$$n \cdot \nabla v_T + \alpha v_T = \alpha\, \delta u_k^j \tag{3.11c}$$

$$n \cdot \nabla v_\rho + \gamma v_\rho = \gamma 16\pi a u_k^3 \delta u_k^j \tag{3.11d}$$

and initial condition

$$v_T(0) = 0 \tag{3.11e}$$

for $v_k^j \stackrel{\text{def}}{=} (v_T, v_\rho) \in X$, where $x_k = (T_k, \rho_k)$.
(2) *Evaluate*

$$J_{xx}(x_k, u_k)(v_k^j, \cdot) + \left\langle e_{xx}(x_k, u_k)(v_k^j, \cdot), \xi_k\right\rangle = v_T + 48\pi\kappa a T_k^2 v_T \xi_{T,k}.$$

(3) *Solve the linearized adjoint system (see system 3.9)*

$$-\partial_t w_T = k\Delta w_T + 16\pi\kappa a T_k^3 w_\rho + v_T - 48\pi\kappa a T_k^2 v_T \xi_{T,k} \tag{3.12a}$$

$$-\frac{\varepsilon^2}{3\kappa}\Delta w_\rho + \kappa w_\rho = \frac{1}{3\kappa}\Delta w_T \tag{3.12b}$$

with boundary conditions

$$k(n\cdot\nabla w_T + \alpha w_T) = 0 \tag{3.12c}$$

$$\varepsilon^2(n\cdot\nabla w_\rho + \gamma w_\rho) + n\cdot\nabla w_T + \gamma w_T = 0 \tag{3.12d}$$

and terminal condition

$$w_T(1) = 0 \tag{3.12e}$$

for $w_k^j \stackrel{\text{def}}{=} (w_T, w_\rho) \in X$.

(4) *Set*

$$q_k^j(t) = \frac{1}{|\partial\Omega|}\int_{\partial\Omega} k\alpha w_T + \frac{\gamma 16\pi a}{3\kappa}u^2(u(w_T+\varepsilon^2 w_\rho) - 3\,\delta u_k^j(\xi_T + \varepsilon^2\xi_\rho))\,ds + \frac{\delta}{|\partial\Omega|}\int_{\partial\Omega}\delta u_k^j + \partial_{tt}\delta u_k^j\,ds.$$

3.3. Numerical Results. The spatial discretization of the PDEs is based on linear finite elements. We use a non-uniform grid with an increasing point density towards the boundary of the medium, consisting of 109 points. The temporal discretization uses a uniform grid consisting of 180 points for the temperature–tracking problem. We employ the implicit backward Euler method to compute the state (T, ρ). The adjoint systems are discretized using a modified implicit Euler backward method taking into account the symmetry of the discrete reduced Hessean [**33**].

The conjugate gradient algorithm was terminated when the norm of the gradient became sufficiently small; to be more precise, in the j-th conjugate gradient step for the computation of the update in Newton step k we stop if the residual r_k^j satisfies

$$\frac{\|r_j^k\|}{\|J'(u^0)\|} \le \min\left\{\left(\frac{\|\hat{J}'(u^k)\|}{\|\hat{J}'(u^0)\|}\right)^p, q\,\frac{\|\hat{J}'(u^k)\|}{\|\hat{J}'(u^0)\|}\right\} \quad \text{or} \quad j \ge 100. \tag{3.13}$$

Note, that p determines the order of the outer Newton algorithm, such that p should be chosen in the open interval $(1, 2)$. The value of q is important for the first step of Newton's method, as for $k = 0$ the norm quotients are all 1; for later steps, the influence of p becomes increasingly dominant. In our numerical experiments, $p = 1.5$ and $q = 0.1$ proved to be a suitable choice.

REMARK 3.5. *In the Newton algorithm, one might use*

$$J_{uu}(u) = \delta(I - \partial_{tt})$$

as a preconditioning operator for the Newton system (ii.a).

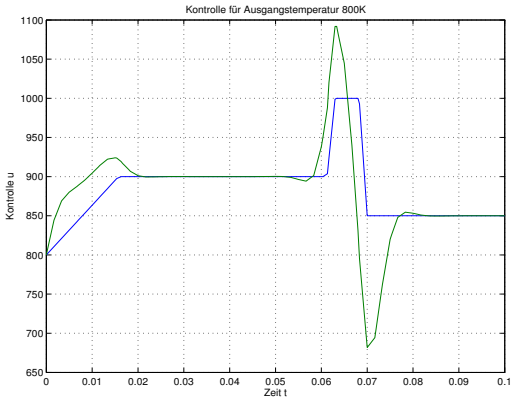


FIGURE 3.3. Unoptimized (blue) and optimized (green) cooling profile.

Now we present numerical results underlining the feasibility of our approach. For a given (time dependent) temperature profile T_d we compute an optimal u such that the temperature of the glass follows the desired profile T_d as good as possible. Such profiles are of great importance in glass manufacturing in order to control at which time, at which place and for how long certain chemical reactions take place, which is essential for the quality of the glass. Intervals of constant temperature allow for lengthy reactions in a controlled manner; short peaks of high temperature trigger reactions that have a high activation energy. Especially, it can be desirable to attain a spatial constant temperature, which is in contradiction to the boundary layers of the temperature due the radiative heat loss over the boundary.

The blue line in Figure 3.3 describes the desired temperature profile $T_d(t)$ which shall be attained homogeneously in space. From the engineering point of view it is an educated guess to use the same profile for the boundary control. Clearly, this leads to deviations which can be seen in the left graphic of Figure 3.4. Our optimal control approach leads now the the green line in Figure 3.3, which yields in turn the improved temperature differences on the right in Figure 3.4. One realizes a significant improvement although we have still a large peak. But note that we require a very sharp jump in the temperature. Due to diffusive part of the equations it is almost impossible to resolve such fast change in the cooling. Finally, let us discuss the influence of the penalizing parameter δ on the convergence of the iterative Newton method. In Table 3.1 and Table 3.2 we compare the number of Newton iterations, the evolution of the cost functional and the residual as well as number of cg iteration in each Newton step. As expected we get e better performance for the "more convex" problem.

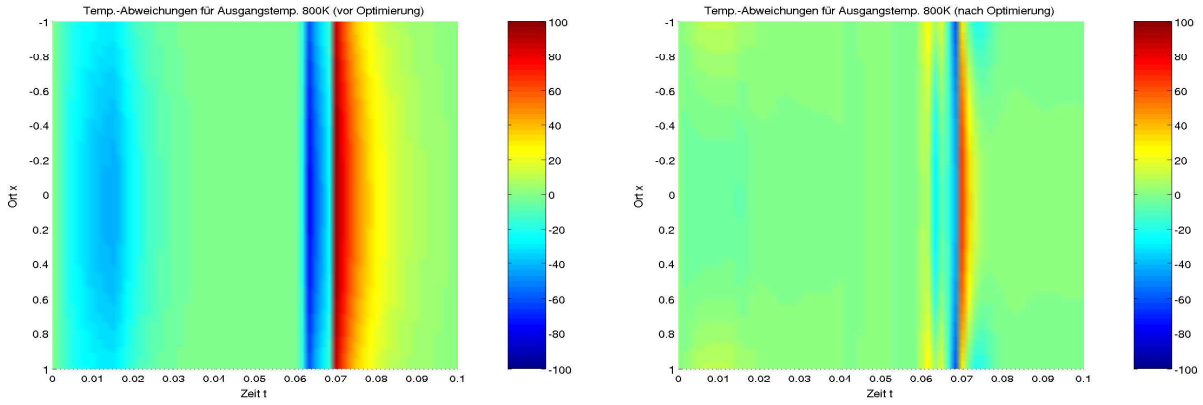


FIGURE 3.4. Temperature differences for the uncontrolled (left) and controlled (right) state.

k	$J(u_k)$	$\|\hat{J}'(u_{k+1})\|_2$	#cg
1	224.7359	$1.605777 \cdot 10^{+01}$	26
2	184.5375	$1.306437 \cdot 10^{+01}$	16
3	142.9351	$1.038065 \cdot 10^{+01}$	14
4	112.5493	$7.985859 \cdot 10^{+00}$	13
5	90.52294	$5.861017 \cdot 10^{+00}$	13
6	74.95062	$3.989118 \cdot 10^{+00}$	14
7	64.45030	$2.357674 \cdot 10^{+00}$	14
8	57.97925	$9.541926 \cdot 10^{-01}$	16
9	54.70802	$4.762934 \cdot 10^{-02}$	17
10	53.96191	$5.101231 \cdot 10^{-04}$	17
11	53.96017	$2.086531 \cdot 10^{-06}$	17
12	53.96017	$1.590937 \cdot 10^{-09}$	25

TABLE 3.1. Convergence statistics for $\delta = 3.5 \cdot 10^{-7}$

k	$J(u_k)$	$\|\hat{J}'(u_{k+1})\|_2$	#cg
1	337.5395	$3.912697 \cdot 10^{+01}$	29
2	254.1703	$2.918320 \cdot 10^{+01}$	27
3	193.4364	$1.978074 \cdot 10^{+01}$	27
4	151.9171	$1.094969 \cdot 10^{+01}$	28
5	126.9592	$2.751367 \cdot 10^{+00}$	29
6	116.2621	$3.163388 \cdot 10^{-02}$	29
7	115.4742	$2.184202 \cdot 10^{-04}$	31
8	115.4741	$4.352735 \cdot 10^{-07}$	37
9	115.4741	$4.542256 \cdot 10^{-09}$	28

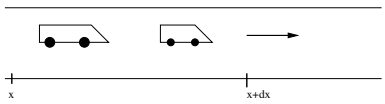
TABLE 3.2. Convergence statistics for $\delta = 3.5 \cdot 10^{-6}$.

4. Optimal Control of Traffic Networks

In this section we study a totally different application, which attained considerable attention during the last years. Everybody using a car knows how annoying traffic jams are. They are in fact not only annoying, but also quite expensive, which explains that there is a strong need for strategies influencing the traffic on our streets in such a way that no traffic jams occur. In cities this can be done by traffic lights and on highways one might use adjustable speed limits. Finally, the long term goal is to influence the individual navigation system of each driver. For this reason many engineers and mathematicians first concentrated on the development of appropriate models for the simulation of traffic flow on highways. There exists a whole hierarchy of models for traffic flow on unidirectional roads, ranging from microscopic over kinetic to macroscopic models [**7**]. Here, we will concentrate on a well-known macroscopic PDE model and its extension to road networks. Since these models need to be able to describe traffic jams, which can be interpreted as a front moving backwards along the road, the reader might already guess that we are heading now for hyperbolic conservation laws. But due to the network structure one has to take special care of the coupling conditions which govern the flow along the junctions and the optimal control problem we want to consider is then related to the traffic management in such large scale networks, which is numerically rather challenging. The forthcoming presentation follows the work of [**35, 36**].

Again we will proceed in several steps. First we derive the model equations for one single road and then we discuss the extension to traffic networks. In the following we will set up the optimal control problem and show an approach for its numerical solution.

4.1. Modeling. We start our modeling by considering a single road. Assume that it has just one lane and all cars are driving in the same direction.



Clearly, this yields a onedimensional problem. Let $\rho(x,t)$ be the density of cars and $v(x,t)$ their mean velocity. The number of cars in the road section $[x_1, x_2 = x_1 + \Delta x]$ at time t is then given by

$$\int_{x_1}^{x_2} \rho(x,t)\, dx \approx \rho \Delta x$$

and the flux of cars in a point x during the time period $[t_1, t_2 = t_1 + \Delta t]$ is given by

$$\int_{t_1}^{t_2} \rho(x,t)\, v(x,t)\, dt \approx \rho v \Delta t.$$

Since no cars are getting lost or are entering this road, i.e. we have no highway entrance, we have the *conservation of cars* which directly yields

> The change of cars on this road section is equal to the difference of the flux of cars leaving or entering over the boundary.

Mathematically, this reads

$$\int_{x_1}^{x_2} \rho(x,t_2)\,dx - \int_{x_1}^{x_2} \rho(x,t_1)\,dx = \int_{t_1}^{t_2} \rho(x_1,t)v(x_1,t)\,dt - \int_{t_1}^{t_2} \rho(x_2,t)v(x_2,t)\,dt.$$

To derive a differential equation we use the main theorem of calculus which gives

$$\rho(x,t_2) - \rho(x,t_1) = \int_{t_1}^{t_2} \frac{\partial}{\partial t}\rho(x,t)\,dt,$$
$$\rho(x_2,t)v(x_2,t) - \rho(x_1,t)v(x_1,t) = \int_{x_1}^{x_2} \frac{\partial}{\partial x}(\rho(x,t)\,v(x,t))\,dx.$$

Now interchanging the order of integration we get

$$\int_{t_1}^{t_2}\int_{x_1}^{x_2} \Big(\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}(\rho v)\Big)\,dx\,dt = 0 \quad \forall t_1,t_2,x_1,x_2.$$

Since this has to hold for each road section and for each time period we can use the variational lemma and get the so-called **continuity equation**

$$\boxed{\partial_t \rho + \partial_x(\rho v) = 0.}$$

To get a closed equation for the density ρ one needs an additional constitutive relation between v and ρ. Now we assume that, the more dense the traffic is, the slower the cars will drive. This can be modelled using the following function

$$v = v(\rho) = v_{\max}\Big(1 - \frac{\rho}{\rho_{\max}}\Big),$$

which gives rise to famous *Lighthill-Whitham model*. Here, $v_{\max}$ is the maximal velocity and $\rho_{\max}$ describes the maximal capacity of the road. Defining the flux function

$$f(\rho) = v_{\max}\,\rho\Big(1 - \frac{\rho}{\rho_{\max}}\Big)$$

we get the equation

$$\partial_t \rho + \partial_x f(\rho) = 0,$$

which is the prototype of a *hyperbolic scalar conservation law* in one space dimension. Note that hyperbolic equations show a totally different behavior compared to e.g. parabolic equations. They allow only for finite speed of propagation and the solutions might be discontinuous. This has to be taken into account for the definition of appropriate network coupling conditions.

4.1.1. *Traffic Networks.* Let us generalize this concept now to traffic networks. A traffic network is nothing else than a directed graph $\mathcal{G} = (V, A)$. A vertex $v \in \mathcal{V}$ describes a junction, and an edge $e \in \mathcal{A}$ describes a road.

On each road $j \in \{1,\dots,|\mathcal{A}|\}$ we have the conservation law

$$\partial_t f\rho_j + \partial_x f_j(\rho_j) = 0, \qquad f_j(\rho_j) = \rho_j \cdot (\rho_{\max,j} - \rho_j), \quad \forall x \in [a_j,b_j], \quad t \in [0,T] \tag{4.1}$$

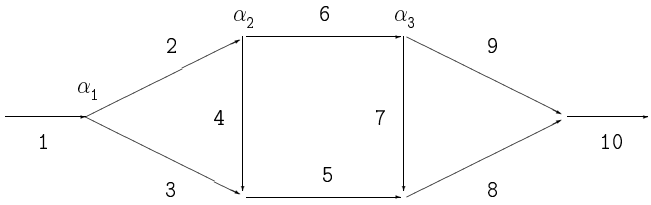


FIGURE 4.1. A Traffic Network

where $\rho_{\max,j}$ describes the maximal capacity of the j–the road, a_j, b_j are the starting and endpoint of the road and T is the time horizon we are considering. For notational simplicity we assume $v_{\max,\mathrm{j}}/\rho_{\max,\mathrm{j}} = 1$.

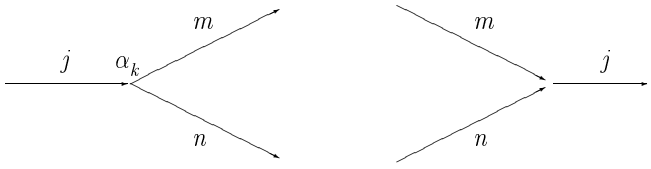
Each equation is further supplemented with the initial condition

$$\rho_j(x,0) = \rho_{j,0}(x), \quad \forall x \in [a_j, b_j]. \tag{4.2}$$

It remains to discuss the coupling conditions at the junctions. We consider a single junction with n roads labelled by $j = 1, \ldots, n$ with end b_j at the junction and m roads labeled by $j = n+1, \ldots, n+m$ with end a_j at the junction. To guarantee the conservation of the numbers of cars, at the junction the following condition is prescribed:

$$\sum_{j=1}^{n} f_j(\rho_j(b_j,t)) = \sum_{j=n+1}^{n+m} f_j(\rho_j(a_j,t)), \quad \forall t \geq 0. \tag{4.3}$$

This corresponds to the well–known Rankine–Hugoniot conditions for hyperbolic equations. One can show, at least if only up to four roads meet at one junction, that there exists a solution to this network problem which we cannot make precise here. However, the above condition does not guarantee the uniqueness of solutions on the network. This drawback can be overcome by additional conditions, for details we refer to [**36**]. We restrict the following discussion to the cases of three connected roads. Then, by composition of such junctions we can easily model all other kinds of possible junctions. There are two possibilities of junctions with a total of three connected roads: either one road disperses into two roads or two roads merging into one road.



Junction of the First Type with number k Junction of the Second Type

To get a unique solution we follow the idea of Coclite and Piccoli [**22**] and prescribe further coupling conditions at the junctions, which is described in the following example.

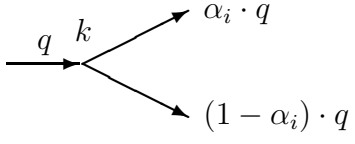
EXAMPLE 4.1. *Let us assume we have a network with just one junction k. Then we have for the solution $\rho(x,t) = (\rho_1, \rho_2, \rho_3)(x,t)$ at a junction k*

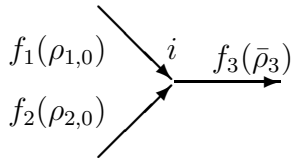
$$(\bar\rho_1, \bar\rho_2, \bar\rho_3)(t) := \begin{pmatrix} \rho_1(x = b_1, t) \\ \rho_2(x = a_2, t) \\ \rho_3(x = a_3, t) \end{pmatrix} \qquad (\bar\rho_1, \bar\rho_2, \bar\rho_3)(t) := \begin{pmatrix} \rho_1(x = b_1, t) \\ \rho_2(x = b_2, t) \\ \rho_3(x = a_3, t) \end{pmatrix}$$

If we pose certain restrictions on $\bar\rho(t), t > 0$, it turns out that $\bar\rho(t)$ is independent of time, i.e. $\bar\rho(t) \equiv \bar\rho$. Selecting the unique real values $\bar\rho_j \in \mathbb{R}$ for all roads $j = 1, 2, 3$, we can obtain a weak solution $\rho_j(x,t)$, for all $x \in [a_j, b_j]$, $j = 1, 2, 3$, and for all $t \in \mathbb{R}^+$ at the junction k by solving the following Riemann problems:

$$\begin{aligned} \text{For } j = 1, \dots, n: \qquad & \partial_t \rho_j + \partial_x f(\rho_j) = 0 \\ & \rho_j(x,0) = \begin{cases} \rho_{j,0}, & x < b_j, \\ \bar\rho_j & x = b_j, \end{cases} \\ \text{For } j = n+1, \dots, n+m: \qquad & \partial_t \rho_j + \partial_x f(\rho_j) = 0 \\ & \rho_j(x,0) = \begin{cases} \rho_{j,0} & x > a_j, \\ \bar\rho_j & x = a_j, \end{cases} \end{aligned}$$

Now, Coclite and Piccoli devise a method to obtain these unique $\bar\rho_j$. They introduce control parameters $\alpha \in (0,1)$ at each junction. Then, the additional coupling conditions read in the two different cases





$$\begin{aligned} f_2(\bar\rho_2) &= \alpha_i f_1(\rho_{1,0}), \\ f_3(\bar\rho_3) &= (1-\alpha_i) f_1(\rho_{1,0}) \end{aligned}$$

for dispersing roads, or just

$$f_3(\bar\rho_3) = f_1(\rho_{1,0}) + f_2(\rho_{2,0})$$

for merging roads.

It holds that $0 \le \alpha_i \le 1$ and the parameter α_i just distributes the flux from one incoming to the two outgoing roads. This will lateron give us the possibility to control the flow in the network. The choice of the values α_i influences directly $\bar\rho_j$. A different value $\bar\rho_j$ yields a different Riemann Problem at a dispersing junction. In turn, this yields a different network solution ρ. It is most convenient to write the unique values $\bar\rho_j$ in terms of a function U_j which just depends on the parameters $\rho_1(b_i, t)$ and α_i

for junctions of the first type and as a function of $\rho_1(b_i,t)$, $\rho_2(b_i,t)$ for junctions of the second type, i.e.

$$\bar\rho_{2,i} = U_2(\rho_1(b_i,t),\alpha_i),$$
$$\bar\rho_{3,i} = U_3(\rho_1(b_i,t),\alpha_i),$$

as well as

$$\bar\rho_{3,i} = U_3(\rho_1(b_i,t),\rho_2(b_i,t)).$$

4.1.2. *Macroscopic ODE models.* Before we turn our attention to the optimization problem, let us shortly discuss the numerical effort of for solving the forward problem. Since one has to solve a Riemann problem for each road and there might be thousands of roads, a solution of a PDE traffic network model will be very time consuming and cannot be done in real–time, even with appropriate schemes. Considering the optimal control problem we will run into troubles since each optimization step usually requires several simulations of the governing equations. Therefore, we present now a simplified model obtained by a spatial discretization of the PDE. To be more precise, based on the averaged density evolution of the traffic on each road, we perform a simple finite spatial discretization of (4.1) and obtain an ODE model. For notational simplicity we drop the subscripts for a_j, b_j and $L_j = b_j - a_j$ in the following.

Integrating (4.1) over the intervals $[a,d]$ and $[d,b]$, $a < d = \frac{a+b}{2} < b$ we obtain,

$$\partial_t \rho_j^{(a)}(t) = -\frac{2}{L}\left(f(\rho_j(d,t)) - f(\rho_j(a,t))\right), \tag{4.4a}$$

$$\partial_t \rho_j^{(b)}(t) = \frac{2}{L}\left(f(\rho_j(d,t)) - f(\rho_j(b,t)))\right), \tag{4.4b}$$

where $L = b - a$ is the length of the road. Here, we use the spatial approximations defined via

$$\rho_j^{(a)}(t) := \frac{2}{L}\int_a^d \rho_j(x,t)dx \quad \text{and} \quad \rho_j^{(b)}(t) := \frac{2}{L}\int_d^b \rho_j(x,t)dx.$$

Note that (4.4) contains additional unknowns. Thus, we assume for $\rho_j(d,t)$, that the mean value is a reasonable approximation and set

$$\rho_j(d,t) = \frac{1}{2}\left(\rho_j^{(a)}(t) + \rho_j^{(b)}(t)\right). \tag{4.5a}$$

Also the initial conditions are obtained by averaging

$$\rho_{j,0}^{(a)} = \frac{2}{L}\int_a^d \rho_{j,0}(x)dx \quad \text{and} \quad \rho_{j,0}^{(b)} = \frac{2}{L}\int_d^b \rho_{j,0}(x)dx. \tag{4.5b}$$

Finally, we obtain the values $\rho_j(a,t)$ and $\rho_j(b,t)$ by the previous coupling conditions, i.e., we define (see Example 4.1)

$$\overline{\rho}_j^a(t) = U_a^j(\rho_j^{(a)}(t), \rho_k^{(a/b)}(t), \rho_l^{(a/b)}(t)), \tag{4.5c}$$

$$\overline{\rho}_j^b(t) = U_b^j(\rho_j^{(b)}(t), \rho_r^{(a/b)}(t), \rho_s^{(a/b)}(t), \alpha), \tag{4.5d}$$

where a and b are chosen for outgoing and ingoing roads at the junctions respectively. For the above formulas let us assume that road j connects two junctions. Altogether, the equations (4.4) and (4.5) constitute a well defined ODE system. For the time discretization, we use a fixed time step τ. We use a Lax–Friedrichs discretization of the time derivative, since the naive Euler discretization would yield oscillating solutions. We emphasize that τ has to satisfy the CFL condition, since the above discretization can be seen as a (very coarse) finite–difference scheme for a conservation law which is using only three spatial points. Hence, we require that τ fulfills

$$\tau \leq \frac{L}{2 \max_\rho f'(\rho)}. \tag{4.6}$$

Finally, we obtain the discretized following ODE system for a road j connected to two junctions, where τ is chosen as in (4.6).

$$\rho_j^{(a)}(t+\tau) = \left(\frac{\overline{\rho}_j^a(t) + \rho_j^{(b)}(t)}{2}\right) - \frac{2\tau}{L}\left(f(\rho_j^{(b)}(t)) - f(\overline{\rho}_j^a(t))\right), \tag{4.7}$$

$$\rho_j^{(b)}(t+\tau) = \left(\frac{\rho_j^{(a)}(t) + \overline{\rho}_j^b(t)}{2}\right) + \frac{2\tau}{L}\left(f(\rho_j^{(a)}(t)) - f(\overline{\rho}_j^b(t))\right), \tag{4.8}$$

$$\overline{\rho}_j^a(t) = U_a^j(\rho_j^{(a)}(t), \rho_k^{(a/b)}(t), \rho_l^{(a/b)}(t)), \tag{4.9}$$

$$\overline{\rho}_j^b(t) = U_b^j(\rho_j^{(b)}(t), \rho_r^{(a/b)}(t), \rho_s^{(a/b)}(t), \alpha). \tag{4.10}$$

REMARK 4.1. *Note that this system is different from any discretization of the partial differential equation* (4.1) *due to the approximation of $\rho_j(d,t)$ and due to the definition of the boundary values $\overline{\rho}_j^{a,b}(t)$. Nevertheless, the ODE model uses the functions $U_{a,b}^j(\cdot)$ of the PDE model. Hence, also the ODE model inherits the property of traffic jams moving backwards through the junction.*

4.2. Optimization. We assume in the following that traffic can be distributed at certain dispersing junctions of the network. In terms of our model, we have a percentage $0 \leq \alpha_i \leq 1$ for each dispersing junction $i = 1, \ldots, M$. These values are the control parameters to optimize the flow in the network. In practical applications the value of α_i is just a recommendation and might be given for example by detour suggestions in the car–navigation systems or signs at the corresponding highway intersections. For simplicity we assume, that the traffic is actually distributed according to the value of α_i. Of course, there are situations where not all cars follow the recommendations and a more sophisticated model has to take into account random behavior at the junction. Hence, we have a total of M real valued controls $\vec{\alpha} = (\alpha_1, \ldots, \alpha_M)$ and the set of admissible controls is given by $S = [0,1]^M$. In the following we assume a network geometries with one inflow and one outflow arc. Further, the inflow profile $\rho_0(t)$ and a time horizon $T > 0$ is given. At each dispersing junction $i = 1, \ldots, n$ of the network we apply a control $\alpha_i \in [0,1]$ which appears in the functions U_b^j and control the distribution of the flux on the outgoing roads.

A measure for the utilization of a single road j of the network is the time and space averaged density $\int_0^T \int_{a_j}^{b_j} \rho(x,t)dxdt$. Hence, an objective functional is

$$J(\vec{\alpha}; T, \rho_0) = \sum_{j=1}^{I} \int_0^T \int_a^b \rho_j(x,t)dxdt. \tag{4.11}$$

It is easy to verify, that in the case of a single inflow arc j_0 and outflow arc j_I and sufficiently regular solutions ρ_j,

$$J(\vec{\alpha}; T, \rho_0) = \int_0^T f_{j_0}(\rho(a_{j_0}, t))dt - \int_0^T f_{j_I}(\rho(b_j, t))dt. \tag{4.12}$$

We are interested in controls $\vec{\alpha}$ such that the functional $J(\vec{\alpha})$ is minimized and give the precise optimization problem below.

REMARK 4.2. *The functional $J(\vec{\alpha}; T, \rho_0)$ is popular in the traffic engineering community. According to (4.12), it measures the possible maximal flow passing the network depending on routing decisions at the junctions. Since the flux functions are concave, high densities are related to small velocities v_j, i.e., $\rho_j v_j = f_j(\rho_j)$. Therefore, minimizing (4.11) yields a traffic situation with a large average speed. Similarly, the functional J penalizes backwards moving waves in the network. These waves can be interpreted as traffic jams.*

To obtain an approximation of J for the ODE model of the previous sections, we consider the discretized and space averaged objective function J_2, i.e.,

$$J_2(\alpha; T, \rho_0) = \sum_{t=1}^{T} \sum_{j=1}^{I} \frac{L_j}{2} \tau \left(\rho_j^{(a)}(t) + \rho_j^{(b)}(t) \right). \tag{4.13}$$

and the minimization problem

$$\min_{\vec{\alpha}} J_2 \text{ subject to } 0 \le \alpha_i \le 1,\ i = 1, \ldots, n, \text{ and } (4.7 - 4.10). \tag{4.14}$$

4.2.1. *Adjoint Equations.* The adjoint equations for the PDE model given by (4.1) can be easily derived using the general calculus. This yields

$$\begin{aligned} \partial_t \mu_j + f_j'(\rho_j)\partial_x \mu_j &= \rho_j, \quad x \in [a_j, b_j], t \in [0, T], \\ \mu_j(b_j, t) &= u_j^*(t), \quad t \in [0, T], \\ \mu_j(x, T) &= 0, \quad x \in [a_j, b_j]. \end{aligned}$$

The only problem which we still have to tackle is to find the adjoint boundary and junction conditions $u_j^*(t)$. These are well defined and given by

$$u_m^*(t) = \alpha_m \mu_r(a_r, t) + (1 - \alpha_m)\mu_s(a_s, t) \tag{4.15}$$

for a dispersing junction or by

$$u_p^*(t) = \mu_r(a_r, t), \quad u_q^*(t) = \mu_r(a_r, t) \tag{4.16}$$

for a merging junction.

REMARK 4.3. *Like in the parabolic case we are able to transform the adjoint equation to an equation forward in time.*

Further, we want to derive the optimality conditions for ODE based model (4.14). The gradient and adjoint equation are used in the numerical solution of (4.14) later on. We consider the general minimization problem

$$\min_{\alpha\in\mathbb{R}^n} f(\alpha) \text{ subject to } (4.18) \tag{4.17}$$

wherein the state equation is given by

$$y_t = F_t(y_{t-1},\alpha) \quad \text{for} \quad t = 1,\dots,T, \tag{4.18}$$

and y_0 given. For each t, F_t is a differentiable, non–linear function from $\mathbb{R}^m \times \mathbb{R}^n$ to $\mathbb{R}^m$. Further the differentiable objective function, $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, has the following form,

$$f = \sum_{t=1}^{T} f_t(y_t,\alpha). \tag{4.19}$$

The optimality system can be derived as follows. Let $g_t \in \mathbb{R}^n$ be the gradient of f as a function of control variables α. To obtain g_t, we differentiate (4.18) and use the differentials $u = d\alpha \in \mathbb{R}^n$, $z = dy \in \mathbb{R}^m$,

$$z_t = (F_t)'_y(y_{t-1},\alpha)z_{t-1} + (F_t)'_\alpha(y_{t-1},\alpha)u_t,\ z_0 = 0. \tag{4.20}$$

By (4.19)

$$df = \sum_{t=1}^{T}(\nabla_y f_t(y_t,u_t),z_t)_m + \sum_{t=1}^{T}(\nabla_\alpha f_t(y_t,u_t),u_t)_n. \tag{4.21}$$

To obtain the adjoint equation we eliminate z as follows:

(1) Setting $G_t = (F_t)'_y(y_{t-1},\alpha)$, $\quad H_t = (F_t)'_\alpha(y_{t-1},\alpha)$, $\quad \gamma_t = \nabla_y f_t(y_t,u_t)$ and $h_t = \nabla_\alpha f_t(y_t,u_t)$. Multiply each linearized state equation in (4.20) by a vector $p_t \in \mathbb{R}^m$ and summing up we have

$$0 = -(p_T,z_T) + \sum_{t=1}^{T-1}(p_t,z_t)_m + \sum_{t=1}^{T-1}(G_{t+1}^\top p_{t+1},z_t)_m + \sum_{t=1}^{T}(H_t^\top p_t,u_t)_n. \tag{4.22}$$

(2) Adding this to the expression of df,

$$df = (-p_T+\gamma_T,z_T)_m + \sum_{t=1}^{T-1}(-p_t + G_{t+1}^\top p_{t+1} + \gamma_t,z_t)_m + \sum_{t=1}^{T}(H_t^\top p_t + h_t,u_t)_n. \tag{4.23}$$

(3) Choosing p such that the coefficients of z_t vanish

$$p_T = \gamma_T,\ p_t = G_{t+1}^\top p_{t+1} + \gamma_t \quad \text{for} \quad t = T-1,\dots,1. \tag{4.24}$$

Then we obtain the gradient in desired form

$$g_t = H_t^\top p_t + h_t \quad \text{for} \quad t = 1,\dots,T. \tag{4.25}$$

Equation (4.24) is called the **adjoint equation** and equation (4.25) is called the **gradient equation**.

We apply the general discussion to the minimization problem (4.14). Considering the ODE–model (4.7-4.10) and the objective functional (4.13), we denote by

$$y_t^j = \begin{pmatrix} \rho_j^{(a)}(t) \\ \rho_j^{(b)}(t) \end{pmatrix} = \begin{pmatrix} y_t^{1,j} \\ y_t^{2,j} \end{pmatrix}, \quad \forall j = 1, \dots, I. \tag{4.26}$$

Hence, $m = 2I$ where I is the number of roads and n is the number of dispersing junctions in the network. Therefore, the control variable $\alpha \in \mathbb{R}^n$ and the state variables $y_t \in \mathbb{R}^m$ for each t. We can rewrite (4.7-4.10) as

$$\begin{aligned} y_t^{1,j} &= F_t^{1,j}(y_{t-1}^{1,j}, y_{t-1}^{2,j}, \alpha), \\ y_t^{2,j} &= F_t^{2,j}(y_{t-1}^{1,j}, y_{t-1}^{2,j}, \alpha), \end{aligned}$$

$$y_t^j = F_t^j(y_{t-1}^j, \alpha) \quad \text{for} \quad t = 1, \dots, T. \tag{4.27}$$

With these prerequisites the reader will be able to fill in the details for our special application, which will be omitted here.

4.3. Numerical Results. In this section we want to present some numerical results underlining the validity of our approach.

4.3.1. *Comparison of the ODE and the PDE Model on a Sample Network.* First, we compare the PDE and ODE models on a sample network. The PDE model is discretized using a first–order Godunov–scheme on an equidistant grid with $N_x \times N_t$ gridpoints. The objective functional J is discretized using a trapezoid–rule with equidistant spacing. We plot contour lines for J and J_2 objective functional for both models for the sample network in Figure 4.2. The sample network has two controls α_1 and α_2, hence the objective functional J and J_2 can be computed for all possible combinations of the controls. This allows to investigate if the ODE model (4.7-4.10) has similar properties than the full PDE model. We consider two different situations corresponding to a free–flow and a traffic jam situation. In the free–flow case the inflow on road 1 is given by $f_1(\rho_0) = 96\%$ and less than the capacities $M_j = 1$ of each road j. Therefore, no traffic jam can occur independent of the applied controls (α_1, α_2). The contour plots of the objective functionals are given in Figures 4.3. We observe a qualitative correspondence of both models and note that even the optimal controls $(1/2, 0)$ coincide in this case. Next, we consider a situation of congested network by varying the maximal densities on each road. We model this by a reduction of the maximal density and set $M_1 = M_2 = M_4 = M_6 = M_7 = 2$, $M_3 = 1$ and $M_5 = 0.5$. The inflow is again $f_1(\rho_0) = 96\%$. The contour lines of the functionals J_2 and J_{2t} are given in Figure 4.4. The white parts of the plot show correspond to controls (α_1, α_2), where a traffic jam reached the inflow arc. Those traffic jams appear in both the ODE and the PDE model for $\alpha_1 \leq 46\%$. Additionally, the PDE model simulates those jams for $\alpha_1 > 90\%$ in contrast to $\alpha_1 > 95\%$ for the ODE model. For the remaining controls we observe a very similar behavior.

REMARK 4.4. *Note, that the main difference between the two models can be observed in the simulation times. For larger networks we have approximately a factor 30 for one forward simulation.*

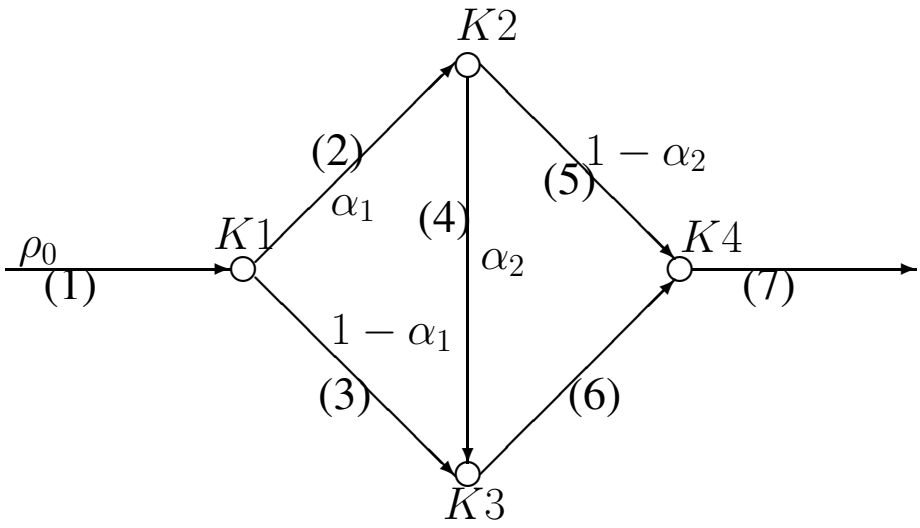


FIGURE 4.2. Sample Network

Consider again a first order descent method, which needs 50 gradient steps, then you can imagine the computational drawback of the PDE model in this case.

4.3.2. *Gradient information.* We consider the network of Figure 4.2. Gradients for the functional J_{2t} and (4.7-4.10) can be obtained either by using the discrete adjoint equations of Section 4.2.1 or by a finite difference approximation for J_{2t} using (4.7-4.10).

For each control α_1 and α_2 we proceed as follows. We fix $\tau = 1/10$. We compute finite differences by one–sided differences with $\Delta\alpha_i = 10^{-1}$. For comparison we compute the adjoints and the gradient by the calculus in Section 4.2.1. We plot the absolute difference between both in Figure 4.5. The gradients differ in order $O(\Delta\alpha_i)$ and vanish at the optimal values $(\frac{1}{2}, 0)$.

Of course, there is major advantage of using the adjoint calculus instead of finite differences. The adjoint calculus yields all derivatives after a single computation of (4.24) and (4.25); whereas for finite differences we have to compute (4.7-4.10) for each control α_i at least twice.

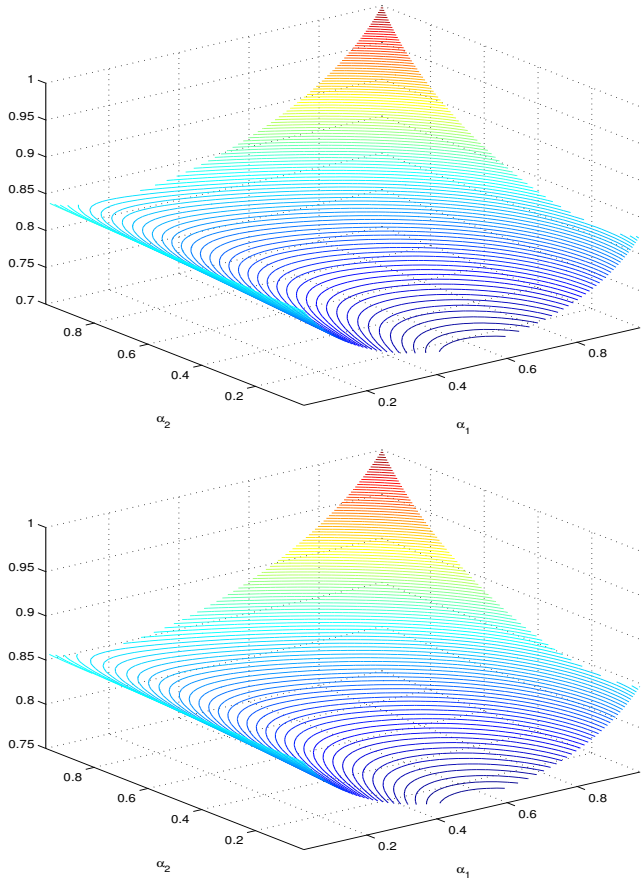


FIGURE 4.3. Contour lines of cost functional J_2, J for the ODE (left) and PDE (right) model, respectively.

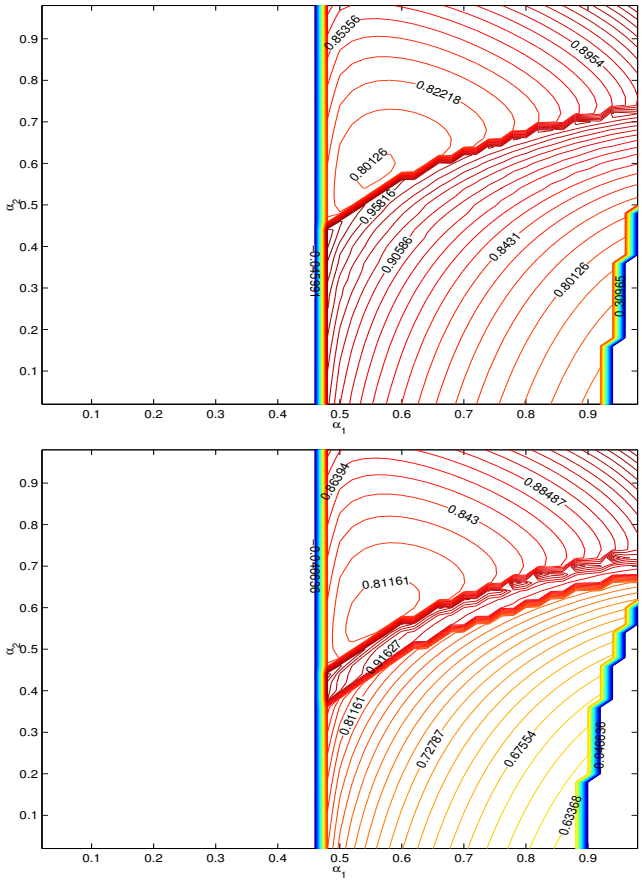


FIGURE 4.4. Contour lines of J_2 and J objective functionals for the ODE (left) and PDE (right) model with occurrence of congestions.

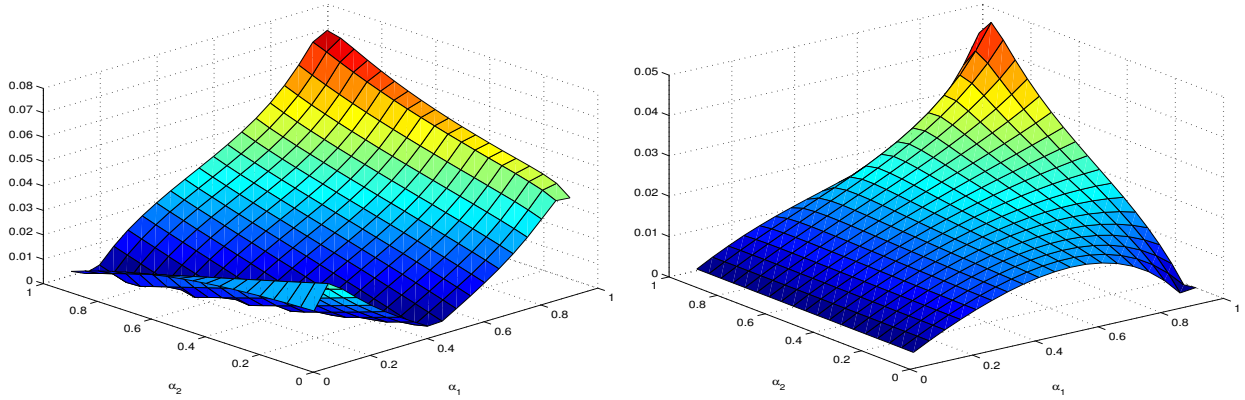


FIGURE 4.5. Difference of adjoint gradient and finite difference approximation gradients.

Bibliography

[1] R.A. Adams: *Sobolev spaces*. Academic press, 1975.

[2] E. L. ALLGOWER, K. BÖHMER, F. A. POTRA, AND W. C. RHEINBOLDT, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[3] H.W. Alt: Lineare Funktionalanalysis, Springer (1999).

[4] W. ALT, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[5] ——, *Discretization and mesh-independence of Newton's method for generalized equations*, in Mathematical programming with data perturbations, Dekker, New York, 1998, pp. 1–30.

[6] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem, Computational Optimization and Applications 23, 201–229 (2002).

[7] N. Bellomo, M. Delitala, and J. Nedelec. On the mathematical theory of vehicular traffic flow fluid dynamic and kinetic modelling. *Math. Mod. Meth. Appl. Sc.*, 12:1801–1843, 2002.

[8] M. Berggren. Approximation of very weak solutions to boundary value problems, SIAM J. Numer. Anal. 42, 860–877 (2004).

[9] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[10] D. P. BERTSEKAS, *Nonlinear Programming (2nd edition)*, Athena Scientific, 1999.

[11] J.F. Bonnans, A. Shapiro: Optimization problems with perturbations: A guided tour. *SIAM Rev.* 40, pp. 228–264, 1998. Springer, 1999.

[12] M. Burger, H.W. Engl, and P. Markowich. Inverse doping problems for semiconductor devices. In T.Tang J.A.Xu L.A.Ying T.F.Chan, Y. Huang, editor, *Recent Progress in Computational and Applied PDEs*, pages 39–54. Kluwer, 2002.

[13] M. Burger, H. W. Engl, P. A. Markowich, and P. Pietra. Identification of doping profiles in semiconductor devices. *Inverse Problems*, 17:1765–1795, 2001.

[14] E. Casas. *L^2 estimates for the finite element method for the Dirichlet problem with singular data*, Numer. Math. 47, 627–632 (1985).

[15] E. Casas. Control of an elliptic problem with pointwise state constraints, SIAM J. Cont. Optim. 4, 1309–1322 (1986).

[16] E. Casas. *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Cont. Optim. 31, 993–1006 (1993).

[17] E. Casas. *Error Estimates for the Numerical Approximation of Semilinear Elliptic Control Problems with Finitely Many State Constraints*, ESAIM, Control Optim. Calc. Var. 8, 345–374 (2002).

[18] E. Casas, and M. Mateos. *Uniform convergence of the FEM. Applications to state constrained control problems*. Comp. Appl. Math. 21 (2002).

[19] E. Casas, M. Mateos, and F. Tröltzsch. Error estimates for the numerical approximation of boundary semilinear elliptic control problems, Report 2003/21, Institut für Mathematik, TU Berlin (2003).

[20] E. Casas and J.P. Raymond. Error estimates for the numerical approximation of Dirichlet Boundary control for semilinear elliptic equations, Preprint (2005).

[21] F. H. CLARKE, *Optimization and nonsmooth analysis*, Wiley, New York, 1983.

[22] G. Coclite and B. Piccoli. Traffic flow on a road network. *To appear in SIAM J. Math. Anal.*

[23] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state constrained elliptic control problem, in preparation (2005).
[24] J. Douglas, 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 (1975).
[25] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, 1983.
[26] P. DEUFLHARD AND F. A. POTRA, *Asymptotic mesh independence of Newton-Galerkin methods via a refined Mysovskiu theorem*, SIAM J. Numer. Anal., 29 (1992), pp. 1395–1412.
[27] A. L. DONTCHEV, W. W. HAGER, AND V. M. VELIOV, *Uniform convergence and mesh independence of Newton's method for discretized variational problems*, SIAM J. Control Optim., 39 (2000), pp. 961–980.
[28] L. C. Evans: *Partial Differential Equations*. American Mathematical Society, 1998.
[29] W. Fang and E. Cumberbatch. Inverse problems for metal oxide semiconductor field-effect transistor contact resistivity. *SIAM J. Appl. Math.*, 52:699–709, 1992.
[30] W. Fang and K. Ito. Reconstruction of semiconductor doping profile from laser-beam-induced current image. *SIAM J. Appl. Math.*, 54:1067–1082, 1994.
[31] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*, (2nd ed.). Springer, 1983.
[32] H. GOLDBERG AND F. TRÖLTZSCH, *On a Lagrange-Newton method for a nonlinear parabolic boundary control problem*, Optimization Methods and Software, 8 (1998), pp. 225–247.
[33] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numer. Math.*, 87(2):247–282, 2000.
[34] W. HACKBUSCH, *Multi-grid methods and applications*, Springer, New York, 1985.
[35] M. Herty. *Mathematics of Traffic Flow Networks*. PhD thesis, TU Darmstadt, 2004.
[36] M. Herty and A. Klar. Simplified dynamics and optimization of large scale traffic networks. *Math. Mod. Meth. Appl. Sc.*, 14(4):579–601, 2004.
[37] M. Hintermüller and K. Kunisch. Path following methods for a class of constrained minimization methods in function spaces, Report RICAM2004-07, RICAM Linz (2004).
[38] M. Hintermüller and K. Kunisch. Feasible and non-interior path following in constrained minimization with low multiplier regularity, Report, Universität Graz (2005).
[39] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semi-smooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[40] M. HINTERMÜLLER AND M. ULBRICH, *A mesh-independence result for semismooth Newton methods*, Math. Programming, 101 (2004), pp. 151–184.
[41] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case, Computational Optimization and Applications 30, 45–63 (2005).
[42] M. Hinze and R. Pinnau. Optimal control of the drift diffusion model for semiconductor devices. In K.-H. Hoffmann, I. Lasiecka, G. Leugering, and J. Sprekels, editors, *Optimal Control of Complex Structures*, volume 139 of *ISNM*, pages 95–106. Birkhäuser, 2001.
[43] M. Hinze and R. Pinnau. An optimal control approach to semiconductor design. *Math. Mod. Meth. Appl. Sc.*, 12(1):89–107, 2002.
[44] M. Hinze and R. Pinnau. Mathematical tools in optimal semiconductor design. *To appear in TTSP*, 2005.
[45] H. JÄGER AND E. W. SACHS, *Global convergence of inexact reduced SQP methods*, Optimization Methods and Software, 7 (1997), pp. 83–110.
[46] J. Jost: *Postmodern Analysis*. Springer, 1998.
[47] C. T. KELLEY, *Iterative methods for optimization*, SIAM, Philadelphia, 1999.
[48] C. T. KELLEY AND E. W. SACHS, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.
[49] D. Kinderlehrer, G. Stampacchia: *Introduction to Variational Inequalities and their Applications*, Academic Press, 1980.

[50] B. KUMMER, *Newton's method for nondifferentiable functions*, in Advances in mathematical optimization, Akademie-Verlag, Berlin, 1988, pp. 114–125.
[51] E. W. Larsen, G. Thömmes, M. Seaid, Th. Götz, and A. Klar. Simplified P_N Approximations to the Equations of Radiative Heat Transfer and applicatons to glass manufacturing. *J. Comp. Phys*, 183(2):652–675, 2002.
[52] W.R. Lee, S. Wang, and K.L. Teo. An optimization approach to a finite dimensional parameter estimation problem in semiconductor device design. *Journal of Computational Physics*, 156:241–256, 1999.
[53] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, second edition, 1989.
[54] P. A. Markowich. *The Stationary Semiconductor Device Equations*. Springer–Verlag, Wien, first edition, 1986.
[55] P. A. Markowich, Ch. A. Ringhofer, and Ch. Schmeiser. *Semiconductor Equations*. Springer–Verlag, Wien, first edition, 1990.
[56] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems, SIAM J. Control Optim. 43, 970–985 (2004).
[57] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control problems of PDEs with regularized pointwise state constraints, Preprint 14, Inst. f. Mathematik, TU Berlin, to appear in Computational Optimization and Applications (2004).
[58] C. Meyer, U. Prüfert, and F. Tröltzsch. On two numerical methods for state-constrained elliptic control problems, Technical Report 5-2005, Institut für Mathematik, TU Berlin (2005).
[59] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[60] M. S. Mock. *Analysis of Mathematical Models of Semiconductor Devices*. Boole Press, Dublin, first edition, 1983.
[61] J. Naumann and M. Wolff. A uniqueness theorem for weak solutions of the stationary semiconductor equations. *Appl. Math. Optim.*, 24:223–232, 1991.
[62] J. Nocedal and S.J. Wright. *Nonlinear optimization*. Springer Series in Operations Research, Springer 1999.
[63] R. Pinnau. Analysis of an optimal boundary control problem for the SP_1 system. *Submitted*, 2005.
[64] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[65] M. Renardy, R. C. Rogers: *An Introduction to Partial Differential Equations*. Springer, 1993.
[66] S.M Robinson: Stability theory for systems of inequalities in nonlinear programming, part II: differentiable nonlinear systems. *SIAM J. Num. Anal.* 13, pp. 497–513, 1976.
[67] S. M. ROBINSON, *Strongly regular generalized equations*, Mathematics of Operations Research, 5 (1980), pp. 43–62.
[68] D.L. Scharfetter and H.K. Gummel. Large signal analysis of a silicon read diode oscillator. *IEEE Trans. Electr. Dev.*, 15:64–77, 1969.
[69] A.H. Schatz. *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids. I: Global estimates*, Math. Comput. 67, No.223, 877–899 (1998).
[70] S. SCHOLTES, *Introduction to piecewise differentiable equations*, technical report no. 53/1994, Universitat Karlsruhe, Institut fur Statistik und Mathematische Wirtschaftstheorie, 1994.
[71] S. Selberherr. *Analysis and Simulation of Semiconductor Devices*. Springer, Wien, New York, 1984.
[72] M. Stockinger, R. Strasser, R. Plasun, A. Wild, and S. Selberherr. A qualitative study on optimized MOSFET doping profiles. In *Proceedings SISPAD 98 Conf.*, pages 77–80, 1998.
[73] S. M. Sze. *Physics of Semiconductor Devices*. Wiley, New York, second edition, 1981.
[74] F. Tröltzsch. Optimale Steuerung mit partiellen Differentialgleichungen (2005).
[75] M. ULBRICH, *Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces*, Habilitationsschrift, Zentrum Mathematik, Technische Universität München, München, Germany, 2001.
[76] ——, *On a nonsmooth Newton method for nonlinear complementarity problems in function space with applications to optimal control*, in Complementarity: Applications, algorithms and extensions (Madison, WI, 1999), Kluwer Acad. Publ., Dordrecht, 2001, pp. 341–360.
[77] M. Ulbrich. Semismooth Newton Methods for Operator Equations in Function Spaces, SIAM J. Optim. 13, 805–841 (2003).

[78] M. ULBRICH AND S. ULBRICH, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.

[79] ——, *A multigrid semismooth Newton method for contact problems in linear elasticity*, tech. rep., Fachbereich Mathematik, Universität Hamburg, 2005. In Vorbereitung.

[80] J. Wloka: *Funktionalanalysis unf ihre Anwendungen*. De Gruyter, 1971.

[81] K. Yosida: *Functional Analysis*. Springer, 1980.

[82] J. Zowe, S. Kurcyusz: Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optimization* 5, pp. 49–62, 1979.