**Lecture Notes, 285J**

---

# Infinite-dimensional Optimization and Optimal Design

Martin Burger
Department of Mathematics, UCLA
martinb@math.ucla.edu
http://www.math.ucla.edu/~martinb/

---

Fall 2003

# Chapter 1

# Introduction

The aim of this course is to provide an overview of modern methods and applications of infinite-dimensional optimization problems. By the term infinite-dimensional optimization we mean the minimization of a functional over an infinite-dimensional space (or subset of an infinite-dimensional space). There are two prototypes of such problems:

(i) The optimization variable is a function or even several functions. Important examples of such problems appear in the calculus of variations, in optimal design and optimal control.

(ii) The optimization variable is a shape or even the shape and topology of an object. Such problems appear in a variety of applications and are usually denoted by the phrases "shape and topology optimization".

Infinite-dimensional optimization problems incorporate some fundamental differences to (finite-dimensional) nonlinear programming problems, but they also share many common properties, as we shall see during this course.

## 1.1 Abstract Formulation

In the most general form, we can write an optimization problem in a topological space $\mathcal{U}$ endowed with some topology $\mathcal{T}$ as

$$J(u) \to \min_{u \in \mathcal{C}}$$

where $\mathcal{C} \subset \mathcal{U}$ and $J : \mathcal{C} \to \overline{\mathbb{R}}$ is the objective functional. By extending the objective functional to U via

$$\tilde{J}(u) := \begin{cases} J(u) & \text{if } u \in \mathcal{C} \\ +\infty & \text{else} \end{cases}$$

we can rewrite this problem as

$$\tilde{J}(u) \to \min_{u \in \mathcal{U}}.$$

The latter formulation is sometimes advantageous for theoretical purpose.

Since it is sufficient for most applications, we restrict our attention mostly to the case when $\mathcal{U}$ is a metric space with metric $d$, in several cases even to Banach or Hilbert spaces.

## 1.2 Examples

In the following we present some typical examples of infinite-dimensional optimization problems, many of them appearing in practical applications.

### 1.2.1 Minimal surfaces

A classical mathematical problem, whose solution is visualized in everyday life is the so-called "minimal surface problem" (or Plateau problem). It consists in seeking surfaces with prescribed boundary that minimize area. Such structures are obtained e.g. in soap bubbles.

In the above setting $\mathcal{U}$ is a space of measurable subsets of $\mathbb{R}^n$, $\mathcal{C}$ is an appropriate subset of $\mathcal{U}$ containing those surfaces with the prescribed boundary, and the functional $J$ is given by

$$J : \mathcal{U} \to \mathbb{R}, \;\; u \to \lambda^n(u),$$

where $\lambda^n$ denotes the $n$-dimensional Lebesgue measure.

The minimal surface problem for graphs consists in minimizing the integral

$$J(u) := \int_\Omega \sqrt{1 + |\nabla u|^2}\, dx,$$

over a set $\mathcal{C}$ of functions $u : \Omega \to \mathbb{R}$ with $u|_{\partial\Omega} = \Phi$. An appropriate function space for u would be the space $\mathcal{U} = BV(\Omega)$ of functions of bounded variation.

### 1.2.2 A Boundary Control Problem

Optimal control problems for ordinary and partial differential equations appear in many applications, an extensive literature covering those problems can be found. We consider a simple example in boundary control for the heat equation. The goal is to achieve a temperature distribution $\theta : \Omega \to \mathbb{R}$ in a room $\Omega$ by controlling the temperature at a part of its boundary over some time (e.g. by a radiator).

The easiest mathematical model for this problem is to use a least-squares functional of the form

$$J(u,v) = \int_\Omega |v(x,T) - \theta(x)|^2 dx, \tag{1.1}$$

where $v$ solves the heat equation

$$\frac{\partial v}{\partial t} = \Delta v \quad \text{in } \Omega \times (0,T) \tag{1.2}$$

with initial values

$$v(x,0) = v_0(x) \quad \text{in } \Omega \tag{1.3}$$

and boundary values

$$v = u \quad \text{on } \Gamma \times (0,T) \tag{1.4}$$

$$\frac{\partial v}{\partial u} = 0 \quad \text{on } (\Omega \backslash \Gamma) \times (0,T), \tag{1.5}$$

i.e. we assume the remaining part of the boundary is isolated.

In the above setting, we can interpret this problem as an optimization problem in the Hilbert space

$$\mathcal{U} = L^2(0,T; H^1(\Omega)) \times L^2(0,T; H^{\frac{1}{2}}(\Gamma))$$

with $\mathcal{C}$ being the subset

$$\mathcal{C} = \{(u, v) \in \mathcal{U} \mid (u, v) \text{ satisfies } (1.2) - (1.5)\}$$

Noticing that the parabolic initial-boundary value problem (1.2) - (1.5) admits a unique solution $v = v(u)$ for each $u$, we can rewrite (1.1) - (1.5) as an unconstrained problem

$$\tilde{J}(u) := J(u, v(u)) \to \min_u$$

where $v(u)$ is implicitly defined by (1.2) - (1.5). We shall discuss such elimination approaches in the section on PDE-constrained optimization.

### 1.2.3 Elliptic PDEs

It is well-known that second-order elliptic partial differential equations are the solution of a variational problem if they are symmetric. E.g., the solution $u$ of

$$\begin{aligned} -div(a\nabla u) + cu &= f \quad \text{in } \Omega \\ \text{subject to the boundary condition} & \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{1.6}$$

is the unique minimizer of the optimization problem.

$$J(u) := \frac{1}{2}\left(\int_\Omega a|\nabla u|^2 + cu^2\right) dx - \int_\Omega fu \; dx \to \min_{u \,\in\, \mathcal{U}}, \tag{1.7}$$

where $\mathcal{U} = H_0^1(\Omega)$. We shall see later that the equation (1.6) can be interpreted as the first-order optimality condition corresponding to (1.7).

### 1.2.4 Image Processing

Image restoration and segmentation is based on variational principles. For denoising blocky images, it is now standard to minimize a total variation functional (introduced by Rudin, Osher, Fatemi) of the form

$$J(u) := \int_\Omega |u - v|^2 dx + \alpha \int_\Omega |\nabla u| dx,$$

where $v$ is the noisy image and $\alpha \in \mathbb{R}^+$ is a regularization parameter. The appropriate function space $\mathcal{U}$ is again $BV(\Omega)$.

For the segmentation of noisy images, it is standard to minimize the Mumford-Shah functional

$$J(u, \Gamma) := \int_\Omega |u - v|^2 dx + \alpha \int_{\Omega\backslash\Gamma} |\nabla u|^2 dx + \beta \int_\Gamma 1 \; ds,$$

where $\Gamma$ is a disjoint union of curves in $\Omega \subset \mathbb{R}^2$ and $u$ is a function in $H^1(\Omega\backslash\Gamma)$. Again, $v$ denotes the noisy image and $\alpha, \beta \in \mathbb{R}^+$ are regularization parameters. The Mumford-Shah functional is a combination of the two prototype problems, since we are looking for a function $u$ as well as for a shape $\Gamma$ (the set of discontinuities of $u$).

### 1.2.5 Semiconductor Design

As an example of the problems appearing in the optimal design of semiconductor devices, we consider the optimization of the doping profile in a unipolar device. Design goals are typically related to the outflow current $I$ over some contact $\Gamma_1 \subset \partial\Omega$, which is given by

$$I = \int_{\Gamma_1} \mu_n(\nabla n - n\nabla V).d\nu,$$

where $\mu_n$ is the electron mobility, $n$ is the electron density and $V$ is the electric potential. A mathematical model relating $n$ and $V$ to the doping profile $C$ is given by the drift-diffusion equations, i.e.

$$\begin{aligned} \lambda^2 \Delta V &= n - C \quad \text{in } \Omega \\ div(\mu_n(\nabla n - n\nabla V)) &= 0 \quad \text{in } \Omega \end{aligned}$$

subject to suitable boundary conditions on $\partial\Omega$. Again, we can implicitly define $n = n(C)$ and $V = V(C)$ and consequently

$$I = I(n, V) = I(n(C), V(C))$$

and write the optimal design problem of the form

$$J(C) := \tilde{J}(I(n(C), V(C))) \to \min_C$$

.

### 1.2.6 Structural Optimization

Problems in structural optimization usually deal with the maximization of stiffness of a material under constrained volume or the minimization of volume under some bounds on the stiffness. We shall describe a typical case of the latter problem in the following, namely volume minimization under local constraints on the stress and on the displacement. A mathematical model for this problem can be obtained as follows: assume that we want to mix two materials $M_1, M_2$ inside the fixed domain $\Omega$. We assume that a part $\Gamma_1$ of $\partial\Omega$ is fixed, whereas a load $g$ is applied to some part $\Gamma_2 \subset \partial\Omega$. If $u : \overline{\Omega} \to \mathbb{R}^3$ and $\sigma : \overline{\Omega} \to \mathbb{R}^{3\times 3}$ denote the elastic displacement vector and the stress tensor, respectively, then these conditions can be formulated as

$$\begin{aligned} u &= 0 \quad \text{on } \Gamma_1, \\ \sigma.n &= g \quad \text{on } \Gamma_2, \end{aligned}$$

where $n$ denotes the outer unit normal. For simplicity we assume that $\partial\Omega = \overline{\Gamma_1} \cup \overline{\Gamma_2}$.

Inside the domain we assume no further forces, so that elastic equilibrium gives the relation

$$div(\sigma) = 0 \quad \text{in } \Omega.$$

Since $\overline{\Omega} = \overline{M_1} \cup \overline{M_2}$ we obtain a constitutive relation of the form

$$\sigma = \begin{cases} \sigma_1(\epsilon) & \text{in } M_1 \\ \sigma_2(\epsilon) & \text{in } M_2 \end{cases}$$

where $\sigma_k$ denotes the stress-strain relation for material $M_k, k = 1, 2$, and $\epsilon$ denotes the stress tensor. In the case of two isotropic linear elastic materials, we obtain

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

and

$$\sigma_k = \lambda_k \epsilon + 2\mu_k \ \mathrm{tr}(\epsilon) I,$$

where $\lambda_k, \mu_k$ are the Lame-parameters of material $M_k$.

The optimization problem can then be formulated as (assuming that $M_1$ is the "expensive" material)

$$\int_{M_1} 1 \ dx \rightarrow \min_{(\mu_k, u_k, \sigma_k, \epsilon_k)_{k=1,2}}$$

subject to the above equilibrium and constitutive relations, a local displacement constraint

$$u^{\min} \leq u \leq u^{\max} \quad \text{in } \Omega$$

and a stress constraint, usually formulated in terms of the von-Mises stress $\sigma_{VM}$ (the largest eigenvalue in magnitude of $\sigma$)

$$\sigma^{\min} \leq \sigma_{VM} \leq \sigma^{\max} \quad \text{in } \Omega.$$

This structural optimization problem is a topology optimization problem, since we want to design the whole topology of material $M_1$. We will discuss such problems at the end of this course.

# Chapter 2

# Infinite-dimensional Optimization

In this chapter we shall discuss basic notions and properties of nonlinear optimization problems by general functional-analytic techniques. We shall adress some fundamental issues such as notions of solutions, their existence, and concepts for derivatives.

## 2.1 Notions of Solutions

Let $\mathcal{U}$ be a topological space and let $J : \mathcal{U} \to \overline{\mathbb{R}}$ be a functional. We are looking for solutions of the problem

$$J(u) \to \min_{u \in \mathcal{U}}.$$

Note that in the case of constraints $u \in C$ we can always set

$$J(u) := +\infty \quad \text{if } u \in C,$$

and hence a minimum can be attained only in $\mathcal{U} \backslash C$ if $J$ is a proper functional, i.e.

$$\exists u_0 \in \mathcal{U} : J(u_0) < +\infty.$$

We distinguish two kinds of solutions, namely local and global minima.

**Definition 2.1.** A point $u \in \mathcal{U}$ is called

(i) *local minimizer*, if there exists an open set $\mathcal{V} \subset \mathcal{U}$ such that $u \in \mathcal{U}$ and

$$J(u) \leq J(v), \quad \forall\, v \in \mathcal{V}.$$

(ii) *global minimizer*, if

$$J(u) \leq J(v), \quad \forall\, v \in \mathcal{U}. \tag{2.1}$$

As we shall see later, local minimizers of smooth functionals can be characterized by conditions on the first and second derivative. For global optima, there is in general no other condition than (2.1). Of course, each global minimizer is also a local one, but in general not vice versa.

## 2.2 Existence of Solutions

In order to obtain existence of solutions for a general optimization problem, two basic properties are needed: *compactness* and *lower semicontinuity*. We recall the definition of the latter.

**Definition 2.2.** Let $(\mathcal{U}, \mathcal{T})$ be a topological space, and let $J : \mathcal{U} \to \overline{\mathbb{R}}$. The functional $f$ is called *lower semicontinuous* at $u \in \mathcal{U}$ if

$$J(u) \leq \sup_{\mathcal{V} \in \mathcal{T}} \inf_{v \in \mathcal{V}} J(v).$$

If $\mathcal{U}$ is a metric space, this definition is equivalent to a characterization by sequences. The functional $J$ is lower semicontinuous at $u$ if

$$J(u) \leq \liminf_{k \to \infty} J(u_k) \tag{2.2}$$

for all sequences $u_k$ converging to $u$.

**Theorem 2.3.** *Let $J : \mathcal{U} \to \overline{\mathbb{R}}$ be lower semicontinuous and let the level set*

$$\{u \in \mathcal{U} \mid J(u) \leq M\}$$

*be non-empty and compact for some $M \in \mathbb{R}$. Then there exists a global minimum of*

$$J(u) \to \min_{u \in \mathcal{U}}$$

*Proof.* Let $\alpha = \inf_{u \in \mathcal{U}} J(u)$. Then there exists a sequence $(u_k)$ such that $J(u_k) \to \alpha$. For $k$ sufficiently large, we must have $J(u_k) \leq M$ and hence, $(u_k)$ is contained in a compact set. Consequently, there exists a subsequence $(u_{k_l})$ such that $u_{k_l} \to \tilde{u}$ for some $\tilde{u} \in \mathcal{U}$. Due to the lower semicontinuity, we obtain

$$\alpha \leq J(\tilde{u}) \leq \liminf_{k \to \infty} J(u_k) = \alpha.$$

Thus, $\tilde{u}$ is a global minimizer. □

From the above proof, one sees that in order to obtain the existence of a minimum, the "sequential lower semicontinuity" defined by (2.2) is sufficient even in the case of a topological space that is non-metrizable.

**Corollary 2.4.** *Under the conditions of Theorem 2.3, the set of global minimizers $G$ is compact.*

*Proof.* Since all global minima lie in the level set $\{J(u) \leq M\}$, we obtain pre-compactness. The closedness of this set follows from the lower semicontinuity, since for each $u$ in the closure of $G$ we have

$$\alpha \leq J(u) \leq \sup_{\mathcal{V} \in \mathcal{T}} \inf_{v \in \mathcal{V}} J(v) \leq \alpha$$

□

For finite-dimensional problems, compactness of level sets is usually caused by boundedness, which is not true for infinite-dimensional problems. A similar property holds for Hilbert spaces (and more general for the dual of a Banach space), which is sometimes called "Eberlein-Smullyan" lemma:

**Lemma 2.5.** *Let $\mathcal{U}$ be a Hilbert space and let $(u_k)$ be a bounded sequence in $\mathcal{U}$. Then there exists a weakly convergent subsequence $(u_{kl})$, i.e.*

$$< v, u_{kl} > \to < v, \tilde{u} > \qquad \forall v \in \mathcal{U},$$

*for some $\tilde{u} \in \mathcal{U}$.*

The analoguous property for dual Banach spaces is:

**Lemma 2.6.** *Let $\mathcal{U} = B^*$ for some Banach space $B$ and let $(u_k)$ be a bounded sequence in $\mathcal{U}$. Then there exists a weak-* convergent subsequence $(u_{kl})$, i.e.*

$$< v, u_{kl} > \to < v, \tilde{u} > \qquad \forall v \in B,$$

*for some $\tilde{u} \in \mathcal{U}$.*

As a consequence of Theorem 2.3 and Lemma 2.5/2.6, existence of a global minimizer is obtained if $J$ is weakly (weak-*) lower semicontinuous and $J(u) \leq M$ implies the boundedness of $u$. This observation leads to the concept of coercivity. A functional is called coercive, if

$$\frac{J(u)}{\|u\|} \to \infty \quad \text{as } \|u\| \to \infty.$$

By similar reasoning as above, one can show that a weakly lower semicontinuous and coercive functional attains its global minimum.

## 2.3 Regularization

Non-existence is of course an undesirable effect for optimization problems, which can arise indeed in applications, in particular if the objective function incorporates design goals only, but no control on the cost of a design.

Consider for example the optimal control problem for the heat equation from Example 1.2.1, now for $\Omega = (0,1)$. In this case we want to find an optimal design $u \in L^2(0,T)$ such that

$$\begin{aligned} v_t - v_{xx} &= 0 \quad \text{in } (0,1) \times (0,T) \\ v &= v_0 \quad \text{in } (0,1) \times \{0\} \\ v &= u \quad \text{in } \{0\} \times (0,T) \\ v_x &= 0 \quad \text{in } \{1\} \times (0,T) \end{aligned}$$

which minimizes

$$J(u) = \int_0^1 |v(x,T) - \theta(x)|^2 \, dx.$$

If $J(u) \leq M$, an application of the triangle inequality shows that

$$\int_0^1 |v(x,T)|^2 \, dx \leq \left( M + \sqrt{\int_0^1 \theta(x)^2 \, dx} \right)^2$$

and thus, $v(x,T)$ is bounded in $L^2(0,1)$. Thus we know that $v$ is a solution of the heat equation with bounded initial value $v(.,0)$, bounded final value $v(.,T)$, and bounded derivative $v_x(1,.)$ at the right boundary point. Due to the ill-posedness of a Cauchy problem for parabolic equations, where initial, final and only part of the value on the spatial boundary are prescribed, the control $u = v(.,0)$ need not be bounded in $L^2(0,T)$. Hence, we cannot even obtain compactness in the weak topology and therefore a solution of this problem need not exist. The situation dffers if we incorporate the cost of the control $u$, given by

$$R(u) := \int_0^T u(t)^2 \, dx = \|u\|^2.$$

We can achieve a solution with limited cost in two ways, either by restricting the class of admissible controls to those satisfying

$$R(u) \leq M$$

for some constant $M > 0$, or by penalizing the original problem to

$$J_\beta(u) = J(u) + \beta R(u) \to \min_u$$

for some parameter $\beta > 0$. In both cases, a level set of $J_\beta$ is compact in the weak topology of $L^2(\Omega)$ and since both $J(u)$ and $R(u)$ are weakly lower semicontinuous, we obtain the existence of a minimizer.

Another important issue is the "robustness of the control", i.e. the stability of a solution (if it exists) with respect to the design goal (with respect to the desired final temperature in our example). A major difference to finite-dimensional problems is that arbitrarily small perturbations of the objective functional (respectively of the design goal $\theta$ in our example) can lead to arbitrarily large differences in the solutions of the optimization problem (even if they always exist). Both the non-existence and the instability cause a need for *regularization*, i.e. the approximation of the ill-posed optimization problem by a nearby stable problem. A standard technique to obtain a regularization of the problem is to add a lower semicontinuous functional $R$ as a penalty to the original problem and to solve

$$J_\beta(u) := J(u) + \beta R(u) \to \min_{u \in \mathcal{U}},$$

with $\beta > 0$ being a regularization parameter. If all level sets of $R$ are compact, then $J_\beta$ has a compact and non-empty level set for each $\beta$, since

$$J_\beta(u) \leq M \qquad \text{implies} \qquad R(u) \leq \frac{M - \inf J}{\beta}.$$

Thus, the regularized functional $J_\beta$ has a minimizer for $\beta > 0$. For the regularized functional we can also obtain a stability result with respect to perturbations in the functional $J$. Let

$J_k$ be a sequence of lower semicontinuous functionals converging uniformly to $J$ on compact subsets of $\mathcal{U}$. (This is the case for the functionals

$$J_k := \int |v(x,T) - \theta_k|^2 \, dx$$

with $\theta_k$ being perturbed design goals converging to $\theta$ as $k \to \infty$.) Moverover, let $\mathcal{U}_k$ denote a global minimizer of $J_k + \beta R$. Then $(u_k)$ has a convergent subsequence $(u_{kl})$, and the limit of each convergent subsequence of $(u_k)$ is a global minimizer of $J + \beta R$.

To prove this statement, assume without restriction of generality that $J(u) \geq 0$ and $J_k(u) \geq 0$ for all $u \in \mathcal{U}$. Then, for each $k$ we obtain

$$J_k(u_k) + \beta R(u_k) \leq J_k(\tilde{u}) + \beta R(\tilde{u}) \tag{2.3}$$

where $\tilde{u}$ is a global minimizer of $J + \beta R$. Since $J_k(\tilde{u}) \to J(\tilde{u})$ there exists a sequence $(\epsilon_k)$ of positive real numbers such that $J_k(\tilde{u}) + \beta R(\tilde{u}) \leq J(\tilde{u}) + \beta R(\tilde{u}) + \epsilon_k$ and $\epsilon_k \to 0$. Thus, there exists a positive number

$$M \geq J(\tilde{u}) + \beta R(\tilde{u}) + \epsilon_k$$

and due to (2.3)

$$R(u_k) \leq \frac{M}{\beta}$$

for all $k$. Consequently, the sequence $(u_k)$ is contained in a compact set because of the properties of $R$, and this implies the existence of a convergent subsequence. If $(u_{kl})$ is a convergent subsequence, then due to (2.3) and due to the uniform convergence of $J_k$ on compact sets we obtain for the limit $\hat{u}$

$$\begin{aligned} J(\hat{u}) + \beta R(\hat{u}) &\leq \lim_{l\to\infty} (J_{k_l}(u_{k_l}) + \beta R(u_{k_l})) \\ &\leq \lim_{l\to\infty} (J_{k_l}(\tilde{u}) + \beta R(\tilde{u})) \\ &\leq \lim_{l\to\infty} (J(\tilde{u}) + \beta R(\tilde{u}) + \epsilon_{k_l}) \\ &= J(\tilde{u}) + \beta R(\tilde{u}) \\ &= \inf_u (J(u) + \beta R(u)) \end{aligned}$$

Hence, $\hat{u}$ is a global minimizer of $J + \beta R$.

Using similar techniques, one can also study the limit $\beta \to 0$. If the limit problem $J(u) \to \min_{u\in\mathcal{U}}$ admits a global minimizer $u_0$ then a global minimizer $u_\beta$ of $J + \beta R$ satisfies

$$\begin{aligned} \beta R(u_\beta) &\leq J(u_\beta) - J(u_0) + \beta R(u_\beta) \\ &\leq \beta R(u_0) \end{aligned}$$

and hence $R(u_\beta) \leq R(u_0)$ for each $\beta > 0$. By compactness this implies the existence of a convergent subsequence $(u_{\beta_k})$ and the limit of each convergent subsequence is a global minimizer of $J$.

Vice versa, if there exists no global minimizer of $J$, then $R(u_{\beta_k})$ cannot be bounded for any subsequence $(u_{\beta_k})$, since otherwise one would obtain the existence of a global minimizer of $J$ in the limit $\beta_k \to 0$. Hence, if there exists no global minimizer of the limit problem, it follows that

$$R(u_\beta) \to +\infty$$

as $\beta \to 0$. I.e., one can decide from the asymptotic behaviour of $R(u_\beta)$ whether the (ill-posed) limit problem admits a global minimizer or not.

Note that all the above statements hold only for global minimizers, one cannot expect similar results for local minimizers in general, which can also be seen in simple finite-dimensional examples.

A typical choice for the regularization functional in Hilbert spaces is given by $R(u) = \|u\|^2$, which satisfies the above compactness condition in the weak topology.

## 2.4 Derivatives

Derivatives of objective functionals and constraints are needed for several reasons: first of all, derivatives are needed to deduce local optimality conditions. Secondly, derivatives appear in almost any local optimization procedure as well as in methods to control step sizes and in stopping rules.

We start with derivatives for general nonlinear operators in Banach spaces. Let $F : \mathcal{U} \to \mathcal{V}$. The simplest type of derivative we can define is the derivative with respect to $t$ of the one-dimensional function

$$f_v(t) := F(u + tv).$$

**Definition 2.7.** Let $F$ be a continuous nonlinear operator acting between the Banach spaces $\mathcal{U}$ and $\mathcal{V}$. Then the directional derivative of $\mathcal{F}$ at point $u$ in direction $v$ is defined as

$$d\mathcal{F}(u; v) := \lim_{t \downarrow 0} \frac{F(u + tv) - F(u)}{t} = f_v'(t)|_{t=0}$$

if the limit on the right-hand side exists. If the directional derivatives $d\mathcal{F}(u; v)$ exist for all $v \in \mathcal{U}$, then the operator $F$ is called Gateâux-differentiable at $u$ and $d\mathcal{F}(u; .)$ is called Gateâux-derivative.

If in addition, $d\mathcal{F}(u; .) : \mathcal{U} \to \mathcal{V}$ is a continuous linear operator, then $F$ is called Fréchet-differentiable with Fréchet-derivative

$$F'(u)v := dF(u; v) \quad \forall v \in \mathcal{U}$$

For objective functionals, i.e., nonlinear operators $J : \mathcal{U} \to \mathbb{R}$, the Fréchet-derivative $J'(u)$., if it exists, is a continuous linear form. Hence, $J'(u)$. can be identified with an element in the dual space $\mathcal{U}^*$. For a Hilbert space $\mathcal{U}$, $J'(u)$ can even be identified with an element of $\mathcal{U}$ due to standard duality.

In an inductive way, we can define higher derivatives. In particular, the second Fréchet-derivative of an operator is given as a bilinear form, defined by

$$F''(u)(v, w) := \lim_{t \downarrow 0} \frac{F'(u + tw)v - F'(u)v}{t}.$$

In the following we present some examples of computing derivatives, in particular for the problems introduced in Chapter 1.

### 2.4.1 Minimal Surfaces

We start by computing the derivative of the functional

$$J(u) := \int_\Omega \sqrt{1+|\nabla u|^2}\ dx,$$

to be minimized by minimal surfaces in graph form. This functional can be defined on the Banach space $\mathcal{U} = BV(\Omega)$ with the norm

$$\|u\|_{BV(\Omega)} := \|u\|_{L^1(\Omega)} + |u|_{BV(\Omega)}$$

and

$$|u|_{BV} := \sup_{g\in C_0^\infty(\Omega)} \int_\Omega u.div\ g\ dx$$

If $u$ is sufficiently smooth, then

$$|u|_{BV} = \int_\Omega |\nabla u|\ dx.$$

The variation in direction $v$ is given by

$$J(u+tv) = \int_\Omega \sqrt{1+|\nabla u + t\nabla v|^2}\ dx$$

If $u \in C_0^\infty(\Omega)$, we can compute the limit $t \to 0$ as

$$\begin{aligned} J'(u)v &= \int_\Omega \frac{\nabla u.\nabla v}{\sqrt{1+|\nabla u|^2}} dx \\ &= -\int v\ div\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right) dx. \end{aligned}$$

Obviously, $J'(u).$ is a continuous linear functional on $BV(\Omega)$ for $u \in C_0^2(\Omega)$, but not for arbitrary $u \in BV(\Omega)$. Thus, the Fréchet-derivative exists only on a dense subspace.

### 2.4.2 Boundary Control

Our second example is the boundary control problem introduced in section 1.2.2. For simplicity, we restrict our attention to the one-dimensional case $\Omega = (0,1)$. The derivative of the regularized objective

$$J(u,v) = \int_0^1 |v(x,T) - \theta(x)|^2\ dx + \beta \int_0^T u(t)^2\ dx$$

is given by

$$\begin{aligned} J'(u,v)(g,h) &= \lim_{s\downarrow 0} \frac{1}{s}\Big( \int_0^1 |v(x,T) + sh(x,T) - \theta(x)|^2 - \int_0^1 |v(x,T)-\theta(x)|^2 dx \\ &\quad + \beta \int_0^T (u(t)+sg(t))^2 dt - \beta \int_0^T u(t)^2 dt\Big) \\ &= \lim_{s\downarrow 0} \Big( 2\int_0^1 (v(x,T)-\theta(x))h(x,T)dx + 2\beta \int_0^T u(t)g(t)dt \\ &\quad + s\int_0^1 h(x,T)^2 dx + \beta s \int_0^T g(t)^2 dt\Big). \end{aligned}$$

Since $\int_0^1 h(x,T)^2 dx$ and $\int_0^T g(t)^2 dt$ are bounded for $h \in C(0,T;L^2(\Omega))$ and $g \in L^2(0,T)$, the last two terms tend to zero, and hence

$$J'(u,v)(g,h) = 2\int_0^1 (v(x,T) - \theta(x))h(x,T)dx + 2\beta \int_0^T u(t)g(t)dt.$$

Using the Cauchy-Schwarz inequality one can check that $J'(u,v)(.,.)$ is a bounded linear functional.

The derivatives of the equation constraint can be computed by introducing the operator

$$e : (u,v) \mapsto \begin{pmatrix} v_t - v_{xx} \\ v(.,0) - v_0 \\ v(0,.) - u \\ v_x(1,.) \end{pmatrix}$$

The equation constraint can be written as $e(u,v) = 0$. Since $e$ is affinely linear, its derivative is given by

$$e'(u,v)(g,h) = e(g,h) - e(0,0) = \begin{pmatrix} h_t - h_{xx} \\ h(.,0) \\ h(0,.) - g \\ h_x(1,.) \end{pmatrix}.$$

## 2.5 Optimality Conditions

In the following we shall derive necessary and sufficient conditions for local minima based on derivatives.

### 2.5.1 Unconstrained Problems

We start our analysis with unconstrained problems, i.e. we investigate

$$J(u) \to \min_{u \in \mathcal{U}},$$

where $J$ is continuously Fréchet-differentiable on $\mathcal{U}$.
Under this condition, we have the following *necessary first-order condition*:

**Proposition 2.8.** *Let $J$ be Fréchet-differentiable and let $\overline{u}$ be a local minimizer. Then $J'(\overline{u}) = 0$.*

*Proof.* Since $\overline{u}$ is a local minimizer, the inequality

$$J(\overline{u} + tv) - J(\overline{u}) \geq 0$$

holds for all $v \in \mathcal{U}$ and $t \in \mathbb{R}^+$ sufficiently small. Thus, we also obtain

$$J'(\overline{u})v = \lim_{t \downarrow 0} \frac{J(\overline{u} + tv) - J(\overline{u})}{t} \geq 0$$

for all $v \in \mathcal{U}$. Since $J'(\overline{u}).$ is a linear functional, we obtain

$$0 \leq J'(\overline{u})(-v) = -J'(\overline{u})v \leq 0$$

and hence, $J'(\overline{u})v = 0$ for all $v \in \mathcal{U}$. □

Note that the last step of the proof cannot be carried out if $J$ is only Gateâux-differentiable at $\overline{u}$. In this case we obtain only $dJ(\overline{u}; v) \geq 0$ for all $v \in \mathcal{U}$.

Of course, the above necessary condition does not characterize a local minimizer in general. In particular, each local maximizer satisfies the same condition. In order to deduce sufficient optimality conditions, we need second derivatives of the objective functional. The main idea of second order *sufficient optimality conditions* is that local convexity around a stationary point $\overline{u}$ implies that $\overline{u}$ is a local minimizer.

**Proposition 2.9.** *Let $J$ be twice continuously Fréchet-differentiable and let $\overline{u}$ be a stationary point of $J$ (i.e., $J(\overline{u}) = 0$ ). Let $J''(\overline{u})$ be a positive definite bilinear form, i.e.*

$$J''(\overline{u})(v, v) \geq \alpha \|v\|^2$$

*for all $v \in \mathcal{U}$ and some $\alpha \in \mathbb{R}^+$ independent of $v$. Then $\overline{u}$ is a local minimizer of $J$.*

*Proof.* Since $J$ is twice continuously Fréchet-differentiable, the Taylor expansion

$$J(v) = J(\overline{u}) + J'(\overline{u})(v - \overline{u}) + \frac{1}{2} J''(\overline{u})(v - \overline{u}, v - \overline{u}) + o(\|v - \overline{u}\|^2)$$

holds in a neighborhood of $\overline{u}$. In particular, there exists an $\epsilon > 0$ such that

$$J(v) \geq J(\overline{u}) + J'(\overline{u})(v - \overline{u}) + \frac{1}{2} J''(\overline{u})(v - \overline{u}, v - \overline{u}) - \frac{\alpha}{4} \|v - \overline{u}\|^2$$

if $\|v - \overline{u}\| < \epsilon$.
Inserting $J'(\overline{u}) = 0$ and the positive definiteness of $J''(\overline{u})$ we obtain

$$J(v) \geq J(\overline{u}) + \frac{\alpha}{4} \|v - \overline{u}\|^2,$$

for $\|v - \overline{u}\| < \epsilon$ and in particular

$$J(v) > J(\overline{u})$$

if $v \neq \overline{u}$. Thus, $\overline{u}$ is a local minimizer. □

### 2.5.2 Constrained Problems

In the following, we derive optimality conditions for constrained optimization problems, i.e. we consider

$$J(u) \to \min_u$$

subject to $u \in \mathcal{C}$. The derivation of local optimality conditions for general $\mathcal{C}$ is impossible, since $\mathcal{C}$ might include isolated points, which are always local minimizers. We therefore restrict our attention to special classes of constraints such as convex constraint sets $\mathcal{C}$ or equality constraints. If $\mathcal{C}$ is closed, we can formulate local optimality conditions in terms of the tangential cone, which consists of all directions being tangential to $\partial\mathcal{C}$ at a point $u$ or pointing into the interior of $\mathcal{C}$.

**Definition 2.10.** Let $\mathcal{C}$ be closed. For $u \in \mathcal{C}$, the tangential cone $T_{\mathcal{C}}(u)$ at $u$ is defined by

$$T_{\mathcal{C}}(u) = \{v \in \mathcal{U} \mid \exists \epsilon > 0, \forall\ 0 \leq t \leq \epsilon\ \exists w(t) \in \mathcal{C} : \|u + tv - w(t)\| = o(t)\}$$

Note that for $u \in int\ \mathcal{C}$, the tangential cone is just $T_{\mathcal{C}}(u) = \mathcal{U}$, since $w(t) = u + tv \in \mathcal{C}$ for $t$ sufficiently small.
The first-order optimality conditions for a problem with closed constraint set just states that the objective may not decrease locally for any tangential direction:

**Proposition 2.11.** *Let $J : \mathcal{U} \to \mathbb{R}$ be continuously Fréchet-differentiable and let $\overline{u}$ be a local minimizer of*

$$J(u) \to \min_{u \in \mathcal{C}}$$

*for a closed set $\mathcal{C}$. Then*

$$J'(\overline{u})v \geq 0 \quad \forall v \in T_{\mathcal{C}}(\overline{u}).$$

*Proof.* Let $v \in T_{\mathcal{C}}(\overline{u})$. Then, for each $t \in [0, \epsilon]$, we have

$$\begin{aligned} J(w(t)) &= J(\overline{u} + tv) + o(t) \\ &= J(\overline{u}) + tJ'(\overline{u})v + o(t). \end{aligned}$$

Assume that $J'(\overline{u})v < 0$, then for $t$ sufficiently small,

$$J(w(t)) \leq J(\overline{u}) + tJ'(\overline{u})v - \frac{t}{2}J'(\overline{u})v < J(\overline{u}),$$

which is a contradiction to $\overline{u}$ being a local minimizer. Hence, $J'(\overline{u})v \geq 0$ for all $v \in T_{\mathcal{C}}(\overline{u})$. □

For second order optimality conditions, we need in addition the convexity of $\mathcal{C}$.

**Proposition 2.12.** *Let $\mathcal{C}$ be closed and convex, and let $J : \mathcal{U} \to \mathbb{R}$ be twice continuously Fréchet-differentiable. If $\overline{u} \in \mathcal{C}$ satisfies*

$$J'(\overline{u})v \geq 0 \quad \forall v \in T_{\mathcal{C}}(\overline{u})$$

*and, for some $\beta > 0$*

$$J''(\overline{u})(v, v) \geq \beta \|v\|^2 \quad \forall v \in T_{\mathcal{C}}(\overline{u}),$$

*then $\overline{u}$ is a local minimizer of $J$ in $\mathcal{C}$.*

*Proof.* Let $w \in \mathcal{C}$ and $\|w - \overline{u}\|$ be sufficiently small. Then, due to convexity of $\mathcal{C}$, $\overline{u} + tv \in \mathcal{C}$ for $t \in [0, 1]$, with $v = w - \overline{u}$. In particular, $v \in T_{\mathcal{C}}(\overline{u})$, and hence

$$J(w) \geq J(\overline{u}) + J'(\overline{u})v + \frac{1}{2}J''(\overline{u})(v, v) - \frac{\beta}{4}\|v\|^2$$

for $\|v\|$ sufficiently small.
Thus,

$$J(w) \geq J(\overline{u}) + \frac{\beta}{4}\|v\|^2 > J(\overline{u}),$$

for all $w \in \mathcal{C}$ with $\|w - u\|$ sufficiently small, which implies that $\overline{u}$ is a local minimizer of $J$ in $\mathcal{C}$. □

An alternative way of dealing with constraints are Lagrange multipliers. In the following, let $E : \mathcal{U} \to \mathcal{V}$, and $I : \mathcal{U} \to \mathcal{W}$, where $\mathcal{V}$ is a Banach space and $\mathcal{W}$ is an ordered Banach space, with order relation $\preceq$.
We consider the problem

$$J(u) \to \min_{u \in \mathcal{U}}$$

subject to

$$\begin{aligned} E(u) &= 0 \\ I(u) &\preceq 0. \end{aligned}$$

Corresponding to the equality constraints in $\mathcal{V}$ and the inequality constraints in $\mathcal{W}$, we introduce Lagrangian variables $p \in \mathcal{V}^*$ and $q \in \mathcal{W}^*$ and the Lagrange functional

$$\mathcal{L}(u; p, q) = J(u) + < p, E(u) > + < q, I(u) > .$$

Now we consider the "dual problem", namely the maximization of $\mathcal{L}$ with respect to $p$ and $q$.

Assume first that $E(u) = 0$. Then $< p, E(u) >= 0$ for all $p$, and hence

$$\sup_{p \in \mathcal{V}^*} < p, E(u) > = 0$$

If $E(u) \neq 0$, then we can find $p \in \mathcal{V}^*$ such that

$$< p, E(u) > \ > 0,$$

and hence, for $t \to +\infty$ in $\mathbb{R}^+$, $p_t := tp$,

$$< p_t, E(u) >= t \ < p, E(u) > \to \ +\infty.$$

I.e.,

$$\sup_{p \in \mathcal{V}^*} < p, E(u) >= \begin{cases} 0 & \text{if } E(u) = 0 \\ +\infty & \text{else.} \end{cases}$$

Similar reasoning applies to the inequality constraint if we restrict the Lagrangian variable to be positive. Note that the natural order on $\mathcal{W}^*$ is given by

$$q \succeq 0 \Leftrightarrow (< q, w >\geq 0, \quad \forall w \in \mathcal{W}, w \succeq 0)$$

Assume that $q \succeq 0$ and $I(u) \preceq 0$. Then $< q, I(u) > \leq 0$, and thus,

$$\sup_{q \in \mathcal{W}^*, q \succeq 0} < q, I(u) >= 0.$$

If $< q_0, I(u) > \ > 0$ for some $q_0 \succeq 0$, then $q_t := tq_0$, $t \in \mathbb{R}^+$ satisfies $< q_t, I(u) >= t \ < q_0, I(u) > \to +\infty$ as $t \to \infty$ and hence,

$$\sup_{q \in \mathcal{W}^*, q \succeq 0} < q, I(u) >= +\infty$$

Summing up, we obtain that

$$\sup_{p \in \mathcal{V}^*, q \in \mathcal{W}^*, q \succeq 0} \mathcal{L}(u; p, q) = \begin{cases} J(u) & \text{if } E(u) = 0, I(u) \preceq 0 \\ +\infty & \text{else.} \end{cases}$$

As a direct consequence, we may conclude that

$$\inf_{u\in\mathcal{U},E(u)=0,I(u)\preceq 0} J(u) = \inf_{u\in\mathcal{U}} \sup_{p\in\mathcal{V}^*,q\in\mathcal{W}^*,q\succeq 0} \mathcal{L}(u;p,q).$$

If the inf and sup are attained, we can apply local optimality for unconstrained problems to deduce that

$$\frac{\partial\mathcal{L}}{\partial u}(\overline{u};\overline{p},\overline{q}) = 0$$
$$\frac{\partial\mathcal{L}}{\partial p}(\overline{u};\overline{p},\overline{q}) = 0$$

in a local minimum of $J$. Moreover, we can apply the above tangential cone condition for constrained optimization to deduce that

$$\frac{\partial\mathcal{L}}{\partial q}(\overline{u};\overline{p},\overline{q})q \leq 0$$

for all $q$ satisfying

$$< q, w >\geq 0 \quad \text{if } < \overline{q}, w >= 0.$$

The main problem in infinite-dimensional constrained optimization is that, opposed to the finite-dimensional case, the sup with respect to $p$ and $q$ need not be attained and in such a case the local optimality conditions need not be valid. In order to understand the existence problem for Lagrangian variables, we investigate the optimality condition

$$0 = \frac{\partial\mathcal{L}}{\partial u}(\overline{u};\overline{p},\overline{q})v = J'(u)v + < p, E'(u)v > + < q, I'(u)v >$$

For simplicity, we start with pure equality constraints, i.e.

$$0 = J'(\overline{u})v + < p, E'(u)v >$$

for all $v \in \mathcal{U}$. This condition is equivalent to

$$J'(\overline{u}) + E'(\overline{u})^* p = 0,$$

which can be interpreted as an equation for $p$. Existence for this linear equation is guaranteed, if $E'(\overline{u})^* : \mathcal{V}^* \to \mathcal{U}^*$ is a surjective, bounded linear operator. If we have an additional inequality condition, then the optimality becomes

$$J'(\overline{u}) + E'(\overline{u})^* p + I'(u)^* q = 0$$

and we can find Lagrangian variables $\overline{p}$ and $\overline{q} \succeq 0$ if the operator

$$\begin{pmatrix} E'(\overline{u})^* \\ I'(\overline{u})^* \end{pmatrix} : \mathcal{V}^* \times \{q \in \mathcal{W}^* | q \succeq 0\} \to \mathcal{U}^*$$

is surjective.

The system of optimality conditions for $(\overline{u},\ \overline{p},\ \overline{q})$ is often called Karush-Kuhn-Tucker (KKT) system. We can summarize the above result in

**Proposition 2.13.** *Let $(\overline{u}, \overline{p}, \overline{q})$ be a local solution of*

$$\mathcal{L}(u; p, q) \to \min_{u \in \mathcal{U}} \max_{p \in \mathcal{V}^*, q \in \mathcal{W}^*, q \succeq 0},$$

*then it satisfies*

$$\begin{aligned} J'(\overline{u}) + E'(\overline{u})^* \overline{p} + I'(\overline{u})^* \overline{q} &= 0 \\ E(\overline{u}) &= 0 \\ I'(\overline{u}) &\preceq 0 \\ \overline{q} &\succeq 0 \end{aligned}$$

*Moreover, $\overline{u}$ and $\overline{q}$ satisfy the so-called complementarity condition $< \overline{q}, I'(\overline{u}) >= 0$.*

In order to obtain sufficient second-order optimality conditions, we can use second derivatives of the Lagrange functional. Note that second derivatives with respect to $p$ and $q$ vanish, i.e.

$$\frac{\partial^2 \mathcal{L}}{\partial p^2} = 0, \frac{\partial^2 \mathcal{L}}{\partial q^2} = 0, \frac{\partial^2 \mathcal{L}}{\partial p \partial q} = 0.$$

The mixed second derivatives

$$\frac{\partial^2 \mathcal{L}}{\partial u \partial p} = E'(u), \frac{\partial^2 \mathcal{L}}{\partial u \partial q} = I'(u)$$

are the linearization of the constraints. The important part for the sufficient condition is the second derivative with respect to $u$,

$$\frac{\partial^2 \mathcal{L}}{\partial u^2} = J''(u)(.,.) + < E''(u)(.,.), p > + < I''(u)(.,.), q > .$$

**Proposition 2.14.** *Let $\overline{u}$ be a local minimum of $J(u) \to \min_{u \in \mathcal{U}}$ subject to $E(u) = 0$, $I(u) \leq 0$. Let $(\overline{p}, \overline{q})$ be Lagrangian variables such that the KKT-conditions are satisfied for $(\overline{u}, \overline{p}, \overline{q})$, and if*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\overline{u}; \overline{p}, \overline{q})(v, v) \geq \alpha \|v\|^2$$

*with some constant $\alpha \in \mathbb{R}^+$ for all $u \in \mathcal{U}$. Then $\overline{u}$ is a local minimizer of $J(u) \to \min$ subject to $E(u) = 0$ and $I(u) \preceq 0$.*

*Proof.* Let $w \in \mathcal{U}$ with $E(w) = 0$, $I(w) \preceq 0$ and $\|w - \overline{u}\|$ sufficiently small. Then

$$J(w) = J(\overline{u}) + J'(\overline{u})(w - \overline{u}) + \frac{1}{2} J''(\overline{u})(w - \overline{u}, w - \overline{u}) + o(\|w - \overline{u}\|^2).$$

Due to the KKT-condition,

$$-J'(\overline{u})(w - \overline{u}) =< E'(\overline{u})(w - \overline{u}), \overline{p} > + < I'(\overline{u})(w - \overline{u}), \overline{q} > .$$

Thus,

$$\begin{aligned} J(w) &= \frac{1}{2} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\overline{u}; \overline{p}, \overline{q})(w - \overline{u}, w - \overline{u}) \\ &\quad + \left\langle E'(\overline{u})(w - \overline{u}) + \frac{1}{2} E''(\overline{u})(w - \overline{u}, w - \overline{u}), \overline{p} \right\rangle \\ &\quad + \left\langle I'(\overline{u})(w - \overline{u}) + \frac{1}{2} I''(\overline{u})(w - \overline{u}, w - \overline{u}), \overline{q} \right\rangle \\ &\quad + o(\|w - \overline{u}\|^2). \end{aligned}$$

Due to $E(\overline{u}) = 0$, $< I(\overline{u}), \overline{q} >= 0$, we obtain

$$E'(\overline{u})(w - \overline{u}) + \frac{1}{2}E''(\overline{u})(w - \overline{u}, w - \overline{u}) = o(\|w - \overline{u}\|^2)$$

and

$$< I'(\overline{u})(w - \overline{u}) + I''(\overline{u})(w - \overline{u}, w - \overline{u}), \overline{q} >= o(\|w - \overline{u}\|^2).$$

Inserting this identity in the above expansion for $J(w)$, we deduce

$$J(w) = J(\overline{u}) + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\overline{u}; \overline{p}, \overline{q})(w - \overline{u}, w - \overline{u}) + o(\|w - \overline{u}\|^2),$$

which implies

$$J(w) > J(\overline{u})$$

if $w \neq \overline{u}$, $\|w - \overline{u}\|$ sufficiently small. □

## 2.6 Applications

In the following we apply the techniques introduced above to some of the examples introduced in Section 1.2.

### 2.6.1 Boundary Control

We consider again the one-dimensional boundary control problem

$$J(u) := \int_0^1 (v(x,T) - \theta(x))^2 dx + \alpha \int_0^1 u(t)^2 dt \to \min$$

subject to

$$\begin{aligned} v_t - v_{xx} &= 0 \quad \text{in } (0,1) \times (0,T) \\ v &= v_0 \quad \text{in } (0,1) \times \{0\} \\ v &= u \quad \text{in } \{0\} \times (0,T) \\ v_x &= 0 \quad \text{in } \{1\} \times (0,T) \end{aligned}$$

As we have seen above, existence of a solution $v \in L^2(0,T; H^1(\Omega))$, $u \in L^2(0,T)$ is guaranteed for $\alpha > 0$. We now compute the derivative of the reduced functional

$$\tilde{J}(u) = J(u, v(u)),$$

using the implicit function theorem, i.e.

$$\begin{aligned} \tilde{J}'(u)w &= \frac{\partial J}{\partial u}(u, v(u)), w + \frac{\partial J}{\partial v}(u, v(u))v'(u)w \\ &= 2\alpha \int uw \; dt + 2 \int (u(x,T) - \theta(x))v'(x,T)dx \end{aligned}$$

The derivative $v' = v'(u)w$ can be computed from the equality constraint $E(u, v(u)) = 0$ as

$$\frac{\partial E}{\partial v}(u, v(u))v' = -\frac{\partial E}{\partial u}(u, v(u))$$

For the heat equation as equality constraint, this implies that $v'$ solves

$$\begin{aligned} v'_t - v'_{xx} &= 0 \quad \text{in } (0,1) \times (0,T) \\ v' &= 0 \quad \text{in } (0,1) \times \{0\} \\ v' &= w \quad \text{in } \{0\} \times (0,T) \\ v'_x &= 0 \quad \text{in } \{1\} \times (0,T) \end{aligned}$$

Hence, the computation of each directional derivative requires the solution of an initial-boundary value problem for the heat equation. The computation of the gradient $\tilde{J}'(u)$ requires the solution of an initial boundary value problem for each variation $w$.
As we shall see in the following (and in a more general way in the section on PDE-constrained optimization below), there exists a more efficient way of computing the gradient $\tilde{J}'(u)$, the so-called *adjoint-method*. The main idea of this approach is to introduce an adjoint equation (related to $E'(u)^*$), in this case a function $\varphi : (0,1) \times (0,T) \to \mathbb{R}$ solving

$$\begin{aligned} -\varphi_t - \varphi_{xx} &= 0 \quad \text{in } (0,1) \times (0,T) \\ \varphi &= v(.,T) - \theta \quad \text{in } (0,1) \times \{T\} \\ \varphi &= 0 \quad \text{in } \{0\} \times (0,T) \\ \varphi_x &= 0 \quad \text{in } \{1\} \times (0,T) \end{aligned}$$

By applying integration by parts we can then compute

$$\begin{aligned} & \int_0^1 \Big(u(x,T) - \theta(x)\Big)\ v'(x,T)\ dx \\ = & \int_0^1 \varphi(x,T)\ v'(x,T)\ dx \\ = & \int_0^1 \varphi(x,0)\ \underbrace{v'(x,0)}_{=0} dx + \int_0^T \int_0^1 \frac{\partial}{\partial t}\Big(\varphi(x,t)\ v'(x,t)\Big)\ dx\ dt \\ = & \int_0^T \int_0^1 \Big(\varphi_t(x,t)\ v'(x,t) + \varphi(x,t)\ v'_t(x,t)\Big)\ dx\ dt \\ = & \int_0^T \int_0^1 \Big(-\varphi_{xx}(x,t)\ v'(x,t) + \varphi(x,t)\ v'_{xx}(x,t)\Big)\ dx\ dt \\ = & \int_0^T \int_0^1 \Big(\varphi_x(x,t)\ v'_x(x,t) - \varphi_x(x,t)\ v'_x(x,t)\Big)\ dx\ dt \\ + & \int_0^T \Big(-\varphi_x(x,t)\ v'(x,t) + \varphi(x,t)\ v'_x(x,t)\Big)\Big|_0^1\ dt \\ = & \int_0^T \varphi_x(0,t)\ w(t)\ dt. \end{aligned}$$

Hence, the derivative of $\tilde{J}$ is given by

$$\tilde{J}'(u)w = 2\int_0^T \Big(\varphi_x(0,t) + \alpha\ u(t)\Big)\ w(t)\ dt$$

and thus, we can identify the linear functional $\tilde{J}'(u) : L^2(0,T) \to \mathbb{R}$ with $\tilde{J}'(u) = \varphi_x|_{x=0} + \alpha u$. I.e., by using the adjoint method we can compute the whole gradient $\tilde{J}'(u)$ by solving a single parabolic initial-boundary value problem (with reversed time direction).

The first-order optimality condition in this case is simply

$$u(t) = -\frac{1}{\alpha}\varphi_x(0,t).$$

In order to test second-order optimality, we consider again the formula

$$\tilde{J}'(u)w = 2\int_0^1 \left(\Big(v(x,T) - \theta(x)\Big)\ v'(x,T) + \alpha u(x)w(x)\right) dx$$

and compute its variation with respect to $u$. Since $v'$ is independent of $u$, we obtain $\frac{\partial}{\partial u}(v') \equiv 0$, and therefore

$$\begin{aligned}\tilde{J}''(u)(w,w) &= 2\int_0^1 \Big((v'(x,T))^2 + \alpha\ w(x)^2\Big)\ dx \\ &\geq 2\ \alpha \int_0^1 w(x)^2\ dx = 2\ \alpha\|w\|^2.\end{aligned}$$

Consequently, each stationary point is a local minimizer. As we shall see below, this problem is strictly convex, and there is only one stationary point, and this point is a global minimizer.

### 2.6.2 Elliptic PDEs

Consider the quadratic variational problem

$$J(u) = \int \left(\frac{1}{2}\Big(a(x)\ |\nabla u(x)|^2 + c(x)\ u(x)^2\Big) - \varphi(x)\ u(x)\right) dx \to \min_{u\in\mathcal{U}}$$

with $\Omega \subset \mathbb{R}^d$,

$$\mathcal{U} = H_0^1(\Omega) := \{u \in L^2(\Omega) | \nabla u \in L^2(\Omega)^d, u|_{\partial\Omega} = 0\}.$$

Assume that $a \in L^\infty(\Omega)$, $c \in L^\infty(\Omega)$ and $a(x) \geq a_0, \quad c(x) \geq 0$ a.e. in $\Omega$, for some $a_0 > 0$. Then $J(u) \leq M$ implies $\int |\nabla u|^2 dx \leq \frac{M}{a_0}$, from which we can deduce compactness in the weak topology of $H_0^1(\Omega)$. Moreover, one can show that $J$ is weakly lower semicontinuous on $H_0^1(\Omega)$, which implies the existence of a minimizer. The derivative of $J$ can be computed as

$$J'(u)v = \int_\Omega (a\nabla u\nabla v + cuv - \varphi v)\ dx$$

and the first-order optimality condition

$$J'(\overline{u})v = 0 \quad \forall v \in \mathcal{V},$$

implies that a minimizer of $J$ is a weak solution of the elliptic differential equation

$$\begin{aligned} -div(a\nabla u) + cu &= \varphi \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

As a consequence, we can obtain the existence of a weak solution for this PDE. Using convexity arguments we will later show uniqueness of a weak solution, too.
This example demonstrates that infinite-dimensional optimization theory can be used to prove existence and uniqueness of partial differential equations. This technique is used in particular for nonlinear problems.

### 2.6.3 Stokes Problem

Let $\Omega \subset \mathbb{R}^d$ be a smooth, bounded domain and

$$J(u) := \frac{1}{2}\int_\Omega |\nabla u|^2 \, dx - \int_\Omega fu \, dx,$$

to be minimized in $\mathcal{U} := H_0^1(\Omega)^d$, subject to the constraint $div(u) = 0 \quad a.e.$ in $\Omega$.

As in the previous section, we obtain the weak lower semicontinuity of $J$ and compactness of its level sets. Moreover, the set of constraints is weakly closed in $H_0^1(\Omega)$. Thus, there exists a minimizer for this problem. In order to obtain the first-order optimality conditions, we introduce the Lagrange functional

$$\begin{aligned} \mathcal{L} &: H_0^1(\Omega)^d \times L^2(\Omega) \to \mathbb{R}, \\ \mathcal{L}(u;p) &= J(u) + \int_\Omega p \; div \; u \; dx \end{aligned}$$

A solution of the variational problem satisfies

$$\begin{aligned} \int \nabla u.\nabla v \; dx + \int p \; div \; v \; dx &= \int f \; v \; dx \\ \int q \; div \; u \; dx &= 0 \end{aligned}$$

for all $v \in H_0^1(\Omega)^d$, $p \in L^2(\Omega)$. An application of Gauss' Theorem finally shows that $(u, p)$ is a weak solution of

$$\begin{aligned} -\Delta u - \nabla p &= f, \\ div \; u &= 0, \end{aligned}$$

the so-called *Stokes problem*.

### 2.6.4 Variational Inequalities

Many obstacle problems and some free boundary problems are modeled by variational inequalities. In the simplest case, consider the functional

$$J(u) := \int_\Omega \Big( \frac{1}{2}(a|\nabla u|^2 + cu^2) - fu \Big) \, dx$$

on $H_0^1(\Omega)$, where $a$ and $c$ are as in Section 2.6.2. Now we minimize the functional $J$ subject to the constraint

$$u \in \mathcal{K}$$

where $\mathcal{K}$ is a closed, convex set. By standard reasoning we may deduce the existence of a minimizer. In order to obtain the first order optimality conditions, we inspect the tangential cone for a convex constraint set $\mathcal{K}$. Let $v \in \mathcal{T}_{\mathcal{K}}(u)$ for some $u \in \mathcal{K}$. Then, for each $t$ sufficiently small, there exists $w(t) \in \mathcal{K}$ such that

$$\|w(t) - u - tv\| = o(t) \tag{2.4}$$

In particular, for each $w \in \mathcal{K}$, we have that $w(t) := u + t(w-u) \in \mathcal{K}$ and thus $w - u \in \mathcal{T}_{\mathcal{K}}(u)$. The first order optimality conditions are given by

$$J'(u)v \geq 0 \quad \forall v \in \mathcal{T}_{\mathcal{K}}(u),$$

which implies in particular

$$J'(u)(w-u) \geq 0 \quad \forall w \in \mathcal{K}. \tag{2.5}$$

Now assume that only (2.5) holds and let $v \in \mathcal{T}_{\mathcal{K}}(u)$, with $w(t) \in \mathcal{K}$ satisfying (2.4). Then

$$J'(u)v = J'(u)\left(\frac{w(t)-u}{t}\right) + \frac{o(t)}{t} \geq \frac{o(t)}{t} \to 0.$$

Thus, the first-order optimality condition is equivalent to

$$J'(u)(w-u) \geq 0 \quad \forall v \in \mathcal{K}$$

in this case.

As a simple example, we consider the obstacle problem

$$\mathcal{K} = \{u \in H_0^1(\Omega) | u(x) \geq g(x) \text{ for a.e. } x\}$$

The first-order optimality condition is given by

$$\int \Big( a\nabla u \nabla(w-u) + cu(w-u) - f(w-u) \Big)\, dx \geq 0, \qquad \forall\ w \in \mathcal{K}.$$

If we assume that $u$ is sufficiently smooth and apply Gauss' Theorem, we obtain

$$\int \Big( -div\ (a\nabla u) + cu - f)(w-u) \Big)\, dx \geq 0$$

If $u(\overline{x}) - g(\overline{x}) > 0$ at a point $\overline{x} \in \mathbb{R}$, then we can find local perturbations $w = u + h$ and $w = u - h$ with $h \geq 0$, $supp(h) \subset B_R(\overline{x})$. Thus,

$$\frac{1}{|B_R|} \int_{B_R(\overline{x})} (-div\ (a\nabla u) + cu - f)h\ dx \geq 0$$
$$\frac{1}{|B_R|} \int_{B_R(\overline{x})} (-div\ (a\nabla u) + cu - f)h\ dx \leq 0$$

and in the limit, this implies

$$-div\ (a\nabla u(\overline{x})) + cu(\overline{x}) - f(\overline{x}) = 0.$$

If $u(\overline{x}) = g(\overline{x})$, then we can only find perturbations $w = u + h$, with $h(\overline{x}) \geq 0$. Thus, with a similar procedure as above, we can only deduce

$$-div\ (a\nabla u(\overline{x})) + cu(\overline{x}) + f(\overline{x}) \geq 0.$$

I.e., a smooth solution of the variational inequality satisfies

$$\begin{aligned} -div\ (a\nabla u) + cu - f &\geq 0 \\ v - g &\geq 0 \\ (-div\ (a\nabla u) + cu - f)(u - g) &= 0 \end{aligned}$$

in $\Omega$.

### 2.6.5 Calculus of Variations

A classical application are the so-called Euler equations in the calculus of variations. Let $g : \mathbb{R}^2 \times [0,1] \to \mathbb{R}$ be a smooth function and define

$$J(u) := \int_0^1 g(u(t), u'(t), t)\ dt.$$

Each smooth local miminizer of $J$ has to satisfy the first-order optimality condition

$$0 = J'(u)v = \int_0^1 \left( \frac{\partial g}{\partial u} v(t) + \frac{\partial g}{\partial u'} v'(t) \right)\ dt.$$

If we restrict our attention to those $v$ satisfying $v(0) = v(1) = 0$, we can integrate by parts and deduce

$$0 = \int_0^1 \left( \frac{\partial g}{\partial u} - \frac{d}{dt} \frac{\partial g}{\partial u'} \right) v(t)\ dt$$

for all variations $v$. Hence, $u$ is a solution of the Euler equation

$$\frac{d}{dt} \frac{\partial g}{\partial u'}(u(t), u'(t), t) = \frac{\partial g}{\partial u}(u(t), u'(t), t),$$

a nonlinear second order differential equation.

### 2.6.6 Optimal Design

The typical structure of optimal design problems is given by

$$J(v, u) \to \min$$

subject to an equation constraint

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

which usually admits a unique solution $v$ (the state) for given $u$ (the design). Therefore it is reasonable to assume that $\frac{\partial E}{\partial v}$ is a continuous linear operator with continuous inverse.

Using the Lagrangian

$$\mathcal{L}(v, u; p) := J(v, u) + < E(v, u), p >$$

we obtain first-order optimality conditions of the form

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial v} \ . \ = \frac{\partial J}{\partial v} \ . \ + \left\langle \frac{\partial E}{\partial v}(v,u) \ . \ , p \right\rangle \\ 0 &= \frac{\partial \mathcal{L}}{\partial u} \ . \ = \frac{\partial J}{\partial u} \ . \ + \left\langle \frac{\partial E}{\partial u}(v,u) \ . \ , p \right\rangle \\ 0 &= \frac{\partial \mathcal{L}}{\partial p} = E(u,v) \end{aligned}$$

Note that due to the above assumption on $\frac{\partial E}{\partial v}$, we may compute the Lagrangian variable

$$p = -\left(\frac{\partial E^*}{\partial v}\right)^{-1} \frac{\partial J}{\partial v}$$

and thus,

$$\frac{\partial J}{\partial u} = \frac{\partial E^*}{\partial u}\left(\frac{\partial E^*}{\partial v}\right)^{-1} \frac{\partial J}{\partial v}.$$

This is exactly the same optimality condition that we would obtain by eliminating $v = v(u)$ and using the implicit function theorem to obtain the derivative of

$$\tilde{J}(u) := J(v(u), v).$$

As an example, we consider the semiconductor design problem from Section 1.2.5. Here, the state variable is given by $v = (V, n)$ and the design variable is the doping profile $C$. As the objective functional, we consider

$$J(n,V,C) := \frac{1}{2}\left|\mu_n \int_{\Gamma_0} (\nabla n - n\nabla V).d\nu - I^*\right|^2 + \frac{\alpha}{2}\int_\Omega \left(|\nabla C^* - C|^2 + |C^* - C|^2\right) \, dx.$$

The corresponding design goal is to obtain an outflow current on a contact $\Gamma_0 \subset \partial\Omega$ close to $I^*$, with a doping profile $C$ as close as possible to a reference profile $C^*$.

The Lagrangian in this case is given by (assuming all constants equal one)

$$\begin{aligned} \mathcal{L}(n,V,C;p^1,p^2) = J(n,V,C) + \int_\Omega \Big(&\nabla V.\nabla p^1 + np^1 - Cp^1 \\ &+(\nabla n - n\nabla V).\nabla p^2\Big) \, dx + \int_{\Gamma_0} (\nabla n - n\nabla V)p^2 \ .d\nu \end{aligned}$$

One can show that for small applied voltages (which means appropriate boundary conditions for $V$ with small Dirichlet value), the derivative $\frac{\partial E}{\partial v} = \frac{\partial E}{\partial (V,n)}$ has a continuous inverse. The derivatives are given by

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial n}\ \tilde{n} &= \Big(\int_{\Gamma_0} (\nabla \tilde{u} - \tilde{n}\nabla V) \ .d\nu\Big)\Big(\int_{\Gamma_0} (\nabla n - n\nabla V) \ .d\nu - I^*\Big) \\ &+ \int_\Omega \Big(\tilde{u}p^1 + (\nabla \tilde{u} - \tilde{n}\nabla V).\nabla p^2\Big) \, dx + \int_{\Gamma_0} \Big(\nabla \tilde{n} - \tilde{n}\nabla V 1\Big)p^2.d\nu \\ \frac{\partial \mathcal{L}}{\partial V}\ \tilde{V} &= -\int_{\Gamma_0} (n\nabla \tilde{V}) \ ds\Big(\int_{\Gamma_0} (\nabla n - n\nabla V) - I^*\Big) \\ &+ \int_\Omega \Big(\nabla \tilde{V}\nabla p^1 - n\nabla \tilde{V}\nabla p^2\Big) \, dx - \int_{\Gamma_0} (n\nabla \tilde{V})p^2.d\nu \\ \frac{\partial \mathcal{L}}{\partial C}\ \tilde{C} &= \alpha\int_\Omega (C - C^*)\tilde{C} \, dx + \int \alpha(\nabla C - \nabla C^*).\nabla \tilde{C} \, dx - \int_\Omega \tilde{C}p^2 \, dx. \end{aligned}$$

Using Gauss' Theorem, we can deduce the equivalent system of PDEs for $p^1, p^2$ and $C$ given by

$$\begin{aligned} -\Delta p^2 - \nabla V.\nabla p^2 + \tilde{n} - div\ (p^1 - u\nabla p^2) &= 0 \\ -\alpha\Delta(C - C^*) + \alpha(C - C^*) - p^2 &= 0 \end{aligned}$$

with a boundary condition

$$p^2 = \frac{1}{\Gamma_0}\Big(I^* - \int_{\Gamma_0} (\nabla n - n\nabla V).d\nu\Big).$$

Together with the original model, we obtain a nonlinear system of five nonlinear partial differential equations.

# Chapter 3

# Convexity

Convex problems are an important and interesting class of minimization problems, they exhibit several advantageous properties.

**Definition 3.1.** Let $\mathcal{U}$ be a topological vector space. A functional $J : \mathcal{U} \to \mathbb{R}$ is called convex, if for all $\alpha \in [0,1], u, v \in \mathcal{U}$:

$$J(\alpha u + (1-\alpha)v) \leq \alpha J(u) + (1-\alpha)J(v)$$

A set $\mathcal{C} \subset \mathcal{U}$ is called convex, if for all $\alpha \in [0,1], u, v \in \mathcal{C}$:

$$\alpha u + (1-\alpha)v \in \mathcal{C}.$$

An optimization problem

$$J(u) \to \min_{u \in \mathcal{C}}$$

is called convex, if both $J$ and $\mathcal{C}$ are convex.

A fundamental property of convex problems is that any local minimizer is also a global one.

**Proposition 3.2.** *Let $\overline{u}$ be a local minimizer of a convex optimization problem*

$$J(u) \to \min_{u \in \mathcal{C}}.$$

*Then $\overline{u}$ is a global minimizer.*

*Proof.* Assume that $\overline{u}$ is no global minimizer, i.e. there exists $\hat{u} \in \mathcal{C}$ with $J(\hat{u}) < J(\overline{u})$. Due to convexity of $\mathcal{C}$

$$u_\alpha := \alpha\hat{u} + (1-\alpha)\overline{u} \in \mathcal{C}$$

for all $\alpha \in [0,1]$, and due to convexity of $J$:

$$J(u_\alpha) \leq \alpha J(\hat{u}) + (1-\alpha)J(\overline{u}) < J(\overline{u}).$$

Since $u_\alpha \to \overline{u}$ as $\alpha \to 0$, this is a contradiction to $\overline{u}$ being a local minimizer. □

In order to avoid the constraint set $\mathcal{C}$ in the following, we introduce again the functional $\tilde{J}$,

$$\tilde{J}(u) := \begin{cases} J(u) & \text{if } u \in \mathcal{C} \\ +\infty & \text{else} \end{cases}$$

and consider the unconstrained problem for $\tilde{J}$.

If the functional $J$ is smooth, convexity can be characterized in terms of the second derivative:

**Proposition 3.3.** *Let $J : \mathcal{C} \to \mathbb{R}$ be twice continuously Fréchet-differentiable and let $\mathcal{C}$ be open and convex. Then $J$ is convex on $M$ if and only if*

$$J''(u)(v,v) \geq 0 \quad \forall u \in \mathcal{C}, v \in \mathcal{U}. \tag{3.1}$$

*Proof.* Let $u_1, u_2 \in \mathcal{C}$ be arbitrary and consider the function $g : [0,1] \to \mathbb{R}, \alpha \mapsto J(\alpha u_1 + (1-\alpha)u_2)$. A straight-forward computation shows that

$$g''(\alpha) = J''(\alpha u_1 + (1-\alpha)u_2)(v,v)$$

with $v = u_1 - u_2$. If (3.1) holds, then $g''(\alpha) \geq 0$ for any $\alpha \in [0,1]$. Using a simple Taylor expansion, we can compute

$$\begin{aligned} g(\alpha) &= g(0) + \int_0^\alpha g'(s)\ ds \\ &= g(0) + \alpha g'(0) + \int_0^\alpha \int_0^s g''(t)\ dt\ ds \end{aligned}$$

and,

$$\begin{aligned} g(\alpha) &= g(1) - \int_\alpha^1 g'(s)\ ds \\ &= g(1) - (1-\alpha)g'(1) - \int_\alpha^1 \int_s^1 g''(t)\ dt\ ds \end{aligned}$$

Taking a convex combination of these two expansions, we deduce

$$\begin{aligned} g(\alpha) &= (1-\alpha)g(0) + \alpha g(1) + \alpha(1-\alpha)(g'(0) - g'(1)) \\ &+ (1-\alpha)\int_0^\alpha \int_0^s g''(t)\ dt\ ds + \alpha \int_\alpha^1 \int_s^1 g''(t)\ dt\ ds \\ &= (1-\alpha)g(0) + \alpha g(1) - \alpha(1-\alpha)\int_0^1 g''(t)\ dt \\ &+ (1-\alpha)\int_0^\alpha (\alpha - t)g''(t)\ dt + \alpha \int_\alpha^1 (1-\alpha-t)g''(t)\ dt \\ &= (1-\alpha)g(0) + \alpha g(1) - (1-\alpha)\int_0^\alpha t g''(t)\ dt - \alpha \int_\alpha^1 t g''(t)\ dt \\ &\leq (1-\alpha)g(0) + \alpha g(1) \end{aligned}$$

Hence, $J$ is convex, if $g'' \geq 0$. Assume now that $J$ is convex and let $\alpha = \frac{1}{2}$,

$$u_1 := u + tv, u_2 := u - tv$$

for $t \geq 0$. Then

$$0 \leq \frac{1}{2t^2}\Big(J(u+tv) + J(u-tv) - 2J(u)\Big),$$

for all $t \geq 0$, and if $J$ is twice Fréchet-differentiable, the limit $t \to 0$ exists and implies

$$0 \leq J''(u)(v,v).$$

□

If $J$ is not smooth, we can employ convexity to prove some fundamental properties:

**Proposition 3.4.** *Let $J : \mathcal{U} \to \mathbb{R}$ be convex, where $\mathcal{U}$ is a Banach space. If $J$ is locally bounded around $u$ then $J$ is lower semicontinuous at $u$.*

*Proof.* Let $u_k \to u$. For each $\epsilon > 0$, we can find a sequence $\alpha_k$ such that $\left\| \frac{u-u_k}{\alpha_k} \right\| \leq \epsilon$ and $\alpha_k \to 0$ for $k \to \infty$. Moreover, for $k$ sufficiently large, we have $\|u_k - u\| \leq \epsilon$. Let $\epsilon$ be such that $J$ is bounded in $\overline{B_{2\epsilon}(u)}$ and define

$$v_k := u_k + \frac{u - u_k}{\alpha_k} \in \overline{B_{2\epsilon}(u)}.$$

Due to convexity

$$\begin{aligned} J(u) &\leq \alpha_k J(v_k) + (1-\alpha_k)J(u_k) \\ &\leq 2\alpha_k c + J(u_k), \end{aligned}$$

where $c$ is a bound for $J$ in $\overline{B_{2\epsilon}(u)}$. Hence, we obtain

$$\begin{aligned} J(u) &\leq \liminf_{k\to\infty}(2\alpha_k c + J(u_k)) \\ &= \liminf_{k\to\infty} J(u_k) \end{aligned}$$

which implies lower semicontinuity of $J$ at $u$. □

The above result of convexity and local boundedness implying lower semicontinuity is similar to a classical result for linear operators, where local boundedness implies continuity. This relation is even closer for least-squares functionals

$$J(u) = \frac{1}{2}\|Au - f\|^2,$$

with $A$ being a linear operator. In general, roughly speaking convexity in optimization plays the same role as linearity in solving equations.

## 3.1 Subgradients

Another advantageous property of convex functionals is the possibility to define a generalized gradient. Assume first that $J$ is twice continuously Fréchet-differentiable. As we have shown above, the second derivative $J''(u)$ is positive definite, and therefore

$$\begin{aligned} J(w) &= J(u) + J'(u)(w-u) + \int_0^1 J''(u + t(w-u))(w-u, w-u)\, dt \\ &\geq J(u) + J'(u)(w-u). \end{aligned}$$

I.e., we obtain a global estimate for $J$ using the local values $J(u)$ and $J'(u)$ only. This global estimate is the starting point for the definition of the *subgradient*:

**Definition 3.5.** Let $\mathcal{U}$ be a Banach space and $J : \mathcal{U} \to \mathbb{R}$ be convex. Then the subgradient $\partial J$ at a point $u$ is defined as:

$$\partial J(u) := \{p \in \mathcal{U}^* | J(w) \geq J(u) + < p, w - u >, \forall w \in \mathcal{U}\}.$$

Note that the subgradient is now a set of elements in $\mathcal{U}^*$ instead of a single element. This can be seen in the simple example of $J : \mathbb{R} \to \mathbb{R}, u \mapsto |u|$. If $u > 0$, then $J(u) = u$ and we can find $w_1, w_2$ satisfying $0 < w_1 < u < w_2$. Then $p \in \partial J(u)$ implies

$$(1 - p)(w_1 - u) \geq 0 \quad \text{and} \quad (1 - p)(w_2 - u) \geq 0$$

which is equivalent to $p \leq 1 \leq p$. On the other hand, the estimate

$$|w| \geq u + w - u = w$$

holds for all $w \in \mathbb{R}$. Thus, $\partial J(u) = \{1\}$.

In an analogous way we can show that $\partial J(u) = \{-1\}$ for $u < 0$. It remains to consider $u = 0$, where $p \in \partial J(u)$ is equivalent to

$$|w| \geq pw \quad \forall w \in \mathbb{R}.$$

This inequality is satisfied if and only if $|p| \leq 1$. Hence, we have shown that for $J(u) = |u|$,

$$\partial J(u) = \begin{cases} \{+1\} & \text{if } u > 0 \\ [-1, 1] & \text{if } u = 0 \\ \{-1\} & \text{if } u < 0 \end{cases}$$

From this example one may conjecture that the subgradient is only set-valued at points $u$, where $J$ is not differentiable.

**Proposition 3.6.** *If $J : \mathcal{U} \to \mathbb{R}$ is convex, and $J$ is Fréchet-differentiable at $u$, then*

$$\partial J(u) = \{J'(u)\}.$$

*Proof.* Let $p \in \partial J(u)$. Then for each $t > 0$

$$\begin{aligned} \frac{J(u + tv) - J(u)}{t} &\geq < p, v > \\ \frac{J(u - tv) - J(u)}{t} &\geq - < p, v >, \end{aligned}$$

which implies in the limit $t \to 0$

$$< p, v > \leq J'(u)v \leq < p, v >$$

for all $v \in \mathcal{U}$. Hence, the linear functionals defined by $p$ and $J'(u)$ coincide, i.e. $p = J'(u)$ in $\mathcal{U}^*$. □

The subgradient can be used to obtain a local optimality condition, which is necessary and sufficient for convex problems.

**Theorem 3.7.** *Let $\mathcal{U}$ be a Banach space and let $J : \mathcal{U} \to \mathbb{R}$ be convex. Then each local minimum is a global minimum. Moreover, $\overline{u} \in \mathcal{U}$ is a minimizer if and only if*

$$0 \in \partial J(\overline{u}).$$

*Proof.* We have shown above that each local minimum is a global one. If $0 \in \partial J(\overline{u})$, then

$$\begin{aligned} J(w) &\geq J(\overline{u}) + < 0, w - \overline{u} > \\ &= J(\overline{u}) \end{aligned}$$

and thus, $\overline{u}$ is a global minimizer. Assume that $0 \notin J(\overline{u})$. Then there exists a $w \in \mathcal{U}$ with

$$J(w) < J(\overline{u}) + < 0, w - \overline{u} >= J(\overline{u}),$$

and hence $\overline{u}$ cannot be a minimizer □

## 3.2 Duality

A frequently used technique for convex problems is duality, which is based on replacing the optimization problem by an equivalent problem in the dual space $\mathcal{U}^*$ involving a dual functional. For this sake, we introduce convex conjugates.

**Definition 3.8.** Let $J : \mathcal{U} \to \overline{\mathbb{R}}$ (not necessarily convex). Then the convex conjugate (or polar) function $J^* : \mathcal{U}^* \to \overline{\mathbb{R}}$ is defined by

$$J^*(p) = \sup_{u \in \mathcal{U}} \Big( < p, u > - J(u) \Big).$$

As an example, consider the indicator function of a convex set $K$,

$$\chi_K(u) := \begin{cases} 0 & \text{if } u \in K \\ +\infty & \text{if } u \notin K. \end{cases}$$

This function is added to $J$, when transforming a constrained problem to an unconstrained problem. A simple calculation shows

$$J^*(p) = \sup_{u \in K} < p, u >,$$

i.e. $J^*$ is the support function of $K$. For a second example, let $\mathcal{U}$ be a Hilbert space and

$$J(u) = \frac{1}{2}\|u\|^2.$$

Then the supremum over $< p, u > - J(u)$ is attained at $\overline{u} = p$, and the corresponding function value is

$$J^*(p) =< p, p > - \frac{1}{2}\|p\|^2 = + \frac{1}{2}\|p\|^2,$$

i.e. $J^* = J$.

$J^*$ is always convex, even for general $J$:

**Proposition 3.9.** *Let $\mathcal{U}$ be a Banach space and $J : \mathcal{U} \to \overline{\mathbb{R}}$. Then $J^*$ is convex.*

*Proof.* For $p, q \in \mathcal{U}^*$, we have

$$\begin{aligned} J^*(\alpha p + (1-\alpha)q) &= \sup_{u\in\mathcal{U}} \Big(\alpha < p, u > +(1-\alpha) < q, u > -J(u)\Big) \\ &= \sup_{u\in\mathcal{U}} \Big(\alpha(< p, u > -J(u)) + (1-\alpha)(< q, u > -J(u))\Big) \\ &\leq \sup_{u,v\in\mathcal{U}} \Big(\alpha(< p, u > -J(u)) + (1-\alpha)(< q, v > -J(v))\Big) \\ &= \alpha \sup_{u\in\mathcal{U}} \Big( < p, u > -J(u)\Big) + (1-\alpha) \sup_{v\in\mathcal{U}} \Big( < q, v > -J(v)\Big) \\ &= \alpha J^*(p) + (1-\alpha)J^*(q) \end{aligned}$$

□

By iterating the definition, we also obtain the bipolar

$$J^{**} = (J^*)^* : \mathcal{U}^{**} \to \overline{\mathbb{R}}.$$

If $\mathcal{U}$ is a reflexive Banach space, then $\mathcal{U}^{**} = \mathcal{U}$ and we can consider the difference between the functionals $J$ and $J^{**}$. Note that $J^{**}$ is convex and therefore cannot equal $J$ unless $J$ is convex.

**Proposition 3.10.** *Let $\mathcal{U}$ be a reflexive Banach space. Then $J^{**}$ is the maximal convex functional below $J$ (also called convex envelope), i.e. $J^{**}(u) \leq J(u)$, $\forall u \in \mathcal{U}$ and $F(u) \leq J^{**}(u)$, $\forall u \in \mathcal{U}$, if $F(u) \leq J(u)$, $\forall u \in \mathcal{U}$, and $F$ is convex. In particular, $J^{**} = J$ if and only if $J$ is convex.*

*Proof.* We start by computing

$$\begin{aligned} J^{**}(u) &= \sup_{p\in\mathcal{U}^*} \Big( < p, u > -J^*(p)\Big) \\ &= \sup_{p\in\mathcal{U}^*} \left( < p, u > - \sup_{v\in\mathcal{U}} \Big( < p, v > -J(v)\Big)\right) \\ &= \sup_{p\in\mathcal{U}^*} \inf_{v\in\mathcal{U}} \Big( < p, u - v > +J(v)\Big). \end{aligned}$$

Since for any $p \in \mathcal{U}^*$

$$\begin{aligned} \inf_{v\in\mathcal{U}} \Big( < p, u - v > +J(v)\Big) &\leq < p, u - u > +J(u) \\ &= J(u), \end{aligned}$$

we may conclude that

$$J^{**}(u) \leq J(u).$$

Now assume that $F$ is convex, and let $q \in \partial F(u)$ for $u \in \mathcal{U}$. Then $F(v) \geq F(u) - < q, v - u >$ and hence,

$$\begin{aligned} F^{**}(u) &= \sup_{p\in\mathcal{U}^*} \inf_{v\in\mathcal{U}} \Big( < p, u - v > +F(v)\Big) \\ &\geq \sup_{p\in\mathcal{U}^*} \inf_{v\in\mathcal{U}} \Big( < p - q, u - v > +F(u)\Big) \\ &\geq \inf_{v\in\mathcal{U}} \Big( < q - q, u - v > +F(u)\Big) = F(u) \end{aligned}$$

Thus $F(u) \geq F^{**}(u) \geq F(u)$, and consequently $F^{**}(u) = F(u)$. If $F(u) \leq J(u)$ then

$$\begin{aligned} F(u) = F^{**}(u) &= \sup_{p \in \mathcal{U}^*} \inf_{v \in \mathcal{U}} \Big( < p, u - v > + F(v) \Big) \\ &\leq \sup_{p \in \mathcal{U}^*} \inf_{v \in \mathcal{U}} \Big( < p, u - v > + J(v) \Big) \\ &= J^{**}(v), \end{aligned}$$

which implies the assertion. □

In order to use duality, we introduce a parametrized familiy $\Phi : \mathcal{U} \times Y \to \overline{\mathbb{R}}$ such that

$$\Phi(u, 0) = J(u), \quad \forall u \in \mathcal{U}.$$

For each $p \in Y$ we now consider

$$\Phi(u, p) \to \min_{u \in \mathcal{U}} \quad (P_p)$$

which is just the minimization of $J$ for $p = 0$. If we denote by $\Phi^* : V^* \times Y^* \to \overline{\mathbb{R}}$, the convex conjugate

$$\Phi^*(u^*, p^*) = \sup_{u,p} \Big( < w^*, u > + < p^*, p > - \Phi(u, p) \Big)$$

then we can define a dual problem

$$-\Phi^*(0, p^*) \to \max_{p^* \in Y^*} \quad (P^*)$$

with respect to $\Phi$. If we denote by $(P)$ the primal problem $(P_0)$, then

$$-\infty < \sup P^* \leq \inf P < \infty,$$

if $\Phi(., 0)$ and $\Phi^*(0, .)$ are proper. This can be seen easily from

$$\begin{aligned} -\Phi^*(0, p^*) &= \inf_{(u,p)} \Big( \Phi(u, p) - < p^*, p > \Big) \\ &\leq \Phi(u, p) - < 0, p > = \Phi(u, p) \end{aligned}$$

for all $u \in \mathcal{U}$, $p \in Y$, $p^* \in Y^*$. By iteration we can also define a bidual problem

$$\Phi^{**}(u, 0) \to \min_{u \in \mathcal{U}}.$$

The primal and dual problem are linked by a so-called *extremal relation.*

**Theorem 3.11.** *Let $\Phi : V \times Y \to \overline{\mathbb{R}}$ be convex, then the following statements are equivalent.*

*(i)* $(P)$ *and* $(P^*)$ *possess solutions* $\overline{u}$ *and* $\overline{p}^*$*, and* $\inf P = \sup P^*$.

*(ii)* $\Phi(\overline{u}, 0) + \Phi^*(0, \overline{p}^*) = 0$

*(iii)* $(0, \overline{p}^*) \in \partial\Phi(\overline{u}, 0)$, $(\overline{u}, 0) \in \partial\Phi(0, \overline{p}^*)$

*Proof.* Assume (i), then

$$\Phi(\overline{u},0) = \inf P = \sup P^* = -\Phi^*(0,\overline{p}^*)$$

and thus, (ii) holds. If (ii) holds, then

$$\sup P^* \leq \Phi(\overline{u},0) = -\Phi^*(0,\overline{p}^*) \leq \inf P$$

implies (i). The equivalence of (i) and (iii) can be shown using the standard properties of the subgradient. □

There are many different applications of these duality concepts, which can be found e.g. in the monographs by Aubin and by Ekeland and Temam. We shall highlight only a special case, sometimes called Fenchel duality, in the following. Assume that

$$J(u) = F(u) + G(Au),$$

where $F : \mathcal{U} \to \overline{\mathbb{R}}, G : V \to \overline{\mathbb{R}}$, are convex functionals, and $A : \mathcal{U} \to V$ is bounded linear operator.

In this case, we introduce the perturbations

$$\Phi(u,p) := F(u) + G(Au - p).$$

The dual problem is obtained with

$$\Phi^*(0,p^*) = \sup_{u\in\mathcal{U},p\in V} \Big( < p^*, p > -F(u) - G(Au-p)\Big).$$

If, for fixed $u$, we set $q = Au - p$, we obtain

$$\begin{aligned}
\Phi^*(0,p^*) &= \sup_{u\in\mathcal{U}} \sup_{p\in V} \Big( < p^*, p > -F(u) - G(Au-p)\Big) \\
&= \sup_{u\in\mathcal{U}} \sup_{q\in V} \Big( < p^*, Au - q > -F(u) - G(q)\Big) \\
&= \sup_{u\in\mathcal{U}} \sup_{q\in V} \Big( < A^*p^*, u > -F(u) - < p^*, q > -G(q)\Big) \\
&= \sup_{u\in\mathcal{U}} \Big( < A^*p^*, u > -F(u)\Big) + \sup_{q\in V} \Big( < -p^*, q > -G(q)\Big) \\
&= F^*(A^*p^*) + G^*(-p^*)
\end{aligned}$$

Hence, the dual problem is given by

$$-F^*(A^*p^*) - G^*(-p^*) \to \max_{p^*}.$$

We also revisit the extremality condition, which simplifies to

$$\begin{aligned}
0 &= F(\overline{u}) + G(A\overline{u}) + F^*(A^*\overline{p}^*) + G^*(-\overline{p}^*) \\
&= [F(\overline{u}) + F^*(A^*\overline{p}^*) - < A^*\overline{p}^*, \overline{u} >] \\
&+ [G(A\overline{u}) + G^*(-\overline{p}^*) - < -p^*, A\overline{u} >]
\end{aligned}$$

Each term in the square brackets is nonnegative, so the fact that their sum is zero implies that both are zero. Hence

$$\begin{aligned} F(\overline{u}) - < A^*\overline{p}^*, \overline{u} > &= \inf_{u \in \mathcal{U}} F(u) - < A^* p^*, u > \\ G(-\overline{p}^*) - < -\overline{p}^*, A\overline{u} > &= \inf_{p \in V} G(p) - < p, A\overline{u} > . \end{aligned}$$

Therefore, the optimality condition implies

$$\begin{aligned} 0 &\in \partial(F - < A^*\overline{p}^*, . >)(\overline{u}) \\ 0 &\in \partial(G - < ., A\overline{u} >)(\overline{p}^*) \end{aligned}$$

and since

$$\partial(F - < A^*\overline{p}^*, . >)(\overline{u}) = \partial F(\overline{u}) - \{A^*\overline{p}^*\},$$

we conclude

$$A^*\overline{p}^* \in \partial F(\overline{u}),$$

and analogously

$$-\overline{p}^* \in \partial G(\overline{u}).$$

If $F$ and $G$ are convex and locally bounded, one can show that

$$\sup P^* = \inf P,$$

and hence, we may use the above duality.

We apply the results to bounded variation denoising, where

$$J(u) = \frac{1}{2}\int_\Omega |u - f|^2 \, dx + \alpha \sup_{\substack{g \in C_0^\infty, \\ \|g\|_\infty \le 1}} \int_\Omega u \; div \; g \; dx.$$

If we formally set

$$\begin{aligned} F(u) &= \frac{1}{2}\int_\Omega |u - f|^2 \, dx, \\ G(p) &= \int_\Omega \|p\| \, dx, \end{aligned}$$

and $A = \alpha\nabla$, we can compute the conjugates

$$\begin{aligned} F^*(v) &= \sup_{u \in L^2(\Omega)} \int_\Omega (uv - \frac{1}{2}|u - f|^2) \, dx \\ &= \frac{1}{2}\int_\Omega |v + f|^2 \, dx - \frac{1}{2}\int_\Omega f^2 \, dx \end{aligned}$$

and

$$\begin{aligned} G^*(q) &= \sup_{u \in L^2(\Omega)} \Big( \int_\Omega uq - \|u\| \, dx \Big) \\ &= \begin{cases} 0 & \text{if } \|q(x)\| \le 1 \\ -\infty & \text{else} \end{cases} \end{aligned}$$

Thus, the dual problem is given by (noticing that $(\alpha\nabla)^* = -\alpha\ div$) the maximization of

$$\begin{aligned} -J^*(p) &= -F^*(A^*p) - G^*(-p) \\ &= -\frac{1}{2}\int_\Omega |\alpha\ div\ p - f|^2\ dx + \frac{1}{2}\int_\Omega f^2\ dx - \chi_{[.1,1]}(p), \end{aligned}$$

where $\chi_{[-1,1]}$ denotes the indicator function

$$\chi_{[-1,1]}(p) = \begin{cases} 0 & \text{if } \|p(x)\| \leq 1\ a.e. \\ +\infty & \text{else.} \end{cases}$$

This problem is equivalent to the constrained miminization problem

$$\frac{1}{2}\int_\Omega |\alpha\ div\ p - f|^2\ dx \to \min_p$$

subject to

$$-1 \leq \|p(x)\| \leq 1 \text{ for a.e. } x \in \Omega.$$

As noted above, the derivation of the dual problem is purely formal, but it can be made rigorous, if we start from the dual problem and compute its conjugates. One then obtains again the original problem (as the bi-dual) and hence, the optimality conditions still hold. In particular, we obtain optimality relation (using subgradients):

$$u = f - \alpha div\ p.$$

# Chapter 4

# Numerical Methods

In the following, we shall discuss some methods for the numerical solution of infinite-dimensional optimization problems. We start with the fundamental issue of discretizing infinite-dimensional problems. Then we discuss the case of non-constrained optimization of smooth functionals. Finally, we discuss numerical methods for constrained problems.

## 4.1 Discretization

There are two different approaches to the solution of infinite-dimensional optimization problems, which are subject to an ongoing discussion concerning their advantages and disadvantages.

(i) "Discretize-then-optimize": The idea of this approach is to discretize the optimization problem directly, which leads to a nonlinear programming problem. The main advantage of this approach consists in the possibility to use available nonlinear programming methods, its disadvantage is a lack of quantitative approximation results for nonlinear problems in general.

(ii) " Optimize-then-discretize": The idea of this approach is to formulate the optimization method in infinite-dimensional spaces, and use discretization only for the solution of (linear or quadratic) subproblems and for evaluations of the objective functional. The main advantage of this approach is that quantitative estimates for convergence of optimization methods can be combined with error estimates for the discretization of subproblems, to obtain estimates on the total error.

In real-life applications of optimal design or optimal control problems, the discretization strategy is mainly determined by the solvers available for the state equations in general. Since the design and development of computational methods for state equations arising from multi-physics models is complicated and expensive, any optimization procedure should aim at incorporating existing software and avoid subproblems, for which no computational methods are available.

The use of a given discretization strategy leaves at most the discretization of the design variables open. In order to obtain reasonable numerical methods, the discretization of the design variables should match somehow the discretization of the state equation. Consider for example the one-dimensional version of the boundary control problem from section 1.2.2.

If we use time steps $0 < t_1 < t_2 < ... < t_n = T$ to discretize the parabolic equation, it is unreasonable to discretize the control variable $u$ on a completely different grid. Typical possibilities for the discretization of $u$ are to use the same grid points, a subset of these grid points, or the midpoints $\frac{t_1}{2}, \frac{t_1+t_2}{2}, ...\frac{t_{n-1}+t_n}{2}$.

Below we shall discuss several methods incorporating one of the two approaches, in many cases the resulting discrete problems are very similar anyway.

We start with a discussion of methods for unconstrained problems, mainly following the second approach, i.e. we formulate optimization methods in an infinite-dimensional setting (noticing, however, that all results apply in particular to the case of $\mathcal{U} = \mathbb{R}^n$).

## 4.2 Methods for Unconstrained Optimization

In the following we shall assume that $\mathcal{U}$ is a Hilbert space unless further noticed.

### 4.2.1 Gradient Methods

Many dynamical models in physics are based on the idea that a system follows a gradient flow with respect to its energy. Consider for example heat conduction. Here, the thermal energy is given by

$$E(u) = \frac{1}{2}\int |\nabla u|^2 \, dx.$$

The gradient flow is defined by $\frac{\partial u}{\partial t} = -E'(u)$, which yields in this case the heat equation

$$\frac{\partial u}{\partial t} = \Delta u.$$

In order to obtain an optimization method, we can use this idea of introducing a gradient flow in a Hilbert space $\mathcal{U}$, as

$$\frac{\partial u}{\partial t} = -J'(u),$$

where $J'(u) \in \mathcal{U}$ is the element in the Hilbert space that can be associated with the gradient of $J$ at $u$. In other words, we define the evolution by

$$\left\langle \frac{\partial u}{\partial t}, v \right\rangle = -J'(u)v \quad \forall v \in \mathcal{U},$$

where $< . \, , \, . >$ denotes the scalar product in $\mathcal{U}$.

The evolution of the objective corresponding to this gradient flow is given by:

$$\frac{\partial}{\partial t}\Big(J(u)\Big) = J'(u)\frac{\partial u}{\partial t} = -\|\frac{\partial u}{\partial t}\|^2 \leq 0,$$

i.e. the objective is decreasing and $\frac{\partial}{\partial t}\Big(J(u)\Big) = 0$ if and only if $\frac{\partial u}{\partial t} = 0$. Moreover, the gradient flow structure implies that $\frac{\partial u}{\partial t} = 0$ if and only if $J'(u) = 0$, i.e. $u$ is a stationary point. Consequently we can expect the gradient flow to decrease the objective until the evolution arrives at a stationary point.

In order to obtain an iterative optimization algorithm we use an explicit time discretization of the flow, namely

$$u_{k+1} = u_k - \tau_k J'(u_k), \tag{4.1}$$

with $\tau_k > 0$ corresponding to a time step. (4.1) is called the gradient method.

It is obvious that the gradient method only makes sense for appropriate (small) choice of the time step $\tau_k$, since explicit time discretizations with too large time steps are not stable. Since we are only interested in the minimization of the objective, but not in the accurate approximation of the solution of the gradient flow, we can base the step size control purely on the aim of achieving a suitable descent in the objective. A classical way to do this are the so-called Armijo-Goldstein rules. They are based on the rationale of comparing the "effective descent"

$$D_{eff}(\tau) = J(u_k + \tau s) - J(u_k)$$

and the "expected descent"

$$D_{exp}(\tau) = \tau J'(u_k)s$$

where $s = -J'(u_k)$ in this case. For $\tau_k$ sufficiently small, they are related by the Taylor formula

$$D_{eff}(\tau) = D_{exp}(\tau) + o(\tau).$$

Therefore we can test if

$$\alpha D_{exp}(\tau) \leq D_{eff}(\tau) \leq \beta D_{exp}(\tau) \tag{4.2}$$

with constants $0 < \beta < \alpha < 1$.

If

$$D_{eff}(\tau) > \beta D_{exp}(\tau),$$

we can argue that $\tau$ is too large (because this cannot be true for $\tau$ sufficiently small) and decrease $\tau$. If

$$D_{eff}(\tau) < \alpha D_{exp}(\tau),$$

we can argue that $\tau$ could still be increased to obtain a sufficient descent and try a larger $\tau$. We finally accept the step size $\tau_k = \tau$, if it satisfies (4.2). Typical choices for the constants in this rule are

$$\alpha \approx 0.9, \quad \beta \approx 0.1.$$

In order to obtain a termination of the stepsize selection one has to make sure that the strategy of increasing and decreasing $\tau$ are different. E.g. one should not divide by two if the first condition is violated and multiply by two if the second is violated. Typically, one divides by two in the first case, but multiplies by 1.5 in the second case.

We finally notice that the Armijo-Goldstein rule is not restricted to the gradient method, but can be applied to any method that yields a *descent direction*, i.e. an element $s \in \mathcal{U}$ satisfying

$$J'(u_k)s < 0.$$

Using the Armijo-Goldstein rule, one can prove global convergence of the gradient method, i.e. convergence to a stationary point for any starting value.

**Theorem 4.1.** *Let $J$ be twice continuously Fréchet-differentiable and weakly lower semicontinuous on a Hilbert space $\mathcal{U}$. Moreover, let the sets*

$$\{u \in \mathcal{U} | J(u) \leq M\}$$

*be bounded in $\mathcal{U}$ for each $M \in \mathbb{R}$ and empty for $M$ sufficiently small. Then the sequence $(u_k)$ generated by the gradient method with Armijo-Goldstein line search has a weakly convergent subsequence, whose limit is a stationary point.*

*Proof.* First of all, since the gradient method is a descent method, we obtain

$$J(u_k) \leq J(u_0)$$

for all $k \geq 0$. Thus, the sequence $(u_k)$ is bounded and therefore contains a weakly convergent subsequence $(u_{k_l})$ with limit $\overline{u}$. Due to the Armijo-Goldstein rules, we obtain that

$$\begin{aligned}
\sum_{k=0}^{N} \|u_{k+1} - u_k\|^2 &= -\sum_{k=0}^{N} \tau_k J'(u_k)(u_{k+1} - u_k) \\
&\leq +\frac{1}{\beta} \sum_{k=0}^{N} \Big( J(u_k) - J(u_{k+1}) \Big) \\
&= \frac{1}{\beta} \Big( J(u_0) - J(u_{N+1}) \Big) \\
&\leq \frac{1}{\beta} \Big( J(u_0) - \inf_u J(u) \Big) =: p.
\end{aligned}$$

Since $p$ is independent of $N$, we obtain for $N \to \infty$

$$\sum_{l=0}^{\infty} \|u_{k_l+1} - u_{k_l}\|^2 \leq \sum_{k=0}^{\infty} \|u_{k+1} - u_k\|^2 \leq p.$$

Hence, there exists a subsequence of $(u_{k_l})$, without restriction of generality $(u_{k_l})$ itself, such that

$$\|\tau_{k_l} J'(u_{k_l})\| = \|u_{k_l+1} - u_{k_l}\| \to 0.$$

Since $J$ is twice semicontinuously Fréchet-differentiable, there exists a constant $c < 0$ such that

$$J''(u_{k_l})(v, v) \leq c\|v\|^2, \quad \forall v \in \mathcal{U}.$$

Thus, the second Armijo-Goldstein rule implies

$$\begin{aligned}
\alpha J'(u_{k_l})(u_{k_l+1} - u_{k_l}) &\leq J(u_{k_{l+1}}) - J(u_{k_l}) \\
&\leq J'(u_{k_l})(u_{k_l+1} - u_{k_l}) + \frac{c}{2}\|u_{k_l+1} - u_{k_l}\|^2
\end{aligned}$$

Inserting $u_{k_{l+1}} - u_{k_l} = -\tau_{k_l} J'(u_{k_l})$ implies

$$(1 - \alpha)\tau_{k_l} \|J'(u_{k_l})\|^2 \leq \frac{c}{2} \tau_{k_l}^2 \|J'(u_{k_l}\|^2.$$

Hence, either $J'(u_{k_l}) = 0$ or

$$\tau_{k_l} \geq \frac{2(1 - \alpha)}{c}.$$

If $J'(u_{k_l}) = 0$, the algorithm has arrived at a stationary point and will stop, i.e., $u_j = u_{k_l}$ for all $j \geq k_l$, and convergence is trivial. In the second case, $\tau_{k_l}$ is uniformly bounded away from zero, and therefore $\|J'(u_{k_l})\| \to 0$, which implies that $J'(\overline{u}) = 0$, i.e. the limit $\overline{u}$ is a stationary point. □

From the analysis of the gradient method and the Armijo-Goldstein rules, one observes that one could choose $s = -A_k J'(u_k)$ as a search direction instead, where $A_k$ is a positive definite bounded linear operator. In this case $s$ is still a descent direction, since

$$\begin{aligned} J'(u_k)s &= - < J'(u_k), A_k J'(u_k) > \\ &\leq -\lambda(A_k)\|J'(u_k)\|^2, \end{aligned}$$

where $\lambda(A_k)$ is the smallest eigenvalue of $A_k$. Motivated by the sufficient second order optimality condition, we can expect $J''(u_k)$ to be positive definite, and try the choice

$$s = -\Big(J''(u_k)\Big)^{-1} J'(u_k),$$

which is the so-called Newton method. We consider the convergence speed of the Newton method locally around a minimum that satisfies the sufficient second order conditions. The error between the minimizer $\overline{u}$ and the iterate $u_{k+1}$ is given by

$$\begin{aligned} \|\overline{u} - u_{k+1}\|^2 &= < \overline{u} - u_k + u_k - u_{k+1}, \overline{u} - u_{k+1} > \\ &= < \overline{u} - u_k + \Big(J''(u_k)\Big)^{-1} J'(u_k), \overline{u} - u_{k+1} > \end{aligned}$$

If $\|\overline{u} - u_k\|$ is sufficiently small, then

$$J'(\overline{u})v \geq J'(u_k)v + J''(u_k)(v, \overline{u} - u_k) - c\|v\| \ \|\overline{u} - u_k\|^2$$

for some constant $c > 0$. Since $J'(\overline{u}) = 0$, this implies

$$\begin{aligned} &< J''(u_k)^{-1} J'(u_k), \overline{u} - u_{k+1} > \\ &\qquad \leq - < \overline{u} - u_k, \overline{u} - u_{k+1} > + c\|\overline{u} - u_k\|^2 \|J''(u_k)^{-1}(\overline{u} - u_{k+1})\| \\ &\qquad \leq - < \overline{u} - u_k, \overline{u} - u_{k+1} > + \frac{c}{\lambda(J''(u_k))}\|\overline{u} - u_k\|^2 \|\overline{u} - u_{k+1}\| \end{aligned}$$

where $\lambda(J''(u_k))$ denotes the minimal eigenvalue of $J''(u_k)$. Inserting this estimate into the above relation for the error we obtain

$$\|\overline{u} - u_{k+1}\| \leq \frac{c}{\lambda(J''(u_k))}\|\overline{u} - u_k\|^2,$$

i.e. Newton's method is locally quadratic convergent.

Of course, the fast local convergence is an advantage of Newton's method, but there are some severe dawbacks. First of all, the operator $J''(u)$ needs not to be positive definite if $u$ is still far away from the solution and thus, Newton's method need not be a descent method (and consequently might not converge). Secondly, the evaluation of $J''(u)$ may be very expensive, e.g. in optimal control problems. In order to overcome these disadvantages while keeping good local convergence properties, Quasi-Newton methods have been introduced. Their common property is that the search direction is computed as

$$s = -A_k^{-1} J'(u_k),$$

where $A_k$ is a positive definite operator satisfying the "secant condition"

$$A_k(u_k - u_{k-1}) = J'(u_k) - J'(u_{k-1}).$$

For the case $\mathcal{U} = \mathbb{R}$, this defines the secant-method, i.e.

$$A_k = \frac{J'(u_k) - J'(u_{k-1})}{u_k - u_{k-1}}.$$

In particular in infinite-dimensional applications, the secant condition leaves most of the operator $A_k$ undefined and therefore additional criteria have to be introduced. It is common to use $A_0 = I$ as a starting value, i.e. the first step of the method coincides with the gradient method. During the iteration procedure, $A_k$ is constructed by modifying the preceding operator $A_{k-1}$. A somehow minimal modification of the operator would be a symmetric rank-one update, i.e.

$$A_k v = A_{k-1} v + \alpha < p, v > p \qquad \forall v \in \mathcal{U},$$

for some $p \in \mathcal{U},\ \alpha \in \mathbb{R}$. If we insert this modification into the second condition, it yields

$$\begin{aligned} A_k(u_k - u_{k-1}) &= A_{k-1}(u_k - u_{k-1}) + \alpha < p, u_k - u_{k-1} > p \\ &= J'(u_k) - J'(u_{k-1}) \end{aligned}$$

as an equation for $p$. With the notation $v = u_k - u_{k-1}$ and $F = J'(u_k) - J'(u_{k-1}) - A_{k-1}(u_k - u_{k-1})$ we obtain

$$p = F, \quad \alpha = \frac{1}{< F, v >}.$$

Since the term in the denominator can be zero, $p$ is not well-defined in general, so that this approach encounters difficulties. As a second attempt, we try to use a symmetric rank-two update

$$A_k v = A_{k-1} v + \alpha < p, v > p + \beta < q, v > q.$$

A reasonable rank-two update satisfying the secant condition is given by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method

$$\begin{aligned} p = \frac{y}{\|y\|} \quad &, \quad q = \frac{A_{k-1} s}{\|A_{k-1} s\|} \\ \alpha = \frac{\|y\|^2}{< y, s >} \quad &, \quad \beta = -\frac{\|A_{k-1} s\|^2}{< s, A_{k-1} s >} \end{aligned}$$

with $s = u_k - u_{k-1}$, $y = J'(u_k) - J'(u_{k-1})$. Since the condition number of $A_k$ can increase strongly during the BFGS-iteration one either uses a restart with $A_k = I$ after some iterations or applies the limited memory BFGS method (L-BFGS) using

$$A_k v = v + \sum_{j=k-L+1}^{k} \Big( \alpha_j < p_j, v > p_j + \beta_j < q_j, v > q_j \Big)$$

for $k > L$.

### 4.2.2 Application to Optimal Design

In the following we consider a typical optimal design problem of the form

$$J(u, v) \to \min_{(u,v) \in \mathcal{U} \times \mathcal{V}}$$

subject to

$$e(u, v) = 0,$$

where $e : \mathcal{U} \times \mathcal{V} \to \mathcal{W}$ is a Fréchet-differentiable nonlinear operator. Under typical conditions, the equality constraint admits a unique solution $v = v(u)$, if the design variable $u$ is fixed. Therefore, it is reasonable to assume that $\frac{\partial e}{\partial v}$ is continuously invertible for all $(u, v) \in \mathcal{U} \times \mathcal{V}$. Consequently, we may consider the reduced functional

$$\tilde{J}(u) := J(u, v(u)),$$

where $v(u)$ is implicitely defined via

$$e(u, v(u)) = 0.$$

In this case, the chain rule shows that

$$\tilde{J}'(u)w = \frac{\partial J}{\partial u}(u, v(u))w + \frac{\partial J}{\partial v}(u, v(u))[v'(u)w],$$

where $v'(u)w$ is given by

$$v'(u)w = -\left(\frac{\partial e}{\partial v}(u, v(u))\right)^{-1} \frac{\partial e}{\partial u}(u, v(u))w.$$

Similar to the example in optimal control, whose derivative we have computed in Section 2, the computation of $v'(u)w$ involves the solution of a linearized equation for each $w \in \mathcal{U}$, which is very expensive in general. However, one observes that for computing $\tilde{J}'(u)w$, one does not need $v'(u)w$ istelf, but only the linear functional $\frac{\partial J}{\partial v}(u, v(u))[v'(u)w]$. By interpreting $\frac{\partial J}{\partial v}(u, v(u)) : \mathcal{V} \to \mathbb{R}$ as an element of $\mathcal{V}^*$, we can make use of this structure to obtain

$$\begin{aligned}
\frac{\partial J}{\partial v}(u, v(u))v'(u)w &= -\left\langle \frac{\partial J}{\partial v}(u, v(u))v'(u)w, \left(\frac{\partial e}{\partial v}(u, v(u))\right)^{-1} \frac{\partial e}{\partial u}(u, v(u))w \right\rangle \\
&= -\left\langle \frac{\partial e}{\partial u}(u, v(u))^* \left(\frac{\partial e}{\partial v}(u, v(u))^*\right)^{-1} \frac{\partial J}{\partial v}(u, v(u)), w \right\rangle,
\end{aligned}$$

where $\frac{\partial e}{\partial u}(u, v(u))^* : \mathcal{W}^* \to \mathcal{U}^*$, and $\frac{\partial e}{\partial v}(u, v(u))^* : \mathcal{W}^* \to \mathcal{V}^*$ are the adjoint operators corresponding to the partial derivatives of $e$. Hence, the Fréchet-derivative of the reduced functional $\tilde{J}$ is given by

$$\tilde{J}'(u) = \frac{\partial J}{\partial u}(u, v(u)) - \frac{\partial e}{\partial u}(u, v(u))^* \left(\frac{\partial e}{\partial v}(u, v(u))^*\right)^{-1} \frac{\partial J}{\partial v}(u, v(u)).$$

This way of computing the gradient is called the "adjoint method", it only involves the solution of one linear system with the adjoint operator $\frac{\partial e}{\partial v}(u, v(u))^*$ and right-hand side $\frac{\partial J}{\partial v}(u, v(u))$.

### 4.2.3 Multi-level Optimization

As discussed above, the numerical solution of infinite-dimensional optimization methods involves discretization at some point, either a direct discretization of the nonlinear problem or a discretization of the linear equations arising in each step of an iteration procedure. A standard way to discretize optimization problems is the Ritz-method, i.e. one minimizes

$$J(u_h) \to \min_{u_h \subset \mathcal{U}_h}$$

where $\mathcal{U}_h$ is a finite-dimensional subspace of $\mathcal{U}$. This method corresponds to Galerkin-type discretizations of partial differential equations. If $J$ satisfies the standard assumptions for the existence result, i.e. lower semicontinuity and compactness of level sets, then one can show that the discretized solutions converge to the solutions of the original problem. More precisely, if $\mathcal{U}_1 \subset \mathcal{U}_2 \subset ... \subset \mathcal{U}_k \subset ...\mathcal{U}$ is a sequence of finite-dimensional spaces such that

$$\bigcup_{k\in\mathbb{N}} \mathcal{U}_k = \mathcal{U},$$

then a sequence $(u_k)$ of global minimizers of $J$ in $\mathcal{U}_k$ contains a subsequence converging to a global minimizer of $J$ in $\mathcal{U}$. A computational realization of this result is obtained in cascadic multi-level methods. The idea of such methods consists in choosing a (finite) sequence $\mathcal{U}_1 \subset \mathcal{U}_2 \subset ... \subset \mathcal{U}_M \subset \mathcal{U}$ of finite dimensional spaces, and to compute the minimizers subsequently, i.e. we start with $u_1$ being the global minimizer of

$$J(u) \to \min_{u\in\mathcal{U}_1}.$$

Then we start the optimization of

$$J(u) \to \min_{u\in\mathcal{U}_2}$$

at the starting point $u_1 \in \mathcal{U}_1 \subset \mathcal{U}_2$, and so on. Since we use nested spaces, $\mathcal{U}_{k-1} \subset \mathcal{U}_k$, it is clear that

$$J(u_k) \leq J(u_{k-1})$$

and since we expect "convergence" to the limit problem, we may expect that $u_{k-1}$ is close to $u_k$ for $k$ large. If we use an iterative method like gradient descent, Newton's method or Quasi-Newton methods, then a multi-level strategy will typically result in a large number of iterations on the "coarse levels" $\mathcal{U}_k$ ($k$ small), where all computations are cheap anyway, while one needs only few iterations at "fine levels" $\mathcal{U}_k$ ($k$ close to $M$). Usually, considerable computational effort compared to a direct discretization in $\mathcal{U}_M$ can be saved when using a multi-level strategy.

## 4.3 Methods for Constrained Problems

Over the last decades, a variety of methods has been proposed for constrained problems, most of them being restricted to special problem classes and special types of constraints. For the general constrained problem

$$\begin{aligned} J(u) &\to \min_{u\in\mathcal{U}} \\ E(u) &= 0 \\ I(u) &\preceq 0, \end{aligned} \tag{4.3}$$

there are two strategies of constructing iterative methods, namely sequential linear and sequential quadratic programming (called SLP and SQP, respectively).

### 4.3.1 SLP and SQP

The idea of SLP is a rather simple one, it consists of computing a linear approximation to the constrained problem (4.3), i.e. given the iterate $u_k$ one solves

$$\begin{aligned} J(u_k) + J'(u_k)v &\rightarrow \min_{v \in \mathcal{U}} \\ E(u_k) + E(u_k)v &= 0 \\ I(u_k) + I'(u_k)v &\preceq 0. \end{aligned}$$

After discretization, this leads to a standard linear programming problem for $v$, which can be solved using techniques for large-scale linear programming problems (e.g. interior point methods). The new iterate $u_{k+1}$ is then computed as

$$u_{k+1} = u_k + \tau_k v,$$

with appropriate step size $\tau_k$ (ideally $\tau_k = 1$).

One observes that the computation of the update $v$ by this strategy is only possible if there exists a solution of the linear problem, which is equivalent to the boundedness of the constraint set in direction $v = -J'(u_k)$, which one cannot expect in general. To overcome this difficulty, one can introduce bound constraints for $v$ of the form

$$-v^{max} \preceq v \preceq v^{max}$$

(at least for the discretized problem). This constraint avoids too large updates $v$, which is of course also needed for the linear approximation to make sense.

The idea of SQP consists in a quadratic appoximation of the Lagrange functional in each step of the iteration, i.e. one minimizes

$$\begin{aligned} \frac{1}{2} < A_k v, v > + J'(u_k)v + J(u_k) &\rightarrow \min_{v \in \mathcal{U}} \\ E(u_k) + E'(u_k)v &= 0 \\ I(u_k) + I'(u_k)v &\preceq 0. \end{aligned}$$

The new iterate is computed as

$$u_{k+1} = u_k + v.$$

The optimal choice for $A_k$ is given by the second derivative of the Lagrange functional with respect to $u$:

$$A_k = \frac{\partial^2 \mathcal{L}}{\partial u^2}(u_k; p_k, q_k)(v, v),$$

where $p_k$ and $q_k$ are approximations of the Lagrangian variables corresponding to $u_k$. These approximations are usually obtained by computing the Lagrangian variables $\hat{p}, \hat{q}$ of the quadratic problem corresponding to $v$ and iterating

$$\begin{aligned} p_{k+1} &= p_k + \hat{p} \\ q_{k+1} &= q_k + \hat{q}. \end{aligned}$$

In the case of pure equality constraints, this approach is equivalent to the application of Newton's method (for equations) to the KKT-system, as we shall see in the next section.

### 4.3.2 SQP-Methods for Equality-Constrained Problems

In this section we discuss the application of SQP methods to problems with pure equality constraints, i.e.

$$\begin{aligned} J(u) &\rightarrow \min_{u \in \mathcal{U}} \\ E(u) &= 0. \end{aligned}$$

The KKT-system corresponding to this problem is given by

$$\begin{aligned} J'(u) + E'(u)^* p &= 0 \\ E(u) &= 0 \end{aligned}$$

We now consider the KKT-system as a system of nonlinear equations for $u$ and $p$ of the form

$$F(u, p) = 0$$

and apply Newton's method, i.e. we compute new iterates $(u_{k+1}, p_{k+1})$ from $(u_k, p_k)$ by solving

$$F'(u_k, p_k)(u_{k+1} - u_k, p_{k+1} - p_k) + F(u_k, p_k) = 0.$$

For our specific form of $F$ this amounts to solving (in weak form)

$$\begin{aligned} J''(u_k)(u_{k+1} - u_k, v) + < E''(u_k)(u_{k+1} - u_k, v), p_k > & \\ + < E'(u_k)v, p_{k+1} - p_k > &= -J'(u_k)v - < E'(u_k)v, p_k > \\ < E'(u_k)(u_{k+1} - u_k), r > &= - < E(u_k), r > \end{aligned}$$

for all $v \in \mathcal{U}, r \in \mathcal{V}^*$. In terms of the Lagrangian $\mathcal{L}$, this iteration procedure can be written as

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial u^2}(u_k; p_k)(u_{k+1} - u_k, v) + \frac{\partial^2 \mathcal{L}}{\partial u \partial p}(u_k, p_k)(v, p_{k+1} - p_k) &= -\frac{\partial \mathcal{L}}{\partial u}(u_k, p_k)v \\ \frac{\partial^2 \mathcal{L}}{\partial u \partial p}(u_k; p_k)(u_{k+1} - u_k, r) &= -\frac{\partial \mathcal{L}}{\partial p}(u_k, p_k)r. \end{aligned}$$

This system is also the KKT-system corresponding to the quadratic programming problem (for $(u_{k+1}, p_{k+1})$)

$$\frac{1}{2}\mathcal{L}''(u_k; p_k)(u_{k+1} - u_k, u_{k+1} - u_k) + \mathcal{L}'(u_k; p_k)(u_{k+1} - u_k) + \mathcal{L}(u_k, p_k) \rightarrow \min_{u_{k+1} \in \mathcal{U}}$$

subject to $E'(u_k)(u_{k+1} - u_k) = -E(u_k)$. Here we have used the notation $\mathcal{L}'' = \frac{\partial^2 \mathcal{L}}{\partial u^2}$. Hence, the Newton method for the KKT-system is equivalent to an SQP method with $A_k = \mathcal{L}''(u_k, p_k)$. In particular, the locally quadratic convergence of Newton's method for equations implies the locally quadratic convergence of the SQP-method. Similar to Newton's method for unconstrained problems, the SQP approach can encounter difficulties away from the solution, where the positive definiteness of $\mathcal{L}''$ on the nullspace of $E$ is not guaranteed. In such a case, SQP is not necessarily a descent method. There are several possibilities to overcome this difficulty, several modifications can be used in practice, e.g.

(i) Using a BFGS- or L-BFGS-approximation for the Hessian $\mathcal{L}''(u_k, p_k)$

(ii) Using a trust-region SQP strategy, i.e. introducing an additional trust-region bound $\|u_k\| \leq R_k$, which acts like a stabilization of the Hessian.

### 4.3.3 Solving Quadratic Subproblems

Quadratic problems play an important role in nonlinear optimization. In general, we denote by a quadratic problem the minimization of a functional of the form

$$J(u) = \frac{1}{2} < Au, u > + < b, u > + f \rightarrow \min_{u \in \mathcal{U}}$$

subject to linear constraints

$$\begin{aligned} C_E u + d_E &= 0 \\ C_I u + d_I &\preceq 0. \end{aligned}$$

Here $A : \mathcal{U} \rightarrow \mathcal{U}^*, C_E : \mathcal{U} \rightarrow \mathcal{V}$, and $C_I : \mathcal{U} \rightarrow \mathcal{W}$ are continuous linear operators, and $b \in \mathcal{U}^*, f \in \mathbb{R}, d_E \in \mathcal{V}, d_I \in \mathcal{W}$. In the finite-dimensional case, this is the most general class of problems that can be solved exactly. For the sake of simplicity, we assume that $\mathcal{U}, \mathcal{V}$, and $\mathcal{W}$ are Hilbert spaces. We start with the case of pure equality constraints, where the KKT-system of the quadratic problem becomes an indefinite linear system of the form

$$\begin{pmatrix} A & C_E^* \\ C_E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} -b \\ -d_E \end{pmatrix}.$$

In general, we may assume that $A$ is symmetric. Existence for this problem is guaranteed, if the restriction of $A$ on $\mathcal{N}(C_E)$ has a continuous inverse, $\mathcal{R}(B^*)$ is closed, and $d_E \in \mathcal{R}(B)$. If these conditions are satisfied, then $\mathcal{R}(B^*) = \mathcal{N}(C_E)^\perp$, and we can decompose

$$\begin{aligned} b &= b_1 + b_2, \quad b_1 \in \mathcal{N}(C_E), b_2 \in \mathcal{N}(C_E)^\perp \\ u &= u_1 + u_2, \quad Au_1 \in \mathcal{N}(C_E), Au_2 \in \mathcal{N}(C_E)^\perp \end{aligned}$$

and thus, the first equation becomes

$$Au_1 = b_1, \qquad Au_2 + C_E^* p = b_2,$$

from which we can determine $u_1$ uniquely, as well as $p$ for given $u_2$. The value of $u_2$ can be determined uniquely from the second equation in the KKT-system.

If $A$ is not positive semidefinite at least on $\mathcal{N}(C_E)$, a stationary point (even if unique) need not be a local minimizer. In order to guarantee a unique minimizer (which is the only stationary point), one has to assume that there exist constants $\alpha > 0, \beta > 0$ such that

$$< Av, v > \geq \alpha \|v\|^2 \quad \forall v \in \mathcal{N}(C_E), \tag{4.4}$$

$$\inf_{p \neq 0} \sup_{u \neq 0} \frac{< C_E u, p >}{\|u\|.\|p\|} \geq \beta \quad \forall u \in \mathcal{U}, p \in \mathcal{V}. \tag{4.5}$$

We then have the following result

**Theorem 4.2.** *Let $A$ and $C_E$ be bounded linear operators and let (4.4), (4.5) be satisfied. Then the quadratic problem*

$$\frac{1}{2} < Au, u > + < b, u > + f \rightarrow \min_u$$

*subject to*

$$C_E u = d_E$$

*has a unique stationary point, which is a global minimizer.*

*Proof.* Existence of a stationary point and a minimizer is guaranteed since under the above conditions $A$ is regular on $\mathcal{N}(C_E)$. In order to show uniqueness of a stationary point, it suffices to show uniqueness of the homogeneous problem

$$\begin{aligned} Au + C_E^* p &= 0 \\ C_E u &= 0 \end{aligned}$$

In this case we obtain from $u \in \mathcal{N}(C_E)$ that

$$\begin{aligned} 0 &= < Au, u > + < C_E^* p, u > \\ &= < Au, u > + < p, C_E u > \\ &\geq \alpha \|u\|^2 \end{aligned}$$

and hence, $u = 0$. □

Besides existence and uniqueness fot the infinite dimensional problem, the above conditions also guarantee quasi-optimal approximation results for Galerkin-type approximations of the form

$$\begin{aligned} \frac{1}{2} < Au_h, u_h > + < b, u_h > + f &\rightarrow \min_{u_h \in \mathcal{U}_h} \\ < C_E u_h, v_h > &= < d_E, v_h > \quad \forall v_h \in \mathcal{V}_h \end{aligned}$$

where $\mathcal{U}_h \subset \mathcal{U}, \mathcal{V}_h \subset \mathcal{V}$ are finite-dimensional subspaces.

**Theorem 4.3.** *Let $u_h$ be a Galerkin-approximation of a quadratic optimization problem as defined above. Moreover, assume that (4.4) and (4.5) hold for the original as well as for the discretized problem with constants independent of $h$. Then the estimate*

$$\|u - u_h\| + \|p - p_h\| \leq c\Big( \inf_{v_h \in \mathcal{U}_h} \|u - v_h\| + \inf_{p_h \in \mathcal{V}_h} \|p - p_h\| \Big)$$

*holds with some constant $c$ depending on $\alpha$ and $\beta$ only, where $(u, p)$ is the unique solution of the KKT-system of the original problem, and $(u_h, p_h)$ its Galerkin-approximation.*

We finally discuss the numerical solution of indefinite systems of the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix},$$

which is equivalent to the minimization of

$$\frac{1}{2} < Au, u > + < f, u > \rightarrow \min_u$$

subject to

$$Bu = g.$$

We assume that the system is discretized such that $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is symmetric, $B : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and that (4.4) and (4.5) are satisfied.

If $A$ is a regular matrix, then one of the simplest iterative solution strategies for the indefinite system is given by the inexact Uzawa method:

$$\begin{aligned}\hat{A}(u_{k+1}-u_k) &= -B^T p_k - Au_k + f \\ \hat{C}(p_{k+1}-p_k) &= -Bu_{k+1} + g\end{aligned}$$

where $\hat{A}$ is a preconditioner for $A$ and $\hat{C}$ is a preconditioner for the Schur-complement

$$C = -BA^{-1}B^T.$$

In the simple case of $A = \hat{A}, C = \hat{C}$, the iteration reduces to

$$\begin{aligned}Au_{k+1} &= -B^T p_k + f \\ Cp_{k+1} &= -(Bu_{k+1} - Cp_k) + g.\end{aligned}$$

Eliminating $u_{k+1}$ one obtains

$$\begin{aligned}Cp_{k+1} &= -(Cp_k + BA^{-1}f - Cp_k) + g \\ &= g - BA^{-1}f,\end{aligned}$$

and one easily checks, that $(u_{k+1}, p_{k+1})$ is the exact solution of the quadratic problem, i.e. the Uzawa method solves the indefinite problem in one step.

If $A$ is not regular itself, but only conditions (4.4) and (4.5) are satisfied, one can make the first block of the indefinite problem regular by considering an equivalent quadratic problem of the form

$$\frac{1}{2} < Au, u > + < f, u > + \frac{\rho}{2}\|Bu - g\|^2 \rightarrow \min_u$$

subject to

$$Bu = g,$$

for some positive parameter $\rho$. Note that the additional term we add to the objective is zero since we enforce the constraint $Bu = g$. The KKT-system for this problem is given by

$$\begin{pmatrix} (A + \rho B^T B) & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f + B^T g \\ g \end{pmatrix}.$$

The minimal eigenvalue of $(A + \rho B^T B)$ is given by

$$\lambda_{min} = \inf_{u \neq 0} \frac{\langle (A + \rho B^T B)u, u \rangle}{\|u\|^2}$$

Let $u = u_1 + u_2$ with $u_1 \in \mathcal{N}(C_E)^\perp, u_2 \in \mathcal{N}(C_E)$, then

$$\begin{aligned}< (A + \rho B^T B)u, u > &= < Au_2, u_2 > + 2 < Au_1, u_2 > + < Au_1, u_1 > + \rho < Bu_1, Bu_1 > \\ &\geq (1 - \epsilon) < Au_2, u_2 > + (1 - \frac{1}{\epsilon}) < Au_1, u_1 > + \rho < Bu_1, Bu_1 >\end{aligned}$$

for any $\epsilon \in [0, 1]$. Due to

$$\beta \leq \inf_p \sup_u \frac{< Bu, p >}{\|u\| \ \|p\|} \leq \sup_u \frac{< Bu, Bu_1 >}{\|u\| \ \|Bu_1\|} = \frac{\|Bu_1\|}{\|u_1\|}$$

and

$$\alpha\|u_2\|^2 \leq < Au_2, u_2 >$$

we obtain the estimate

$$< A + \rho B^T B u, u > \geq (1-\epsilon)\alpha\|u_2\|^2 + \rho\beta^2\|u_1\|^2 + (1 - \frac{1}{\epsilon})\|A\| \ \|u_1\|^2.$$

Hence, for $\epsilon < 1$ and $\rho > \frac{(\epsilon-1)\|A\|}{\epsilon\beta^2}$, we obtain

$$\big\langle (A + \rho B^T B)u, u \big\rangle \geq \min\Big\{(1-\epsilon)\alpha, \rho\beta^2 + (1 - \frac{1}{\epsilon})\|A\|\Big\}(\|u_1\|^2 + \|u_2\|^2)$$

and consequently, the minimal eigenvalue of $A + \rho B^T B$ is positive.

Besides this possibility of making the first block regular, we can also attempt to apply an inexact Uzawa iteration, which is well-defined as long as $\hat{A}$ is regular. Typically, the Uzawa iteration performs well if $\hat{A} - A$ is positive definite.

Another alternative for $A$ being regular is to solve the Schur-complement equation

$$-BA^{-1}B^T p = g - A^{-1}B^T f$$

directly, e.g. by a conjugate gradient method (note that $BA^{-1}B^T$ is positive definite).

An approach that receives growing attention is to solve the indefinite system directly by a Krylov-subspace method like GMRES and to use techniques such as inexact Uzawa for preconditioning.

### 4.3.4 Active-set QP Methods

If we want to apply SQP methods in presence of inequality constraints, we also have to solve quadratic problems with linear inequality constraints. A method to realize this is provided by the active-set QP strategy. Consider a problem of the form

$$\frac{1}{2} < A^T u, u > + < b, u > \rightarrow \min_u$$

subject to

$$\begin{aligned} C_E u &= d_E \\ C_I u &\leq d_I, \end{aligned}$$

on $\mathcal{U} = \mathbb{R}^n$, where $C_E : \mathbb{R}^n \rightarrow \mathbb{R}^k, C_I : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are appropriate matrices. We define the active set at a point $u \in \mathbb{R}^n$ as

$$\mathcal{A}(u) = \Big\{ i \in \{1, ..., m\} \Big| e_i^T C_I u = 0 \Big\},$$

where $e_i$ denotes the i-th unit vector. I.e. the active set consists of those indices for which the corresponding inequality constraint is satisfied with equality (" the constraint is active").

If we start at some point $u_k$ with active set $\mathcal{A}(u_k)$, we can compute a step $s$ by solving the equality-constrained problem

$$\begin{aligned} \frac{1}{2} < As, s > + < b, s > &\rightarrow \min_u \\ C_E s &= d_E \\ e_i^T C_I s &= e_i^T d_I \quad \forall i \in \mathcal{A}(u_k) \end{aligned}$$

This means, we only use those inequality constraints corresponding to active indices and convert them to equality constraints.

Now we can check the Lagrangian variables $q$ and the new solution $s$ and distinguish several cases:

(i) $s = u_k$ and $q \geq 0$. In this case $u_k$ is already a stationary point and we can stop.

(ii) $s = u_k$, but there exists an active index $i \in \mathcal{A}(u_k)$ such that $e_i^T q < 0$. In this case we set $u_{k+1} := u_k$ and remove $i$ from the active set.

(iii) $s \neq u_k$ and $s$ is feasible, i.e. $C_I s \leq d_I$. Then we set $u_{k+1} = s$ and $\mathcal{A}(u_{k+1}) = \mathcal{A}(u_k)$.

(iv) $s \neq u_k$ and $s$ is not feasible. In this case we set $u_{k+1} = u_k + \alpha(s - u_k)$, where $\alpha \in (0, 1)$ is the maximal value such that $u_k + \alpha(s - u_k)$ is feasible. We can compute this value $\alpha$ as

$$\alpha = \min \left\{ \frac{e_i^T (d_I - C_i u_k)}{e_i^T C_I (s - u_k)} \Big| i \notin \mathcal{A}(u_k), e_i^T C_I (s - u_k) > 0 \right\}.$$

Since these four cases include all possibilities with an update of this kind, we can use this strategy to generate a sequence $(u_k)$, which converges to a solution of the quadratic programming problem.

## 4.4 Penalty and Barrier Methods

Penalty and barrier methods are governed by a completely different philosophy than the methods discussed above. The idea of penalty methods is to penalize the deviation of a constraint by an additional term added to the objective functional instead of enforcing it exactly. An convenient form of penalization consists in minimizing

$$J(u) + \sum_{j=1}^{k} \frac{1}{\epsilon_j} |e_j^T E(u)|^2 + \sum_{j=1}^{m} \frac{1}{\delta_j} |\max\{0, e_j^T I(u)\}|^2$$

with small positive parameters $(\epsilon_j), (\delta_j)$ terching to zero. Such penalties have the advantage of being differentiable, but due to the quadratic parts they allow for large derivations in the constraints. Note that the first order optimality conditions for the penalized problem are given by

$$\begin{aligned} J'(u) + \sum_{j=1}^{k} \lambda_j e_j^T E'(u) + \sum_{j=1}^{m} \mu_j e_j^T I'(u) &= 0 \\ e_j^T E(u) &= \epsilon_j \lambda_j \\ e_j^T I(u) &\leq \delta_j \mu_j \\ \mu_j &\geq 0 \\ \mu_j (e_j^T I(u) - \delta_j \mu_j) &= 0, \end{aligned}$$

i.e. we can interpret the penalization also as a perturbation of the optimality conditions by the terms $\epsilon_j \lambda_j$ and $\delta_j \mu_j$.

The "optimal" type of penalization are exact penalty functions of the form

$$J(u) + \sum_{j=1}^{k} \frac{1}{\epsilon_j} |e_j^T E(u)| + \sum_{j=1}^{m} \frac{1}{\delta_j} |\max\{0, e_j^T I(u)\}|,$$

for which one can prove that the minimizer coincides with the solution of the original constrained problem if $\epsilon_j$ and $\delta_j$ are sufficiently small. The disadvantage of this choice is that the resulting objective functional is non-smooth and difficult to minimize.

While penalty methods allow for a violation of the constraint, barrier methods strictly avoid that the constraint is active. The idea of barrier methods for an inequality constrained problem of the form

$$\begin{aligned} J(u) &\to \min_u \\ I(u) &\leq 0 \end{aligned}$$

is to add a barrier term to the objective, i.e. to minimize

$$J(u) + \epsilon B(u),$$

where $B$ is chosen such that $B(u) = +\infty$, if the inequality is violated. Since we need a smooth term $B$ for the subsequent optimization of the functional $J + \epsilon B$, this implies that $B(u) \to \infty$ as $u$ tends to the boundary of the feasible set. Hence, a minimizer of $J + \epsilon B$ for positive $\epsilon$ must lie in the interior of the feasible set. Therefore, barrier methods are often called "interior-point methods", in particular in large-scale linear programming, where they outperform the classical simplex method in most cases.

A commonly used form of barrier functions are logarithmic barriers of the form

$$B(u) = -\sum_{j=1}^{m} \log\Big(-e_i^T I(u)\Big).$$

The first-order optimality condition for the functional $J + \epsilon B(u)$ then becomes

$$\begin{aligned} 0 &= J'(u) - \epsilon \sum_{j=1}^{m} \frac{e_j^T I'(u)}{e_j^T I(u)} \\ &= J'(u) + \sum_{j=1}^{m} p_j (e_j^T I'(u)) \end{aligned}$$

with

$$p_j := -\frac{\epsilon}{e_j^T I(u)}$$

For this choice of a "dual" variable $p$ we obtain

$$\begin{aligned} J'(u) + \sum_{j=1}^{m} p_j (e_j^T I'(u)) &= 0 \\ I(u) &\leq 0 \\ p &\geq 0 \\ p_j e_j^T I(u) &= -\epsilon. \end{aligned}$$

Thus, we can also interpret interior point methods as perturbations of the optimality condition by $-\epsilon$ with a perturbation of the complementarity condition in this case.

Of course, in order to apply barrier methods, one has to make sure, that the feasible set has nonempty interior, which is a nontrivial task in general. Moreover, to guarantee convergence of an interior-point strategy, one also needs so-called constraint qualification such as e.g. the Mangasarian-Fromowitz condition. This condition states that for every element in the feasible set there exists a vector pointing inside the interior of the feasible set. If this condition is violated, a sequence of interior points can never converge to such an element (which we need in an barrier method for $\epsilon \to 0$). Interior-point strategies are used typically for the solution of linear and quadratic problems. Therefore, the combination with SLP and SQP methods is of particular interest, such approaches are called interior-point SLP or interior-point SQP methods.

### 4.4.1 The Method of Moving Asymptotes

We shall finally present a method for purely inequality constrained problems, which has become increasingly popular in particular in topology optimization. The main idea of the method of moving asymptotes (MMA, originally introduced by Svanberg 1987) is to use convex approximation of the objective functional by logarithmic terms (therefore this approach is sometimes called SCP-sequentially convex programming). For the solution of the arising convex subproblems, duality can be used effectively.

We consider in this section an optimization problem of the form

$$\begin{aligned} J(u) &\to \min_{u\in\mathbb{R}^n} \\ I(u) &\leq 0 \\ u_{\min} \leq u &\leq u_{\max} \end{aligned}$$

where $J : \mathbb{R}^n \to \mathbb{R}, I : \mathbb{R}^n \to \mathbb{R}^m$ are given nonlinear functions and $u_{\min} \in \mathbb{R}^n, u_{\max} \in \mathbb{R}^n$ are given vectors. The method of moving asymptotes is based on the form

$$H(v;u,L,U) = J(u) + \sum_{j=1}^{n} \left( \frac{p_j}{U_j - v_j} + \frac{q_j}{v_j - L_j} \right) - \sum_{j=1}^{n} \left( \frac{p_j}{U_j - u_j} + \frac{q_j}{u_j - L_j} \right),$$

where

$$\begin{aligned} p_j &= (U_j - u_j)^2 \max\left\{ \frac{\partial J}{\partial u_j}(u), 0 \right\} \\ q_j &= (u_j - L_j)^2 \min\left\{ \frac{\partial J}{\partial u_j}(u), 0 \right\}. \end{aligned}$$

The function $v \mapsto H(v;u,L,U)$ has some very nice properties:

(i) $H(u;u,L,U) = J(u)$

(ii) $\frac{\partial H}{\partial v}(u;u,L,U) = J'(u)$

(iii) $H$ is globally convex in $v$ for $L \leq v \leq U$

The statements (i) and (ii) follow immediately from the definition of $H$, which is constructed such that $H$ is the first-order approximations of $J$ at $u$. The statement (iii) follows from

$$\frac{\partial^2 H}{\partial v^2}(v;u,L,U)(w,w) = \sum_{j=1}^{n}\left(\frac{p_j}{(U_j - v_j)^3}w_j^2 + \frac{q_j}{(L_j - v_j)^3}w_j^2\right)$$

and the fact that $\frac{p_j}{U_j - v_j} \geq 0, \frac{q_j}{L_j - v_j} \geq 0$.

In a similar way we can construct approximations to the inequality constraints defined by $I = (I_l)$ and obtain an iterative method with $u^{k+1}$ being the minimizer of the convex problem

$$\tilde{J}(v) := H(v;u_k,L^k,U^k) \to \min_{v\in\mathbb{R}^n}$$

subject to

$$\tilde{I}_l(v) := I_l(u^k) + \sum_{j=1}^{n}\left(\frac{p_{lj}}{U_j^k - v_j} + \frac{q_{lj}}{v_j - L_j^k}\right) - \sum_{j=1}^{n}\left(\frac{p_{lj}}{U_j^k - u_j^k} + \frac{q_{lj}}{u_j^k - L_j^k}\right) \leq 0$$

for $l = 1, ..., m$, and with

$$p_{lj} = (U_j^k - u_j^k)^2 \max\left\{\frac{\partial I_l}{\partial u_j}(u^k), 0\right\},$$
$$q_{lj} = (u_j^k - L_j^k)^2 \min\left\{\frac{\partial I_l}{\partial u_j}(u^k), 0\right\}.$$

Thus, we construct a sequence of convex problems approximating the original one. The choice of the bounds $L^k$ and $U^k$ provides some freedom, but for $k \to \infty$ one should have $L^k \to u^{\min}, U^k \to u^{\max}$.

In order to realize the method of moving asymptotes, it remains to have a strategy for solving the convex subproblem. It turns out that one can efficiently use duality for this task. If we introduce a Lagrangian of the form

$$\mathcal{L}(v,\lambda) = \tilde{J}(v) + \sum_{i=1}^{m}\tilde{I}_l(v)\lambda_l,$$

then $\mathcal{L}$ can be rewritten in the form (dropping the iteration index $k$ in the following)

$$\mathcal{L}(v,\lambda) = -\sum_{l=1}^{m} r_l\lambda_l - r_0 + \sum_{j=1}^{n}\mathcal{L}_j(v_j,\lambda),$$

with $r_0 \in \mathbb{R}$, a vector $(r_l) \in \mathbb{R}^m$, and

$$\mathcal{L}_j(v_j,x) = \frac{p_j + \sum_{l=1}^{m}\lambda_l p_{lj}}{U_j - v_j} + \frac{q_j + \sum_{l=1}^{m}\lambda_l q_{lj}}{v_j - L_j}$$

We can then derive a dual problem of the form

$$w(\lambda) \to \max_{\substack{\lambda\in\mathbb{R}^m \\ \lambda \geq 0}},$$

where

$$\begin{aligned} w(\lambda) &= \min_{L \le x \le U} \mathcal{L}(v, \lambda) \\ &= -r_0 - \sum_{l=1}^{m} r_l \lambda_l + \sum_{j=1}^{n} \Big( \min_{L_j \le v_j \le U_j} \mathcal{L}_j(v_j, \lambda) \Big). \end{aligned}$$

Due to the simple form of $\mathcal{L}_j$, it is easy to compute its minimizer as

$$v_j(\lambda) = \frac{L_j \sqrt{p_j + \sum \lambda_l p_{lj}} + U_j \sqrt{p_j + \sum \lambda_l q_{lj}}}{\sqrt{p_j + \sum \lambda_l p_{lj}} + \sqrt{q_j + \sum \lambda_l q_{lj}}}.$$

Hence , the dual problem is a concave maximization problem with simple nonnegativity constraint, which can be solved efficiently, e.g. by a modified conjugate gradient method. Moreover, for $m << n$, we significantly reduce the computational effort by using duality as above.

# Chapter 5

# Shape Optimization

In this section we shall deal with the solution of optimization problems, where the design variable is a shape or geometry in $\mathbb{R}^d$. Shapes can be considered as sets with regular boundary and therefore we may perform standard set operations like unions or intersections. However, there is no way to make a class of shapes into a linear space in general, but only with severe restrictions. An obvious way of solving a problem in a linear space instead of a problem on a class of shapes is to use parametrization (e.g. as piecewise graphs, by polar coordinates, or locally around a given shape). Since the parametrization is usually represented by a function on a fixed set, one can just minimize over all such functions in an appropriate Hilbert or Banach space. This allows to use standard methods as discussed above, but strongly limits the class of admissible shapes.

## 5.1 Shape Sensitivity Analysis

The main idea of shape sensitivity analysis is to consider "natural deformations" of shapes and inspect the corresponding variations of the objective functional. The general setup in the following is the minimization of

$$J(\Omega) \to \min_{\Omega \in \mathcal{K}},$$

where $\mathcal{K}$ is a suitable class of compact subsets of $\mathbb{R}^d$, with regular boundary.

There are two different ways of deriving shape sensitivities (both leading to the same result), namely via "direct deformations" or via the "speed method". We shall follow the latter, since this approach fits very well to the level set method, which we will discuss below as a possible solution method for shape optimization problems.

Before considering shapes we illustrate the idea of the speed method when applied to Gateaux-derivatives in linear spaces. In order to compute the directional derivative of a functional $J : \mathcal{U} \to \mathbb{R}$, we have so far considered the variation between the values of $J$ at $\overline{u} \in \mathcal{U}$ and at its local deformation $\overline{u} + tv$. Alternatively, we could define $u(t) = \overline{u} + tv$ by

$$\frac{du}{dt} = v, \quad u(0) = \overline{u},$$

which is an initial value problem for an ordinary differential equation in $\mathcal{U}$. Using the chain rule, we can then compute

$$\frac{d}{dt} J(u(t)) = J'(u(t)) \frac{du}{dt} = J'(u(t))v.$$

In particular,

$$\frac{d}{dt}J(u(t))\Big|_{t=0} = J'(\overline{u})v,$$

i.e., we obtain the directional derivative at $\overline{u}$ by evaluating the time derivative of $J(u(t))$ at time $t = 0$.

In a similar way, we can define derivatives of shapes. Let $V : \mathbb{R}^d \to \mathbb{R}^d$ be a given velocity field and define $x(t)$ via

$$\frac{dx}{dt}(t) = V(x(t)), \quad x(0) = \overline{x}, \tag{5.1}$$

for each $\overline{x} \in \mathbb{R}^d$. We can then define the shape sensitivity

$$dJ(\overline{\Omega}; V) := \left(\frac{d}{dt}J(\Omega(t))\right)\Big|_{t=0},$$

where

$$\Omega(t) = \{x(t) \mid x(0) \in \overline{\Omega}\}.$$

Note that the main difference to derivatives in linear spaces is that the deformation defined by the ODE (5.1) is nonlinear, since $V$ depends on $x$ itself.

We start with some examples. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function and define

$$J(\Omega) := \int_\Omega g(x)\ dx.$$

Then, by change of variables

$$\begin{aligned} J(\Omega(t)) &= \int_{\Omega(t)} g(x)\ dx \\ &= \int_\Omega g(x_y(t))|M_y|\ dy \end{aligned}$$

where $x_y(t)$ is defined by

$$\frac{dx_y}{dt}(t) = V(x_y, t), \quad x_y(0) = y \in \Omega$$

and $M_y = \det \frac{\partial x_y}{\partial y}$. Hence, the time derivative can be computed as

$$\frac{d}{dt}J(\Omega(t)) = \int_\Omega \nabla g(x_y)\frac{\partial x_y}{\partial t}|M_y|\ dy + \int_\Omega g(x_y)\frac{\frac{\partial M_y}{\partial t}M_y}{|M_y|}dy.$$

For the derivative of the determinant we have

$$\begin{aligned} \frac{\partial M_y}{\partial t} &= \frac{\partial}{\partial t}\left(\sum_{(i_1,\dots i_d)\in\Pi(d)} (-1)^{i_1+\dots+i_d}\prod_{k=1}^d \frac{\partial (x_y)_k}{\partial y_{i_k}}\right) \\ &= \left(\sum_{(i_k)\in\Pi(d)} (-1)^{\sum i_k}\sum_j \frac{\partial^2 (x_y)_j}{\partial y_{i_j}\partial t}\prod_{l\neq j}^d \frac{\partial (x_y)_l}{\partial y_{i_l}}\right) \\ &= \sum_{(i_k)\in\Pi(d)} (-1)^{\sum i_k}\sum_j \frac{\partial V_j}{\partial y_{i_j}}\prod_{l\neq j}^d \frac{\partial (x_y)_l}{\partial y_{i_l}} \end{aligned}$$

For $t = 0$, we have $\frac{\partial x_y}{\partial y} = I, M_y = 1$, and this implies

$$\frac{\partial M_y}{\partial t} = \sum_j \frac{\partial V_j}{\partial y_j} = div(V)$$

As a consequence, we have

$$\begin{aligned}
\frac{d}{dt} J(\Omega(t))\Big|_{t=0} &= \int_\Omega \left( \nabla g(x_y) \frac{\partial x_y}{\partial t} \right) \Big|_{t=0} dy + \int_\Omega \Big( g(x_y)\ div\ V(x_y) \Big)\Big|_{t=0} dy \\
&= \int_\Omega \Big( \nabla g(y) V(y) + g(y)\ div\ V(y) \Big) dy \\
&= \int_\Omega div\Big( g(y) V(y) \Big) dy \\
&= \int_{\partial\Omega} g(y) V(y).n\ ds,
\end{aligned}$$

where $n$ denotes the unit outer normal on $\partial\Omega$. I.e., the shape sensitivity is a linear functional of $V$ concentrated on $\partial\Omega$. Another key oberservation is that the shape sensitivity

$$J'(\Omega)V := \frac{d}{dt} J(\Omega(t))|_{t=0}$$

depends on $V.n|_{\partial\Omega}$ only, while it is completely independent of the values for $V$ inside $\Omega$ and of its tangential component. Consequently, we may directly consider variations of $\partial\Omega$ with a velocity $V = V_n n$, where $V_n$ is a scalar speed function. The shape sensitivity then becomes

$$J'(\Omega)V_n = \int_{\partial\Omega} g V_n\ ds.$$

The statement that the shape sensitivity is a linear functional of $V.n$ holds for very general classes of objective functionals, it is usually known as the "Hadamard-Zolésio Structure Theorem". The independence of the shape sensitivity on tangential components is clear from geometric intuition, since those components correspond to a change of parametrization only. The independence on values of $V$ in the interior of $\Omega$ seems obvious, too, since they do not change the domain of integration in the objective functional.

In most typical applications of shape optimization, the objective functional depends on a state variable $u$ that satisfies a partial differential equation related to $\Omega$. This relation can arise in several ways, e.g.

(i) $u$ solves a partial differential equation in a domain $\Omega \subset\subset D$, and $\partial\Omega$ is the discontinuity set for some of the parameters. A simple example is the optimal design of two conductive materials, where the conductivity $a$ takes two different values, i.e.,

$$a(x) = \begin{cases} a_1 & x \in \Omega \\ a_2 & x \in D \backslash \Omega. \end{cases}$$

A typical shape optimization problem consists in the optimization of some functional $J(\Omega) = \tilde{J}(u_\Omega)$, where $u_\Omega$ solves

$$-div\ (a \nabla u_\Omega) = 0.$$

(ii) $u$ solves a partial differential equation in $\Omega$ and satisfies a boundary condition on $\partial\Omega$.

(iii) $u$ solves a partial differential equation on the surface of $\partial\Omega$.

The general structure of such problems is

$$J(\Omega) = \tilde{J}(u_\Omega, \Omega) \to \min_\Omega$$

subject to

$$e(u_\Omega, \Omega) = 0,$$

where $e$ denotes the partial differential equation. In this case we have to use the chain rule and an implicit function theorem to compute the shape sensitivity. Let $\Omega(t)$ be as above and let $u(t)$ denote the solution of

$$e(u(t), \Omega(t)) = 0$$

with $\Omega(t)$ given. Then the shape sensitivity of $J$ is given by

$$\begin{aligned} J'(\Omega)V &= \frac{d}{dt} J(\Omega(t))\Big|_{t=0} \\ &= \frac{d}{dt}\Big(\tilde{J}(u(t), \Omega(t))\Big)\Big|_{t=0} \\ &= \frac{\partial \tilde{J}}{\partial u}(u(0), \Omega(0))u'(0) + \frac{\partial \tilde{J}}{\partial \Omega}(u(0), \Omega(0))V. \end{aligned}$$

Here $\frac{\partial \tilde{J}}{\partial u}$ denotes the (Gateaux-)derivative of $\tilde{J}$ with respect to $u$ (for $\Omega$ fixed) and $\frac{\partial \tilde{J}}{\partial \Omega}$ denotes the shape sensitivity of $\tilde{J}$ with respect to $\Omega$ (for $u$ fixed). Due to the chain rule we obtain for $u'(0) = \frac{d}{dt}u(t)|_{t=0}$ the equation

$$0 = \frac{d}{dt} e(u(t), \Omega(t)) = \frac{\partial e}{\partial u}(u(t), \Omega(t))u'(t) + \frac{\partial e}{\partial \Omega}(u(t), \Omega(t))V.$$

Here, $\frac{\partial e}{\partial \Omega}(u, \Omega(t))V = \frac{d}{dt} e(u, \Omega(t))$, for $u$ fixed, i.e., it means a generalization of shape sensitivities from functionals to operators. The function $u' = u'(0)$ is usually called "shape derivative".

We shall discuss the computation of shape derivatives for two examples. First, consider the maximization of current for a conductive material. The objective is given by

$$J(\Omega) = -\int_\Gamma a \frac{\partial u_\Omega}{\partial n}\, ds,$$

where $\Gamma \subset D, \Omega \subset\subset D$ and $u$ solves

$$-div\ (a\nabla u) = f, \quad \text{in } D$$

with homogeneous boundary values $u = 0$ on $\partial D$. Here, $f$ is a given function and $a$ is defined as above, i.e.

$$a(x) = \begin{cases} a_1 & x \in \Omega \\ a_2 & x \in D\backslash\Omega. \end{cases}$$

The shape sensitivity is then given by (note that $\Omega \subset\subset D$ and thus $a = a_2$ on $\Gamma \subset \partial D$)

$$J'(\Omega)V = -\int_\Gamma a_2 \frac{\partial u'}{\partial n} ds,$$

where $u'$ is the shape derivative corresponding to the above state equation. In order to compute the shape derivative $u'$, we consider the state equation in its weak form, i.e. we seek $u \in H_0^1(D)$ satisfying

$$\int_D a\nabla u\nabla v \; dx = \int_D fv \; dx \quad \forall v \in H_0^1(D)$$

We can write the left-hand side as

$$< v, e(u,\Omega) >= \int_D a_2\nabla u\nabla v \; dx + \int_\Omega (a_1 - a_2)\nabla u\nabla v \; dx.$$

The derivative with respect to $u$ is given by

$$\frac{\partial e}{\partial u}(u,\Omega)u' = \int_D a_2\nabla u'\nabla v \; dx + \int_\Omega (a_1 - a_2)\nabla u'\nabla v \; dx = \int_D a\nabla u'\nabla v \; dx.$$

In order to compute the derivative with respect to $\Omega$, we can use the above results on shape sensitivities for the functional $\int_\Omega g \; dx$, now with $g = (a_1 - a_2)\nabla u.\nabla v$. Thus,

$$\frac{\partial e}{\partial \Omega}(u,\Omega)V = \int_{\partial\Omega} \Big((a_1 - a_2)\nabla u.\nabla v\Big)V.n \; ds \quad \forall v \in H_0^1(D).$$

As for standard optimal design problems, we can also employ the adjoint method to compute the shape sensitivity. For this sake, let $u^* \in H_0^1(D)$ be the unique weak solution of

$$\int_\Gamma a_2 \frac{\partial w}{\partial n} dx = \int_D a\nabla w\nabla u^* \; dx \quad \forall w \in H_0^1(D).$$

Then we obtain

$$-\int_\Gamma a_2 \frac{\partial u'}{\partial n} ds = -\int_D a\nabla u'\nabla u^* \; dx = \int_{\partial\Omega} \Big((a_1 - a_2)\nabla u.\nabla u^*\Big)V.n \; ds,$$

i.e., the shape sensitivity is again a functional of $V.n$ concentrated on $\partial\Omega$.

Our second example is the shape derivative for a state equation with Dirichlet boundary condition, i.e.

$$\begin{aligned} \Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

It is easy to show that

$$\Delta u' = 0 \quad \text{in } \Omega.$$

For the boundary condition, let $y \in \partial\Omega$ and let $\frac{dx}{dt}(t) = V(x(t)), x(0) = y$. Then $u(x(t)) = 0$ for all $t$ and thus

$$\frac{d}{dt}u(x(t)) = u'(x(t)) + \nabla u(x(t)).V(x(t)) = 0.$$

Hence, $u'$ satisfies

$$u' = -\nabla u.V \quad \text{on } \partial\Omega.$$

We finally notice that second derivatives, so-called shape Hessians can be computed by applying the same technique as for shape sensitivities to $J'(\Omega)V$, now with variations due to a second velocity $W$.

## 5.2 Level Set Methods

Level set methods recently received growing attention in shape optimization due to their capabilities of solving shape optimization problems without parametrizations. The main idea of the level set method is to represent a shape as

$$\Omega(t) = \{\phi(.,t) < 0\},$$

where $\phi : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$ is a suitable continuous function, ideally the signed distance function to $\partial\Omega$ (i.e., equal to the distance between $x$ and $\partial\Omega$ if $x \in \mathbb{R}^d \backslash \Omega$, and equal to the negative distance if $x \in \Omega$). For an appropriate $\phi$ we have that

$$\partial\Omega(t) = \{\phi(.,t) = 0\}.$$

Now consider the motion of points in $\Omega(t)$ by $\frac{dx}{dt} = V(x)$. Then we obtain from the chain rule for $x(t) \in \partial\Omega(t)$

$$0 = \frac{d}{dt}\phi(x(t),t) = \frac{\partial\phi}{\partial t} + V.\nabla\phi = 0,$$

i.e., $\phi$ can be determined by solving a transport equation. As we have seen above, the most interesting case is the one of a motion in normal direction on $\partial\Omega(t)$, i.e., $V = V_n.n$. In order to use such a velocity in the level set method, we have to express the normal in terms of the level set function $\phi$. Assume that $\{\tilde{x}(s,t)|s \in (-\epsilon,\epsilon)\}$ is an arc on $\partial\Omega(t)$, locally parametrized by $s$ around $x(t) = \tilde{x}(0,t)$. Then

$$0 = \frac{d}{ds}\phi(\tilde{x}(s,t),t) = \nabla\phi(\tilde{x}(s,t),t)\frac{\partial\tilde{x}}{\partial s}.$$

Since $\frac{\partial\tilde{x}}{\partial s}$ can be any tangential direction, we obtain that $\nabla\phi$ is a normal direction, and one obtains the unit normal as

$$n(s,t) = \frac{\nabla\phi}{|\nabla\phi|}(\tilde{x}(s,t),t).$$

Using these formulas together with the transport equation for $\phi$, we obtain the Hamilton-Jacobi equation

$$\frac{\partial\phi}{\partial t} + V_n|\nabla\phi| = 0 \tag{5.2}$$

for $\phi$. One can show that the motion of $\Omega(t)$ is determined by

$$\Omega(t) = \{\phi(.,t) < 0\}$$

if $\phi$ is a solution of (5.2) in $\mathbb{R}^d \times \mathbb{R}^+$ where $V_n$ is an arbitrary extension from $\{\phi(.,0) < 0\}$ to $\mathbb{R}^d$.

For further details and applications of the level set method we refer to the monograph by Osher and Fedkiw.

## 5.3 Computing Shape Sensitivities by Level Set Methods

Using the level set method, we can formally compute shape sensitivities in a simple way. Consider again the functional

$$J(\Omega) = \int_\Omega g(x)\ dx$$

and let $\partial\Omega(t)$ move with normal speed $V_n$. Then we obtain

$$\begin{aligned} J(\Omega(t)) &= \int_{\{\phi(.,t)<0\}} g(x)\ dx \\ &= \int_{\mathbb{R}^d} H(-\phi(x,t))g(x)\ dx, \end{aligned}$$

where $H$ denotes the Heaviside function

$$H(p) = \begin{cases} 1 & \text{if } p > 0 \\ 0 & \text{else.} \end{cases}$$

Since the derivative of the Heaviside function is the Dirac-delta-distribution, we obtain formally

$$\begin{aligned} \frac{d}{dt}J(\Omega(t)) &= \int_{\mathbb{R}^d} -H'(-\phi(x,t))\frac{\partial\phi}{\partial t}(x,t)\ g(x)\ dx \\ &= \int_{\mathbb{R}^d} \delta(\phi(x,t))|\nabla\phi(x,t)|\ V_n\ g(x)\ dx \end{aligned}$$

Now we apply the co-area formula, i.e.

$$\int_{\mathbb{R}^d} A(\phi(x))\ B(x)\ |\nabla\phi(x)|\ dx = \int_{\mathbb{R}} A(p) \int_{\{\phi=p\}} B(x)\ ds(x)\ dp.$$

This implies

$$\begin{aligned} \frac{d}{dt}J(\Omega(t))\Big|_{t=0} &= \int_{\mathbb{R}^d} \delta(\phi(x,0))\ g(x)\ V_n(x)\ |\nabla\phi(x,0)|\ dx \\ &= \int_{\mathbb{R}} \delta(p) \int_{\{\phi=p\}} g(x)\ V_n(x)\ ds\ dp \\ &= \int_{\{\phi=0\}} g(x)\ V_n(x)\ ds(x) \\ &= \int_{\partial\Omega} g\ V_n\ ds, \end{aligned}$$

i.e., we recover the above formula for the shape sensitivity.

In a similar way we can compute the shape sensitivity of the functional

$$J(\Omega) = \int_{\partial\Omega} g\ ds$$

For this sake we use again the $\delta$-distribution and the coarea formula to deduce

$$\begin{aligned} J(\Omega(t)) &= \int_{\{\phi(.,t)=0\}} g(x)\ ds(x) \\ &= \int_{\mathbb{R}} \delta(p) \int_{\{\phi(.,t)=p\}} g(x)\ ds(x)\ dp \\ &= \int_{\mathbb{R}^d} \delta(\phi(x,t))\ g(x)\ |\nabla\phi(x,t)|\ dx \end{aligned}$$

Thus, we can try to compute the time derivative as

$$\begin{aligned}
\frac{d}{dt}J(\Omega(t)) &= \int_{\mathbb{R}^d} g\left(\delta'(\phi)|\nabla\phi|\phi_t + \delta(\phi)\frac{\nabla\phi\nabla\phi_t}{|\nabla\phi|}\right)\,dx \\
&= \int_{\mathbb{R}^d} g\left(\frac{\nabla\delta(\phi)\nabla\phi}{|\nabla\phi|}\phi_t + \delta(\phi)\frac{\nabla\phi\nabla\phi_t}{|\nabla\phi|}\right)\,dx \\
&= \int_{\mathbb{R}^d} \delta(\phi)\left(-div\,\left(g\frac{\nabla\phi}{|\nabla\phi|}\phi_t\right) + g\frac{\nabla\phi\nabla\phi_t}{|\nabla\phi|}\right)\,dx \\
&= -\int_{\mathbb{R}^d} \delta(\phi)\left(\frac{\nabla g.\nabla\phi}{|\nabla\phi|}.\phi_t + g\;div\,\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\phi_t\right)\,dx \\
&= \int_{\mathbb{R}^d} \delta(\phi)|\nabla\phi|\;V_n\,\left(\nabla g\frac{\nabla\phi}{|\nabla\phi|} + g\;div\,\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right)\,dx \\
&= \int_{\{\phi=0\}} V_n\left(\nabla g\frac{\nabla\phi}{|\nabla\phi|} + g\;div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right)\,ds
\end{aligned}$$

One observes that on $\partial\Omega = \{\phi = 0\}$ we have

$$n = \frac{\nabla\phi}{|\nabla\phi|}, \quad \kappa = div\; n = div\left(\frac{\nabla\phi}{|\nabla\phi|}\right),$$

where $n$ is the unit normal and $\kappa$ is the mean curvature. Thus,

$$J'(\Omega)V_n = \int_\Gamma V_n\left(\frac{\partial g}{\partial n} + g\;\kappa\right)\,ds.$$

We finally notice that the above strategy of removing the term $\delta'(\phi)$ by rewriting

$$\delta'(\phi)|\nabla\phi| = \nabla\delta(\phi)\frac{\nabla\phi}{|\nabla\phi|}$$

and applying Gauss' Theorem can be used for general functionals (e.g. for second derivatives of the functional $J$ above). In this way, we always obtain a term of the form

$$-\delta(\phi)\;div\left(\frac{\nabla\phi}{|\nabla\phi|}\right),$$

i.e., the mean curvature on $\{\phi = 0\} = \partial\Omega$. In particular, we can rewrite all derivatives as surface integrals on $\partial\Omega$, involving only natural geometric quantities like the normal $n$ or the curvature $\kappa$, its normal derivative $\frac{\partial\kappa}{\partial n}$, etc. It is a good advice to check all quantities that one obtains by computing shape sensitivities in this way with respect to their geometric meaning. If some terms do not have a geometric interpretation, then most likely the calculation was wrong.

## 5.4 Numerical Solution

In order to obtain computational methods for shape optimization problems we can again employ the level set method. In principle, we can apply any of the optimization methods discussed in chapter 4, once we know how to compute derivatives. The major difference is the

way we update the design variable. In the setting of chapter 4, we have computed a search direction $S$ to obtain

$$u_{k+1} = u_j + \tau_k S.$$

Obviously, we cannot use the same strategy in shape optimization, since a formula like

$$\Omega_{k+1} = \Omega_k + \tau_k S$$

does not make sense for shapes $\Omega_k$. However, there is a natural update offered by the speed method. First we notice that the update for a design variable $u$ in a Hilbert space can be rewritten as

$$u_{k+1} = u(\tau_k), \quad \frac{du}{dt} = S, \quad u(0) = u_k.$$

As in the context of shape derivatives, the corresponding speed method for shapes gives

$$\Omega_{k+1} = \left\{ x(t_k) \;\middle|\; \frac{dx}{dt} = S,\ x(0) \in \Omega_k \right\}.$$

Since the motion depends only on the normal velocity on $\partial\Omega$, we can define the update also via the level set method as

$$\begin{aligned} \Omega_{k+1} &= \{\phi(.,\tau_k) < 0\} \\ \frac{\partial \phi}{\partial t} + S_n|\nabla\phi| &= 0 \quad \text{in } (0,\tau_k) \\ \{\phi(.,0\} &= \Omega_k, \end{aligned}$$

where $S_n$ is the normal component of the update $S$. Hence, the iterative method is characterized by choosing a normal update. Below, we shall detail some possible ways for choosing this update.

We start with a gradient-type method. One observes that for optimization in Hilbert spaces, the gradient method is characterized by choosing the update $S$ via

$$< S, v >= -J'(u)v \quad \forall v \in \mathcal{U}.$$

We can now write an analogous formula for the update $S_n$, namely

$$< S_n, V_n >= -J'(\Omega)V_n \quad \forall V_n \in \mathcal{U},$$

where $\mathcal{U}$ is a suitable Hilbert space for which we have several possibilities. We start with the simple choice $\mathcal{U} = L^2(\partial\Omega)$, i.e.,

$$< S_n, V_n >= \int_{\partial\Omega} S_n V_n \; ds.$$

As we have seen above, one can usually write the shape sensitivity in the form

$$J'(\Omega)V_n = \int_{\partial\Omega} hV_n \; ds$$

(with $h = g$ for $J(\Omega) = \int_\Omega g \; dx$, and $h = \frac{\partial g}{\partial n} + g\kappa$ for $J(\Omega) = \int_{\partial\Omega} g \; ds$). Thus, the equation for $S_n$ becomes

$$\begin{aligned} \int_{\partial\Omega} S_n V_n \; ds =< S_n, V_n > &= -J'(\Omega)V_n \\ &= -\int_{\partial\Omega} h \; V_n \; ds \quad \forall V_n \in L^2(\partial\Omega) \end{aligned}$$

which is equivalent to choosing $S_n = -h$.

Another interesting Hilbert space is $H^1(\partial\Omega)$. The scalar product in this space is given by

$$\begin{aligned} < S_n, V_n > &= \int_{\partial\Omega} (\nabla_s S_n \nabla_s \ V_n + S_n V_n) \ ds \\ &= \int_{\partial\Omega} V_n(-\Delta_s S_n + S_n) \ ds, \end{aligned}$$

where $\Delta_s$ denotes the gradient with respect to the surface variable $S$ on $\partial\Omega$ and $\Delta_s$ is the surface Laplacian. Consequently, the update $S_n$ can be computed by solving the Laplace-Beltrami equation

$$-\Delta_s S_n - S_n = h$$

on $\partial\Omega$ (note that we do not need a boundary condition, since the boundary of the surface $\partial\Omega$ is empty).

In general, we can write a Hilbert space scalar product as

$$< S_n, V_n >= \int_{\partial\Omega} (AS_n) V_n \ ds,$$

where $A$ is a positive definite operator. Thus, we may choose any search direction of the form

$$S_n = -A^{-1}h,$$

where $A$ is a positive definite operator. Since

$$J'(\Omega)S_n = - < S_n, S_n >= -\|S_n\|^2,$$

this yields a descent direction and we can use line search techniques to find a reasonable $\tau_k$.

In a similar way to gradient methods we can derive Newton-type methods, for which $S_n$ is chosen solving

$$J''(\Omega)(S_n, V_n) = -J'(\Omega)V_n, \quad \forall V_n \in \mathcal{U}.$$

# Chapter 6

# Topology Optimization

Topology optimization is a generalization of shape optimization, where one does not only want to design the shape of objects, but the total topology, in particular by introducing holes. Using the level set method it is sometimes possible to change the topology by splitting or merging connected components, but the kind of topological changes that can occur is very limited. Therefore we shall disuss some alternative approaches in this section.

## 6.1 Topological Derivatives

The approach that is most closely related to shape optimization uses topological derivatives as a criterion to introduce holes in addition to shape derivatives for moving shapes. The topological derivative measures the first-order variation of the objective when introducing an infinitesimal hole, usually limited to a spherical shape. I.e., the topological derivative of a functional $J$ at topology $\Omega \subset \mathbb{R}^d$ with respect to a variation at $x \in \mathbb{R}^d$ is given by

$$d_T J(\Omega; x) = \lim_{R \downarrow 0} \frac{J(\Omega \backslash B_R(x)) - J(\Omega)}{|B_R(x)|}.$$

One observes that for $d_T J(\Omega; x) < 0$ the nucleation of a small hole centered at $x$ is favorable, since

$$J(\Omega \backslash B_R(x)) < J(\Omega)$$

for $R$ sufficiently small. Thus, one can combine the use of the topological derivative with shape optimization techniques, e.g. by alternating the nucleation of holes and the motion of the arising shapes.

We consider a simple example: Let

$$J(\Omega) = \int_D f(u_\Omega)\ dx,$$

where $f : \mathbb{R} \to \mathbb{R}$ is a smooth given function, and $u_\Omega \in H_0^1(\Omega)$ solves

$$-\Delta u_\Omega = \chi_\Omega \quad \text{in } D \supset\supset \Omega.$$

Then the topological derivative is given by

$$d_T J(\Omega; \overline{x}) = \int_D f'(u_\Omega) u'\ dx,$$

where

$$u' = \lim_{R \downarrow 0} \frac{u_{\Omega \backslash B_R(\overline{x})} - u_\Omega}{|B_R(\overline{x})|}.$$

Since

$$-\Delta(u_{\Omega \backslash B_R(\overline{x})} - u_\Omega) = \chi_{\Omega \backslash B_R(\overline{x})} - \chi_\Omega = -\chi_{B_R(\overline{x})}$$

we obtain

$$-\Delta u' = -\delta(\overline{x}),$$

and hence, $u' = -G(.;\overline{x})$, where $G$ is the Green function of the Laplace operator on $D$.

Using topological derivatives leads to a method with clear geometric interpretation, but it suffers from two major drawbacks in general. First of all, it is difficult to switch between topological and shape derivatives in an automatic way. Secondly, the topological derivative has difficulties to handle surface functionals. Consider e.g., the case of

$$J(\Omega) = \int_\Omega g \; dx + \int_{\partial\Omega} 1 \; ds.$$

Then

$$\begin{aligned} J(\Omega \backslash B_R(\overline{x})) - J(\Omega) &= \int_{B_R(\overline{x})} g \; dx + \int_{\partial B_R(\overline{x})} 1 \; ds \\ &= \int_{B_R(\overline{x})} g \; dx + 2\pi R. \end{aligned}$$

Thus,

$$\frac{J(\Omega \backslash B_R(\overline{x})) - J(\Omega)}{R^2\pi} = \frac{2\pi}{R} + \frac{\int_{B_R(\overline{x})} g \; dx}{R^2\pi} = o\left(\frac{1}{R}\right),$$

and the limit $R \to 0$ always gives $+\infty$, i.e., the topological derivative cannot generate a hole. We shall therefore consider alternative approaches in the following sections.

## 6.2 Phase-Field Methods

In this section we consider functionals of the form

$$\begin{aligned} J(\Omega) &= G(\chi_\Omega) + \alpha \int |\nabla \chi_\Omega| \; dx \\ &= G(\chi_\Omega) + \alpha \int_{\partial\Omega} 1 \; ds, \end{aligned}$$

where $\chi_\Omega$ denotes the indicator function of the set $\Omega$. Then one can try to approximate the minimization with respect to the signed distance function by the minimization of

$$\tilde{J}(u) = G(u) + \alpha \int_{\mathbb{R}^d} \left( \epsilon |\nabla u|^2 + \frac{1}{\epsilon} W(u) \right) \; dx$$

with respect to $u \in H_0^1(\Omega)$, where $\epsilon > 0$ is a small parameter and $W$ is a double-well potential with minima at $u = 0, u = 1$, e.g.

$$W(u) = u^2(1-u)^2.$$

One can show that the functional $\tilde{J}$ converges to the original functional $J$ as $\epsilon \to 0$ (in an appropriate sense).

One can interpret the $\epsilon$-dependent terms in $\tilde{J}$ as penalizations: the term $\frac{1}{\epsilon}W(u)$ favors the values $u = 0$ and $u = 1$ and causes the convergence to indicator functions as $\epsilon \to 0$. The term $\epsilon|\nabla u|^2$ penalizes oscillations in $u$ and causes the boundedness of the perimeter $\int |\nabla u|\, dx$ as $\epsilon \to 0$.

The phase-field method allows to use standard otpimization techniques in the Hilbert space $H_0^1(\Omega)$. Moreover, the parameter $\epsilon$ can be used to obtain a continuation strategy, i.e., one can start the optimization procedure by computing a minimizer of $\tilde{J}$ for large $\epsilon = \epsilon_1$, where the problem is globally convex, use the result as a starting value for the minimization with $\epsilon = \epsilon_2 < \epsilon_1$, and so on. In this way one can compute global minima of $\tilde{J}$, although this functional is non-convex for small $\epsilon$ in general.

## 6.3 Homogenization Methods

An alternative approach to the relaxation of topology optimization problems are homogenization methods. The scope of homogenization is to find effective properties of fine mixtures of materials. Consider for example a conductivity problem

$$-div\ (a\nabla u) = f, \quad \text{in } D,$$

where $a$ is a piecewise constant function taking two different values $a_1$ and $a_2$ (depending on the actual material). Then homogenization leads to an effective equation of the form

$$-div\ (A(\rho)\nabla u) = f, \quad \text{in } D,$$

where $A$ is a nonlinear function with $a_1 \le A(\rho) \le a_2$, and $\rho : D \to [0, 1]$ is a material density. In certain cases, in particular for periodic media, the effective modulus $A$ is known. In other cases, phenomenological material interpolation schemes are used, e.g., the SIMP ( Simple Isotropic Material with Penalization) model

$$A(\rho) = a_1 + (a_2 - a_1)\rho^p$$

with $p > 1$ (often $p = 3$) or the RAMP (Rational Approximation of Material Properties) model

$$A(\rho) = a_1 + \frac{(a_2 - a_1)\rho}{1 + q(1 - \rho)},$$

with $q > 0$, are used. In typical applications, such material interpolation schemes favour optimal densities $\rho$ with values close to 0 or 1, and therefore the solution can often be interpreted as a material distribution.

# Appendix A

# References and Further Reading

In the following we provide some links to literature related to the topics treated in this class or providing a basis for further investigations in this subjects.

### Chapter 1

General introductions to optimization problems can be found e.g. in Pedregal [Pe04], Zeidler [Ze85]

Applications of infinite-dimensional optimization can be found in the following monographs:

- Optimal Control: [Ba84, Be88, Be98, HoLaLeSpTr02, HoLeTr98, NeTi94, Tr96]
- Inverse Problems and Optimal Design: [EnMcL93, BiCoCoSa97a, BiCoCoSa97b, HoHoSch00]
- Calculus of Variations: [Bu89, Da89, JoLi98]
- Structural Optimization: [BeSi03, HaMa03]
- Optimization in Fluid Mechanics: [MoPi01]

### Chapter 2

An extensive source for theoretical aspects of optimization problems in Hilbert and Banach spaces is the book by Zeidler [Ze85]. More recent books on this subject are the ones by Jahn [Ja96] and Pedregal [Pe04].

Discussions of weak lower semicontinuity of functionals, with particular focus on calculus of variations, can be found in the monographs by Buttazzo [Bu89], Dacorogna [Da82, Da89], Giusti [Gi03], and Jost, Li [JoLi98].

Regularization theory for ill-posed problems is presented in detail in the book by Engl, Hanke, and Neubauer [EnHaNe96].

### Chapter 3

By now, a classical reference of convex analysis, which also was the basis of chapter 3 in this lecture notes, is the book by Ekeland and Temam [EkTe99]. Other extensive sources on convex analysis are the monographs by Aubin [Au98], who also discusses other equilibrium type problems like those arising in game theory, and Rockafellar [Ro97].

Monographs on convex analysis with a particular focus on nonsmooth analysis are those by Borwein and Lewis [BoLe00] and Clarke et. al. [Cl90, ClLeStWo98].

### Chapter 4

Classical books on numerical methods in optimization are the ones by Dennis, Schnabel [DeSch96] and Gill, Murray, Wright [GiMuWr81, GiMuWr91]. An extensive and modern introduction to this subject is given in the book by Nocedal and Wright [NoWr99], as well as by Bonnans et. al. [BoGiLeSa03] and Gould, Leyffer [GoLe03].

Moreover, several monographs deal with special numerical methods:

- Kelley [Ke99] discusses iterative methods for unconstrained problems and problems with simple bound constraints.
- Conn, Gould, and Toint [CoGoTo00] provide an extensive overview of trust-region methods.
- Sequential quadratic programming is discussed by Murray [Mu97a, Mu97b] and Boggs, Tolle [BoTo95].
- Interior point methods are discussed in the monographs by Jansen [Ja97], Nesterov and Nemirovskii [NeNe94], and Renegar [Re01], as well as in the book by Wright [Wr97], who gives a good introduction to modern primal-dual interior point methods. An easy-to-read introduction is the review article by Forsgren, Gill, and Wright [FoGiWr02].
- Special numerical methods for optimal control problems are discussed by Betts [Be98] and Neittanmäki, Tiba [NeTi94].
- The method of moving asymptotes is discussed in the papers by Svanberg [Sva87, Sva01].

### Chapter 5

General introduction to shape optimization are the monographs by Sokolowski, Zolesio [SoZo92], and Haslinger, Mäkinen [HaMa03]. The book by Delfour and Zolesio [DeZo01] includes several chapter on shape sensitivity analysis.

An applied introduction to shape optimization with particular emphasis on problems arising in fluid dynamics is the book by Mohamadi and Pironneau [MoPi01]. This book also discusses classical numerical methods based on parametrization of shapes, an approach which is also presented in the book by Laporte and LeTallec [LaLeT03].

A survey on level set methods in shape optimization is provided by Burger and Osher [BuOs04].

### Chapter 6

Applied introductions to topology optimization are the books by Bendsoe, Sigmund [BeSi03] and Haslinger, Neittanmäki [HaNe96]. The monograph by Allaire [Al02] provides a detailed discussion of the homogenization method in topology optimization.

Sigmund and Pettersson [SiPe98] provide as discussion of difficulties that can arise in the discretization of topology optimization problems. The phase-field approach for topology optimization is discussed by Bourdin and Chambolle [BoCha03].

# Bibliography


[Al02] G.Allaire, *Shape Optimization by the Homogenization Method* (Springer, New York, 2002).

[Au98] J.P.Aubin, *Optima and equilibria. An introduction to nonlinear analysis* (Springer, Berlin, 1998).

[Ba84] V.Barbu, *Optimal Control of Variational Inequalities* (Pitman, Boston, 1984).

[BeSi03] M.P.Bendsoe, O.Sigmund, *Topology Optimization. Theory, Methods and Applications* (Springer, Berlin, 2003).

[Be88] A.Bensoussan, *Perturbation Methods in Optimal Control* (Wiley/Gauthier-Villars, Chichester, Montrouge, 1988).

[Be98] J.T.Betts, *Practical Methods for Optimal Control using Nonlinear Programming* (SIAM, Philadelphia, 2001).

[BiCoCoSa97a] L.T.Biegler, T.F.Coleman, A.R.Conn, F.N.Santosa, eds., *Large-scale optimization with applications. Part I. Optimization in inverse problems and design* (Springer, New York, 1997).

[BiCoCoSa97b] L.T.Biegler, T.F.Coleman, A.R.Conn, F.N.Santosa, eds., *Large-scale optimization with applications. Part II. Optimal design and control* (Springer, New York, 1997).

[BoTo95] P.T.Boggs, J.W.Tolle, *Sequential quadratic programming*, Acta numerica 1995, 1–51.

[BoGiLeSa03] J.F.Bonnans, J.C.Gilbert, C.Lemarechal, C.A.Sagastizabal, *Numerical Optimization. Theoretical and Practical Aspects* (Springer, Berlin, 2003).

[BoLe00] J.M.Borwein, A.S.Lewis, *Convex Analysis and Nonlinear Optimization. Theory and examples* (Springer, New York, 2000).

[BoCha03] B.Bourdin, A.Chambolle, *Design-dependent loads in topology optimization.* ESAIM Control Optim. Calc. Var. **9** (2003), 19–48.

[BuOs04] M.Burger, S.J.Osher, *A survey on level set methods for inverse problems and optimal design*, CAM Report 04-02 (UCLA, 2004).

[Bu89] G.Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations* (Longman Scientific, Harlow, 1989).

[Cl90] F.H.Clarke, *Optimization and Nonsmooth Analysis* (SIAM, Philadelphia, 1990).

[ClLeStWo98] F.H.Clarke, Y.S.Ledyaev, R.J.Stern, P.R.Wolenski, *Nonsmooth Analysis and Control Theory* (Springer, New York, 1998).

[CoGoTo00] A.R.Conn, N.I.Gould, P.L.Toint, *Trust-region Methods* (SIAM, Philadelphia, 2000).

[Da82] B.Dacorogna, *Weak Continuity and Weak Lower Semicontinuity of Nonlinear Functionals* (Springer, Berlin, New York, 1982).

[Da89] B.Dacorogna, *Direct Methods in the Calculus of Variations* (Springer, Berlin, 1989).

[DeZo01] M.C.Delfour, J.P.Zolesio, *Shapes and Geometries. Analysis, Differential Calculus, and Optimization* (SIAM, Philadelphia, 2001).

[DeSch96] J.E.Dennis, R.B.Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (SIAM, Philadelphia, 1996).

[EkTe99] I.Ekeland, R.Tmam, *Convex Analysis and Variational Problems* (SIAM, Philadelphia, 1999).

[EnMcL93] H.W.Engl, J.McLaughlin, *Inverse Problems and Optimal Design in Industry* (Teubner, Stuttgart, 1994).

[EnHaNe96] H.W.Engl, M.Hanke, A.Neubauer, *Regularization of Inverse Problems* (Kluwer, Dordrecht, 1996).

[FoGiWr02] A.Forsgren, P.E.Gill, M.H.Wright, *Interior methods for nonlinear optimization*, SIAM Review 44 (2002), 525–597.

[GiMuWr81] P.E.Gill, W.Murray, M.H.Wright, *Practical Optimization* (Academic Press, London, New York, 1981).

[GiMuWr91] P.E.Gill, W.Murray, M.H.Wright, *Numerical Linear Algebra and Optimization* (Addison-Wesley, Redwood City, 1991).

[Gi03] E.Giusti, *Direct Methods in the Calculus of Variations* (World Scientific, River Edge, 2003).

[GoLe03] N.I.M.Gould, S.Leyffer, *An Introduction to Algorithms for Nonlinear Optimization* (Springer, Berlin, 2003).

[HoLaLeSpTr02] K.H.Hoffmann, I.Lasiecka, G.Leugering, J.Sprekels, F.Tröltzsch, eds., *Optimal Control of Complex Structures* (Birkhäuser, Basel, 2002).

[HoHoSch00] K.H.Hoffmann, R.H.W.Hoppe, V.Schulz, eds., *Fast Solution of Discretized Optimization Problems* (Birkhäuser, Basel, 2000).

[HoLeTr98] K.H.Hoffmann, G.Leugering, F.Tröltzsch, eds., *Optimal Control of Partial Differential Equations* (Birkhäuser, Basel, 1998).

[HaMa03] J.Haslinger, R.A.E.Mäkinen, *Introduction to Shape Optimization. Theory, Approximation, and Computation* (SIAM, Philadelphia, 2003).

[HaNe96] J.Haslinger, P.Neittaanmäki, *Finite Element Approximation for Optimal Shape, Material and Topology Design* (Wiley, Chichester, 1996).

[Ja96] J.Jahn, *Introduction to the Theory of Nonlinear Optimization* (Springer, Berlin, 1996).

[Ja97] B.Jansen, *Interior Point Techniques in Optimization. Complementarity, Sensitivity and Algorithms* (Kluwer, Dordrecht, 1997).

[JoLi98] J.Jost, X.Li, *Calculus of Variations* (Cambridge University Press, Cambridge, 1998).

[Ke99] C.T.Kelley, *Iterative Methods for Optimization* (SIAM, Philadelphia, 1999).

[LaLeT03] E.Laporte, P.LeTallec, *Numerical Methods in Sensitivity Analysis and Shape Optimization* (Birkhäuser, Boston, 2003).

[MoPi01] B.Mohammadi, O.Pironneau, *Applied Shape Optimization for Fluids* (Clarendon Press, Oxford University Press, New York, 2001).

[Mu97a] W.Murray, *Sequential quadratic programming methods for large-scale problems. Computational issues in high performance software for nonlinear optimization* , Comput. Optim. Appl. 7 (1997), 127–142.

[Mu97b] W.Murray, *Some Aspects of Sequential Quadratic Programming Methods. Large-scale Optimization with Applications, Part II* (Springer, New York, 1997).

[NeTi94] P.Neittaanmäki, D.Tiba, *Optimal Control of Nonlinear Parabolic Systems. Theory, Algorithms, and Applications* (Marcel Dekker, New York, 1994).

[NeNe94] Y.Nesterov, A.Nemirovskii, *Interior-point Polynomial Algorithms in Convex Programming* (SIAM, Philadelphia, 1994).

[NoWr99] J.Nocedal, S.J.Wright, *Numerical Optimization* (Springer, New York, 1999).

[Pe04] P.Pedregal, *Introduction to Optimization* (Springer, New York, 2004).

[Re01] J.A.Renegar, A Mathematical View of Interior-point Methods in Convex Optimization (SIAM, Philadelphia, 2001).

[Ro97] R.T.Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1997).

[SiPe98] O.Sigmund, J.Petersson, *Numerical instabilities in topology optimization: A survey of procedures dealing with checkerboards, mesh-dependencies, and local minima*, Struct. Optim. 16 (1998), 68-75.

[SoZo92] J.Sokolowski, J.P,Zolesio, *Introduction to Shape Optimization. Shape Sensitivity Analysis* (Springer, Berlin, 1992).

[Sva87] K.A.Svanberg, *The method of moving asymptotes—a new method for structural optimization*, Int. J. Numer. Meth. Engrg. 24 (1987), 359–373.

[Sva01] K.A.Svanberg, *A class of globally convergent optimization methods based on conservative convex separable approximations*, SIAM J. Optim. 12 (2001), 555–573.

[Tr96] J.L.Troutman, *Variational Calculus and Optimal Control* (Springer, New York, 1996).

[Wr97] S.J.Wright, *Primal-dual Interior-point Methods* (SIAM, Philadelphia, 1997).

[Ze85] E.Zeidler, *Nonlinear Functional Analysis and its Applications. III. Variational Methods and Optimization* (Springer, New York, 1985).