Technische Universität Darmstadt
Fachbereich Mathematik

Optimization with Partial Differential Equations

Stefan Ulbrich

Summer 2011

with contributions by Michael Ulbrich.

Contents

Preface

These notes contain in part material from the lecture notes by M. Hinze, R. Pinnau, M. Ulbrich, S. Ulbrich for the autumn school *Modelling and optimization with partial differential equations* (Hamburg, September 26–30, 2005).

In the current version of these lecture notes, only the contributions of M. Ulbrich and S. Ulbrich have been used.

Chapter 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. Many of the techniques that we will develop can also be applied to optimization problems with other constraints than PDEs, e.g., ordinary differential equations (ODEs) or partial differntial-algebraic equations (PDAEs).

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 bounded below on the feasible set W_{ad} of (1.3) and that (1.3) has a feasible point. Then

$$-\infty < f^* = \inf_{w \in W_{ad}} f(w).$$

We consider a minimizing sequence $(w^k) \subset W_{ad}$, i.e., $E(w^k) = 0$, $C(w^k) \le 0$, $f(w^k) \to f^*$. 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

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

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:

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

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.

1.2.1 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.

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

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

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

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

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

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.

1.2.2 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.

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

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*.

1.2.3 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:

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

Here, $\mathrm{div}\, v = \sum_i (v_i)_{x_i}$ 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

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

and the volume constraint

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

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

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

Chapter 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., [Al99], [Jo98], [ReRo93], [Wl71], [Yo80].

2.1 Banach and Hilbert spaces

2.1.1 Basic definitions

Definition 2.1.1 (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 2.1.2

1. *The function space*

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

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

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

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 2.1.3 (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 H, 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 2.1.4 *Let $\emptyset \neq \Omega \subset \mathbb{R}^n$ 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 2.1.5 *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 2.1.6 *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 2.1.7 *For bounded Ω the space $C(\bar{\Omega})$ is separable (the polynomials with rational coefficients are dense by Weierstraß's approximation theorem).*

2.1.2 Linear operators and dual space

Obviously, linear partial differential operators define linear mappings between function spaces. We recall the following definition.

Definition 2.1.8 (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*

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

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 2.1.9 *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 2.1.10 *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 2.1.11 (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 2.1.9) 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 2.1.12 (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 (we will introduce this later).

Definition 2.1.13 *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 $\|A^*\|_{Y^*,X^*} = \|A\|_{X,Y}$.

2.2 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

2.2.1 Lebesgue spaces

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

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

2.2.2 Lebesgue measurable functions and Lebesgue integral

Definition 2.2.1 *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 2.2.2 *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 2.2.3 *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 2.2.4 *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 2.2.5 *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 $f \ge 0$ 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 2.2.6 (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* $f 1_A$ *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 $d\mu(x)$. $\Box$

2.2.3 Definition of Lebesgue spaces

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

Definition 2.2.7 *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*

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

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

$$u \sim v \ \text{ in } L^p(\Omega) \ \ :\Longleftrightarrow \ \ \|u - v\|_{L^p(\Omega)} = 0 \ \overset{\text{by Lemma 2.2.8}}{\Longleftrightarrow} \ \ u = v \text{ a.e.}$$

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 $p \in [1, \infty]$. $\Box$

We collect several important facts of Lebesgue spaces.

Lemma 2.2.8 *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 $\int_{\mathbb{R}^n} |w|^p dx = 0$. $\Box$

Theorem 2.2.9 (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 2.2.10 (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

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

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

Theorem 2.2.11 *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*

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

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

2.2.4 Density results and convergence theorems

A fundamental result is the following:

Theorem 2.2.12 (Dominated convergence theorem) *Let $\Omega \in \mathcal{B}_n$. Assume that $f_k : \Omega \to \mathbb{R}$ are measurable with*

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

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

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

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

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

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

Lemma 2.2.13 *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 2.2.14 *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.

2.2.5 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 $\alpha \in \mathbb{N}_0^n$, $|\alpha \leq k$, the identity holds (integrate $|\alpha|$-times by parts)

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

This motivates the definition

Definition 2.2.15 *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*

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

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

Remark

1. By Lemma 2.2.14, (2.2) determines the weak derivative $D^\alpha u \in L^1_{loc}(\Omega)$ uniquely.
2. For $u \in C^k(\bar{\Omega})$ and $\alpha \in \mathbb{N}_0^n$, $|\alpha| \leq k$, the classical derivative $w = D^\alpha u$ satisfies (2.1) and thus (2.2). Hence, the weak derivative is consistent with the classical derivative. □

2.2.6 Regular domains and integration by parts

For $k \in \mathbb{N}_0$ and $\beta \in (0, 1]$ let

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

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

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

Of course, 1-Hölder continuity is Lipschitz continuity.

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

Definition 2.2.16 ($C^{k,\beta}$-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 2.2.17 *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with $C^{0,1}$-boundary. Then for all $u, v \in C^1(\bar{\Omega})$*

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

2.2.7 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 2.2.18 *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*

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

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 2.2.19 • *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 2.2.20 *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 2.2.21 *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 2.2.22 If Ω has Lipschitz-boundary then $W_0^{k,p}(\Omega)$ 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 determined up to a set of measure zero, the definition of boundary values requires care. We resolve the problem by defining a *trace operator*.

Theorem 2.2.23 *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 $\partial\Omega$.*

2.2.8 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 2.2.24 (Poincaré's inequality)
Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then there exists a constant $C > 0$ with

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

2.2.9 Sobolev imbedding theorem

Sobolev spaces are embedded in classical spaces:

Theorem 2.2.25 *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 [Al99], [Ad75], [Ev98]. $\Box$

Example 2.2.26 *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 $n \leq 3$.*

2.2.10 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 2.2.27 *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

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

□

2.3 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 [Al99], [Ev98], [ReRo93], [Tr05], [Wl71].

2.3.1 Weak solutions of the Poisson equation

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 2.2.22. 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 2.3.1 *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

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

is linear and bounded:

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

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. $\square$

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 2.3.2 *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

$$\begin{aligned}V &= H_0^1(\Omega),\\ a(y,v) &= \int_\Omega \nabla y \cdot \nabla v\,dx, \quad y,v \in V, &(2.14)\\ F(v) &= (f,v)_{L^2(\Omega)}, \quad v \in V. &(2.15)\end{aligned}$$

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

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

Remark Since $a(y, \cdot) \in V^*$ for all $y \in V$ and $y \in V \mapsto a(y, \cdot) \in V^*$ is continuous 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

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

□

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

Lemma 2.3.3 (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, y 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 2.3.4 *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 2.3.3. 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

$$a(y,y) = \int_\Omega \nabla y \cdot \nabla y \, dx = |y|^2_{H^1_0(\Omega)} \geq \frac{1}{C^2} \|y\|^2_{H^1_0(\Omega)} = \frac{1}{C^2} \|y\|^2_V$$

which shows the V-coercivity (2.20).

Finally, the definition of F in (2.15) yields

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

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

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 $\Omega \subset \mathbb{R}^n$ is open and bounded with $C^{0,1}$-boundary, $f \in L^2(\Omega)$ 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 2.3.5 *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 2.3.6 *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 [Tr05].

Theorem 2.3.7 *Let the assumptions of the previous theorem hold and let $r > n/2$, $s > n-1$, $n \ge 2$. 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*

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

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 [KS80].

2.3.2 Weak solutions of uniformly elliptic equations

More generally, we can consider general second order elliptic PDEs of the form

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

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

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

with the bilinear form

$$a(y,v) = \int_\Omega \sum_{i,j=1}^{n}(a_{ij}\, y_{x_i} v_{x_j} + c_0\, y\, v)\, dx.$$

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

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

2.3.3 An existence and uniqueness result for semilinear elliptic equations

We finally state an existence and uniqueness result for a uniformly elliptic semilinear equation

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

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

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

Under these assumptions the theory of maximal monotone operators and a technique of Stampacchia can be applied to extend Theorem 2.3.7 to the semilinear elliptic equation (2.30), see for example [Tr05].

Theorem 2.3.8 *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 $r > n/2$, $s > n-1$, $2 \leq n \leq 3$.*

If $d(\cdot,0) = 0$ and $b(\cdot,0) = 0$ 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*

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

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

If more generally $d(\cdot,0) \in L^r(\Omega)$ and $b(\cdot,0) \in L^s(\partial\Omega)$ then there exists a constant $C_\infty > 0$ with

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq C_\infty(\|f - d(\cdot,0)\|_{L^r(\Omega)} + \|g - b(\cdot,0)\|_{L^s(\partial\Omega)}),$$

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

2.4 Gâteaux- and Fréchet Differentiability

We extend the notion of differentiability to operators between Banach spaces.

Definition 2.4.1 *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,

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

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

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

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

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

Chapter 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 3.1.1 *Let X be a normed 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 3.1.2 *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 sequentially weakly closed, i.e., for every sequence $(x_k) \subset C$ with $x_k \rightharpoonup x$ one has $x \in C$.

Definition 3.1.3 *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 3.1.4 (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 [Al99], [Yo80].

Theorem 3.1.5 (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 3.1.6 *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 3.1.7 *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 $x_k \rightharpoonup x$ and $A \in \mathcal{L}(X,Y)$ we see that $Ax_k \rightharpoonup Ax$. Since (x_k) is bounded (Theorem 3.1.2), 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 $(Ax_k)_{K'} \rightharpoonup Ax$ and $(Ax_k)_{K'} \to y$ we must have $y = Ax$, which is a contradiction. $\Box$

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 3.2.1

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 3.2.2 *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 3.2.3 *Let assumption 3.2.1 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 3.1.4 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 3.1.5 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 3.3.1

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, i.e., $(y_k,u_k)\rightharpoonup (y,u)$ in $Y\times U$ implies $E(y_k,u_k)\rightharpoonup E(y,u)$ in Z.*
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 3.3.2 *To show 4. for the semilinear state equation*

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

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, $y_k\rightharpoonup y$ weakly in Y implies $y_k\to y$ strongly in $L^5(\Omega)$ and this implies (see below) $y_k^3\to y^3$ strongly in $L^{5/3}(\Omega)=L^{5/2}(\Omega)^*\subset Y^*$ (note that $Y\subset L^{5/2}(\Omega)$), 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, for appropriate $t\in[0,1]$ we have

$$|b^3-a^3|=3|(a+t(b-a))^2(b-a)|\le 3\max(|a|^2,|b|^2)|b-a|\le 3(|a|^2+|b|^2)|b-a|.$$

Therefore,

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

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

$$\begin{aligned}\|y_k^3 - y^3\|_{L^{5/3}} &\le \|y_k^2|y_k - y|\|_{L^{5/3}} + \|y^2|y_k - y|\|_{L^{5/3}} \le (\|y_k\|_{L^5}^2 + \|y\|_{L^5}^2)\|y_k - y\|_{L^5} \\ &\to 2\|y\|_{L^5}^2 \cdot 0 = 0 \quad \textit{as } y_k \to y \textit{ in } L^5(\Omega).\end{aligned}$$

We summarize: $y_k \rightharpoonup y$ *in* Y *implies* $y_k \to y$ *in* $L^5(\Omega)$. *From this it follows that* $y_k^3 \to y^3$ *in* $L^{5/3}(\Omega)$ *which implies* $y_k^3 \to y^3$ *in* $Y^* = Z$. *Hence, 4. follows, since the remaining linear operators in* $E(y,u)$ *are bounded.*

3.4 Applications

3.4.1 Distributed control of elliptic equations

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

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

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 2.3.4 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 2.3.4, $A \in \mathcal{L}(Y, Y^*)$ has a bounded inverse. Therefore, Assumption 3.2.1 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).

Chapter 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 an 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, the derivative of $\hat{f}$ can be computed much cheaper by solving a single adjoint equation.

4.2 Adjoint approach

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

$$\langle \hat{f}'(u), s \rangle_{U^*,U} = \langle L'_y(y(u), u, p), y'(u)s \rangle_{Y^*,Y} + \langle L'_u(y(u), u, p), s \rangle_{U^*,U}. \tag{4.4}$$

Now we choose a special $p = p(u) \in Z^*$, namely such that the *adjoint equation* holds

$$L'_y(y(u), u, p) = 0. \tag{4.5}$$

To write the adjoint equation in a concrete form, we note that for all $d \in Y$

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

Therefore,

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

and the adjoint equation (4.5) reads

Adjoint Equation:

$$E'_y(y(u), u)^* p = -f'_y(y(u), u). \tag{4.6}$$

Completely analogous we obtain

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

Now, choosing the *adjoint state* $p = p(u) \in Z^*$ according to the adjoint equation (4.6), we obtain from (4.4) that

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

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

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

4.3 Application to a linear-quadratic optimal control problem

We consider the linear-quadratic optimal control problem

$$\begin{aligned} \min_{(y,u)\in Y\times U} \quad & f(y,u) \stackrel{\text{def}}{=} \frac{1}{2}\|Qy-q_d\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 \\ \text{subject to} \quad & Ay+Bu=g, \quad u\in U_{ad}, \;\; y\in Y_{ad} \end{aligned} \tag{4.7}$$

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, let Assumption 3.2.1 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

$$\begin{aligned} & f_y'(y,u)=(Qy-q_d,Q\cdot)_H=\langle Qy-q_d,Q\cdot\rangle_{H^*,H}=\langle Q^*(Qy-q_d),\cdot\rangle_{Y^*,Y}=Q^*(Qy-q_d), \\ & f_u'(y,u)=\alpha(u,\cdot)_U=\alpha u. \end{aligned}$$

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

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

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

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

4.3.1 Application to distributed control of an elliptic equation

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

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

The appropriate spaces are

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

and we assume

$$l, r \in U, \;\; l \le r, \quad y_d \in L^2(\Omega), \quad \alpha > 0, \quad y_a \in L^2(\partial\Omega), \quad \gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \ge 0.$$

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

$$y \in Y, \quad a(y,v) = (\gamma u, v)_{L^2(\Omega)} + ((\beta/\kappa)y_a, v)_{L^2(\partial\Omega)} \quad \forall\, v \in Y = H^1(\Omega) \tag{4.8}$$

with

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

By the existence and uniqueness result of Theorem 2.3.6 a induces an operator $A \in \mathcal{L}(Y, Y^*)$, which has a bounded inverse.

Hence, we set $Z = Y^*$, $H = L^2(\Omega)$ and

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

Then, we arrive at a linear quadratic problem of the form (4.7) that satisfies Assumption 3.2.1.

Adjoint approach

Variant 1: Determine the adjoint operators

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 $B^*v = -\gamma v$.

Now, since $Qy = y$, we have This means that

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

Moreover,

$$f'_u(y,u) = \alpha(u, \cdot)_{L^2(\Omega)} = \alpha u.$$

Taking all together, the adjoint equation thus reads

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

which is the weak form of

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

The adjoint gradient representation then is

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

Variant 2: Work directly with the Lagrangian

If the PDE constraint is given in weak form, it is often more convenient to work directly with the Lagrangian.

The operator $E : Y \times U \mapsto Z = Y^*$ is given by the weak formulation (4.8), i.e., for all $p \in Z^* = Y$ we have

$$\langle p, E(y,u) \rangle_{Z^*,Z} = a(y,p) - (\gamma u, p)_{L^2(\Omega)} - ((\beta/\kappa) y_a, p)_{L^2(\partial\Omega)}.$$

Hence, the Lagrangian has the form

$$L(y,u,p) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} a(y,p) - (\gamma u, p)_{L^2(\Omega)} - ((\beta/\kappa) y_a, p)_{L^2(\partial\Omega)}.$$

The adjoint equation is now

$$L'_y(y(u), u, p) = 0 \quad \iff \quad \langle L'_y(y(u), u, p), v \rangle_{Y^*,Y} = 0 \quad \forall\, v \in Y.$$

Inserting the Lagrangian, the adjoint equation reads

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

This is exactly the same adjoint equation as in Variant 1 (note that $a(v,p) = a(p,v)$. The reduced derivative is now given by

$$\hat{f}'(u) = L'_u(y(u), u, p) = \alpha(u, \cdot)_{L^2(\Omega)} - (\gamma p, \cdot)_{L^2(\Omega)} = (\alpha u - \gamma p, \cdot)_{L^2(\Omega)} = \alpha u - \gamma p,$$

where we have used the identification $U = U^*$ in the last equality.

4.4 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

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

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

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

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 (4.7) 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.

Chapter 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 5.1.1 *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,

$$f(w) - f(\bar{w}) = \frac{(1 - t)f(\bar{w}) + tf(w) - f(\bar{w})}{t} \geq \frac{f(\bar{w} + t(w - \bar{w})) - f(\bar{w})}{t} \overset{t\to 0^+}{\longrightarrow} \langle f'(\bar{w}), w - \bar{w}\rangle_{W^*,W}.$$

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 $(w_k)_K \rightharpoonup \bar{w}$. 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. $\square$

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 5.1.2 *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:*

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

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 $w_1, w_2 \in W$, $w_1 \ne w_2$, 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 5.1.1, 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 5.1.1 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):

$$\begin{aligned} (v - P(v), P(w) - P(v))_W &\leq 0, \\ (w - P(w), P(v) - P(w))_W &\leq 0. \end{aligned}$$

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,

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

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,

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

□

Lemma 5.1.3 *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:*

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

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

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

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

By Lemma 5.1.2 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 5.1.2 b) yields

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

□

Corollary 5.1.4 *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 5.2.1

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 3.2.1 ensures Assumption 5.2.1, since $E'_y(y,u)=A$.

5.2.1 A general first order optimality condition

Now consider problem (5.8) and let Assumption 5.2.1 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 5.2.2 *Let Assumption 5.2.1 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 5.1.1. □

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

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

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

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

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

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

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 $\nabla \hat{f}(\bar{u})\,(u - \bar{u}) \geq 0$ a.e. for all $u \in U_{ad}$ and thus

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

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

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

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 $a \leq \bar{u} \leq b$ and $\bar{z}_a, \bar{z}_b \geq 0$ hold trivially. Furthermore,

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

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 5.1.3 to prove the equivalence of i) and iv). □

5.2.2 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 5.2.2.

Corollary 5.2.4 *Let* $(\bar{y}, \bar{u})$ *an optimal solution of the problem* (5.8) *and let Assumption 5.2.1 hold. Then there exists an adjoint state (or Lagrange multiplier)* $\bar{p} \in Z^*$ *such that the following optimality conditions hold*

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

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

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

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

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

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

5.2.3 Applications

General linear-quadratic problem

We apply the result to the linear-quadratic problem

$$\begin{aligned}\min_{(y,u)\in Y\times U} \quad & f(y,u):=\frac12\|Qy-q_d\|_H^2+\frac{\alpha}{2}\|u\|_U^2\\ \text{subject to} \quad & Ay+Bu=g, \quad u\in U_{ad}\end{aligned} \tag{5.21}$$

under Assumption 3.2.1. Then

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

and Corollary 5.2.4 is applicable. We only have to compute L_y' and L_u' for the Lagrange function

$$\begin{aligned}L(y,u,p)&=f(y,u)+\langle p,Ay+Bu-g\rangle_{Z^*,Z}\\ &=\frac12(Qy-q_d,Qy-q_d)_H+\frac{\alpha}{2}(u,u)_U+\langle p,Ay+Bu-q\rangle_{Z^*,Z}.\end{aligned}$$

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

$$A\bar y+B\bar u=g, \tag{5.24}$$
$$A^*\bar p=-Q^*(Q\bar y-q_d), \tag{5.25}$$
$$\bar u\in U_{ad}, \quad (\alpha\bar u+B^*\bar p,u-\bar u)_U\ge 0 \quad \forall\, u\in U_{ad}. \tag{5.26}$$

Distributed control of elliptic equations

We consider next the distributed optimal control of a steady temperature distribution with boundary temperature zero

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

where

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

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

$$U = H = L^2(\Omega), \quad Y = H^1_0(\Omega), \quad Z = Y^*, \quad g = 0, \quad Q = I_{Y,H},$$

and

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

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 5.2.4 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 5.2.3 we can summarize

Theorem 5.2.5 *If* $(\bar{y}, \bar{u})$ *is an optimal solution of* (5.27) *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 conditions hold in the weak sense.*

$$\begin{aligned}
&-\Delta \bar{y} = \gamma \bar{u}, \quad \bar{y}|_{\partial\Omega} = 0, \\
&-\Delta \bar{p} = -(\bar{y} - y_d), \quad \bar{p}|_{\partial\Omega} = 0, \\
&\alpha \bar{u} - \gamma \bar{p} + \bar{z}_b - \bar{z}_a = 0, \\
&\bar{u} \geq a, \quad \bar{z}_a \geq 0, \quad \bar{z}_a\,(\bar{u} - a) = 0, \\
&\bar{u} \leq b, \quad \bar{z}_b \geq 0, \quad \bar{z}_b\,(b - \bar{u}) = 0.
\end{aligned}$$

Distributed control of semilinear elliptic equations

We consider next the distributed optimal control of a semilinear elliptic PDE:

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

where

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

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 2.2.25 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 Ḧolder's inequality:

Lemma 5.2.6 *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 $p = 2$:

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

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,

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

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

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

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

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

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 $E'_y(y,u)v = Av + 2y^2 v$: Let $(\bar{y}, \bar{u}) \in Y \times U$ be an optimal solution. Then by Corollary 5.2.4 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 5.2.3 we can summarize

Theorem 5.2.7 *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.*

$$\begin{aligned}
&-\Delta\bar{y} = \gamma\bar{u}, \quad \bar{y}|_{\partial\Omega} = 0,\\
&-\Delta\bar{p} + 3\bar{y}^2\bar{p} = -(\bar{y} - y_d), \quad \bar{p}|_{\partial\Omega} = 0,\\
&\alpha\bar{u} - \gamma\bar{p} + \bar{z}_b - \bar{z}_a = 0,\\
&\bar{u} \ge a, \quad \bar{z}_a \ge 0, \quad \bar{z}_a\,(\bar{u} - a) = 0,\\
&\bar{u} \le b, \quad \bar{z}_b \ge 0, \quad \bar{z}_b\,(b - \bar{u}) = 0.
\end{aligned}$$

5.3 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

$$G(w) := \begin{pmatrix} E(w) \\ C(w) \end{pmatrix} \in \{0\} \times \mathcal{K}_C =: \mathcal{K}.$$

5.3.1 A basic first order optimality condition

Let $\bar{w}$ be a local solution of (5.35). To develop an extension of Theorem 5.2.2, we define the cone of feasible directions as follows.

Definition 5.3.1 *Let $W_{ad} \subset W$ be nonempty. The* tangent cone *of W_{ad} at $w \in W_{ad}$ is defined by*

$$T(W_{ad}; w) = \left\{ s \in W \;:\; \exists\, \eta_k > 0, w_k \in W_{ad} : \; \lim_{k\to\infty} w_k = w, \;\; \lim_{k\to\infty} \eta_k (w_k - w) = s \right\}.$$

Then we have the following optimality condition.

Theorem 5.3.2 *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 $\eta_k(w_k - \bar{w}) \to s$. $\Box$

5.3.2 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 5.3.3 *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 $s \in T(W_{ad}; \bar{w})$. $\square$

In general, (5.37) can be ensured if $\bar{w}$ satisfies the

Regularity Condition of Robinson:

$$0 \in \operatorname{int}\left(G(\bar{w}) + G'(\bar{w})\left(\mathcal{C} - \bar{w}\right) - \mathcal{K}\right). \tag{5.39}$$

We have the following important and deep result by Robinson [Ro76].

Theorem 5.3.4 *Robinson's regularity condition* (5.39) *implies the constraint qualification* (5.37).

Proof: See [Ro76, Thm. 1, Cor. 2]. $\square$

5.3.3 Karush-Kuhn-Tucker conditions

Using Robinson's regularity condition, we can write the optimality condition (5.38) in a more explicit form.

Theorem 5.3.5 *(Zowe and Kurcyusz [ZK79])*
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

$$G(\bar{w}) \in \mathcal{K}, \tag{5.40}$$

$$\bar{q} \in \mathcal{K}^\circ := \{q \in V^* \, : \, \langle q, v \rangle_{V^*,V} \leq 0 \quad \forall\, v \in \mathcal{K}\}, \tag{5.41}$$

$$\langle \bar{q}, G(\bar{w}) \rangle_{V^*,V} = 0, \tag{5.42}$$

$$\bar{w} \in \mathcal{C}, \quad \langle f'(\bar{w}) + G'(\bar{w})^* \bar{q}, w - \bar{w} \rangle_{W^*,W} \geq 0 \quad \forall\, w \in \mathcal{C}. \tag{5.43}$$

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*

$$\bar{w} \in \mathcal{C}, \quad \langle L_w'(\bar{w}, \bar{q}), w - \bar{w} \rangle_{W^*,W} \geq 0 \quad \forall\, w \in \mathcal{C}. \qquad (5.43)$$

Proof: Under Robinson's regularity condition (5.39), a separation argument can be used to derive (5.41)–(5.43), see [ZK79]. □

A similar result can be shown if $\mathcal{K}$ is a closed convex set instead of a closed convex cone, see [BS98], but then (5.41), (5.42) have a more complicated structure.

5.3.4 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

$$0 \in \text{int} \left(\begin{pmatrix} 0 \\ C(\bar{y}) \end{pmatrix} + \begin{pmatrix} E_y'(\bar{w}) & E_u'(\bar{w}) \\ C'(\bar{y}) & 0 \end{pmatrix} \begin{pmatrix} Y \\ U_{ad} - \bar{u} \end{pmatrix} - \begin{pmatrix} 0 \\ \mathcal{K}_C \end{pmatrix} \right). \tag{5.45}$$

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 5.2.4. □

We show next that the following Slater-type condition implies Robinson's regularity condition (5.45).

Lemma 5.3.6 *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*

$$\begin{aligned}
&E_y'(\bar{w})(\tilde{y} - \bar{y}) + E_u'(\bar{w})(\tilde{u} - \bar{u}) = 0,\\
&C(\bar{y}) + C'(\bar{y})(\tilde{y} - \bar{y}) \in \text{int}(\mathcal{K}_C)
\end{aligned}$$

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 $Z \times V$. $\square$

5.3.5 Applications

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 2.3.7 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?).

Here, $Ay \in L^2(\Omega)$ has to be understood in the sense that $Ay \in (H^1(\Omega))^*$ can be represented in the form $Ay = (f, \cdot)_(L^2(\Omega))$ with some $f \in L^2(\Omega)$.

Then $A : Y \mapsto L^2(\Omega) =: Z$ is bounded and by Theorem 2.3.7 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 5.3.6 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

$$\begin{aligned}
&A\bar{y} + B\bar{u} = 0, \quad \bar{y} \geq 0,\\
&\bar{\lambda} \in \mathcal{K}_C^\circ, \quad \langle \bar{\lambda}, \bar{y} \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = 0,\\
&(\bar{y} - y_d, v)_{L^2(\Omega)} + (\bar{p}, Av)_{L^2(\Omega)} + \langle \bar{\lambda}, v \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = 0 \quad \forall\, v \in Y,\\
&(\alpha\bar{u} - \gamma\bar{p}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall\, u \in U.
\end{aligned}$$

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.

$$\lambda \in \mathcal{K}_C^\circ \quad \Longleftrightarrow$$

$$\langle \lambda, v \rangle_{C(\bar{\Omega})^*, C(\bar{\Omega})} = -\int_\Omega v(x)\, d\mu_\Omega(x) - \int_{\partial\Omega} v(x)\, d\mu_{\partial\Omega}(x) \quad \text{with nonneg. measures } \mu_\Omega, \mu_{\partial\Omega}.$$

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 6

Generalized Newton methods

The aim of this chapter is to lay the ground for fast locally convergent methods that are applicable to the (constrained) optimization of complex systems. Newton's method or variants of it are at the heart of the most efficient methods of nonlinear optimization. Newton's method is applicable to systems of equations

$$G(x) = 0. \tag{6.1}$$

Here $G : X \to Y$ must be sufficiently smooth. The classical Newton's method requires G to be continuously F-differentiable, but as we will see, semismoothness of G is sufficient.

To get the link from optimization problems to the equation (6.1), we note the following:

- If $\bar{w}$ is a local solution of

  $$\min\ f(w)$$

  and $f : W \to \mathbb{R}$ is continuously differentiable, then it satisfies the optimality condition

  $$f'(\bar{w}) = 0.$$

  This results in (6.1) with $G : W \to W^*$, $G(w) = f'(w)$.

- Let $(\bar{y}, \bar{u})$ be a local solution of

  $$\min\ f(y, u) \quad \text{s.t.} \quad E(y, u) = 0$$

  with $f : Y \times U \to \mathbb{R}$ and $E : Y \times U \to Z$ continuously F-differentiable. Assume that $E_y'(\bar{y}, \bar{u})$ is boundedly invertible. Then there exists by Corollary 5.2.4 a Lagrange multiplier (adjoint state) $\bar{p} \in Z^*$ such that the following optimality condition holds:

  $$\begin{aligned} f_y'(\bar{y}, \bar{u}) + E_y'(\bar{y}, \bar{u})^* \bar{p} &= 0, \\ f_u'(\bar{y}, \bar{u}) + E_u'(\bar{y}, \bar{u})^* \bar{p} &= 0, \\ E(\bar{y}, \bar{u}) &= 0. \end{aligned}$$

Defining

$$\begin{aligned} &G : Y \times U \times Z^* \to Y^* \times U^* \times Z, \\ &G(y,u,\mu) = \begin{pmatrix} f'(y,u) + E'_{(y,u)}(y,u)^*\mu \\ E(y,u) \end{pmatrix} = \begin{pmatrix} L'_{(y,u)}(y,u,p) \\ E(y,u) \end{pmatrix}, \end{aligned}$$

we arrive at an operator equation of the form (6.1).

- Let $(\bar{y}, \bar{u})$ be a local solution of

$$\min\ f(y,u) \quad \text{s.t.} \quad E(y,u) = 0, \ \ u \in U_{ad}$$

with a Hilbert space U, $\emptyset \neq U_{ad} \subset U$ closed, convex and $f : Y \times U \to \mathbb{R}$ and $E : Y \times U \to Z$ continuously F-differentiable, . Assume that $E'_y(\bar{y}, \bar{u})$ is boundedly invertible. Then there exists by Corollary 5.2.4 a Lagrange multiplier (adjoint state) $\bar{p} \in Z^*$ such that the following optimality condition holds:

$$\begin{aligned} f'_y(\bar{y},\bar{u}) + E'_y(\bar{y},\bar{u})^*\bar{p} = 0, \\ \bar{u} - P(\bar{u} - \beta(f'_u(\bar{y},\bar{u}) + E'_u(\bar{y},\bar{u})^*\bar{p})) = 0, \\ E(\bar{y},\bar{u}) = 0, \end{aligned}$$

were P is the projection onto U_{ad} and $\beta > 0$ is arbitrary. Note, however, that the projection is Lipschitz-continuous, but non-differentiable.

6.1 A general superlinear convergence result

Consider the operator equation (6.1) with $G : X \to Y$, X, Y Banach spaces.

A general Newton-type method for (6.1) has the form

Algorithm 6.1.1 (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*

$$M_k s = -G(x^k), \tag{6.2}$$

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

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \le \gamma \|d^k\|_X \quad \forall\, k \text{ with } \|d^k\|_X \text{ sufficiently small.} \tag{6.3}$$

2. (x^k) converges q-superlinearly to x^* iff

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \to 0. \tag{6.4}$$

3. (x^k) convergences with q-order $1+\alpha > 1$ iff

$$\|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d^k\|_X \to 0. \tag{6.5}$$

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**:

$$\|M_k^{-1}\|_{Y,X} \le C \quad \forall\, k \ge 0. \tag{6.6}$$

2. **Approximation condition**:

$$\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X = o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \to 0. \tag{6.7}$$

or

$$\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d^k\|_X \to 0. \tag{6.8}$$

We obtain

Theorem 6.1.2 *Consider the operator equation* (6.1) *with $G : X \to Y$, where X and Y are Banach spaces. Let (x^k) be generated by the generalized Newton method (Alg.* 6.1.1*). Then:*

1. *If x^0 is sufficiently close to x^* and* (6.3) *holds then $x^k \to x^*$ q-linearly with rate γ.*
2. *If x^0 is sufficiently close to x^* and* (6.4) *(or* (6.6) *and* (6.7)*) holds then $x^k \to x^*$ q-superlinearly.*
3. *If x^0 is sufficiently close to x^* and* (6.5) *holds (or* (6.6) *and* (6.8)*) then $x^k \to x^*$ q-superlinearly with order $1 + \alpha$.*

Proof: 1. Let $\delta > 0$ be so small that (6.3) holds for all x^k with $\|d^k\|_X < \delta$. Then, for x^0 satisfying $\|x^0 - x^*\|_X < \delta$, we have

$$\begin{aligned}\|x^1 - x^*\|_X &= \|d^1\|_X = \|M_0^{-1}(G(x^* + d^0) - G(x^*) - M_0 d^0)\|_X \le \gamma \|d^0\|_X \\ &= \gamma \|x^0 - x^*\|_X < \delta.\end{aligned}$$

Inductively, let $\|x^k - x^*\|_X < \delta$. Then

$$\begin{aligned}\|x^{k+1} - x^*\|_X &= \|d^{k+1}\|_X = \|M_k^{-1}(G(x^* + d^k) - G(x^*) - M_k d^k)\|_X \\ &\le \gamma \|d^k\|_X = \gamma \|x^k - x^*\|_X < \delta.\end{aligned}$$

Hence, we have

$$\|x^{k+1} - x^*\|_X \le \gamma \|x^k - x^*\|_X \quad \forall\, k \ge 0.$$

2. Fix $\gamma \in (0,1)$ and let $\delta > 0$ be so small that (6.3) 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, (6.4) 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 (6.2) 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 (6.4) (or the conditions (6.6) and (6.7)) 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 [DS83] 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.

6.2 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

$$\|G'(x^k)^{-1}\|_{Y,X} \le C \quad \forall\, k \ge 0.$$

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',

$$\begin{aligned}
\|G(x^* + d^k) - G(x^*) - M_k d^k\|_X &= \|G(x^* + d^k) - G(x^*) - G'(x^* + d^k)d^k\|_X \\
&\le \|G(x^* + d^k) - G(x^*) - G'(x^*)d^k\|_X + \|(G'(x^*) - G'(x^* + d^k))d^k\|_X \\
&= o(\|d^k\|_X) + \|G'(x^*) - G'(x^* + d^k)\|_{X,Y}\|d^k\|_X = o(\|d^k\|_X).
\end{aligned}$$

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 6.2.1 *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.* 6.1.1 *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 6.2.2 *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.*

6.3 Semismooth Newton methods

If G is nonsmooth, the question arises if a suitable substitute for G' can be found. We follow [Ul01, Ul03] here; a related approach can be found in [HIK03]. Thinking at subgradients of convex functions, which are set-valued, we consider set-valued generalized differentials $\partial G : X \rightrightarrows \mathcal{L}(X, Y)$. 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 [Ul01, Ul03]:

Definition 6.3.1 (Semismoothness) *Let $G : X \to Y$ be a continuous operator between Banach spaces. Furthermore, let be given the set-valued mapping $\partial G : X \rightrightarrows \mathcal{L}(X, Y)$ 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 6.3.2 *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:

$$\begin{aligned}
&\|G(x + d) - G(x) - G'(x + d)d\|_Y \leq \\
&\quad \leq \|G(x + d) - G(x) - G'(x)d\|_Y + \|G'(x)d - G'(x + d)d\|_Y \\
&\quad \leq o(\|d\|_X) + \|G'(x) - G'(x + d)\|_{X,Y} \|d\|_X = o(\|d\|_X).
\end{aligned}$$

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 For locally Lipschitz-continuous functions $G : \mathbb{R}^n \to \mathbb{R}^m$, the standard choice for ∂G is Clarke's generalized Jacobian:

$$\partial^{cl} G(x) = \text{conv}\left\{M \;:\; x^k \to x,\ G'(x^k) \to M,\ G \text{ differentiable at } x^k\right\}. \tag{6.9}$$

This definition is justified since G' exists almost everywhere on $\mathbb{R}^n$ by Rademacher's theorem (which is a deep result).

Remark The classical definition of semismoothness for functions $G : \mathbb{R}^n \to \mathbb{R}^m$ [Mi77, QS93] is equivalent to $\partial^{cl}G$-semismoothness, where $\partial^{cl}G$ is Clarke's generalized Jacobian defined in (6.9), in connection with directional differentiability of G.

Next, we give a concrete example of a semismooth function:

Example Consider $\psi : \mathbb{R} \to \mathbb{R}$, $\psi(x) = P_{[a,b]}(x)$, then Clarke's generalized derivative is

$$\partial^{cl}\psi(x) = \begin{cases} 0 & x < a \text{ or } x > b, \\ 1 & a < x < b, \\ \text{conv}\{0,1\} = [0,1] & x = a \text{ or } x = b. \end{cases}$$

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 6.3.2, 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 [Ul01]:

Theorem 6.3.3 *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

$$\|h(d)\|_Y = \|G_2(x + d) - G_2(x)\|_Y \leq L_2 \|d\|_Y.$$

Hence, for $M_1 \in \partial G_1(G_2(x + d))$ and $M_2 \in \partial G_2(x + d)$, we obtain

$$\begin{aligned}
&\|G_1(G_2(x + d)) - G_1(G_2(x)) - M_1 M_2 d\|_Z = \\
&\quad = \|G_1(y + h(d)) - G_1(y) - M_1 h(d) + M_1(G_2(x + d) - G_2(x) - M_2 d)\|_Z \\
&\quad \leq \|G_1(y + h(d)) - G_1(y) - M_1 h(d)\|_Y + \|M_1\|_{Y,Z} \|G_2(x + d) - G_2(x) - M_2 d\|_Y
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{M \in \partial G(x+d)} \|G(x + d) - G(x) - Md\|_Z \\
&\quad \leq \sup_{M_1 \in \partial G_1(y+h(d))} \|G_1(y + h(d)) - G_1(y) - M_1 h(d)\|_Y \\
&\quad\quad + C \sup_{M_2 \in \partial G_2(x+d)} \|G_2(x + d) - G_2(x) - M_2 d\|_Y \\
&\quad = o(\|h(d)\|_Y) + o(\|d\|_X) = o(\|d\|_X).
\end{aligned}$$

□

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

$$\|M^{-1}\|_{Y,X} \le C \quad \forall\, M \in \partial G(x) \;\; \forall\, x \in X,\; \|x - x^*\|_X < \delta. \tag{6.10}$$

Under these two assumptions, the following generalized Newton method for semismooth operator equations is q-superlinearly convergent:

Algorithm 6.3.4 (Semismooth Newton's method)

0. *Choose $x^0 \in X$ (sufficiently close to the solution x^*.)*

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

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 6.1.2:

Theorem 6.3.5 *Let $G : X \to Y$ be continuous and ∂G-semismooth at a solution x^* of* (6.1)*. Furthermore, assume that the regularity condition* (6.10) *holds. Then there exists $\delta > 0$ such that for all $x^0 \in X$, $\|x^0 - x^*\|_X < \delta$, the semismooth Newton method (Alg.* 6.3.4*) 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 (6.10) implies (6.6) as long as x^k is close enough to x^*. Furthermore, the semismoothness of G at x^* ensures the q-superlinear approximation property (6.7).

In the case of α-order semismoothness, the approximation property with order $1 + \alpha$ holds.

Therefore, Theorem 6.1.2 yields the assertions. $\Box$

6.4 Semismooth Newton methods in function spaces

In section 6.3 we introduced the concept of semismoothness for nonsmooth operators and developed superlinearly convergent generalized Newton methods for semismooth operator equations. We will now show that optimality conditions can be rewritten as semismooth equations.

Let $\Omega \subset \mathbb{R}^n$ be measurable with $0 < |\Omega| < \infty$. We consider the problem

$$\min_{(y,u)\in Y\times L^2(\Omega)} f(y,u) \quad E(y,u) = 0, \quad a \le u \le b \quad \text{a.e. on } \Omega.$$

The optimality conditions are

$$\begin{aligned} f'_y(\bar y,\bar u) + E'_y(\bar y,\bar u)^*\bar p = 0,\\ \bar u - P_{[a,b]}(\bar u - \beta(f'_u(\bar y,\bar u) + E'_u(\bar y,\bar u)^*\bar p)) = 0,\\ E(\bar y,\bar u) = 0, \end{aligned}$$

were $P_{[a,b]}$ is the projection onto U_{ad} and $\beta > 0$ is arbitrary. or alternatively, the reduced problem

$$\min_{u\in L^2(\Omega)} \hat f(u) \quad a \le u \le b \quad \text{a.e. on } \Omega$$

with $\hat f : L^2(\Omega) \to \mathbb{R}$ 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 without restriction the problem

$$\min_{u\in L^2(\Omega)} \hat f(u) \quad l \le u \le r \quad \text{a.e. on } \Omega \tag{6.11}$$

with constants $l < r$. Let $U = L^2(\Omega)$ and $\mathcal{S} = \{u \in L^2(\Omega) : l \le u \le r\}$. 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

$$u \in \mathcal{S}, \quad (\nabla \hat f(u), v - u)_{L^2} \ge 0 \quad \forall\, v \in \mathcal{S}.$$

We now use the projection P_S onto $\mathcal{S}$, which is given by

$$P(v)(x) = P_{[l,r]}(v(x)), \quad x \in \Omega.$$

Then the optimality conditions can be written as

$$\Phi(u) := u - P(u - \beta\nabla \hat f(u)) = 0, \tag{6.12}$$

where $\beta > 0$ is arbitrary, but fixed. Note that, since P conincides with the pointwise projection onto $[l, r]$, we have

$$\Phi(u)(x) = u(x) - P_{[l,r]}(u(x) - \beta\nabla \hat f(u)(x)).$$

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 $P_{[l,r]}(v(\cdot))$ can be defined.

In fact, the following can be proved:

Theorem 6.4.1 *Let $\Omega \subset \mathbb{R}^n$ be bounded and $q \in (2, \infty)$. Then the operator*

$$\Psi : L^q(\Omega) \to L^2(\Omega), \quad \Psi(u)(x) = P_{[l,r]}(u(x)),$$

is $\partial\Psi$-semismooth with

$$\partial\Psi(u) = \{g \cdot I \ : \ g(x) = 1 \text{ if } u(x) \in (l,r),\ g(x) = 0 \text{ if } u(x) \notin [l,r], g(x) \in [0,1] \text{ if } u(x) \in \{l,r\}\}\,.$$

Proof: Let $u, s \in L^q(\Omega)$ be arbitrary. Let $gI \in \partial\Psi(u+s)$ be arbitrary.

If $u(x) \notin \{l,r\}$ and $|s(x)| < \text{dist}(u(x), \{l,r\})$, then $t \mapsto \Psi(u+ts)(x)$, $t \in [0,1]$ is linear and thus we have

$$\Psi(u+s)(x) - \Psi(u)(x) - g(x)s(x) = 0.$$

If $u(x) = l$ and $s(x) < r - l$ or $u(x) = r$ and $s(x) > l - r$ then again

$$\Psi(u+s)(x) - \Psi(u)(x) - g(x)s(x) = 0.$$

In all other cases we have

$$|\Psi(u+s)(x) - \Psi(u)(x) - g(x)s(x)| \le 2|s(x)|.$$

Hence, we have for all $M \in \partial\Psi(u+s)$ and all $\epsilon > 0$

$$\begin{aligned}\|\Psi(u+s) - \Psi(u) - Ms\|_{L^2} &\le \big\|2s\, 1_{\{x:\ |s(x)|<\max(r-l,\text{dist}(u(x),\{l,r\}))\}}\big\|_{L^2} \\ &\le \|2s\|_{L^q} \big\|1_{\{x:\ |s(x)|<\max(r-l,\text{dist}(u(x),\{l,r\}))\}}\big\|_{L^{2q/(q-2)}}.\end{aligned}$$

Now $\|s\|_{L^q} \to 0$ implies $s \to 0$ almost everywhere. Therefore

$$\big\|1_{\{x:\ |s(x)|<\max(r-l,\text{dist}(u(x),\{l,r\}))\}}\big\|_{L^{2q/(q-2)}} \to 0$$

as $\|s\|_{L^q} \to 0$. $\square$

We now return to the operator Φ defined in (6.12). To be able to prove the semismoothness of $\Phi : L^2 \to L^2$ definied in (6.12), we need some kind of smoothing property of the mapping

$$u \mapsto u - \beta\nabla\hat{f}(u).$$

Therefore, we assume that ∇f has the following structure:

$$\begin{aligned}
&\text{There exist } \beta > 0 \text{ and } q > 2 \text{ such that}\\
&\nabla \hat{f}(u) = \alpha u + B(u),\\
&B : L^2(\Omega) \to L^q(\Omega) \text{ continuously F-differentiable.}
\end{aligned} \tag{6.13}$$

This assumption implies that B is locally Lipschitz continuous. In fact,

$$\begin{aligned}
\|B(u) - B(v)\|_{L^q} &\leq \int_0^1 \|B'(v + t(u-v))(u-v)\|_{L^q}\, dt\\
&\leq \int_0^1 \|B'(v + t(u-v))\|_{L^2,L^q}\, dt\, \|u - v\|_{L^2}.
\end{aligned}$$

Remark This structure is met by many optimal control problems, see, e.g., the optimal heating problem in the second part of section 4.3. There, we obtained

$$\nabla \hat{f}(u) = \alpha u - \gamma p(u),$$

with $\alpha > 0$, $\gamma \in L^\infty(\Omega)$ and $L^2(\Omega) \ni u \mapsto p(u) \in H^1(\Omega)$ continuous affine linear. Thus, using the Sobolev embedding theorems, we obtain that for appropriate $q > 2$, the operator

$$B : u \mapsto -\gamma p(u)$$

defines a continuous affine linear mapping from L^2 to L^q as required.

If we now choose $\beta = 1/\alpha$, then we have

$$\Phi(u) = u - P_{[l,r]}(u - (1/\alpha)(\alpha u + B(u))) = u - P_{[l,r]}(-(1/\alpha)B(u)).$$

Example: Distributed control of elliptic equations

We consider for example

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

where

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

We choose as above

$$U = L^2(\Omega), \quad Y = H_0^1(\Omega), \quad Z = Y^*.$$

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

Let $(\bar{y}, \bar{u}) \in Y \times U$ be an optimal solution. Then by Corollary 5.2.4 and (5.22), (5.23) the optimality system in the form (5.18)–(5.20) reads

$$\begin{aligned} a(\bar{y}, v) - (\gamma \bar{u}, v)_{L^2(\Omega)} &= 0 \quad \forall\, v \in Y, \\ (\bar{y} - y_d, v)_{L^2\Omega} + a(\bar{p}, v) &= 0 \quad \forall\, v \in Y, \\ a \le \bar{u} \le b, \quad (\alpha \bar{u} - \gamma \bar{p}, u - \bar{u})^2_L(\Omega) &\ge 0, \quad \forall\, u \in U,\ a \le u \le b. \end{aligned}$$

Moreover, we have

$$\nabla \hat{f}(u) = \alpha u - \gamma p(u),$$

where $p = p(u) \in Y$ solves the adjoint equation

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

□

We obtain:

Theorem 6.4.2 *Consider the problem* (6.11) *with* $l < r$ *and let* $\hat{f} : L^2(\Omega) \to L^2(\Omega)$ *satisfy condition* (6.13). *Then, for* $\beta = 1/\alpha$, *the operator* Φ *in the reformulated optimality conditions* (6.12) *is* $\partial\Phi$*-semismooth with*

$$\begin{aligned} &\partial\Phi : L^2(\Omega) \rightrightarrows \mathcal{L}(L^2(\Omega), L^2(\Omega)), \\ &\partial\Phi(u) = \Big\{ M\ ;\ M = I + \frac{g}{\alpha} \cdot B'(u),\ g \in L^\infty(\Omega), \\ &\qquad\qquad g(x) \in \partial^{cl} P_{[l,r]}(-1/\alpha B(u)(x)) \text{ for a.a. } x \in \Omega \Big\}. \end{aligned}$$

Here,

$$\partial P_{[l,r]}(t) \begin{cases} 0 & t < l \text{ or } t > r, \\ 1 & l < t < r, \\ [0,1] & t = l \text{ or } t = r. \end{cases}$$

Proof: By the chain rule, the smoothness of $B : L^2 \to L^q$ and the semismoothness of $\Psi : L^q \to L^2$, $\Psi(u)(x) = P_{[l,r]}(u(x))$, we see that Φ is semismooth with respect to the stated generalized differential. □

For the applicability of the semismooth Newton method (Alg. 6.3.4) 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 [Ul01, Ul01a].

Chapter 7

Globalization for problems with simple constraints

We develop now globalized descent methods for simply constrained problems of the form

$$\min f(w) \quad \text{s.t.} \quad w \in S \tag{7.1}$$

with W a Hilbert space, $f : W \to \mathbb{R}$ continuously F-differentiable, and $S \subset W$ closed and convex. Optimality conditions for this type of problems have already been considered in 5.1.

Example 7.0.3 *A scenario 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:

$$P_S(w)(x) = P_{[a(x),b(x)]}(w(x)) = \max(a(x), \min(w(x), b(x))).$$

In the case of control constraints, the globalization techniques of this chapter can be combined with the semismooth Newton method of the last chapter to obtain a globally convergent method that converges locally superlinearly.

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 $\{w^k + \sigma s^k \,:\, \sigma \geq 0\}$, we perform a line search along the projected path

$$\left\{P_S(w^k + \sigma s^k) \,:\, \sigma \geq 0\right\},$$

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 7.0.4 *Consider*

$$S = \left\{w \in \mathbb{R}^2 \,:\, w_1 \geq 0,\ w_1 + w_2 \geq 3\right\}, \quad f(w) = 5w_1^2 + w_2^2.$$

Then, at $w^k = (1,2)^T$*, we have* $\nabla f(w^k) = (10,4)^T$*. Since* f *is convex quadratic with minimum* $\bar{w} = 0$*, 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 $\sigma \geq 0$*, 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.*

7.1 Projected gradient method

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 [Ke99] for a detailed discussion. Here, we only consider the projected gradient method.

Algorithm 7.1.1 (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 such that $f(P_S(w^k + \sigma_k s^k)) < f(w^k)$.*
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 7.1.2 *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

$$\begin{aligned} -\sigma(\nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &= (w_\sigma^k - w^k, P_S(w_\sigma^k) - w^k)_W \\ &= (w_\sigma^k - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &= (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &\quad + \underbrace{(w_\sigma^k - P_S(w_\sigma^k), P_S(w_\sigma^k) - P_S(w^k))_W}_{\ge 0 \text{ by Lemma 5.1.2, b)}} \\ &\ge (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &= \|P_S(w_\sigma^k) - w^k\|_W^2. \end{aligned}$$

Next, we use

$$\|v_\sigma^k - w^k\|_W \le \|P_S(w_\sigma^k) - w^k\|_W.$$

Hence,

$$\begin{aligned}(\nabla f(v_\sigma^k) - \nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &\le \|\nabla f(v_\sigma^k) - \nabla f(w^k)\|_W \|P_S(w_\sigma^k) - w^k\|_W \\ &\le L\|v_\sigma^k - w^k\|_W^\alpha \|P_S(w_\sigma^k) - w^k\|_W \\ &\le L\|P_S(w_\sigma^k) - w^k\|_W^{1+\alpha}.\end{aligned}$$

□

We now consider the following

Projected Armijo rule:

Choose the maximum $\sigma_k \in \{1, 1/2, 1/4, \ldots\}$ for which

$$f(P_S(w^k + \sigma_k s^k)) - f(w^k) \le -\frac{\gamma}{\sigma_k}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^2.$$

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 (7.1) are

$$w \in S, \quad (\nabla f(w), v - w)_W \ge 0 \quad \forall\, v \in S.$$

By Lemma 5.1.2, 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 7.1.3 *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 successfully.*

Proof: We proceed as in the proof of Lemma 7.1.2 and obtain (we have not assumed Hölder continuity of ∇f here)

$$f(P_S(w_\sigma^k)) - f(w^k) \le \frac{-1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W).$$

It remains to show that, for all small $\sigma > 0$,

$$\frac{\gamma - 1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W) \le 0$$

But this follows easily from (Lemma 5.1.2 e)):

$$\frac{\gamma - 1}{\sigma}\|P_S(w_\sigma^k) - w^k\|_W^2 \le \underbrace{(\gamma - 1)\|p(w^k)\|_W}_{<0}\|P_S(w_\sigma^k) - w^k\|_W.$$

□

Theorem 7.1.4 *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 7.1.1 with the projected Armijo rule and assume that $f(w^k)$ is bounded below. Furthermore, let ∇f be α-order Hölder continuous on*

$$N_0^\rho = \{w + s \,:\, f(w) \le f(w^0),\ \|s\|_W \le \rho\}$$

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 $\|p^k\|_W \ge \varepsilon$ for all $k \in K$.

By construction we have that $f(w^k)$ is monotonically decreasing and by assumption the sequence is bounded below. For all $k \in K$, we obtain

$$f(w^k) - f(w^{k+1}) \ge \frac{\gamma}{\sigma_k}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^2 \ge \gamma\sigma_k\|p^k\|_W^2 \ge \gamma\sigma_k\varepsilon^2,$$

where we have used the Armijo condition and Lemma 5.1.2 e). This shows $(\sigma_k)_K \to 0$ and $(\|P_S(w^k + \sigma_k s^k) - w^k\|_W)_K \to 0$.

For large $k \in K$ we have $\sigma_k \le 1/2$ and therefore, the Armijo condition did not hold for the step size $\sigma = 2\sigma_k$. Hence,

$$\begin{aligned} -\frac{\gamma}{2\sigma_k}\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 &\le f(P_S(w^k + 2\sigma_k s^k)) - f(w^k) \\ &\le -\frac{1}{2\sigma_k}\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 + L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}. \end{aligned}$$

Here, we have applied Lemma 7.1.2 and the fact that by Lemma 5.1.2 e)

$$\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W \leq 2\|P_S(w^k + \sigma_k s^k) - w^k\|_W \stackrel{K\ni k\to\infty}{\longrightarrow} 0.$$

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,

$$(1-\gamma)\varepsilon \leq L\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{\alpha} \leq L2^{\alpha}\|P_S(w^k + \sigma_k s^k) - w^k\|_W^{\alpha} \stackrel{K\ni k\to\infty}{\longrightarrow} 0.$$

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 [Ke99, Be99]. In an L^2 setting, the superlinear convergence of projected Newton methods was investigated by Kelley and Sachs in [KS94].

Bibliography

[Ad75] R.A. Adams: *Sobolev spaces*. Academic press, 1975.

[Al99] H.W. Alt: *Lineare Funktionalanalysis*. Springer, 1999.

[Be99] D. P. Bertsekas, *Nonlinear Programming (2nd edition)*, Athena Scientific, 1999.

[BS98] J.F. Bonnans, A. Shapiro: Optimization problems with perturbations: A guided tour. *SIAM Rev.* 40, pp. 228–264, 1998. Springer, 1999.

[DS83] J.E. Dennis, R.B. Schnabel: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia, 1996.

[Ev98] L. C. Evans: *Partial Differential Equations*. American Mathematical Society, 1998.

[HIK03] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semi-smooth Newton method. *SIAM J. Optim.*, 13(3):865888, 2003.

[Jo98] J. Jost: *Postmodern Analysis*. Springer, 1998.

[Ke99] C. T. Kelley, *Iterative methods for optimization*, SIAM, Philadelphia, 1999.

[KS94] C. T. Kelley and E. W. Sachs, Multilevel algorithms for constrained compact fixed point problems, *SIAM J. Sci. Comput.*, 15 (1994), pp. 645–667.

[KS80] D. Kinderlehrer, G. Stampacchia: *Introduction to Variational Inequalities and their Applications*, Academic Press, 1980.

[Mi77] R. Mifflin: Semismooth and semiconvex functions in constrained optimization, *SIAM J. Control Optim.*, 15 (1977) 957972.

[QS93] L. Qi, J. Sun: A nonsmooth version of Newtons method, *Math. Programming*, 58 (1993), 353367.

[ReRo93] M. Renardy, R. C. Rogers: *An Introduction to Partial Differential Equations*. Springer, 1993.

[Ro76] 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.

[Tr05] F. Tröltzsch: *Optimale Steuerung partieller Differentialgleichungen*. Vieweg, 2005.

[Ul01] M. Ulbrich: Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces, Habilitation, Fakultät für Mathematik, Technische Universität München, June 2001.

[Ul01a] M. Ulbrich: On a Nonsmooth Newton Method for Nonlinear Complementarity Problems in Function Space with Applications to Optimal Control, in M. C. Ferris, O. L. Mangasarian, and J.-S. Pang (eds.), Complementarity: Applications, Algorithms and Extensions, Kluwer Academic Publishers, 2001, pp. 341-360.

[Ul03] M. Ulbrich: Semismooth Newton Methods for Operator Equations in Function Spaces, *SIAM J. Optim.*, 13 (2003), pp. 805-842.

[Wl71] J. Wloka: *Funktionalanalysis unf ihre Anwendungen*. De Gruyter, 1971.

[Yo80] K. Yosida: *Functional Analysis*. Springer, 1980.

[ZK79] J. Zowe, S. Kurcyusz: Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optimization* 5, pp. 49–62, 1979.