Lecture Notes
Infinite Dimensional Optimization

Roland Herzog*

2025-01-26

*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany (roland.herzog@iwr.uni-heidelberg.de, https://scoop.iwr.uni-heidelberg.de/team/roland-herzog).

Material for approximately 25 lectures, 14 weeks.

In these lecture notes we use colored markup for **definitions** and alerts.

Expert Knowledge: topic

A block like this contains further information that are not subject to examination.

Contents

Chapter 0 Introduction

We will consider in this class optimization problems of the following kind:

$$\begin{aligned} &\text{Minimize} && f(x), \quad \text{where } x \in X \\ &\text{subject to} && h(x) = 0. \end{aligned}$$

In this problem, $f\colon X \to \mathbb{R}$ is called the **objective function** and $h\colon X \to Y$ is the **equality constraint**. The **optimization variable** x is sought in some **optimization space** X.

Inequality constraints may be added to the above problem, either

- explicitly in the form $g(x) \leqslant 0$ or, more generally, in the form $g(x) \in K \subseteq Z$,
- or implicitly, by imposing $x \in C \subseteq X$ or allowing f to take values in $\mathbb{R} \cup \{\infty\}$.

Often, K is a cone and C is a convex set.

What are reasonable choices for the "spaces" X, Y, Z?

(1) To define the notion of global minimizers, no structure at all is required, so X, Y, Z can be general sets.

(2) To define the notion of local minimizers, the space X of optimization variables must carry a topology since we require the concept of neighborhoods.

(3) Statements about the existence of global minimizers build on notions of continuity and compactness.[1] Therefore, topological spaces are required for this purpose as well.

(4) To formulate first-order optimality conditions, we need to be able to differentiate. A convenient setting for this are normed linear spaces.

(5) For algorithmic purposes, derivatives need to be converted into directions, e. g., directions of largest/smallest directional derivatives over the unit sphere. For this purpose, normed linear spaces or even Hilbert spaces, are convenient.

Based on these considerations, we will consider only **normed linear spaces** over the field of real numbers $\mathbb{R}$ (§ 2).[2]

We may anticipate a couple of differences compared to optimization over finite-dimensional linear spaces, as well as a number of questions that we will may want to to answer throughout the course:

(1) Different norms on an infinite-dimensional linear space are, in general, not equivalent to each other.

(2) How do we differentiate functions defined on infinite-dimensional normed linear space?

[1] Compare, for instance, the Weierstrass extreme value theorem: a continuous function $f\colon X \to \mathbb{R}$ attains its minimum (and its maximum) on a compact set $K \subseteq X$; see also Theorem 5.1.

[2] We use the term **linear space** instead of the synonymous **vector space**.

(3) Can we formulate optimization algorithms on infinite-dimensional spaces?

(4) If so, then when and how do we discretize in order to realize them numerically?

§ 1 Motivating Examples

Example 1.1 (Brachistochrone problem)**.**
In a 1696 article, Johann Bernoulli posted the following problem:

> Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time?

This problem is known as the **Brachistochrone problem** (ancient Greek: *βράχιστος χρόνος*). In modern terms, it can be formulated as follows. Suppose that the points have coordinates $A = (0, 0)$ and $B = (b_1, b_2)$ with $b_2 \geqslant 0$. Let $g > 0$ denote the gravitational constant.

We are seeking a function $\gamma\colon [0, b_1] \to \mathbb{R}$ whose graph defines the curve from A to B. Using the principle of conservation of (potential plus kinetic) energy, we may express the speed of the particle at horizontal position x in terms of its height $\gamma(x)$. Skipping the details, this eventually leads to the following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(\gamma) := \int_0^{b_1} \frac{\sqrt{1+\gamma'(x)^2}}{\sqrt{2\,g\,\gamma(x)}}\,\mathrm{d}x, \quad \text{where } \gamma \in X \\ \text{s.t.} \quad & \gamma(0) = 0 \\ \text{and} \quad & \gamma(b_1) = b_2 \\ \text{as well as} \quad & \gamma \geqslant 0 \text{ on } [0, b_1]. \end{aligned} \tag{1.1}$$

Here X is a suitable vector space of functions $\gamma\colon [0, b_1] \to \mathbb{R}$, e. g., $X = C^1(0, b_1) \cap C([0, b_1])$, the space of continuous functions on $[0, b_1]$ whose restriction to the open interval $(0, b_1)$ is continuously differentiable. An alternative is the **Sobolev space** $X = H^1(0, b_1)$ of square integrable functions with square integrable weak derivative on $(0, b_1)$.[3]

(**Quiz 1.1:** Does the gravitational constant impact optimal curves?) One can show that the (unique) minimizer of (1.1) satisfies a first-order necessary optimality condition, which comes in the form of a differential equation:

$$\frac{1}{2}\sqrt{\frac{1+\gamma'(x)^2}{\gamma(x)^3}} + \frac{\mathrm{d}}{\mathrm{d}x}\frac{\gamma'(x)}{\sqrt{\gamma(x)\left(1+\gamma'(x)^2\right)}} = 0.$$

The solutions of this equation satisfy

$$\gamma(x)\left(1+\gamma'(x)^2\right) = C \quad \text{in } (0, b_1) \tag{1.2}$$

for some $C > 0$, and it has infinite slope initially:

$$\lim_{x \searrow 0} \gamma'(x) = \infty.$$

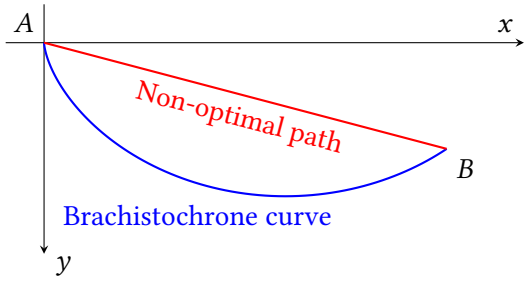
[3] We will introduce Sobolev spaces later; see § 2.6.

The unique solution is given by the curve

$$t \mapsto \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = C \begin{pmatrix} t - \sin(t) \\ 1 - \cos(t) \end{pmatrix} \quad \text{for } t \in [0, T], \tag{1.3}$$

where $C > 0$ and $T \in (0, 2\pi]$ are determined by the conditions $x(T) = b_1$ and $y(T) = b_2$.

This curve is a segment of a cycloid with radius C. △



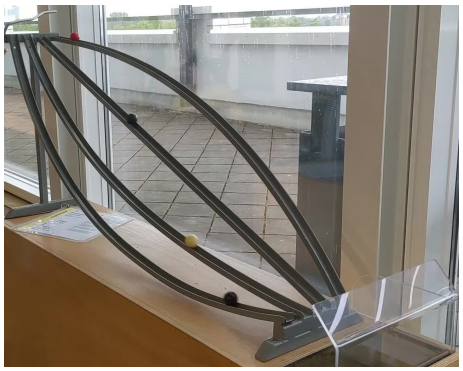


Figure 1.1: Some non-optimal curve $\gamma\colon [0, b_1] \to \mathbb{R}$ from A to B (left) as well as the unique global minimizer of the Brachistochrone problem (1.1), given by the segment of a cycloid (left). Image of an experimental device on display at Technoseum Mannheim (right), shot by Roland Herzog.

Remark 1.2 (on the Brachistochrone problem)**.**
The first-order optimality condition of the Brachistochrone problem come in the form of a differential equation (1.2). This is typical for optimization problems whose variables are functions and whose objectives involve derivatives of those functions. As a result, minimizers may be more regular than suggested by the optimization space X. This is indeed the case in the Brachistochrone problem (1.1), where the unique minimizer turns out to be a $C^\infty(0, b_1)$-function. △

Expert Knowledge: The origins of the calculus of variations

The Brachistochrone problem belongs to a class of problems referred to as calculus of variations, where optimization variables are functions and objectives are typically integrals involving values of the function and its derivative(s). This term was coined in 1766 by Leonhard Euler. The first-order optimality conditions for calculus of variations problems are referred to as **Euler-Lagrange equations**.

Newton's problem of minimal resistance from 1687 is considered the first problem of this type, and the Brachistochrone problem (1696) is second. That problem attracted the attention of Johann Bernoulli's brother Jakob, as well as of Isaac Newton, Gottfried Leibniz, Ehrenfried Walther von Tschirnhaus and Guillaume de l'Hôpital, who all turned in solutions.

Example 1.3 (Fermat's principle in optics)**.**

Suppose that $n\colon \mathbb{R}^2 \to \mathbb{R}_{>0}$ is the material dependent refractive index of an optical material. Let $\gamma\colon [0, b_1] \to \mathbb{R}$ denote a function whose graph defines a curve through this material. Then the optical length of this curve is defined by

$$\int_0^{b_1} n(x, \gamma(x)) \sqrt{1 + \gamma'(x)^2}\, \mathrm{d}x.$$

Fermat's principle stipulates that the path a ray of light will take minimizes the optical length. Suppose that the end points of that path are $A = (0, 0)$ and $B = (b_1, b_2)$. Then we obtain the following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(\gamma) := \int_0^{b_1} n(x, \gamma(x)) \sqrt{1 + \gamma'(x)^2}\, \mathrm{d}x, \quad \text{where } \gamma \in X \\ \text{s. t.} \quad & \gamma(0) = 0 \\ \text{and} \quad & \gamma(b_1) = b_2. \end{aligned} \tag{1.4}$$

In the particular case where the refractive index is piecewise constant on slabs, the unique global minimizer of (1.4) satisfies **Snell's law**, which states that the incident angles θ_+, θ_- (measured against the normal) of two neighboring slabs satisfy the relation $n_+ \sin(\theta_+) = n_- \sin(\theta_-)$, see Figure 1.2.

Similar as in Example 1.1, every minimizer satisfies a first-order optimality condition that amounts to a differential equation:

$$-\frac{n(x, \gamma(x))\, \gamma'(x)}{\sqrt{1 + \gamma'(x)^2}} + n_y(x, \gamma(x)) \sqrt{1 + \gamma'(x)^2} = 0.$$

In this case, however, the discontinuous coefficient n may limit the regularity of an optimal path. Again, for piecewise constant refractive index, an optimal curve will be piecewise linear with discontinuous derivative at optical interfaces. △

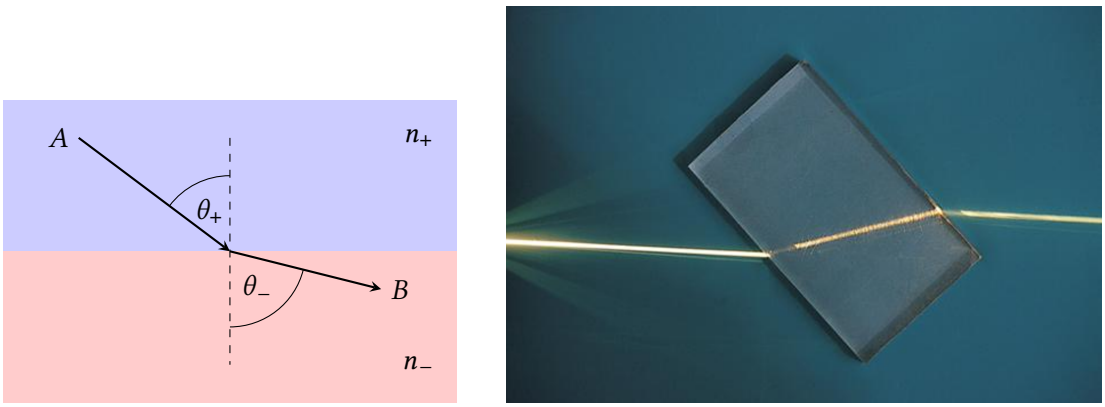


Figure 1.2: Illustratrion of Snell's law of refraction (left) as a special case of Example 1.3. Image (right) obtained from https://en.wikipedia.org/wiki/Refraction, released into the public domain by creator ajizai.

End of Class 1

Example 1.4 (signal denoising)**.**

Suppose a signal $s\colon [0,T] \to \mathbb{R}$ is given.[4] In case the signal is noisy, we may formulate an optimization problem to try and find a denoised signal $y\colon [0,T] \to \mathbb{R}$:

$$\text{Minimize} \quad f(y) := \int_0^T |y(t) - s(t)|^2 \,\mathrm{d}t + \beta \int_0^T |\dot{y}(t)|^2 \,\mathrm{d}t, \quad \text{where } y \in X. \tag{1.5}$$

The dot denotes the time derivative. A suitable function space for this problem is the Sobolev space $X = H^1(0,T)$.

The second term in the objective penalizes "fast variations" in the signal. The parameter $\beta > 0$ balances the two summands in the objective and thus determines the degree of denoising.

We will be able to show later that the first-order optimality conditions for (1.5) involve the second-order differential equation

$$-\beta\, \ddot{y}(t) + y(t) = s(t), \tag{1.6}$$

which shows that the minimizer will indeed be a smoothed version of the noisy signal s. More precisely, we can expect the solution to gain two orders of differentiation compared to the data s. In particular, the solution will not admit any discontinuties. Therefore, one often prefers a "less powerful" regularization term, such as the **total variation** of the function y. We will come back to this type of problem in the context of image denoising problems in **??**. △

Example 1.5 (crane trolley optimal control problem)**.**
Consider a load on rope of length ℓ hanging from a crane trolley system (Figure 1.3). We denote the position of the trolley relative to the origin by s. The position of the load relative to the trolley is denoted by z. The trolley has mass M and the load has mass m. A controllable force u acts on the trolley.

This system is described by a second-order differential equation for the positions (s,z). It can be derived by working out Newton's law, force equals mass times acceleration. We convert it here to a first-order system of differential equations in terms of $x = (s, \dot{s}, z, \dot{z})$, where the dot denotes the time derivative. Assuming small angles θ, the differential equations can be taken as linear and the system reads

$$\begin{pmatrix} \dot{s} \\ \ddot{s} \\ \dot{z} \\ \ddot{z} \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{m}{M}\frac{g}{\ell} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{m+M}{M}\frac{g}{\ell} & 0 \end{bmatrix}}_{=:A} \begin{pmatrix} s \\ \dot{s} \\ z \\ \dot{z} \end{pmatrix} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ \frac{1}{M} \end{bmatrix}}_{=:B} u \tag{1.7}$$

or, in short, $\dot{x} = A\,x + B\,u$. Notice that we have omitted the (t) argument everywhere for brevity.

We wish to steer the system from an initial state $x(0) = (0,0,0,0)^\mathsf{T}$ to a terminal state $x(T) = (E,0,0,0)^\mathsf{T}$ in as short a time T as possible. This leads us to the preliminary optimization problem

$$\begin{aligned} \text{Minimize} \quad & \int_0^T 1 \,\mathrm{d}t, \quad \text{where } (u,x,T) \in U \times X \times \mathbb{R} \\ \text{s.t.} \quad & \dot{x} = A\,x + B\,u \quad \text{in } [0,T] \\ \text{and} \quad & x(0) = (0,0,0,0)^\mathsf{T} \\ \text{and} \quad & x(T) = (E,0,0,0)^\mathsf{T} \\ \text{as well as} \quad & T > 0. \end{aligned} \tag{1.8}$$

[4] Think, for instance, of an audio signal sampled with a certain frequency, say, 48 kHz into a piecewise constant function.

This preliminary problem formulation has some issues. Due to the terminal time T being an optimization variable, we cannot fix function spaces for the **control** u and the **state** x since they depend on T.

There is, however, an easy remedy to this. We can renormalize the unknown time interval $[0, T]$ to the fixed interval $[0, 1]$. Replacing the unknowns x and u by their counterparts on the fixed interval, the dynamics need to be rescaled and the problem becomes

$$\begin{aligned} \text{Minimize} \quad & \int_0^1 T \, \mathrm{d}t, \quad \text{where } (u, x, T) \in U \times X \times \mathbb{R} \\ \text{s. t.} \quad & \dot{x} = \frac{1}{T}(A\,x + B\,u) \quad \text{in } [0, 1] \\ \text{and} \quad & x(0) = (0, 0, 0, 0)^\mathsf{T} \\ \text{and} \quad & x(1) = (E, 0, 0, 0)^\mathsf{T} \\ \text{as well as} \quad & T > 0. \end{aligned} \tag{1.9}$$

We can now fix suitable function spaces[5], e. g., $U = L^2(0, 1)$ and $X = H^1(0, 1)^4$. A problem such as (1.9), in which a **state** function x depends on the choice of the **control** function u through a differential equation, is termed an **optimal control problem**. We will see more of these in Chapter 1.

Unfortunately, problem (1.9) as stated will not have a solution. (**Quiz 1.2:** Can you see why?) We may fix this by imposing bounds on the control function, e. g., by adding the pointwise inequality constraints

$$u(t) \in [-u_{\max}, u_{\max}],$$

with some $u_{\max} > 0$ to problem (1.9), or by adding a cost term such as

$$\beta \int_0^1 |u(t)| \, \mathrm{d}t$$

to the objective. △

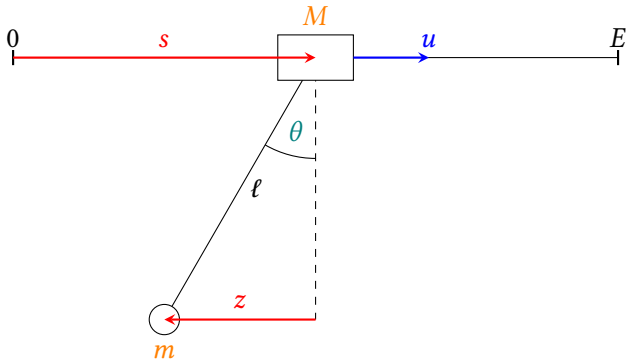


Figure 1.3: Illustration of the crane trolley problem (Example 1.5).

[5]Again, we will introduce these Lebesgue and Sobolev spaces later; see §§ 2.5 and 2.6.

§ 2 Normed Linear Spaces

In this section we recap the notion of a normed linear space. We will also introduce Lebesgue and Sobolev spaces as our prime examples of normed linear spaces.

Definition 2.1 (linear space)**.**
An algebraic structure $(V, +, \cdot)$ with two operations[6]

$$\begin{aligned} +&: V \times V \to V \quad \text{(addition)} \\ \cdot&: \mathbb{R} \times V \to V \quad \text{(S-multiplication)} \end{aligned}$$

is said to be a **linear space** over the field of real numbers $\mathbb{R}$ if

(*i*) $(V, +)$ is an Abelian group.

(*ii*) The S-multiplication satisfies the mixed distributive laws

$$\begin{aligned} \alpha\,(u + v) &= (\alpha\,u) + (\alpha\,v) \\ (\alpha + \beta)\,v &= (\alpha\,v) + (\beta\,v) \end{aligned}$$

as well as the mixed associative law

$$(\alpha\,\beta)\,v = \alpha\,(\beta\,v)$$

for all $\alpha, \beta \in \mathbb{R}$ and $u, v \in V$. Moreover, the neutral element $1 \in \mathbb{R}$ w.r.t. multiplication in $\mathbb{R}$ is also neutral w.r.t. S-multiplication:

$$1\,v = v.$$

△

All linear spaces will be over the field of real numbers $\mathbb{R}$ and we will not explicitly mention that. We already anticipated that in order to be able to differentiate functions $f : V \to \mathbb{R}$ or, more generally, $f : V \to W$, we will require linear spaces to be normed.

Definition 2.2 (normed linear space)**.**
Suppose that V is a linear space.

(*i*) A map $\|\cdot\| : V \to \mathbb{R}$ is said to be a **norm on** V if the following conditions hold:

$$\|u\| \geqslant 0, \quad \text{and } \|u\| = 0 \implies u = 0 \qquad \textbf{positive definiteness} \tag{2.1a}$$

$$\|\alpha\,u\| = |\alpha|\,\|u\| \qquad \textbf{absolute homogeneity} \tag{2.1b}$$

$$\|u + v\| \leqslant \|u\| + \|v\| \qquad \textbf{triangle inequality} \text{ or } \textbf{subadditivity} \tag{2.1c}$$

for all $u, v \in V$ and all $\alpha \in \mathbb{R}$.

(*ii*) The pair $(V, \|\cdot\|)$ is said to be a **(real) normed linear space** or **normed vector space**. △

[6] The dot $\cdot$ for S-multiplication is usually not written, just as the multiplication symbol in $\mathbb{R}$ is usually not written.

Expert Knowledge: from topological to normed linear spaces

We have the inclusions

- Every normed linear space is a metric space.
- Every metric space is a topological space.

A topological space is defined by a collection of its subsets that are called the open sets. Topological spaces admit notions of convergence and limits, closure and compactness of sets, as well as notions of continuity of functions.

Metric spaces are spaces with a notion of distance. The metric induces a topology.

Normed spaces are spaces with a notion of length. The norm induces a metric.

We will not discuss general topological spaces in full generality but restrict ourselves to normed linear spaces.

§ 2.1 Open and Closed Sets

Definition 2.3 (balls, spheres, open sets, closed sets)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space.

(*i*) For $\varepsilon > 0$, the set

$$B_\varepsilon(x) := \{y \in V \mid \|y - x\| < \varepsilon\}$$

is said to be the **open ε-ball** about x of radius ε. In particular, $B_1(0)$ is termed the **open unit ball**.

(*ii*) A point $x \in E$ of a subset $E \subseteq V$ is said to be an **interior point** of E if there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq E$. The subset of interior points of E is called the **interior** of E and it is denoted by $\operatorname{int} E$.

(*iii*) A set $U \subseteq V$ is said to be **open** if every $x \in U$ is an interior point of U, i. e., if $\operatorname{int} U = U$.

(*iv*) A set $A \subseteq V$ is said to be **closed** if its complement $V \setminus A$ is open.

(*v*) For $\varepsilon > 0$, the set

$$\overline{B_\varepsilon(x)} := \{y \in V \mid \|y - x\| \leqslant \varepsilon\}$$

is said to be the **closed ε-ball** about x of radius ε. In particular, $\overline{B_1(0)}$ is termed the **closed unit ball**.

(*vi*) The **closure** of a subset $E \subseteq V$ is

$$\operatorname{cl} E := \bigcap \big\{A \subseteq V \,\big|\, A \text{ is closed and } E \subseteq A\big\}. \tag{2.2}$$

(*vii*) The **boundary** of a subset $E \subseteq V$ is $\partial E := \operatorname{cl} E \setminus \operatorname{int} E$, i. e., the closure minus the interior of E.

(*viii*) The set

$$\partial B_\varepsilon(x) := \{y \in V \mid \|y - x\| = \varepsilon\}$$

is said to be the **ε-sphere** about x of radius ε. In particular, $\partial B_1(0)$ is termed the **unit sphere** of V. △

It is not difficult to show that the interior of a set is open and the closure of a set is closed. In fact, a set E is open if and only if $E = \operatorname{int} E$, and a set A is closed if and only if $A = \operatorname{cl} A$. Also, a set A is closed if and only if $A = \partial A$. The boundary of a set is also closed. (**Quiz 2.1:** Can you show this?)

The following result was inserted after the class.

Lemma 2.4 (characterization of the closure[7])**.**

Suppose that $(V, \|\cdot\|)$ is a normed linear space and $E \subseteq V$. Then

$$\begin{aligned} \operatorname{cl} E &= \{y \in V \mid \text{for any } \varepsilon > 0 \text{ there exists } x \in E \text{ such that } \|x - y\| < \varepsilon\} \\ &= \{y \in V \mid \text{for any } \varepsilon > 0, B_\varepsilon(y) \cap E \neq \emptyset\} \\ &= \{y \in V \mid \text{there exists a sequence } \left(x^{(k)}\right) \text{ in } E \text{ converging to } y\}. \end{aligned} \tag{2.3}$$

Proof. □

The following lemma **(inserted after the class)** confirms that the nomenclature and symbols related to balls and spheres is meaningful:

Lemma 2.5 (openness, closedness, boundary of balls and spheres)**.**

Suppose that $(V, \|\cdot\|)$ is a normed linear space.

(*i*) Open balls $B_\varepsilon(x)$ are open sets.

(*ii*) Closed balls $\overline{B_\varepsilon(x)}$ are closed sets.

(*iii*) Open balls and closed balls are related via

$$\overline{B_\varepsilon(x)} = \operatorname{cl} B_\varepsilon(x) \quad \text{and} \quad B_\varepsilon(x) = \operatorname{int} \overline{B_\varepsilon(x)}. \tag{2.4}$$

(*iv*) Spheres and balls are related via

$$\partial B_\varepsilon(x) = \partial\big(B_\varepsilon(x)\big) = \partial\big(\overline{B_\varepsilon(x)}\big). \tag{2.5}$$

Proof. □

End of Class 2

End of Week 1

[7] We can read this result as "The closure of a set E consists of the **accumulation points** of E."

§ 2.2 Banach Spaces

Since norms furnish a linear space with a topology, they also bring about a notion of convergence.

Definition 2.6 (convergent sequence, Cauchy sequence)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space.

(*i*) A sequence[8] $\left(x^{(k)}\right)$ in V is said to **converge to** $x \in V$ in case $\|x^{(k)} - x\| \to 0$ in $\mathbb{R}$. We then write $x^{(k)} \to x$ or $\lim_{k\to\infty} x^{(k)} = x$ and call x a **limit point** or **limit** of the sequence $\left(x^{(k)}\right)$.
In other words, $x^{(k)} \to x$ means: for every $\varepsilon > 0$ there exists an index k_ε such that $\|x^{(k)} - x\| < \varepsilon$ holds for all $k \geqslant k_\varepsilon$.

(*ii*) A sequence $\left(x^{(k)}\right)$ in V is said to **converge** if there exists some $x \in V$ such that $x^{(k)} \to x$.

(*iii*) A sequence $\left(x^{(k)}\right)$ in V is said to be a **Cauchy sequence** in V if, for every $\varepsilon > 0$, there exists an index k_ε such that $\|x^{(k)} - x^{(\ell)}\| < \varepsilon$ holds for all $k, \ell \geqslant k_\varepsilon$. △

Lemma 2.7 (properties of convergent sequences)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space and that $\left(x^{(k)}\right)$ is a sequence in V.

(*i*) Suppose that $\left(x^{(k)}\right)$ converges. Then its limit is unique.

(*ii*) Suppose that $\left(x^{(k)}\right)$ converges. Then it is a Cauchy sequence.

Proof. This proof is addressed in homework problem 2.3. □

The converse of statement (*ii*) is not true in general. Therefore, spaces in which it is true deserve special mention:

Definition 2.8 (complete normed linear space, Banach space, complete subset)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space.

(*i*) The space $(V, \|\cdot\|)$ is said to be **complete** or a **Banach space** if every Cauchy sequence in V converges.

(*ii*) A subset $A \subseteq V$ is said to be **complete** if every Cauchy sequence in A converges to a limit in A. △

The following result was inserted after the class.

Lemma 2.9 (in Banach spaces, completeness is closedness)**.**
Suppose that $(V, \|\cdot\|)$ is a Banach space. The $A \subseteq V$ is complete if and only if A is closed.

Proof. This proof is addressed in homework problem 2.2. □

The following result was inserted after the class.

[8] The exact index set of a sequence does not matter. We will allow any interval of the integers $\mathbb{Z}$ which is bounded below but not bounded above. In other words, any subset of $\mathbb{Z}$ of the form $\{k_0, k_0 + 1, k_0 + 2, \ldots\}$.

Lemma 2.10 (complete sets are closed)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space and $E \subseteq V$. If E is complete, then E is closed.

Proof. Suppose that $\left(x^{(k)}\right)$ is a sequence in E converging to some $x \in V$. Then this sequence is a Cauchy sequence in E. Since E is complete, $\left(x^{(k)}\right)$ converges to a limit $y \in E$. By uniqueness of the limit, we have $x = y \in E$. By the characterization (2.3) of the closure, we have $E = \operatorname{cl} E$. □

§ 2.3 Comparison of Norms

We wish to be able to compare two different norms on the same linear space. The following definition allows us to do that.

Definition 2.11 (partial ordering of norms)**.**
Suppose that V is a linear space and that $\|\cdot\|_a$ and $\|\cdot\|_b$ are two norms on V.

(*i*) The norm $\|\cdot\|_a$ is said to be **weaker** than the norm $\|\cdot\|_b$ if there exists a constant $c > 0$ such that

$$\|x\|_a \leqslant c\,\|x\|_b \quad \text{holds for all } x \in V. \tag{2.6}$$

In this case, we also say that $\|\cdot\|_b$ is **stronger** than $\|\cdot\|_a$. We write $\|\cdot\|_a \lesssim \|\cdot\|_b$ or $\|\cdot\|_b \gtrsim \|\cdot\|_a$.

(*ii*) The norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are said to be **equivalent** if both $\|\cdot\|_a \lesssim \|\cdot\|_b$ and $\|\cdot\|_b \lesssim \|\cdot\|_a$ hold, i. e., if there exist constants $c_1, c_2 > 0$ such that

$$c_1\,\|x\|_a \leqslant \|x\|_b \leqslant c_2\,\|x\|_a \quad \text{holds for all } x \in V. \tag{2.7}$$

△

The following result was corrected.

Lemma 2.12 (openness, closedness, completeness and the Cauchy property are preserved under weaker norms)**.**
Suppose that V is a linear space and that $\|\cdot\|_a$ and $\|\cdot\|_b$ are two norms on V such that $\|\cdot\|_a \lesssim \|\cdot\|_b$. Then the following hold:

(*i*) For any open ball $B_\varepsilon^{\|\cdot\|_a}(x)$ in the weaker norm $\|\cdot\|_a$, there exists an open ball $B_\delta^{\|\cdot\|_b}(x)$ in the stronger norm $\|\cdot\|_b$ such that $B_\delta^{\|\cdot\|_b}(x) \subseteq B_\varepsilon^{\|\cdot\|_a}(x)$.

(The stronger norm has the smaller balls and more open sets.)

(*ii*) If $U \subseteq V$ is open in the weaker norm $\|\cdot\|_a$, then U is open in the stronger norm $\|\cdot\|_b$.

(The stronger norm defines the finer topology.)

(*iii*) If $A \subseteq V$ is closed in the weaker norm $\|\cdot\|_a$, then A is closed in the stronger norm $\|\cdot\|_b$.

(*iv*) If $E \subseteq V$ is bounded in the stronger norm $\|\cdot\|_b$, then E is bounded in the weaker norm $\|\cdot\|_a$.

(*v*) If $K \subseteq V$ is totally bounded in the stronger norm $\|\cdot\|_b$, then K is totally bounded in the weaker norm $\|\cdot\|_a$.

(*vi*) If $K \subseteq V$ is compact in the stronger norm $\|\cdot\|_b$, then K is compact in the weaker norm $\|\cdot\|_a$.

(*vii*) If $\left(x^{(k)}\right)$ converges in the stronger norm $\|\cdot\|_b$, then $\left(x^{(k)}\right)$ converges in the weaker norm $\|\cdot\|_a$ (to the same limit point).

(*viii*) If $\big(x^{(k)}\big)$ is a Cauchy sequence in the stronger norm $\|\cdot\|_b$, then $\big(x^{(k)}\big)$ is a Cauchy sequence in the weaker norm $\|\cdot\|_a$.

Proof. This proof is addressed in homework problem 3.1. □

Theorem 2.13 (in finite-dimensional normed linear spaces, all norms are equivalent)**.**
Suppose that V is a finite-dimensional linear space. If $\|\cdot\|_a$ and $\|\cdot\|_b$ are two norms on V, then $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent.

Proof. Suppose that $\{v^{(1)},\dots,v^{(n)}\}$ is a basis of V. Then every $x \in V$ can be uniquely written as $x = \sum_{j=1}^{n} x_j\, v^{(j)}$. The map $x \mapsto \|x\|_\infty := \left\|\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right\|_\infty = \max\big\{|x_1|,\dots,|x_n|\big\}$ is a norm on V.

It is enough to prove that the norms $\|\cdot\|_a$ and $\|\cdot\|_\infty$ are equivalent norms on V since equivalence of norms is an equivalence relation.

Step 1: To show $\|\cdot\|_a \lesssim \|\cdot\|_\infty$, we estimate:

$$\begin{aligned}\|x\|_a &= \Big\|\sum_{j=1}^{n} x_j\, v^{(j)}\Big\|_a \\ &\leqslant \sum_{j=1}^{n} |x_j|\, \|v^{(j)}\|_a \\ &\leqslant \|x\|_\infty \sum_{j=1}^{n} \|v^{(j)}\|_a \\ &=: c\, \|x\|_\infty.\end{aligned}$$

Step 2: We show that $\|\cdot\|_\infty \lesssim \|\cdot\|_a$.

Suppose that this is not the case. Then there exists a sequence $\big(x^{(k)}\big)$ in V such that $\|x^{(k)}\|_\infty > k\, \|x^{(k)}\|_a$. We can assume that $\|x^{(k)}\|_\infty = 1$ holds. (**Quiz 2.2:** Why?)

On the other hand, for all $j = 1,\dots,n$, the j-th coefficients $\big\{x_j^{(k)} \,\big|\, k \in \mathbb{N}\big\}$ belong to the compact interval $[-1,1]$. Therefore, we can find a subsequence $x^{(k^{(\ell)})}$ such that $x_j^{(k^{(\ell)})}$ converges to some x_j^* for all $j = 1,\dots,n$. Moreover, for at least one index $j_0 \in \{1,\dots,n\}$, we have $\big|x_{j_0}^{(k^{(\ell)})}\big| = 1$ for infinitely many indices $\ell \in \mathbb{N}$. We pass to this subsequence without re-labeling it. This shows $|x_{j_0}^*| = 1$ by continuity of the absolute value function.

We define $x^* := \sum_{j=1}^{n} x_j^*\, v^{(j)}$. The estimate

$$\begin{aligned}\|x^*\|_a &\leqslant \|x^* - k^{(\ell)}\|_a + \|k^{(\ell)}\|_a \\ &\leqslant c\, \|x^* - k^{(\ell)}\|_\infty + \frac{1}{k^{(\ell)}} \qquad \text{by step 1} \\ &\to 0 + 0 \quad \text{as } \ell \to \infty\end{aligned}$$

shows $x^* = 0$, i. e., all coefficients x_j^* are zero. This contradicts $|x_{j_0}^*| = 1$. □

Note: As a consequence of this theorem, we do not necessarily need to specify the norm when we talk about a finite-dimensional linear space. In particular, all norms on $\mathbb{R}$ are equivalent, with the absolute value $|\cdot|$ as the standard norm.

As a consequence of Theorem 2.13, we can show:

Lemma 2.14 (finite-dimensional subspaces are complete and thus closed)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space. Every finite-dimensional subspace $Y \subseteq V$ is complete and thus closed.

Proof. Suppose that $\{y^{(1)}, \dots, y^{(n)}\}$ is a basis of Y. By Theorem 2.13, the norms $\|\cdot\|$ and $\|\cdot\|_\infty$ are equivalent on Y, where $\|x\|_\infty := \max\{|x_1|, \dots, |x_n|\}$ when $x = \sum_{j=1}^{n} x_j y^{(j)}$.

Suppose now that $\left(x^{(k)}\right)$ is a Cauchy sequence in Y. The elements of $\left(x^{(k)}\right)$ have a representation

$$x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} y^{(j)}.$$

Then for any $j = 1, \dots, n$, the sequence $\{x_j^{(k)}\}$ is a Cauchy sequence in $(\mathbb{R}, |\cdot|)$. Therefore, $x_j^{(k)} \to x_j^*$ for some $x_j^* \in \mathbb{R}$. We thus obtain

$$x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} y^{(j)} \to \sum_{j=1}^{n} x_j^* y^{(j)} \in Y.$$

This shows that $\left(x^{(k)}\right)$ converges in Y. Therefore, Y is a complete subset of V and thus closed by Lemma 2.10. □

Note: In particular, if V itself is finite-dimensional, then it is complete and thus closed.

§ 2.4 Compactness

Compactness of sets plays a major role in topology, analysis, and also in optimization.

Definition 2.15 (compact, sequentially compact and totally bounded sets)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space and $E \subseteq V$ is some subset.

(*i*) A collection $\left(U_i\right)_{i \in I}$ of open subsets $U_i \subseteq V$ is said to be an **open cover** of E if $E \subseteq \bigcup_{i \in I} U_i$ holds.

(*ii*) A subset $K \subseteq V$ is said to be **compact** if every open cover $\left(U_i\right)_{i \in I}$ of K contains a finite subcover, i. e., there exist a finite number of indices $i_1, \dots, i_N \in I$ such that $K \subseteq \bigcup_{j=1}^{N} U_{i_j}$.

(*iii*) A subset $K \subseteq V$ is said to be **sequentially compact** if every sequence $\left(x^{(k)}\right)$ in K contains a convergent subsequence whose limit belongs to K.[9]

[9] Stated equivalently, $\left(x^{(k)}\right)$ has an accumulation point in K.

(*iv*) A subset $K \subseteq V$ is said to be **totally bounded** if for any $\varepsilon > 0$, there exist finitely many $x^{(1)}, \dots, x^{(N)} \in K$ such that $\{B_\varepsilon(x^{(1)}), \dots, B_\varepsilon(x^{(N)})\}$ covers K. △

The verification of compactness via Definition 2.15 (*ii*) can be cumbersome. The following results can help.

Lemma 2.16 (compact sets are closed and bounded)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space and $K \subseteq V$ is a compact subset. Then K is closed and bounded.

Proof. We prove both properties independently.

Step 1: We show that K is closed.

The statement is true when $K = V$ (**Quiz 2.3:** Is it clear to you?), so suppose $K \subsetneq V$ from now on. Suppose that $z \in V \setminus K$ is a point of the complement of K. We need to show that there exists an open ball $B_\varepsilon(z) \subseteq V \setminus K$.

For any $x \in K$, define $\varepsilon_x := \frac{1}{2}\|x - z\|$. In view of $z \notin K$ and the positive definiteness of the norm, we have $\varepsilon_x > 0$. The open balls $B_{\varepsilon_x}(x)$ and $B_{\varepsilon_x}(z)$ are disjoint since for any point y in their intersection, the triangle inequality would imply the contradiction

$$\|x - z\| \leqslant \|x - y\| + \|y - z\| < \varepsilon_x + \varepsilon_x = \|x - z\|.$$

The sets $\{B_{\varepsilon_x}(x) \mid x \in K\}$ form an open cover of K. Since K is compact, finitely many of these suffice, say, those with center points $x^{(1)}, \dots, x^{(N)} \in K$. As we noticed above, $B_{\varepsilon_{x^{(j)}}}(x^{(j)})$ and $B_{\varepsilon_{x^{(j)}}}(z)$ are disjoint for all $j = 1, \dots, N$. Let $\varepsilon := \min\{\varepsilon_{x^{(1)}}, \dots, \varepsilon_{x^{(N)}}\}$. Then $B_\varepsilon(z)$ is disjoint from all $B_{\varepsilon_{x^{(j)}}}(x^{(j)})$ and hence from K.

Step 2: We show that K is bounded.

Fix $x \in V$ arbitrarily and consider the open balls $\{B_i(x) \mid i \in \mathbb{N}\}$. Since every element of K has a finite distance from the point x, this collection of open balls covers K. Since K is compact, a finite number of these suffice, say,

$$\{B_{i^{(1)}}(x), \dots, B_{i^{(N)}}(x)\}.$$

These being balls with the same center, one of them is largest, say, $B_{i^{(*)}}(x)$, which alone covers K. □

Theorem 2.17 (in normed linear spaces, the notions of compact and sequentially compact sets coincide)**.**

Suppose that $(V, \|\cdot\|)$ is a normed linear space and $K \subseteq V$ is some subset. Then the following statements are equivalent:

(*i*) K is compact.

(*ii*) K is sequentially compact.

(*iii*) K is complete and totally bounded.

Proof. Statement (i) $\Rightarrow$ statement (ii): Suppose that $\left(x^{(n)}\right)$ is a sequence in K that does not possess a convergent subsequence with limit in K. In other words, $\left(x^{(n)}\right)$ does not have an accumulation point in K. Therefore, for any $x \in K$, there exists $\varepsilon_x > 0$ such that $x^{(k)} \in B_x(\varepsilon_x)$ holds only for finitely many indices k. The sets $\{B_{\varepsilon_x}(x) \mid x \in K\}$ form an open cover of K. By the compactness of K, there exists a finite subcover

$$\left\{B_{\varepsilon_{x^{(1)}}}(x^{(1)}), \ldots, B_{\varepsilon_{x^{(N)}}}(x^{(N)})\right\}$$

of K. By construction, $x^{(k)} \in B_{\varepsilon_{x^{(i)}}}(x^{(i)})$ holds only for finitely many indices k. That is, $x^{(k)} \in \bigcup_{i=1}^{N} B_{\varepsilon_{x^{(i)}}}(x^{(i)})$ also holds only for finitely many indices k. Therefore, finally, $x^{(k)} \in K$ also holds only for finitely many indices k. This contradicts $\left(x^{(n)}\right)$ being a sequence in K.

Statement (ii) $\Rightarrow$ statement (iii): Suppose now that K is sequentially compact. Then, by definition, every sequence in K contains a convergent subsequence whose limit belongs to K. In particular, this is true for any Cauchy sequence in K, hence K is complete.

To show that K is totally bounded, suppose that $\varepsilon > 0$. If $K = \emptyset$, nothing is to be done, so suppose $K \neq \emptyset$. Pick a point $x^{(1)} \in K$. In case $K \subseteq B_\varepsilon(x^{(1)})$, we are done. Otherwise, pick a point $x^{(2)} \in K \setminus B_\varepsilon(x^{(1)})$. In case $K \subseteq B_\varepsilon(x^{(1)}) \cup B_\varepsilon(x^{(2)})$, we are done. Otherwise, continue in the same way. If this process produced an infinite sequence $\left(x^{(k)}\right)$, its members would satisfy $\|x^{(k)} - x^{(\ell)}\| \geqslant \varepsilon$ for all $k \neq \ell$. Therefore, this sequence in K cannot have a convergent subsequence, contradicting the assumption that K is sequentially compact. Consequently, the process above terminates after finitely many steps, showing $K \subseteq \bigcup_{i=1}^{N} B_{\varepsilon_{x^{(i)}}}(x^{(i)})$. That is, K is totally bounded.

Statement (iii) $\Rightarrow$ statement (i): We proceed by contradiction. Suppose that $\left(U_i\right)_{i \in I}$ is an open cover of K that does not possess a finite subcover.

Since K is totally bounded, K can be covered by a finite number of open balls of radius 1 with centers in K. For at least one of these, say, $B_1(x^{(0)})$, the intersection $B_1(x^{(0)}) \cap K$ cannot be covered by a finite subfamily of $\left(U_i\right)_{i \in I}$. (Otherwise, K itself could be covered by a finite subfamily of $\left(U_i\right)_{i \in I}$, which we assumed is not the case.)

Now consider $B_1(x^{(0)}) \cap K$. As a subset of K, this set is again totally bounded and thus can be covered by a finite number of open balls of radius $1/2$ with centers in $B_1(x^{(0)}) \cap K$. Again, for at least one of these, say, $B_{1/2}(x^{(2)})$, the intersection $B_{1/2}(x^{(2)}) \cap K$ cannot be covered by a finite subfamily of $\left(U_i\right)_{i \in I}$. (Otherwise, $B_1(x^{(0)}) \cap K$ itself could be covered by a finite subfamily of $\left(U_i\right)_{i \in I}$, which we know is not the case.)

Repeating this process, we obtain a sequence of balls $B_{2^{-k}}(x^{(k)})$, for none of which $B_{2^{-k}}(x^{(k)}) \cap K$ is covered by a finite subfamily of $\left(U_i\right)_{i \in I}$. The centers satisfy $x^{(k+1)} \in B_{2^{-k}}(x^{(k)}) \cap K$. Therefore, the sequence $\left(x^{(k)}\right)$ is a Cauchy sequence in K since $\|x^{(k)} - x^{(\ell)}\| < 2^{1-k}$ holds for all $\ell \geqslant k$. (**Quiz 2.4:** Can you fill in the details?) Since K was assumed to be a complete subset of V, this Cauchy sequence converges and its limit x^* belongs to K.

This implies that x^* belongs to some member of the family $\left(U_i\right)_{i \in I}$, say, $x \in U_{i^*}$. Since U_{i^*} is open, there exists $\varepsilon > 0$ such that

$$x^* \in B_\varepsilon(x^*) \subseteq U_{i^*}$$

holds. We can find an index $N \in \mathbb{N}$ such that $2^{-N} < \varepsilon/2$ and

$$\|x^{(N)} - x^*\| < \frac{\varepsilon}{2}$$

holds. Consequently, for any $y \in B_{2^{-N}}(x^{(N)})$, we have

$$\|y - x^*\| \leqslant \|y - x^{(N)}\| + \|x^{(N)} - x^*\| < 2^{-N} + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which means that we have

$$B_{2^{-N}}(x^{(N)}) \subseteq B_\varepsilon(x^*) \subseteq U_{i^*}.$$

This, however, contradicts the fact that for none of the balls $B_{2^{-k}}(x^{(k)})$, the intersection $B_{2^{-k}}(x^{(k)}) \cap K$ can be covered by a finite subfamily of $(U_i)_{i\in I}$.

Consequently, the assumption that there exists an open cover $(U_i)_{i\in I}$ of K that does not possess a finite subcover, cannot be true. This shows that K is compact. □

The notion of compactness is very strong in infinite-dimensional normed linear spaces. As a consequence, only "few" sets are compact.

Theorem 2.18 (compactness of the unit ball)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space. Then the following statements are equivalent:

(*i*) The closed unit ball $\overline{B_1(0)}$ is compact.

(*ii*) The unit sphere $\partial B_1(0)$ is compact.

(*iii*) $\dim(V)$ is finite.

Notice that this theorem holds independently of which particular norm is chosen on the linear space V!

The proof of Theorem 2.18 uses the following result:

Lemma 2.19 (**Riesz lemma**)**.**
Suppose that $(V, \|\cdot\|)$ is a normed linear space. Moreover, let $Y \subsetneq V$ be a closed proper subspace of V. Then for any $\theta \in (0, 1)$, there exists $x_\theta \in V$ of unit norm $\|x_\theta\| = 1$ such that

$$\theta \leqslant \|x_\theta - y\| \quad \text{for all } y \in Y. \tag{2.8}$$

Note: Read this as: "You can find a vector x_θ on the unit sphere that is at least the distance θ away from any point in the subspace Y." This result is sometimes written equivalently as[10]

$$\theta \leqslant \operatorname{dist}_Y(x_\theta) \leqslant 1.$$

Proof. Pick any $v \in V \setminus Y$ and define $R := \inf\{\|v - y\| \,|\, y \in Y\}$. By Lemma 2.4, $\operatorname{dist}_Y(x) = 0$ if and only if $x \in \operatorname{cl} Y$. Therefore, we have $R = \operatorname{dist}_Y(v) > 0$. Due to $\theta < 1$, we can find $y_\theta \in Y$ such that

$$0 < \|v - y_\theta\| \leqslant \frac{R}{\theta} \tag{2.9}$$

holds. We define

$$x_\theta := \frac{v - y_\theta}{\|v - y_\theta\|}.$$

[10]The **distance** of a point x to a set Y in a normed linear space is defined as $\operatorname{dist}_Y(x) := \inf\{\|x - y\| \,|\, y \in Y\}$.

Then we have $\|x_\theta\| = 1$ and, for any $y \in Y$,

$$\begin{aligned}\|x_\theta - y\| &= \left\| \frac{v - y_\theta}{\|v - y_\theta\|} - y \right\| \\ &= \frac{1}{\|v - y_\theta\|} \big\| v - \underbrace{\big(y_\theta + \|v - y_\theta\|\, y\big)}_{\in Y} \big\| \\ &\geqslant \frac{R}{\|v - y_\theta\|}.\end{aligned}$$

Together with (2.9), this proves (2.8). □

End of Class 4

Proof of Theorem 2.18:
Item *(i)* ⇒ item *(iii)*: When the closed unit ball $\overline{B_1(0)}$ is compact, then it is also totally bounded by Theorem 2.17. Thus, it can be covered by finitely many balls of radius 1/2:

$$\overline{B_1(0)} \subseteq \bigcup_{i=1}^{N} B_{1/2}(y^{(i)}).$$

Define $Y := \operatorname{span}\{y^{(1)}, \dots, y^{(N)}\}$. Then by Lemma 2.14, Y is a closed subspace of V.

Suppose that $Y \subseteq V$ is a *proper* subspace. The Riesz lemma 2.19 then implies that there exists $x_\theta \in V$ of unit norm such that $\operatorname{dist}_Y(x_\theta) \geqslant \theta := \frac{3}{4}$. Moreover, x_θ belongs to one of the covering balls, say, $B_{1/2}(y^{(j)})$. Therefore, we have

$$\operatorname{dist}_Y(x_\theta) \leqslant \|x_\theta - y^{(j)}\| < \frac{1}{2},$$

which contradicts $\operatorname{dist}_Y(x_\theta) \geqslant \frac{3}{4}$. Therefore, $Y = V$ and $\dim(V)$ is finite.

Item *(ii)* ⇒ item *(iii)*: The proof is the same as above.

Item *(iii)* ⇒ item *(i)*: The closed unit ball $\overline{B_1(0)}$ is a clearly a closed subset of V. Suppose that $\dim(V) = n \in \mathbb{N}_0$ and that $\{v^{(1)}, \dots, v^{(n)}\}$ is a basis of V. Then V is complete by Lemma 2.14. When we show that $\overline{B_1(0)}$ is totally bounded w.r.t. $\|\cdot\|$, then it is compact by Theorem 2.17. By the equivalence of norms (Theorem 2.13), we may equivalently show that $\overline{B_1(0)}$ is totally bounded w.r.t. $\|\cdot\|_\infty$.

Suppose that $\|\cdot\|_\infty \leqslant c\,\|\cdot\|$ holds for $c > 0$. Consider $\varepsilon > 0$. We claim that

$$\overline{B_1(0)} \subseteq \overline{B_c^{\|\cdot\|_\infty}(0)} \subseteq \bigcup_{\substack{q \in \varepsilon\mathbb{Z}^n \\ \|q\|_\infty \leqslant c + \varepsilon/2}} B_\varepsilon^{\|\cdot\|_\infty}\Big(\sum_{j=1}^{n} q_j\, v^{(j)}\Big),$$

holds. Notice that the right-hand side is a finite union of open balls of radius ε. The first inequality is clear. For the second inequality, consider a point $x \in \overline{B_c^{\|\cdot\|_\infty}(0)}$, whose coordinates x_j then satisfy $|x_j| \leqslant c$. For $j = 1, \dots, n$, find $q_j \in \varepsilon\mathbb{Z}$ closest to x_j. This implies $|x_j - q_j| \leqslant \varepsilon/2$ and thus

$$\Big\| x - \sum_{j=1}^{n} q_j\, v^{(j)} \Big\|_\infty \leqslant \frac{\varepsilon}{2} < \varepsilon.$$

In other words, x belongs to the open ball $B_\varepsilon^{\|\cdot\|_\infty}\Big(\sum_{j=1}^{n} q_j\, v^{(j)}\Big)$. Due to $|x_j| \leqslant c$, we will have $|q_j| \leqslant c+\varepsilon/2$.

This proves the claim.

Item (iii) $\Rightarrow$ item (ii): The proof is the same as above. □

Note: The proof item (iii) $\Rightarrow$ item (i) can be easily extended to show that every bounded set in a finite-dimensional normed linear space is totally bounded.

Remark 2.20 (there is nothing special about *unit* balls)**.**
For any $r > 0$, the closed ball $\overline{B_r(0)}$ is compact if and only if $\dim(V)$ is finite. The same holds for spheres. △

§ 2.5 Lebesgue Spaces

Literature: Rudin, 1987, Chapter 3

Lebesgue spaces are prominent examples of Banach spaces. All references to a measure will mean the Lebesgue measure on $\mathbb{R}^d$. We will state results in this subsection without proof.

Definition 2.21 (Lebesgue spaces)**.**
Suppose that $\Omega \subseteq \mathbb{R}^n$ is an open set and $p \in [1, \infty)$.

- (i) A measurable function $f\colon \Omega \to \mathbb{R}$ is said to be **Lebesgue integrable of index p** or simply **p-integrable** if $|f|^p$ is integrable on Ω.
- (ii) A measurable function $f\colon \Omega \to \mathbb{R}$ is said to be **essentially bounded** if it is bounded except on a set of measure zero.
- (iii) Two measurable functions $f, g\colon \Omega \to \mathbb{R}$ are said to be **equivalent** if they coincide except on a set of measure zero.
- (iv) The **Lebesgue space** $L^p(\Omega)$ is defined as the set of equivalence classes[11] of measurable functions $f\colon \Omega \to \mathbb{R}$ that are Lebesgue integrable of index p:

$$L^p(\Omega) := \big\{[f] \,\big|\, f\colon \Omega \to \mathbb{R} \text{ is Lebesgue integrable of index } p\big\}. \tag{2.10}$$

- (v) The **Lebesgue space** $L^\infty(\Omega)$ is defined as the set of equivalence classes of measurable functions $f\colon \Omega \to \mathbb{R}$ that are essentially bounded:

$$L^\infty(\Omega) := \big\{[f] \,\big|\, f\colon \Omega \to \mathbb{R} \text{ is essentially bounded}\big\}. \tag{2.11}$$

△

It is customary to denote the equivalence class of a function f by f itself. We will do so from now on.

[11] The construction is that of a quotient space: we begin with the vector space of p-integrable functions and factor out the subspace of functions which are almost everywhere zero. Recall that "**almost everywhere**" means "except on a set of measure zero".

Theorem 2.22 (Lebesgue spaces as Banach spaces).
Suppose that $\Omega \subseteq \mathbb{R}^n$ is an open set.

(*i*) For $p \in [1, \infty)$, the Lebesgue space $L^p(\Omega)$ is a Banach space when equipped with the norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f|^p\right)^{1/p}. \tag{2.12}$$

(*ii*) The Lebesgue space $L^\infty(\Omega)$ is a Banach space when equipped with the norm

$$\|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x\in\Omega} |f(x)| := \inf\big\{M \geqslant 0 \,\big|\, |f(x)| \leqslant M \text{ for almost all } x \in \Omega\big\}. \tag{2.13}$$

(*iii*) For any $p \in [1, \infty]$, the triangle inequality $\|f+g\|_{L^p(\Omega)} \leqslant \|f\|_{L^p(\Omega)} + \|g\|_{L^p(\Omega)}$ for all $f, g \in L^p(\Omega)$ is called the **Minkowski inequality**.

Example 2.23 (functions in L^p).

(*i*) On $\Omega = \mathbb{R}$, non-zero constant functions belong to $L^\infty(\mathbb{R})$ but not to any $L^p(\mathbb{R})$ with $p < \infty$.

(*ii*) On $\Omega = (-1, 1)$, the absolute power function $x \mapsto |x|^\alpha$ belongs to $L^p((-1,1))$ if and only if $\alpha\, p > -1$.[12] For instance, the inverse square root function $x \mapsto 1/\sqrt{|x|} = x^{-1/2}$ belongs to $L^p((-1,1))$ if and only if $p < 2$.

(*iii*) More generally, on the open unit ball $\Omega = B_1(0) \subseteq \mathbb{R}^d$, the function $x \mapsto |x|^\alpha$ belongs to $L^p(\Omega)$ if and only if $\alpha\, p > -d$ holds.[13] △

Lemma 2.24 (Hölder's inequality).
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set. Moreover, let $p, q \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{q} = 1$.[14] For all $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$, the product $f\, g$ belongs to $L^1(\Omega)$, and the estimate

$$\|f\, g\|_{L^1(\Omega)} \leqslant \|f\|_{L^p(\Omega)}\, \|g\|_{L^q(\Omega)} \tag{2.14}$$

holds. Inequality (2.14) is known as **Hölder inequality**.

Lemma 2.25 (comparison of norms on Lebesgue spaces).
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set. For $1 \leqslant p \leqslant q \leqslant \infty$, the space $L^q(\Omega)$ is a subspace of $L^p(\Omega)$ (and a proper subspace if $1 \leqslant p < q \leqslant \infty$). Moreover, the L^q-norm is stronger than the L^p-norm:

$$\|f\|_{L^p(\Omega)} \leqslant |\Omega|^{\frac{q-p}{pq}}\, \|f\|_{L^q(\Omega)} \quad \text{for all } f \in L^q(\Omega), \tag{2.15}$$

where $|\Omega|$ denotes the Lebesgue measure (d-dimensional volume) of Ω. When $q = \infty$, the expression $\frac{q-p}{pq}$ is to be understood as $1/p$ (for $p < \infty$) or as 0 (for $p = \infty$).

Note: Lemma 2.25 states that the higher the index of a Lebesgue space on a bounded domain, the smaller the space and the stronger the norm.

[12] With the convention that $\alpha\, \infty = \infty$ for $\alpha > 0$ and $\alpha\, \infty = -\infty$ for $\alpha < 0$ as well as $0\, \infty = 0$.
[13] Here $|\cdot|_2$ denotes the Euclidean norm on $\mathbb{R}^d$.
[14] Such numbers p, q are called **conjugate exponents**. The convention here is that $1/\infty = 0$ so that 1 and ∞ are conjugate.

Example 2.26 (comparison of norms on Lebesgue spaces).
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set.

(*i*) $\|f\|_{L^1(\Omega)} \leqslant |\Omega|^{1/2}\|f\|_{L^2(\Omega)}$ for all $f \in L^2(\Omega)$.

(*ii*) $\|f\|_{L^2(\Omega)} \leqslant |\Omega|^{1/2}\|f\|_{L^\infty(\Omega)}$ for all $f \in L^\infty(\Omega)$.

(*iii*) $\|f\|_{L^1(\Omega)} \leqslant |\Omega|\|f\|_{L^\infty(\Omega)}$ for all $f \in L^\infty(\Omega)$. △

End of Class 5

End of Week 3

§ 2.6 Sobolev Spaces

Lebesgue spaces are not sufficient to deal with optimization problems whose objective functions involve derivatives of the unknown, as is the case in the Brachistochrone problem (Example 1.1), Fermat's principle in optics (Example 1.3), the signal denoising problem (Example 1.4), and the optimal control example (Example 1.5). Sobolev spaces are the natural setting for such problems. In brief, they consist of functions whose derivatives up to a certain order are in a Lebesgue space. The notion of derivative is meant in a *weak sense*.

Derivatives of multivariate functions are conveniently described using multi-indices.

Definition 2.27 (multi-index).
Suppose $d \in \mathbb{N}$.

(*i*) A **multi-index** of length d is a tuple $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$.

(*ii*) The **order** of a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$ is defined as $|\alpha| := \alpha_1 + \dots + \alpha_d$.

(*iii*) We associate with a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$ the derivative operator $D^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$. The **order** of D^α is defined as $|\alpha|$.

(*iv*) In particular, we have $D^{(0,\dots,0)} = \mathrm{id}$ and

$$D_i := \frac{\partial}{\partial x_i} = D^{(0,\dots,0,1,0,\dots,0)}$$

for $i = 1, \dots, d$. △

Definition 2.28 (function spaces $C^k(\Omega)$, $C_c^k(\Omega)$ and $C^k(\operatorname{cl}\Omega)$).
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set.

(*i*) For $f\colon \Omega \to \mathbb{R}$, the set

$$\operatorname{supp} f := \operatorname{cl}\{x \in \Omega \mid f(x) \neq 0\} \tag{2.16}$$

is called the **support** of f.

(*ii*) For $k \in \mathbb{N}_0$, the set of all ***k*-times continuously partially differentiable functions** on Ω is denoted by $C^k(\Omega)$. This means that all partial derivatives of order $\leqslant k$ exist and are continuous functions on Ω.

Moreover, $C^\infty(\Omega) := \bigcap_{k\in\mathbb{N}_0} C^k(\Omega)$ is the set of all **infinitely often continuously partially differentiable functions** on Ω.

(*iii*) For $k \in \mathbb{N}_0$, the set $C_c^k(\Omega)$ consists of all **functions $f \in C^k(\Omega)$ with compact support**, i. e., $\operatorname{supp} f$ is a compact subset of Ω.

Moreover, $C_c^\infty(\Omega) \coloneqq \bigcap_{k\in\mathbb{N}_0} C_c^k(\Omega)$ is the set of all **infinitely often continuously partially differentiable functions** on Ω **with compact support**.

(*iv*) For $k \in \mathbb{N}_0$, $C^k(\operatorname{cl}\Omega)$ denotes the set of all k-times continuously partially differentiable functions $f\colon \Omega \to \mathbb{R}$ such that all partial derivatives of order $\leqslant k$ extend continuously to $\operatorname{cl}\Omega$. △

Note: Given $k \in \mathbb{N}_0$ and $f \in C^k(\Omega)$, the support of all partial derivatives $D^\alpha f$ of order $|\alpha| \leqslant k$ is contained in the support of f.

Lemma 2.29 (properties of derivatives)**.**

Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set. Moreover, suppose that $\alpha \in \mathbb{N}_0^d$ is a multi-index of order $m \in \mathbb{N}_0$ and $k \geqslant m$. Then the following holds:

(*i*) The derivative operator D^α is well-defined as a map

$$D^\alpha\colon C^k(\Omega) \to C^{k-m}(\Omega).$$

(*ii*) The order of differentiation does not matter, i. e., for any decomposition of the multi-index $\alpha = \beta + \gamma$, we have

$$D^\alpha f = D^\beta(D^\gamma f) = D^\gamma(D^\beta f)$$

for all $f \in C^k(\Omega)$.

Proof. Statement (*i*) follows immediately from the fact that higher-order partial derivatives are derivatives of lower-order partial derivatives. Statement (*ii*) is a consequence of the commutativity of partial derivatives by Schwarz' theorem. □

Example 2.30 (function spaces $C^k(\Omega)$ and $C^k(\operatorname{cl}\Omega)$)**.**

(*i*) For $\Omega = (0,1)$, the function $x \mapsto 1/x$ belongs to $C^\infty(\Omega)$ but not to $C(\operatorname{cl}\Omega)$ since it does not extend continuously to 0.

(*ii*) For $\Omega = (0,1)$, the function $x \mapsto \sqrt{x}$ belongs to $C^\infty(\Omega)$ and to $C(\operatorname{cl}\Omega)$ but not to $C^1(\operatorname{cl}\Omega)$ since the derivative $1/(2\sqrt{x})$ does not extend continuously to 0. △

Lemma 2.31 (integration by parts)**.**

Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set and $f \in C^1(\Omega)$. Then for any $i = 1, \dots, d$, we have

$$\int_\Omega (D_i f)\, g \,\mathrm{d}x = -\int_\Omega f\, (D_i g) \,\mathrm{d}x \quad \text{for all } g \in C_c^1(\Omega). \tag{2.17}$$

Note: The supports of both integrands are compact subsets of Ω and the integrands are continuous, so that the integrals are well-defined.

Proof. Suppose that $C = (a_1, b_1) \times \cdots \times (a_d, b_d) \subseteq \mathbb{R}^d$ is an open and bounded box containing the compact set $\operatorname{supp} g$. Define

$$\Phi(x) := \begin{cases} f(x)\, g(x) & \text{if } x \in \Omega, \\ 0 & \text{if } x \in \mathbb{R}^d \setminus \Omega. \end{cases}$$

Then by the product rule, $\Phi \in C^1(\mathbb{R}^d)$ and $\operatorname{supp} \Phi \subseteq \operatorname{supp} g \subseteq C$ and thus also $\operatorname{supp} D_1\Phi \subseteq C$.

For notational convenience, we consider only the case $i = 1$. By the fundamental theorem of calculus, we have

$$\int_{a_1}^{b_1} D_1\Phi(x_1, x_2, \ldots, x_d)\ \mathrm{d}x_1 = \Phi(b_1, x_2, \ldots, x_d) - \Phi(a_1, x_2, \ldots, x_d)$$

for any $x_2, \ldots, x_d \in \mathbb{R}$. Plugging in the definition of Φ, this amounts to

$$\begin{aligned} &\int_{a_1}^{b_1} (D_1 f)(x_1, x_2, \ldots, x_d)\, g(x_1, x_2, \ldots, x_d)\ \mathrm{d}x_1 + \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_d)\, (D_1 g)(x_1, x_2, \ldots, x_d)\, \mathrm{d}x_1 \\ &\quad = f(b_1, x_2, \ldots, x_d) \underbrace{g(b_1, x_2, \ldots, x_d)}_{=0} - f(a_1, x_2, \ldots, x_d) \underbrace{g(a_1, x_2, \ldots, x_d)}_{=0}. \end{aligned}$$

Notice that the right-hand side is zero since the points where g is being evaluated are outside of $\operatorname{supp} g$. We now see that

$$\begin{aligned} &\int_\Omega (D_1 f)\, g\ \mathrm{d}x + \int_\Omega f\, (D_1 g)\ \mathrm{d}x \\ &\quad = \int_C (D_1 f)\, g\ \mathrm{d}x + \int_C f\, (D_1 g)\ \mathrm{d}x \\ &\quad = \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} (D_1 f)(x_1, x_2, \ldots, x_d)\, g(x_1, x_2, \ldots, x_d)\ \mathrm{d}x_1 \cdots \mathrm{d}x_d \\ &\qquad + \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_d)\, (D_1 g)(x_1, x_2, \ldots, x_d)\ \mathrm{d}x_1 \cdots \mathrm{d}x_d \quad \text{by Fubini's theorem} \\ &\quad = 0, \end{aligned}$$

which concludes the proof. □

By induction, we can generalize Lemma 2.31 to higher-order derivatives:

Corollary 2.32 (integration by parts for higher-order derivatives)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set and $f \in C^k(\Omega)$. Then for any multi-index $\alpha \in \mathbb{N}_0^d$ of order $k \in \mathbb{N}_0$, we have

$$\int_\Omega (D^\alpha f)\, g\ \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f\, (D^\alpha g)\ \mathrm{d}x \quad \text{for all } g \in C_c^k(\Omega). \tag{2.18}$$

Formula (2.18) describes properties of classical derivatives for sufficiently smooth functions. These properties serve as a motivation for the definition of a more general notion of derivative, applicable to a much larger class of functions.

Definition 2.33 (weak derivative)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set.

(*i*) For any $A \subseteq \Omega$, the **characteristic function** $\chi_A \colon \Omega \to \mathbb{R}$ is defined as $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ if $x \notin A$.

(*ii*) The set $L^1_{\mathrm{loc}}(\Omega)$ denotes the set of all (equivalence classes of) functions $f \colon \Omega \to \mathbb{R}$ such that $f\,\chi_K \in L^1(\Omega)$ for all compact subsets $K \subseteq \Omega$.

(*iii*) Suppose that $f \in L^1_{\mathrm{loc}}(\Omega)$ and $\alpha \in \mathbb{N}_0^d$ is a multi-index. A function $w \in L^1_{\mathrm{loc}}(\Omega)$ is called the α-th **weak derivative** of f if

$$\int_\Omega f\, D^\alpha \varphi \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega w\, \varphi \, \mathrm{d}x \quad \text{for all } \varphi \in C_c^\infty(\Omega) \tag{2.19}$$

holds. In this case, we write $w = D^\alpha f$. △

Note: For any $\varphi \in C_c^\infty(\Omega)$, both v and $D^\alpha v$ have compact support in Ω. The function class $L^1_{\mathrm{loc}}(\Omega)$ is therefore a natural setting so that the integrals in (2.19) are well-defined.

Remark 2.34 (weak derivative)**.**

(*i*) The α-th weak derivative $D^\alpha f$ of a function $f \in L^1_{\mathrm{loc}}(\Omega)$ is unique (if it exists).

(*ii*) The existence of a weak derivative $D^\alpha f$ does not imply the existence of weak derivatives $D^{\alpha'} f$ for multi-indices $\alpha' \leqslant \alpha$.

(*iii*) If both $D^\alpha f$ and $D^{\alpha+\beta} f$ exist, then $D^{\alpha+\beta} f = D^\beta (D^\alpha f)$.

(*iv*) If both $D^\alpha f$ and $D^\beta (D^\alpha f)$ exist, then $D^{\alpha+\beta} f = D^\beta (D^\alpha f)$. △

Example 2.35 (weak derivative)**.**
The function $f \colon \Omega := (-1, 1) \to \mathbb{R}$ defined by $f(x) = |x|$ has the weak first-order derivative

$$w(x) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0. \end{cases}$$

But f does not have a weak second-order derivative in $L^1_{\mathrm{loc}}(\Omega)$. △

We can now define the Sobolev spaces.

Definition 2.36 (Sobolev spaces)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set. The **Sobolev space** of differentiability index $k \in \mathbb{N}_0$ and index $p \in [1, \infty]$ is defined as

$$W^{k,p}(\Omega) := \left\{ f \in L^p(\Omega) \,\middle|\, D^\alpha f \in L^p(\Omega) \text{ for all } |\alpha| \leqslant k \right\}. \tag{2.20}$$

△

We re-iterate that the elements of a Sobolev space are actually equivalence classes of functions but we continue to use the simplified notation. For $k = 0$, the Sobolev spaces agree with the Lebesgue spaces: $W^{0,p}(\Omega) = L^p(\Omega)$

Remark 2.37 (alternative definition of Sobolev spaces)**.**
For $k \in \mathbb{N}_0$ and $p \in [1, \infty)$, Sobolev spaces can be defined alternatively via a process of completion:

$$H^{k,p}(\Omega) := \mathrm{cl}\big(C^\infty(\Omega) \cap W^{k,p}(\Omega)\big) \text{ w.r.t. the norm } \|\cdot\|_{W^{k,p}(\Omega)}.$$

The paper Meyers, Serrin, 1964 with the title "$H = W$" shows that $H^{k,p}(\Omega) = W^{k,p}(\Omega)$; see also Adams, Fournier, 2003, Theorem 3.17. △

Theorem 2.38 (Sobolev spaces as Banach spaces)**.**
Suppose that $\Omega \subseteq \mathbb{R}^n$ is an open set.

(*i*) For $k \in \mathbb{N}_0$ and $p \in [1, \infty)$, the Sobolev space $W^{k,p}(\Omega)$ is a Banach space when equipped with the norm

$$\|f\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \leqslant k} \int_\Omega |D^\alpha f|^p \right)^{1/p} = \left(\sum_{|\alpha| \leqslant k} \|D^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p}. \tag{2.21}$$

(*ii*) The Sobolev space $W^{k,\infty}(\Omega)$ is a Banach space when equipped with the norm

$$\|f\|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| \leqslant k} \operatorname*{ess\,sup}_{x \in \Omega} |D^\alpha f(x)| = \max_{|\alpha| \leqslant k} \|D^\alpha f\|_{L^\infty(\Omega)}. \tag{2.22}$$

Example 2.39 (Sobolev spaces)**.**
By Example 2.35, the function defined by $f(x) = |x|$ on $\Omega = (-1, 1)$ belongs to $W^{1,\infty}(\Omega)$. However, it does not belong to any $W^{2,p}(\Omega)$ for any $p \in [1, \infty]$. △

End of Class 6

§ 3 Inner Product Spaces

In this section we introduce the notion of an inner product space, which is a concept more specific than a normed linear space.

Definition 3.1 (inner product space)**.**
Suppose that V is a linear space.

(*i*) A map $(\cdot, \cdot) : V \times V \to \mathbb{R}$ is said to be an **inner product on V** if the following conditions hold:

$$(\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 (u_1, v) + \alpha_2 (u_2, v) \quad \textbf{linearity in the first argument} \tag{3.1a}$$
$$(u, \alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 (u, v_1) + \alpha_2 (u, v_2) \quad \textbf{linearity in the second argument} \tag{3.1b}$$
$$(u, v) = (v, u) \quad \textbf{symmetry} \tag{3.1c}$$
$$(u, u) \geqslant 0, \quad \text{and } (u, u) = 0 \implies u = 0 \quad \textbf{positive definiteness} \tag{3.1d}$$

for all $u, u_1, u_2, v, v_1, v_2 \in V$ and all $\alpha_1, \alpha_2 \in \mathbb{R}$.

(*ii*) The pair $(V, (\cdot, \cdot))$ is said to be a **(real) inner product space**.

(*iii*) Two vectors $u, v \in V$ are said to be **orthogonal** if $(u, v) = 0$. △

In brief, an inner product is a bilinear form on V that is symmetric and positive definite.

An inner product induces a norm on the linear space under consideration:

Lemma 3.2 (inner product induces norm)**.**
Suppose that $(V, (\cdot,\cdot))$ is an inner product space. Then

$$\|u\| \coloneqq \sqrt{(u,u)} \tag{3.2}$$

defines a norm on V.

Proof. The proof is part of homework problem 4.2. □

Definition 3.3 (Hilbert space)**.**
An inner product space $(V, (\cdot,\cdot))$ is said to be a **Hilbert space** if the norm induced by the inner product is complete (see Definition 2.8). △

Note: In other words, a Hilbert space is a Banach space whose norm is induced by an inner product.

Example 3.4 (Hilbert space)**.**

(*i*) In $\mathbb{R}^n$, inner products are in a bijective correspondence with symmetric positive definite matrices. Every inner product on $\mathbb{R}^n$ has the form

$$(u,v)_M = u^\mathsf{T} M\, v$$

for some symmetric positive definite matrix $M \in \mathbb{R}^{n\times n}$. The induced norm is then given by

$$\|u\|_M = \sqrt{u^\mathsf{T} M\, u}.$$

(*ii*) Every finite-dimensional inner product space is a Hilbert space.

(*iii*) Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set. The Lebesgue space $L^2(\Omega)$ carries the inner product

$$(f,g)_{L^2(\Omega)} \coloneqq \int_\Omega f\, g \,\mathrm{d}x, \tag{3.3}$$

which induces the norm (2.12) for $p = 2$. Since $L^2(\Omega)$ is complete, it is a Hilbert space.

(*iv*) More generally, suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set and $k \in \mathbb{N}_0$. The Sobolev space $W^{k,2}(\Omega)$ carries the inner product

$$(f,g)_{W^{k,2}(\Omega)} \coloneqq \sum_{|\alpha|\leqslant k} \int_\Omega (D^\alpha f)\,(D^\alpha g)\,\mathrm{d}x, \tag{3.4}$$

which induces the norm (2.21) for $p = 2$. Since $W^{k,2}(\Omega)$ is complete, it is a Hilbert space. It is customary to denote the inner product space $W^{k,2}(\Omega)$ by $H^k(\Omega)$. In particular, $H^0(\Omega) = L^2(\Omega)$. △

Lemma 3.5 (Cauchy-Schwarz inequality)**.**
Suppose that $(V, (\cdot,\cdot))$ is an inner product space. Then for all $u, v \in V$, we have

$$|(u,v)| \leqslant \|u\|\,\|v\|. \tag{3.5}$$

Equality holds if and only if u and v are linearly dependent, i. e., $\alpha\, u + \beta\, v = 0$ and not both α and β are zero.

Proof. When $v = 0$, then (3.5) holds with equality, and $\{u, v\}$ is linearly dependent.

For the rest of the proof, assume $v \neq 0$. For $\beta \in \mathbb{R}$ we have

$$\begin{aligned} 0 &\leqslant (u - \beta\, v, u - \beta\, v) && \text{due to positive definiteness} \\ &= (u,u) - 2\,\beta\,(u,v) + \beta^2\,(v,v) && \text{due to bilinearity and symmetry.} \end{aligned}$$

Here $(v,v) > 0$ due to positive definiteness, and we set $\beta := \frac{(u,v)}{(v,v)}$. This implies

$$\begin{aligned} 0 &\leqslant (u,u) - 2\,\frac{(u,v)}{(v,v)}\,(u,v) + \frac{(u,v)^2}{(v,v)} \\ &= (u,u) - \frac{(u,v)^2}{(v,v)}. \end{aligned}$$

Multiplication by $(v,v) > 0$ yields (3.5).

We have to investigate when equality holds in (3.5). We can continue to assume $v \neq 0$. When $\{u, v\}$ is linearly dependent, then we have $u = \delta v$ for some $\delta \in \mathbb{R}$. Bilinearity then implies $(u,u) = \delta^2\,(v,v)$ and

$$(u,v)^2 = \big(\delta\,(v,v)\big)^2 = \delta^2\,(v,v)^2 = (u,u)\,(v,v),$$

hence equality (3.5).

Conversely, suppose that equality holds in (3.5), i. e., $(u,v)^2 = (u,u)\,(v,v)$. Setting $\beta := \frac{(u,v)}{(v,v)}$ and applying the same manipulations as above, we find that

$$\begin{aligned} 0 &= (u,u) - \frac{(u,v)^2}{(v,v)} \\ &= (u,u) - 2\,\beta\,(u,v) + \beta^2\,(v,v) \\ &= (u - \beta\, v, u - \beta\, v). \end{aligned}$$

The positive definiteness implies $u - \beta\, v = 0$, and thus $\{u, v\}$ is linearly dependent. □

§ 4 Continuous Functions

The continuity of functions between normed linear spaces can be defined via sequences.

Definition 4.1 (continuity)**.**
Suppose that X and Y are normed linear spaces. A map $F\colon X \to Y$ is said to be **continuous at** $x \in X$ if for all sequences $\big(x^{(k)}\big)$ in X with $x^{(k)} \to x$ in X, we have $F(x^{(k)}) \to F(x)$ in Y. It is said to be **continuous** (**on** X) if it is continuous at every $x \in X$. △

Lemma 4.2 (equivalent definition of continuity)**.**
Suppose that X and Y are normed linear spaces. A map $F\colon X \to Y$ is continuous if and only if for all $x \in X$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x - y\|_X < \delta$ implies $\|F(x) - F(y)\|_Y < \varepsilon$.

Proof. □

§ 4.1 Linear Operators

Linear maps between normed linear spaces are of particular importance.

Definition 4.3 (linear operator, bounded linear operator)**.**
Suppose that X and Y are normed linear spaces.

(*i*) A function $A\colon X \to Y$ is said to be a **linear map** or **linear operator** if

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A(x_1) + \alpha_2 A(x_2) \tag{4.1}$$

holds for all $x_1, x_2 \in X$ and all $\alpha_1, \alpha_2 \in \mathbb{R}$.

(*ii*) A linear operator $A\colon X \to Y$ is said to be **bounded** if there exists $C \geqslant 0$ such that

$$\|A(x)\|_Y \leqslant C\,\|x\|_X \quad \text{for all } x \in X. \tag{4.2}$$

The number

$$\|A\|_{\mathcal{L}(X,Y)} \coloneqq \inf\big\{C \geqslant 0 \,\big|\, (4.2) \text{ holds}\big\} \tag{4.3}$$

is called the **operator norm** of A.

△

Note: It is easy to see that the interval $\{C \geqslant 0 \mid (4.2) \text{ holds}\}$ is closed, and thus the infimum in (4.3) is actually a minimum.

Lemma 4.4 (alternative definitions of the operator norm)**.**
Suppose that X and Y are normed linear spaces and $A\colon X \to Y$ is a bounded linear operator. The operator norm satisfies

$$\|A\|_{\mathcal{L}(X,Y)} = \sup_{\|x\|_X = 1} \|A(x)\|_Y = \sup_{\|x\|_X \leqslant 1} \|A(x)\|_Y = \sup_{x \neq 0} \frac{\|A(x)\|_Y}{\|x\|_X}. \tag{4.4}$$

Lemma 4.5 (boundedness is continuity)**.**
Suppose that X and Y are normed linear spaces and $A\colon X \to Y$ is a linear operator. Then the following statements are equivalent:

(*i*) A is continuous at 0.

(*ii*) A is continuous on X.

(*iii*) A is Lipschitz continuous.

(*iv*) A is bounded.

Proof. The proof is part of homework problem 5.3. □

End of Class 7

End of Week 4

Convergence in the operator norm implies pointwise convergence:

Lemma 4.6 (convergence in the operator norm implies pointwise convergence)**.**
Suppose that X and Y are normed linear spaces and $\left(A^{(k)}\right)$ is a sequence of bounded linear operators $X \to Y$. If $A^{(k)}$ converges to $A \in \mathcal{L}(X,Y)$ in the operator norm, then $A^{(k)}(x)$ converges to $A(x)$ for all $x \in X$.

Proof. The proof is part of homework problem 5.2. □

Lemma 4.7 (existence of unbounded operators)**.**
Suppose that X and Y are normed linear spaces with $\dim(Y) \geqslant 1$. Then the following statements are equivalent:

(*i*) X is finite-dimensional.

(*ii*) Every linear operator $A\colon X \to Y$ is continuous.

Proof. Statement (*i*) $\Rightarrow$ statement (*ii*): Suppose that $\dim(X) = n \in \mathbb{N}_0$ and that $\{v^{(1)}, \dots, v^{(n)}\}$ is a basis of X. For any $x \in X$, we can write $x = \sum_{j=1}^{n} x_j\, v^{(j)}$ and thus $A(x) = \sum_{j=1}^{n} x_j\, A(v^{(j)})$. We estimate

$$\|A(x)\|_Y = \Big\|\sum_{j=1}^{n} x_j\, A(v^{(j)})\Big\|_Y \leqslant \|x\|_\infty \sum_{j=1}^{n} \|A(v^{(j)})\|_Y =: C\,\|x\|_\infty,$$

where $C \geqslant 0$ is a constant. By Theorem 2.13, the norms $\|\cdot\|_\infty$ and $\|\cdot\|_X$ are equivalent, and thus A is continuous.

$\neg$ Statement (*i*) $\Rightarrow$ $\neg$ statement (*ii*): Suppose that X is infinite-dimensional, i. e., at least of countable dimension. Suppose that $(v^{(i)})_{i\in I}$ is a basis for X. Without loss of generality, $\mathbb{N} \subseteq I$. Pick a non-zero element $y \in Y$ and define the linear operator $A\colon X \to Y$ by $A(v^{(k)}) = k\,\|v^{(k)}\|_X\, y$ for $k \in \mathbb{N}$, and $A(v^{(i)}) = 0$ for $i \in I \setminus \mathbb{N}$. Then A is not bounded since $\|A(x^{(k)})\|_Y = k\,\|y\|_Y$ for all $k \in \mathbb{N}$. □

The set of all linear operators $X \to Y$ forms itself a linear space, which we denote by $L(X,Y)$. Addition and scalar multiplication are defined pointwise. The subset of *bounded* linear operators forms a subspace:

Theorem 4.8 (subspace of bounded linear operators)**.**
Suppose that X and Y are normed linear spaces.

(*i*) The set of all bounded linear operators $X \to Y$ is a linear subspace of the space of all linear operators $X \to Y$. We denote it by $\mathcal{L}(X,Y)$.

(*ii*) The operator norm (4.3) is a norm on $\mathcal{L}(X,Y)$.

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

(*iv*) If $\mathcal{L}(X, Y)$ is a Banach space and $\dim(X) \geqslant 1$, then Y is a Banach space.

Proof. Statement (*i*) and statement (*ii*): We use the subspace criterion to show that $\mathcal{L}(X, Y)$ is a linear subspace of $L(X, Y)$. The zero operator is bounded, so $\mathcal{L}(X, Y)$ is nonempty. With $A \in \mathcal{L}(X, Y)$, we have $\alpha A \in \mathcal{L}(X, Y)$ since

$$\|\alpha A\|_{\mathcal{L}(X,Y)} = \sup_{\|x\|_X=1} \|\alpha A(x)\|_Y = \sup_{\|x\|_X=1} |\alpha| \, \|A(x)\|_Y = |\alpha| \sup_{\|x\|_X=1} \|A(x)\|_Y.$$

This proves the absolute homogeneity of the operator norm. Also, for $A, B \in \mathcal{L}(X, Y)$, we have

$$\begin{aligned} \|A + B\|_{\mathcal{L}(X,Y)} &= \sup_{\|x\|_X=1} \|A(x) + B(x)\|_Y \\ &\leqslant \sup_{\|x\|_X=1} \|A(x)\|_Y + \|B(x)\|_Y \\ &\leqslant \sup_{\|x\|_X=1} \|A(x)\|_Y + \sup_{\|x\|_X=1} \|B(x)\|_Y \\ &= \|A\|_{\mathcal{L}(X,Y)} + \|B\|_{\mathcal{L}(X,Y)} \end{aligned}$$

and thus $A + B \in \mathcal{L}(X, Y)$ and the triangle inequality holds. Finally, $\|A\|_{\mathcal{L}(X,Y)} \geqslant 0$ is clear, and $\|A\|_{\mathcal{L}(X,Y)} = 0$ implies $\|A(x)\|_Y = 0$ for all $x \in X$, and thus $A = 0$, the zero element of $\mathcal{L}(X, Y)$.

Statement (*iii*): Suppose that Y is a Banach space and that $\left(A^{(k)}\right)$ is a Cauchy sequence in $\mathcal{L}(X, Y)$. That is, for every $\varepsilon > 0$, there exists an index k_ε such that $\|A^{(k)} - A^{(\ell)}\| < \varepsilon$ holds for all $k, \ell \geqslant k_\varepsilon$.

Step 1: We construct the candidate $A \colon X \to Y$ for the limit of $A^{(k)}$.

For any fixed $x \in X$, we have

$$\|A^{(k)}(x) - A^{(\ell)}(x)\|_Y = \left\|[A^{(k)} - A^{(\ell)}](x)\right\|_Y \leqslant \|A^{(k)} - A^{(\ell)}\|_{\mathcal{L}(X,Y)} \, \|x\|_X.$$

Therefore, the sequence $\left(A(x)^{(k)}\right)$ is Cauchy in Y. Since Y is complete, we can define the pointwise limit $A(x) \coloneqq \lim_{k\to\infty} A_k(x)$.

Step 2: We show that A is linear.

For any $x, y \in X$ and $\alpha, \beta \in \mathbb{R}$, we have

$$\begin{aligned} A(\alpha\, x + \beta\, y) &= \lim_{k\to\infty} A^{(k)}(\alpha\, x + \beta\, y) && \text{by definition of } A \\ &= \lim_{k\to\infty} \left[\alpha\, A^{(k)}(x) + \beta\, A^{(k)}(y)\right] && \text{by linearity of } A^{(k)} \\ &= \alpha \lim_{k\to\infty} A^{(k)}(x) + \beta \lim_{k\to\infty} A^{(k)}(y) && \text{by linearity of the limit, all limits exist} \\ &= \alpha\, A(x) + \beta\, A(y) && \text{by definition of } A. \end{aligned}$$

Step 3: We show that A is bounded.

For any $x \in X$, we have

$$\begin{aligned} \|A(x)\|_Y &\leqslant \|A(x) - A^{(k)}(x)\|_Y + \|A^{(k)}(x)\|_Y && \text{by the triangle inequality} \\ &\leqslant \|A(x) - A^{(k)}(x)\|_Y + \|A^{(k)}\|_{\mathcal{L}(X,Y)} \, \|x\|_X. \end{aligned}$$

Since every Cauchy sequence is bounded (**Quiz 4.1:** Can you prove it?), we have

$$\leqslant \|A(x) - A^{(k)}(x)\|_Y + C\,\|x\|_X.$$

By letting $k \to \infty$, we find that $\|A(x)\|_Y \leqslant C\,\|x\|_X$, with C independent of x. That is, A is bounded.

Step 4: We show that $A^{(k)} \to A$ in $\mathcal{L}(X,Y)$.

Let $\varepsilon > 0$. Since $\left(A^{(k)}\right)$ is Cauchy, there exists k_ε such that $\|A^{(k)} - A^{(\ell)}\|_{\mathcal{L}(X,Y)} < \varepsilon$ for all $k, \ell \geqslant k_\varepsilon$. Now let $x \in X$ be arbitrary. We estimate

$$\begin{aligned}\|A^{(k)}(x) - A^{(\ell)}(x)\|_Y &\leqslant \|A^{(k)} - A^{(\ell)}\|_{\mathcal{L}(X,Y)}\,\|x\|_X \\ &\leqslant \varepsilon\,\|x\|_X \quad \text{for all } k, \ell \geqslant k_\varepsilon.\end{aligned}$$

Passing to the limit $\ell \to \infty$, we obtain

$$\|A^{(k)}(x) - A(x)\|_Y \leqslant \varepsilon\,\|x\|_X \quad \text{for all } k \geqslant k_\varepsilon.$$

This shows $\|A^{(k)} - A\|_{\mathcal{L}(X,Y)} \leqslant \varepsilon$ for all $k \geqslant k_\varepsilon$, i. e., $A^{(k)} \to A$ in $\mathcal{L}(X,Y)$.

Statement *(iv)*: Suppose that $\mathcal{L}(X,Y)$ is a Banach space and $\dim(X) \geqslant 1$. Then there exists a non-zero bounded linear map $f\colon X \to \mathbb{R}$. (**Quiz 4.2:** How do we see this?) In particular, we have $f(x_0) = 1$ for some $x_0 \in X$.

Now define a family $(A_y)_{y\in Y}$ of bounded linear operators $X \to Y$ by

$$A_y(x) \coloneqq f(x)\,y \quad \text{for all } x \in X.$$

Notice that $y \mapsto A_y$ is a linear map $Y \to \mathcal{L}(X,Y)$. Every A_y is indeed bounded since

$$\|A_y(x)\|_Y = |f(x)|\,\|y\|_Y \leqslant \|f\|_{\mathcal{L}(X,\mathbb{R})}\,\|y\|_Y\,\|x\|_X$$

and thus $\|A_y\|_{\mathcal{L}(X,Y)} \leqslant \|f\|_{\mathcal{L}(X,\mathbb{R})}\,\|y\|_Y$. Suppose now that $\left(y^{(k)}\right)$ is a Cauchy sequence in Y. Then

$$\|A_{y^{(k)}} - A_{y^{(\ell)}}\|_{\mathcal{L}(X,Y)} = \|A_{y^{(k)}-y^{(\ell)}}\|_{\mathcal{L}(X,Y)} \leqslant \|f\|_{\mathcal{L}(X,\mathbb{R})}\,\|y^{(k)} - y^{(\ell)}\|_Y$$

and therefore, $A_{y^{(k)}}$ is a Cauchy sequence in $\mathcal{L}(X,Y)$. Since $\mathcal{L}(X,Y)$ is complete, there exists a limit $A \in \mathcal{L}(X,Y)$. But this and Lemma 4.6 imply

$$y^{(k)} = A_{y^{(k)}}(x_0) \to A(x_0) \in Y,$$

and thus $\left(y^{(k)}\right)$ converges, i. e., Y is complete. □

End of Class 8

End of Week 5

§ 4.2 Continuous Embeddings

Definition 4.9 (continuous embedding, isomorphism)**.**
Suppose that X and Y are normed linear spaces.

(*i*) An injective linear map $A\colon X \to Y$ that is also bounded is said to be a **continuous embedding** of X into Y. In this case, the space X is said to be **continuously embedded** into Y.

(*ii*) A bijective linear map $A\colon X \to Y$ that is also bounded and whose inverse is bounded is said to be an **isomorphism** of X onto Y. In this case, the spaces X and Y are said to be **isomorphic**.

(*iii*) An isomorphism $A\colon X \to Y$ such that $\|A(x)\|_Y = \|x\|_X$ for all $x \in X$ is said to be an **isometric isomorphism** or an **isometry** of X onto Y. In this case, the spaces X and Y are said to be **isometric**. △

Remark 4.10 (continuous embedding, isomorphism)**.**

(*i*) In many cases, $X \subseteq Y$ algebraically as a subspace, and we consider the linear inclusion map $i\colon X \to Y$ with $i(x) = x$, which is clearly injective. Notice that the inclusion map is continuous if and only if $\|x\|_Y = \|i(x)\|_Y \leqslant C\,\|x\|_X$, i. e., if and only if $\|\cdot\|_Y$ is weaker on X than $\|\cdot\|_X$. We denote the continuous embedding of X into Y by $X \hookrightarrow Y$.

(*ii*) A surjective linear map $A\colon X \to Y$ is an isomorphism if and only if there exist constants $c, C > 0$ such that

$$c\,\|x\|_X \leqslant \|A(x)\|_Y \leqslant C\,\|x\|_X \quad \text{for all } x \in X$$

holds.

(*iii*) A surjective linear map $A\colon X \to Y$ is an isometry if and only if

$$\|x\|_X = \|A(x)\|_Y \quad \text{for all } x \in X$$

holds.

(*iv*) Two isomorphic normed linear spaces X and Y cannot be distinguished in terms of their structure, up to the equivalence of norms. Two isometric normed linear spaces X and Y cannot be distinguished at all. △

Example 4.11 (continuous embeddings)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set. Then we have the following continuous embeddings of Sobolev spaces:

$$\begin{array}{ccccccccc}
L^\infty(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & L^2(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & L^1(\Omega) \\
\uparrow & & & & \uparrow & & & & \uparrow \\
W^{1,\infty}(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & W^{1,2}(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & W^{1,1}(\Omega) \\
\uparrow & & & & \uparrow & & & & \uparrow \\
W^{2,\infty}(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & W^{2,2}(\Omega) & \hookrightarrow & \cdots & \hookrightarrow & W^{2,1}(\Omega) \\
\vdots & & & & \vdots & & & & \vdots
\end{array}$$

The inclusions in horizontal direction rely on the boundedness of Ω, while the inclusions in vertical direction hold for any open set Ω. Moreover, there are further embeddings in "north-westerly" direction due to the Sobolev embedding theorem, which allow differentiability to be traded for higher integrability indices. △

§ 4.3 The Dual Space

Definition 4.12 (algebraic and topological dual spaces)**.**

Suppose that X is a normed linear space.

(*i*) The **algebraic dual space** of X is the linear space

$$X' := L(X, \mathbb{R}) \tag{4.5}$$

of all linear maps $X \to \mathbb{R}$, also known as **linear functionals** on X.

(*ii*) The **topological dual space** of X is the linear space

$$X^* := \mathcal{L}(X, \mathbb{R}) \tag{4.6}$$

of continuous (bounded) linear functionals on X. △

Clearly, X^* is a linear subspace of X'. It is, in fact a proper subspace, if and only if X is infinite-dimensional (Lemma 4.7). Since $\mathbb{R}$ is complete, X^* is always a Banach space by Theorem 4.8. Since we use the absolute value as the norm on $\mathbb{R}$, the dual space X^* is equipped with the operator norm

$$\|f\|_{X^*} = \sup_{\|x\|_X=1} |f(x)|.$$

Given $f \in X^*$ and $x \in X$, we often use the notation

$$\langle f, x \rangle_{X^*,X} := f(x).$$

The bracket $\langle \cdot, \cdot \rangle_{X^*,X}$ is a bilinear form on $X^* \times X$ and it is called the **dual pairing** of X and X^*. In the future, we will often simply say **dual space** instead of **topological dual space** since we will not use the algebraic dual space much.

Example 4.13 (dual spaces)**.**

(*i*) Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set. Moreover, let $p \in [1, \infty)$ and $q \in (1, \infty]$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then the dual space of $L^p(\Omega)$ is isometrically isomorphic to $L^q(\Omega)$.
In this representation of $L^p(\Omega)^*$, the dual pairing is given by

$$\langle f, g \rangle := \int_\Omega f\, g \, \mathrm{d}x \quad \text{for } f \in L^p(\Omega) \text{ and } g \in L^q(\Omega). \tag{4.7}$$

(*ii*) The dual space of $L^\infty(\Omega)$ does not have a similarly simple representation. It is isometrically isomorphic to the space of finitely additive signed measures on Ω that are absolutely continuous w.r.t. the Lebesgue measure; see for instance Dunford, Schwartz, 1988, Theorem IV.8.16. △

§ 4.4 The Dual Space of a Hilbert Space

The ability to represent the dual of a normed linear space as concretely as for L^p spaces is a rather special property of a normed linear space. However, it is always possible for Hilbert spaces.

Theorem 4.14 (Riesz representation theorem)**.**
Suppose that H is a Hilbert space. Then the dual space H^* of H is isometrically isomorphic to H itself, via the isomorphism

$$\Phi\colon H \ni u \mapsto (u,\cdot)_H \in H^*. \tag{4.8}$$

Moreover, the norm of H^* (i. e., the operator norm of $f \in \mathcal{L}(H,\mathbb{R})$) is induced by the inner product

$$(f,g)_{H^*} \coloneqq \left(\Phi^{-1}(f),\Phi^{-1}(g)\right)_H = \langle f\,,\Phi^{-1}(g)\rangle_{H^*,H} = \langle g\,,\Phi^{-1}(f)\rangle_{H^*,H}. \tag{4.9}$$

Proof. We break the proof down into several steps.

Step 1: We show that $\Phi\colon H \to H'$ is linear.

First of all, $\Phi(u) \in H'$ for all $u \in H$ since $\Phi(u) = (u,\cdot)_H$ and the inner product is linear in the second argument.

For $u, v, w \in H$ and $\alpha, \beta \in \mathbb{R}$, we have

$$\begin{aligned}
\langle \Phi(\alpha\,u + \beta\,v)\,, w\rangle &= (\alpha\,u + \beta\,v, w)_H && \text{by definition of } \Phi\\
&= \alpha\,(u,w)_H + \beta\,(v,w)_H && \text{by linearity of the inner product in the first argument}\\
&= \alpha\,\langle \Phi(u)\,, w\rangle + \beta\,\langle \Phi(v)\,, w\rangle && \text{by definition of } \Phi.
\end{aligned}$$

This shows $\Phi(\alpha\,u + \beta\,v) = \alpha\,\Phi(u) + \beta\,\Phi(v)$, so Φ is linear.

Step 2: We show that $\Phi\colon H \to H^*$ holds.

For $u \in H$ and $v \in H$, we have

$$|\langle \Phi(u)\,, v\rangle| = |(u,v)_H| \leqslant \|u\|_H\,\|v\|_H \quad \text{by the Cauchy-Schwarz inequality.}$$

Therefore, $\Phi(u)$ is a bounded linear functional on H with $\|\Phi(u)\|_{H^*} \leqslant \|u\|_H$.

Step 3: We show that $\|\Phi(u)\|_{H^*} = \|u\|_H$ for all $u \in H$.

For $u \in H$, we have

$$|\langle \Phi(u)\,, u\rangle| = |(u,u)_H| = (u,u)_H = \|u\|_H^2,$$

which shows $\|\Phi(u)\|_{H^*} \geqslant \|u\|_H$.

Step 4: We show that Φ is surjective.[15] (By Remark 4.10 *(iii)* this implies that Φ is an isometric isomorphism.)

Suppose that $f \in H^*$ is given. When $f = 0$, we can simply choose $u = 0$ since $\Phi(0) = 0$, which holds for any linear map. Now suppose $f \neq 0$. Consider the kernel (nullspace) of f,

$$\ker(f) \coloneqq \{v \in H \mid f(v) = 0\}.$$

[15] This is the main step in the proof, where the completeness of H is crucial.

It is not difficult to see that $\ker(f)$ is a closed subspace of H, and it is not equal to H since $f \neq 0$. One can show that, as a consequence, there exists $v \in H$ such that $f(v) \neq 0$ that is orthogonal to $\ker(f)$.[16] Without loss of generality, we can assume that $\|v\|_H = 1$.

We now choose $u := f(v)\, v$ and show $\Phi(u) = f$, so that Φ is surjective. Indeed, we have

$$\|u\|_H = \|f(v)\, v\|_H = |f(v)|\, \|v\|_H = |f(v)|$$

and

$$f(u) = f(f(v)\, v) = f(v)\, f(v) = |f(v)|^2 = \|u\|_H^2.$$

For any $w \in H$, this implies

$$\begin{aligned}
\langle \Phi(u)\,, w\rangle &= (u, w)_H && \text{by definition of } \Phi \\
&= \Big(u, w - \frac{f(w)}{\|u\|_H^2}\, u\Big)_H + \Big(u, \frac{f(w)}{\|u\|_H^2}\, u\Big)_H \\
&= \Big(u, w - \frac{f(w)}{\|u\|_H^2}\, u\Big)_H + \frac{f(w)}{\|u\|_H^2}(u, u)_H \\
&= \Big(u, w - \frac{f(w)}{\|u\|_H^2}\, u\Big)_H + f(w).
\end{aligned}$$

The second factor in the inner product belongs to $\ker(f)$, since

$$\begin{aligned}
f\Big(w - \frac{f(w)}{\|u\|_H^2}\, u\Big) &= f(w) - \frac{f(w)}{\|u\|_H^2} f(u) && \text{by linearity of } f \\
&= f(w) - \frac{f(w)}{\|u\|_H^2} \|u\|_H^2 && \text{since } f(u) = \|u\|_H^2 \\
&= 0.
\end{aligned}$$

But since v is orthogonal to $\ker(f)$, so is $u = f(v)\, v$. This proves

$$\langle \Phi(u)\,, w\rangle = f(w) \quad \text{for all } w \in H,$$

whence $\Phi(u) = f$.

Step 5: We show that (4.9) defines an inner product that induces the norm of H^*.

First of all, we have by definition of Φ and the symmetry of $(\cdot,\cdot)_H$ that

$$\langle f\,, \Phi^{-1}(g)\rangle_{H^*,H} = \big(\Phi^{-1}(f), \Phi^{-1}(g)\big)_H = \big(\Phi^{-1}(g), \Phi^{-1}(f)\big)_H = \langle g\,, \Phi^{-1}(f)\rangle_{H^*,H}$$

and so the equalities in (4.9) hold. Defining now

$$(f, g)_{H^*} := \big(\Phi^{-1}(f), \Phi^{-1}(g)\big)_H$$

and the linearity of Φ^{-1} then show that $(\cdot,\cdot)_{H^*}$ is a symmetric bilinear form on H^*. it is also positive definite since Φ^{-1} is a bijection. □

End of Class 9

[16] The proof would require more machinery, including the parallelogram identity for inner products and subsequently the existence of orthogonal projections onto closed and convex subsets (in particular, onto closed subspaces) in Hilbert spaces.

§ 5 Existence Theorems for Global Minimizers

In this section we will discuss sufficient conditions for minimizers of optimization problems in normed linear spaces to exist. We begin with the well known

Theorem 5.1 (Weierstrass extreme value theorem)**.**
Suppose that V is a normed linear space and $K \subseteq V$ is compact. Moreover, suppose that $f\colon K \to \mathbb{R}$ is continuous. Then $f(K) \subseteq \mathbb{R}$ is compact. As a consequence, f attains its minimum (and its maximum) on K.

Proof. We will show that $f(K)$ is sequentially compact, which is equivalent to compactness due to Theorem 2.17. Suppose that $\left(r^{(k)}\right)$ is a sequence in $f(K)$. That is, there exists a sequence $\left(x^{(k)}\right)$ in K such that $f(x^{(k)}) = r^{(k)}$. Since K is (sequentially) compact, there exists a subsequence $\left(x^{(k^{(\ell)})}\right)$ such that $x^{(k^{(\ell)})} \to x^* \in K$ as $\ell \to \infty$. Due to the continuity of f (Definition 4.1), $f\left(x^{(k^{(\ell)})}\right) \to f(x^*)$, and since $x^* \in K$, we have $f(x^*) \in f(K)$. This shows that $f(K)$ is sequentially compact.

As a compact set, $f(K) \subseteq \mathbb{R}$ is closed and bounded, i. e., $\inf\{f(x) \mid x \in K\}$ and $\sup\{f(x) \mid x \in K\}$ are finite. Due to the closedness, inf and sup are actually attained. □

So Weierstrass' theorem is the same as in $V = \mathbb{R}^n$. However, it is rarely applicable in infinite-dimensional normed linear spaces V, because the choice of compact subsets $K \subseteq V$ is quite limited. This is hinted at by the fact that even unit balls in infinite-dimensional normed linear spaces are not compact (Theorem 2.18). For instance, in $L^p(\Omega)$, one can precisely characterize the compact subsets.

Theorem 5.2 (compact subsets of $L^p(\Omega)$, **Kolmogorov-Riesz theorem**)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set and $K \subseteq L^p(\Omega)$. Then the following statements are equivalent:

(*i*) K is compact in $L^p(\Omega)$.

(*ii*) K is closed, bounded and **equicontinuous**.

For a proof, see for instance Adams, Fournier, 2003, Theorem 2.32. The definition of equicontinuity makes use of the shift-operator $\tau_h\colon L^p(\Omega) \to L^p(\Omega)$ for $h \in \mathbb{R}^d$, defined by $f \mapsto \tau_h f := f(\cdot + h)\,\chi_\Omega$.[17] **Equicontinuity** means that for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\|\tau_h f - f\|_{L^p(\Omega)} (\varepsilon$ for all $f \in K$ and all $h \in \mathbb{R}^d$ with $|h|_2 < \delta$.

Example 5.3 (non-compactness of L^p-functions with bound constraints)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open and bounded set. Moreover, let $a, b \in \mathbb{R}$ be such that $a < b$. Then the set

$$A := \{f \in L^p(\Omega) \mid a \leqslant f(x) \leqslant b \text{ for a.a. } x \in \Omega\} \tag{5.1}$$

is closed and bounded in $L^p(\Omega)$, but it not compact.

[17] It is easy to see that τ_h indeed maps $L^p(\Omega)$ into itself and has operator norm $\leqslant 1$.

To see this, we discuss for simplicity the case where $\Omega = (0,1) \subseteq \mathbb{R}$ is an open and bounded interval. Consider the sequence $\left(f^{(k)}\right)$ defined by

$$f^{(k)}(x) := \begin{cases} 0 & \text{if the } k\text{-th binary digit (after the decimal) of } x \text{ is } 0, \\ 1 & \text{if the } k\text{-th binary digit (after the decimal) of } x \text{ is } 1. \end{cases}$$

In other words, $f^{(k)}$ is the characteristic function of a union of disjoint intervals of length 2^{-k}. Then we have

$$\|f^{(k)} - f^{(\ell)}\|_{L^p(\Omega)}^p = \frac{1}{2} \quad \text{for all } k \neq \ell.$$

Therefore, no subsequence of $\left(f^{(k)}\right)$ is a Cauchy sequence. △

We would need to add further conditions to the functions in (5.1) to obtain a compact subset of $L^p(\Omega)$. Some possibilities are monotonicity (for $\Omega \subseteq \mathbb{R}$), convexity or concavity, or additional smoothness (such as $f \in W^{1,1}(\Omega)$).

The following example is a demonstration that global minimizers may fail to exist in infinite-dimensional normed linear spaces in the absence of compactness.

Example 5.4 (non-existence of global minimizers[18])**.**
On $\Omega = \mathbb{R}$, consider the function $g \in L^2(\Omega)$ defined by $g(x) := \exp(-x^2)$ and the problem

$$\begin{aligned} \text{Minimize} \quad & J(f) := \int_\Omega f(x)\, g(x)\, \mathrm{d}x, \quad \text{where } f \in L^2(\Omega) \\ \text{subject to} \quad & f \geqslant 0 \quad \text{a.e. in } \Omega \\ \text{and} \quad & \|f\|_{L^2(\Omega)} = 1. \end{aligned}$$

This problem has the feasible set

$$F := \{f \in L^2(\Omega) \mid f \geqslant 0 \text{ a.e. in } \Omega \text{ and } \|f\|_{L^2(\Omega)} = 1\},$$

which is not compact. The objective $f \mapsto J(f)$ is continuous on $L^2(\Omega)$ (**Quiz 5.1:** Why?) and bounded below by 0. In fact, for all $f \in F$, we have $J(f) > 0$.

Considering the sequence of characteristic functions $f^{(k)} = \chi_{[k,k+1]}$ shows

$$J(f^{(k)}) = \int_k^{k+1} \exp(-x^2)\, \mathrm{d}x \to 0 \quad \text{as } k \to \infty.$$

Therefore, the infimum of J on F is 0, but it is not attained. △

As a remedy, we may resort to a different topology on normed linear spaces. Broadly speaking, when we have fewer open sets and thus fewer open covers of a set, we have a "better chance" of compactness.

[18] example communicated by Gerd Wachsmuth (BTU Cottbus)

§ 5.1 The Weak Topology on a Normed Linear Space

Definition 5.5 (weakly open sets, weakly convergent sequences)**.**
Suppose that V is a normed linear space.

(*i*) A set $U \subseteq V$ is said to be **weakly open** if for all $x \in U$ there exist $\varepsilon > 0$, $n \in \mathbb{N}$ and $f^{(1)}, \dots, f^{(n)} \in V^*$ such that

$$\left\{ y \in V \,\middle|\, |\langle f^{(i)}, y - x\rangle| < \varepsilon \text{ for } i = 1, \dots, n \right\} \subseteq U. \tag{5.2}$$

(*ii*) A sequence $\left(x^{(k)}\right)$ in V is said to be **weakly convergent** to $x \in V$ if for all $f \in V^*$, we have

$$\lim_{k\to\infty} \langle f, x^{(k)}\rangle = \langle f, x\rangle.$$

In this case we write $x^{(k)} \rightharpoonup x$. △

The collection of weakly open sets in V is called the **weak topology** on $(V, \|\cdot\|_V)$. For a clearer distinction, we may refer to the norm topology on V as the **strong topology**. Similarly, we may speak of **strongly convergent sequences**.

One can show that the weak limit of a sequence is unique.

Theorem 5.6 (weak topology in finite-dimensional normed linear spaces)**.**
Suppose that V is a finite-dimensional normed linear space. Then the weak topology on V coincides with the strong topology.

Proof. We will show below in Theorem 5.8 that every weakly open set in V is open in the strong topology. Therefore, we only need to show that every strongly open set is weakly open. So suppose that $U \subseteq V$ is strongly open and $x \in U$. Then there exists $r > 0$ such that $B_r(x) \subseteq U$. By Definition 5.5, we need to show that there exist $\varepsilon)0$, $n \in \mathbb{N}$ and $f^{(1)}, \dots, f^{(n)} \in V^*$ such that (5.2)

$$U' := \left\{ y \in V \,\middle|\, |\langle f^{(i)}, y - x\rangle| < \varepsilon \text{ for } i = 1, \dots, n \right\} \subseteq U$$

holds.

Suppose that $\{v^{(1)}, \dots, v^{(n)}\}$ is a basis of V and that $x = \sum_{i=1}^n x_i v^{(i)}$. We denote by $f^{(i)}$ the coordinate map $V \ni x \mapsto x_i \in \mathbb{R}$, which is linear and, thanks to the finite dimensionality of V, continuous (Lemma 4.7). For any $y \in V$, we find

$$\begin{aligned}
\|y - x\|_V &= \left\| \sum_{i=1}^n (y_i - x_i)\, v^{(i)} \right\|_V \\
&= \left\| \sum_{i=1}^n \langle f^{(i)}, y - x\rangle\, v^{(i)} \right\|_V \\
&\leqslant \sum_{i=1}^n \left\| \langle f^{(i)}, y - x\rangle\, v^{(i)} \right\|_V \\
&\leqslant \sum_{i=1}^n |\langle f^{(i)}, y - x\rangle| \max\left\{ \|v^{(i)}\|_V \,\middle|\, i = 1, \dots, n \right\} \\
&= C \sum_{i=1}^n |\langle f^{(i)}, y - x\rangle|.
\end{aligned}$$

Consequently, when we choose

$$U' := \left\{ y \in V \,\middle|\, |\langle f^{(i)}, y - x \rangle| < \varepsilon \text{ for } i = 1, \dots, n \right\}$$

with $\varepsilon := \frac{r}{Cn}$, then we have $U' \subseteq B_r(x) \subseteq U$. □

Note: In particular, the strong and weak topologies on $\mathbb{R}$ coincide. In general, the finite dimension is sufficient, but not necessary for the weak and strong topologies to coincide. A prominent example if the space ℓ^1 of absolutely summable sequences.

End of Class 10

End of Week 6

Remark 5.7 (weak topology)**.**

(*i*) The norm on V enters Definition 5.5 only through the dual space V^*. (Recall that the norm determines which linear functionals are continuous.)

(*ii*) When $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on V, then both induce the same weak topology on V.

(*iii*) The weak topology is not, in general, induced by a norm or metric (not "metrizable"). Therefore, there is in general no notion of "distance" in the weak topology. In addition, we cannot define the notion of **weak continuity** via weakly convergent sequences.

(*iv*) Our Definition 5.5 of weakly convergent sequences is compatible with the general notion of convergence in topological spaces, i. e., for all weak neighborhoods U of the limit x, there exists $k_0 \in \mathbb{N}$ such that $x^{(k)} \in U$ for all $k \geqslant k_0$. △

Theorem 5.8 (relation between the weak and strong topologies)**.**

Suppose that V is a normed linear space.

(*i*) Every weakly open set in V is open in the strong topology.

(*ii*) Every strongly convergent sequence is weakly convergent (to the same limit).

(*iii*) Suppose that $f : V \to \mathbb{R}$ is **weakly continuous**, i. e., continuous in the weak topology.[19] Then f is continuous in the strong topology as well.

(*iv*) Suppose that $f : V \to \mathbb{R}$ is **weakly sequentially continuous**, i. e., $x^{(k)} \rightharpoonup x$ implies $f(x^{(k)}) \to f(x)$.[20] Then f is continuous in the strong topology as well.

Proof. Statement (*i*): Suppose that $U' \subseteq V$ is weakly open and $x \in U'$. Then there exist $\varepsilon > 0$, $n \in \mathbb{N}$ and $f^{(1)}, \dots, f^{(n)} \in V^*$ such that (5.2) holds. We set

$$r := \min\left\{ \frac{\varepsilon}{\cancel{2}\, \|f^{(i)}\|_{V^*} \cancel{+1}} \,\middle|\, i = 1, \dots, n \right\}. \tag{5.3}$$

[19] This means that pre-images of (weakly) open sets in $\mathbb{R}$ are weakly open in V.

[20] Here we use the fact that the weak and strong topologies on $\mathbb{R}$ coincide so we do not have to distinguish between weak-weak sequential continuity and weak-strong sequential continuity.

Then we have for $y \in B_r(x)$:

$$\begin{aligned} |\langle f^{(i)}, y - x\rangle|_V &\leqslant \|f^{(i)}\|_{V^*} \|y - x\|_V \\ &< \|f^{(i)}\|_{V^*} r \\ &\leqslant \varepsilon. \end{aligned}$$

This implies $B_r(x) \subseteq U'$, so U' is open in the strong topology.

Statement (ii): Suppose that $\|x^{(k)} - x\|_V \to 0$ as $k \to \infty$. When $f \in V^*$, then this implies

$$\langle f, x^{(k)} - x\rangle \leqslant \|f\|_{V^*} \|x^{(k)} - x\|_V \to 0 \quad \text{as } k \to \infty,$$

so $\lim_{k\to\infty}\langle f, x^{(k)}\rangle = \langle f, x\rangle$, which means $x^{(k)} \rightharpoonup x$.

Statement (iii): Suppose that $U \subseteq \mathbb{R}$ is open. Then $f^{-1}(U)$ is weakly open, so $f^{-1}(U)$ is open in the strong topology as well. This means that f is strongly continuous.

Statement (iv): We can show the strong continuity of f using sequences. Suppose that $x^{(k)} \to x$, then also $x^{(k)} \rightharpoonup x$ by statement (ii) and the claim follows. □

The following result simplifies the proof of convergence or weak convergence of sequences in normed linear spaces:

Lemma 5.9 (convergence principle)**.**
Suppose that X is a normed linear space and that $\left(x^{(k)}\right)$ is a sequence in X.

(i) The following are equivalent:
  (a) $x^{(k)} \to x$.
  (b) Every subsequence of $\left(x^{(k)}\right)$ contains a subsequence that converges to x strongly.

(ii) The following are equivalent:
  (a) $x^{(k)} \rightharpoonup x$.
  (b) Every subsequence of $\left(x^{(k)}\right)$ contains a subsequence that converges to x weakly.

Proof. This proof is addressed in homework problem 7.1. □

Remark 5.10 (further properties)**.**
Suppose that V is a normed linear space.

(i) The weak topology on V is the weakest topology so that all strongly continuous linear functionals (elements of V^*) remain continuous.

(ii) Weakly convergent sequences are bounded.[21]

(iii) Suppose that $x^{(k)} \rightharpoonup x$ in V and $f^{(k)} \to f$ in V^*. Then $\langle f^{(k)}, x^{(k)}\rangle \to \langle f, x\rangle$. △

For linear maps between normed linear spaces, the notions of weak, weak sequential and strong continuity coincide.

[21] This follows from the Banach-Steinhaus theorem (uniform boundedness principle).

Lemma 5.11 ((weak, sequential) continuity of linear operators)**.**
Suppose that X and Y are normed linear spaces and $A \in L(X,Y)$ is a linear map. Then the following statements are equivalent:

(*i*) A is continuous (bounded) w.r.t. the strong topologies, i. e., $A \in \mathcal{L}(X,Y)$.

(*ii*) A is weakly continuous, i. e., for every weakly open $V \subseteq Y$, the pre-image $A^{-1}(V)$ is weakly open in X.

(*iii*) A is weakly sequentially continuous, i. e., $x^{(k)} \rightharpoonup x$ implies $A(x^{(k)}) \rightharpoonup A(x)$.

Proof. Statement (*i*) $\Rightarrow$ statement (*ii*). It is enough to show that the pre-images of weak neighborhoods of $0 \in Y$ are weak neighborhoods of $0 \in X$. Suppose that $V \subseteq Y$ is a weak neighborhood of $0 \in Y$. That is, there exist $\varepsilon > 0$, $n \in \mathbb{N}$ and $f^{(1)}, \dots, f^{(n)} \in Y^*$ such that

$$V_0 := \big\{ y \in Y \,\big|\, |\langle f^{(i)}, y\rangle| < \varepsilon \text{ for } i = 1, \dots, n \big\}.$$

We claim that

$$U_0 := \big\{ x \in X \,\big|\, |\langle f^{(i)} \circ A, x\rangle| < \varepsilon \text{ for } i = 1, \dots, n \big\}$$

is a weak neighborhood of 0 contained in the pre-image of V_0. This shows that $A^{-1}(V_0)$ is itself a weak neighborhood of $0 \in X$.

Indeed, $f^{(i)} \circ A$ is linear and continuous, i. e., an element of X^*, and thus U_0 is a weak neighborhood of 0. Moreover,

$$\begin{aligned} U_0 &= \big\{ x \in X \,\big|\, |\langle f^{(i)} \circ A, x\rangle| < \varepsilon \text{ for } i = 1, \dots, n \big\} \\ &= \big\{ x \in X \,\big|\, |\langle f^{(i)}, A(x)\rangle| < \varepsilon \text{ for } i = 1, \dots, n \big\} \\ &\subseteq A^{-1}\big\{ y \in Y \,\big|\, |\langle f^{(i)}, y\rangle| < \varepsilon \text{ for } i = 1, \dots, n \big\}, \end{aligned}$$

or $U_0 \subseteq A^{-1}(V_0)$.[22]

Statement (*ii*) $\Rightarrow$ statement (*iii*). Suppose that $x^{(k)} \rightharpoonup x$. Suppose that $V \subseteq Y$ is some weak neighborhood of $A(x)$. Then, by the assumption of weak continuity, $A^{-1}(V)$ is a weak neighborhood of x. Thus, there exists $k_0 \in \mathbb{N}$ such that $x^{(k)} \in A^{-1}(V)$ for all $k \geqslant k_0$. Consequently, $A(x^{(k)}) \in V$ for all $k \geqslant k_0$. This shows that $A(x^{(k)}) \rightharpoonup A(x)$.

Statement (*iii*) $\Rightarrow$ statement (*i*). We argue by contradiction. Suppose that A is not bounded, i. e., not continuous at 0 (Lemma 4.5). We can find a sequence $x^{(k)}$ in X such that $\|x^{(k)}\|_X = 1$ and $\|A(x^{(k)})\|_Y \geqslant k^2$. By rescaling, we may assume $\|x^{(k)}\|_X = 1/k \to 0$ and $\|A(x^{(k)})\|_Y \geqslant k$. Since strong convergence implies weak convergence, we have $x^{(k)} \rightharpoonup 0$. By the assumption of weak sequential continuity, we have $A(x^{(k)}) \rightharpoonup A(0) = 0$. But this implies that $\|A(x^{(k)})\|_Y$ is bounded, which is a contradiction. □

End of Class 11

Corollary 5.12 ((weak, sequential) continuity of linear functionals)**.** Suppose that X is a normed linear space and $f \in L(X,\mathbb{R})$ is a linear functional. Then the following statements are equivalent:

(*i*) f is continuous (bounded) w.r.t. the strong topology, i. e., $f \in \mathcal{L}(X,\mathbb{R}) = X^*$.

[22] To clarify this, suppose that $x \in X$ satisfies $|\langle f^{(i)}, A(x)\rangle| < \varepsilon$ for $i = 1, \dots, n$. Then clearly $A(x) \in V_0$, i. e., $x \in A^{-1}(V_0)$.

(*ii*) f is weakly continuous, i. e., for every open $V \subseteq \mathbb{R}$, the pre-image $A^{-1}(V)$ is weakly open in X.

(*iii*) f is weakly sequentially continuous, i. e., $x^{(k)} \rightharpoonup x$ implies $f(x^{(k)}) \rightharpoonup f(x)$.

Proof. The result follows directly from Lemma 5.11, taking into account that on $\mathbb{R}$, the strong and weak topologies coincide by Theorem 5.6. □

The remaining results in this subsection simplify tremendously the verification of weakly sequentially closed sets and weakly sequentially lower semi-continuous functionals. They combine a geometric and a topological assumption.

Theorem 5.13 (convex closed sets are weakly sequentially closed)**.**
Suppose that X is a normed linear space and $A \subseteq X$ is convex and (strongly) closed. Then A is **weakly sequentially closed**, i. e., for any sequence $x^{(k)}$ in A

$$x^{(k)} \rightharpoonup x \quad \Rightarrow \quad x \in A. \tag{5.4}$$

The proof of Theorem 5.13 uses a version of the Hahn-Banach separation theorem. We refer the interested reader, e. g., to Werner, 2007, Theorem III.3.8 or Barbu, Precupanu, 2012.

Theorem 5.14 (convex continuous functionals are weakly sequentially lower semi-continuous)**.**
Suppose that X is a normed linear space and $f : X \to \mathbb{R}$ is a convex and (strongly) continuous functional. Then f is **weakly sequentially lower semi-continuous**, i. e., for any sequence $x^{(k)}$ in X, we have

$$x^{(k)} \rightharpoonup x \quad \Rightarrow \quad \liminf_{k\to\infty} f(x^{(k)}) \geqslant f(x). \tag{5.5}$$

Theorem 5.14 can be shown by combining the result of Theorem 5.13 with the following lemma.

Lemma 5.15 (characterization of weak sequential lower semi-continuity)**.**
Suppose that X is a normed linear space and $f : X \to \mathbb{R}$ is a functional. Then the following are equivalent:

(*i*) f is weakly sequentially lower semi-continuous.

(*ii*) The epigraph $\operatorname{epi} f$ is weakly sequentially closed.

(*iii*) The sublevel sets $S_\alpha := \{x \in X \,|\, f(x) \leqslant \alpha\}$ are weakly sequentially closed (possibly empty) for all $\alpha \in \mathbb{R}$.

Proof. This proof is addressed in homework problem 7.2. □

Example 5.16 (weakly sequentially lower semi-continuous functionals)**.**

(*i*) On a normed linear space X, every norm is weakly sequentially lower semi-continuous since it is, of course, continuous, and convex by the triangle inequality:

$$\|\alpha\, x + (1-\alpha)\, y\| \leqslant \alpha\, \|x\| + (1-\alpha)\, \|y\|$$

for alle $\alpha \in [0,1]$ and $x, y \in X$.

(*ii*) In infinite-dimensional normed linear spaces, the norm is, in general, not weakly sequentially continuous. Consider as an example the orthonormal system

$$u^{(k)} := \frac{1}{\sqrt{k}} \sin(k\,x)$$

on the Hilbert space $L^2((0,\pi))$. Then we have

$$\|u^{(k)}\|_{L^2((0,\pi))} = 1 \quad \text{for all } k \in \mathbb{N}.$$

Moreover, $u^{(k)} \rightharpoonup 0$ in $L^2((0,\pi))$. To see this, consider $f \in L^2((0,\pi))^*$, representing an element of the dual space. Then $(f, u^{(k)})$ is the sequence of Fourier coefficients of f, and Parseval's identity shows

$$\sum_{k=1}^{\infty} \left|(f, u^{(k)})\right|^2 = \|f\|^2_{L^2((0,\pi))}.$$

Therefore, $(f, u^{(k)}) \to 0$ as $k \to \infty$. △

§ 5.2 Reflexivity

Recall that we motivated the concept of the weak topology on a normed linear space in order to obtain more compact sets compared to the strong topology. Our hope was to find that, e. g., L^p functions subject to bound constraints (Example 5.3) form a weakly (sequentially) compact subset.

Analogous as in Definition 2.15, a subset $K \subseteq V$ in a normed linear space V is said to be **weakly sequentially compact** if every sequence $\left(x^{(k)}\right)$ in K contains a weakly convergent subsequence whose limit belongs to K.

The following example shows that, in general, we may still not obtain weak sequential compactness.[23]

Example 5.17 (not weakly sequentially compact set of L^1 functions with bound constraints)**.**
The set

$$A := \left\{u \in L^1(\mathbb{R}) \,\middle|\, 0 \leqslant u \leqslant 1 \text{ a.e. in } \mathbb{R}\right\}$$

is not weakly sequentially compact in $L^1(\mathbb{R})$. To see this, we will show that the sequence $u^{(k)} := \chi_{[k,k+1]}$ in A does not contain a weakly convergent subsequence.

Suppose that $\left(u^{(k^{(\ell)})}\right)$ is a subsequence of $\left(u^{(k)}\right)$. We will exhibit a linear functional $f \in L^1(\mathbb{R})^*$ such that $\langle f, u^{(k^{(\ell)})}\rangle$ does not converge. In other words, $u^{(k^{(\ell)})}$ does not converge weakly. The topological dual space of $L^1(\mathbb{R})$ is isometrically isomorphic to $L^\infty(\mathbb{R})$. We choose $f \in L^\infty(\mathbb{R})$ as follows:

$$f := \sum_{\ell=1}^{\infty} \chi_{[k^{(\ell)},k^{(\ell)}+1]}(-1)^\ell.$$

This means that f alternatingly takes the values ±1 on unit-length intervals starting at the indices of the subsequence $k^{(\ell)}$. We obtain

$$\langle f, u^{(k^{(\ell)})}\rangle = (-1)^\ell,$$

which indeed does not converge. △

[23] By the way, the Eberlein–Šmulian theorem shows that for weakly closed subsets of Banach spaces, weak compactness and weak sequential compactness are the same.

Despite the failure of weak sequential compactness in Example 5.17, there is only one additional property missing to fix the issue.

Definition 5.18 (reflexive normed linear space)**.**
Suppose that X is a normed linear space with dual space X^*.

(*i*) We denote the **bidual space** of X, i. e., the dual space of X^*, by X^{**}.

(*ii*) Given $x \in X$, we consider the map

$$X^* \ni f \mapsto F_x(f) := \langle f, x \rangle \in \mathbb{R}. \tag{5.6}$$

Notice that F_x is a bounded linear functional on X^* thanks to the estimate

$$|F_x(f)| = |\langle f, x \rangle| \leqslant \|f\|_{X^*} \|x\|_X.$$

In other words, $F_x \in X^{**}$ holds with $\|F_x\|_{X^{**}} \leqslant \|x\|_X$. Moreover, the map

$$X \ni x \mapsto i_{X^{**} \leftarrow X}(x) := F_x \in X^{**}$$

is obviously linear, and, as we saw, continuous. The Hahn-Banach theorem can be used to show that it is also injective and an isometry. Therefore, we call $i \colon X \to X^{**}$ the **canonical embedding** or **canonical isometric embedding**; compare Definition 4.9.

(*iii*) The normed linear space X is said to be **reflexive** if the canonical embedding $X \hookrightarrow X^{**}$ is surjective (i. e., an isometric isomorphism of X and X^{**}). △

Note: A reflexive normed linear space is necessarily a Banach space since X^{**} is a dual space.

End of Class 12

End of Week 7

Example 5.19 (reflexivity of Lebesgue and Sobolev spaces)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is open.

(*i*) The Lebesgue space $L^p(\Omega)$ is reflexive if and only if $p \in (1, \infty)$ holds.

(*ii*) The Sobolev space $W^{k,p}(\Omega)$ is reflexive if and only if $p \in (1, \infty)$ holds. △

Lemma 5.20 (Hilbert spaces are reflexive)**.**
Every Hilbert space X is reflexive.

Proof. This proof is addressed in homework problem 7.3. □

The utility of reflexivity for us is that it simplifies the verification of weak sequential compactness.

Theorem 5.21 (characterization of weakly sequentially compact sets in reflexive spaces)**.**
Suppose that X is a reflexive normed linear space and $A \subseteq X$. Then the following are equivalent:

(*i*) A is weakly sequentially compact.

(*ii*) A is bounded and weakly sequentially closed.

Corollary 5.22 (in reflexive spaces, convex, closed and bounded sets are weakly sequentially compact)**.**

Suppose that X is a reflexive normed linear space and $A \subseteq X$ is convex, closed and bounded. Then A is weakly sequentially compact.

Proof. Since A is convex and closed, it is weakly sequentially closed by Theorem 5.13. The result follows from Theorem 5.21. □

Example 5.23 (weakly sequentially compact sets)**.**

(*i*) The closed unit ball $\overline{B_1(0)}$ (and other closed balls as well) in a reflexive normed linear space is convex, closed and bounded and thus weakly sequentially compact.

(*ii*) Suppose that $\Omega \subseteq \mathbb{R}^d$ is open. Consider $p \in (1, \infty)$ and $a, b \in \mathbb{R}$ with $a \leqslant b$, the set

$$A := \{f \in L^p(\Omega) \mid a \leqslant f(x) \leqslant b \text{ for a.a. } x \in \Omega\}$$

is convex, closed and bounded in $L^p(\Omega)$ and thus weakly sequentially compact. △

Corollary 5.24 (bounded sequences in reflexive spaces contain weakly convergent subsequences)**.**

Suppose that X is a reflexive normed linear space. Then every bounded sequence $x^{(k)}$ in X contains a weakly convergent subsequence.[24]

Proof. By definition, the bounded sequence $\left(x^{(k)}\right)$ is contained in some closed ball $\overline{B_r(0)}$, which is weakly sequentially compact by Example 5.23. This means that $\left(x^{(k)}\right)$ contains a weakly convergent subsequence (whose weak limit belongs to $\overline{B_r(0)}$). □

§ 5.3 Existence Theorems Using Weak Sequential Compactness

We can now state a general existence result for optimization problems in reflexive normed linear spaces. In comparison to Weierstrass' theorem, we relax the condition of compactness to weak sequential compactness. On the other hand, we tighten the condition of lower semi-continuity to weak sequential lower semi-continuity.

Definition 5.25 (radially unbounded function)**.**

Suppose that X is a normed linear space. A function $f\colon X \to \mathbb{R}$ is called **radially unbounded** if for any sequence $x^{(k)}$ in X,

$$\|x^{(k)}\|_X \to \infty \quad \Rightarrow \quad f(x^{(k)}) \to \infty. \tag{5.7}$$

△

[24] Indeed, the converse also holds: Suppose that X is a Banach space such that every bounded sequence in X contains a weakly convergent subsequence. Then X is reflexive. See for instance Heuser, 1992, Satz 60.6.

We can now prove an existence result for optimization problems in reflexive normed linear spaces that is sometimes referred to as the **direct method** of the calculus of variations. We consider the problem

$$\text{Minimize} \quad f(x), \quad \text{where } x \in X_{\text{ad}} \tag{5.8}$$

with X_{ad} denotes the **admissible set**.

Theorem 5.26 (existence result in reflexive spaces)**.**
Suppose that X is a reflexive normed linear space and that $X_{\text{ad}} \neq \emptyset$ is weakly sequentially closed. Moreover, let $f : X \to \mathbb{R}$ be weakly sequentially lower semi-continuous. If

(*i*) f is radially unbounded, or

(*ii*) X_{ad} is bounded (i. e., weakly sequentially compact by Theorem 5.21),

then the optimization problem (5.8) possesses at least one global minimizer.

Proof. We define $f^* := \inf\{f(x) \mid x \in X_{\text{ad}}\} \in \mathbb{R} \cup \{-\infty\}$ to be the infimal value of (5.8). Suppose that $\left(x^{(k)}\right) \subseteq X_{\text{ad}}$ is a minimizing sequence, i. e., we have $f(x^{(k)}) \searrow f^*$.

Step 1: The sequence $\left(x^{(k)}\right)$ is bounded.

This is clear if X_{ad} is assumed bounded.

Otherwise f is radially unbounded. Suppose that $\left(x^{(k)}\right)$ is not bounded. Then there exists a subsequence $\left(x^{(k^{(\ell)})}\right)$ such that $\left\|x^{(k^{(\ell)})}\right\| \to \infty$. Since f is radially unbounded, $f(x^{(k^{(\ell)})}) \to \infty$, which contradicts the assumption $f(x_k) \to f^* \in \mathbb{R} \cup \{-\infty\}$.

Step 2: We construct the candidate for the global minimizer $x^* \in X_{\text{ad}}$.

By Corollary 5.24, the bounded sequence $\left(x^{(k)}\right)$ contains a weakly convergent subsequence $\left(x^{(k^{(\ell)})}\right)$, whose weak limit we denote by x^*. Since X_{ad} is weakly sequentially closed, $x^* \in X_{\text{ad}}$.

Step 3: We show that $f(x^*) = f^*$.

We have

$$\begin{aligned} f^* &= \inf\{f(x) \mid x \in X_{\text{ad}}\} && \text{by definition of } f^* \\ &\leqslant f(x^*) && \text{since } x^* \in X_{\text{ad}} \\ &\leqslant \liminf_{\ell \to \infty} f\left(x^{(k^{(\ell)})}\right) && \text{by weak sequential lower semi-continuity of } f \\ &= \lim_{\ell \to \infty} f\left(x^{(k)}\right) && \text{since } \lim_{\ell \to \infty} f\left(x^{(k)}\right) \text{ exists} \\ &= f^*. \end{aligned} \tag{5.9}$$

This shows that $f(x^*) = f^*$ and thus $f^* \in \mathbb{R}$, hence $x^* \in X_{\text{ad}}$ is a global minimizer of (5.8). □

Remark 5.27 (existence result in reflexive spaces)**.**
The weak sequential closedness of the admissible set X_{ad} and the weak sequential lower semi-continuity of f may be verified with the help of convexity:

(*i*) If $X_{\text{ad}} \neq \emptyset$ is convex and closed, then it is weakly sequentially closed by Theorem 5.13.

(*ii*) If f is convex and continuous, then it is weakly sequentially lower semi-continuous by Theorem 5.14.

(*iii*) If, in addition, f is strictly convex and X_{ad} is convex, then the global minimizer is unique: Suppose that $x^* \neq x^{**}$ are two distinct global minimizers, then

$$f\left(\frac{x^* + x^{**}}{2}\right) < \frac{1}{2} f(x^*) + \frac{1}{2} f(x^{**}) = f(x^*) = f(x^{**}).$$

Due to the convexity of X_{ad}, the midpoint $(x^* + x^{**})/2$ is contained in X_{ad}, contradicting the global optimality of x^* and x^{**}. △

Example 5.28 (orthogonal projection onto closed convex sets in Hilbert spaces)**.**
Suppose that H is a Hilbert space and that $C \subseteq H$ is nonempty, convex and closed. Given $y \in H$, the problem

$$\text{Minimize} \quad \|x - y\|_H, \quad \text{where } x \in C \tag{5.10}$$

has a unique solution, which is called the **orthogonal projection** of y onto C. △

End of Class 13

Chapter 1 Optimal Control of Partial Differential Equations

§ 6 Introduction

An optimal control problem is a particular type of optimization problem where the optimization variables belong to some space of functions. The optimization variables can be split into a state variable (we typically call y) and a control variable (we typically call u). These variables are coupled through a partial differential equation (PDE) that we formulate as an abstract constraint $e(y,u)=0$ for the moment. We thus obtain an optimization problem of the form

$$\begin{aligned} \text{Minimize} \quad & J(y,u), \quad \text{where } (y,u) \in (Y,U) \\ \text{subject to (s. t.)} \quad & e(y,u)=0 \\ \text{and} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{6.1}$$

Often, the state can be expressed as a function of the control, i. e., $y=G(u)$ with the **control-to-state map** G. This allows us to consider the **reduced formulation** of the optimal control problem, in which the control variable is the only optimization variable:

$$\begin{aligned} \text{Minimize} \quad & f(u) \coloneqq J(G(u),u), \quad \text{where } u \in U \\ \text{s. t.} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{6.2}$$

The reformulation leaves global and local minimizers intact, as shown in the following lemma.

Lemma 6.1. Suppose that Y and U are normed linear spaces.

(i) Suppose that $G\colon U_{\text{ad}} \to Y$ provides, for any $u \in U_{\text{ad}}$, the unique solution $y=G(u)$ of the constraint $e(y,u)=0$.

  (a) If (y^*,u^*) is a global minimizer of the original problem (6.1), then u^* is a global minimizer of the reduced problem (6.2).

  (b) If u^* is a global minimizer of the reduced problem (6.2), then $(G(u^*),u^*)$ is a global minimizer of the original problem (6.1).

(ii) Suppose in addition that $G\colon U_{\text{ad}} \to Y$ is continuous on U_{ad}.

  (a) If (y^*,u^*) is a local minimizer of the original problem (6.1), then u^* is a local minimizer of the reduced problem (6.2).

  (b) If u^* is a local minimizer of the reduced problem (6.2), then $(G(u^*),u^*)$ is a local minimizer of the original problem (6.1).

Proof. □

We will begin with a class of problems that fit into the following framework that could be designated as **linear-quadratic optimal control problems**:

Example 6.2 (framework for linear-quadratic problems)**.**

Suppose that Y and U are Hilbert spaces and that the control-to-state map $G\colon U \to Y$ is linear and bounded. Moreover, suppose that H is another Hilbert space and $E\colon Y \to H$ is a bounded linear (observation) operator. Consider the following reduced optimal control problem:

$$\begin{aligned} \text{Minimize} \quad & f(u) := \frac{1}{2}\|E\,G\,u - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2, \quad \text{where } u \in U \\ \text{s. t.} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{6.3}$$

Here $U_{\text{ad}} \subseteq U$ is the **admissible set** of controls. One refers to the first term in the objective as a **tracking term** and $z \in H$ is the **target observation** or **desired observation**. The second term in the objective is called a **cost cost term** and the number $\gamma \geqslant 0$ is the **control cost parameter**. It balances both terms in the objective, which are usually competing.

One often combines the control-to-state map G with the observation map E into the **control-to-observation** (or **control-to-observable**) map $S := E \circ G\colon U \to H$. △

We can use Theorem 5.26 to prove the existence/uniqueness of a global minimizer for problem (6.3). Notice that the objective in (6.3) is convex on U. When $U_{\text{ad}} \subseteq U$ is a convex set, then (6.3) is a convex optimization problem. This means that we do not need to worry about local minimizers, since every local minimizer is also a global minimizer.

Theorem 6.3 (existence theorem for linear-quadratic problems)**.**

Suppose that U_{ad} is nonempty, convex and closed and that

(*i*) $\gamma > 0$ holds for the control cost parameter, or

(*ii*) U_{ad} is bounded.

Then the optimal control problem (6.3) possesses a global minimizer. In case $\gamma > 0$, the global minimizer is also unique.

Proof. We verify that Theorem 5.26 is applicable. The control space U is reflexive as a Hilbert space. The admissible set U_{ad} is convex and closed, hence weakly sequentially closed by Theorem 5.13. The objective is convex and continuous, hence weakly sequentially lower semi-continuous by Theorem 5.14. When $\gamma > 0$ holds, then the objective is radially unbounded. Theorem 5.26 now implies the existence of a global minimizer. When $\gamma > 0$, then the objective is in fact strictly convex, and thus the global minimizer is unique. □

§ 7 Floor-Heating Problem

In this section we consider a first example of an optimal control problem for a partial differential equation (PDE). One of the simplest PDEs is the Poisson equation, which models the stationary state of a diffusion process. In terms of the diffusion of thermal energy, the equation reads

$$-\operatorname{div}\big(\kappa\,\nabla T(x)\big) = q(x) \quad \text{in } \Omega. \tag{7.1}$$

Here $\Omega \subseteq \mathbb{R}^3$ is an open bounded set that represents the domain of the heat distribution. Moreover, κ denotes the thermal conductivity (unit: $\mathrm{W\,m^{-1}\,K^{-1}}$), T is the temperature (unit: K), and q is the heat source (unit: $\mathrm{W\,m^{-3}}$). In physical terms, the equation states that the divergence of the heat flux $\kappa\,\nabla T$ equals the heat source q. When κ is spatially constant, then it can be pulled out and (7.1) becomes

$$-\kappa\,\Delta T(x) = q(x) \quad \text{in } \Omega,$$

where

$$\Delta T \coloneqq \sum_{i=1}^{d} \frac{\partial^2 T}{\partial x_i^2}$$

denotes the Laplace operator and $d \in \mathbb{N}$ is the dimension of the domain Ω.

The heat equation (7.1) needs to be accompanied by boundary conditions that describes the heat exchange with the environment. We consider here a so-called **Robin boundary condition** that states that the heat flux across the boundary is proportional to the temperature difference between the domain and the environment:

$$\kappa\,\frac{\partial}{\partial n} T(x) = \alpha(x)\,\big(T_\infty(x) - T(x)\big) \quad \text{on } \Gamma \coloneqq \partial\Omega. \tag{7.2}$$

Here n is the outer unit normal vector on the boundary γ. Moreover, α is the heat transfer coefficient (unit: $\mathrm{W\,m^{-2}\,K^{-1}}$) that depends on the material of the boundary and the environment. Finally, T_∞ is the temperature of the environment (unit: K).

We now formulate an optimal control problem in which the temperature T is the state variable and will be denoted by y by consistency with the general framework. The heat source q serves as the control variable and will be denoted by u. The control will act on a "subdomain" (any measurable subset) $\Omega_{\mathrm{ctrl}} \subseteq \Omega$ and will be set to zero in $\Omega \setminus \Omega_{\mathrm{ctrl}}$.

The objective features a tracking term that measures the deviation of the temperature from a desired temperature distribution y_d on another subdomain (measurable subset) $\Omega_{\mathrm{obs}} \subseteq \Omega$. In addition, we have a control cost term that penalizes non-zero values of the heat source.

Overall, we obtain the following optimal control problem:

$$\begin{aligned}
\text{Minimize} \quad & \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_{\mathrm{obs}})} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega_{\mathrm{ctrl}})} \\
\text{s. t.} \quad & \begin{cases} -\operatorname{div}\big(\kappa\,\nabla y\big) = \chi_{\mathrm{ctrl}}\,u & \text{in } \Omega \\ \kappa\,\dfrac{\partial}{\partial n} y = \alpha\,(y_\infty - y) & \text{on } \Gamma \end{cases} \\
\text{and} \quad & u \in U_{\mathrm{ad}} \coloneqq \{u \in L^2(\Omega_{\mathrm{ctrl}}) \mid u_a \leqslant u \leqslant u_b \text{ a.e. in } \Omega_{\mathrm{ctrl}}\}.
\end{aligned} \tag{7.3}$$

As a motivation for problem (7.3), we may think of a scenario where the heat source is realized through floor heating panels which can be controlled pointwise. The heating panels occupy the subset Ω_{ctrl}. The control bounds u_a and u_b model technical constraints on the heat source. The objective expresses the desire to maintain a desired temperature distribution y_d in a certain area $\Omega_{\text{obs}} \subseteq \Omega$.

In order to put problem (7.3) into the framework of § 6 we will have to discuss the functional analytic setting of the heat equation and the boundary condition.

End of Class 14

End of Week 8

§ 7.1 Weak Formulation of the Heat Equation

The heat equation in (7.1) is in its so-called **strong formulation**. In order to satisfy (7.1) in a pointwise sense, the temperature distribution $y\colon \Omega \to \mathbb{R}$ would need to be twice differentiable, or at least twice differentiable in the weak sense (Definition 2.33). This is often too restrictive. We therefore proceed to the **weak formulation** of (7.1)–(7.2) with the intention of generalizing the concept of a solution. To this end, we multiply (7.1) with a test function $v \in H^1(\Omega)$ and integrate over the domain Ω, obtaining

$$-\int_\Omega \operatorname{div}(\kappa \nabla y)\, v \,\mathrm{d}x = \int_\Omega q\, v \,\mathrm{d}x.$$

We now perform integration by parts[1] on the left-hand side, which yields

$$\int_\Omega \kappa \nabla y \cdot \nabla v \,\mathrm{d}x - \int_\Gamma \kappa \frac{\partial y}{\partial n} v \,\mathrm{d}s = \int_\Omega q\, v \,\mathrm{d}x.$$

Plugging in the boundary condition (7.2), i. e.,

$$\kappa \frac{\partial}{\partial n} y = \alpha\,(y_\infty - y) \quad \text{on } \Gamma := \partial\Omega$$

yields the **weak formulation** of (7.1)–(7.2):

$$\begin{aligned} &\text{Find} && y \in H^1(\Omega) \\ &\text{s. t.} && \int_\Omega \kappa \nabla y \cdot \nabla v \,\mathrm{d}x + \int_\Gamma \alpha\, y\, v \,\mathrm{d}s = \int_\Omega q\, v \,\mathrm{d}x + \int_\Gamma \alpha\, y_\infty\, v \,\mathrm{d}s \quad \text{for all } v \in H^1(\Omega). \end{aligned} \tag{7.4}$$

In comparison with the strong formulation, we now only need the temperature (state) y to have first-order instead of second-order weak derivatives. One derivative has been passed on to the test function v by the integration by parts.

Later on we are going to plug in that the right-hand side has the form $q = \chi_{\text{ctrl}}\, u$ but we keep it open at the moment.

[1] The special case $v \equiv 1$ is known as the Gauss divergence theorem, $\int_\Omega \boldsymbol{F} \,\mathrm{d}x = \int_\Gamma \boldsymbol{F} \cdot n \,\mathrm{d}s$. We have $\boldsymbol{F} = \kappa \nabla y$.

§ 7.2 The Trace Operator

We need to discuss how the term $\int_\Gamma \alpha\, y\, v \, \mathrm{d}s$ in (7.4) is to be interpreted. To this end, we first need to assume a certain regularity for the boundary Γ of our bounded open set Ω.

Definition 7.1 (Lipschitz boundary)**.**
Suppose that $\Omega \subset \mathbb{R}^d$, $d \geqslant 2$ is open and bounded and that $\Gamma := \partial\Omega$ is its topological boundary. The boundary Γ is said to be **Lipschitz**[2] if Γ can be written as the finite union of pieces Γ_j, such that Γ_j is the graph of a Lipschitz continuous function defined on a $(d-1)$-dimensional cube in a suitable coordinate system. Moreover, Ω may locally lie only on one side of Γ. △

For a more detailed definition of Lipschitz boundaries, see for instance Tröltzsch, 2010, Chapter 2.2.2.

The representation of parts of the boundary as the graphs of Lipschitz continuous functions allows us to define a Lebesgue measure on the boundary.[3] We denote the integration w.r.t. this measure by $\mathrm{d}s$ to distinguish it from the volume measure $\mathrm{d}x$.

Functions in $C(\operatorname{cl}\Omega)$ naturally have a trace operator $\tau\colon C(\operatorname{cl}\Omega) \to C(\Gamma)$ defined for them since the restriction of a continuous function is continous. (**Quiz 7.1:** Can you prove that?) Also, the trace operator is linear. The following theorem shows that τ extends to the Sobolev spaces $W^{1,p}(\Omega)$.

Theorem 7.2 (Sobolev trace theorem (Gagliardo, 1957, Evans, 1998, Chapter 5.5))**.**
Suppose that $\Omega \subset \mathbb{R}^d$, $d \geqslant 2$ is open and bounded with Lipschitz boundary Γ. Then the trace map $\tau\colon C(\operatorname{cl}\Omega) \cap W^{1,p}(\Omega) \to C(\Gamma)$ extends in a unique way to a continuous linear operator $\tau\colon W^{1,p}(\Omega) \to L^p(\Gamma)$ for $1 \leqslant p < \infty$.

The fact that the case $p = \infty$ is not covered is not problematic since the trace operator actually has much better properties than suggested by Theorem 7.2.

Theorem 7.3 (Mapping properties of the trace map (Nečas, 2012, Theorem 2.4.2, Tröltzsch, 2010, Theorem 7.2))**.**
Suppose that $\Omega \subset \mathbb{R}^d$, $d \geqslant 2$ is open and bounded with Lipschitz boundary Γ. Moreover, let $1 \leqslant p \leqslant \infty$. Then the trace map has the following mapping properties:

(*i*) In case $p < d$:
$$\tau\colon W^{1,p}(\Omega) \to L^r(\Gamma) \quad \text{for all } 1 \leqslant r \leqslant \frac{(d-1)\,p}{d-p}.$$
Note: The upper bound is $\geqslant p$ and thus this result strengthens Theorem 7.2.

(*ii*) In case $p = d$:
$$\tau\colon W^{1,p}(\Omega) \to L^r(\Gamma) \quad \text{for all } 1 \leqslant r < \infty.$$

(*iii*) In case $p > d$:
$$\tau\colon W^{1,p}(\Omega) \to C(\Gamma).$$

Remark 7.4 (on the trace map)**.**

[2] Likewise, we say that Ω is a set with Lipschitz boundary.
[3] see for instance Nečas, 2012, Chapter 2.4

(*i*) Important special cases of Theorem 7.3 comprise $\tau\colon H^1(\Omega) \to L^r(\Gamma)$ for all $r \in [1,\infty)$ in dimension $d = 2$ and for all $r \in [1,4]$ in dimension $d = 3$.

(*ii*) The trace map cannot be continuously extended to any $L^p(\Omega)$. In order words, L^p functions do not necessarily have a well-defined trace.

(*iii*) The trace map $\tau\colon W^{1,p}(\Omega) \to L^r(\Gamma)$ or $\tau\colon W^{1,p}(\Omega) \to C(\Gamma)$ (as in Theorem 7.3) is not surjective. That is, there exist functions in $L^r(\Gamma)$ or $C(\Gamma)$ that are not the trace of any function in $W^{1,p}(\Omega)$.

(*iv*) The trace map is also not injective. For instance, all functions in $W^{1,p}(\Omega)$ that are zero near the boundary have the same (zero) trace. △

With the help of the trace map we can now understand the term $\int_\Gamma \alpha\, y\, v \,\mathrm{d}s$ in (7.4). In fact, the trace of the $H^1(\Omega)$-functions y and v are well-defined in $L^r(\Gamma)$ for some $r \geqslant 2$ (depending on the dimension d). The triple product $\alpha\, y\, v \,\mathrm{d}s$ then belongs to $L^1(\Gamma)$ and the integral is well-defined (using Hölder's inequality Lemma 2.24), under an appropriate integrability assumption on the coefficient α. For instance, $\alpha \in L^\infty(\Gamma)$ would be sufficient in any dimension. In dimension $d = 3$, $\alpha \in L^2(\Gamma)$ would be sufficient since $y, v \in L^4(\Gamma)$ and $1/4 + 1/4 + 1/2 = 1$. In dimension $d = 2$, $\alpha \in L^{1+\varepsilon}(\Gamma)$ would be sufficient (for some $\varepsilon > 0$) since $y, v \in L^r(\Gamma)$ holds for any $1 \leqslant r < \infty$.

§ 7.3 The Lax-Milgram Lemma

The left-hand side of our example problem's weak formulation (7.4) is a (symmetric) bilinear form on $H^1(\Omega) \times H^1(\Omega)$:

$$a(y,v) := \int_\Omega \kappa\, \nabla y \cdot \nabla v \,\mathrm{d}x + \int_\Gamma \alpha\, y\, v \,\mathrm{d}s. \tag{7.5a}$$

The right-hand side

$$F(v) := \int_\Omega q\, v \,\mathrm{d}x + \int_\Gamma \alpha\, y_\infty\, v \,\mathrm{d}s \tag{7.5b}$$

defines a linear form on $H^1(\Omega)$.

The Lax-Milgram lemma provides sufficient conditions under which an abstract weak formulation (also known as **variational formulation**)

$$\begin{aligned} &\text{Find} && y \in H \\ &\text{s.t.} && a(y,v) = F(v) \quad \text{for all } v \in H \end{aligned} \tag{7.6}$$

possesses a unique solution.

Theorem 7.5 (Lax-Milgram lemma)**.**
Suppose that H is a Hilbert space and that $a\colon H \times H \to \mathbb{R}$ is a (not necessarily symmetric) bilinear form with the following properties: there exist $\alpha_0, \beta_0 > 0$ such that

$$|a(y,v)| \leqslant \alpha_0 \|y\|_H \|v\|_H \quad \textbf{boundedness} \tag{7.7a}$$

$$a(y,y) \geqslant \beta_0 \|y\|_H^2 \quad \textbf{ellipticity} \tag{7.7b}$$

holds for all $y, v \in H$. Then, for any $F \in H^*$, problem (7.6) possesses a unique solution $y \in H$. This solution satisfies the **a-priori estimate** $\|y\|_H \leqslant \frac{1}{\beta_0} \|F\|_{H^*}$. That is, the linear right-hand-side-to-solution map

$$H^* \ni F \mapsto y \in H$$

is bounded with boundedness constant $1/\beta_0$.

Just like for linear forms, one can show that the boundedness of a bilinear form (on normed linear spaces) is equivalent to its continuity. On a finite dimensional normed linear space, the ellipticity of a bilinear form is equivalent to its positive definiteness. On infinte-dimensional spaces, however, ellipticity is a stronger property than positive definiteness.

Proof of Theorem 7.5. **Step** 1: We associate with the bilinear form a a bounded linear operator $A\colon H \to H^*$ by setting $A\,y := a(y,\cdot)$.

Indeed, A maps into H^* since we have for any $y \in H$

$$|a(y,v)| \leqslant \alpha_0 \,\|y\|_H \,\|v\|_H$$

and thus

$$\|A\,y\|_{H^*} = \|a(y,\cdot)\|_{H^*} \leqslant \alpha_0 \,\|y\|_H.$$

The mapping $H \ni y \mapsto a(y,\cdot) \in H^*$ is linear since a is bilinear, and thus $A\colon H \to H^*$ is linear. The above estimate $\|A\,y\|_{H^*} \leqslant \alpha_0 \,\|y\|_H$ confirms that $A \in \mathcal{L}(H, H^*)$.

Step 2: We formulate the variational problem (7.6) as the linear equation $A\,y = F$ with $A\colon H \to H^*$.

Step 3: We further reformulate the problem as a fixed-point problem $y = T_\delta\, y$ in the space H, where $\delta > 0$ is a parameter to be determined.

To this end, we use the Riesz isomorphism $R\colon H \to H^*$ and define

$$H \ni y \mapsto T_\delta\, y := y - \delta\, R^{-1}\,(A\,y - F) \in H.$$

Then $y = T_\delta\, y$ is equivalent to $A\,y = F$.

Step 4: We show that T_δ is a contraction for small enough $\delta > 0$.

We estimate

$$\begin{aligned}
\|T_\delta\, y - T_\delta\, z\|_H^2 &= \|(y-z) - \delta\, R^{-1}\,(A\,y - A\,z)\|_H^2 \\
&= \|y-z\|_H^2 - 2\,\delta\,\left(R^{-1}\,(A\,y - A\,z),\, y-z\right) + \delta^2\,\|R^{-1}\,(A\,y - A\,z)\|_H^2 \\
&= \|y-z\|_H^2 - 2\,\delta\, a(y-z,\,y-z) + \delta^2\,\|(A\,y - A\,z)\|_{H^*}^2 \\
&\leqslant \left[1 - 2\,\delta\,\beta_0 + \delta^2\alpha_0^2\right]\|y-z\|_H^2.
\end{aligned}$$

When we choose $\delta \in (0, \frac{2\,\beta_0}{\alpha_0^2})$, the factor in square brackets is less than one. Banach's fixed point theorem now implies the existence of a unique fixed point $y = T_\delta\, y$, i. e., the existence of a unique solution of $A\,y = F$ or $a(y,v) = F(v)$ for all $v \in V$.

Step 5: We plug in $v = y$ into the variational equation to obtain

$$\beta_0\,\|y\|_H^2 \leqslant a(y,y) = F(y) \leqslant \|F\|_{H^*}\,\|y\|_H.$$

This shows $\|y\|_H \leqslant \frac{1}{\beta_0}\,\|F\|_{H^*}$.

Finally, since $A\,y = F$ is a linear equation, the right-hand-side-to-solution map $H^* \ni F \mapsto y \in H$ is linear, and the above estimate shows that its operator norm is bounded by $1/\beta_0$. □

End of Class 15

End of Week 9

§ 7.4 Control-to-State Map and Reduced Formulation of the Floor-Heating Problem

We are now in a position to discuss the unique solvability of the heat equation (7.4) under appropriate assumptions.

Assumption 7.6 (for the heat equation (7.4))**.**

(*i*) Suppose that $\Omega \subseteq \mathbb{R}^d$ is a bounded open, connected set with Lipschitz boundary Γ.[4]

(*ii*) The thermal conductivity coefficient $\kappa \in L^\infty(\Omega)$ satisfies $\kappa \geqslant \kappa_0 > 0$ a.e. in Ω for some $\kappa_0 > 0$.

(*iii*) The heat transfer coefficient $\alpha \in L^\infty(\Gamma)$ satisfies $\alpha \geqslant 0$ a.e. in Γ but $\|\alpha\|_{L^\infty(\Gamma)} \neq 0$. △

(**Quiz 7.2:** Why would $\alpha \equiv 0$ not lead to a well-posed problem?)

In order to show the ellipticity of the bilinear form (7.5a), we require the following lemma.

Lemma 7.7 (generalized Friedrich's inequality)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is a bounded domain with Lipschitz boundary Γ. Moreover, suppose that $\Gamma_1 \subseteq \Gamma$ is part of the boundary with positive boundary measure.[5] Then there exists a constant $C > 0$ such that

$$\begin{aligned} \|y\|^2_{H^1(\Omega)} &\leqslant C\left(\int_\Omega |\nabla y|^2 \,\mathrm{d}x + \frac{1}{|\Gamma_1|}\left(\int_{\Gamma_1} y \,\mathrm{d}s\right)^2\right) \\ &\leqslant C\left(\int_\Omega |\nabla y|^2 \,\mathrm{d}x + \int_{\Gamma_1} y^2 \,\mathrm{d}s\right) \\ &= C\left(\|\nabla y\|^2_{L^2(\Omega)} + \|y\|^2_{L^2(\Gamma_1)}\right). \end{aligned} \tag{7.8}$$

Proof. We will not be showing the inequality itself. We will only convince ourselves of the second inequality. Indeed, the triangle and Hölder's inequalities give

$$\int_{\Gamma_1} y \,\mathrm{d}s \leqslant \int_{\Gamma_1} |y| \,\mathrm{d}s = \|y\|_{L^1(\Gamma_1)} \leqslant |\Gamma_1|^{1/2} \|y\|_{L^2(\Gamma_1)}.$$

Squaring this inequality gives

$$\left(\int_{\Gamma_1} y \,\mathrm{d}s\right)^2 \leqslant |\Gamma_1| \,\|y\|^2_{L^2(\Gamma_1)},$$

which shows the second inequality. □

We can now show the well-posedness of the heat equation (7.4) under Assumption 7.6. We consider $C > 0$ a generic constant that may change from line to line.

Lemma 7.8 (well-posedness of the heat equation (7.4))**.**
Suppose that Assumption 7.6 holds.

(*i*) The bilinear form (7.5a) is bounded and elliptic on $H^1(\Omega) \times H^1(\Omega)$.

[4] An open connected set is often referred to as a **domain**. So we require here that $\Omega \subseteq \mathbb{R}^d$ is a bounded domain with Lipschitz boundary.

[5] For example, Γ_1 can be a relatively open subset of the boundary.

(*ii*) For every $q \in L^2(\Omega)$ and $y_\infty \in L^2(\Gamma)$, the linear form (7.5b) is bounded on $H^1(\Omega)$. It satisfies

$$|F(v)| \leqslant \|q\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|\alpha\|_{L^\infty(\Gamma)} \|y_\infty\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)}$$

and thus there exists a constant $C > 0$ such that

$$\|F\|_{H^1(\Omega)^*} \leqslant \|q\|_{L^2(\Omega)} + C \|\alpha\|_{L^\infty(\Gamma)} \|y_\infty\|_{L^2(\Gamma)}.$$

(*iii*) Consequently, the variational formulation (7.6) has a unique solution $y \in H^1(\Omega)$ that satisfies the a-priori estimate

$$\|y\|_{H^1(\Omega)} \leqslant C \left(\|q\|_{L^2(\Omega)} + \|y_\infty\|_{L^2(\Gamma)} \right).$$

Proof. Assumption (*i*): We define $\overline{\kappa} := \|\kappa\|_{L^\infty(\Omega)}$ and $\overline{\alpha} := \|\alpha\|_{L^\infty(\Gamma)}$. We estimate

$$\begin{aligned}
|a(y,v)| &\leqslant \left| \int_\Omega \kappa \, \nabla y \cdot \nabla v \, \mathrm{d}x \right| + \left| \int_\Gamma \alpha \, y \, v \, \mathrm{d}s \right| && \text{by the triangle inequality} \\
&\leqslant \overline{\kappa} \int_\Omega |\nabla y \cdot \nabla v| \, \mathrm{d}x + \overline{\alpha} \int_\Gamma |y \, v| \, \mathrm{d}s && \text{by the triangle inequality for integrals} \\
&\leqslant \overline{\kappa} \int_\Omega |\nabla y|_2 |\nabla v|_2 \, \mathrm{d}x + \overline{\alpha} \int_\Gamma |y| \, |v| \, \mathrm{d}s && \text{by the Cauchy-Schwarz inequality} \\
&\leqslant \overline{\kappa} \, \|\nabla y\|_{L^2(\Omega)} \, \|\nabla v\|_{L^2(\Omega)} + \overline{\alpha} \, \|y\|_{L^2(\Gamma)} \, \|v\|_{L^2(\Gamma)} \\
&\leqslant \overline{\kappa} \, \|y\|_{H^1(\Omega)} \, \|v\|_{H^1(\Omega)} + C \, \overline{\alpha} \, \|y\|_{H^1(\Omega)} \, \|v\|_{H^1(\Omega)}.
\end{aligned}$$

The last inequality follows from the definition of the H^1-norm and the boundedness of the trace operator $H^1(\Omega) \to L^2(\Gamma)$. So the boundedness constant of the bilinear form has the form $\alpha_0 = \overline{\kappa} + C \, \overline{\alpha}$.

For the ellipticity, we observe that Assumption 7.6 implies that there exists $\delta > 0$ and a subset $\Gamma_1 \subseteq \Gamma$ with positive measure such that $\alpha \geqslant \delta$ holds a.e. on Γ_1. **(Quiz 7.3:** How do you prove that?) We can now estimate

$$\begin{aligned}
a(y,y) &= \int_\Omega \kappa \, |\nabla y|_2^2 \, \mathrm{d}x + \int_\Gamma \alpha \, |y|^2 \, \mathrm{d}s \\
&\geqslant \kappa_0 \int_\Omega |\nabla y|_2^2 \, \mathrm{d}x + \delta \int_{\Gamma_1} |y|^2 \, \mathrm{d}s && \text{by monotonicity of the integral} \\
&\geqslant \min\{\kappa_0, \delta\} \left(\|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Gamma_1)}^2 \right) \\
&\geqslant \frac{\min\{\kappa_0, \delta\}}{C} \, \|y\|_{H^1(\Omega)}^2 && \text{by the generalized Friedrich's inequality Lemma 7.7.}
\end{aligned}$$

So the ellipticity constant of the bilinear form has the form $\beta_0 = \frac{\min\{\kappa_0, \delta\}}{C}$.

Assumption (*ii*): We estimate

$$\begin{aligned}
|F(v)| &\leqslant \left| \int_\Omega q \, v \, \mathrm{d}x \right| + \left| \int_\Gamma \alpha \, y_\infty \, v \, \mathrm{d}s \right| && \text{by the triangle inequality} \\
&\leqslant \|q\|_{L^2(\Omega)} \, \|v\|_{L^2(\Omega)} + \overline{\alpha} \, \|y_\infty\|_{L^2(\Gamma)} \, \|v\|_{L^2(\Gamma)} && \text{by the Hölder inequality} \\
&\leqslant \|q\|_{L^2(\Omega)} \, \|v\|_{H^1(\Omega)} + C \, \overline{\alpha} \, \|y_\infty\|_{L^2(\Gamma)} \, \|v\|_{H^1(\Omega)}.
\end{aligned}$$

The last inequality follows again from the definition of the H^1-norm and the boundedness of the trace operator $H^1(\Omega) \to L^2(\Gamma)$. We thus have

$$\|F\|_{H^1(\Omega)^*} \leqslant \|q\|_{L^2(\Omega)} + C\,\|\alpha\|_{L^\infty(\Gamma)}\,\|y_\infty\|_{L^2(\Gamma)}$$

as claimed.

Assumption *(iii)*: The Lax-Milgram theorem 7.5 now confirms that the variational problem (7.6) has a unique solution $y \in H^1(\Omega)$ for every heat source $q \in L^2(\Omega)$ and outside temperature $y_\infty \in L^2(\Gamma)$. This solution satisfies the a-priori estimate

$$\|y\|_{H^1(\Omega)} \leqslant \frac{1}{\beta_0}\,\|F\|_{H^1(\Omega)^*},$$

i. e., it is of the form

$$\|y\|_{H^1(\Omega)} \leqslant C\,\big(\|q\|_{L^2(\Omega)} + \|y_\infty\|_{L^2(\Gamma)}\big).$$ □

We recall that the control u in the floor heating problem enters only into the heat source q, while we keep the outside temperature y_∞ fixed. We also recall that the heat source q is determined by the control via $q = \chi_{\text{ctrl}}\,u$ with $u \in L^2(\Omega_{\text{ctrl}})$. Thanks to Lemma 7.8, we can now introduce the control-to-state map

$$G\colon L^2(\Omega_{\text{ctrl}}) \ni u \mapsto y \in H^1(\Omega), \tag{7.9}$$

where y is the unique solution to (7.4) with heat source $q = \chi_{\text{ctrl}}\,u$ and outside temperature y_∞.

In fact, since the heat equation (7.4) is linear in the state variable y, a superposition principle applies. Let us denote by y_0 the unique solution of the heat equation (7.4) with heat source $q = 0$. Moreover, let us denote by G_0 the control-to-solution map of (7.4) with $y_\infty = 0$. Then we see that

$$G(u) = G_0(u) + y_0$$

holds, i. e., the control-to-state map is an affine function of the control u.

In order to fit the reduced form of our floor heating optimal control problem (7.3) into the general quadratic framework (6.3), we need to introduce the linear and bounded observation map $E\colon H^1(\Omega) \to L^2(\Omega_{\text{obs}})$, simply be the embedding $H^1(\Omega) \to L^2 1(\Omega)$ and restriction to the subdomain Ω_{obs}, i. e.,

$$E\,y := y|_{\Omega_{\text{obs}}}.$$

With these settings, the floor heating problem (7.3), i. e.,

$$\begin{aligned}
\text{Minimize}\quad & \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_{\text{obs}})} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega_{\text{ctrl}})} \\
\text{s. t.}\quad & \begin{cases} -\operatorname{div}\big(\kappa\,\nabla y\big) = \chi_{\text{ctrl}}\,u & \text{in } \Omega \\ \qquad \kappa\,\dfrac{\partial}{\partial n}\,y = \alpha\,(y_\infty - y) & \text{on } \Gamma \end{cases} \\
\text{and}\quad & u \in U_{\text{ad}} := \{u \in L^2(\Omega_{\text{ctrl}}) \mid u_a \leqslant u \leqslant u_b \text{ a.e. in } \Omega_{\text{ctrl}}\}.
\end{aligned}$$

with the state equation understood in weak form (7.4), can be cast into the general quadratic form (6.3), i. e.,

$$\begin{aligned}
\text{Minimize}\quad & f(u) := \frac{1}{2}\|E\,G(u) - z\|^2_H + \frac{\gamma}{2}\|u\|^2_U, \quad \text{where } u \in U \\
\text{s. t.}\quad & u \in U_{\text{ad}}.
\end{aligned} \tag{7.10}$$

The control space is $U = L^2(\Omega_{\text{ctrl}})$ and the observation space is $H = L^2(\Omega_{\text{obs}})$. In fact, in (6.3) we assumed a linear (not affine) control-to-state map G, but this is not a restriction since we could also write the problem in the form

$$\begin{aligned} \text{Minimize} \quad & f(u) := \frac{1}{2}\|E\,G_0(u) + E\,y_0 - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2, \quad \text{where } u \in U \\ \text{s. t.} \quad & u \in U_{\text{ad}}. \end{aligned}$$

We may now invoke the existence theorem 6.3 for linear-quadratic problems to derive the following result. In addition to Assumption 7.6, we formulate:

Assumption 7.9 (for the floor heating problem (7.3))**.**

(*i*) Suppose that $\Omega_{\text{ctrl}} \subseteq \Omega$ and $\Omega_{\text{obs}} \subseteq \Omega$ are measurable subsets of Ω.

(*ii*) Suppose that the desired temperature $y_d \in L^2(\Omega_{\text{obs}})$.

(*iii*) Suppose that $u_a \in \mathbb{R} \cup \{-\infty\}$ and $u_b \in \mathbb{R} \cup \{\infty\}$ are given constants with $u_a \leqslant u_b$.

(*iv*) Suppose that $\gamma \geqslant 0$. If one or both of the bounds u_a and u_b are infinite, then suppose $\gamma > 0$. △

Theorem 7.10. Suppose that Assumption 7.6 and Assumption 7.9 hold. Then the floor heating problem (7.3) in its reduced form (7.10) possesses at least one (globally optimal) solution $u \in U_{\text{ad}}$. In case $\gamma > 0$, the solution is unique.

End of Class 16

Proof. Assumption 7.6 ensures that the control-to-state map G is well-defined so that the reduced form (7.10) is defined. The set U_{ad} is nonempty, closed and convex. When both bounds are finite, U_{ad} is also bounded; otherwise, we have $\gamma > 0$. The result now follows from the existence theorem 6.3 for linear-quadratic problems. □

Remark 7.11 (types of optimal control problems)**.**
The floor heating problem (7.3) is an example of a linear-quadratic optimal control problem with **distributed control**, since the control acts through the heat source in (part of) the domain Ω. The problem also features **distributed observation**, since the observation is taken in (part of) the domain Ω.

Other optimal control problems might consider **boundary control**, where the control acts through the boundary of the domain. This could be either through the right-hand side of a Robin boundary condition as in **Robin boundary control**

$$\kappa \frac{\partial}{\partial n} y = \alpha\,(u - y),$$

or through the right-hand side of a Neumann boundary condition as in **Neumann boundary control**

$$\kappa \frac{\partial}{\partial n} y = u,$$

or even through the right-hand side of a Dirichlet boundary condition as in **Dirichlet boundary control**

$$y = u.$$

Of course, the control may possibly act only on a part of the boundary.

In terms of the observation, we may also consider **boundary observation** problems, by taking the observation map as $E = \tau$ (the trace), or by taking flux observations on the boundary as in $E = \kappa \frac{\partial}{\partial n}$.

In any case, one needs to carefully choose the control and observation spaces and study the well-posedness of the control-to-state and the observation maps in order to obtain a well-defined problem. △

§ 8 Differentiability in Normed Linear Spaces

Our next goal is to derive first-order optimality conditions for linear-quadratic problems (6.3) and the (reduced) floor-heating problem (7.10) in particular.

Throughout this section, suppose that U and V are normed linear spaces, and $F\colon U \to V$ any map.

We consider three differentiability concepts of increasing strength:

Definition 8.1 (directional differentiability)**.**
Consider a point $u \in U$ and a direction $\delta u \in U$. If it exists, the limit

$$F'(u;\delta u) := \lim_{t \searrow 0} \frac{F(u + t\,\delta u) - F(u)}{t} \quad \text{in } V \tag{8.1}$$

is called the **(one-sided) directional derivative** of F at the point u in the direction δu, and F is said to be **directionally differentiable** at u in the direction of δu. △

Example 8.2 (directional differentiability)**.**

(*i*) The map $F\colon \mathbb{R} \to \mathbb{R}$ given by $F(u) = \max\{u, 0\}$ is everywhere directionally differentiable in all directions. In fact, we have

$$F'(u;\delta u) = \begin{cases} \delta u & \text{if } u > 0, \\ 0 & \text{if } u < 0, \\ \max\{\delta u, 0\} & \text{if } u = 0. \end{cases}$$

(*ii*) The map $F\colon \mathbb{R} \to \mathbb{R}$ given by $F(u) = |u|$ is everywhere directionally differentiable in all directions. In fact, we have

$$F'(u;\delta u) = \begin{cases} \delta u & \text{if } u > 0, \\ -\delta u & \text{if } u < 0, \\ |\delta u| & \text{if } u = 0. \end{cases}$$ △

These examples show that even if all directional derivatives exist at a point u, the map $\delta u \mapsto F'(u;\delta u)$ is not necessarily linear. However, it is not difficult to see that directional derivatives are positively homogeneous, i. e., if $F'(u;\delta u)$ exists, then the directional derivative also exists for all $\alpha\,\delta u$ with $\alpha \geqslant 0$, and we have

$$F'(u;\alpha\,\delta u) = \alpha\,F'(u;\delta u) \quad \text{for all } \alpha \geqslant 0.$$

Definition 8.3 (Gâteaux differentiability)**.**
Consider a point $u \in U$. If the directional derivatives $F'(u;\delta u)$ exist for all directions $\delta \in U$ and if

$$F'(u;\delta u) = A\,\delta u \tag{8.2}$$

holds for some bounded linear operator $A \in \mathcal{L}(U,V)$, then A is said to be the **Gâteaux derivative** of F at the point u, and F is said to be **Gâteaux differentiable** at u. We also write $F'_G(u) = A$ in this case. △

Notice that even in finite dimensional spaces, Gâteaux differentiability is more than the existence of all partial derivatives (Jacobian matrix), i. e., the (two-sided) directional differentiability w.r.t. all directions from some basis of U. The reason is that the existence of these partial derivatives does not guarantee that the directional derivative in any other direction also exist. And even when they do exist, they are not necessarily linear.

Note: For $f : U \to \mathbb{R}$, the Gâteaux derivative A is an element of $\mathcal{L}(U,\mathbb{R}) = U^*$, the dual space of U.

Definition 8.4 (Fréchet differentiability)**.**
Consider a point $u \in U$. F is said to be **Fréchet differentiable** at u if there exists a bounded linear operator $A \in \mathcal{L}(U,V)$ and a map $r : U \to V$ such that

$$F(u+\delta u) - F(u) = A\,\delta u + r(\delta u) \quad \text{for all } \delta u \in U \tag{8.3a}$$

holds, where the **remainder function** satisfies

$$\lim_{\delta u \to 0} \frac{\|r(\delta u)\|_V}{\|\delta u\|_U} = 0. \tag{8.3b}$$

In this case, we write $F'(u) = A$ for the **Fréchet derivative**. △

Note: When a function F is Fréchet differentiable at u, then F is Gâteaux differentiable at u and $F'(u) = F'_G(u)$ holds.

End of Class 17

End of Week 10

Remark 8.5 (differentiability concepts)**.**

(*i*) Fréchet differentiability amounts to Gâteaux differentiability plus the property (8.3b) about the remainder function. It is the strongest form of differentiability.

(*ii*) Often one starts by proving directional differentiability properties. If all directional derivatives exist, one can check for linearity and continuity w.r.t. the direction, i. e., Gâteaux differentiability. Once Gâteaux differentiability is established, one can check for the remainder property, i. e.,

$$\lim_{\delta u \to 0} \frac{\|F(u+\delta u) - F(u) - A\,\delta u\|_V}{\|\delta u\|_U} = 0.$$

We will use this procedure in Example 8.7 below.

(*iii*) In particular in infinite-dimensional spaces, the Gâteaux (or Fréchet) derivative $A \in \mathcal{L}(U,V)$ is best described by specifying the action of A on directions, i. e., $A\,\delta u$. In case $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$, one may of course continue to encode A as the Jacobian matrix, which describes A in terms of its action on the standard basis of $\mathbb{R}^n$ and represents the image in terms of the standard basis of $\mathbb{R}^m$.

(*iv*) If we have the continuous embeddings $U_1 \hookrightarrow U$ and $V \hookrightarrow V_1$, and if F is directionally or Gâteaux or Fréchet differentiable at $u \in U$, then the same holds for F as a map $U_1 \to V_1$. △

Example 8.6 (Derivatives)**.**
Suppose that U, V are normed linear spaces.

(*i*) Every bounded linear operator $A \in \mathcal{L}(U,V)$ is everyhwere Fréchet differentiable with derivative $A'(u) = A$ and zero remainder, due to

$$A\,(u + \delta u) = A\,u + A\,\delta u + 0.$$

(*ii*) Similarly, every affine operator $A(u) = A_0 u + b$ with $A_0 \in \mathcal{L}(U,V)$ and $b \in V$ is everywhere Fréchet differentiable with derivative $A'(u) = A_0$ and zero remainder.

(*iii*) In a Hilbert space H, $f(u) = \|u\|_H^2 = (u,u)_H$ is everywhere Fréchet differentiable. The derivative at u is given by

$$f'(u)\,\delta u = 2\,(u, \delta u)_H.$$

(*iv*) Suppose that U and H are Hilbert spaces, $S \in \mathcal{L}(U,H)$ and $z \in H$. Then

$$f(u) = \|S\,u - z\|_H^2$$

is everywhere Fréchet differentiable. The derivative at u is given by

$$f'(u)\,\delta u = 2\,\big(S\,u - z, S\,\delta u\big)_H.$$ △

Let us discuss a more complex example.

Example 8.7 (Derivative of a nonlinear point evaluation)**.**
Suppose that $U = C([0,1])$ is endowed with the maximum norm

$$\|u\|_{C([0,1])} = \max\big\{|u(x)| \,\big|\, x \in [0,1]\big\}$$

and $f(u) = \sin(u(1)) \in V = \mathbb{R}$. We begin by investigating the directional differentiability at an arbitrary $u \in U$ in any direction $\delta u \in U$:

$$\begin{aligned}
\lim_{t \searrow 0} \frac{f(u + t\,\delta u) - f(u)}{t} &= \lim_{t \searrow 0} \frac{\sin\big((u + t\,\delta u)(1)\big) - \sin\big(u(1)\big)}{t} \\
&= \lim_{t \searrow 0} \frac{\sin\big((u(1) + t\,\delta u(1))\big) - \sin\big(u(1)\big)}{t} \\
&= \frac{\mathrm{d}}{\mathrm{d}t} \sin\big(u(1) + t\,\delta u(1)\big)|_{t=0} \\
&= \cos\big(u(1)\big)\,\delta u(1).
\end{aligned}$$

This shows that f is directionally differentiable everywhere and in any direction with

$$f'(u;\delta u) = \cos\big(u(1)\big)\,\delta u(1).$$

The directional derivative is a linear function of the direction that is also continuous since

$$\big|\cos\big(u(1)\big)\,\delta u(1)\big| = \big|\cos\big(u(1)\big)\big|\,|\delta u(1)| \leqslant \big|\cos\big(u(1)\big)\big|\,\|\delta u\|_{C([0,1])}.$$

This shows that f is everywhere Gâteaux differentiable with Gâteaux derivative

$$f_G'(u)\,\delta u = \cos\big(u(1)\big)\,\delta u(1).$$

Note: As stated in Remark 8.5, the derivative cannot be written without reference to the direction δu.

We finally show that f is indeed everywhere Fréchet differentiable. To this end, we use Taylor's theorem for the C^2 function sin, which shows that

$$\sin(x+h) = \sin(x) + \cos(x)\,h - \frac{1}{2}\sin(\xi)\,h^2$$

for any $x, h \in \mathbb{R}$ and some $\xi \in \mathbb{R}$ between x and $x+h$. Using this with $x = u(1)$ and $h = \delta u(1)$ shows

$$\begin{aligned}\frac{\big|\sin\big(u(1)+\delta u(1)\big) - \sin\big(u(1)\big) - \cos\big(u(1)\big)\,\delta u(1)\big|}{\|\delta u\|_{C([0,1])}} \\ = \frac{1}{2}\frac{\big|\sin(\xi)\,(\delta u(1))^2\big|}{\|\delta u\|_{C([0,1])}}\end{aligned}$$

with some ξ between $u(1)$ and $u(1)+\delta u(1)$. Due to the boundedness of the sin function on $\mathbb{R}$, this is

$$\begin{aligned}&\leqslant \frac{1}{2}\frac{|\delta u(1)|^2}{\|\delta u\|_{C([0,1])}} \\ &\leqslant \frac{1}{2}\frac{\|\delta u\|_{C([0,1])}^2}{\|\delta u\|_{C([0,1])}} \\ &= \frac{1}{2}\|\delta u\|_{C([0,1])},\end{aligned}$$

which converges to 0 as $\delta u \to 0$ in $C([0,1])$. This confirms that f is everywhere Fréchet differentiable. △

§ 9 Particularities in Hilbert Spaces

§ 9.1 Derivatives and Gradients

When considering a differentiable functional $f\colon U \to \mathbb{R}$ on a normed linear space, its derivative $f'(u)$ at $u \in U$ belongs to $\mathcal{L}(U,\mathbb{R}) = U^*$. In a Hilbert space H, we may represent $f'(u) \in H^*$ using the Riesz map (Theorem 4.14).

Definition 9.1 (Gradient)**.**
Suppose that H is a Hilbert space and $f\colon H \to \mathbb{R}$ a differentiable functional. Let $f'(u)$ be the Fréchet derivative of f at $u \in H$. Then the Riesz representer $\Phi^{-1}(f'(u)) \in H$ of $f'(u) \in H^*$ is said to be the **gradient** of f at u, denoted by $\nabla f(u) \in H$. △

The derivative $f'(u) \in H^*$ and the gradient $\nabla f(u) \in H$ are related by

$$f'(u)\,\delta u = (\nabla f(u), \delta u)_H. \tag{9.1}$$

Note: The derivative does not depend on the inner product, but the gradient does. More precisely, when we replace the inner product on a Hilbert space by an equivalent one[6], the gradient will change in general while the derivative does not. It is easy to see that the negative gradient is the unique (up to positive scaling) direction of steepest descent, i. e., it solves the problem

$$\text{Minimize} \quad \frac{f'(u)\,\delta u}{\|\delta u\|_H} \quad \text{where } \delta u \in H \setminus \{0\}.$$

Example 9.2 (derivatives and gradients in finite-dimensional Hilbert spaces)**.**
Consider the Hilbert space $H = \mathbb{R}^n$. We represent the dual space H^* with $\mathbb{R}^n$ as well, where the duality pairing is given by

$$\langle f, u\rangle_{H^*,H} = f^{\mathsf{T}} u.$$

(1) When we endow H with the Euclidean inner product $(u, v) = u^{\mathsf{T}} v$, then the Riesz isometric isomorphism and its inverse are given by

$$\Phi\colon H = \mathbb{R}^n \ni u \mapsto \Phi(u) := u \in H^* = \mathbb{R}^n$$
$$\Phi^{-1}\colon H^* = \mathbb{R}^n \ni f \mapsto \Phi^{-1}(f) := f \in H = \mathbb{R}^n$$

since we have

$$\langle f, u\rangle_{H^*,H} = f^{\mathsf{T}} u = (f, u)_{\mathrm{id}} \quad \text{for all } u \in H.$$

(2) When instead we use the inner product represented by the symmetric and positive definite matrix M on H, then the Riesz isomorphism becomes

$$\Phi\colon H = \mathbb{R}^n \ni u \mapsto \Phi(u) := M\,u \in H^* = \mathbb{R}^n$$
$$\Phi^{-1}\colon H^* = \mathbb{R}^n \ni f \mapsto \Phi^{-1}(f) := M^{-1} f \in H = \mathbb{R}^n$$

since we have

$$\langle f, u\rangle_{H^*,H} = f^{\mathsf{T}} u = f^{\mathsf{T}} M^{-1} M\,u = (M^{-1} f)^{\mathsf{T}} M\,u = (M^{-1} f, u)_M \quad \text{for all } u \in H.$$

△

Example 9.3 (gradient of squared norm)**.**
We know from Example 8.6 that in a Hilbert space H, $f(u) = \|u\|_H^2$ is everywhere Fréchet differentiable with derivative

$$f'(u)\,\delta u = 2\,(u, \delta u)_H.$$

Therefore, the gradient is given by $\nabla f(u) = 2\,u \in H$. △

End of Class 18

End of Week 11

[6] i. e., the induced norms are equivalent

§ 9.2 Dual and Adjoint Operators

Definition 9.4 (dual operator)**.**
Suppose that U and V are normed linear spaces and $A \in \mathcal{L}(U, V)$ is a bounded linear operator. The **dual operator** $A^* \in \mathcal{L}(V^*, U^*)$ is defined by the relation $A^* g = g \circ A$ for any $g \in V^*$. That is, we have

$$\langle A^* g, u \rangle_{U^*,U} = \langle g, A u \rangle_{V^*,V} \quad \text{for all } u \in U. \tag{9.2}$$

△

Lemma 9.5 (norm of the dual operator)**.**
Suppose that U and V are normed linear spaces, $A \in \mathcal{L}(U, V)$ is a bounded linear operator and $A^* \in \mathcal{L}(V^*, U^*)$ is its dual operator. Then

$$\|A^*\|_{\mathcal{L}(V^*,U^*)} = \|A\|_{\mathcal{L}(U,V)}$$

holds.

While $\|A^*\|_{\mathcal{L}(V^*,U^*)} \leqslant \|A\|_{\mathcal{L}(U,V)}$ is easy to see, the proof of the reverse inequality uses the Hahn-Banach theorem.

Example 9.6 (dual operator in finite-dimensional spaces)**.**
Suppose that $A \in \mathbb{R}^{m\times n}$ is a matrix, which we identify with the linear map $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ via $x \mapsto A x$. Then $A^* = A^\mathsf{T}$ holds, since we have

$$\langle g, A x \rangle_{(\mathbb{R}^m)^*,\mathbb{R}^m} = g^\mathsf{T} A x = (A^\mathsf{T} g)^\mathsf{T} x = \langle A^\mathsf{T} g, x \rangle_{(\mathbb{R}^n)^*,\mathbb{R}^n}.$$

That is, A^T satisfies the relation (9.2) that characterizes the dual map A^* of A. △

When H_1 and H_2 are Hilbert spaces, we may compose the dual operator $A^* \in \mathcal{L}(H_2^*, H_1^*)$ of $A \in \mathcal{L}(H_1, H_2)$ with the Riesz isomorphism of H_1 and the inverse Riesz isomorphism of H_2 so that the duality pairings in (9.2) are replaced by inner products:

Definition 9.7 (adjoint operator)**.**
Suppose that H_1 and H_2 are Hilbert spaces and $A \in \mathcal{L}(H_1, H_2)$ is a bounded linear operator. The **adjoint operator** $A^\circ \in \mathcal{L}(H_2, H_1)$ is defined by the relation $A^\circ = (\Phi_{H_1^* \leftarrow H_1})^{-1} \circ A^* \circ \Phi_{H_2^* \leftarrow H_2}$. That is, we have

$$(A^\circ v, u)_{H_1} = (v, A u)_{H_2} \quad \text{for all } u \in H_1, v \in H_2. \tag{9.3}$$

△

Corollary 9.8 (norm of the adjoint operator)**.**
Suppose that H_1 and H_2 are Hilbert spaces, $A \in \mathcal{L}(H_1, H_2)$ is a bounded linear operator and $A^\circ \in \mathcal{L}(H_2, H_1)$ is its adjoint operator. Then

$$\|A^\circ\|_{\mathcal{L}(H_2,H_1)} = \|A\|_{\mathcal{L}(H_1,H_2)}$$

holds.

Example 9.9 (adjoint operators)**.**

(*i*) Suppose that $A \in \mathbb{R}^{m\times n}$ is a matrix, which we identify with the linear map $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ via $x \mapsto A\,x$. When we endow $\mathbb{R}^n$ and $\mathbb{R}^m$ with the Euclidean inner products, then $A^\circ = A^* = A^\mathsf{T}$ holds.

(*ii*) When instead we use the inner products represented by the symmetric and positive definite matrices M on $\mathbb{R}^m$ and N on $\mathbb{R}^n$, then we obtain $A^\circ = N^{-1}A^\mathsf{T}M$.

(*iii*) Suppose that U and H are Hilbert spaces, $S \in \mathcal{L}(U, H)$ and $z \in H$. Consider

$$f(u) = \|S\,u - z\|_H^2$$

with Fréchet derivative

$$f'(u)\,\delta u = 2\,\big(S\,u - z, S\,\delta u\big)_H,$$

compare Example 8.6. Using the adjoint $S^\circ \in \mathcal{L}(H, U)$, we can reformulate this as

$$f'(u)\,\delta u = 2\,\big(S^\circ(S\,u - z), \delta u\big)_U.$$

This shows

$$\nabla f(u) = 2\,S^\circ(S\,u - z). \qquad \triangle$$

§ 10 Optimality Conditions for the Floor-Heating Problem

In this section we will derive necessary and sufficient optimality conditions for the floor heating problem (7.3). In fact, we are going to derive these conditions for the reduced problem

$$\begin{aligned} \text{Minimize} \quad & f(u) := \frac{1}{2}\|E\,G(u) - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2, \quad \text{where } u \in U \\ \text{s.t.} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{7.10}$$

The reason to consider the reduced problem is that the PDE constraint has been eliminated using the control-to-state map, so the discussion of constraint qualifications in infinite-dimensional spaces is avoided. Notice that we are going to treat the remaining constraint $u \in U_{\text{ad}}$ as an abstract constraint, without assigning Lagrange multipliers.

Theorem 10.1 (necessary and sufficient optimality conditions)**.**
Suppose that U is a normed linear space, $U_{\text{ad}} \subseteq U$ is convex and $f\colon U \to \mathbb{R}$ is everywhere directionally differentiable. We consider the abstract problem

$$\begin{aligned} \text{Minimize} \quad & f(u), \quad \text{where } u \in U \\ \text{s.t.} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{10.1}$$

(*i*) Suppose that $u^* \in U_{\text{ad}}$ is a locally optimal solution of (10.1). Then the variational inequality

$$f'(u^*; u - u^*) \geqslant 0 \quad \text{for all } u \in U_{\text{ad}} \tag{10.2}$$

holds.

(*ii*) Suppose that f is convex on U_{ad}. Then every $u^* \in U_{\text{ad}}$ satisfying the variational inequality (10.2) is a global minimizer.

Proof. Statement (*i*): Suppose that $u^* \in U_{\text{ad}}$ is a locally optimal solution of (10.1) and $u \in U_{\text{ad}}$ is arbitrary. Then any convex combination $\alpha\, u + (1-\alpha)\, u^*$ belongs to U_{ad} for any $\alpha \in [0,1]$, and for sufficiently small $\alpha > 0$, local optimality implies

$$0 \leqslant \frac{f(\alpha\, u + (1-\alpha)\, u^*) - f(u^*)}{\alpha}.$$

Passing to the limit $\alpha \searrow 0$ now shows

$$0 \leqslant f'(u^*; u - u^*).$$

Statement (*ii*): Suppose now that f is convex on U_{ad} and $u \in U_{\text{ad}}$ is arbitrary. Then for any $\alpha \in [0,1]$, we have

$$f(\alpha\, u + (1-\alpha)\, u^*) \leqslant \alpha\, f(u) + (1-\alpha)\, f(u^*)$$

and thus

$$\frac{f(u^* + \alpha\,(u - u^*)) - f(u^*)}{\alpha} \leqslant f(u) - f(u^*).$$

By passing to the limit $\alpha \searrow 0$ we obtain

$$f'(u^*; u - u^*) \leqslant f(u) - f(u^*).$$

By assumption, the left-hand side is $\geqslant 0$ due to the variational inequality (10.2). This shows $0 \leqslant f(u) - f(u^*)$. Since $u \in U_{\text{ad}}$ was arbitrary, u^* is globally optimal. □

We can now apply this result to the reduced problem (7.10).

Corollary 10.2 (necessary and sufficient optimality conditions for (7.10))**.**
Suppose that U, Y, H are Hilbert spaces, the observation operator $E \in \mathcal{L}(Y, H)$ is linear and bounded, and the control-to-state operator $G\colon U \to Y$ is affine and continuous. Denote by G_0 the linear part of G and set $S := E \circ G$ and $S_0 := E \circ G_0$. Moreover, let $z \in H$, $\gamma \geqslant 0$ and $U_{\text{ad}} \subseteq U$ be convex. Then $u^* \in U_{\text{ad}}$ is a global minimizer for (7.10) if and only if the variational inequality

$$\big(\nabla f(u^*), u - u^*\big)_U = \big(\underbrace{S_0^\circ}_{\text{adjoint of } S_0}(S(u^*) - z) + \gamma\, u^*, u - u^*\big)_U \geqslant 0 \quad \text{for all } u \in U_{\text{ad}} \tag{10.3}$$

holds.

End of Class 19

Proof. The objective function

$$f(u) = \frac{1}{2}\|S(u) - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2$$

is Fréchet differentiable by Example 8.6 with derivative

$$\begin{aligned} f'(u)\,\delta u &= \big(S(u) - z, S'(u)\,\delta u\big)_H + \gamma\,(u, \delta u)_U \\ &= \big(S(u) - z, S_0\,\delta u\big)_H + \gamma\,(u, \delta u)_U \\ &= \big(S_0^\circ(S(u) - z), \delta u\big)_U + \gamma\,(u, \delta u)_U \end{aligned}$$

and gradient

$$\nabla f(u) = S_0^\circ(S(u) - z) + \gamma\, u.$$

The objective is also convex. (**Quiz 10.1:** Is this clear?) The result thus follows from Theorem 10.1. □

We recall that the floor-heating problem (7.3) is expressed by the choices $U = L^2(\Omega_{\text{ctrl}})$, $Y = H^1(\Omega)$ and $H = L^2(\Omega_{\text{obs}})$, as well as the control-to-state map $y = G(u)$ given by the unique solution of the weak PDE formulation (7.6):

$$\begin{aligned} &\text{Find} \quad y \in H^1(\Omega) \\ &\text{s.t.} \quad \int_\Omega \kappa \nabla y \cdot \nabla v \, \mathrm{d}x + \int_\Gamma \alpha \, y \, v \, \mathrm{d}s = \int_{\Omega_{\text{ctrl}}} u \, v \, \mathrm{d}x + \int_\Gamma \alpha \, y_\infty \, v \, \mathrm{d}s \quad \text{for all } v \in H^1(\Omega). \end{aligned}$$

The linear part G_0 of G is obtained by replacing y_∞ by 0. The observation map $E\colon H^1(\Omega) \to L^2(\Omega_{\text{obs}})$ is a combined embedding and restriction, $E\, y := y|_{\Omega_{\text{obs}}}$.

We need to clarify what $S_0^\circ = (E \circ G_0)^\circ = G_0^\circ \circ E^\circ$ means. We recall that this adjoint map is defined through

$$(h, S_0\, u)_{L^2(\Omega_{\text{obs}})} = (S_0^\circ\, h, u)_{L^2(\Omega_{\text{ctrl}})} \quad \text{for all } u \in L^2(\Omega_{\text{ctrl}}) \text{ and } h \in L^2(\Omega_{\text{obs}}). \tag{10.4}$$

Lemma 10.3 (representation of S_0°).
The adjoint $S_0^\circ\colon L^2(\Omega_{\text{obs}}) \to L^2(\Omega_{\text{ctrl}})$ of the linear part of the control-to-observation map $S_0\colon L^2(\Omega_{\text{ctrl}}) \to L^2(\Omega_{\text{obs}})$ is given by

$$S_0^\circ\, h = p|_{\Omega_{\text{ctrl}}},$$

where $p \in H^1(\Omega)$ is the unique weak solution of

$$\begin{aligned} -\operatorname{div}(\kappa \nabla p) &= \chi_{\text{obs}}\, h \quad \text{in } \Omega \\ \kappa \frac{\partial}{\partial n} p + \alpha\, p &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{10.5}$$

Proof. Suppose that $u \in L^2(\Omega_{\text{ctrl}})$ and $h \in L^2(\Omega_{\text{obs}})$ are arbitrary. We know that $y = S_0\, u$ is given by the unique solution $y \in H^1(\Omega)$ of

$$\int_\Omega \kappa \nabla y \cdot \nabla v \, \mathrm{d}x + \int_\Gamma \alpha \, y \, v \, \mathrm{d}s = \int_{\Omega_{\text{ctrl}}} u \, v \, \mathrm{d}x \quad \text{for all } v \in H^1(\Omega)$$

and its restriction to Ω_{obs}. On the other hand, the unique weak solution $p \in H^1(\Omega)$ of (10.5) satisfies

$$\int_\Omega \kappa \nabla p \cdot \nabla v \, \mathrm{d}x + \int_\Gamma \alpha \, p \, v \, \mathrm{d}s = \int_{\Omega_{\text{obs}}} h \, v \, \mathrm{d}x \quad \text{for all } v \in H^1(\Omega).$$

When we plug in $v = p$ as test function into the first equation and $v = y$ into the second, we see that the left-hand sides agree, and so the right-hand sides must agree as well. This shows

$$\int_{\Omega_{\text{ctrl}}} u \, p \, \mathrm{d}x = \int_{\Omega_{\text{obs}}} h \, y \, \mathrm{d}x,$$

which is the defining equation that shows $S_0^\circ\, h = p|_{\Omega_{\text{ctrl}}}$. □

These findings suggest the following procedure to evaluate the gradient $\nabla f(u) = S_0^\circ(S(u) - z) + \gamma\, u$:

Algorithm 10.4 (evaluation of $\nabla f(u)$ for the floor-heating problem).

1: Evalute $y := S(u)$ by solving the weak form of the state equation

$$\begin{aligned} -\operatorname{div}(\kappa \nabla y) &= \chi_{\text{ctrl}}\, u && \text{in } \Omega \\ \kappa \frac{\partial}{\partial n} y &= \alpha\,(y_\infty - y) && \text{on } \Gamma. \end{aligned}$$

2: Evaluate $p := -S_0^\circ(S(u) - z) = -S_0^\circ(y - z)$ by solving the weak form of the **adjoint equation**[7]

$$\begin{aligned} -\operatorname{div}(\kappa \nabla p) &= -\chi_{\text{obs}}\,(y - z) && \text{in } \Omega \\ \kappa \frac{\partial}{\partial n} p + \alpha\, p &= 0 && \text{on } \Gamma. \end{aligned}$$

3: Assemble the gradient

$$\nabla f(u) = -p|_{\Omega_{\text{ctrl}}} + \gamma\, u.$$

We refer to y as the **state associated with the control** u and to p as the **adjoint state associated with** u.

We can thus formulate the necessary and sufficient optimality conditions in Corollary 10.2 as

$$\int_{\Omega_{\text{ctrl}}} (\gamma\, u^* - p)\,(u - u^*)\, \mathrm{d}x \geqslant 0 \quad \text{for all } u \in U_{\text{ad}}. \tag{10.6}$$

Remark 10.5 (interpretation of the adjoint state)**.**

(*i*) With (10.5) at hand, we see that the left-hand side of (10.4), i. e., solving a PDE, then multiply the solution with h and integrate, is replaced by solving the adjoint PDE with data h, then multiply the solution with h and integrate.

(*ii*) Finding the form of the adjoint PDE (10.5) was little intuitive. In what follows, we will derive the adjoint PDE by other, yet informal, means. We will also see that the adjoint state p can be interpreted as the Lagrange multiplier pertaining to the state equation (PDE constraint). △

§ 10.1 Informal Derivation of the Adjoint PDE

In this subsection we proceed in an informal way. We define the Lagrange function by adjoining the PDE constraint (the state equation) in weak form to the objective and derive optimality conditions of KKT type. The attribute "informal" comes from the fact that we do not discuss or verify constraint qualifications. So the arguments in this subsection do not constitute a rigorous derivation of optimality conditions.

We define the **Lagrange function** (or **Lagrangian**) for the floor-heating problem (7.3) as

$$\begin{aligned} \mathcal{L}(y, u, p) := {} & \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_{\text{obs}})} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega_{\text{ctrl}})} \\ & + \int_\Omega \kappa \nabla y \cdot \nabla p \, \mathrm{d}x + \int_\Gamma \alpha\, y\, p \, \mathrm{d}s - \int_{\Omega_{\text{ctrl}}} u\, p \, \mathrm{d}x - \int_\Gamma \alpha\, y_\infty\, p \, \mathrm{d}s. \end{aligned} \tag{10.7}$$

[7] It will become clear in § 10.1 why we introduce p with an additional minus sign and then use $-p$ in the next step.

We do not add terms for the control constraints $u \in U_{\text{ad}}$ but keep them as abstract constraints. We expect the KKT conditions to be of the following form:

$$\begin{aligned} \mathcal{L}_y(y,u,p)\,\delta y &= 0 \quad \text{for all } \delta y \in H^1(\Omega) \\ \mathcal{L}_u(y,u,p)\,(\widetilde{u}-u) &\geqslant 0 \quad \text{for all } \widetilde{u} \in U_{\text{ad}} \\ \mathcal{L}_p(y,u,p)\,\delta p &= 0 \quad \text{for all } \delta p \in H^1(\Omega). \end{aligned}$$

The first equality amounts to

$$\mathcal{L}_y(y,u,p)\,\delta y = \int_{\Omega_{\text{obs}}} (y-y_d)\,\delta y + \int_\Omega \kappa\,\nabla \delta y \cdot \nabla p \,\mathrm{d}x + \int_\Gamma \alpha\,\delta y\, p \,\mathrm{d}s = 0.$$

Upon inspection, we see that this is precisely the weak form of the adjoint equation

$$\begin{aligned} -\operatorname{div}\bigl(\kappa\,\nabla p\bigr) &= -\chi_{\text{obs}}\,(y-y_d) \quad && \text{in } \Omega \\ \kappa\,\frac{\partial}{\partial n} p + \alpha\, p &= 0 && \text{on } \Gamma. \end{aligned}$$

(Here it becomes clear why we introduced $p := -S_0^\circ(y-z)$ with the minus sign.)

Second, we see

$$\mathcal{L}_u(y,u,p)\,\delta u = \gamma \int_{\Omega_{\text{ctrl}}} u\,\delta u \,\mathrm{d}x - \int_{\Omega_{\text{ctrl}}} \delta u\, p \,\mathrm{d}x$$

holds, so that $\mathcal{L}_u(y,u,p)\,(\widetilde{u}-u) \geqslant 0$ amounts to the variational inequality (10.6)

$$\int_{\Omega_{\text{ctrl}}} (\gamma\, u - p)\,(\widetilde{u}-u) \,\mathrm{d}x \geqslant 0,$$

and thus the Riesz representer of $\mathcal{L}_u(y,u,p)$ is the gradient $\nabla f(u)$.

Finally, it is easy to see that $\mathcal{L}_p(y,u,p)\,\delta p = 0$ for all $\delta p \in H^1(\Omega)$ is the weak formulation of the state equation.

Remark 10.6 (interpretation of the variational inequality as an orthogonal projection).
When $\gamma > 0$, we can interpret the variational inequality (10.6) as an orthogonal projection formula. Dividing by γ, we obtain

$$\int_{\Omega_{\text{ctrl}}} (u - \gamma^{-1} p)\,(\widetilde{u}-u) \,\mathrm{d}x \geqslant 0 \quad \text{for all } \widetilde{u} \in U_{\text{ad}},$$

and this variational inequality is well-known to present the necessary and sufficient optimality conditions of the $L^2(\Omega_{\text{ctrl}})$-orthogonal projection problem of $\gamma^{-1} p|_{\Omega_{\text{ctrl}}}$ onto U_{ad}:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|u - \gamma^{-1} p\|^2_{L^2(\Omega_{\text{ctrl}})}, \quad \text{where } u \in L^2(\Omega_{\text{ctrl}}) \\ \text{s.t.} \quad & u \in U_{\text{ad}}. \end{aligned}$$

We can write this as

$$u = \operatorname{proj}_{U_{\text{ad}}}\bigl(\gamma^{-1} p|_{\Omega_{\text{ctrl}}}\bigr) \quad \text{w.r.t. the } L^2(\Omega_{\text{ctrl}})\text{-inner product.}$$

When U_{ad} is given by bound constraints, i. e.,

$$U_{\text{ad}} = \{u \in L^2(\Omega_{\text{ctrl}}) \mid u_a \leqslant u \leqslant u_b \text{ a.e. in } \Omega_{\text{ctrl}}\},$$

then this orthogonal projection can be carried out pointwise:

$$u(x) = \operatorname{proj}_{[u_a,u_b]}\big(\gamma^{-1}p(x)\big) \quad \text{w.r.t. the } \mathbb{R}\text{-inner product,}$$

or simply

$$u = \min\big\{\max\big\{\gamma^{-1}p, u_a\big\}, u_b\big\}, \tag{10.8}$$

which is to be understood pointwise a.e. in Ω_{ctrl}. This continues to hold when $u_a = -\infty$ and/or $u_b = \infty$. △

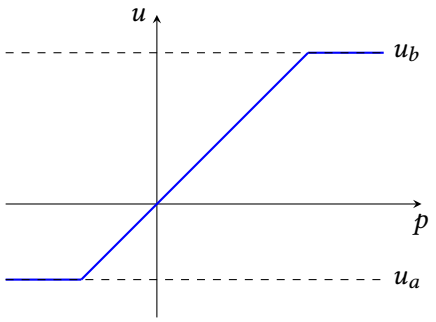


Figure 10.1: Illustration of the relation (10.8).

End of Class 20

End of Week 12

§ 10.2 Introduction of Inequality Constraint Multipliers

In this section we will present yet another equivalent variant of the optimality system. We assume pointwise bound constraints, i. e.,

$$U_{\text{ad}} = \{u \in L^2(\Omega_{\text{ctrl}}) \mid u_a \leqslant u \leqslant u_b \text{ a.e. in } \Omega_{\text{ctrl}}\}.$$

We now also adjoin these constraints (in addition to the state equality constraint) to the objective and we obtain the **extended Lagrange function** (or **extended Lagrangian**) for the floor-heating problem (7.3):

$$\begin{aligned}\mathcal{L}(y,u,p,\mu_a,\mu_b) \coloneqq{}& \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_{\text{obs}})} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega_{\text{ctrl}})} \\ &+ \int_\Omega \kappa\,\nabla y \cdot \nabla p \,\mathrm{d}x + \int_\Gamma \alpha\, y\, p \,\mathrm{d}s - \int_{\Omega_{\text{ctrl}}} u\, p \,\mathrm{d}x - \int_\Gamma \alpha\, y_\infty\, p \,\mathrm{d}s \\ &+ \int_{\Omega_{\text{ctrl}}} \mu_a\,(u_a - u) \,\mathrm{d}x + \int_{\Omega_{\text{ctrl}}} \mu_b\,(u - u_b) \,\mathrm{d}x.\end{aligned} \tag{10.9}$$

Differentiation w.r.t. the control now gives

$$\begin{aligned}\mathcal{L}_u(y,u,p,\mu_a,\mu_b)\,\delta u &= \gamma\int_{\Omega_{\text{ctrl}}} u\,\delta u\;\mathrm{d}x - \int_{\Omega_{\text{ctrl}}}\delta u\,p\;\mathrm{d}x - \int_{\Omega_{\text{ctrl}}}\mu_a\,\delta u\;\mathrm{d}x + \int_{\Omega_{\text{ctrl}}}\mu_b\,\delta u\;\mathrm{d}x\\ &= \int_{\Omega_{\text{ctrl}}}(\gamma\,u - p + \mu_b - \mu_a)\,\delta u\;\mathrm{d}x\end{aligned}$$

and the optimality system contains $\mathcal{L}_u(y,u,p,\mu_a,\mu_b)\,\delta u = 0$ for all $\delta u \in L^2(\Omega_{\text{ctrl}})$, i. e.,

$$\gamma\,u - p + \mu_b - \mu_a = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}.$$

Since Lagrange multipliers for inequality constraints are non-negative, we add $u_a \geqslant 0$ and $u_b \geqslant 0$ a.e. in Ω_{ctrl} to the optimality system. In other words, we obtain

$$\begin{aligned}\mu_a &= [\gamma\,u - p]^+ \quad \text{a.e. in } \Omega_{\text{ctrl}}\\ \mu_b &= [\gamma\,u - p]^- \quad \text{a.e. in } \Omega_{\text{ctrl}},\end{aligned}$$

where $[\cdot]^+$ and $[\cdot]^-$ denote the pointwise positive and negative parts of a function, respectively. Notice that μ_a, μ_b belong to $L^2(\Omega_{\text{ctrl}})$. (**Quiz 10.2:** Is this clear?) To complete the optimality conditions of KKT type, we require the complementarity conditions

$$\begin{aligned}\int_{\Omega_{\text{ctrl}}}\mu_a\,(u_a - u)\;\mathrm{d}x &= 0,\\ \int_{\Omega_{\text{ctrl}}}\mu_b\,(u - u_b)\;\mathrm{d}x &= 0.\end{aligned}$$

Although the derivation of the KKT conditions was again obtained on an informal basis (avoiding the discussion of constraint qualifications), we indeed obtain yet another equivalent form of the optimality conditions.

Theorem 10.7 (various equivalent forms of the necessary and sufficient optimality conditions)**.** Suppose that Assumption 7.6 and Assumption 7.9 hold. Moreover, suppose that $u \in U_{\text{ad}}$ holds, with associated state $y \in H^1(\Omega)$ satisfying the weak form of the state equation

$$\begin{aligned}-\operatorname{div}(\kappa\,\nabla y) &= \chi_{\text{ctrl}}\,u && \text{in } \Omega\\ \kappa\,\frac{\partial}{\partial n}y &= \alpha\,(y_\infty - y) && \text{on } \Gamma\end{aligned}$$

and associated adjoint state $p \in H^1(\Omega)$ satisfying the weak form of the adjoint equation

$$\begin{aligned}-\operatorname{div}(\kappa\,\nabla p) &= -\chi_{\text{obs}}\,(y - y_d) && \text{in } \Omega\\ \kappa\,\frac{\partial}{\partial n}p + \alpha\,p &= 0 && \text{on } \Gamma.\end{aligned}$$

Then the following are equivalent:

(i) u is a global minimizer of (7.3).

(ii) $u \in U_{\text{ad}}$ satisfies the variational inequality

$$\int_{\Omega_{\text{ctrl}}}(\gamma\,u - p)\,(\widetilde{u} - u)\;\mathrm{d}x \geqslant 0 \quad \text{for all } \widetilde{u} \in U_{\text{ad}}.$$

(*iii*) There exist inequality constraint multipliers $\mu_a, \mu_b \in L^2(\Omega_{\text{ctrl}})$ such that

$$\gamma\, u - p + \mu_b - \mu_a = 0$$

and the complementarity system

$$\mu_a \geqslant 0, \quad u_a - u \leqslant 0, \quad \int_{\Omega_{\text{ctrl}}} \mu_a\,(u_a - u)\,\mathrm{d}x = 0 \tag{10.10a}$$

$$\mu_b \geqslant 0, \quad u - u_b \leqslant 0, \quad \int_{\Omega_{\text{ctrl}}} \mu_b\,(u - u_b)\,\mathrm{d}x = 0 \tag{10.10b}$$

holds, with the inequalities understood almost everywhere.

Proof. Statement (*i*) $\Leftrightarrow$ statement (*ii*): We proved in Corollary 10.2 that statement (*i*) is equivalent to the variational inequality $\big(\nabla f(u), \widetilde{u} - u\big)_U \geqslant 0$ holds for all $\widetilde{u} \in U_{\text{ad}}$. Since $\nabla f(u) = \gamma\, u - p$ holds by (10.6), the equivalence is proved.

Statement (*ii*) $\Rightarrow$ statement (*iii*): We define

$$\mu_a = [\gamma\, u - p]^+ \quad \text{and} \quad \mu_b = [\gamma\, u - p]^- \quad \text{in } \Omega_{\text{ctrl}}.$$

Then $\mu_a \geqslant 0$ and $\mu_b \geqslant 0$ holds a.e. in Ω_{ctrl}. Since $u \in U_{\text{ad}}$ by assumption, we also have $u_a - u \leqslant 0$ and $u - u_b \leqslant 0$ a.e. in Ω_{ctrl}. It remains to verify the complementarity condition.

To make the following arguments rigorous, we work with arbitrary but fixed representers (in the equivalence classes of a.e. equal functions) of the quantities u and p and define μ_a and μ_b based on these representers. We define the set

$$\mathcal{A}^- := \{x \in \Omega_{\text{ctrl}} \,|\, \mu_a(x) > 0\} = \{x \in \Omega_{\text{ctrl}} \,|\, \gamma\, u(x) - p(x) > 0\}.$$

Therefore, $u > \gamma^{-1} p$ holds on $\mathcal{A}^-$. On the other hand, we have

$$u = \min\big\{\max\big\{\gamma^{-1} p, u_a\big\}, u_b\big\}$$

on Ω_{ctrl} by (10.8); compare Figure 10.1. This is only possible if $u = u_a$ holds on $\mathcal{A}^-$. On the complement of $\mathcal{A}^-$, we have $\mu_a = 0$.

Therefore, we obtain $\mu_a\,(u_a - u) = 0$ everywhere on Ω_{ctrl}, and thus

$$\int_{\Omega_{\text{ctrl}}} \mu_a\,(u_a - u)\,\mathrm{d}x = 0$$

as claimed. The proof of

$$\int_{\Omega_{\text{ctrl}}} \mu_b\,(u - u_b)\,\mathrm{d}x = 0$$

proceeds in the same way.

Statement (*iii*) $\Rightarrow$ statement (*ii*): First of all, $u \in U_{\text{ad}}$ holds by assumption. As in the previous step, we fix representers and distinguish two subsets:

$$\mathcal{A}^- := \{x \in \Omega_{\text{ctrl}} \,|\, \mu_a(x) > 0\},$$
$$\mathcal{A}^+ := \{x \in \Omega_{\text{ctrl}} \,|\, \mu_b(x) > 0\}.$$

On $\mathcal{A}^-$ we must have $u = u_a$. This follows from the following Lemma 10.8. Therefore, for any $\widetilde{u} \in U_{\text{ad}}$, we have

$$\begin{aligned}\int_{\Omega_{\text{ctrl}}} \mu_a\,(\widetilde{u}-u)\;\mathrm{d}x &= \int_{\mathcal{A}^-} \mu_a\,(\widetilde{u}-u)\;\mathrm{d}x + \int_{\Omega_{\text{ctrl}}\setminus\mathcal{A}^-} \mu_a\,(\widetilde{u}-u)\;\mathrm{d}x\\ &= \int_{\mathcal{A}^-} \underbrace{\mu_a}_{>0}\,(\underbrace{\widetilde{u}-u_a}_{\geqslant 0})\;\mathrm{d}x + \int_{\Omega_{\text{ctrl}}\setminus\mathcal{A}^-} \underbrace{\mu_a}_{=0}\,(\underbrace{\widetilde{u}-u}_{\in\mathbb{R}})\;\mathrm{d}x\\ &\geqslant 0.\end{aligned}$$

Similarly, on $\mathcal{A}^+$ we must have $u = u_b$. Therefore, for any $\widetilde{u} \in U_{\text{ad}}$, we have

$$\begin{aligned}\int_{\Omega_{\text{ctrl}}} \mu_b\,(\widetilde{u}-u)\;\mathrm{d}x &= \int_{\mathcal{A}^+} \mu_b\,(\widetilde{u}-u)\;\mathrm{d}x + \int_{\Omega_{\text{ctrl}}\setminus\mathcal{A}^+} \mu_b\,(\widetilde{u}-u)\;\mathrm{d}x\\ &= \int_{\mathcal{A}^+} \underbrace{\mu_b}_{>0}\,(\underbrace{\widetilde{u}-u_b}_{\leqslant 0})\;\mathrm{d}x + \int_{\Omega_{\text{ctrl}}\setminus\mathcal{A}^+} \underbrace{\mu_b}_{=0}\,(\underbrace{\widetilde{u}-u}_{\in\mathbb{R}})\;\mathrm{d}x\\ &\leqslant 0.\end{aligned}$$

This implies

$$\int_{\Omega_{\text{ctrl}}} (\gamma\,u-p)\,(\widetilde{u}-u)\;\mathrm{d}x = \int_{\Omega_{\text{ctrl}}} (\mu_a-\mu_b)\,(\widetilde{u}-u)\;\mathrm{d}x \geqslant 0$$

for all $\widetilde{u} \in U_{\text{ad}}$, i. e., statement (ii) is satisfied. □

The following lemma was added after class.

Lemma 10.8 (non-vanishing non-negative functions are bounded away from zero on sets of positive measure)**.**
Suppose that $\Omega \subseteq \mathbb{R}^d$ is a measurable set and that $f \in L^1(\Omega)$ is non-negative. The following are equivalent:

(i) $f = 0$ a.e. in Ω.
(ii) $\int_\Omega f\;\mathrm{d}x = 0$.

Proof. Statement (i) $\Rightarrow$ statement (ii): This is clear.

$\neg$ Statement (i) $\Rightarrow$ $\neg$ statement (ii): Suppose that $f > 0$ on a subset of Ω of positive measure.

Step 1: We show that there exists a subset Ω_0 of positive measure and $\delta > 0$ such that $f(x) \geqslant \delta > 0$ holds on Ω_0. Indeed, suppose that this is not the case. Then for every $n \in \mathbb{N}$, the set

$$\Big\{x \in \Omega \,\Big|\, f(x) \geqslant \frac{1}{n}\Big\}$$

has zero measure. Consequently, the union

$$\bigcup_{n\in\mathbb{N}} \Big\{x \in \Omega \,\Big|\, f(x) \geqslant \frac{1}{n}\Big\} = \{x \in \Omega \,|\, f(x) > 0\}$$

has zero measure as well. This contradicts the assumption that $f > 0$ holds on a subset of positive measure.

Step 2: As a consequence, we obtain

$$\int_\Omega f \,\mathrm{d}x \geqslant \int_{\Omega_0} f \,\mathrm{d}x \geqslant \int_{\Omega_0} \delta \,\mathrm{d}x = |\Omega_0|\,\delta > 0,$$

i. e., ¬ statement (*ii*). □

Remark 10.9 (uniqueness of inequality constraint multipliers)**.**

(*i*) As for box constraints for vectors in finite-dimensional vector spaces, the complementarity condition (10.10) in integral form is equivalent to the pointwise complementarity conditions

$$\mu_a \geqslant 0, \quad u_a - u \leqslant 0, \quad \mu_a\,(u_a - u) = 0 \tag{10.11a}$$
$$\mu_b \geqslant 0, \quad u - u_b \leqslant 0, \quad \mu_b\,(u - u_b) = 0 \tag{10.11b}$$

a.e. in Ω_{ctrl}.

(*ii*) In case $u_a < u_b$, the inequality constraint multipliers μ_a and μ_b are unique. This follows from the fact that both inequalities cannot be simultaneously active on a subset of Ω_{ctrl} of positive measure.

(*iii*) However, in case $u_a = u_b$, then the multipliers are not unique. For any pair (μ_a, μ_b) satisfying the complementarity conditions, $(\mu_a + c, \mu_b + c)$ is also a solution for any $c \geqslant 0$. △

For algorithmic purposes, it may be useful to combine μ_a and μ_b into a single (signed) multiplier

$$\mu := \mu_b - \mu_a. \tag{10.12}$$

This combined multiplier $\mu \in L^2(\Omega_{\text{ctrl}})$ is always uniqu, even when $u_a = u_b$. Its positive and negative parts act as the original multipliers for the upper and lower bound, respectively.

End of Class 23

§ 11 Algorithms for Reduced Linear-Quadratic Problems Without Inequality Constraints

In this section we discuss some optimization methods to solve optimal control problems such as the floor heating problem (7.3) in their reduced formulation, with and without inequality constraints. Since these are infinite-dimensional problems in a Hilbert space U, we will formulate the solution methods in Hilbert spaces. We will then briefly discuss possible discretizations for the numerical realization on a computer.

We are considering problems as in (6.3), i. e.,

$$\text{Minimize} \quad f(u) := \frac{1}{2}\|\underbrace{E\,G}_{=S}\,u - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2, \quad \text{where } u \in U$$

with Hilbert spaces U and H and $S \in \mathcal{L}(U, H)$.

The following example shows that we can further simplify the notation for this problem and write the objective as

$$f(u) = \underbrace{\frac{1}{2}(A\,u, u)_U}_{\text{quadratic}} - \underbrace{(b, u)_U}_{\text{linear}} + \underbrace{c}_{\text{constant}} \tag{11.1}$$

Let us show that this form can indeed be achieved. We can write the objective as

$$\begin{aligned} f(u) &= \frac{1}{2}\|S\,u - z\|_H^2 + \frac{\gamma}{2}\|u\|_U^2 \\ &= \frac{1}{2}(S\,u, S\,u)_H - (z, S\,u)_H + \frac{1}{2}(z, z)_H + \frac{\gamma}{2}\|u\|_U^2 \\ &= \frac{1}{2}(S^\circ S\,u, u)_U - (S^\circ z, u)_U + \frac{1}{2}(z, z)_H + \frac{\gamma}{2}\|u\|_U^2 \\ &= \frac{1}{2}\big((S^\circ S + \gamma\,\mathrm{id})\,u, u\big)_U - (S^\circ z, u)_U + \frac{1}{2}(z, z)_H. \end{aligned}$$

Hence we have

$$A = S^\circ S + \gamma\,\mathrm{id} \in \mathcal{L}(U), \quad b = S^\circ z \in U, \quad c = \frac{1}{2}(z, z)_H \in \mathbb{R}. \tag{11.2}$$

Lemma 11.1. The endomorphism A as in (11.2) is self-adjoint and positive semi-definite. When $\gamma > 0$, then A is positive definite.

Proof. □

We will work with (11.1) under the assumption of self-adjointness and positive definiteness of A. The minimization of (11.1) is a convex problem with a unique solution, characterized by

$$\nabla f(u) = A\,u - b = 0. \tag{11.3}$$

The quantity

$$r := A\,u - b \in U$$

is called the **residual** associated with the control u.

Although (11.3) is a linear (operator) equation, it cannot be solved by a direct method (such as Gaussian elimination) due to the infinite dimension of the space U. We therefore consider iterative methods to minimize (11.1), or equivalently, solve (11.3).

We will discuss two such methods. Beforehand, we derive some useful identities.

Lemma 11.2 (useful identities)**.**
For the quadratic objective (11.1), we have the following identities:

$$f(u + \alpha\,d) = f(u) + \alpha\,\big(\underbrace{A\,u - b}_{=\nabla f(u)}, d\big)_U + \frac{\alpha^2}{2}(A\,d, d)_U \tag{11.4a}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(u + \alpha\,d) = \big(\nabla f(u + \alpha\,d), d\big)_U. \tag{11.4b}$$

Proof. □

§ 11.1 Gradient Descent Algorithm

The gradient method, also known as method of steepest descent, is defined by choosing as search direction the negative gradient of the objective function, and by selecting the exact step size (also known as Cauchy step size) in the search direction, i. e., the minimizer of the quadratic objective function along the search direction. This can be written as the following algorithm.

Algorithm 11.3 (gradient descent algorithm for unconstrained quadratic problems)**.**
Input: initial guess $u^{(0)} \in U$
Input: relative tolerance $\varepsilon > 0$
Output: approximate minimizer of (11.1)

Set $r^{(0)} := \nabla f(u^{(0)}) = A\,u^{(0)} - b \in U$
Set $\delta^{(0)} := \|r^{(0)}\|_U^2$
Set $n := 0$
while $\delta^{(n)} > \varepsilon^2 \delta^{(0)}$ **do**
    Set $q^{(n)} := A\,r^{(n)}$
    Set $\alpha^{(n)} := \frac{\delta^{(n)}}{(q^{(n)}, r^{(n)})_U}$ // evaluate the step size
    Set $u^{(n+1)} := u^{(n)} - \alpha^{(n)}\, r^{(n)}$ // update the control
    Set $r^{(n+1)} := r^{(n)} - \alpha^{(n)}\, q^{(n)}$ // update the residual
    Set $\delta^{(n+1)} := \|r^{(n+1)}\|_U^2$
    Set $n := n + 1$
end while
return $u^{(n)}$

The gradient method is a member of the class of descent methods. Its search direction and the n-th iterate $u^{(n)}$ is given by $d^{(n)} = -r^{(n)} = -\nabla f(u^{(n)})$. The step size $\alpha^{(n)}$ minimizes the one-dimensional function $\alpha \mapsto f(u^{(n)} + \alpha\, d^{(n)})$. The descent achieved in the n-th step is given by

$$f(u^{(n+1)}) - f(u^{(n)}) = -\frac{\alpha^{(n)}}{2} \|r^{(n)}\|_U^2.$$

The stopping criterion implemented in Algorithm 11.3 represents a reduction in the norm of the initial residual:

$$\frac{\|r^{(n)}\|_U}{\|r^{(0)}\|_U} \leqslant \varepsilon.$$

The gradient descent method is globally convergent to the unique minimizer of (11.1). Its speed of convergence is naturally measured in the A-norm on the Hilbert space U:

$$\|u\|_A := (A\,u, u)_U^{1/2}. \tag{11.5}$$

Due to the assumptions of boundedness and positive definiteness, the A-norm is equivalent to the original norm in the Hilbert space U. Specifically for $A = S^* S + \gamma\,\mathrm{id}$ as in (11.2), we find

$$\gamma\,\|u\|_U^2 \leqslant (A\,u, u)_U \leqslant \|A\|_{\mathcal{L}(U)}\,\|u\|_U^2,$$

which shows that the eigenvalues of the self-adjoint operator A belong to the interval $[\gamma, \|A\|_{\mathcal{L}(U)}]$. This follows as in finite dimensions from considering the Rayleigh quotient. The estimate implies that the **condition number** κ of $A \in \mathcal{L}(U)$ satisfies

$$\kappa \leqslant \frac{\|A\|_{\mathcal{L}(U)}}{\gamma}.$$

Theorem 11.4 (convergence estimates for the gradient descent method)**.**
Suppose that U is a Hilbert space, $A \in \mathcal{L}(U)$ is self-adjoint and positive definite, $b \in U$ and $c \in \mathbb{R}$. Then the gradient descent method (Algorithm 11.3) with the stopping criterion turned off converges to the unique minimizer of the quadratic objective (11.1), regardless of the initial value $u^{(0)}$. In terms of the condition number κ of A, the following convergence estimates hold:

$$f(u^{(k+1)}) - f(u^*) \leqslant \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \big(f(u^{(k)}) - f(u^*)\big) \tag{11.6a}$$

$$\|u^{(k+1)} - u^*\|_A \leqslant \left(\frac{\kappa - 1}{\kappa + 1}\right) \|u^{(k)} - u^*\|_A \tag{11.6b}$$

and consequently

$$f(u^{(k)}) - f(u^*) \leqslant \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \big(f(u^{(0)}) - f(u^*)\big) \tag{11.6c}$$

$$\|u^{(k)} - u^*\|_A \leqslant \left(\frac{\kappa - 1}{\kappa + 1}\right)^{k} \|u^{(0)} - u^*\|_A. \tag{11.6d}$$

Moreover, the objective values $f(u^{(k)})$ and thus the norm of the error $\|u^{(k)} - u^*\|_A$ are monotonically decreasing.

The proof is exactly the same as in finite dimensions; see, e. g., Herzog, 2023, Theorem 4.8.

Remark 11.5 (comparison to finite-dimensional gradient methods)**.**

(*i*) The gradient descent algorithm (Algorithm 11.3) and its convergence properties are exactly the same as in finite dimensions.

(*ii*) The condition number κ which governs the convergence estimate depends on the norm (inner product) of the Hilbert space U. In finite dimensions, one is free to choose this norm since all norms are equivalent. In infinite dimensions, one usually starts with a natural norm, such as the L^2-norm. However, one may replace this norm by an equivalent one if necessary in order to reduce the condition number.

(*iii*) Algorithm 11.3 requires the following operations:

- vector space operations in U,
- evaluations of inner products in U,
- evaluations of $A\,r$ for $r \in U$.

Even after discretization, the latter is usually realized in a "matrix-free" way, i. e., using a routine that evaluates $A\,r$ for given vector r, without explicitly forming the matrix A.

(*iv*) In the context of the optimal control problem where $A = S^{\circ}S + \gamma\,\mathrm{id}$, the evaluation of $A\,r$ amounts to solving one state equation and one adjoint equation; compare Lemma 10.3. △

End of Class 22

End of Week 13

§ 11.2 Conjugate Gradient Algorithm

The typical inefficient zig-zaggging pattern of the directions $d^{(k)}$ is a consequence of the fact that gradient descent is a memoryless method. That is, we could restart the method at any iterate and it would produce the same iterates, whether restarted or not. This is where the **conjugate gradient method** (**CG method**, introduced in Hestenes, Stiefel, 1952) takes a different turn. It works with search directions $d^{(k)}$ which are pairwise A-orthogonal (also known as A-conjugate), and builds a memory of previously visited directions.

Definition 11.6 (Conjugate directions)**.**
Suppose that U is a Hilbert space, $A \in \mathcal{L}(U)$ is self-adjoint and positive definite. A set of non-zero vectors $\{d^{(0)}, \dots, d^{(k)}\} \subseteq U$ is termed A**-conjugate** if

$$\big(A\,d^{(i)}, d^{(j)}\big)_U = 0 \quad \text{for } 0 \leqslant i, j \leqslant k, \quad i \neq j. \qquad \triangle$$

In other words, A-conjugate vectors are pairwise orthogonal w.r.t. the A-inner product. In particular, $\{d^{(0)}, \dots, d^{(k)}\}$ is a linearly independent set. (**Quiz 11.1:** Can you prove that?)

The CG method is a member of the class of **conjugate direction methods**. We begin by describing the properties of a generic conjugate direction method first before we particularize to the CG method. A conjugate direction method chooses its search directions $d^{(0)}, d^{(1)}, \dots$ so that they are A-conjugate, and the iterates satisfy

$$u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)}. \tag{11.7}$$

The step size $\alpha^{(k)}$ is the Cauchy step size, which minimizes the one-dimensional quadratic polynomial

$$\alpha \mapsto f(u^{(k)} + \alpha\, d^{(k)}).$$

That is, we have

$$\alpha^{(k)} := -\frac{\big(r^{(k)}, d^{(k)}\big)_U}{\big(A\,d^{(k)}, d^{(k)}\big)_U}. \tag{11.8}$$

As in the gradient descent method, the residuals satisfy the recursion

$$r^{(k+1)} = r^{(k)} + \alpha^{(k)} A\,d^{(k)}. \tag{11.9}$$

Conjugate direction methods have the remarkable property that the sequence of one-dimensional minimizations in the A-conjugate directions $d^{(0)}, d^{(1)}, \dots$ is equivalent to the minimization over the entire affine subspace $u^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots\}$. This is shown in the following result.

Lemma 11.7 (Properties of conjugate direction methods).
Suppose that U is a Hilbert space, $A \in \mathcal{L}(U)$ is self-adjoint and positive definite. Given an initial guess $u^{(0)}$ and a set $\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$, $k \geqslant 1$ of A-conjugate search directions, suppose that the iterates $u^{(0)}, \dots, u^{(k)}$ are generated according to (11.7) with Cauchy step size (11.8). Then the following holds.

(*i*)
$$\left(r^{(k)}, d^{(i)}\right)_U = 0 \quad \text{for all } i = 0, 1, \dots, k-1. \tag{11.10}$$

(*ii*) $u^{(k)}$ minimizes f over the affine subspace $u^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$.

Proof. We can show statement (*i*) via induction over k. For $k = 1$,

$$\begin{aligned}
\left(r^{(1)}, d^{(0)}\right)_U &= \left(A\,u^{(1)} - b, d^{(0)}\right)_U && \text{by definition of the residual}\\
&= \left(A\,u^{(0)} + \alpha^{(0)} A\,d^{(0)} - b, d^{(0)}\right)_U && \text{by (11.7)}\\
&= \left(r^{(0)}, d^{(0)}\right)_U + \alpha^{(0)} \left(d^{(0)}, A\,d^{(0)}\right)_U && \text{by definition of the residual}\\
&= 0 && \text{since } \alpha^{(0)} \text{ is the Cauchy step size (11.8).}
\end{aligned}$$

The induction step assumes $\left(r^{(k-1)}, d^{(i)}\right)_U = 0$ for all $i = 0, 1, \dots, k-2$ and proceeds as follows.

$$\begin{aligned}
\left(r^{(k)}, d^{(k-1)}\right)_U &= \left(r^{(k-1)} + \alpha^{(k-1)} A\,d^{(k-1)}, d^{(k-1)}\right)_U && \text{by the residual recursion (11.9)}\\
&= 0 && \text{since } \alpha^{(k-1)} \text{ is the Cauchy step size (11.8).}
\end{aligned}$$

For the remaining search directions $d^{(i)}$, $i = 0, 1, \dots, k-2$ we have

$$\begin{aligned}
\left(r^{(k)}, d^{(i)}\right)_U &= \left(r^{(k-1)} + \alpha^{(k-1)} A\,d^{(k-1)}, d^{(i)}\right)_U && \text{by the residual recursion (11.9)}\\
&= \underbrace{\left(r^{(k-1)}, d^{(i)}\right)_U}_{=0 \text{ by assumption}} + \alpha^{(k-1)} \underbrace{\left(d^{(k-1)}, A\,d^{(i)}\right)_U}_{=0 \text{ due to } A\text{-conjugacy}}\\
&= 0.
\end{aligned}$$

For statement (*ii*) we consider the function $h\colon \mathbb{R}^k \to \mathbb{R}$

$$h(\sigma) := f\left(u^{(0)} + \sum_{j=0}^{k-1} \sigma_j\, d^{(j)}\right).$$

h is strongly convex (**Quiz 11.2:** Why?), and the unique minimizer σ^* is characterized by

$$\frac{\partial h(\sigma^*)}{\partial \sigma_i} = \left(\nabla f\left(u^{(0)} + \sum_{j=0}^{k-1} \sigma_j^*\, d^{(j)}\right), d^{(i)}\right)_U = 0, \quad i = 0, \dots, k-1. \tag{11.11}$$

However, we already know that it is the iterate

$$u^{(k)} = u^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)}\, d^{(j)} \in u^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$$

which satisfies (11.11), since

$$\left(\nabla f\left(u^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)}\, d^{(j)}\right), d^{(i)}\right)_U = \left(\nabla f(u^{(k)}), d^{(i)}\right)_U = \left(r^{(k)}, d^{(i)}\right)_U = 0$$

holds for all $i = 0, \dots, k-1$, as shown in statement (*i*). □

Differences to the gradient descent method are highlighted.

Algorithm 11.8 (conjugate gradient algorithm for unconstrained quadratic problems)**.**
Input: initial guess $u^{(0)} \in U$
Input: relative tolerance $\varepsilon > 0$
Output: approximate minimizer of (11.1)

1: Set $r^{(0)} := \nabla f(u^{(0)}) = A\,u^{(0)} - b \in U$
2: Set $d^{(0)} := -r^{(0)}$ // initialize the search direction with the negative gradient direction
3: Set $\delta^{(0)} := \|r^{(0)}\|_U^2$
4: Set $n := 0$
5: **while** $\delta^{(n)} > \varepsilon^2 \delta^{(0)}$ **do**
6: Set $q^{(n)} := A\,d^{(n)}$
7: Set $\alpha^{(n)} := \frac{\delta^{(n)}}{(q^{(n)}, d^{(n)})_U}$ // evaluate the step size
8: Set $u^{(n+1)} := u^{(n)} + \alpha^{(n)}\, d^{(n)}$ // update the control
9: Set $r^{(n+1)} := r^{(n)} + \alpha^{(n)}\, q^{(n)}$ // update the residual
10: Set $\delta^{(n+1)} := \|r^{(n+1)}\|_U^2$
11: Set $\beta^{(n+1)} := \frac{\delta^{(n+1)}}{\delta^{(n)}}$
12: Set $d^{(n+1)} := -r^{(n+1)} + \beta^{(n+1)}\, d^{(n)}$
13: Set $n := n + 1$
14: **end while**
15: **return** $u^{(n)}$

Remark 11.9 (on Algorithm 11.8)**.**

(*i*) Line 12 A-orthogonalizes the negative gradient direction $-r^{(n+1)}$ against the most recent search direction $d^{(n)}$. In fact, one can show that this is enough for the entire sequence $d^{(0)}, d^{(1)}, \ldots$ to be A-conjugate. This phenomenon, known as **short-term recurrence**, is possible due to the self-adjointness of A.

(*ii*) The conjugate thus keeps a memory of previously visited directions, although this memory is mainly implicit. As shown in Algorithm 11.8, we can implement the method with a constant amount of storage.

(*iii*) The implementation of the CG method is very similar to the steepest descent method (Algorithm 11.3). The only (but significant!) difference lies in the fact that we A-orthogonalize the steepest descent direction against $d^{(k)}$ before we use it as the new search direction $d^{(k+1)}$. The initial search direction $d^{(0)}$ is the steepest descent direction for f at $u^{(0)}$. Consequently, the iterate $u^{(1)}$ is the same for the conjugate gradient method and the steepest descent method with Cauchy step size.

(*iv*) The name **conjugate gradient method** is a bit of a misnomer, since it is not the gradients which are A-conjugate, but rather the search directions $d^{(k)}$. △

We now establish a convergence result for the conjugate gradient method, and to compare it to the steepest descent method with Cauchy step size. A major difference is that we will not obtain a result about the reduction of the error from iteration to iteration, but rather a result about the reduction of the error compared with its initial value.

Theorem 11.10 (Convergence of Algorithm 11.8, compare Theorem 11.4)**.**
Suppose that U is a Hilbert space, $A \in \mathcal{L}(U)$ is self-adjoint and positive definite. Specifically, suppose that

$$\alpha \|u\|_U^2 \leqslant (A\,u, u)_U \leqslant \beta \|u\|_U^2$$

holds for all $u \in U$. Then the conjugate gradient descent method (Algorithm 11.3) with the stopping criterion turned off converges to the unique minimizer of the quadratic objective (11.1), regardless of the initial value $u^{(0)}$. In terms of the condition number $\kappa = \beta/\alpha$, we have the estimates[8]

$$f(u^{(k)}) - f(u^*) \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \left(f(u^{(0)}) - f(u^*)\right) \tag{11.12a}$$

$$\|u^{(k)} - u^*\|_A \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k} \|u^{(0)} - u^*\|_A, \tag{11.12b}$$

Moreover, the objective values $f(u^{(k)})$ and thus the norm of the error $\|u^{(k)} - u^*\|_A$ are monotonically decreasing.

The proof is again the same as in finite dimensions; see, e. g., Herzog, 2023, Theorem 4.19.

End of Class 23

§ 12 Algorithms for Reduced Linear-Quadratic Problems With Inequality Constraints

In this section, we consider the minimization of (11.1) subject to additional inequality constraints for the control variable u. In order for these constraints to be meaningful, we assume that the control space is $U = L^2(\Omega_{\text{ctrl}})$ and we consider

$$\begin{aligned} \text{Minimize} \quad & f(u) = \frac{1}{2}(A\,u, u)_U - (b, u)_U + c \\ \text{s. t.} \quad & u \in U_{\text{ad}} = \{u \in L^2(\Omega_{\text{ctrl}}) \mid u_a \leqslant u \leqslant u_b \text{ a.e. in } \Omega_{\text{ctrl}}\} \end{aligned} \tag{12.1}$$

with $u_a \leqslant u_b$ where $u_a \in \mathbb{R} \cup \{-\infty\}$ and $u_b \in \mathbb{R} \cup \{\infty\}$. As in § 11, we assume that $A \in \mathcal{L}(U)$ is self-adjointness and positive definite.

An obvious choice for an algorithm to solve this problem is the projected gradient method; see, e. g., Tröltzsch, 2010, Chapter 2.12. However, since the gradient descent method is only Q-linearly convergent even in the unconstrained setting, we will not consider it here.

Instead, we describe an **active-set method**, which estimates the subsets of the control space where the constraints are active. That is, the method works with iteration-dependent active/inactive sets $\mathcal{A}^+$, $\mathcal{A}^-$ and $\mathcal{I}$, which form a decomposition of the control domain Ω_{ctrl}. Notice that with active sets deemed known, the problem reduces to

$$\begin{aligned} \text{Minimize} \quad & f(u) = \frac{1}{2}(A\,u, u)_U - (b, u)_U + c \\ \text{s. t.} \quad & u = u_b \quad \text{on } \mathcal{A}^+ \\ \text{and} \quad & u = u_a \quad \text{on } \mathcal{A}^-, \end{aligned} \tag{12.2}$$

[8] compare (11.6c), (11.6d)

which is essentially an unconstrained problem for $u \in L^2(\mathcal{I})$.

We now need to specify how we determine the active sets $\mathcal{A}^+$ and $\mathcal{A}^-$. To this end, we consider an active set strategy that was originally developed in Bergounioux, Ito, Kunisch, 1999. Its name, the **primal-dual active set strategy**, derives from the fact that both the primal variable u as well as the dual variable, the (signed) multiplier μ, are used to determine the active sets.

The resulting algorithm is given as follows.

Algorithm 12.1 (primal-dual active set algorithm for bound-unconstrained quadratic problems)**.**
Input: initial guess $u^{(0)} \in U = L^2(\Omega_{\text{ctrl}})$
Input: initial guess $\mu^{(0)} \in U = L^2(\Omega_{\text{ctrl}})$
Input: active-set parameter $c > 0$
Output: approximate minimizer of (12.1)

1: Set $n := 0$
2: **while** not converged **do**
3: Determine the active sets

$$\begin{aligned}
\mathcal{A}_+^{(n+1)} &:= \left\{ x \in \Omega_{\text{ctrl}} \,\middle|\, \mu^{(n)} + c\,(u^{(n)} - u_b) > 0 \right\} \\
\mathcal{A}_-^{(n+1)} &:= \left\{ x \in \Omega_{\text{ctrl}} \,\middle|\, \mu^{(n)} + c\,(u^{(n)} - u_a) < 0 \right\} \\
\mathcal{I}^{(n+1)} &:= \Omega_{\text{ctrl}} \setminus \left(\mathcal{A}_+^{(n)} \cup \mathcal{A}_-^{(n)} \right)
\end{aligned}$$

4: Solve the problem (12.2) for $u^{(n+1)}$ with active sets $\mathcal{A}_\pm^{(n+1)}$ (and no constraints on $\mathcal{I}^{(n+1)}$)
5: Set $\mu^{(n+1)} := b - A\,u^{(n+1)}$
6: Set $n := n + 1$
7: **end while**
8: **return** $u^{(n)}$ and $\mu^{(n)}$

The iterates $(u^{(n)}, \mu^{(n)})$ of Algorithm 12.1 for $n \geqslant 1$ satisfy

$$\mu^{(n)} \left(u_a - u^{(n)}\right) \left(u^{(n)} - u_b\right) = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}$$

since the subset on which $u_a < u^{(n)} < u_b$ holds is necessarily a subset of $\mathcal{I}^{(n)}$, so the optimality conditions for (12.2) imply $\mu^{(n)} = 0$ there. However, the primal iterates $u^{(n)}$ are not necessarily feasible, i. e., they generally fail to satisfy the constraints $u_a \leqslant u^{(n)} \leqslant u_b$ on the inactive set $\mathcal{I}^{(n)}$. Moreover, the multiplier may have the wrong sign on some subset.

It turns out that as soon as the active sets coincide in two consecutive iterations, the unique solution of (12.1) has been found:

Lemma 12.2 (exact stopping)**.**

Suppose that $\mathcal{A}_-^{(n)} = \mathcal{A}_-^{(n+1)}$ and $\mathcal{A}_+^{(n)} = \mathcal{A}_+^{(n+1)}$ holds for some $n \in \mathbb{N}_0$. Then $u^{(n+1)}$ is the unique minimizer of (12.1), and $\mu^{(n+1)}$ is the corresponding (signed) Lagrange multiplier.

Proof. We need to verify that $(u^{(n+1)}, \mu^{(n+1)})$ satisfies the necessary and sufficient optimality conditions (compare Theorem 10.7)

$$A\,u - b + \mu = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}, \tag{12.3a}$$

$$[\mu]^- \geqslant 0, \quad u_a - u \leqslant 0, \quad [\mu]^-\,(u_a - u) = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}, \tag{12.3b}$$

$$[\mu]^+ \geqslant 0, \quad u - u_b \leqslant 0, \quad [\mu]^+\,(u - u_b) = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}. \tag{12.3c}$$

Equation (12.3a) holds by construction for $(u^{(n+1)}, \mu^{(n+1)})$. It remains to verify the feasibility of $u^{(n+1)}$ as well as the sign conditions $\mu^{(n+1)} = 0$ on $\mathcal{I}^{(n+1)}$, $\mu^{(n+1)} > 0$ on $\mathcal{A}_+^{(n+1)}$, and $\mu^{(n+1)} < 0$ on $\mathcal{A}_-^{(n+1)}$.

On $\mathcal{I}^{(n+1)}$, we have $\mu^{(n+1)} = 0$ by construction. Moreover, we have

$$\mu^{(n)} + c\,(u^{(n)} - u_a) \geqslant 0 \quad \text{on } \mathcal{I}^{(n+1)}$$

since $\mathcal{I}^{(n+1)}$ is disjoint from $\mathcal{A}_-^{(n+1)}$. Due to $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$, the result of the previous step implies $\mu^{(n)} = 0$ on $\mathcal{I}^{(n+1)}$ so that $u^{(n)} - u_a \geqslant 0$ on $\mathcal{I}^{(n+1)}$ follows. By a similar reasoning, we also conclude $u^{(n)} - u_b \leqslant 0$ on $\mathcal{I}^{(n+1)}$.

On $\mathcal{A}_+^{(n)}$, we have $u^{(n)} = u_b$, and $\mathcal{A}_+^{(n+1)} = \mathcal{A}_+^{(n)}$ implies

$$\mu^{(n)} = \mu^{(n)} + c\,(u^{(n)} - u_b) > 0 \quad \text{on } \mathcal{A}_+^{(n+1)}.$$

Similarly, we obtain

$$\mu^{(n)} = \mu^{(n)} + c\,(u^{(n)} - u_a) < 0 \quad \text{on } \mathcal{A}_-^{(n+1)}. \qquad \square$$

We now revisit the proof of global convergence of Algorithm 12.1 that was given in Kunisch, Rösch, 2002 for the unilaterally constrained case. In that case, $\mathcal{A}_-^{(n)} = \emptyset$ holds throughout the algorithm.

Theorem 12.3 (global convergence of Algorithm 12.1)**.**
Suppose that Ω_{ctrl} is a bounded domain in $\mathbb{R}^d$. Consider Hilbert spaces U, H with $U = L^2(\Omega_{\text{ctrl}})$ and $S \in \mathcal{L}(U, H)$. Suppose $A = S^\circ S + \gamma\,\mathrm{id}$, $u_a = -\infty$ and $u_b \in \mathbb{R}$ holds. Finally, assume $\gamma > 2\,\|S\|^2_{\mathcal{L}(U,H)}$ for the control cost parameter. Then there exists a choice for the active-set parameter $c > 0$ satisfying

$$\gamma + \varepsilon \;<\; c \;<\; \gamma - \frac{\gamma^2}{\varepsilon} + \frac{\gamma^2}{\|S\|^2_{\mathcal{L}(U,H)}} \tag{12.4}$$

for some $\varepsilon > 0$. For any c satisfying (12.4), the primal-dual active set algorithm (Algorithm 12.1) converges to the unique minimizer of (12.1), regardless of the initial value $u^{(0)}$.

Sketch of the proof. In the proof we write $\|S\|$ in place of $\|S\|_{\mathcal{L}(U,H)}$.

Step 1: We prove that $\gamma > 2\,\|S\|^2$ is sufficient to satisfy (12.4).

Consider the function

$$\varepsilon \mapsto \varepsilon + \frac{\gamma^2}{\varepsilon} - \frac{\gamma^2}{\|S\|^2}.$$

It is smooth, bounded below and radially unbounded on $\mathbb{R}_{>0}$ and thus attains a minimum. Since it is also strictly convex, the minimizer is unique and characterized by a zero derivative, which happens at $\varepsilon = \gamma$. The function value there is

$$v^* := 2\,\gamma - \frac{\gamma^2}{\|S\|^2}$$

and $v^* < 0$ holds due to the assumption.

We have shown that there exists some $\varepsilon > 0$ satisfying

$$\begin{aligned} & \varepsilon + \frac{\gamma^2}{\varepsilon} - \frac{\gamma^2}{\|S\|^2_{\mathcal{L}(U,H)}} < 0 \\ \Rightarrow \quad & \gamma + \varepsilon < \gamma - \frac{\gamma^2}{\varepsilon} + \frac{\gamma^2}{\|S\|^2_{\mathcal{L}(U,H)}}, \end{aligned}$$

which proves the assertion.

Step 2: We prove that

$$f(u^{(n)}) - f(u) = -\frac{1}{2}\big(A\,(u - u^{(n)}), u - u^{(n)}\big)_{L^2(\Omega_{\text{ctrl}})} + \big(\mu^{(n)}, u - u^{(n)}\big)_{L^2(\mathcal{A}_+^{(n)})} \tag{12.5}$$

holds for any $u \in L^2(\Omega_{\text{ctrl}})$.

Step 3: Consider the augmented Lagrangian function

$$\begin{aligned} & \mathcal{L}_c(u,\mu) := f(u) + \frac{1}{2\,c}\big\|[\mu + c\,(u - u_b)]^+\big\|^2_{L^2(\Omega_{\text{ctrl}})} - \frac{1}{2\,c}\|\mu\|^2_{L^2(\Omega_{\text{ctrl}})} \\ \text{as well as} \quad & \mathcal{L}_c^+(u,\mu) := \mathcal{L}_c(u, [\mu]^+). \end{aligned} \tag{12.6}$$

We prove using (12.4) that in any iteration of Algorithm 12.1, we either have $\mathcal{L}_c^+(u^{(n)}, \mu^{(n)}) < \mathcal{L}_c^+(u^{(n+1)}, \mu^{(n+1)})$, or else $u^{(n+1)} = u^{(n)}$ holds and the algorithm has converged to the solution. □

Remark 12.4 (on Algorithm 12.1)**.**

(*i*) In every iteration of Algorithm 12.1, we need to set u to one of the bounds on $\mathcal{A}_\pm$ and solve an unconstrained linear-quadratic problem for $u \in L^2(\mathcal{I})$. This can be achieved by a relatively straightforward modification of the conjugate gradient algorithm 11.8 that is instructed to work only on the subspace $L^2(\mathcal{I})$.

(*ii*) Under the additional assumption that $S \in \mathcal{L}(U,H)$ is a **compact operator**[9], it can be shown that the iterates of Algorithm 12.1 $u^{(n)}$ converge *strongly* to the unique solution.

(*iii*) Interestingly, the primal-dual active set algorithm 12.1 can be viewed as a generalized form of Newton's method; see Hintermüller, Ito, Kunisch, 2002. To see this, one rewrites the optimality conditions (12.3) using the **complementarity function**[10] $\Psi(a,b) = \max\{a, c\,b\}$ as

$$A\,u - b + \mu = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}, \tag{12.7a}$$

$$\mu - \min\{0,\ \mu + c\,(u - u_a)\} - \max\{0,\ \mu + c\,(u - u_b)\} = 0 \quad \text{a.e. in } \Omega_{\text{ctrl}}. \tag{12.7b}$$

When $A = S^\circ S + \gamma\,\mathrm{id}$ holds, where S satisfies a certain smoothing condition (typically satisfied for control-to-observation maps involving a PDE), and $c := \gamma$ is set, then the left-hand side of the root-finding problem (12.7) can be shown to admit a generalized form of differentiability suitable for Newtwon's method; see Ito, Kunisch, 2002. This can be used to show that Algorithm 12.1 converges locally superlinearly. △

[9] A linear map $S \in \mathcal{L}(U,H)$ between normed linear spaces is said to be **compact** if for any bounded subset $V \subseteq U$, the closure of $S(V) \subseteq H$ is a compact set. Equivalently, for every bounded sequence $(u^{(n)})$, the sequence $S(u^{(n)})$ contains a convergent subsequence.

[10] Generally, a **complementarity function** $\Psi\colon \mathbb{R}^2 \to \mathbb{R}$ is a function satisfying $\Phi(a,b) = 0 \Leftrightarrow a \leqslant 0,\ b \leqslant 0,\ a\,b = 0$.

End of Class 24

End of Week 14

Index

Bibliography

Adams, R.; J. Fournier (2003). *Sobolev Spaces*. 2nd ed. New York: Academic Press.

Barbu, V.; T. Precupanu (2012). *Convexity and Optimization in Banach Spaces*. 4th ed. Springer Monographs in Mathematics. Springer, Dordrecht. DOI: 10.1007/978-94-007-2247-7.

Bergounioux, M.; K. Ito; K. Kunisch (1999). "Primal-dual strategy for constrained optimal control problems". *SIAM Journal on Control and Optimization* 37.4, pp. 1176–1194. DOI: 10.1137/s0363012997328609.

Dunford, N.; J. T. Schwartz (1988). *Linear Operators. Part I: General Theory*. Wiley Classics Library. John Wiley & Sons, Inc., New York.

Evans, L. C. (1998). *Partial Differential Equations*. Vol. 19. Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society. DOI: 10.1090/gsm/019.

Gagliardo, E. (1957). "Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili". *Rendiconti del Seminario Matematico della Università di Padova* 27, pp. 284–305. URL: http://www.numdam.org/item/RSMUP_1957__27__284_0/.

Herzog, R. (2023). *Nonlinear Optimization*. Lecture notes. URL: https://scoop.iwr.uni-heidelberg.de/teaching/2023ss/lecture-nonlinear-optimization/.

Hestenes, M. R.; E. Stiefel (1952). "Methods of conjugate gradients for solving linear systems". *Journal of Research of the National Bureau of Standards* 49, 409–436 (1953). DOI: 10.6028/jres.049.044.

Heuser, H. (1992). *Funktionalanalysis*. 3rd ed. Mathematische Leitfäden. Theorie und Anwendung. Stuttgart: B. G. Teubner, p. 696.

Hintermüller, M.; K. Ito; K. Kunisch (2002). "The primal-dual active set strategy as a semismooth Newton method". *SIAM Journal on Optimization* 13.3, pp. 865–888. DOI: 10.1137/s1052623401383558.

Ito, K.; K. Kunisch (2002). "Semi-smooth Newton methods for variational inequalities of the first kind". *RAIRO Modélisation Mathématique et Analyse Numérique* 37, pp. 41–62. DOI: 10.1051/m2an:2003021.

Kunisch, K.; A. Rösch (2002). "Primal-dual active set strategy for a general class of constrained optimal control problems". *SIAM Journal on Optimization* 13.2, pp. 321–334. DOI: 10.1137/s1052623499358008.

Meyers, N.; J. Serrin (1964). "H=W". *Proceedings of the National Academy of Sciences* 51, pp. 1055–1056. DOI: 10.1073/pnas.51.6.1055.

Nečas, J. (2012). *Direct Methods in the Theory of Elliptic Equations*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-10455-8.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill.

Tröltzsch, F. (2010). *Optimal Control of Partial Differential Equations*. Vol. 112. Graduate Studies in Mathematics. Providence: American Mathematical Society. DOI: 10.1090/gsm/112.

Werner, D. (2007). *Funktionalanalysis*. 8th ed. Berlin/Heidelberg: Springer. DOI: 10.1007/978-3-662-55407-4.