Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Optimization of Complex Systems

Lecture Notes in WiSe2023

5. Oktober 2022

Winnifried Wollner

winnifried.wollner@uni-hamburg.de
Universität Hamburg,
MIN Fakultät,
Fachbereich Mathematik

Preface

These notes accompany the lecture “Optimization of Complex Systems” in Winter 2022/2023. They have advanced from lecture notes from my lextures on “Optimization in Function Spaces” in Summer 2018 and Winter 2019/2020 at TU Darmstadt. and are based on lecture notes “Optimization in Function Spaces” by Prof. Dr. Stefan Ulbrich, TU Darmstadt and “Optimization in Banach Spaces” by Prof. Dr. Michael Ulbrich, TU München.

They are made available to participants for their private educational use. Please do not distribute them any further.

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

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

Calculus of Variations Dacorogna [2008]

Sobolev-spaces Adams and Fournier [2003]

Functional analysis (and function spaces) Alt [2002], Ciarlet [2013], Rudin [1991], Werner [2005]

Optimization with PDE constraints De los Reyes [2015], Hinze et al. [2009], Ito and Kunisch [2008], Tröltzsch [2005].

It is expected by the participants to have some background in functional analysis, the main results needed in the lecture are summarized in Appendix A if you need a reminder.

You may send comments and corrections to

winnifried.wollner@uni-hamburg.de

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

Winnifried Wollner

Hamburg, October 2022.

Inhaltsverzeichnis

1. Introduction

1.1. Introduction

In this lecture, we are interested in problems of the type

$$f(x) \to \min \qquad \text{s.t.} \quad x \in X^{\text{ad}} \tag{1.1}$$

where $f : X \to \mathbb{R}$ is a functional acting on a Banach space X and $X^{\text{ad}} \subset X$ denotes the admissible set.

In many cases the unknown variable denotes a function and the corresponding Banach space encodes the properties of the considered functions. There are many examples where such problems occure, to give only a few of them:

- Modelling of elastic materials, damage, fracture
- Optimal design (shape and topology) of buildings, ships, aircrafts, cars
- Optimal control of robot movements, space craft trajectories, heating and cooling processes (wine fermentation, steel hardening, ...)
- Inverse problems, e.g., seismic inversion, data assimilation in weather and climate models, tomography

1.2. Why Should we Consider the Function Space Case?

We will highlight this with the very basic question of existence of solutions. The following is well known from 'Calculus 2'

Theorem 1.2.1. *Let $f : X^{\text{ad}} \subset \mathbb{R}^n \to \mathbb{R}$ be continuous on the bounded closed set X^{ad}. Then (1.1) has at least one solution.*

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

Unfortunately, this is no longer true in infinite dimensions, as is shown by the following example

Example 1.2.2. (Counterexample) Consider the problem

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

with

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

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

$$x_k(s) = \begin{cases} -1 & s < -1/k, \\ ks & -1/k \le s \le 1/k, \\ 1 & s \ge 1/k, \end{cases}$$

satisfies

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

Hence it holds

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

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

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

Hence, we may expect that the right choice of the space where minimizers are searched for is crucial but also that we need an improved theory for the analysis.

1.2.1. Lavrentiev Phenomenon

It should be noted that, in general, not only the existence of a minimizer is depending on the choice of the spaces. Even the infimum values of the functional may depend on the chosen spaces. This is obvious, of course, since on any one-dimensional subspace not containing a minimizing sequence for the functional this is true. But infact, such defects can be a lot more subtle as it is demonstrated by the so called Lavrentiev phenomenon, Lavrentieff [1927] which stated that certain minimal functional values can not be approximated by Lipschitz continuous functions.

We illustrate this with the following example. Consider the functional

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

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

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

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

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

and $u(t_0) = v(t_0)$ due to the regularity (and boundary values) of u and v. **draw sketch** Hence for $t \in [0, t_0]$ and $\xi \in \mathbb{R}$ it holds ($|u^3 - t| \geq |v^3 - t|$, since $u^3 \leq v^3 \leq t/8 \leq t$)

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

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

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

Hence we obtain

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

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

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

Which shows the claim.

2. Existence of Solutions

We will now discuss how to obtain the existence of solutions to minimization problems in Banach spaces.

There are essentially two possibilities to cure the defect seen in Example 1.2.2. One is by adding assumptions on the feasible set while the other adds assumptions for the functions.

In both cases we assume that our problem is given in the form

$$\min \quad f(x) \qquad G(x) \in K \tag{P}$$

where

Assumption 2.0.1.

1. $G : X \to Z$ is a continuous operator between two (real) Banach spaces,
2. $f : X \to \mathbb{R}$ is lower semicontinuous.
3. $K \subset Z$ is closed and convex,

as a reminder:

Definition 2.0.2. A function $f : X \to \mathbb{R}$ is lower semicontinuous, if for any convergent sequence $x^k \in X$ it holds

$$\liminf_{k\to\infty} f(x^k) \geq f(\lim_{k\to\infty} x^k),$$

or, as an equivalent geometric condition, if for any $\alpha \in \mathbb{R}$ the sets

$$\{x \in X \mid f(x) \leq \alpha\}$$

are (sequentially) closed.

It is clear by this, that the admissible set

$$X^{\mathrm{ad}} = \{x \in X \mid G(x) \in K\}$$

is closed since G is continuous and K is closed.

The notation in (P) contains all the usual conditions, e.g.,

$$h(x) = 0, \quad g(x) \in C, \quad x \in M$$

for continuous operators $h : X \to Z_1$, $g : X \to Z_2$ for Banach spaces Z_1, Z_2, a closed convex cone $C \in Z_2$, and a closed convex set $M \subset X$ can be modeled in this setting by choosing $Z = Z_1 \times Z_2 \times X$ with $K = \{0\} \times C \times M$ and $G(x) = (h(x), g(x), x)^T$.

The constraint $F(x) \in C$ for a closed convex cone is an abstract inequality constraint, as the usual component-wise inequality $g(x) \leq 0$ in $\mathbb{R}^m$ from nonlinear optimization can also be written as the inclusion

$$g(x) \in C = (-\infty, 0]^m.$$

2.1. Existence by Compactness

Now, we show the first alternative of generalizing the existence Theorem 1.2.1.

Theorem 2.1.1. *Let Assumption 2.0.1 be satisfied and assume that there is a point $x^0 \in X^{\mathrm{ad}}$ such that the level-set*

$$\mathcal{N}(x^0) = \{x \in X^{\mathrm{ad}} \mid f(x) \leq f(x^0)\}$$

is compact. Then Problem (P) has at least one global solution.

Beweis. Let $x^k \in \mathcal{N}(x^0)$ be a minimizing sequence, i.e.,

$$\lim_{k\to\infty} f(x^k) = \inf_{x \in X^{\mathrm{ad}}} f(x) =: f^*.$$

By compactness of $\mathcal{N}(x^0) \subset X^{\mathrm{ad}}$, there exists a convergent subsequence $x^k \to \bar{x}$. Lower semicontinuity asserts

$$f^* = \lim_{k\to\infty} f(x^k) \geq f(\lim_{k\to\infty} x^k) = f(\bar{x}) \geq f^*$$

and thus $\bar{x}$ is a global minimizer. □

Remark 2.1.2. Notice, that level-set $\mathcal{N}(x^0)$ is always closed for a lower semicontinuous function, so the essential assumption of Theorem 2.1.1 is that in addition the set $\mathcal{N}(x^0)$ is precompact.

While this appears to be very natural (we only use the Bolzano-Weierstrass Theorem in finite-dimensions to see the equivalence to bounded and closed), the characterization of compact sets in infinite dimensional spaces is a bit more cumbersome. For instance on a bounded domain (open and connected) $\Omega \subset \mathbb{R}^d$ a subset $M \subset C(\overline{\Omega})$ is relatively compact ($\operatorname{cl} M$ is compact) if and only if the functions in M are uniformly bounded, i.e.,

$$\sup_{f \in M} \|f\|_\infty < \infty$$

and equicontinuous, i.e., for any $\varepsilon > 0$ there is some δ such that

$$\sup_{f \in M} |f(x) - f(y)| \leq \varepsilon, \qquad \forall \|x - y\| \leq \delta$$

as it is asserted by the Arzelà-Ascoli theorem.

2.2. Existence by Weak-Convergence

To avoid checking for compactness (which is often not given) an alternative lies in enforcing existence of minimizers by increasing the conditions on the operators f and G.

To this end, we recall from functional analysis

Definition 2.2.1. We say that a sequence $x^k \in X$ converges weakly to $x \in X$ (written as $x^k \rightharpoonup x$) if

$$\langle x^*, x^k \rangle \to \langle x^*, x \rangle \qquad \forall x^* \in X^*$$

where $X^* = \mathscr{L}(X; \mathbb{R})$ denotes the dual space to X.

Before we continue, we collect a few useful facts:

Theorem 2.2.2. *Let X be a real Banach space, then it holds:*

1. *any closed (norm-topology) and convex set is weak sequentially closed (Banach-Saks-Mazur theorem),*
2. *if X is reflexive, then any bounded sequence $x^k \in X$ contains a weakly convergent subsequence (Banach-Eberlein-Šmulian theorem),*
3. *a weakly convergent sequence is bounded,*
4. *if $f : X \to \mathbb{R}$ is convex and lower semicontinuous, then f is weakly lower semiconti-*

nuous (w.l.s.c), i.e., for any $x^k \rightharpoonup x$ it holds

$$\liminf_{k\to\infty} f(x^k) \geq f(x).$$

In particular, it is 'easy' to check whether a sequence contains a weakly convergent subsequence!

With this we can now formulate our second existence result

Theorem 2.2.3. *Assume that*

1. *$G : X \to Z$ is weakly sequentially continuous on the **reflexive** Banach space X to the Banach space Z, i.e.,*

$$x_k \rightharpoonup x \qquad \text{then} \qquad G(x_k) \rightharpoonup G(x),$$

2. *$f : X \to \mathbb{R}$ is weakly lower semicontinuous,*
3. *$K \subset Z$ is closed and convex,*
4. *there is a point $x^0 \in X^{\mathrm{ad}}$ such that the level-set*

$$\mathcal{N}(x^0) = \{x \in X^{\mathrm{ad}} \mid f(x) \leq f(x^0)\}$$

is bounded.

Then Problem (P) has at least one global solution.

Beweis. As in the proof of Theorem 2.1.1, we select a minimizing sequence $x^k \in \mathcal{N}(x^0)$ with

$$\lim_{k\to\infty} f(x^k) = \inf_{x\in X^{\mathrm{ad}}} f(x) = f^*.$$

By boundedness of $\mathcal{N}(x^0)$ and Theorem 2.2.2-2. this sequence contains a subsequence with weak limit $\bar{x}$. By the assumed weak continuity $G(x^k) \rightharpoonup G(\bar{x})$ and since K is closed and convex Theorem 2.2.2-1. asserts $\bar{x} \in X^{\mathrm{ad}}$.

Now, f is w.l.s.c and thus

$$f^* = \lim_{k\to\infty} f(x^k) \geq f(\bar{x}) \geq f^*$$

and thus $\bar{x}$ is a global minimizer. □

This leaves us with the remaining problem to see which functions are weak continuous or lower semicontinuous. Since every convergent sequence is also weakly convergent this is a stronger requirement than continuity. One possibility is utilizing Theorem 2.2.2-4.

Further, in many situations it is helpful to utilize compactness properties of mappings.

Definition 2.2.4. A continuous linear operator $A: X \to Y$ between two Banach spaces is called compact, if for any bounded sequence $x^k \in X$ the sequence $Ax^k \in Y$ contains a convergent subsequence.

In particular, a Banach space X is compactly embedded in a Banach space X_0 if the embedding $\mathrm{id}: X \to X \subset X_0$ is compact.

Lemma 2.2.5. *If a linear operator $A: X \to Y$ is compact, and $x^k \rightharpoonup x$ in X then $Ax^k \to Ax$ in Y.*

Beweis. Exercise! □

To show how these techniques can be applied, we consider the following example.

Example 2.2.6. Consider the problem

$$\begin{aligned} \min \quad & J(q,u) = \frac{1}{2}\|u-u^d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|q\|^2_{L^2(\partial\Omega)} \\ \text{s.t.} \quad & -\Delta u + u^3 = 0 \qquad \text{on } \Omega, \\ & \partial_n u + u = q \qquad \text{on } \partial\Omega, \\ & a \le q \le b \qquad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\Omega \subset \mathbb{R}^d$, $d = 2,3$ a Lipschitz-domain and $a, b \in L^2(\partial\Omega)$ with w.l.o.g. $a \le 0 \le b$.

Then we can define $Q = L^2(\partial\Omega)$, $U = H^1(\Omega)$, and

$$Q^{\mathrm{ad}} = \{q \in Q \,|\, a \le q \le b\}.$$

For the equality constraint, we consider the weak form to define

$$\begin{aligned} & E: (q,u) \in Q \times U \to U^* \\ & \langle E(q,u), v\rangle := (\nabla u, \nabla v)_{L^2(\Omega)^n} + (u^3, v)_{L^2(\Omega)} + (u-q, v)_{L^2(\partial\Omega)}. \end{aligned}$$

Although it can be shown, that for any given $q \in Q$ the equation $E(q,u) = 0$ admits a unique solution, we will not use this fact. Instead, we just notice, that the embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ is continuous (Sobolev-embedding theorem) and that $H^1(\Omega) \hookrightarrow L^5(\Omega)$ is compact (Rellich-Kondrachov) if $d = 2,3$.

We can now start to check the prerequisites of the existence Theorem 2.2.3. To this end, we fix $X = Q \times U$, $Z = Q \times U^*$, $K = Q^{\mathrm{ad}} \times \{0\}$, $G(x) = G(q,u) = (q, E(q,u))^T$, $f(x) = f(q,u) = J(q,u)$.

Now, we check the prerequisites.

1. First, we check the weak continuity of G. To this end, let $x^k = (q^k, u^k) \rightharpoonup x = (q, u)$ be a weak convergent sequence. for the first component of G nothing is to be shown. For the second component, we notice, that all continuous linear terms immediately converge by definition as, e.g.,

$$(\nabla u^k, \nabla v) \to (\nabla u, \nabla v)$$

by definition. For the nonlinear term, we recall that $H^1(\Omega) \subset L^5(\Omega)$ is compact, and hence $u^k \to u$ (strongly) in $L^5(\Omega)$. As a consequence $(u^k)^3 \to u^3$ in $L^{5/3}(\Omega)$, because the inequality $(a, b \in \mathbb{R})$

$$|a^3 - b^3| = |a^2 + ab + b^2||a - b| \leq \frac{3}{2}|a^2 + b^2||a - b|$$

shows via Hölder's inequality with $\frac{2}{5} + \frac{1}{5} = \frac{3}{5}$

$$\|(u^k)^3 - u^3\|_{L^{5/3}} \leq \frac{3}{2}\Big(\|(u^k)^2\|_{L^{5/2}} + \|u^2\|_{L^{5/2}}\Big)\|u^k - u\|_{L^5} \to 0.$$

Now, $U \subset L^{5/2}(\Omega) = L^{5/3}(\Omega)^*$ and thus, for any $v \in U = H^1(\Omega)$,

$$((u^k)^3 - u^3, v)_{L^2(\Omega)} \leq \|(u^k)^3 - u^3\|_{L^{5/3}}\|v\|_{H^1(\Omega)} \to 0$$

which shows

$$G(x^k) \rightharpoonup G(x).$$

Further, X is a Hilbert-space and thus reflexive.

2. To see weak lower semi-continuity of f, we refer to Theorem 2.2.2-4.
3. Clearly K is closed and convex.
4. To see, the last point, let $x^0 = (0, 0)$ be given. Clearly $x^0 \in X^{\text{ad}}$. Now, to see boundedness, we notice that for $\|q^k\|_{L^2(\partial\Omega)} \to \infty$ it is

$$f(x^k) \geq \frac{\alpha}{2}\|q^k\|^2_{L^2(\partial\Omega)} \to \infty$$

and hence the q-component remains bounded. (This is alternatively clear from the bounds $a \leq q \leq b$)

For the u-component assume, that $(q,u) \in Q \times U$ solves $E(q,u) = 0$. Then testing the weak equation with $v = u$ one obtains

$$\begin{aligned} \|\nabla u\|^2_{L^2(\Omega)^n} + \|u\|^4_{L^4(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} &= (q,u)_{L^2(\partial\Omega)} \\ &\leq \|q\|_{L^2(\partial\Omega)} \|u\|_{L^2(\partial\Omega)} \\ &\leq \frac{1}{2}\|q\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\|u\|^2_{L^2(\partial\Omega)} \end{aligned}$$

and consequently

$$\|\nabla u\|^2_{L^2(\Omega)^n} + \frac{1}{2}\|u\|^2_{L^2(\partial\Omega)} \leq \frac{1}{2}\|q\|^2_{L^2(\partial\Omega)}.$$

Using a generalized Poincaré-Friedrichs inequality, i.e.,

$$\|u\|^2_{L^2(\Omega)} \leq c_\Omega\Big(\|\nabla u\|^2_{L^2(\Omega)^n} + \frac{1}{2}\|u\|^2_{L^2(\partial\Omega)}\Big),$$

shows

$$\|u\|^2_{H^1(\Omega)} \leq c_\Omega \|q\|^2_{L^2(\partial\Omega)}$$

and thus $\mathcal{N}(0)$ is bounded.

Consequently, Theorem 2.2.3 asserts the existence of at least one solution.

3. Optimality Conditions

Analogously to the finite dimensional case, we will try to characterize local minimizers by conditions on the derivatives of the involved functionals. Hence we start with a recollection on differentiability in Banach-spaces.

3.1. Differentiability

Let X, Y be two Banach spaces, and $F : X \to Y$ be some mapping.

Definition 3.1.1. For given $x \in X$ and $\delta x \in X$ assume that the limit

$$F'(x;\delta x) = \lim_{t \downarrow 0} \frac{F(x + t\delta x) - F(x)}{t}$$

exists, then $F'(x;\delta x)$ is called the *directional derivative* of F at x in direction δx.

Note that the directional derivative is not necessarily linear, e.g., consider $F : \mathbb{R} \to \mathbb{R}$, $F(x) = |x|$ which is directionally differentiable in $x = 0$ but $F'(0;1) = 1 = F'(0;-1)$.

Definition 3.1.2. Assume that $F'(x;\delta x)$ exists for all $\delta x \in X$ and that there exists an operator $A \in \mathscr{L}(X, Y)$ such that

$$A\delta x = F'(x;\delta x) \quad \forall\, \delta x \in X$$

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

Definition 3.1.3. F is called *Fréchet-differentiable* at x if there exists a function $\omega : X \to Y$ with $\omega(\delta x) \in o(\|\delta x\|_X)$ such that

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

In this case $F'(x)$ is called the *Fréchet-derivative*.

The notion $\omega(\delta x) \in o(\|\delta x\|_X)$ means

$$\lim_{\delta x \to 0} \frac{\|\omega(\delta x)\|_Y}{\|\delta x\|_X} = 0$$

Example 3.1.4.

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

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

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

$$F'(x) = 2x \qquad \in \mathscr{L}(H, \mathbb{R}) = H^*.$$

(Exercise)

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

$$\begin{aligned} F'(x; \delta x)(s) &= \lim_{t \downarrow 0} \frac{1}{t} \Big(\sin(x(s) + t\delta x(s)) - \sin(x(s)) \Big) \\ &= \cos(x(s)) \delta x(s) \end{aligned}$$

noting that the convergence is uniform w.r.t. $s \in [0,1]$ due to the uniformly bounded second derivative of sin. Now we know from Example A.3.5 that the operator given by pointwise multiplication with $\cos(x(s))$ is in $\mathscr{L}(X, X)$ and hence sin is Gâteaux-differentiable with $\sin' = \cos$. To obtain Fréchet-differentiability we note that

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

and hence

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

i.e., the mapping is Fréchet-differentiable.

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

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

Now we can define a sequence of functions

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

and obtain

$$\|\omega(\delta x_\varepsilon)\|_p^p = \int_0^1 |\omega(\delta x_\varepsilon(s))|^p \, \mathrm{d}s = \int_0^\varepsilon (1-\sin(1))^p \, \mathrm{d}s = (1-\sin(1))^p \varepsilon.$$

This gives

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

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

Remark 3.1.5. The last example is not a pathological case, in fact, if $\psi\colon \mathbb{R} \to \mathbb{R}$ generates a Nemyzkii-operator $\Psi\colon W^{1,p}(\Omega) \to W^{1,p}(\Omega)$ with $1 \le p < \infty$, then this operator is Fréchet-differentiable iff Ψ is continuous and **affine-linear**, see, e.g., Appell and Zabrejko [1990].

Theorem 3.1.6. *Let X, Y, Z be Banach spaces and let $\mathcal{O}_X \subset X$ and $\mathcal{O}_Y \subset Y$ be open. Then the following is true:*

1. *(Sum-rule)*
   If $f, g\colon \mathcal{O}_X \to Y$ are Gâteaux or Fréchet-differentiable then so is $\lambda f + g$ with $\lambda \in \mathbb{R}$ and it holds
   $$(\lambda f + g)'(x) = \lambda f'(x) + g'(x).$$

2. *(Mean-value theorem)*
   If $f : \mathcal{O}_X \to Y$ is Gâteaux-differentiable and the interval $I = [x_0, x_1] = \{\lambda x_0 + (1-\lambda)x_1 \mid \lambda \in [0,1]\}$ is contained in $\mathcal{O}_X$, then
   $$\|f(x_0) - f(x_1)\|_Y \le \sup_{\xi \in I} \|f'(\xi)\|_{\mathcal{L}(X,Y)} \|x_0 - x_1\|_X$$

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

4. *(Chain rule)*
Let $f : \mathcal{O}_X \to Y$ and $g : \mathcal{O}_Y \to Z$ with $f(\mathcal{O}_X) \subset \mathcal{O}_Y$ be Fréchet-differentiable in x_0 and $f(x_0)$, then $g \circ f$ is Fréchet-differentiable in x_0 and it holds

$$(g \circ f)'(x_0) = g'(f(x_0)) \circ f'(x_0).$$

5. *(Implicit function theorem)*
Let $F : X \times Y \supset \mathcal{O}_X \times \mathcal{O}_Y \to Z$ be continuously differentiable with $F(x_0, y_0) = 0$. Further, assume that the derivative of $y \mapsto F(x_0, y)$ is an isomorphism from Y to Z at $y = y_0$. Then there exists neighborhoods $N(x_0)$ of x_0 and $N(y_0)$ of y_0 such that for any $x \in N(x_0)$ the equation $F(x,y) = 0$ admits a unique solution $y = f(x) \in N(y_0)$. The mapping $f : N(x_0) \to Y$ defined by this is continuously differentiable.

3.2. First Order Necessary Optimality Conditions

We will now derive some optimality conditions for solutions of (P) and set

$$X^{\mathrm{ad}} = \{x \in X \mid G(x) \in K\}.$$

Definition 3.2.1. We say that $\overline{x} \in X^{\mathrm{ad}}$ is a *local solution* to the problem (P), with respect to the norm on X, if there exists $r > 0$ such that

$$f(\overline{x}) \leq j(x) \quad \forall\, x \in X^{\mathrm{ad}}, \|x - \overline{x}\|_X \leq r.$$

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

Analogous to the finite dimensional case, we will define a cone of tangent directions as

Definition 3.2.2. For a set $X^{\mathrm{ad}} \subset X$ in a Banach space X, we define the *Bouligand*-cone (or *sequential tangent cone*, or *contingent*-cone) at $\overline{x}$ as follows

$$\begin{aligned} T(X^{\mathrm{ad}}, \overline{x}) :&= \{d \in X \mid d = \lim_{k\to\infty} \tfrac{1}{t_k}(x^k - \overline{x}), t_k \downarrow 0, x^k \in X^{\mathrm{ad}}\} \\ &= \{d \in X \mid d = \lim_{k\to\infty} \eta_k(x^k - \overline{x}), \eta_k > 0, \overline{x} \leftarrow x^k \in X^{\mathrm{ad}}\}. \end{aligned}$$

This cone is always closed (Exercise)!

With this, we have the following

Theorem 3.2.3. *Let $f : O_X \subset X \to \mathbb{R}$ be defined on an open neighborhood of $X^{\mathrm{ad}} \subset X$, and let $\overline{x} \in X^{\mathrm{ad}}$ be a local solution of (P). If f is Fréchet-differentiable in $\overline{x}$ then*

$$f'(\overline{x})d \geq 0 \quad \forall\, d \in T(X^{\mathrm{ad}}, \overline{x}).$$

Beweis. Let $d \in T(X^{\mathrm{ad}}, \overline{x})$ be given and t_k and $x^k \in X^{\mathrm{ad}}$ be the approximating sequences. Then it holds

$$f'(\overline{x})d = \lim_{k\to\infty} \frac{1}{t_k} f'(\overline{x})(x^k - \overline{x}) = \lim_{k\to\infty} \frac{1}{t_k}\Big(f(x^k) - f(\overline{x}) - \omega(x^k - \overline{x})\Big).$$

Using $x^k \to \overline{x}$ and local optimality of $\overline{x}$, we get

$$f'(\overline{x})d \geq \lim_{k\to\infty} \frac{1}{t_k}\omega(x^k - \overline{x}) = \lim_{k\to\infty} \frac{\|x^k - \overline{x}\|_X}{t_k} \frac{\omega(x^k - \overline{x})}{\|x^k - \overline{x}\|_X} = \|d\|_X \lim_{k\to\infty} \frac{\omega(x^k - \overline{x})}{\|x^k - \overline{x}\|_X} = 0.$$

□

As in finite-dimensions, this characterization is not considered to be sufficient, since the set $T(X^{\mathrm{ad}}; \overline{x})$ is hard to describe in general. Thus we will try to replace it by simpler settings.

Theorem 3.2.4 (The Convex Case)**.** *Let $X^{\mathrm{ad}} \subset X$ be convex, and $f : X^{\mathrm{ad}} \to \mathbb{R}$ be directionally differentiable. Then for any local solution $\overline{x} \in X^{\mathrm{ad}}$ it holds variational inequality (VI)*

$$f'(\overline{x}; x - \overline{x}) \geq 0 \quad \forall\, x \in X^{\mathrm{ad}}. \tag{3.1}$$

Further, if f is convex on X^{ad} then any $\overline{x} \in X^{\mathrm{ad}}$ which satisfies (3.1) is a global solution to (P).

Beweis. The proof for the necessary condition is analogous to the proof of Theorem 3.2.3. For any $x \in X^{\mathrm{ad}}$, we set $d = x - \overline{x}$ and observe that for any $t \in [0,1]$ $\overline{x} + td = (1-t)\overline{x} + tx \in X^{\mathrm{ad}}$ and thus

$$f'(\overline{x})(x - \overline{x}) = \lim_{t\downarrow 0} \frac{f(\overline{x} + td) - f(\overline{x})}{t} \geq 0.$$

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

$$\frac{f(\overline{x} + \lambda(x - \overline{x})) - f(\overline{x})}{\lambda} \leq f(x) - f(\overline{x})$$

and thus

$$f(x) - f(\overline{x}) \geq \delta f(\overline{x}; x - \overline{x})$$

which shows the assertion. □

Remark 3.2.5. The convex case is indeed a generalization of Theorem 3.2.3, since for $\overline{x} \in X^{\mathrm{ad}}$ it holds

$$T(X^{\mathrm{ad}}; \overline{x}) = \mathrm{cl}\{d \in X \mid d = \lambda(x - \overline{x}), \lambda > 0, x \in X^{\mathrm{ad}}\}$$

for any convex set X^{ad}.

As in finite-dimensions, for general functionals and sets, we would like to replace $T(X^{\mathrm{ad}}; \overline{x})$ by a simpler object.

Definition 3.2.6. Let $G : X \to Z$ be Fréchet-differentiable at $\overline{x} \in X^{\mathrm{ad}}$. Then we define the *linearizing cone* at $\overline{x}$ as

$$T_l(G, K, \overline{x}) = \{d \in X \mid G'(\overline{x})d \in T(K, G(\overline{x}))\}.$$

note, that we keep the Tangent-cone to the convex set K, as it is 'easy' to calculate following Remark 3.2.5. As in finite-dimensions, we have the inclusion

Lemma 3.2.7. *Let G be Fréchet-differentiable at $\overline{x} \in X^{\mathrm{ad}}$. Then it holds*

$$T(X^{\mathrm{ad}}, \overline{x}) \subset T_l(G, K, \overline{x}).$$

Beweis. Let $d \in T(X^{\mathrm{ad}}, \overline{x})$ be given with corresponding sequences $t_k \downarrow 0$ and $X^{\mathrm{ad}} \ni x^k \to \overline{x}$. Then

$$\frac{G(x^k) - G(\overline{x})}{t_k} = G'(\overline{x})\frac{x^k - \overline{x}}{t_k} + \frac{\omega(x^k - \overline{x})}{\|x^k - \overline{x}\|_X}\frac{\|x^k - \overline{x}\|_X}{t_k} \to G'(\overline{x})d.$$

Further, $K \ni G(x^k) \to G(\overline{x}) \in K$, and thus

$$G'(\overline{x})d = \lim_{k \to \infty} \frac{G(x^k) - G(\overline{x})}{t_k} \in T(K, G(\overline{x})).$$

□

Hence, in order to replace the Bouligand-cone by the linearizing cone, we need to assume

Definition 3.2.8. A point $\overline{x} \in X^{\mathrm{ad}}$ satisfies the *Abadie Constraint Qualification* (ACQ), iff it holds

$$T_l(G, K, \overline{x}) \subset T(X^{\mathrm{ad}}, \overline{x}). \tag{ACQ}$$

As in finite dimensions, we would like to have more easily accessible conditions to verify if (ACQ) holds.

Definition 3.2.9. A point $\overline{x} \in X^{\mathrm{ad}}$ is said to satisfy *Robinson's Constraint Qualification* (RCQ) Robinson [1976], iff it holds

$$0 \in \operatorname{int}(G(\overline{x}) + G'(\overline{x})X - K). \tag{RCQ}$$

Before we can show, that (RCQ) is in fact a constraint qualification, we need a preparatory result on so called *metric regularity*, see, e.g., [Bonnans and Shapiro, 1998, Proposition 3.3] for the statement or [Bonnans and Shapiro, 2000, Theorem 2.87] for the proof.

Lemma 3.2.10. *Let $G\colon X \to Z$ be continuously Fréchet-differentiable near $\overline{x} \in X^{\mathrm{ad}} = G^{-1}(K)$, where as before $K \subset Z$ is closed and convex. Then assuming (RCQ) to hold at $\overline{x}$, there exist constants $c, \delta > 0$ such that*

$$\operatorname{dist}(x, G^{-1}(K - z)) \leq c \operatorname{dist}(G(x) + z, K)$$

holds for all $x \in B^X_\delta(\overline{x}) \subset X$ and $z \in B^Z_\delta(0) \subset Z$. As usual in the above, we have

$$G^{-1}(K - z) = \{x \in X \mid G(x) + z \in K\}, \qquad \operatorname{dist}(b, A) = \inf_{a \in A} \|a - b\|.$$

Remark 3.2.11. While we do not provide the technical proof of this result –compare[Bonnans and Shapiro, 2000, Theorem 2.87]–, we would like to motivate it in case the stronger (Exercise) constraint qualification that $G'(\overline{x})$ is an isomorphism holds. As we will only need the case $z = 0$, we will fix $z = 0$ for this. Since $G'(\overline{x})$ is an isomorphism, we obtain from the implicit function Theorem 3.1.6-5. that there exists neighborhoods $B^X_{2\delta_x}(\overline{x}) \subset X$ and $B^Z_{2\delta_z}(G(\overline{x})) \subset Z$ such that $G\colon B^X_{2\delta_x}(\overline{x}) \to B^Z_{2\delta_z}(G(\overline{x}))$ is invertible and that

$$\sup_{x \in B^X_{2\delta_x}(\overline{x})} \|G'(x)\|_{L(X;Z)} + \sup_{z \in B^Z_{2\delta_z}(G(\overline{z}))} \|(G^{-1})'(z)\|_{L(Z;X)} \leq C.$$

draw sketch Now, let $x \in B^X_{\delta_x}(\overline{x})$ and assume (potentially shrinking δ_x) that $G(B^X_{\delta_x}(\overline{x})) \subset$

$B^Z_{\delta_z}(G(\overline{x}))$. Then, taking a minimizing sequence $z^k \in K$ of the distance to $G(x)$, i.e.,

$$\|G(x)-z^k\|_Z \to \inf_{k\in K}\|G(x)-k\|_Z$$

we notice, that since $G(\overline{x}) \in K$ we can assume that $z^k \in B^Z_{2\delta_z}(G(\overline{x}))$ and thus define $x^k = G^{-1}(z^k)$. Then we obtain, with the mean-value Theorem 3.1.6-2

$$\begin{aligned}\operatorname{dist}(x, G^{-1}(K)) &\le \|x-x^k\|_X \\ &= \|G^{-1}(G(x))-G^{-1}(z^k)\|_X \\ &\le \sup_{z\in B^Z_{2\delta_z}(G(\overline{x}))}\|(G^{-1})'(z)\|_{L(Z,X)}\|G(x)-z^k\|_Z \\ &\le C\|G(x)-z^k\|_Z \to C\operatorname{dist}(G(x),K).\end{aligned}$$

We are now in position to show that (RCQ) is a constraint qualification.

Theorem 3.2.12. *If G is continuously Fréchet-differentiable near $\overline{x} \in X^{\mathrm{ad}}$ and (RCQ) holds at $\overline{x}$. Then (ACQ) is satisfied at $\overline{x}$, i.e., $T(X^{\mathrm{ad}}, \overline{x}) = T_l(G, K, \overline{x})$.*

Beweis. In view of Lemma 3.2.7, it is sufficient to show $T_l(G,K,\overline{x}) \subset T(X^{\mathrm{ad}},\overline{x})$. Hence, let $d \in T_l(G,K,\overline{x})$ be given, and let $y = G'(\overline{x})d \in T(K, G(\overline{x}))$. By definition there exists $z^k \in K$ with $z^k \to G(\overline{x})$ and $t_k \downarrow 0$ such that

$$y^k = \frac{z^k - G(\overline{x})}{t_k} \to y.$$

draw sketch

Taylor-expansion gives

$$\begin{aligned}G(\overline{x}+t_k d) &= G(\overline{x}) + t_k G'(\overline{x})d + \omega(t_k d) \\ &= z^k + (G(\overline{x})-z^k) + t_k y + \omega(t_k d) \\ &= z^k + t_k(y-y^k) + \omega(t_k d),\end{aligned}$$

with $\omega(t_k d) \in o(t_k\|d\|_X)$. Taking $x = \overline{x} + t_k d$ and $z = 0$ in Lemma 3.2.10 (and k sufficiently large so that $x \in B^X_\delta(\overline{x})$), we get, using $z^k \in K$, that

$$\begin{aligned}\operatorname{dist}(\overline{x}+t_k d, X^{\mathrm{ad}}) &\le c\operatorname{dist}(G(\overline{x}+t_k d), K) \\ &= c\operatorname{dist}(z^k + t_k(y-y^k) + \omega(t_k d), K) \\ &\le c\|(z^k + t_k(y-y^k)+\omega(t_k d)) - z^k\|_z \\ &= c\|t_k(y-y^k)+\omega(t_k d)\|_Z.\end{aligned}$$

Consequently, for k large enough, there is $x^k \in X^{\mathrm{ad}}$ such that

$$\|(\overline{x} + t_k d) - x^k\|_X \leq c\|t_k(y - y^k) + \omega(t_k d)\|_Z + \frac{t_k}{k}.$$

Dividing by t_k asserts

$$\left\| \frac{x^k - \overline{x}}{t_k} - d \right\|_X \leq c\|y - y^k\|_Z + \frac{\|\omega(t_k d)\|_Z}{t_k} + \frac{1}{k} \to 0.$$

Hence, we have found $x^k \in X^{\mathrm{ad}}$ with $\frac{x^k - \overline{x}}{t_k} \to d$ and thus $d \in T(X^{\mathrm{ad}}, \overline{x})$. □

Before we continue, we would like to state a few additional constraint qualifications that are potentially easier to verify.

Definition 3.2.13. We say, that a point $\overline{x} \in X^{\mathrm{ad}}$ satisfies the *linearized Slater* constraint qualification if there exists $d \in X$ such that

$$G(\overline{x}) + G'(\overline{x})d \in \mathrm{int}(K). \tag{3.2}$$

Clearly, this condition requires that $\mathrm{int}(K) \neq \emptyset$.

Lemma 3.2.14. *Assume that* $\mathrm{int}(K) \neq \emptyset$*, then* (3.2) *is equivalent to* (RCQ).

Beweis. Let (3.2) be satisfied. Then, for $\delta > 0$ sufficiently small, it is $G(\overline{x}) + G'(\overline{x})d + B_\delta^Z(0) \subset K \subset Z$. Consequently, it holds

$$B_\delta^Z(0) \subset G(\overline{x}) + G'(\overline{x})d - K \subset G(\overline{x}) + G'(\overline{x})X - K$$

and thus (RCQ) holds.

Conversely, if (3.2) is violated, then the convex sets $G(\overline{x}) + G'(\overline{x})X$ and $\mathrm{int}(K)$ have an empty intersection. By the Hahn-Banach theorem, there exists a separating hyperplane, i.e., $z^* \in Z^* \setminus \{0\}$ such that for any $d \in X$ and $k \in K$ it holds

$$\langle z^*, G(\overline{x}) + G'(\overline{x})d \rangle_{Z^*,Z} \geq \langle z^*, k \rangle_{Z^*,Z}.$$

Now, let $z \in Z$ be such that $\langle z^*, z \rangle < 0$. Then for any $t > 0$, $d \in X$ and $k \in K$ it is

$$\langle z^*, G(\overline{x}) + G'(\overline{x})d - k \rangle_{Z^*,Z} \geq 0 > \langle z^*, tz \rangle.$$

Consequently, $tz \notin G(\overline{x}) + G'(\overline{x})X - K$ and since $tz \to 0$ (RCQ) can not hold. □

Remark 3.2.15. It is easy to see, that if $G'(\overline{x}) \in \mathscr{L}(X;Z)$ is surjective, then (RCQ) holds. This corresponds to the well known linear independence constraint qualification.

Finally, we want to see, that (RCQ) corresponds to the well known Mangasarian Fromovitz constraint qualification (MFCQ) in finite dimensions.

To see this, consider the NLP

$$\min f(x) \quad \text{s.t.} \quad g(x) \leq 0, \quad h(x) = 0$$

with differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$. In our notation, this means $X = \mathbb{R}^n$, $Z = \mathbb{R}^m \times \mathbb{R}^p$, $G(x) = \begin{pmatrix} g(x) \\ h(x) \end{pmatrix}$, $K = (-\infty, 0]^m \times \{0\}^p$. Now, we want to see, that (RCQ) is equivalent to the MFCQ

$$\operatorname{rank} \nabla h(\overline{x}) = p, \qquad \text{there exists } d \in \mathbb{R}^n : \nabla h(\overline{x})^T d = 0, \quad \nabla g_i(\overline{x}) d < 0 \text{ whenever } g_i(\overline{x}) = 0.$$

"$\Rightarrow$" Now, let $\overline{x} \in X^{\mathrm{ad}}$ be given, and (RCQ) be satisfied, i.e.,

$$0 \in \operatorname{int} \left\{ \begin{pmatrix} g(\overline{x}) + \nabla g(\overline{x})^T d - v \\ \nabla h(\overline{x})^T d \end{pmatrix} \,\middle|\, d \in \mathbb{R}^n, v \in \mathbb{R}^m, v \leq 0 \right\}.$$

The lower line implies that $\nabla h(\overline{x})$ is surjective, i.e., $\operatorname{rank} \nabla h(\overline{x}) = p$. For the first line, let $\delta > 0$ be sufficiently small, and set $w \in \mathbb{R}^m$ with $w_i = -\delta$ if $g_i(\overline{x}) = 0$ and $w_i = 0$ otherwise. Then, by (RCQ) there is $d \in \mathbb{R}^m$, and $v \leq 0$ solving

$$\begin{pmatrix} w \\ 0 \end{pmatrix} = \begin{pmatrix} g(\overline{x}) + \nabla g(\overline{x})^T d - v \\ \nabla h(\overline{x})^T d \end{pmatrix}$$

hence $\nabla h(\overline{x})^T d = 0$ and for all i with $g_i(\overline{x}) = 0$ it is

$$\nabla g_i(\overline{x})^T d = w_i + v_i = -\delta + v_i \leq -\delta < 0$$

and hence MFCQ holds.

"$\Leftarrow$" Conversely, let MFCQ hold. Rescaling d from MFCQ, we can assert

$$\begin{aligned} \nabla h(\overline{x})^T d &= 0 \\ g_i(\overline{x}) + \nabla g_i(\overline{x})^T d &< -2\delta \end{aligned}$$

for all i and a sufficiently small $\delta > 0$. Now, we can pick $\rho > 0$ such that for any $d_0 \in B_\rho^{\mathbb{R}^n}(0)$ it is

$$g_i(\overline{x}) + \nabla g_i(\overline{x})^T (d + d_0) < -\delta.$$

Since $\nabla h(\overline{x})$ has rank p there is some $\varepsilon \in (0, \delta]$ such that

$$B_\varepsilon^{\mathbb{R}^p}(0) \subset \nabla h(\overline{x})^T B_\rho^{\mathbb{R}^n}(0).$$

Now, let $z_1 \in \mathbb{R}^m$ and $z_2 \in \mathbb{R}^p$ with $\|z_i\| < \varepsilon$ be arbitrary. Then there is $d_0 \in B_\rho^{\mathbb{R}^n}(0)$ such that

$$z_2 = \nabla h(\overline{x})^T d_0 = \nabla h(\overline{x})^T (d + d_0)$$

and

$$-(z_1)_i + [g_i(\overline{x}) + \nabla g_i(\overline{x})^T (d + d_0)] < \varepsilon - \delta \leq 0.$$

Hence we can set $v = -z_1 + g(\overline{x}) + \nabla g(\overline{x})^T (d + d_0) \in (-\infty, 0]^m$ and have

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} g(\overline{x}) + \nabla g(\overline{x})^T (d + d_0) - v \\ \nabla h(\overline{x})^T (d + d_0) \end{pmatrix}$$

and thus (RCQ) holds.

3.3. The Karush-Kuhn-Tucker conditions

Similar to the finite-dimensional NLP case, we would like to utilize the simpler structure of the linearizing cone $T_l(G, K, \overline{x})$ to derive optimality conditions of Karush-Kuhn-Tucker (KKT) type.

Before we can state the KKT-conditions, we need to generalize the sign conditions we are used to from NLPs.

Definition 3.3.1. Let $C \subset Z$ be a cone. Then we define the *polar cone* C° of C by

$$C^\circ = \{z^* \in Z^* \mid \langle z^*, z \rangle_{Z^*,Z} \leq 0 \ \forall z \in C\}.$$

The polar cone is always closed and convex (exercise!).

Theorem 3.3.2 (KKT-Conditions)**.** *Let X, Z be Banach-spaces and $K \subset Z$ be closed and convex. Further, let $\overline{x}$ be a local solution of (P) at which (RCQ) is satisfied (and of course where $f : X \to \mathbb{R}$ and $G : X \to Z$ are Fréchet-differentiable). Then there exists a Lagrange multiplier $\overline{\lambda} \in Z^*$, i.e.,*

$$\begin{aligned} f'(\overline{x}) + G'(\overline{x})^* \overline{\lambda} &= 0, \\ G(\overline{x}) &\in K, \\ \overline{\lambda} &\in T(K, G(\overline{x}))^\circ. \end{aligned}$$

Moreover, the set

$$\Lambda(\overline{x}) = \{\lambda \in T(K, G(\overline{x}))^\circ \mid f'(\overline{x}) + G'(\overline{x})^*\lambda = 0\}$$

is closed, convex, and bounded.

Before we come to the proof of the KKT-conditions, we need to collect a few preparatory results.

Clearly, (RCQ) is equivalent to the existence of $\delta > 0$ such that the ball $B_\delta^Z(0) \subset Z$ satisfies

$$B_\delta^Z(0) \subset G'(\overline{x})X - (K - G(\overline{x})).$$

Multiplication with $\lambda > 0$ implies that

$$Z = G'(\overline{x})X - \operatorname{cone}(K - G(\overline{x})) \tag{3.3}$$

where

$$\operatorname{cone}(K - G(\overline{x})) = \{d \in Z \mid d = \lambda(k - G(\overline{x})) \text{ for some } \lambda \geq 0, k \in K\}$$

is the *radial cone* of K at $G(\overline{x})$. Hence (RCQ) implies (3.3). In fact it can be shown, e.g., [Zowe and Kurcyusz, 1979, Theorem 2.1] or [Ito and Kunisch, 2008, Theorem 1.4], that both conditions are equivalent.

Due to Remark 3.2.5 this shows that (RCQ) also implies

$$Z = G'(\overline{x})X - T(K, G(\overline{x})). \tag{3.4}$$

We will later need to use this representation for 'small' $z \in Z$ and need to assert, that these small values can be reached by small values in $x \in X$ and $k \in T(K, G(\overline{x}))$. To this end, we require the following:

Theorem 3.3.3 (Generalized Open Mapping Theorem)**.** *Let X, Z be Banach-spaces and $\Psi : X \rightrightarrows Z$ be a closed convex multifunction, i.e., its graph*

$$\operatorname{graph}\Psi = \{(x, z) \in X \times Z \mid z \in \Psi(x)\}$$

is closed and convex. Then for any $z \in \operatorname{int}(\Psi(X))$, all $x \in \Psi^{-1}(z)$ and any $\delta > 0$ it is

$$z \in \operatorname{int}\Psi(B_\delta^X(x)).$$

For a proof, see [Bonnans and Shapiro, 2000, Theorem 2.70].

We can now apply this to the set valued mapping corresponding to (3.4), i.e., $\Psi : X \times Z \rightrightarrows Z$ with

$$\Psi(x, k) = \begin{cases} \{G'(\overline{x})x - k\} & k \in T(K, G(\overline{x})), \\ \emptyset & \text{otherwise.} \end{cases}$$

Indeed, this mapping is closed, since for $(x^n,k^n,z^n)\in \operatorname{graph}\Psi$ with $(x^n,k^n,z^n)\to(x,k,z)$ it holds $k^n\in T(K,G(\overline{x}))$ and $z^n=G'(\overline{x})x^n-k^n$. Thus, since $T(K,G(\overline{x}))$ is closed, $k\in T(K,G(\overline{x}))$ and further

$$k=\lim_{n\to\infty}k^n=\lim_{n\to\infty}\Big(G'(\overline{x})x^n-z^n\Big)=G'(\overline{x})x-z$$

and thus $(x,k,z)\in\operatorname{graph}\Psi$. Convexity of $\operatorname{graph}\Psi$, i.e., for any $(x_i,k_i)\in X\times Z$ $(i=1,2)$ and any $t\in[0,1]$ it is

$$t\Psi(x_1,k_1)+(1-t)\Psi(x_2,k_2)\subset\Psi(t(x_1,k_1)+(1-t)(x_2,k_2)),$$

follows immediately, since x,k appear linearly in Ψ whenever $\Psi(x,k)\neq\emptyset$, convexity of $T(K,G(\overline{x}))$, and $\emptyset+M=\emptyset$ for any set M.

Now, if (3.4) holds we have $\Psi(X\times Z)=Z$, and $\Psi(0,0)=0$. The generalized open mapping Theorem 3.3.3 yields $0\in\operatorname{int}\Psi(B_1^X(0)\times B_1^Z(0))$. More explicitly written, this is

$$0\in\operatorname{int}\Big(G'(\overline{x})B_1^X(0)-(T(K,G(\overline{x}))\cap B_1^Z(0))\Big). \tag{3.5}$$

It is also true, that (3.5) implies (3.4) which can be seen by multiplication with $t>0$.

Hence, we have

$$\text{(RCQ)}\quad\Leftrightarrow\quad\text{(3.3)}\quad\Rightarrow\quad\text{(3.4)}\quad\Leftrightarrow\quad\text{(3.5)}$$

Beweis. *[Of Theorem 3.3.2]* As usual, existence of Lagrange multipliers is based on a Hahn-Banach separation type argument, where the separating hyperplane provides the desired multiplier.

To this end, we define $M\subset\mathbb{R}\times Z$ as

$$M=\{(t,z)\in\mathbb{R}\times Z\,|\,\exists d\in X:\,t\geq f'(\overline{x})d,\ G'(\overline{x})d-z\in T(K,G(\overline{x}))\}.$$

draw sketch, and discuss expectation on M in 1d Since t,z,h appear only linearly, and $T(K,G(\overline{x}))$ is closed and convex this set M is convex and closed.

Further, we notice, that $(0,0)\in M$ is a boundary point of M. To see this, let $\tau>0$ be given; and we show $(-\tau,0)\notin M$. Assume for contradiction, that $(-\tau,0)\in M$. Then there would exist some $d\in X$ such that

$$f'(\overline{x})d\leq-\tau<0,\qquad\text{and}\qquad G'(\overline{x})d\in T(K,G(\overline{x})).$$

The latter means $d\in T_l(G,K,\overline{x})$ and the necessary optimality condition in Theorem 3.2.3 and Theorem 3.2.12 give

$$f'(\overline{x})d\geq0$$

contradicting our assumption.

Next, we need to see that $\operatorname{int}M\neq\emptyset$ so that we can utilize Hahn-Banach's theorem to separate $\operatorname{int}M$ from $(0,0)$. This can be done by (RCQ) and the implied condition (3.5).

Indeed (3.5) asserts that there is $\delta > 0$ such that for any $z \in B_\delta^Z(0)$ there exists $d \in B_1^X(0)$ and $k \in T(K, G(\overline{x})) \cap B_1^Z(0)$ such that

$$G'(\overline{x})d - z = k \in T(K, G(\overline{x})).$$

Moreover, for any $t \geq \|f'(\overline{x})\|_{X^*}$ it is by definition

$$f'(\overline{x})d \leq \|f'(\overline{x})\|_{X^*}\|d\|_X \leq \|f'(\overline{x})\|_{X^*} \leq t$$

and hence

$$[\|f'(\overline{x})\|_{X^*}, \infty) \times B_\delta^Z(0) \subset M.$$

Now, Hahn-Banach asserts the existence of $(\alpha, \lambda) \in (\mathbb{R} \times Z^*) \setminus \{(0,0)\}$ such that

$$\alpha t + \langle \lambda, z \rangle_{Z^*,Z} \geq 0 \quad \forall (t,z) \in M.$$

In more detail this means, for any $\tau > 0$, $d \in X$, $k \in T(K, G(\overline{x}))$ it holds

$$\alpha(f'(\overline{x})d + \tau) + \langle \lambda, G'(\overline{x})d - k \rangle_{Z^*,Z} \geq 0.$$

We now need to show that $\alpha > 0$.

We start by fixing $d = 0$, $k = 0$, $\tau \to \infty$ to get $\alpha \geq 0$. Now, to exclude the case $\alpha = 0$, assume for contradiction that $\alpha = 0$ we have

$$\langle \lambda, G'(\overline{x})d - k \rangle_{Z^*,Z} \geq 0 \quad \forall d \in X, k \in T(K, G(\overline{x}))$$

by (3.4) $G'(\overline{x})d - k$ can take any value in Z and thus $\lambda = 0$ contradicting $(\alpha, \lambda) \neq (0,0)$.

Now, we can rescale the multiplier, to get $(1, \overline{\lambda}) = (1, \frac{1}{\alpha}\lambda)$, to have

$$(f'(\overline{x})d + \tau) + \langle \overline{\lambda}, G'(\overline{x})d - k \rangle_{Z^*,Z} \geq 0 \qquad \forall \tau \geq 0, d \in X, k \in T(K, G(\overline{x})).$$

Taking $d = 0$, $\tau = 0$, and $k \in T(K, G(\overline{x}))$ arbitrary gives

$$\overline{\lambda} \in T(K, G(\overline{x}))^\circ.$$

Now, taking $k = 0$, $\tau = 0$ and $d \in X$ arbitrary shows

$$f'(\overline{x})d + \langle \overline{\lambda}, G'(\overline{x})d \rangle_{Z^*,Z} \geq 0 \qquad \forall d \in X$$

or equivalently, by definition of $G'(\overline{x})^*$

$$\langle f'(\overline{x}) + G'(\overline{x})^*\overline{\lambda}, d \rangle_{Z^*,Z} \geq 0 \qquad \forall d \in X.$$

Since $d \in X$ is arbitrary this shows

$$f'(\overline{x}) + G'(\overline{x})^*\overline{\lambda} = 0$$

and thus $\overline{\lambda}$ is the desired multiplier.

To see, the claim on $\Lambda(\overline{x})$, we note that closedness and convexity follow immediately from these properties of $T(K, G(\overline{x}))^\circ$ and the fact that λ appears only linearly in the confining equation.

To see boundedness, let $\lambda \in \Lambda(\overline{x})$ be given, picking $(1, \lambda) \in \mathbb{R} \times Z^*$ it holds for arbitrary $d \in X$, $k \in T(K, G(\overline{x}))$, and $\tau \geq 0$ that

$$\begin{aligned}(f'(\overline{x})d + \tau) + \langle \lambda, G'(\overline{x})d - k\rangle_{Z^*,Z} \\ = \langle f'(\overline{x}) + G'(\overline{x})^*\lambda, d\rangle_{Z^*,Z} + \tau - \langle \lambda, k\rangle_{Z^*,Z} \\ = \tau - \langle \lambda, k\rangle_{Z^*,Z} \geq 0\end{aligned}$$

by definition of $T(K, G(\overline{x}))^\circ$.

Hence the normal $(1, \lambda)$ supports M at $(0, 0)$. Further, we know that for some $\delta > 0$ we have $[\|f'(\overline{x})\|_{X^*}, \infty) \times B_\delta^Z(0) \subset M$. Now, choosing $z_\lambda \in B_\delta^Z(0)$ with $\langle \lambda, z_\lambda\rangle_{Z^*,Z} \leq -\frac{\delta}{2}\|\lambda\|_{Z^*}$ and $\tau = \|f'(\overline{x})\|_{X^*}$ we obtain

$$0 \leq 1 \cdot \|f'(\overline{x})\|_{X^*} + \langle \lambda, z_\lambda\rangle_{Z^*,Z} \leq \|f'(\overline{x})\|_{X^*} - \frac{\delta}{2}\|\lambda\|_{Z^*}$$

and thus

$$\|\lambda\|_{Z^*} \leq \frac{2}{\delta}\|f'(\overline{x})\|_{X^*}.$$

□

Remark 3.3.4. Utilizing the *Lagrange* function

$$\mathscr{L} : X \times Z^* \to \mathbb{R}; \quad \mathscr{L}(x, \lambda) = f(x) + \langle \lambda, G(x)\rangle_{Z^*,Z}$$

the KKT-conditions can be written in very short notation since

$$\mathscr{L}_x(x, \lambda) = f'(x) + G'(x)^*\lambda.$$

In fact, not only can we characterize solutions to

$$f'(\overline{x})d \geq 0 \qquad \forall\, d \in T_l(G, K, \overline{x}).$$

by the KKT-conditions provided (RCQ) holds in such a solution, but also the converse holds true.

Lemma 3.3.5. *If $\overline{x}$ is a KKT-point of (P), then*

$$f'(\overline{x})d \geq 0 \qquad \forall\, d \in T_l(G, K, \overline{x}).$$

Beweis. If $d \in T_l(G,K,\overline{x})$ then $G'(\overline{x})d \in T(K,G(\overline{x}))$ and consequently, for any $\lambda \in \Lambda(\overline{x}) \subset T(K,G(\overline{x}))^\circ$ it holds

$$f'(\overline{x})d = -\langle \lambda, G'(\overline{x})d\rangle_{Z^*,Z} \geq 0.$$

□

3.3.1. Cone Constraints

We will discuss the case, if we add the assumption, that the convex, closed set $K \subset Z$ is in addition a *cone*, i.e., for any $z \in K$ and $\lambda \geq 0$ it is

$$\lambda z \in K.$$

If this cone is, in addition, *salient* (im deutschen 'regulär' oder 'spitz'), i.e.,

$$z \in K \text{ and } -z \in K \quad \Leftrightarrow \quad z = 0,$$

then we can define a partial ordering by

$$z \leq_K 0 \quad \Leftrightarrow \quad z \in K,$$

or for any $z_1, z_2 \in Z$

$$z_1 \leq_K z_2 \quad \Leftrightarrow \quad z_1 - z_2 \in K$$

Hence, in this case the constraint $G(x) \in K$ can be seen as a generalized inequality constraint. In order to rewrite the KKT-conditions 3.3.2 in this case, we note the following.

Lemma 3.3.6. *If $K \subset Z$ is a closed, convex cone, then for any $\overline{z} \in K$ it is*

$$T(K,\overline{z})^\circ = K^\circ \cap \{\overline{z}\}^\perp$$

where the annihilator is given by

$$\{\overline{z}\}^\perp = \{\lambda \in Z^* \mid \langle \lambda, \overline{z}\rangle_{Z^*,Z} = 0\}.$$

Beweis. "$\supset$" Let $\lambda \in K^\circ \cap \{\overline{z}\}^\perp$ be given. Then for any $d \in T(K,\overline{z})$, with corresponding sequence $t_k \downarrow 0$, $z^k \in K$, it is

$$\langle \lambda, d\rangle_{Z^*,Z} \leftarrow \langle \lambda, \frac{(z^k - \overline{z})}{t_k}\rangle_{Z^*,Z} = \tfrac{1}{t_k}\langle \lambda, z^k\rangle_{Z^*,Z} - \tfrac{1}{t_k}\langle \lambda, \overline{z}\rangle_{Z^*,Z} = \tfrac{1}{t_k}\langle \lambda, z^k\rangle_{Z^*,Z} \leq 0.$$

Hence, $\lambda \in T(K,\overline{z})^\circ$.

"$\subset$" Conversely, let $\lambda \in T(K,\overline{z})^\circ$ be given.

We start by showing $\lambda \in K^\circ$. For any $z \in K$ it is

$$z + \overline{z} = 2\left(\tfrac{1}{2}z + \tfrac{1}{2}\overline{z}\right) \in K$$

since K is a convex cone. By Remark 3.2.5 it is thus

$$z = (z + \overline{z}) - \overline{z} \in T(K, \overline{z})$$

hence $\lambda \in K^\circ$.

Now, to see that $\lambda \in \{\overline{z}\}^\perp$, we note that analogously, $\overline{z} = 2\overline{z} - \overline{z} \in T(K, \overline{z})$ and $-\overline{z} = 0 - \overline{z} \in T(K, \overline{z})$. As a consequence

$$\langle \lambda, \pm\overline{z} \rangle_{Z^*,Z} \leq 0$$

and thus $\lambda \in \{\overline{z}\}^\perp$. □

We have thus learned, that for a closed, convex cone, the second part of the KKT-conditions

$$G(\overline{x}) \in K, \quad \overline{\lambda} \in T(K, G(\overline{x}))^\circ$$

can equivalently be written as a *complementarity condition*

$$G(\overline{x}) \in K, \quad \overline{\lambda} \in K^\circ, \quad \langle \lambda, G(\overline{x}) \rangle_{Z^*,Z} = 0.$$

3.3.2. Slater Condition

Definition 3.3.7. Let $K \subset Z$ be a closed, convex cone, and $G\colon X \to Z$ be convex with respect to, the partial order given by, K, i.e., for any $t \in [0,1]$ and all $x_1, x_2 \in X$ it holds

$$G((1-t)x_1 + tx_2) - (1-t)G(x_1) - tG(x_2) \in K.$$

We say that the *Slater*-condition holds if there exists some $\hat{x} \in X^{\mathrm{ad}}$ such that

$$G(\hat{x}) \in \operatorname{int}(K).$$

Theorem 3.3.8. *Let K be a closed, convex cone, and $G\colon X \to Z$ Fréchet-differentiable and convex w.r.t K. Further, assume that (P) satisfies the Slater-condition. Then any $\overline{x} \in X^{\mathrm{ad}}$ satisfies (RCQ).*

Beweis. Indeed, by convexity, of G, we have for any $t \in (0,1)$

$$K \ni \frac{G(\overline{x} + t(\hat{x} - \overline{x})) - (1-t)G(\overline{x}) - tG(\hat{x})}{t} = G(\overline{x}) + \frac{G(\overline{x} + t(\hat{x} - \overline{x})) - G(\overline{x})}{t} - G(\hat{x}).$$

Taking the limit $t \downarrow 0$ shows

$$G(\overline{x}) + G'(\overline{x})(\hat{x} - \overline{x}) \in K + G(\hat{x}).$$

Now, the Slater-condition asserts, that for some $\varepsilon > 0$ it is $G(\hat{x}) + B_\varepsilon^Z(0) \subset K$ and hence

$$G(\overline{x}) + G'(\overline{x})(\hat{x} - \overline{x}) + B_\varepsilon^Z(0) \subset K + G(\hat{x}) + B_\varepsilon^Z(0) \subset K + K = K,$$

where the last equality follows as in the previous Section 3.3.1 from the fact that K is a convex cone. We have thus shown, that the Slater-condition implies the linearized Slater condition (3.2). Thus we have shown the assertion by Lemma 3.2.14. □

Before we continue, we will discuss the KKT-conditions for two examples.

Example 3.3.9. First, we consider a control constrained optimal control problem:

$$\min\ J(q,u) \quad \text{s.t.} \begin{cases} E(q,u) = 0, \\ q \in Q^{\text{ad}} := \{p \in Q = L^2(\Omega) \mid a \le p \le b\}. \end{cases}$$

Here $E\colon Q \times U \to W$ and $J\colon Q \times U \to \mathbb{R}$ are continuously differentiable, U, W are Banach-spaces, Ω is a bounded domain, and $a \le b \in L^2(\Omega)$.

As we have seen in the exercises, the set Q^{ad} has no interior in $L^2(\Omega)$. Hence we have to work with the (RCQ) directly. To this end, let us assume that $(\overline{q}, \overline{u})$ is a local solution in which

$$E_u'(\overline{q}, \overline{u}) \text{ is surjective, i.e., } E_u'(\overline{q}, \overline{u})U = W.$$

Then, (RCQ) is satisfied. To see this, let $X = Q \times U$, $Z = W \times Q$, $K = \{0\} \times Q^{\text{ad}}$, $\overline{x} = (\overline{q}, \overline{u})$, and $G(q,u) = (E(q,u), q)^T \in Z$. Then we have

$$G(\overline{x}) + G'(\overline{x})X - K = \left\{ \begin{pmatrix} E_u'(\overline{q}, \overline{u})v + E_q'(\overline{q}, \overline{u})p \\ \overline{q} + p - r \end{pmatrix} \,\middle|\, v \in U, p \in Q, r \in Q^{\text{ad}} \right\}.$$

Now, for arbitrary $z = (w, q) \in Z$, let $p = q$, $r = \overline{q}$ and select $v \in U$ such that $E_u'(\overline{q}, \overline{u})v = w - E_q'(\overline{q}, \overline{u})p$, which is possible since E_u' is surjective. This shows, that

$$z \in G(\overline{x}) + G'(\overline{x})X - K.$$

Since $z \in Z$ was arbitrary, (RCQ) is shown.

The KKT-conditions 3.3.2 ensure the existence of $\overline{\lambda} = (\overline{\nu}, \overline{\mu}) \in W^* \times Q^* = W^* \times Q$ such that

$$E_q'(\overline{q}, \overline{u})^* \overline{\mu} = -J_q'(\overline{q}, \overline{u}), \quad E_u'(\overline{q}, \overline{u})^* \overline{\nu} = -J_u'(\overline{q}, \overline{u}), \quad E(\overline{q}, \overline{u}) = 0, \quad \overline{u} \in Q^{\text{ad}}, \quad \overline{\mu} \in T(Q^{\text{ad}}, \overline{q})^\circ.$$

Now, we would like to extract sign conditions from this optimality conditions. To this end, we need the pointwise representation

$$T(Q^{\mathrm{ad}},\overline{q}) = \{d \in L^2(\Omega) \,|\, d\big|_{\overline{q}=a} \geq 0, d\big|_{\overline{q}=b} \leq 0\}. \tag{3.6}$$

To see that this representation holds, we start by taking any d of the form on the right of (3.6). We can define

$$d^k = \frac{q^k - \overline{q}}{t_k}$$

with $t_k \downarrow 0$ and

$$q^k = \min(\max(\overline{q} + t_k d), a), b) \in Q^{\mathrm{ad}}.$$

Now, we note that almost all $x \in \Omega$ are simultaneous Lebesgue points of d, a, b, and $\overline{q}$, cf., Theorem A.6.5. Hence for any such x it is immediately clear that $d^k(x) \to d(x)$. Further, by Lipschitz continuity of min and max we get

$$|d^k(x)| \leq |d(x)|$$

and Lebesgue's dominated convergence theorem shows $d = \lim_{k\to\infty} d^k \in T(Q^{\mathrm{ad}},\overline{q})$.

The reverse inclusion is simple to see, i.e., if

$$T(Q^{\mathrm{ad}},\overline{q}) \ni d = \lim_{k\to\infty} d^k = \lim_{k\to\infty} \frac{q^k - \overline{q}}{t_k},$$

with $q^k \in Q^{\mathrm{ad}}$, then clearly for almost any $x \in \{x \in \Omega \,|\, \overline{q}(x) = a(x)\}$ it is $d^k(x) \geq 0$ and hence the same holds true for d. Analogously, the sign condition on $\{x \in \Omega \,|\, \overline{q}(x) = b(x)\}$ is obtained.

Now, (3.6) implies

$$T(Q^{\mathrm{ad}},\overline{q})^\circ = \{\mu \in L^2(\Omega) \,|\, \mu\big|_{\overline{q}=a} \leq 0, \mu\big|_{\overline{q}=b} \geq 0, \mu\big|_{a<\overline{q}<b} = 0\}.$$

This shows the expected sign conditions on the multiplier, i.e.,

$$\mu(x) \begin{cases} \leq 0 & \text{a.e. on } \{\overline{q} = a\}, \\ \geq 0 & \text{a.e. on } \{\overline{q} = b\}, \\ = 0 & \text{otherwise.} \end{cases}$$

Example 3.3.10. In the second example, we want to consider so called state constraints.

$$\min\ J(q,u) \qquad \text{s.t.}\ \begin{cases} Au = Bq + b, \\ u - \psi \in U^{\mathrm{ad}} := \{v \in C(\overline{\Omega}) \,|\, v \leq 0\}. \end{cases}$$

For this, we make the following assumptions

1. $\Omega \subset \mathbb{R}^n$, $n = 1,2,3$ is a bounded Lipschitz domain,
2. $b \in L^2(\Omega)$, $\psi \in C(\overline{\Omega})$, $B \in \mathscr{L}(Q, L^2(\Omega))$, Q a Banach space,
3. $U = H_0^1(\Omega) \cap H^2(\Omega) \subset C(\overline{\Omega})$ (embedding ok for $n = 1,2,3$.)
4. $A \in \mathscr{L}(H_0^1(\Omega), H^{-1}(\Omega))$ is a second order elliptic operator. In particular $A^{-1} \in \mathscr{L}(H^{-1}(\Omega), H_0^1(\Omega))$ exists.
5. A is H^2 regular, i.e., $A^{-1} \in \mathscr{L}(L^2(\Omega), U)$ (and $A \in \mathscr{L}(U, L^2(\Omega))$). This can be asserted, if the domain has a sufficiently smooth ($C^{1,1}$ or convex) boundary, and the coefficients in A are $W^{1,\infty}(\Omega) = C^{0,1}(\Omega)$.
6. $J : Q \times U \to \mathbb{R}$ is continuously differentiable.

In this situation, we can not use the same trick as in the control constrained case, i.e., pick something suitable for the inequality constrained component in the (RCQ) and then correct in the equality component since B is not surjective, in general. Hence, we have to work in a stronger topology which allows for interior points – thus we picked $U \subset C(\overline{\Omega})$.

Now, let $(\overline{q}, \overline{u})$ be a local solution in which the Slater-type condition

$$\text{there exists } \hat{q} \in Q, \hat{u} \in U \text{ satisfying } A\hat{u} = B\hat{q} + b,\ \hat{u} < \psi \text{ on } \overline{\Omega}$$

holds. To see that now (RCQ) holds, we set $X = Q \times U$, $Z = L^2(\Omega) \times C(\overline{\Omega})$, $K = \{0\} \times U^{\mathrm{ad}}$, and

$$G(x) = \begin{pmatrix} Au - Bq - b \\ Iu - \psi \end{pmatrix}.$$

Notice, that in contrast to the previous example, the inequality constraint is considered in a different topology than the natural one. Thus we add the embedding $I : Y \to C(\overline{\Omega})$.

To see (RCQ), i.e.,

$$0 \in \operatorname{int}\left\{ \begin{pmatrix} Av - Bp \\ (I\overline{u} - \psi) + Iv - w \end{pmatrix} \middle| v \in U, p \in Q, w \in U^{\mathrm{ad}} \right\}$$

we notice, that, by Weierstrass's extreme value theorem, there exists $\varepsilon > 0$ such that

$$I\hat{u} - \psi \leq -\varepsilon < 0.$$

By continuity of the solution map $A^{-1}\colon L^2(\Omega) \to U$, we have that for $\delta > 0$ small enough and any $p \in B_\delta^{L^2(\Omega)}(0)$

$$IA^{-1}(B(\hat{q}) + b + p) = I\hat{u} + IA^{-1}p \in B_{\varepsilon/2}^{C(\overline{\Omega})}(0) + I\hat{u}$$

since $\hat{u} = A^{-1}B(\hat{q} + b)$.

Now, let $z = (q,u) \in Z = L^2(\Omega) \times C(\overline{\Omega})$ with $\|q\|_Q \le \delta$ and $\|u\|_\infty \le \varepsilon/2$ be arbitrary. We can set $v = A^{-1}(B(\hat{q} - \overline{q}) + q)$ and $p = \hat{q} - \overline{q}$ asserting $Av - Bp = q$. Further, by definition of ε, δ, it holds

$$\begin{aligned}(I\bar{u} - \psi) + Iv &= I\bar{u} - \psi + IA^{-1}(B(\hat{q} - \overline{q}) + b - b + q)\\ &= I\bar{u} - \psi + IA^{-1}(B(\hat{q}) + b + q) - I\bar{u}\\ &= IA^{-1}q + I\hat{u} - \psi\\ &\le IA^{-1}q - \varepsilon\\ &\le \frac{\varepsilon}{2} - \varepsilon = -\frac{\varepsilon}{2} < 0\end{aligned}$$

and hence $(I\bar{u} - \psi) + Iv \in \operatorname{int} U^{\mathrm{ad}}$ showing that (RCQ) holds.

Hence in the selected minimizer the KKT-conditions 3.3.2 hold, i.e., there exists $\overline{\lambda} = (\overline{\nu}, \overline{\mu}) \in L^2(\Omega) \times C(\overline{\Omega})^*$ such that

$$\begin{aligned} A^*\overline{\nu} + I^*\overline{\mu} &= -J_u'(\overline{q}, \overline{u}),\\ J_q'(\overline{q}, \overline{u}) - B^*\overline{\nu} &= 0,\\ A\overline{u} &= B\overline{q} + b,\\ I\overline{u} - \psi &\le 0,\\ \overline{\mu} &\in T(U^{\mathrm{ad}}, I\overline{u} - \psi)^\circ = (U^{\mathrm{ad}})^\circ \cap (I\overline{u} - \psi)^\perp,\end{aligned}$$

where the last identity follows from Section 3.3.1.

To understand what I^* in the first, so called *adjoint*, equation means, we write for any $u \in U$

$$\langle I^*\overline{\mu}, u\rangle_{U^*,U} = \langle \overline{\mu}, Iu\rangle_{C^*,C} = \langle \overline{\mu}, u\rangle_{C^*,C}$$

Hence one often simply writes

$$A^*\overline{\nu} + \overline{\mu} = -J_u'(\overline{q}, \overline{u})$$

dropping the explicit notion of I and I^*.

To understand the complementarity condition

$$\overline{u} - \psi \le 0, \quad \overline{\mu} \in (U^{\mathrm{ad}})^\circ, \quad \langle \overline{\mu}, \overline{u} - \psi\rangle_{C^*,C} = 0,$$

we remark, that it can be shown, that $C(\overline{\Omega}) \simeq M(\overline{\Omega})$ is the space of real, regular Borel measures on $\overline{\Omega}$ with

$$\|\overline{\mu}\|_{M(\overline{\Omega})} = |\mu|(\overline{\Omega}) = \sup_{\|\nu\|_\infty \leq 1, \nu \in C(\overline{\Omega})} \int_{\overline{\Omega}} \nu \, \mathrm{d}\overline{\mu}.$$

With this the complementarity conditions can be written as

$$\overline{u} - \psi \leq 0, \quad \overline{\mu} \geq 0, \quad \overline{\mu}(\{\overline{u} - \psi < 0\}) = 0.$$

As a final remark, in many cases, $J : Q \times U \to \mathbb{R}$ is also differentiable as a mapping on a larger space $Q \times \widetilde{U} \supset Q \times U$. Then improved regularity of $\overline{\nu}$ can be shown, however this is limited by the appearance of $\overline{\mu}$ in the adjoint equation.

3.4. Sufficient optimality conditions

3.4.1. The convex case

Theorem 3.4.1. *Let X, Z be Banach-spaces, $f : X \to \mathbb{R}$ be convex and Fréchet-differentiable, and $G : X \to Z$ be Fréchet-differentiable and convex w.r.t the closed, convex cone $K \subset Z$. Then, any point $\overline{x} \in X^{\mathrm{ad}}$ satisfying the KKT-conditions is a global solution of* (P) .

Beweis. Exercise.

□

3.4.2. Second order sufficient conditions

As in finite dimensions, it is desirable to obtain sufficient conditions that are close to the necessary ones. It is thus necessary to consider growth conditions only on a very small set of critical directions. This is done by the following definition.

Definition 3.4.2. For $\overline{x} \in X^{\mathrm{ad}}$ we define the *critical cone* by

$$\begin{aligned} C(\overline{x}) &= \{d \in X \mid G'(\overline{x})d \in T(K, G(\overline{x})),\ f'(\overline{x})d \leq 0\} \\ &= \{d \in T_l(K, G, \overline{x}) \mid f'(\overline{x})d \leq 0\}. \end{aligned}$$

Clearly, if $\overline{x}$ is a stationary point satisfying (ACQ), then no directions d with $f'(\overline{x})d < 0$ exist, and the definition reduces to

$$C(\overline{x}) = \{d \in X \mid G'(\overline{x})d \in T(K, G(\overline{x})),\ f'(\overline{x})d = 0\}.$$

In finite dimensions, one would now formulate a sufficient condition by requiring

$$f'(\overline{x})d \geq 0 \quad \forall d \in T_l(K, G, \overline{x})$$

and

$$f''(\overline{x})(d,d) > 0 \quad \forall d \in C(\overline{x}) \setminus \{0\}$$

however, this is **not enough** in function spaces! To understand this, consider the following example.

Example 3.4.3. Let $X = Z = l_2$ the Hilbert-space of square summable sequences, i.e.,

$$\|x\|_{l_2}^2 = \sum_{i=0}^{\infty} x_i^2.$$

Consider $K = \{x \in l_2 \,|\, x_i \geq 0\}$, $G(x) = x$, and $f(x) = (c,x)_{l_2} - (x,x)_{l_2}$ where $c \in l_2$ with $c_i > 0$ for all i is fixed.

Then at $\overline{x} = 0$ it holds

$$T(K, G(\overline{x})) = T(K, 0) = K.$$

Further, for $d \in K \setminus \{0\}$ it is

$$f'(\overline{x})d = (c,d)_{l_2} = \sum_{i=0}^{\infty} c_i d_i > 0$$

since $c_i > 0$, $d_i \geq 0$, and $d_j > 0$ at least for one j.

Consequently, $C(\overline{x}) = \{0\}$, i.e., all directions $d \in T(K,0) \setminus \{0\}$ are ascent directions. But $\overline{x}$ is no local minimum of f on $X^{\mathrm{ad}} = K$. To see this, let $x^k = (2\delta_{ik}c_i)_{i\in\mathbb{N}}$. Then $\|x^k - \overline{x}\|_{l_2} = \|x^k\|_{l_2} = 2c_k \to 0$ as $k \to \infty$, and

$$f(x^k) = 2c_k^2 - 4c_k^2 = -2c_k^2 < 0 = f(\overline{x}).$$

Hence, the above sketched conditions can not be sufficient.

This shows, that the cone of critical directions is too small in infinite dimensions. Thus we consider the enlarged sets

Definition 3.4.4. For any $\eta \geq 0$ and $\overline{x} \in X^{\mathrm{ad}}$, we define the *approximate critical cone*

$$\begin{aligned} C_\eta(\overline{x}) &= \{d \in X \,|\, G'(\overline{x})d \in T(K, G(\overline{x})),\ f'(\overline{x})d \leq \eta\|d\|_X\} \\ &= \{d \in T_l(K, G, \overline{x}) \,|\, f'(\overline{x})d \leq \eta\|d\|_X\}. \end{aligned}$$

As in finite dimensions, we need to assert that the boundary of the feasible set is not too 'wild', and can reasonably be approximated by the linearizing cone. This will be asserted by the condition that for $X^{\mathrm{ad}} \ni x \to \overline{x}$ it is

$$\operatorname{dist}(x-\overline{x}, T_l(K,G,\overline{x})) = o(\|x-\overline{x}\|_X).$$

This is indeed the case, given that (RCQ) holds.

Lemma 3.4.5. *Let (RCQ) hold at $\overline{x} \in X^{\mathrm{ad}}$, then there exists a map $h\colon X^{\mathrm{ad}} \to T_l(K,G,\overline{x})$ such that*

$$\|h(x)-(x-\overline{x})\|_X = o(\|x-\overline{x}\|_X) \quad \text{for } X^{\mathrm{ad}} \ni x \to \overline{x}.$$

Beweis. For any $x \in X^{\mathrm{ad}}$ the Fréchet-differentiability of G asserts

$$G(x) = G(\overline{x}) + G'(\overline{x})(x-\overline{x}) + \omega(x)$$

with $\omega(x) = o(\|x-\overline{x}\|_X)$. By (3.5) there is $\delta > 0$ such that

$$\operatorname{cl} B_\delta^Z(0) \subset G'(\overline{x})B_1^X(0) - (T(K,G(\overline{x})) \cap B_1^Z(0)).$$

Hence, we can write

$$\delta \frac{\omega(x)}{\|\omega(x)\|_Z} = G'(\overline{x})\hat{s}(x) - \hat{k}(x)$$

with suitable $\hat{s}(x) \in B_1^X(0) \subset X$ and $\hat{k}(x) \in T(K,G(\overline{x})) \cap B_1^Z(0)$. Rescaling with $\frac{\|\omega(x)\|_Z}{\delta}$ asserts that

$$\omega(x) = G'(\overline{x})s(x) - k(x)$$

with $s(x) \in X$ and $k(x) \in T(K,G(\overline{x}))$ and

$$\|s(x)\|_X \le \frac{\|\omega(x)\|_Z}{\delta} = o(\|x-\overline{x}\|_X), \qquad \|k(x)\|_Z \le \frac{\|\omega(x)\|_Z}{\delta} = o(\|x-\overline{x}\|_X).$$

Now, we set $h(x) = x - \overline{x} + s(x)$, and obtain the desired approximation property

$$\|h(x)-(x-\overline{x})\|_X = \|s(x)\|_X = o(\|x-\overline{x}\|_X)$$

and the feasibility $G'(\overline{x})h(x) \in T(K,G(\overline{x}))$ from

$$\begin{aligned} G'(\overline{x})h(x) &= G'(\overline{x})(x-\overline{x}) + G'(\overline{x})s(x) \\ &= G'(\overline{x})(x-\overline{x}) + \omega(x) + k(x) \\ &= G(x) - G(\overline{x}) + k(x) \\ &\in K - G(\overline{x}) + T(K,G(\overline{x})) \\ &\subset T(K,G(\overline{x})). \end{aligned}$$

□

With this we can now show the sufficient conditions

Theorem 3.4.6 (Second order sufficient conditions). *Let X, Z be Banach-spaces, $K \subset Z$ closed and convex, $f : X \to \mathbb{R}$ and $G : X \to Z$ be twice Fréchet-differentiable. Further, let $\overline{x} \in X^{\mathrm{ad}}$ satisfy (RCQ) and KKT-conditions 3.3.2 with multiplier $\overline{\lambda}$, i.e.,*

$$\begin{aligned} f'(\overline{x}) + G'(\overline{x})^* \overline{\lambda} &= 0, \\ G(\overline{x}) &\in K, \\ \overline{\lambda} &\in T(K, G(\overline{x}))^\circ. \end{aligned}$$

If in addition, the second order condition

$$\mathscr{L}''_{xx}(\overline{x}, \overline{\lambda})(d,d) \geq \gamma \|d\|_X^2 \quad \forall\, d \in C_\eta(\overline{x})$$

holds for some $\gamma > 0$ and $\eta > 0$. Then $\overline{x}$ is an isolated local solution and there exists $\kappa, \delta > 0$ such that

$$f(x) - f(\overline{x}) \geq \kappa \|x - \overline{x}\|_X^2 \quad \forall\, x \in X^{\mathrm{ad}} \cap B_\delta^X(\overline{x}).$$

Beweis. Let $\delta > 0$ be sufficiently small, to be fixed later. For any $x \in X^{\mathrm{ad}} \cap B_\delta^X(\overline{x})$, set $d = x - \overline{x}$. By Lemma 3.4.5 we can write

$$d = h(x) + \omega(x), \qquad \omega(x) = o(\|d\|_X)$$

with $h(x) \in T_l(K, G, \overline{x})$. In particular, it holds

$$\|h(x)\|_X = \|d\|_X + o(\|d\|_X).$$

We start assuming, $h(x) \in C_\eta(\overline{x})$. Since $G(x) - G(\overline{x}) \in T(K, G(\overline{x}))$ it is $\langle \lambda, G(x) - G(\overline{x}) \rangle_{Z^*,Z} \leq 0$ and thus

$$\begin{aligned} f(x) - f(\overline{x}) &\geq \mathscr{L}(x, \overline{\lambda}) - \mathscr{L}(\overline{x}, \overline{\lambda}) \\ &= \mathscr{L}'_x(\overline{x}, \overline{\lambda})d + \frac{1}{2}\mathscr{L}''_{xx}(\overline{x}, \overline{\lambda})(d,d) + o(\|d\|_X^2) \\ &= \frac{1}{2}\mathscr{L}''_{xx}(\overline{x}, \overline{\lambda})(h(x) + \omega(x), h(x) + \omega(x)) + o(\|d\|_X^2) \\ &\geq \frac{1}{2}\mathscr{L}''_{xx}(\overline{x}, \overline{\lambda})(h(x), h(x)) - c\|h(x)\|_X \|\omega(x)\|_X - c\|\omega(x)\|_X^2 + o(\|d\|_X^2) \\ &\geq \frac{\gamma}{2}\|h(x)\|_X^2 - c\|h(x)\|_X \|\omega(x)\|_X + o(\|d\|_X^2) \\ &\geq \frac{\gamma}{2}\|h(x)\|_X^2 - \frac{\gamma}{4}\|h(x)\|^2 - 2c^2\|\omega(x)\|_X^2 + o(\|d\|_X^2) \\ &\geq \frac{\gamma}{4}\|h(x)\|^2 + o(\|d\|_X^2) \\ &\geq \frac{\gamma}{4}\|d\|^2 + o(\|d\|_X^2) \\ &\geq \frac{\gamma}{8}\|d\|^2 \end{aligned}$$

once δ is small enough to assert $|o(\|d\|_X^2)| \leq \frac{\gamma}{8}\|d\|_X^2$.

Otherwise, $h(x) \notin C_\eta(\overline{x})$, and thus

$$\begin{aligned} f(x) - f(\overline{x}) &= f'(\overline{x})d + o(\|d\|_X) \\ &= f'(\overline{x})h(x) + f'(\overline{x})\omega(x) + o(\|d\|_X) \\ &\geq \eta\|h(x)\|_X + o(\|d\|_X) \\ &= \eta\|d\|_X + o(\|d\|_X) \\ &\geq \eta\|d\|_X^2 \end{aligned}$$

once δ is sufficiently small.

□

4. Adjoint Approach

Many optimization problems are of the form $X = Q \times U$ with

$$\min_{(q,u)} J(q,u) \quad \text{s.t.} \quad E(q,u) = 0,\ (q,u) \in X^{\text{ad}} \tag{4.1}$$

where $J : Q \times U \to \mathbb{R}$ is the cost functional, $E : Q \times U \to Z$ is the so called state equation, and X^{ad} collects additional constraints on *control* q and *state* u. Assuming, that for each $q \in Q$ the state equation $E(q,u) = 0$ admits a unique solution $u = u(q)$, with the thus defined *solution operator* $q \mapsto u(q)$, one sees, that (4.1) is equivalent to

$$\min_{q} j(q) := J(q,u(q)) \quad \text{s.t.} \quad q \in Q^{\text{ad}} := \{q \in Q \,|\, (q,u(q)) \in X^{\text{ad}}\}. \tag{4.2}$$

When deriving optimality conditions for this problem, one needs to calculate, e.g., the derivative j'. By the chain rule, this can be calculated as

$$j'(q) = J'_q(q,u(q)) + J'_u(q,u(q)) \circ u'(q).$$

However, u' is only implicitly defined, and moreover for computations it may be inconvenient to evaluate this formula for multiple directions. Hence we will discuss how to obtain an alternative so called *adjoint* formula for the derivative which is very efficient in computations.

4.1. Adjoint Approach for the First Derivative

To derive a formula for the involved derivatives, we define the Lagrangian $\mathscr{L} : Q \times U \times Z^* \to \mathbb{R}$ by

$$\mathscr{L}(q,u,z) = J(q,u) - \langle E(q,u), z \rangle.$$

By definition of $u(q)$, it holds

$$j(q) = \mathscr{L}(q,u(q),z)$$

We now let $z = z(q)$ depend on q and calculate

$$\begin{aligned} j'(q) &= \mathscr{L}'_q(q,u(q),z(q)) + \mathscr{L}'_u(q,u(q),z(q)) \circ u'(q) + \mathscr{L}'_z(q,u(q),z(q)) \circ z'(q) \\ &= \mathscr{L}'_q(q,u(q),z(q)) + \mathscr{L}'_u(q,u(q),z(q)) \circ u'(q) \end{aligned}$$

since due to the choice $u = u(q)$ it holds $\mathscr{L}'_z(q,u(q),z(q)) = E(q,u) = 0$. Following this idea, we define $z = z(q)$ to solve

$$\mathscr{L}'_u(q,u(q),z(q)) = 0$$

the so called *adjoint equation*. In more detail, $z = z(q)$ is the solution of the equation

$$E'_u(q,u(q))^* z = J'_u(q,u(q)).$$

We have thus shown, that for this choice of $z(q)$ it is

$$j'(q) = \mathscr{L}'_q(q,u(q),z(q)) = J'_q(q,u(q)) - E'_q(q,u(q))^* z(q).$$

Remark 4.1.1. We notice, that the first order optimality conditions for (4.1) with $X^{\mathrm{ad}} = X$ (given (RCQ)) are (notice that the sign of z can be chosen arbitrary!)

$$\begin{aligned} \mathscr{L}'_u(q,u,z) &= J'_u(q,u) - E'_u(q,u)^* z = 0, \\ \mathscr{L}'_q(q,u,z) &= J'_q(q,u) - E'_q(q,u)^* z = 0, \\ E(q,u) &= 0. \end{aligned}$$

This is exactly the optimality condition

$$j'(q) = 0$$

for the reduced problem (4.2) noting that unique solvability of the adjoint equation implies (RCQ), compare Example 3.3.9.

In summary, to obtain the derivative $j'(q)$ one has to proceed as follows

1. Solve the state equation
$$E(q,u) = 0$$
to get $u = u(q) \in U$.
2. Solve the adjoint equation
$$E'_u(q,u)^* z = J'_u(q,u)$$
to get $z = z(q) \in Z^*$.
3. Evaluate
$$j'(q) = J'_q(q,u) - E'_q(q,u)^* z.$$

Remark 4.1.2. Notice, that the sign of the multiplier z is arbitrary, i.e., we could also have used the Lagrangian
$$J(q,u) + \langle E(q,u), z \rangle.$$
to obtain z with the opposite sign.

Example 4.1.3. To clarify this, let us consider the example

$$\min J(q,u) = \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2$$
$$\text{s.t.} \begin{cases} (\nabla u, \nabla\varphi) = (q,\varphi) & \forall \varphi \in H_0^1(\Omega) \\ u \in H_0^1(\Omega) & q \in L^2(\Omega) \end{cases}$$

on a domain $\Omega \subset \mathbb{R}^n$ and $\alpha > 0$. It follows from Riesz-representation theorem A.3.7 (or Lax-Milgram A.3.8), that the equation admits for any given $q \in L^2(\Omega)$ a unique solution $u = u(q) \in H_0^1(\Omega) \subset L^2(\Omega)$. This defines the control-to-state map $S\colon L^2(\Omega) \to L^2(\Omega)$, and the reduced problem

$$\min_{q\in L^2(\Omega)} j(q) := J(q,Sq).$$

Then, by direct calculation, we get the derivative in any direction $\delta q \in L^2(\Omega)$ as

$$j'(q)\delta q = (Sq-u^d, S\delta q) + \alpha(q,\delta q).$$

Hence to calculate the L^2-gradient $\nabla_{L^2} j(q)$ we would need to compute for a Schauder-basis δq_i, $i \in \mathbb{N}$ the coefficients

$$(\nabla_{L^2} j(q), \delta q_i) = (Sq-u^d, S\delta q_i) + \alpha(q,\delta q_i).$$

where for each index i an application of the operator S (one PDE solve) is needed. With a simple trick this can be avoided, noting that

$$(\nabla_{L^2} j(q), \delta q_i) = (S^*(Sq-u^d), \delta q_i) + \alpha(q,\delta q_i).$$

can easily be solved to get $\nabla_{L^2} j(q) = S^*(Sq-u^d) + \alpha q$ only needing two PDE solves, one for S and one for S^*. Indeed in the above formalism, we have the state equation

$$u = Sq \qquad \text{solving} \qquad 0 = \mathscr{L}'_z(q,u,z) = -(\nabla u, \nabla\cdot) + (q,\cdot) \in H_0^1(\Omega)^*,$$

the adjoint equation

$$z = S^*(u-u^d) \qquad \text{solving} \qquad 0 = \mathscr{L}'_u(q,u,z) = -(\nabla\cdot, \nabla z) + (u-u^d,\cdot) \in H_0^1(\Omega)^*,$$

and the gradient representation

$$\nabla_{L^2} j(q) = \nabla_{q,L^2} \mathscr{L}(q,u,z) = z + \alpha q.$$

4.2. Adjoint Formulas for the Second Derivatives

Similarly, adjoint formulas for the second derivatives can be obtained (assuming that all involved mappings are twice differentiable). In this case one is not interested in obtaining a formula for the Hessian $j''(q) \in \mathscr{L}(Q, \mathscr{L}(Q;\mathbb{R}))$ but rather in the product 'Hessian times direction' $j''(q)\delta q \in \mathscr{L}(Q;\mathbb{R})$ which is needed, e.g., in the iterative solution of the 'Newton equation' $j''(q)\delta q = -j'(q)$.

To this end, we look at the directional derivatives. We obtain, for two directions δq, τq and the abbreviations $\delta u = u'(q)\delta q$, $\tau u = u'(q)\tau q$, $\delta\tau u = u''(q)(\tau q, \delta q)$ and the same for $\delta z = z'(q)\delta q$, $\tau z = z'(q)\tau q$, and $\delta\tau z = z''(q)(\tau q, \delta q)$, the following representation for the hessian

$$\begin{aligned} j''(q)(\delta q, \tau q) = {} & \mathscr{L}''_{qq}(x)(\delta q, \tau q) + \mathscr{L}''_{qu}(x)(\delta q, \tau u) + \mathscr{L}''_{qz}(x)(\delta q, \tau z) \\ & + \mathscr{L}''_{uq}(x)(\delta u, \tau q) + \mathscr{L}''_{uu}(x)(\delta u, \tau u) + \mathscr{L}''_{uz}(x)(\delta u, \tau z) \\ & + \mathscr{L}''_{zq}(x)(\delta z, \tau q) + \mathscr{L}''_{zu}(x)(\delta z, \tau u) + \mathscr{L}''_{zz}(x)(\delta z, \tau z) \\ & + \mathscr{L}'_{u}(x)(\delta\tau u) + \mathscr{L}'_{z}(x)(\delta\tau z). \end{aligned}$$

By definition of $u(q)$ and $z(q)$ the last two terms vanish. Further the term $\mathscr{L}''_{zz} = 0$ since $\mathscr{L}$ is linear in z. Now, we can group these terms in order to determine a *tangent equation* to define $\delta u \in U$ such that

$$\mathscr{L}''_{qz}(x)(\delta q, \varphi) + \mathscr{L}''_{uz}(x)(\delta u, \varphi) = 0 \quad \forall \varphi \in U.$$

as well as a *dual-for-hessian* problem to determine $\delta z \in U$ such that

$$\mathscr{L}''_{qu}(x)(\delta q, \varphi) + \mathscr{L}''_{uu}(x)(\delta u, \varphi) + \mathscr{L}''_{zu}(x)(\delta z, \varphi) = 0 \quad \forall \varphi \in U.$$

In operator notation, we can calculate the directional derivative $j''(q)\delta q$ of $j'(q)$ as follows

1. Solve the state equation
$$E(q,u) = 0$$
to get $u = u(q) \in U$.

2. Solve the adjoint equation
$$E'_u(q,u)^* z = J'_u(q,u)$$
to get $z = z(q) \in Z^*$.

3. Solve the tangent equation
$$E'_u(q,u)\delta u = -E'_q(q,u)\delta q$$
to get $\delta u \in U$.

4. Solve the auxiliary adjoint (dual-for-hessian) equation

$$E'_u(q,u)^*\delta z = \mathscr{L}''_{uu}(q,u,z)\delta u + \mathscr{L}_{qu}(q,u,z)\delta q$$

   to get $\delta z \in Z^*$.

5. Evaluate

$$j''(q)\delta q = \mathscr{L}''_{qq}(q,u,z)\delta q + \mathscr{L}''_{uq}(q,u,z)\delta u - E'_q(q,u)^*\delta z.$$

5. Methods

5.1. Unconstrained Case

We start our discussion, with the unconstrained problem

$$\min_{x \in X} f(x) \tag{5.1}$$

on a Hilbert spaces X with continuously differentiable $f : X \to \mathbb{R}$.

We will start with a few definitions

Definition 5.1.1. We say that a method converges globally, if for any initial value x^0 the sequence x^k converges to a stationary point $\overline{x}$ of (5.1), i.e., a point such that $f'(\overline{x}) = 0$.

Definition 5.1.2. We say that the convergence $x^k \to \overline{x}$ is *linear* with rate c if

$$\|x^{k+1} - \overline{x}\|_X \leq c\|x^k - \overline{x}\|_X.$$

It converges *super-linear* if

$$\|x^{k+1} - \overline{x}\|_X \leq c_k\|x^k - \overline{x}\|_X.$$

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

$$\|x^{k+1} - \overline{x}\|_X \leq c\|x^k - \overline{x}\|_X^\alpha.$$

There is now a variety of algorithm available, we will only consider the simplest case – the generalization of the well-known steepest descent method.

5.1.1. Gradient descent

As a first (and most simple) choice for the direction one can consider the method of steepest descent. To this end we recall that we can define the gradient of f at x^k, using Riesz representation Theorem A.3.7, as follows. Find $\nabla f(x^k) \in X$ that satisfies

$$(\nabla f(x^k), d)_X = f'(x^k)d \quad \forall\, d \in X. \tag{5.2}$$

It is then immediately clear that $d^k = -\nabla f(x^k)$ is a descent direction

$$f'(x^k)d^k = (\nabla f(x^k), d^k)_X = -\|\nabla f(x^k)\|_X^2 \leq 0.$$

Further if $f'(x^k)d^k = 0$ then $\nabla f(x^k) = 0$ and thus the necessary conditions for a minimizer are fulfilled.

As in the finite dimensional case, a suitable step-length selection is needed, and we choose Armijo-Backtracking, i.e., for $\beta \in (0,1)$ and $\gamma \in (0, 1/2)$, we let $l \in \mathbb{N}_0$ be minimal such that for $t_k := \beta^l$ the inequality

$$f(x^k + t_k d^k) \leq f(x^k) + t_k \gamma f'(x^k)(d^k) \qquad (\mathscr{A})$$

is satisfied. Indeed such an index l, and thus t_k exists for any continuously differentiable f (proof as in finite dimensions).

With this, we obtain the gradient descent Algorithm 5.1.

Algorithm 5.1 (Gradient descend). Choose $x^0 \in X$ and let $k = 0$.
Calculate $d_0 = -\nabla f(x^0)$
while $\nabla f(x^k) \neq 0$ **do**
Determine t_k via $(\mathscr{A})$ with Backtracking.
$x^{k+1} = x^k + t_k d^k$
$d^{k+1} = -\nabla f(x^{k+1})$
$k \leftarrow k + 1$
end while

Remark 5.1.3. 1. Clearly, in practical computations, the Algorithm 5.1 needs to be run with a more suitable stopping criterion. Moreover, as an evaluation of ∇f is typically impossible inexact variants of the algorithm need to be considered.

We can now show the global convergence of the gradient descent Algorithm 5.1 similar to the finite dimensional case.

Theorem 5.1.4. *For any initial value x^0 Algorithm 5.1 either terminates finitely with $\nabla f(x^k) = 0$ or it creates an infinite sequence x^k such that $f(x^{k+1}) < f(x^k)$ and each accumulation point $\overline{x}$ is a stationary point of f, i.e., $\nabla f(\overline{x}) = 0$.*

Beweis. If the algorithm terminates after finitely many iterations there is nothing to show.

Otherwise, we have an infinite sequence $x^k \in X$. Let $(x^k)_{k\in\mathscr{K}}$ be a convergent subsequence with limit $\overline{x}$. By construction $f(x^k)$ is monotone decreasing, see, $(\mathscr{A})$ together with the definition (5.2) and $d^k = -\nabla f(x^k)$. We obtain

$$\lim_{\mathscr{K}\ni k\to\infty} f(x^k) = \lim_{k\to\infty} f(x^k) = f(\overline{x}).$$

Utilizing telescope sums, we obtain from the Armijo condition (𝒜)

$$\begin{aligned}
-\infty < f(\overline{x}) - f(x^0) &= \lim_{k\to\infty} f(x^k) - f(x^0) \\
&= \sum_{k=0}^{\infty} (f(x^{k+1}) - f(x^k)) \\
&= \sum_{k=0}^{\infty} (f(x^k + t_k d^k) - f(x^k)) \\
&\leq \sum_{k=0}^{\infty} \gamma t_k f'(x^k)(d^k) \\
&= -\gamma \sum_{k=0}^{\infty} t_k \|\nabla f(x^k)\|_X^2 \\
&\leq 0
\end{aligned}$$

and consequently

$$t_k \|\nabla f(x^k)\|_X^2 \to 0.$$

Now, if $\nabla f(x^k) \to 0$ on $k \in \mathcal{K}$ nothing is to show. Otherwise, possibly taking again a subsequence, $\lim_{\mathcal{K}\ni k\to\infty} t_k = 0$, w.l.o.g $t_k < 1$ for all $k \in \mathcal{K}$. Since l is minimal such that $t_k = \beta^l$ satisfies (𝒜), we know

$$f(x^k + \beta^{-1} t_k x^k) - f(x^k) > \gamma \beta^{-1} t_k f'(x^k)(d^k)$$

and since $\beta^{-1} t_k \to 0$ as well, we conclude by the one-dimensional mean-value theorem the existence of $\theta_k \in (0,1)$ such that

$$\begin{aligned}
-\|\nabla f(\overline{x}))\|_X^2 &= \lim_{\mathcal{K}\ni k\to\infty} -\|\nabla f(x^k)\|_X^2 \\
&< -\gamma \lim_{\mathcal{K}\ni k\to\infty} \|\nabla f(x^k)\|_X^2 \\
&= \gamma \lim_{\mathcal{K}\ni k\to\infty} f'(x^k)(d^k) \\
&\leq \lim_{\mathcal{K}\ni k\to\infty} \frac{f(x^k + \beta^{-1} t_k d^k) - f(x^k)}{\beta^{-1} t_k} \\
&= \lim_{\mathcal{K}\ni k\to\infty} f'(x^k + \theta_k \beta^{-1} t_k d^k)(d^k) \\
&= \lim_{\mathcal{K}\ni k\to\infty} -f'(x^k + \theta_k \beta^{-1} t_k d^k)(\nabla f(x^k)) \\
&= -f'(\overline{x})(\nabla f(\overline{x})) \\
&= -\|\nabla f(\overline{x})\|^2.
\end{aligned}$$

By the strict inequality, we get a contradiction, implying that $\|\nabla f(\overline{x})\| = 0$. □

Remark 5.1.5. 1. Notice, that although the theorem looks identical to the finite dimensional case it is much weaker in that it is less clear when convergent subsequences occur. This is due to the fact that for an infinite dimensional space bounded and closed sets are no longer compact. Hence although $\alpha\|x^k\|^2 \leq f(x^k) < f(x^0)$ implies boundedness of x^k this sequence does not necessarily contain a convergent subsequence.

2. However, strong convergence can be asserted if the derivative is monotone in the sense that there exists an $\varepsilon > 0$ such that

$$(\nabla f(x) - \nabla f(p), x - p)_X \geq \varepsilon \|x - p\|_X^2 \qquad \forall x, p \in X.$$

Such conditions are fulfilled, e.g., if $\widehat{S}$ is linear, i.e., $\widehat{S} = S' = S$, or near stationary points that satisfy the second order sufficient conditions $f''(\overline{x})(d,d) > \varepsilon\|d\|_X^2$ (different ε).

We will now see that, indeed, the last remark is true.

Theorem 5.1.6. *Let $f : X \to \mathbb{R}$ be continuously differentiable and assume that the set*

$$N := \{x \in X \mid f(x) < f(x^0)\}$$

is bounded and nonempty. Further let $\nabla f(x)$ be uniformly continuous on conv(N) *and monotone in the sense, that there exists $\varepsilon > 0$ such that*

$$(\nabla f(x) - \nabla f(p), x - p)_X \geq \varepsilon \|x - p\|_X^2$$

is true for all $x, p \in X$. Then, Algorithm 5.1 either terminates after finitely many iterations, or it generates a sequence x^k converging to the only stationary point $\overline{x} \in N \subset X$

Beweis. If the algorithm terminates after finitely many iterations there is nothing to show. Otherwise, we first show that there is at most one stationary point on N. To see this, let $\overline{x}_1, \overline{x}_2 \in X$ be stationary point, i.e., $\nabla f(\overline{x}_i) = 0$. Then by monotonicity it holds

$$0 = (\nabla f(\overline{x}_1) - \nabla f(\overline{x}_2), \overline{x}_1 - \overline{x}_2)_X \geq \varepsilon \|\overline{x}_1 - \overline{x}_2\|^2,$$

and thus $\overline{x}_1 = \overline{x}_2$.

Now, we need to show, that the sequence generated by the algorithm converges. We note, that since $\nabla f(x)$ is uniformly continuous on N there exists a constant C such that $\|\nabla f(x)\|_X \leq C$ on the bounded set conv(N); this is particularly true for the directions $d^k = -\nabla f(x^k)$. By the mean-value Theorem 3.1.6 f is bounded on conv(N). Now, analogously to the proof of

Theorem 5.1.4, we assert that the sequence of function values converges (to $\overline{f}$) and thus

$$-\infty < \overline{f} - f(x^0) = -\gamma \sum_{k=0}^{\infty} t_k \|\nabla f(x^k)\|_X^2 \leq 0.$$

and consequently $t_k\|\nabla f(x^k)\|_X^2 \to 0$. If we would know that x^k converges we would be done now, but we don't! Now, we show $\lim_{k\to\infty} \|d^k\|_X = 0$. To this end, assume for contradiction that this is not the case, i.e., there exists a subsequence $\mathscr{K}$ and $\varepsilon > 0$ such that $\|d^k\|_X \geq \varepsilon$ for all $k \in \mathscr{K}$. By the above convergence $t_k \to 0$ for $\mathscr{K} \ni k \to \infty$ and w.l.o.g $t_k < 1$. Now, for $k \in \mathscr{K}$, by our Armijo-Backtracking, we know that $\beta^{-1}t_k \leq 1$ does not satisfy ($\mathscr{A}$), i.e.,

$$f(x^k + \beta^{-1}t_k d^k) - f(x^k) > \beta^{-1}t_k \gamma f'(x^k)d^k$$

and thus by the mean-value theorem for the mapping $t \mapsto x^k + t_k d^k$ there exists a sequence $\theta_k \in (0,1)$ such that

$$\begin{aligned}
-\|d^k\|_X^2 &= f'(x^k)d^k \\
&< \gamma f'(x^k)d^k \\
&\leq \frac{f(x^k + \beta^{-1}t_k d^k) - f(x^k)}{\beta^{-1}t_k} \\
&= f'(x^k + \theta_k\beta^{-1}t_k d^k)d^k \\
&= f'(x^k)d^k + \Big(f'(x^k + \theta_k\beta^{-1}t_k d^k) - f'(x^k)\Big)d^k.
\end{aligned}$$

Since $\theta_k\beta^{-1}t_k d^k \to 0$ uniform-continuity and boundedness of d^k imply that the second summand converges to zero and we obtain

$$-\liminf_{\mathscr{K}\ni k} \|d^k\|_X^2 = \liminf_{\mathscr{K}\ni k} f'(x^k)d^k < \gamma \liminf_{\mathscr{K}\ni k} f'(x^k)d^k \leq \liminf_{\mathscr{K}\ni k} f'(x^k)d^k + 0 = -\liminf_{\mathscr{K}\ni k} \|d^k\|_X^2.$$

Since $\liminf_{\mathscr{K}\ni k} \|d^k\|_X \neq 0$ this is a contradiction and thus $\|d^k\|_X \to 0$.

Now, monotonicity gives for any k, l that

$$\varepsilon\|x^k - x^l\|_X^2 \leq (\nabla f(x^k) - \nabla f(x^l), x^k - x^l)_X \leq \|\nabla f(x^k) - \nabla f(x^l)\|_X \|x^k - x^l\|_X$$

and thus

$$\varepsilon\|x^k - x^l\|_X \leq 2\max(\|\nabla f(x^k)\|_X, \|\nabla f(x^l)\|_X) \to 0 \qquad (k,l \to \infty).$$

Consequently, the sequence x^k is converging to some limit $\overline{x}$ (with $\nabla f(\overline{x}) = 0$), and it is

$$f(\overline{x}) = \lim_{k\to\infty} f(x^k) \leq f(x^1) < f(x^0)$$

and thus $\overline{x} \in N$. □

Analogous to the finite dimensional case, the above algorithm is not very fast. To see this, we recall the following result:

Lemma 5.1.7. *Let $f(x) = (Hx,x)_X + (b,x)_X$, for a symmetric positive definite, and bounded operator $H \in \mathscr{L}(X,X)$ and $x \in X$. Assume that the spectrum $\sigma(H)$ satisfies $\sigma(H) \subset [\lambda,\Lambda]$ for some constants $0 < \lambda \leq \Lambda < \infty$. We define the condition number*

$$\kappa = \frac{\Lambda}{\lambda}.$$

Then the gradient descend method with exact line-search converges linearly in the norm induced by H with rate $\frac{(\kappa-1)}{(\kappa+1)}$, e.g.,

$$\|x^{k+1} - \overline{x}\|_H \leq \frac{(\kappa-1)}{(\kappa+1)} \|x^k - \overline{x}\|_H$$

where

$$\|x\|_H^2 = (Hx,x)_X.$$

Beweis. The proof in finite dimensions can be found in any introductory script to numerical analysis. For the infinite dimensional case, see, e.g., Daniel [1967]. □

5.1.2. Newton Methods

In a next step, we consider the second derivative of our functional; of course assuming that f is twice-continuously differentiable.

Analogously to the gradient, we can define a Hessian-operator $H = H(x) = \nabla^2 f(x) \in \mathscr{L}(X,X)$ by the relation

$$(Hd_1,d_2)_X := f''(x)[d_1,d_2].$$

Near a point $\overline{x}$ satisfying the second order sufficient condition

$$\nabla f(\overline{x}) = 0, \qquad (H(\overline{x})d,d)_X \geq \varepsilon \|d\|_X^2 \quad \forall d \in X$$

one can assume that H remains positive definite. This is analogous to the finite dimensional case, however the gap between the necessary condition

$$(H(\overline{x})d,d)_X \geq 0 \qquad \forall d \in X$$

is larger in the infinite-dimensional case.

As in the finite-dimensional case, the equation $f'(\overline{x}) = 0$ and Taylor-expansion

$$-\nabla f(x) = \nabla f(\overline{x}) - \nabla f(x) \approx \nabla^2 f(x)(\overline{x} - x)$$

motivates the local Newton-iteration

Algorithm 5.2 (Local Newton method). Choose $x^0 \in X$ and let $k = 0$.
while $\nabla f(x^k) \neq 0$ **do**
Calculate d_k satisfying $\nabla^2 f(x^k) d_k = -\nabla f(x^k)$
$x^{k+1} = x^k + d^k$
$k \leftarrow k + 1$
end while

Convergence of the local Newton-method

Assuming that the Functional f is twice-continuously differentiable, we obtain the usual local convergence theorem:

Theorem 5.1.8. *Let $f : X \to \mathbb{R}$ be twice continuously Fréchet-differentiable and $\overline{x} \in X$ be a local minimizer satisfying the second order sufficient condition*

$$f''(\overline{x})[d,d] \geq \mu \|d\|_X^2 \qquad \forall d \in X.$$

Then there exists $\delta > 0$ such that

1. *For any $x \in B_\delta(\overline{x})$ it is*

$$f''(x)[d,d] \geq \frac{\mu}{2} \|d\|_X^2 \qquad \forall d \in X.$$

2. *$\overline{x}$ is the only stationary point on $B_\delta(\overline{x}) := \{x \in X \mid \|x - \overline{x}\|_X < \delta\}$.*
3. *For any $x^0 \in B_\delta(\overline{x})$ the local Newton-method (with $\mathrm{TOL} = 0$) is well-defined and either terminates with $x^k = \overline{x}$ or creates a sequence that converges q-superlinearly towards $\overline{x}$.*

Beweis.

1. This is clear since $f'' \in \mathscr{L}(X; \mathscr{L}(X; \mathbb{R}))$ is continuous.
2. Assume that there is a second stationary point $\hat{x} \in B_\delta(\overline{x})$. Then with $d = \hat{x} - \overline{x}$ it holds

$$f'(\hat{x})d - f'(\overline{x})d = f''(\overline{x} + \xi d)[d,d]$$

for some $\xi \in (0,1)$ by the mean value theorem for the mapping $t \mapsto f'(\overline{x} + td)d$ (note that $\overline{x} + td \in B_\delta(\overline{x})$ by convexity of the norm). By the lower bound from 1. it is thus

$$0 = f'(\hat{x})d - f'(\overline{x})d \geq \frac{\mu}{2} \|d\|_X^2$$

and thus $\hat{x} = \overline{x}$.

3. Now, we need to see that the iteration is well-defined. To this end, let $x^k \in B_\delta(\overline{x})$ be given and let $H_k = H(x^k)$ be the Hessian-operator. Then the Newton-step $H_k d^k = -\nabla f(x^k)$ has a unique solution d^k. This follows directly from the fact, that due to 1. the functional
$$d \mapsto f_k(d) := \frac{1}{2}(H_k d, d)_X + (\nabla f(x^k), d)_X$$
is monotone in the sense of Theorem 5.1.6 (and strictly convex) and thus there is a unique minimizer d^k equivalently given by the Newton-equation.

   It remains to show that the sequence x^k remains in the ball $B_\delta(\overline{x})$ and converges superlinearly towards $\overline{x}$. To this end, we notice that
$$\begin{aligned}\frac{\mu}{2}\|x^{k+1}-\overline{x}\|_X^2 &\le f''(x^k)[x^{k+1}-\overline{x}, x^{k+1}-\overline{x}]\\ &= (H_k(x^{k+1}-\overline{x}), x^{k+1}-\overline{x})_X\\ &= (H_k(x^k+d^k-\overline{x}), x^{k+1}-\overline{x})_X\\ &= (H_k(x^k-\overline{x}), x^{k+1}-\overline{x})_X - (\nabla f(x^k), x^{k+1}-\overline{x})_X\\ &= (H_k(x^k-\overline{x}), x^{k+1}-\overline{x})_X + (\nabla f(\overline{x})-\nabla f(x^k), x^{k+1}-\overline{x})_X\\ &\le \|\nabla f(\overline{x})-\nabla f(x^k)+H_k(x^k-\overline{x})\|_X \|x^{k+1}-\overline{x}\|_X.\end{aligned}$$

   Consequently, noting that $H_k = f''(x^k)$, we get
$$\|x^{k+1}-\overline{x}\|_X \le \frac{2}{\mu}\sup_{x\in B_{\|x^k-\overline{x}\|_X}(\overline{x})}\|f''(x)-H_k\|_{\mathcal{L}(X;\mathcal{L}(X;\mathbb{R}))}\|x^k-\overline{x}\|_X$$
   by the mean-value Theorem 3.1.6 applied to the mapping $x \mapsto f'(x) - H_k x$, see also [Ciarlet, 2013, Theorem 7.2.2]. By continuity of f'' (and the definition of H_k) there exists some δ, such that
$$\sup_{x\in B_\delta(\overline{x})}\|f''(x)-H_k\|_{\mathcal{L}(X;\mathcal{L}(X;\mathbb{R}))} < \frac{\mu}{2}$$
   and thus linear-convergence of $x^k \to \overline{x}$ follows if $x^0 \in B_\delta(\overline{x})$ for sufficiently small δ. Superlinear convergence is immediate, since indeed, from linear convergence $\|x^k - \overline{x}\|_X \to 0$ it follows that
$$\sup_{x\in B_{\|x^k-\overline{x}\|_X}(\overline{x})}\|f''(x)-H_k\|_{\mathcal{L}(X;\mathcal{L}(X;\mathbb{R}))} \to 0.$$

□

Remark 5.1.9.

1. As in finite dimensions for Lipschitz-continuous second derivatives, quadratic convergence follows.

2. Indeed, as in finite dimensions, inexact solution of the linear-system are ok, as long as the inexactness tends to zero sufficiently fast. (Exercise!)

5.2. Constrained Case

In the constrained case, one can generalize first order methods such as the Gradient descent algorithm 5.1 to work for constrained problems. This can be done for instance by replacing the search directions $x^{k+1} = x^k + t_k d^k$ with the curve

$$x^{k+1} = P_{X^{\text{ad}}}(x^k + t_k d^k)$$

where $P_{X^{\text{ad}}}$ denotes a suitable projection onto the feasible set. However such methods will not be any faster than the Gradient descent algorithm 5.1 and moreover requires the evaluation of a projection in each step. Further, any such problem can be approximated by suitable penalty or barrier methods, see, e.g., Schiela and Wollner [2011], Schiela [2006] for barrier methods or Hintermüller et al. [2014], Ulbrich [2011] for penalty approaches.

Instead of these approaches, we will discuss a generalization of Newton's method which is capable to give at least superlinear convergence for many constrained optimization problems.

5.2.1. Reformulation of the KKT-conditions as a nonsmooth equation

Definition 5.2.1. Let $K \subset Z$ be a closed convex set. A function $\Psi : Z \times Z^* \to \mathbb{R}$ satisfying

$$\Psi(z,\lambda) = 0 \quad \Leftrightarrow \quad z \in K, \lambda \in T(K,z)^\circ$$

is called a *complementarity function*.

It is clear, that any such function Ψ would allow us to reformulate the KKT-conditions as an equation. That in fact such a complementarity function exists can be seen by the following

Lemma 5.2.2. *Let $Z \simeq Z^*$ be a Hilbert space, $K \subset Z$ be closed and convex. Then for any $\sigma > 0$ the following is equivalent for $z, \lambda \in Z$*

1. $z \in K$, $\lambda \in T(K,z)^\circ$
2. $z = P_K(z + \sigma\lambda)$

where $P_K : Z \to Z$ is the projection onto K given by

$$P_K(v) = \operatorname{argmin}_{k \in K} \|k - v\|_Z.$$

Beweis. By definition, 2. means

$$\min_{k\in K}\frac{1}{2}\|k-(z+\sigma\lambda)\|_Z^2 := \min_{k\in K}\varphi(k)=\varphi(z).$$

Now, $\varphi'(k)=k-(z+\sigma\lambda)$, and thus $\varphi'(z)=-\sigma\lambda$ and the first order necessary conditions from Theorem 3.2.3 show

$$z\in K,\qquad (-\sigma\lambda,d)\geq 0\quad \forall d\in T(K,z)$$

which is 1. Since φ is convex the conditions are also sufficient and hence 1. implies 2., too. □

We have thus seen that to find KKT-pairs $(\overline{x},\overline{\lambda})\in X\times Z$, where Z is a Hilbert space, we can equivalently search for solutions of the system

$$\begin{pmatrix} f'(x)+G'(x)^*\lambda \\ G(x)-P_K(G(x)+\sigma\lambda)\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

where $\sigma>0$ is arbitrary.

To this end, we would like to use a Newton method. Unfortunately, while we may hope that f,G are twice Fréchet-differentiable, and thus the first line has one derivative, differentiability of the second line can not be assumed in general. This is shown by the following example

Example 5.2.3. Consider $X=Z=L^2(\Omega)$, $G(x)=x$ and

$$K=\{x\in L^2(\Omega)\,|\,a\leq x\leq b\}$$

for some $a\leq b\in L^\infty(\Omega)$.

As we have seen in the exercises, for any $x\in K$ it holds

$$\begin{aligned} T(K,G(x))^\circ &= T(K,x)^\circ \\ &= \left\{\lambda\in L^2(\Omega)\,\middle|\,\lambda\begin{cases}\leq 0 & \text{on } \{x=a\},\\ \geq 0 & \text{on } \{x=b\},\\ =0 & \text{on } \{a<x<b\}.\end{cases}\right\}\end{aligned}$$

The projection onto the admissible set is given by

$$P_K(x)=\min(\max(a,x),b),$$

the pointwise projection onto the interval given by a, b. Now, since P_K is induced by a non-differentiable mapping $\mathbb{R}\to\mathbb{R}$ it is clear, that it is not differentiable.

Hence, we need to see whether we can generalize the Newton-method to less regular mappings.

5.2.2. Generalized Newton Methods

Definition 5.2.4. Let $F\colon \operatorname{dom}(F) \subset X \to Y$ be a given mapping between two Banach spaces. We call F *slant* differentiable at $x \in X$ if there exists an open neighborhood $\mathcal{N}(x)$ and a mapping $G\colon \mathcal{N}(x) \to \mathcal{L}(X,Y)$ such that

$$\lim_{\|d\|_X \downarrow 0} \frac{\|F(x+d) - F(x) - G(x+d)d\|_Y}{\|d\|_X} = 0.$$

Before we continue with our analysis we will state the first result which establishes the local fast convergence of the generalized Newton method.

Theorem 5.2.5. *Let $\overline{x}$ be a solution to $F(x) = 0$. Assume that F is slant differentiable with derivative G. If G is non singular and $\|G(x)^{-1}\|_{\mathcal{L}(Y,X)} \leq M$ on $\mathcal{N}(\overline{x})$ then the Newton iteration*

$$x_{k+1} = x_k - G(x_k)^{-1} F(x_k)$$

converges q-superlinearly to $\overline{x}$ if the initial value x_0 is close enough to $\overline{x}$.

Beweis. Let $r > 0$ be such that $B_r(\overline{x}) \subset \mathcal{N}(\overline{x})$ then by definition of slant differentiability it holds

$$\|F(x_k) - F(\overline{x}) - G(x_k)(x_k - \overline{x})\|_Y \leq c_r \|x_k - \overline{x}\|_X$$

for all $x_k \in B_r(\overline{x})$ and a constant $c_r \to 0$ as $r \to 0$.

We initialize the iteration such that $x_0 \in B_{r_0}(\overline{x})$ such that $c_{r_0} < 1/M$. Now, by definition of the iteration it holds

$$\begin{aligned}
\|x_{k+1} - \overline{x}\|_X &= \|x_k - \overline{x} - G(x_k)^{-1} F(x_k)\|_X \\
&\leq \|G(x_k)^{-1}\|_{\mathcal{L}(Y,X)} \|G(x_k)(x_k - \overline{x}) - F(x_k) + F(\overline{x})\|_Y \\
&\leq c_{\|x_k - \overline{x}\|_X} M \|x_k - \overline{x}\|_X.
\end{aligned}$$

By induction it follows that $c_{\|x_k - \overline{x}\|_X} M \leq c_{r_0} M < 1$. Thus we have

$$\|x_{k+1} - \overline{x}\|_X \leq M^{k+1} \prod_{l=0}^{k} c_{\|x_l - \overline{x}\|_X} \leq (M c_{r_0})^{k+1} \to 0.$$

The super-linear convergence then follows as $c_{\|x_k - \overline{x}\|_X} \to 0$ with $k \to \infty$. □

Before we continue to apply this to our model problem we will start with some examples

Example 5.2.6. 1. Any bounded linear operator A is slant differentiable with $G(x) = A$.

2. In a Hilbert space H the mapping $F(u) = \|u\|_H$ is slant differentiable (exercise).

3. The map $\max(0,\cdot)$ defined as

$$\max(0,\cdot)\colon L^r(\Omega) \to L^p(\Omega), \quad u \mapsto \max(0,u)$$

is slant differentiable for all $1 \le p < r \le \infty$. Its slant derivative $G(u) \in \mathscr{L}(L^r(\Omega), L^p(\Omega))$ is defined by pointwise multiplication with the function

$$g_u(x) = \begin{cases} 1 & u(x) > 0 \\ \delta & u(x) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\delta \in \mathbb{R}$ is arbitrary. This means

$$(G(u)\delta u)(x) = g_u(x)\delta u(x)$$

We will only show that this is true for the choice $\delta = 0$. Then, we can write this as

$$G(u)\delta u = \chi_{\mathscr{A}(u)}\delta u$$

where $\mathscr{A}(u) = \{x \in \Omega | u(x) > 0\}$. χ_M is called characteristic function of the set M and is defined as

$$\chi_M = \begin{cases} 1 & x \in M \\ 0 & x \notin M \end{cases}$$

To see slant differentiability let $u, \delta u \in L^r(\Omega)$ be given. Then we denote the remainder by

$$\omega_{u,\delta u}(x) = \max(0, u(x) + \delta u(x)) - \max(0, u(x)) - g_{u+\delta u}(x)\delta u(x).$$

A simple calculation shows

$$|\omega_{u,\delta u}(x)| \begin{cases} \le |u(x)| & (u(x) + \delta u(x))u(x) < 0, \\ \le |u(x)| & (u(x) + \delta u(x)) = 0,\ u(x) \ne 0 \\ = 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

From (5.3), we obtain that we only need to consider the set

$$\Omega_0^{\delta u} = \{x \in \Omega \,|\, u(x) \ne 0, u(x)(u(x) + \delta u(x)) \le 0\}.$$

Now, there are two 'subsets' to consider. One $\Omega_\varepsilon^{\delta u}$ where u is large, compared to ε, and thus a sign change can only occur if δu is also large; hence we expect this measure of this set to vanish as $\delta u \to 0$. And another set $\Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}$, where u is small. On this set we have trouble since it depends on both ε and δu, and thus we put it into a fixed set $\widetilde{\Omega}_\varepsilon^u \supset \Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}$ where we need to show that as $\varepsilon \to 0$ this set vanishes.

In more detail, we define the following subsets $\Omega_\varepsilon \subset \Omega_0$ as follows

$$\Omega_\varepsilon^{\delta u} = \{x \in \Omega \,|\, |u| \geq \varepsilon, u(x)(u(x) + \delta u(x)) \leq 0\}.$$

Now, $|u(x)| \geq \varepsilon$ on $\Omega_\varepsilon^{\delta u}$ and thus to get the second defining inequality we have in addition $|\delta u(x)| \geq \varepsilon$ on $\Omega_\varepsilon^{\delta u}$. Hence it holds ($r < \infty$)

$$\|\delta u\|_{L^r(\Omega)} \geq \|\delta u\|_{L^r(\Omega_\varepsilon^{\delta u})} \geq \varepsilon |\Omega_\varepsilon^{\delta u}|^{1/r}.$$

In particular, for any fixed $\varepsilon > 0$ it holds

$$\lim_{\|\delta u\|_{L^r(\Omega)} \to 0} |\Omega_\varepsilon^{\delta u}| = 0. \tag{5.4}$$

Which is true for $r = \infty$ as well.

We continue by defining an other set

$$\widetilde{\Omega}_\varepsilon^u = \{x \in \Omega \,|\, 0 < |u(x)| \leq \varepsilon\} \supset \Omega_0^{\delta u} \setminus \Omega_\varepsilon^{\delta u}.$$

For this it holds

$$\widetilde{\Omega}_\varepsilon^u \subset \widetilde{\Omega}_{\varepsilon'}^u, \quad 0 < \varepsilon \leq \varepsilon', \qquad \text{and} \qquad \bigcap_{\varepsilon > 0} \widetilde{\Omega}_\varepsilon^u = \emptyset.$$

This implies

$$\lim_{\varepsilon \to 0} |\widetilde{\Omega}_\varepsilon^u| = 0. \tag{5.5}$$

By Hölder's inequality, we get for any function $f \in L^r(\Omega)$, and $1 \leq p < r \leq \infty$, it holds

$$\|f\|_p \leq \|1\|_s \|f\|_r = |\Omega|^{1/s} \|f\|_r, \quad s = \begin{cases} \frac{pr}{r-p} & r < \infty \\ p & r = \infty. \end{cases}$$

Then, we use (5.3) to conclude

$$\begin{aligned}\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}} &\leq \frac{1}{\|\delta u\|_{L^r}}\|u(x)\|_{L^p(\Omega_0^{\delta u})}\\ &\leq \frac{1}{\|\delta u\|_{L^r}}\left\{\|u(x)\|_{L^p(\Omega_\varepsilon^{\delta u})}+\|u(x)\|_{L^p(\Omega_0^{\delta u}\setminus\Omega_\varepsilon^{\delta u})}\right\}\\ &\leq \frac{1}{\|\delta u\|_{L^r}}\left\{|\Omega_\varepsilon^{\delta u}|^{1/s}\|u(x)\|_{L^r(\Omega_\varepsilon^{\delta u})}+|\widetilde{\Omega}_\varepsilon^{u}|^{1/s}\|u(x)\|_{L^r(\Omega_0^{\delta u}\setminus\Omega_\varepsilon^{\delta u})}\right\}\\ &\leq \frac{c}{\|\delta u\|_{L^r}}\|u(x)\|_{L^r(\Omega_0^{\delta u})}\Big(|\Omega_\varepsilon^{\delta u}|^{1/s}+|\widetilde{\Omega}_\varepsilon^{u}|^{1/s}\Big).\end{aligned}$$

Now, we note that on $\Omega_0^{\delta u}$ it holds $|\delta u|\geq|u|$ and hence

$$\begin{aligned}\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}} &\leq \frac{c}{\|\delta u\|_{L^r}}\|\delta u\|_{L^r}\Big(|\Omega_\varepsilon^{\delta u}|^{1/s}+|\widetilde{\Omega}_\varepsilon^{u}|^{1/s}\Big)\\ &\leq c\Big(|\Omega_\varepsilon^{\delta u}|^{1/s}+|\widetilde{\Omega}_\varepsilon^{u}|^{1/s}\Big).\end{aligned}$$

Now, due to (5.5), the second summand can be made arbitrarily small independent of δu. Further, due to (5.4), the first summand goes to zero which shows the slant differentiability, e.g.

$$\frac{\|\omega_{u,\delta u}\|_{L^p}}{\|\delta u\|_{L^r}}\to 0,\quad (\|\delta u\|_{L^r}\to 0).$$

Remark 5.2.7. Finally, we like to note that the restriction to consider $\max(0,u)$ as a mapping from L^r into L^p with $p<r$ is not just for convenience, but in fact the function

$$g_u(x)\begin{cases}1 & u(x)\geq 0\\ 0 & \text{otherwise}\end{cases}$$

is not a slant derivative if $r=p$ (exercise).

Now, we would like to apply our knowledge to the KKT-conditions given at the beginning of this section. But before we can do so we first need the following chain rule

Theorem 5.2.8. *Let $f:\operatorname{dom}(f)\subset X\to Y$ be continuously Fréchet differentiable and $g:Y\to Z$ be slant differentiable at $f(x)$ with derivative G. Assume that $\|G\|_{\mathscr{L}(Y,Z)}\leq c$ in $\mathscr{N}(f(x))$. Then $F=g\circ f$ is slant differentiable in x and the slant derivative is given by $G(f(x+d))f'(x+d)\in\mathscr{L}(X,Z)$ for d sufficiently small.*

Beweis. □

5.2.3. Semi-smoothness

On careful inspection of Theorem 5.2.5, we notice that while a nice convergence property holds if a suitable derivative is chosen, the choice is not local at the moment since we require that G is a slant derivative at the solution $\overline{x}$. To allow for a reasonable selection of $G(x^k)$, we need to restrict the set of functions to allow for a selection of $G(x^k)$ based on local information, only.

This will lead to so called semi-smooth Newton methods, see, e.g., Ito and Kunisch [2008], Ulbrich [2011].

Definition 5.2.9. Let $F : X \to Y$ be a continuous operator between Banach spaces X, Y. And let $\partial F : X \rightrightarrows \mathscr{L}(X,Y)$ be a set-valued mapping with $\partial F(x) \neq \emptyset$ for all $x \in X$. Then F is called ∂F-semi-smooth at $x \in X$, if

$$\lim_{\|d\|_X \to 0} \sup_{G \in \partial F(x+d)} \frac{\|F(x+d) - F(x) - Gd\|_Y}{\|d\|_X} = 0.$$

Further, F is called ∂F-semi-smooth of order $\alpha > 0$ at $x \in X$, if

$$\limsup_{\|d\|_X \to 0} \sup_{G \in \partial F(x+d)} \frac{\|F(x+d) - F(x) - Gd\|_Y}{\|d\|_X^{1+\alpha}} < \infty.$$

Algorithm 5.3 (Local semi-smooth Newton method). Choose $x^0 \in X$, and let $k = 0$.
while $F(x^k) \neq 0$ **do**
  Choose $G_k \in \partial F(x^k)$
  Calculate d^k satisfying $G_k d^k = -F(x^k)$
  Set $x^{k+1} \leftarrow x^k + d^k$
  $k \leftarrow k+1$
end while

Corollary 5.2.10. *Let $\overline{x}$ be a solution to $F(x) = 0$. Assume that F is ∂F-semi-smooth, and there exists $\delta > 0$ and $M > 0$ such that for any $x \in B_\delta^X(\overline{x})$ any $G \in \partial F(x)$ is invertible and satisfies*

$$\|G^{-1}\|_{\mathscr{L}(Y,X)} \leq M.$$

Then the semi-smooth Newton method 5.3 converges superlinearly to $\overline{x}$ if the initial value x_0 is close enough to $\overline{x}$.

Moreover, if F is ∂F-semi-smooth of order $\alpha > 0$, then the convergence is of order $1 + \alpha$.

Beweis. The superlinear convergence is an immediate consequence of Theorem 5.2.5.

The case of smoothness order $\alpha > 0$ is left as an exercise. □

Similar to slant-differentiable operators, semi-smooth operators satisfy some useful calculus rules, cf., Ulbrich [2011].

Theorem 5.2.11. *Let X, Y, Z, X_i, and Y_i be Banach spaces. Then it holds:*

1. *If $F_i : X \to Y_i$, $i = 1, 2$, are ∂F_i-semi-smooth at x then $(F_1, F_2) : X \to Y_1 \times Y_2$ is $(\partial F_1, \partial F_2)$-semi-smooth at x.*
2. *If $F_i : X \to Y$, $i = 1, 2$, are ∂F_i-semi-smooth at x, then $F_1 + F_2$ is $(\partial F_1 + \partial F_2)$-semi-smooth at x.*
3. *If $F_1 : Y \to Z$ and $F_2 : X \to Y$ are ∂F_i-semi-smooth at $F_2(x)$ and x, respectively, ∂F_1 is bounded near $F_2(x)$, and F_2 is Lipschitz-continuous near x, then $F = F_1 \circ F_2$ is ∂F-semi-smooth at x with*

$$\partial F(x) = \{G_1 G_2 \,|\, G_1 \in \partial F_1(F_2(x)), G_2 \in \partial F_2(x)\}.$$

4. *If $F : X \to Y$ is continuously differentiable at x, then F is $\{F'\}$-semi-smooth at x, where $\{F'\}$ denotes the set-valued operator $\{F'\} : X \rightrightarrows \mathscr{L}(X, Y)$ with $\{F'\}(x) = \{F'(x)\}$.*

Beweis. Exercise. □

5.2.4. Semi-Smoothness in Finite Dimensions

We would like to discuss the application of the results towards typical optimization problems. To this end, we need to define the generalized differential. As we have seen in Example 5.2.3, it happens, that the Projection $P_K : Z \to Z$ is given as a Nemytskii operator of a locally Lipschitz-continuous mapping $\psi : \mathbb{R} \to \mathbb{R}$. It is thus useful to construct a generalized derivative of P_K based on a generalized derivative of ψ.

Definition 5.2.12. For a locally Lipschitz-continuous function $F : \mathbb{R}^n \to \mathbb{R}^m$ we define the Clarke generalized Jacobian as

$$\partial^{\mathrm{cl}} F(x) = \mathrm{conv}\{G \,|\, \text{there exists } x^k \to x,\ \mathrm{F} \text{ is differentiable at } x^k, F'(x^k) \to G\}.$$

Indeed, this definition is well-posed since by Rademacher's theorem any Lipschitz-continuous function is almost everywhere differentiable. This result can not be generalized to Fréchet-differentiability in infinite dimensions, but (for locally convex spaces) to Gâteaux derivatives, see, e.g., Bogachev and Mayer-Wolf [1996].

Clearly, if F is continuously differentiable in a point x then $\partial^{\mathrm{cl}} F(x) = \{F'(x)\}$.

Example 5.2.13. Let us consider the function $\psi : \mathbb{R} \to \mathbb{R}$ with $\psi(x) = P_{[a,b]}(x) = \min(\max(a, x), b)$ as given in Example 5.2.3. The Clarke generalized derivative is

$$\partial^{\mathrm{cl}}\psi(x) = \begin{cases} 0 & x \notin [a,b], \\ 1 & x \in (a,b), \\ \mathrm{conv}\{0,1\} = [0,1] & x \in \{a,b\}. \end{cases}$$

Indeed, ψ is then $\partial^{\mathrm{cl}}\psi$-semi-smooth.

5.2.5. Semi-Smoothness of Nemyzkii-Operators

A natural way, to construct a generalized derivative for a Nemyzkii-operator Ψ given by superposition with ψ is by superposition with the generalized derivative of ψ. This is analogous to the differentiable case, where our candidate for a derivative is given by the superposition of the pointwise derivatives.

Definition 5.2.14. Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz-continuous and $\partial^{\mathrm{cl}}\psi(x)$ semi-smooth. For any $1 \le r \le p \le \infty$ we define the superposition operator

$$\Psi : L^p(\Omega)^m \to L^r(\Omega), \quad \Psi(u)(x) = \psi(u(x)) \quad \forall u \in L^p(\Omega), x \in \Omega$$

and the generalized differential

$$\partial\Psi(u) = \{G \mid Gw = g \cdot w, g \in L^\infty(\Omega)^m, g(x) \in \partial^{\mathrm{cl}}\psi(u(x)) \text{ for all } w \in L^p(\Omega)^m \text{ for a.e. } x \in \Omega\}.$$

We note, that $g(x)$ is uniformly bounded by Lipschitz-continuity of ψ.

From Example 5.2.3 one would like to use $p = r = 2$ as then a generalized derivative of $P_K(G(x) + \sigma\lambda)$ could immediately be obtained by the chain-rule from Theorem 5.2.11-3.

As we have to expect from Remark 5.2.7 this is not possible (unless $p = r = \infty$).

Lemma 5.2.15. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be Lipschitz-continuous and **not** affine linear, and $\Omega \subset \mathbb{R}^n$*

a domain. Then, for any $r \in [1, \infty)$ the operator

$$\Psi \colon L^r(\Omega) \ni u \mapsto \psi(u(\cdot)) \in L^r(\Omega)$$

is not $\partial\Psi$-semi-smooth.

Beweis. Let $b \in \mathbb{R}$ be given and choose $g_b \in \partial^{\mathrm{cl}}\psi(b)$. Since ψ is not affine linear, there is $a \in \mathbb{R}$ such that

$$\psi(a) \neq \psi(b) + g_b(a-b)$$

and thus

$$\rho := |\psi(b) - \psi(a) - g_b(b-a)| > 0.$$

Now, let $x_0 \in \Omega$ and $U_\varepsilon = (x_0 - h_\varepsilon, x_0 + h_\varepsilon)^n$ with $h_\varepsilon = \varepsilon^{1/n}/2$. Define

$$u(x) = a, \quad x \in \Omega, \quad d_\varepsilon(x) = \begin{cases} b-a & x \in U_\varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$\|d_\varepsilon\|_{L^r}^r = \int_\Omega |d_\varepsilon(x)|^r \,\mathrm{d}x = \int_{U_\varepsilon} |b-a|^r \,\mathrm{d}x = \varepsilon |b-a|^r.$$

Now, let $g_a \in \partial\psi(a)$ be arbitrary, and define

$$g_\varepsilon(x) = \begin{cases} g_b & x \in U_\varepsilon, \\ g_a & \text{otherwise.} \end{cases}$$

By this construction $G \colon L^r(\Omega) \ni v \mapsto g_\varepsilon \cdot v \in L^r(\Omega)$ is an element of $\partial\Psi(u + d_\varepsilon)$. For arbitrary $x \in \Omega$, we have the remainder

$$|\psi(u(x)+d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x) d_\varepsilon(x)| = \begin{cases} |\psi(b) - \psi(a) - g_b(b-a)| = \rho > 0 & x \in U_\varepsilon, \\ |\psi(a) - \psi(a) - g_b 0| = 0 & \text{otherwise.} \end{cases}$$

We obtain

$$\begin{aligned} \|\Psi(u+d_\varepsilon) - \Psi(u) - G d_\varepsilon\|_{L^r}^r &= \int_\Omega |\psi(u(x)+d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x) d_\varepsilon(x)|^r \,\mathrm{d}x \\ &= \int_{U_\varepsilon} \rho^r \,\mathrm{d}x \\ &= \varepsilon \rho^r \\ &= \frac{\rho^r}{|b-a|^r} \|d_\varepsilon\|_{L^r}^r. \end{aligned}$$

This shows

$$\frac{\|\Psi(u+d_\varepsilon) - \Psi(u) - G d_\varepsilon\|_{L^r}}{\|d_\varepsilon\|_{L^r}} = \frac{\rho}{|b-a|} \neq 0$$

and thus Ψ is not $\partial\Psi$-semi-smooth. □

This shows, that unfortunately we can not simply use the chain-rule for proving semi-smoothness. Instead, we need to utilize the combined mapping, noting that G is not only semi-smooth, but infact Fréchet-differentiable and moreover in many cases maps to a slightly better space L^p with $p > r$.

Theorem 5.2.16. *Let $\Omega \subset \mathbb{R}^n$ be measurable with $0 < |\Omega| < \infty$. Further, let $\psi : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz-continuous and $\partial^{\mathrm{cl}}\psi$-semi-smooth. Let $1 \leq r < p \leq \infty$. Then the Nemyzkii-operator Ψ is $\partial\Psi$-semi-smooth (with the generalized differential given in Definition 5.2.14).*

Beweis. Assume that Ψ is not semi-smooth at $u \in L^p(\Omega)^m$. Then there exists $\varepsilon > 0$, $d^k \to 0$ in $L^p(\Omega)^m$, and a sequence $G_k \in \partial\Psi(u + d^k)$ with

$$\|\psi(u + d^k) - \psi(u) - G_k d^k\|_{L^r} \geq \varepsilon \|d^k\|_{L^p}.$$

By definition it is (for almost every $x \in \Omega$)

$$(\Psi(u + d^k) - \Psi(u) - G_k d^k)(x) = \psi(u(x) + d^k(x)) - \psi(u(x)) - g^k(x) d^k(x)$$

with $g^k(x) \in \partial^{\mathrm{cl}}(u(x) + d^k(x))$. If L is the Lipschitz-constant of ψ it holds

$$\begin{aligned} |\psi(u(x) + d^k(x)) - \psi(u(x)) - g^k(x) d^k(x)| &\leq |\psi(u(x) + d^k(x)) - \psi(u(x))| + |g^k(x) d^k(x)| \\ &\leq L|d^k(x)| + |g^k(x) d^k(x)| \\ &\leq 2L|d^k(x)|. \end{aligned}$$

By convergence $d^k \to 0$ in L^p, we can select a subsequence such that $d^k(x) \to 0$ pointwise almost everywhere. As a consequence of pointwise convergence and the definition $g^k(x) \in \partial^{\mathrm{cl}}\psi(u(x) + d^k(x))$, for any given $\delta > 0$ the sets

$$M_{\delta,k} := \{x \in \Omega \,|\, |\psi(u(x) + d^k(x)) - \psi(u(x)) - g^k(x) d^k(x)| \geq \delta |d^k(x)|\}$$

satisfy $|M_{\delta,k}| \to 0$ as $k \to \infty$. Now, $p > r$ and thus Hölder inequality with $\frac{1}{r} = \frac{1}{p} + \frac{1}{s}$ and thus $\frac{1}{s} = \frac{p-r}{pr} > 0$ (with the suitable interpretation for $p = \infty$) gives

$$\begin{aligned} &\|\Psi(u + d^k) - \Psi(u) - G_k d^k\|_{L^r} \\ &\quad \leq \|\Psi(u + d^k) - \Psi(u) - G_k d^k\|_{L^r(M_{\delta,k}^c)} + \|\Psi(u + d^k) - \Psi(u) - G_k d^k\|_{L^r(M_{\delta,k})} \\ &\quad \leq \delta \|d^k\|_{L^r(M_{\delta,k}^c)} + 2L\|d^k\|_{L^r(M_{\delta,k})} \\ &\quad \leq \delta \|1\|_{L^s(M_{\delta,k}^c)} \|d^k\|_{L^p(M_{\delta,k}^c))} + 2L\|1\|_{L^s(M_{\delta,k})} \|d^k\|_{L^p(M_{\delta,k}))} \\ &\quad \leq \left(\delta |\Omega|^{1/s} + 2L|M_{\delta,k}|^{1/s}\right) \|d^k\|_{L^p}. \end{aligned}$$

Selecting $\delta |\Omega|^{1/s} \leq \varepsilon/2$ gives

$$\frac{\|\Psi(u + d^k) - \Psi(u) - G_k d^k\|_{L^r}}{\|d^k\|_{L^p}} \leq \varepsilon/2 + 2L|M_{\delta,k}|^{1/s} \to \varepsilon/2$$

as $k \to \infty$; contradicting the assumption. □

Corollary 5.2.17. *Let $\psi\colon \mathbb{R}^m \to \mathbb{R}$ be Lipschitz-continuous and $\partial^{\mathrm{cl}}\psi(x)$ semi-smooth. Further, let $1 \le r < p \le \infty$ be given, U be a Banach space, $G\colon U \to L^p(\Omega)^m$ Fréchet-differentiable, and $\Omega \subset \mathbb{R}^n$ be measurable with $0 < |\Omega| < \infty$. Define*

$$\Psi_G\colon U \to L^r(\Omega), \quad \Psi_G(u)(x) = \psi(G(u)(x))$$

for any $u \in U$ and $x \in \Omega$. Further, we define the generalized differential

$$\partial\Psi_G\colon U \rightrightarrows \mathscr{L}(U; L^r(\Omega))$$
$$\partial\Psi_G(u) = \{G \mid Gv = g \cdot (G'(u)v), g \in L^\infty(\Omega)^m, g(x) \in \partial^{\mathrm{cl}}\psi(G(u)(x)) \text{ for a.e. } x \in \Omega\}.$$

Then Ψ_G is $\partial\Psi_G$ semi-smooth.

Beweis. This is an immediate consequence of Theorem 5.2.16 and Theorem 5.2.11 (3. and 4.). □

5.2.6. Application

We now consider an optimization problem to discuss the applicability of the afore discussed theory.

$$\begin{aligned} &\min j(q) \\ &\quad \text{s.t. } q \in L^2(\Omega), \quad a \le q \le b \text{ a.e. on } \Omega. \end{aligned} \tag{5.6}$$

Where $\Omega \subset \mathbb{R}^n$ is some measurable set (Domain, boundary, ...) with $0 < |\Omega| < \infty$. W.l.o.g. $a < b \in \mathbb{R}$, otherwise non-constant bounds could be transformed to constant bounds by substituting $q \mapsto \frac{q-a}{b-a}$. Clearly, in this example $G = \mathrm{Id}$ (asserting (RCQ)) and thus the first order optimality conditions can be written as

$$\begin{aligned} \nabla j(q) + \lambda &= 0, \\ q - P_K(q + \sigma\lambda) &= 0 \end{aligned}$$

following Example 5.2.3, with the projection

$$P_K(f)(x) = \min(\max(a, f(x)), b).$$

Clearly, the two lines can equivalently be written as

$$q - P_K(q - \sigma\nabla j(q)) = 0.$$

As we have seen, $P_K\colon L^2 \to L^2$ will not be semi-smooth. Hence, we need to assume additional structure on $\nabla j(q)$ which can be satisfied in many cases.

Assumption 5.2.18. Assume that there exists $\alpha > 0$ and $p > 2$ such that

1. $\nabla j(q) = \alpha q + H(q)$
2. $H : L^2(\Omega) \to L^p(\Omega)$ is continuously Fréchet-differentiable.

With this assumptions, we can select $\sigma = 1/\alpha$ to have the optimality conditions

$$\Phi(q) = q - P_K(q - \sigma\nabla j(q)) = q - P_K(\frac{-1}{\alpha}H(q)) = 0. \tag{5.7}$$

This operator $\Phi(q) = q - P_K(\frac{-1}{\alpha}H(q))$ is now matching our conditions of Corollary 5.2.17 and is thus semi-smooth.

Hence we have seen

Corollary 5.2.19. *Consider the problem (5.6) with Assumption 5.2.18. Then, for $\sigma = 1/\alpha$ the operator Φ in (5.7) is $\partial\Phi$-semi-smooth with*

$$\partial\Phi : L^2(\Omega) \rightrightarrows \mathscr{L}(L^2(\Omega), L^2(\Omega)),$$
$$\partial\Phi(q) = \{G \mid G = \mathrm{Id} + \frac{g}{\alpha}H'(q), \quad g \in L^\infty(\Omega), \quad g(x) \in \partial^{\mathrm{cl}} P_{[a,b]}(-1/\alpha H(q)(x)) \text{ for a.e. } x \in \Omega\}$$

where

$$\partial^{\mathrm{cl}} P_{[a,b]}(s) = \begin{cases} 0 & x \notin [a,b], \\ 1 & x \in (a,b), \\ [0,1] & \text{otherwise.} \end{cases}$$

For the application of the semi-smooth Newton method 5.3 one now needs to show, that the inverse of all elements in $\partial\Phi(q)$ are uniformly bounded close to $\overline{q}$, i.e.,

$$\|G^{-1}\|_{\mathscr{L}(L^2,L^2)} \leq C \quad \forall G \in \partial\Phi(q), q \in B_\delta^{L^2}(\overline{q}).$$

5.2.7. Control Constrained Optimal Control

We come back to a control constrained setting similar to Example 3.3.9 but with a bit more explicit structure.

$$\min\ J(q,u) = \frac{1}{2}\|u - u^d\|^2 + \frac{\alpha}{2}\|q\|^2$$
$$\text{s.t.} \begin{cases} -\Delta u = q, & \in H^{-1}(\Omega) = H_0^1(\Omega)^* \\ q \in Q^{\mathrm{ad}} := \{p \in Q = L^2(\Omega) \mid a \leq p \leq b\}, \end{cases}$$

on a bounded domain $\Omega \subset \mathbb{R}^2$. Since the operator $-\Delta$ defines an isomorphism from $H_0^1(\Omega) \to H^{-1}(\Omega)$ we can consider the (linear) solution operator $S: H^{-1}(\Omega) \to H_0^1(\Omega)$. Notice, that the embeddings $H_0^1(\Omega) \to L^p(\Omega)$ are compact (in 2d for any $p < \infty$). This gives the reduced problem

$$\begin{aligned} \min\ & j(q) = \frac{1}{2}\|Sq - u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } & \{q \in Q^{\text{ad}} := \{p \in Q = L^2(\Omega) \,|\, a \le p \le b\}. \end{aligned}$$

A short calculation shows that w.r.t the L^2-inner product we have

$$\begin{aligned} \nabla j(q) &= \alpha q + S^*(Sq - u^d) \in L^2(\Omega), \\ \nabla^2 j(q) &= \alpha \,\mathrm{Id} + S^*S \in \mathscr{L}(L^2, L^2). \end{aligned}$$

Hence we identify Assumption 5.2.18 with $H(q) = S^*(Sq - u^d)$ Note that the adjoint $S^*: H^{-1}(\Omega) \to H_0^1(\Omega)$ and hence

$$S^*(Sq - u^d) \in H_0^1(\Omega) \subset L^p(\Omega) \subset L^2(\Omega)$$

for some $p > 2$, and thus it is differentiable as a mapping $L^2(\Omega) \to L^p(\Omega)$.

Hence our optimality system can be written in terms of the semi-smooth equation

$$\Phi(q) = q - P_{[a,b]}\Big(\frac{-1}{\alpha} S^*Sq\Big) = 0.$$

To simplify the following calculations, we select the particular generalized derivative given by selecting

$$g(x) = \begin{cases} 1 & \frac{-1}{\alpha} S^*(Sq(x) - u^d(x)) \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

then the generalized derivative of Φ is given by

$$\partial\Phi(q) = \{\mathrm{Id} + \frac{1}{\alpha}\chi_{\mathscr{I}} S^*S\} \tag{5.8}$$

where $\chi_{\mathscr{I}}$ is the indicator function of the inactive set of the projection $P_{[a,b]}$

$$\mathscr{I} = \{x \in \Omega \,|\, \frac{-1}{\alpha}(S^*Sq)(x) \in [a,b]\}$$

Lemma 5.2.20. *Let $G \in \partial\Phi(q)$ given by (5.8). Then G is invertible and satisfies*

$$\|G^{-1}\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))} \le 2 + \frac{1}{\alpha}\|S^*S\|_{\mathscr{L}(L^2(\Omega,L^2(\Omega)))}.$$

Beweis. To see invertibility of G, we notice that $G = \mathrm{id} + \frac{1}{\alpha}\chi_{\mathscr{I}} S^*S$ is a compact perturbation of the identity and thus Fredholm of index zero and invertibility follows if $\ker G = \{0\}$.

To this end, let $d, F \in L^2(\Omega)$ be given satisfying the equation $Gd = F$, i.e.,

$$F = Gd = \frac{1}{\alpha}\chi_{\mathscr{I}} S^*Sd + d.$$

On the active set $\mathscr{A} = \Omega \setminus \mathscr{I}$ this gives

$$\chi_{\mathscr{A}} d = \chi_{\mathscr{A}} F.$$

Now we use this to determine $d\big|_{\mathscr{I}}$. To this end, we consider the relation

$$\frac{1}{\alpha} S^*S(\chi_{\mathscr{I}} d) + \chi_{\mathscr{I}} d = F - \frac{1}{\alpha} S^*S(\chi_{\mathscr{A}} d)$$

on $\mathscr{I}$.

Then it holds

$$\begin{aligned}\int_\Omega (\chi_{\mathscr{I}} d)^2 \,\mathrm{d}x &\le \frac{1}{\alpha}\int_\Omega (S^*S(\chi_{\mathscr{I}} d)\chi_{\mathscr{I}} d \,\mathrm{d}x + \int_\Omega (\chi_{\mathscr{I}} d)^2 \,\mathrm{d}x \\ &= \int_\Omega (F - \frac{1}{\alpha} S^*S(\chi_{\mathscr{A}} d)(\chi_{\mathscr{I}} d)\,\mathrm{d}x.\end{aligned}$$

By Hölder's inequality, we get

$$\begin{aligned}\|d\|_{L^2(\Omega)} &\le \|\chi_{\mathscr{I}} d\|_{L^2(\Omega)} + \|\chi_{\mathscr{A}} d\|_{L^2(\Omega)} \\ &\le (\|F\|_{L^2(\Omega)} + \frac{1}{\alpha}\|S^*S(\chi_{\mathscr{A}} d)\|_{L^2(\Omega)}) + \|\chi_{\mathscr{A}} d\|_{L^2(\Omega)} \\ &\le 2\|F\|_{L^2(\Omega)} + \frac{1}{\alpha}\|S^*S\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))}\|F\|_{L^2(\Omega)}.\end{aligned}$$

Hence taking $F = 0$ this gives $\ker G = \{0\}$ and thus

$$\|G^{-1}F\|_{L^2(\Omega)} \le 2\|F\|_{L^2(\Omega)} + \frac{1}{\alpha}\|S^*S\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))}\|F\|_{L^2(\Omega)}$$

showing the assertion. □

6. Discretization

To sketch the necessary steps in analysing discrete approximations to optimization problems in function spaces, we consider the following model problem

$$\min J(q,u) \quad \text{s.t. } (6.1).$$

Where the constraint is to find $u \in H_0^1(\Omega)$ such that

$$(a\nabla u, \nabla\varphi) = (f,\varphi) \quad \forall \varphi \in H_0^1(\Omega) \tag{6.1}$$

for some $a \in W^{1,\infty}(\Omega)$ with $0 < \underline{a} \leq a(a) \leq \overline{a}$ on Ω. Then a generic way to discretize this equation is to choose some finite dimensional space V_h (for convenience let $V_h \subset H_0^1(\Omega)$ then V_h is called H_0^1-conform). Then one considers the problem to find $u_h \subset V_h$ such that

$$(a\nabla u_h, \nabla\varphi_h) = (f,\varphi_h) \quad \forall \varphi_h \in V_h. \tag{6.2}$$

In particular, one searches for a solution u_h in the *ansatz space* V_h such that the variational equation is satisfied for all φ_h in the *test space* V_h (Although here they are the same.). The equation (6.2) has a unique solution by the same arguments that are used for (6.1).

Theorem 6.0.1. *Let u be the solution to (6.1) and u_h be the solution to (6.2). Then the following quasi-best-approximation property holds*

$$\|\nabla(u-u_h)\| \leq \frac{\overline{a}}{\underline{a}} \inf_{\varphi_h \in V_h} \|\nabla(u-\varphi_h)\|. \tag{6.3}$$

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

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

□

By choosing a basis $\varphi_h^{(i)}$, $i = 1,\dots,N$ of V_h the problem (6.2) an be rewritten as a linear algebraic problem as follows: Find $\mathbf{u_h} = (u_h^{(i)})_{i=1,\dots,N}$ such that

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

with the stiffness matrix

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

and the right hand side

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

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

6.1. Linear Finite Elements for Elliptic Problems

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

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

1. (Structural regularity or feasibility) Any two different elements T_1, $T_2 \in \mathcal{T}_h$ intersect at most in a vertex or along an entire edge.
2. (Shape regularity) For the radius ρ_T of the incircle and h_T of the circumscribed circle of an element $T \in \mathcal{T}_h$ it holds
$$\max_{T\in\mathcal{T}_h} \frac{h_T}{\rho_T} \le c$$
uniformly in $h \to 0$.
3. (Size regularity) It holds
$$\max_{T\in\mathcal{T}_h} h_T \le c \min_{T\in\mathcal{T}_h} h_T$$
uniformly in T.

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

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

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

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

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

Remark 6.1.1. Of course higher order polynomials are also possible, e.g., piecewise polynomials of degree two. Which then require additional basis functions, for example midpoints of the edges.

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

6.1.1. A Priori Error Estimates

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

Definition 6.1.2. For a function $u \in C(\overline{\Omega})$ we define the Lagrange interpolation $I_h u \colon C(\overline{\Omega}) \to V_h^{(1)}$ as follows

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

Theorem 6.1.3. *Let $\mathcal{T}_h$ satisfy the regularity assumptions above. Then there exists a constant $c > 0$ independent of h such that for $u \in W^{2,2}(\Omega)$ it holds*

$$\|u - I_h u\|_2 + h\|\nabla(u - I_h u)\|_2 \leq ch^2 \|u\|_{2,2}.$$

Beweis. See, e.g., [Brenner and Scott, 2008, Theorem 4.4.20] □

Theorem 6.1.4. *Let $\mathcal{T}_h$ satisfy the regularity assumptions above, and let Ω be convex. Then*

the solution u to (6.1) and u_h to (6.2) satisfy for $f \in L^2(\Omega)$

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

Beweis. From Theorem 6.0.1 and Theorem 6.1.3 we have that

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

and by the elliptic regularity, see, e.g., Gilbarg and Trudinger [2001], it follows

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

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

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

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

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

Then we obtain

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

hence the assertion follows. □

Remark 6.1.5. The above theorem states that the operator $S\colon L^2(\Omega) \to L^2(\Omega)$ given by (6.1) and the discrete operator $S_h\colon L^2(\Omega) \to L^2(\Omega)$ given by (6.2) satisfy

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

or equivalently

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

6.2. Discretization of the Model Problem

We return to the model problem

$$\begin{aligned} \min J(q,u) &= \frac{1}{2}\|u-u^d\|^2 + \frac{\alpha}{2}\|q\|^2 \\ \text{s.t. } & u \in H_0^1(\Omega), \\ & (a\nabla u, \nabla\varphi) = (q,\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \\ & q_{\min} \le q \le q_{\max} \quad \text{a.e. on } \Omega. \end{aligned} \tag{6.4}$$

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

6.2.1. Variational Discretization

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

$$\begin{aligned} \min J(q_h, u_h) = \frac{1}{2}\|u_h - u^d\|^2 + \frac{\alpha}{2}\|q_h\|^2 \\ \text{s.t. } u_h \in V_h, \\ (a\nabla u_h, \nabla \varphi_h) = (q_h, \varphi_h) \quad \forall\, \varphi_h \in V_h, \\ q_{\min} \le q_h \le q_{\max} \quad \text{a.e. on } \Omega. \end{aligned} \tag{6.5}$$

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

Theorem 6.2.1. *For $\alpha > 0$ there exists a unique solution $(\overline{q}_h, \overline{u}_h) = (\overline{q}_h, S_h\overline{q}_h) \in Q^{ad} \times V_h$ to (6.5). This solution is (equivalently) characterized by the existence of $\overline{z}_h \in V_h$ such that the following holds:*

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

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

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

is convex and continuous and thus w.l.s.c. with $j_h(q_h) \to \infty$ as $\|q_h\| \to \infty$. Thus we can apply Theorem 2.2.3 to obtain existence of a solution $(\overline{q}_h, \overline{u}_h)$ to (6.5). Due to strict convexity of j_h this solution is unique.

Now as for j the functional j_h is continuously Fréchet differentiable and hence by Theorem 3.2.3 the solution $\overline{q}_h$ to

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

is equivalently characterized by the variational inequality

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

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

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

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

Remark 6.2.2. We remark that the optimality conditions obtained here are the discretized optimality conditions of the continuous case. This means that it does not matter whether we discretize the optimization problem and compute necessary optimality conditions or if we directly discretize the optimality conditions. This is often refered to as *'optimize-then-discretize = discretize-then-optimize'*. This is not always the case, e.g., in finite difference schemes this need not happen.

Theorem 6.2.3. *Let $(\overline{q},\overline{u})$ be the solution to (6.4) and $(\overline{q}_h,\overline{u}_h)$ be the solution to (6.5). Then there exists a constant (independent of h and α) such that*

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

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

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

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

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

We obtain, using $(S-S_h)^*=S^*-S_h^*$, Young's inequality, and a generic constant c (which may

be different in each line of the inequality),

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

By subtracting the last term, we get

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

which proofs the assertion. □

Remark 6.2.4. It is clear that if Q is a finite dimensional space, e.g., we have only finitely many control variables, then this discretization is sufficient to obtain a finite dimensional problem.

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

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

holds again pointwise almost everywhere, by our usual arguments, yielding that

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

This can then be used to eliminate the control in the optimality conditions given by Theorem 6.2.1. This means that

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

solves (6.5) if and only if $(\overline{u}_h,\overline{z}_h)$ solves the following nonlinear (and non smooth) problem

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

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

This technique has been advertised in Hinze [2005], see also Hinze et al. [2009]

Remark 6.2.5. It was recently shown Gaspoz et al. [2020] that in fact the preasymptotic estimate in Theorem 6.2.3 overestimates the true error. In fact, it was shown, that the quasi-best approximation estimate

$$\|\nabla(u-u_h)\| + \|\nabla(z-z_h)\| \leq c_h \inf_{\varphi_h,\psi_h \in V_h} \Big(\|\nabla(u-\varphi_h)\| + \|\nabla(z-\psi_h)\| \Big) \tag{6.6}$$

with a constant $c_h \to \frac{\bar{a}}{\underline{a}}$ so that asymptotically the same best-approximation property as for the PDE (6.3) holds.

Remark 6.2.6. Of course, next to the implicit, variational, discretization one can choose to discretize the control space as well. Analogous estimates to Theorem 6.2.3 can be shown, by more complicated arguments. However, a quasi-best approximation as in 6.6 will no longer hold, see, e.g., Gaspoz et al. [2020].

A. Functional Analytic Background

A.1. Normed Linear Spaces

Definition A.1.1. A $\mathbb{R}$-linear space V is called normed linear space if there exists a map $\|\cdot\| = \|\cdot\|_V : V \to \mathbb{R}_{\geq 0}$ which satisfies:

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

Definition A.1.2. A sequence $v_k \in V$ in a normed linear space V is called *Cauchy sequence* if for any given ε there exists some $n_0 \in \mathbb{N}$ such that

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

Definition A.1.3. A normed linear space V is called *Banach space* if it is complete, i.e., any Cauchy sequence in V has a limit in V.

Definition A.1.4. A Banach space V whose norm is induced by a scalar product is called a *Hilbert space*. We recall, a scalar product is a positive definite bilinear form, $(\cdot,\cdot)_V = (\cdot,\cdot) : V \times V \to \mathbb{R}$, e.g.,

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

The scalar product induces a norm by the following definition

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

Example A.1.5. An example for an infinite dimensional Banach space is given by the space $C[0,1]$ of continuous functions on the interval $[0,1]$ equipped with the norm

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

Example A.1.6. An example for an infinite dimensional Hilbert space is given by the space l_2 of square summable sequences, e.g., $(x_i)_{i=0}^\infty \in l_2$ if and only if $\sum_{i=0}^\infty x_i^2 < \infty$. The scalar product is given by

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

Remark A.1.7. A norm $\|\cdot\|$ on a normed space V is induced by a scalar product if and only if the parallelogram law is satisfied, e.g., it holds for $v,w \in V$

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

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

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

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

We conclude this section with an important definition.

Definition A.1.8. A normed linear space V is called *separable*, if there exists a countable dense (w.r.t the norm on V) subset of V.

Example A.1.9. A typical example of a separable Banach space is again $C[0,1]$ with the norm $\|\cdot\|_\infty$. Since it is known from calculus lessons, that any continuous function can be approximated uniformly by polynomials (Theorem of Stone-Weierstrass).

A.2. Convexity

Definition A.2.1. Let V be a vector space. A set $C \subset V$ is called *convex*, if for any $\lambda \in (0,1)$ and $v, w \in C$ it holds

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

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

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

Definition A.2.2. Let V be a vector space and $C \subset V$ be convex. A function $f : C \to \mathbb{R} \cup \{\infty\}$ is called convex, if (with the usual definitions for arithmetic with ∞) for any $\lambda \in (0,1)$ and $v, w \in C$ it holds

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

Theorem A.2.3. *Let V be a Hilbert space and $C \subset V$ nonempty, closed, and convex. Then there exists a unique mapping $\mathscr{P}_C : V \to C$ given by*

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

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

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

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

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

This yields

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

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

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

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

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

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

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

an application of the parallelogram law yields as above

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

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

Lemma A.2.4. *Let V be a Hilbert space and $C \subset V$ nonempty, closed, and convex. The element $\mathscr{P}_C$ in Theorem A.2.3 is given equivalently by the variational inequality*

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

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

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

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

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

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

□

A.3. Linear Operators

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

Definition A.3.1. A mapping $A\colon V \to W$ is called linear or *linear operator* if for any $\lambda \in \mathbb{R}$ and $v_1, v_2 \in V$ it holds

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

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

Definition A.3.2. A linear operator $A\colon V \to W$ is *bounded* if there exists a constant c_A independent of v such that

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

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

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

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

Theorem A.3.3. *$\mathscr{L}(V,W)$ is a normed linear space (exercise) with norm*

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

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

Theorem A.3.4. *A linear operator $A\colon V \to W$ is bounded if and only if it is continuous.*

Example A.3.5 (Multiplication operator)**.** Consider $V = C[0,1]$. Let $f \in V$ be an arbitrary function. Then we define a linear operator A by

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

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

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

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

Definition A.3.6. The set $\mathscr{L}(V,\mathbb{R})$ is called (topological) *dual space* to V and is denoted by V^*.

Theorem A.3.7 (Riesz representation)**.** *Let V be a Hilbert space with scalar product $(\cdot,\cdot)$. Then for any bounded linear functional $F \in V^*$ there exists a unique element $f \in V$ such that*

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

Beweis. Assume w.l.o.g. that $F \neq 0$. The set $C = \mathscr{N}(F) = \{x \in H \mid F(x) = 0\}$ is clearly convex, closed and nonempty (Exercise). Hence by Theorem A.2.3 the projection $\mathscr{P}_C$ is well defined.

We choose some $w \in V$ with $F(w) = 1$ and define $\tilde{f} = w - \mathscr{P}_C(w)$. We obtain $F(\tilde{f}) = 1$ and hence $\tilde{f} \neq 0$. From Lemma A.2.4 (and the fact that C is a subspace) we obtain

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

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

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

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

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

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

□

For more general cases one may need the following generalization

Theorem A.3.8 (Lax-Milgram). *Let V be a Hilbert space and $a : V \times V \to \mathbb{R}$ a bilinear form which satisfies for some constants $\alpha, \beta > 0$ that for any $u, v \in V$ it holds*

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

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

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

Further it holds

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

Beweis. Exercise. □

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

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

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

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

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

Definition A.3.9. A Banach space B is called reflexive if i is surjective, e.g., B and B^{**} are isometrically isomorphic.

Example A.3.10.
1. Every Hilbert space is reflexive.
2. The spaces l_p and $L^p(\Omega)$ ($\Omega \subset \mathbb{R}^n$) are reflexive if $1 < p < \infty$.
3. The space $C(\overline{\Omega})$, $L^1(\Omega)$, and $L^\infty(\Omega)$ are not reflexive if $\Omega \subset \mathbb{R}^n$ is open and nonempty.

A.4. Adjoints

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

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

clearly g is linear and because of

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

it holds $g \in U^*$ with

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

Definition A.4.1. For any $A \in \mathscr{L}(U,V)$ the above mapping $f \mapsto g = f \circ A$ defines a linear operator $A^* : V^* \to U^*$, $f \mapsto f \circ A$ the *adjoint* or *dual* operator to A.

It is given by the relation

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

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

For Hilbert spaces there exists a natural choice for the duality pairing given by Theorem A.3.7. This gives the following

Definition A.4.2. Let U and V be Hilbert spaces and $A \in \mathscr{L}(U,V)$. Then we define the

Hilbert space adjoint $A^* \in \mathscr{L}(V,U)$ to A by

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

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

A.5. Weak Convergence

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

Definition A.5.1 (Weak convergence)**.** Let B be a Banach space. A sequence $v_k \in B$ *converges weakly* to some $v \in B$ if

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

This is written as $v_k \rightharpoonup v$.

Definition A.5.2 (Weak* convergence)**.** Let B be a Banach space. A sequence $f_k \in B^*$ *converges weakly** to some $f \in B^*$ if

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

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

Example A.5.3. Consider the Hilbert space $L^2(0,2\pi)$ with the sequence

$$v_k = \sin(kx)$$

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

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

by Parseval's identity (The scalar product is up to a constant the k-th Fourier coefficient.

This means $v_k \rightharpoonup 0$. Moreover we have that

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

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

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

The following important properties hold:

Lemma A.5.4. *Let B be a Banach and H be a Hilbert space.*

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

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

   (The analog property holds for Banach spaces)

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

Definition A.5.5. A set $M \subset B$ of a Banach space B is *weak sequentially closed* if for any sequence $v_k \in M$ with weak limit $v \in B$ it holds $v \in M$.

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

We have the following important properties:

Theorem A.5.6. *Let B be a Banach space.*

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

Theorem A.5.7. *Let B be a Banach space and $J : B \to \mathbb{R}$ be a continuous and convex functional. Then J is weakly lower semicontinuous, e.g., if $v_k \rightharpoonup v$ then*

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

Remark A.5.8. In particular the norm $\|\cdot\| : B \to \mathbb{R}$ is convex and continuous and thus weakly lower semicontinuous.

Definition A.5.9. Let V, W be two Banach spaces. We call a map $A \in \mathscr{L}(V, W)$ compact, if for any bounded sequence $v_k \in V$ there exists a subsequence of Av_k which converges strongly.

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

A.6. Lebesgue and Sobolev spaces

A.6.1. Domain regularity

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

Definition A.6.1. We say that a bounded domain Ω with boundary Γ is of class $C^{k,1}$ if the boundary of Ω is locally the graph of a $C^{k,1}$ function such that Ω is locally on one side of the boundary.

For the technical detail see, e.g., Adams and Fournier [2003], Wloka [1982]. For our purposes the most important regularity will be Lipschitz domains ($C^{0,1}$) domains.

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

A.6.2. Lebesgue spaces

Definition A.6.2. On a domain Ω we define the spaces $L^p(\Omega)$ as follows

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

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

Remark A.6.3. Note that from now on if we write down a norm or scalar product for functions without any indices this will be the $L^2(\Omega)$ norm or scalar product.

Theorem A.6.4. *The dual space to $L^p(\Omega)$ for $1 \le p < \infty$ is isometrically isomorphic to $L^{p'}(\Omega)$ where $\frac{1}{p} + \frac{1}{p'} = 1$.*

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

Theorem A.6.5 (Lebesgue differentiation theorem)**.** *Let $f \in L^1(\Omega)$ for some bounded domain $\Omega \subset \mathbb{R}^n$. Then for almost all $x \in \Omega$ it holds*

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

Beweis. For a proof see Giaquinta et al. [1998]. □

A.6.3. Sobolev spaces

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

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

Definition A.6.6 (Weak derivatives)**.** Let $u \in L^1_{\text{loc}}(\Omega)$ and α a multi-index. If there exists

a function $w \in L^1_{\text{loc}}(\Omega)$ such that

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

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

As usual we denoted by ∂^α the term

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

Example A.6.7. The function $u(x) = |x|$ on $\Omega = (-1,1)$ has the weak derivative

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

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

We can now define the Sobolev spaces

Definition A.6.8. On a bounded domain Ω with given $k \in \mathbb{N}_0$ and $1 \le p \le \infty$, we define

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

Together with the norm

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

for $1 \le p < \infty$ and

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

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

Definition A.6.9. On a bounded domain Ω, we define for $k \in \mathbb{N}_0$ and $1 \le p < \infty$ we define $W_0^{k,p}(\Omega)$ as the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$.

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

Example A.6.10. In the special case of $H^1(\Omega)$, which we will frequently use, the norm has the following form

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

The corresponding scalar product is given as

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

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

Theorem A.6.11 (Boundary Traces)**.** *Let Ω be a bounded Lipschitz domain and $1 \leq p \leq \infty$. Then there exists a unique operator $\tau \in \mathscr{L}(W^{1,p}(\Omega), L^p(\partial\Omega))$ such that*

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

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

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

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

Theorem A.6.12 (Poincaré inequality)**.** *Let Ω be a bounded domain. Then there exists a constant c_Ω depending on Ω only such that*

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

Corollary A.6.13. *The norms $\|\cdot\|_{1,2}$ and $\|\nabla \cdot\|$ are equivalent on $H_0^1(\Omega)$.*

A.6.4. Embedding theorems

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

Theorem A.6.14. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. Let $k_1, k_2 \in \mathbb{N}_0$ and $1 \leq p_1, p_2 < \infty$.*
a) If it holds

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

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

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

b) If it holds

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

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

Theorem A.6.15. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. Let $k_1, k_2 \in \mathbb{N}_0$ with $k_1 \geq 1$ and $1 \leq p < \infty$ and $0 \leq \alpha \leq 1$. If it holds*

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

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

Theorem A.6.16. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. Let $k \in \mathbb{N}$, then the embedding*

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

exists and is an isomorphism.

Example A.6.17. We consider again the case of the space $H^1(\Omega)$. Now we consider the question whether or not these functions are continuous. To this end we note that the relevant inequality gets

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

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

In fact for these dimensions functions in $H^1(\Omega)$ are not necessarily continuous. To see this

let $\Omega = B_1(0) \subset \mathbb{R}^2$ and consider the function $u(x) = \ln(\ln(r^{-1}) + 1)$. A straightforward computation shows that $u \in H^1(\Omega)$ but clearly $u \notin C^0(\overline{\Omega})$.

Literaturverzeichnis

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

Hans W. Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Springer, Berlin – Heidelberg – New York, 2002.

Jürgen Appell and Petr P. Zabrejko. *Nonlinear Superposition Operators*. Cambridge University Press, 1990.

V. I. Bogachev and E. Mayer-Wolf. Some remarks on Rademacher's theorem in infinite dimensions. *Potential Analysis*, 5:23–30, 1996.

J. F. Bonnans and A. Shapiro. Optimization problems with perturbations: A guided tour. *SIAM Rev.*, 40(2):228–264, 1998. doi: 10.1137/S0036144596302644.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, 2000.

S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, New York, 3. edition, 2008.

P. G. Ciarlet. *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

B. Dacorogna. *Direct Methods in the Calculus of Variations*, volume 78 of *Applied Mathematical Sciences*. Springer, second edition, 2008.

J. W. Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.*, 4:10–26, 1967.

J. C. De los Reyes. *Numerical PDE-constrained optimization*. SpringerBriefs in Optimization. Springer, Cham, 2015.

F. Gaspoz, C. Kreuzer, A. Veeser, and W. Wollner. Quasi-best approximation in optimization with PDE constraints. *Inverse Problems*, 36(1), 2020. doi: 10.1088/1361-6420/ab47f3.

M. Giaquinta, G. Modica, and J. Souček. *Cartesian Currents in the Calculus of Variations I*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, Berlin – Heidelberg – New York, 1. edition, 1998.

D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Grundlehren der mathematischen Wissenschaften*. Springer, revised 3. edition, 2001.

M. Hintermüller, A. Schiela, and W. Wollner. The length of the primal-dual path in Moreau-Yosida-based path-following for state constrained optimal control. *SIAM J. Optim.*, 24(1): 108–126, 2014.

M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comp. Optim. Appl.*, 30(1):45–61, 2005.

M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.

K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), 2008.

M. Lavrentieff. Sur quelques problèmes du calcul des variations. *Ann. Mat. Pura Appl.*, 4(1): 7–28, 1927.

Stephen M. Robinson. Stability theory for systems of inequalities. II. Differentiable nonlinear systems. *SIAM J. Numer. Anal.*, 13(4):497–513, 1976.

W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., 1991.

A. Schiela and W. Wollner. Barrier methods for optimal control problems with convex nonlinear gradient state constraints. *SIAM J. Optim.*, 21(1):269–286, 2011.

Anton Schiela. *The Control Reduced Interior Point Method*. Verlag Dr. Hut, München, 2006.

F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen*. Vieweg, 1. edition, 2005.

M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2011.

D. Werner. *Funktionalanalysis*. Springer, Berlin – Heidelberg – New York, 2005.

J. Wloka. *Partielle Differentialgleichungen. Sobolewräume und Randwertaufgabe*. Teubner, Leipzig, 1. edition, 1982.

Jochem Zowe and Stanislaw Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5:49–62, 1979.