Optimal Control of Partial Differential Equations

Peter Philip*

Lecture Notes

Originally Created for the Class of Spring Semester 2007 at HU Berlin,
Revised and Extended for the Class of Spring Semester 2009 at LMU Munich†

October 17, 2013

Contents

*E-Mail: philip@math.lmu.de

†Roughly based on Fredi Tröltzsch [Trö05]. I wish to thank Olaf Klein for pointing out several errors contained in my original manuscript. Any remaining errors contained in the lecture notes are solely my responsibility.

1 Motivating Examples

As motivating examples, we will consider several variants of optimal heating problems. Further examples of applied problems can be found in Sections 1.2 and 1.3 of [Trö05].

1.1 Stationary Optimal Heating Problems

1.1.1 General Setting

The equilibrium distribution of the absolute temperature $y : \Omega \longrightarrow \mathbb{R}^+$ inside a body $\Omega \subseteq \mathbb{R}^3$ (see Fig. 1) is determined by the stationary heat equation

$$-\operatorname{div}(\kappa \nabla y) = f, \tag{1.1}$$

where κ is the body's thermal conductivity, and $f : \Omega \longrightarrow \mathbb{R}_0^+$ represents possible heat sources. In the simplest situation, κ is a positive constant, but, in general, it can depend on both y and on the space coordinate $x \in \Omega$.

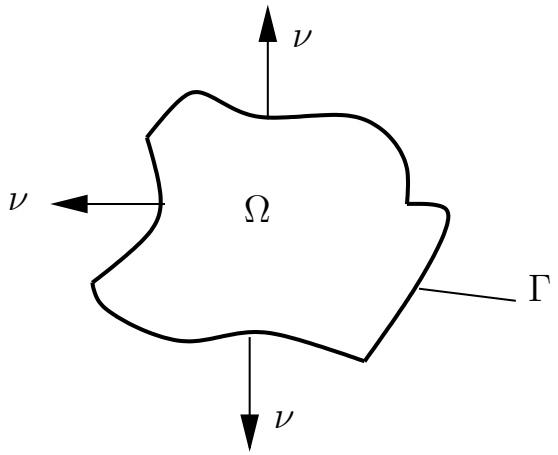

Figure 1: Visualization of the space domain Ω.

To complete the problem description for the determination of the equilibrium temperature distribution in Ω, one still needs to formulate boundary conditions on $\Gamma := \partial\Omega$. The appropriate choice of boundary condition depends on the physical situation to be modeled as well as on what quantity can be physically measured and controlled in the situation of interest. If the temperature on Γ is known, then one will use a *Dirichlet* boundary condition, i.e.

$$y = y_{\mathrm{D}} \quad \text{on } \Gamma, \tag{1.2}$$

where $y_{\mathrm{D}} : \Gamma \longrightarrow \mathbb{R}^+$ is the known temperature on Γ. For example, if the interior of some apparatus is heated in a room temperature environment, then, by choosing Ω sufficiently large, one can ensure that y_{D} is known to be room temperature. In a different situation, the temperature distribution y_{D} on Γ might also be known if it is controlled by means of a heating device.

We are now in a position to formulate some first *optimal control problems*. The general idea is to vary (i.e. control) an input quantity (called the *control*, typically denoted by u) such that some output quantity (called the *state*, typically denoted by y) has a desired property. This desired property is measured according to some function J, called the *objective function* or the *objective functional* (if it is defined on a space of infinite dimension, as is the case when controlling partial differential equations). Usually, the objective functional is formulated such that the desired optimal case coincides with a minimum of J. In general, J can depend on both the control u and on the state y. However, if there exists a unique state for each control (i.e. if there is a map $S : u \mapsto y = S(u)$), then J can be considered as a function of the control alone. We will mostly concentrate on this latter situation, considering partial differential equations that admit a unique solution y for each control u.

When controlling partial differential equations (PDE), the state y is the quantity determined as the solution of the PDE, whereas the control can be an input function prescribed on the boundary Γ (so-called *boundary control*) or an input function prescribed on the volume domain Ω (so-called *distributed control*).

In the context of optimal heating, where the state y is the absolute temperature determined as a solution of the heat equation (1.1), we will now consider one example of boundary control (Sec. 1.1.2) and one example of distributed control (Sec. 1.1.3).

1.1.2 Boundary Control

Consider the case that, for a desired application, the optimal temperature distribution $y_\Omega : \Omega \longrightarrow \mathbb{R}^+$ is known and that heating elements can control the temperature $u := y_D$ at each point of the boundary Γ. The goal is to find u such that the actual temperature y approximates y_Ω. This problem leads to the minimization of the objective functional

$$J(y,u) := \frac{1}{2}\int_\Omega \big(y(x) - y_\Omega(x)\big)^2 \,\mathrm{d}x \;+ \frac{\lambda}{2}\int_\Gamma u(x)^2 \,\mathrm{d}s(x)\,, \tag{1.3}$$

where $\lambda > 0$, and s is used to denote the surface measure on Γ. The second integral in (1.3) is a typical companion of the first in this type of problem. It can be seen as a measure for the expenditure of the control. For instance, in the present example, it can be interpreted as measuring the energy costs of the heating device. In mathematical terms, the second integral in (1.3) has a *regularizing* effect; it is sometimes called a *Tychonoff regularization*. It counteracts the tendency of the control to become locally unbounded and rugged as J approaches its infimum.

Due to physical and technical limitations of the heating device, one needs to impose some restrictions on the control u. Physical limitations result from the fact that any device will be destroyed if its temperature becomes to high or to low. However, the technical limitations of the heating device will usually be much more restrictive, providing upper and lower bounds for the temperatures that the device can impose. Hence, one is led to the *control constraints*

$$a \le u \le b \quad \text{on } \Gamma, \tag{1.4}$$

where $0 < a < b$. Control constraints of this form are called *box constraints*.

If, apart from the control, the system does not contain any heat sources, then $f \equiv 0$ in (1.1) and the entire optimal control problem can be summarized as follows:

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(x) - y_\Omega(x)\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_\Gamma u(x)^2 \,\mathrm{d}s(x)\,, \tag{1.5a}$$

subject to the PDE constraints

$$-\operatorname{div}(\kappa\,\nabla y) = 0 \qquad \text{in } \Omega, \tag{1.5b}$$

$$y = u \qquad \text{on } \Gamma, \tag{1.5c}$$

and control constraints

$$a \le u \le b \quad \text{on } \Gamma. \tag{1.5d}$$

In a slightly more realistic setting, one might only be able to control the temperature on some part Γ_c of Γ with $\Gamma_\mathrm{c} \subsetneq \Gamma$. For example, the goal might be to homogeneously heat a room to a temperature θ_opt, where the heating element has already been installed at a fixed location with boundary Γ_c. In this case, (1.5a) needs to be replaced by a version where the second integral is only carried out over Γ_c. The result is (1.7a) below, where $y_\Omega \equiv \theta_\mathrm{opt}$ was used as well. Since the control is now only given on Γ_c, (1.5c) and (1.5d) also need to be modified accordingly, leading to (1.7c) and (1.7e), respectively. In consequence, one still needs to specify a boundary condition on $\Gamma \setminus \Gamma_\mathrm{c}$. If the surrounding environment is at temperature y_ext, then, according to the Stefan-Boltzmann law of (emitted heat) radiation, the boundary condition reads

$$\nabla y \cdot \nu = \alpha\,(y_\mathrm{ext}^4 - y^4) \quad \text{on } \Gamma \setminus \Gamma_\mathrm{c}, \tag{1.6}$$

where ν denotes the outer unit normal on Γ, and α is a positive constant.

Summarizing this modified optimal control problem yields the following system (1.7):

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(x) - \theta_\mathrm{opt}\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_{\Gamma_\mathrm{c}} u(x)^2 \,\mathrm{d}s(x)\,, \tag{1.7a}$$

subject to the PDE constraints

$$-\operatorname{div}(\kappa\,\nabla y) = 0 \qquad \text{in } \Omega, \tag{1.7b}$$

$$y = u \qquad \text{on } \Gamma_\mathrm{c}, \tag{1.7c}$$

$$\nabla y \cdot \nu = \alpha\,(y_\mathrm{ext}^4 - y^4) \quad \text{on } \Gamma \setminus \Gamma_\mathrm{c}, \tag{1.7d}$$

and control constraints

$$a \le u \le b \quad \text{on } \Gamma_\mathrm{c}. \tag{1.7e}$$

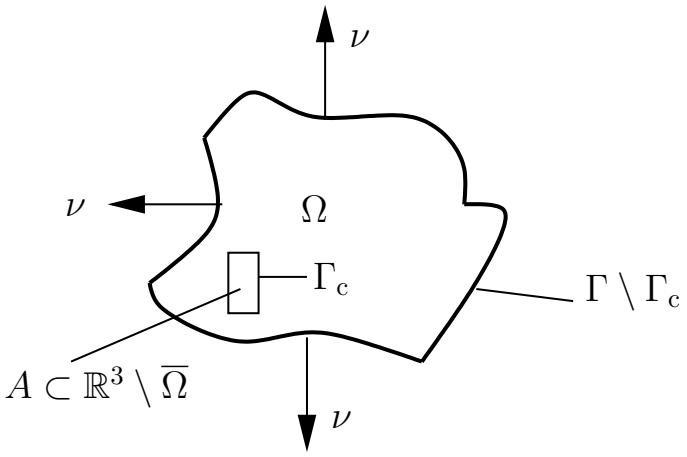

Figure 2: Visualization of the space domain Ω for the optimal heating problem (1.7).

1.1.3 Distributed Control

We now consider the case that we can control the heat sources f inside the domain Ω, setting $f = u$ in (1.1). The control is no longer concentrated on the boundary Γ, but *distributed* over Ω. Such distributed heat sources occur, for example, during electromagnetic or microwave heating.

As u now lives on Ω, the corresponding integration in the objective functional (cf. (1.5a) and (1.7a)) has to be performed over Ω. Similarly, the control constraints now have to be imposed on Ω rather than on the boundary. Thus, keeping the Dirichlet condition from (1.2), the complete optimal control problem reads as follows:

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(x) - y_\Omega(x)\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_\Omega u(x)^2 \,\mathrm{d}x\,, \tag{1.8a}$$

subject to the PDE constraints

$$-\operatorname{div}(\kappa \,\nabla y) = u \qquad \text{in } \Omega, \tag{1.8b}$$

$$y = y_\mathrm{D} \qquad \text{on } \Gamma, \tag{1.8c}$$

and control constraints

$$a \le u \le b \quad \text{on } \Omega. \tag{1.8d}$$

In Sec. 1.1.2 on boundary control, we had considered a second more realistic example, where the goal was to heat a room to a homogeneous temperature θ_opt, but where the control could only be imposed on some strict subset of the boundary. We now consider the analogous situation for distributed control. Here, the control u can usually only be imposed on a strict subset Ω_c of Ω, where Ω_c represents the heating element (see Fig. 3). Then the domain for the second integral in (1.8a) as well as the domain for the control constraints is merely Ω_c. Assuming that there are no uncontrolled heat sources in Ω, (1.8b) has to be replaced by the two equations (1.9b) and (1.9c). If, as in (1.7), one replaces the Dirichlet condition (1.8c) by a Stefan-Boltzmann emission condition,

then one obtains the following modified version of the optimal control problem (1.8):

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(x) - \theta_{\text{opt}}\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_{\Omega_\text{c}} u(x)^2 \,\mathrm{d}x\,, \tag{1.9a}$$

subject to the PDE constraints

$$-\operatorname{div}(\kappa\,\nabla y) = u \qquad \text{in } \Omega_\text{c}, \tag{1.9b}$$
$$-\operatorname{div}(\kappa\,\nabla y) = 0 \qquad \text{in } \Omega\setminus\Omega_\text{c}, \tag{1.9c}$$
$$\nabla y\cdot\nu = \alpha\,(y_{\text{ext}}^4 - y^4) \quad \text{on } \Gamma, \tag{1.9d}$$

and control constraints

$$a \le u \le b \quad \text{on } \Omega_\text{c}. \tag{1.9e}$$

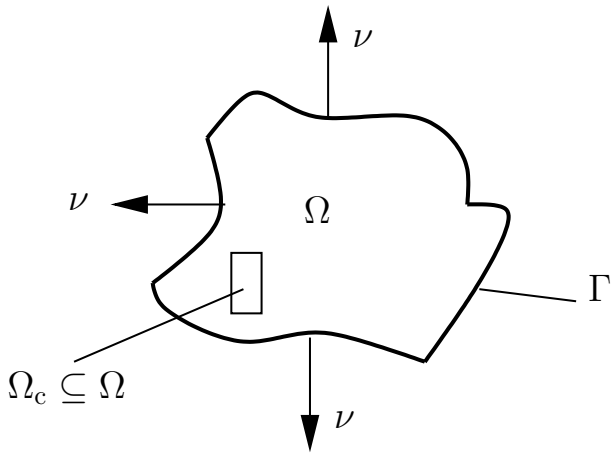

Figure 3: Visualization of the space domain Ω for the distributed control problem (1.9).

1.2 Transient Optimal Heating Problems

1.2.1 General Setting

While (1.1) describes the *equilibrium* temperature distribution inside a body, if the temperature is (still) changing with time t, then it is governed by the transient heat equation

$$\partial_t\, y - \operatorname{div}(\kappa\,\nabla y) = f, \tag{1.10}$$

that merely differs from (1.1) by the presence of the partial derivative with respect to time. Of course, in general, the temperature y and the heat sources f can now depend on time as well as on the space coordinate, i.e. they are defined on the so-called time-space cylinder $[0,T]\times\Omega$, where $T>0$ represents a final time. Here and in the following, we use 0 as the initial time, which is a customary convention. In the transient situation, one needs another condition for a complete problem formulation. One usually starts

the evolution from a known temperature distribution at the initial time $t = 0$, i.e. one starts with an *initial condition*

$$y(0, \cdot) = y_0 \quad \text{in } \Omega, \tag{1.11}$$

where $y_0 : \Omega \longrightarrow \mathbb{R}^+$ is the known initial temperature in Ω.

Each of the optimal control examples considered in Sec. 1.1 can also be considered in a corresponding time-dependent setting. Instead of trying to generate a desired equilibrium temperature y_Ω, one might want to reach y_Ω already for $t = T$. At the same time, it might be possible to vary the control u with time. As in Sec. 1.1, we consider the case, where u controls the temperature on (some part of) the boundary (boundary control, see Sec. 1.2.2), and the case, where u controls the heat sources inside (some part of) Ω (see Sec. 1.2.3).

1.2.2 Boundary Control

The objective functional J (cf. 1.5a) needs to be modified to be suitable for the time-dependent situation. As the temperature y now depends on both time and space, and as the desired temperature field y_Ω should be approximated as good as possible at the final time T, $y(T, \cdot)$ occurs in the first integral in J. The second integral, involving the control u, now needs to be carried out over the time domain as well as over the space domain (see (1.12a)). In the PDE and control constraints, the only modifications are that the constraints are now considered in the respective time-space cylinders and that the initial condition (1.11) is added. Thus, the transient version of (1.5) reads:

$$\text{Minimize} \quad J(y, u) = \frac{1}{2} \int_\Omega \big(y(T, x) - y_\Omega(x)\big)^2 \, \mathrm{d}x + \frac{\lambda}{2} \int_0^T \int_\Gamma u(t, x)^2 \, \mathrm{d}s(x) \, \mathrm{d}t \,, \tag{1.12a}$$

subject to the PDE constraints

$$\partial_t y - \operatorname{div}(\kappa \nabla y) = 0 \quad \text{in } [0, T] \times \Omega, \tag{1.12b}$$

$$y = u \quad \text{on } [0, T] \times \Gamma, \tag{1.12c}$$

$$y(0, \cdot) = y_0 \quad \text{in } \Omega, \tag{1.12d}$$

and control constraints

$$a \le u \le b \quad \text{on } [0, T] \times \Gamma. \tag{1.12e}$$

Similarly, one obtains the following transient version of (1.7):

$$\text{Minimize} \quad J(y, u) = \frac{1}{2} \int_\Omega \big(y(T, x) - \theta_{\text{opt}}\big)^2 \, \mathrm{d}x + \frac{\lambda}{2} \int_0^T \int_{\Gamma_c} u(t, x)^2 \, \mathrm{d}s(x) \, \mathrm{d}t \,, \tag{1.13a}$$

subject to the PDE constraints

$$\partial_t y - \operatorname{div}(\kappa \nabla y) = 0 \quad \text{in } [0, T] \times \Omega, \tag{1.13b}$$

$$y = u \quad \text{on } [0, T] \times \Gamma_c, \tag{1.13c}$$

$$\nabla y \cdot \nu = \alpha \, (y_{\text{ext}}^4 - y^4) \quad \text{on } [0, T] \times (\Gamma \setminus \Gamma_c), \tag{1.13d}$$

$$y(0, \cdot) = y_0 \quad \text{in } \Omega, \tag{1.13e}$$

and control constraints

$$a \le u \le b \quad \text{on } [0,T] \times \Gamma_{\mathrm{c}}. \tag{1.13f}$$

1.2.3 Distributed Control

The way one passes from the stationary to the corresponding transient control problems is completely analogous to the boundary control problems. The first integral in the objective functional J now involves the temperature at the final time T while the second integral is over both space and time. Furthermore, space domains are replaced by time-space cylinders and the initial condition (1.11) is added. Thus, one obtains the transient version of (1.8):

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(T,x) - y_\Omega(x)\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_0^T \int_\Omega u(t,x)^2 \,\mathrm{d}x \,\mathrm{d}t\,, \tag{1.14a}$$

subject to the PDE constraints

$$\partial_t y - \operatorname{div}(\kappa \nabla y) = u \quad \text{in } [0,T] \times \Omega, \tag{1.14b}$$

$$y = y_{\mathrm{D}} \quad \text{on } [0,T] \times \Gamma, \tag{1.14c}$$

$$y(0,\cdot) = y_0 \quad \text{in } \Omega, \tag{1.14d}$$

and control constraints

$$a \le u \le b \quad \text{on } [0,T] \times \Omega. \tag{1.14e}$$

Analogously, one obtains the transient version of (1.9):

$$\text{Minimize} \quad J(y,u) = \frac{1}{2}\int_\Omega \big(y(T,x) - \theta_{\mathrm{opt}}\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_0^T \int_{\Omega_{\mathrm{c}}} u(t,x)^2 \,\mathrm{d}x \,\mathrm{d}t\,, \tag{1.15a}$$

subject to the PDE constraints

$$\partial_t y - \operatorname{div}(\kappa \nabla y) = u \quad \text{in } [0,T] \times \Omega_{\mathrm{c}}, \tag{1.15b}$$

$$\partial_t y - \operatorname{div}(\kappa \nabla y) = 0 \quad \text{in } [0,T] \times \Omega \setminus \Omega_{\mathrm{c}}, \tag{1.15c}$$

$$\nabla y \cdot \nu = \alpha\,(y_{\mathrm{ext}}^4 - y^4) \,\text{on } [0,T] \times \Gamma, \tag{1.15d}$$

$$y(0,\cdot) = y_0 \quad \text{in } \Omega, \tag{1.15e}$$

and control constraints

$$a \le u \le b \quad \text{on } [0,T] \times \Omega_{\mathrm{c}}. \tag{1.15f}$$

2 Convexity

It is already well-known from finite-dimensional optimization that a unique minimum of the objective function can, in general, not be expected if the objective function is not convex. On the other hand, as we will see in Th. 2.17, strict convexity of the

objective function does guarantee uniqueness. Therefore, according to the properties of the considered objective functional, optimal control problems are often classified into convex and nonconvex problems.

We start by reviewing the basic definitions of convex sets and functions in Sec. 2.1, where we will also study sufficient conditions for functions to be convex. These will later be useful to determine the convexity properties of objective functionals.

In Sec. 2.2, we provide the relevant results regarding the relation between the uniqueness of extreme values of functions and the functions' convexity properties.

2.1 Basic Definitions and Criteria for Convex Functions

Definition 2.1. A subset C of a real vector space X is called *convex* if, and only if, for each $(x, y) \in C^2$ and each $0 \leq \alpha \leq 1$, one has $\alpha\, x + (1-\alpha)\, y \in C$.

Lemma 2.2. *Let X, Y be real vector spaces, let $C_1 \subseteq X$, $C_2 \subseteq Y$ be convex. Then $C_1 \times C_2$ is a convex subset of $X \times Y$.*

Proof. Let $(x_1, x_2) \in C_1^2$, $(y_1, y_2) \in C_2^2$, $\alpha \in [0, 1]$. Then

$$\alpha\,(x_1, y_1) + (1-\alpha)\,(x_2, y_2) = \big(\alpha\, x_1 + (1-\alpha)\, x_2,\, \alpha\, y_1 + (1-\alpha)\, y_2\big) \in C_1 \times C_2 \quad (2.1)$$

as $\alpha\, x_1 + (1-\alpha)\, x_2 \in C_1$ and $\alpha\, y_1 + (1-\alpha)\, y_2 \in C_2$ due to the convexity of C_1 and C_2, respectively. ■

Definition 2.3. Let C be a convex subset of a real vector space X.

(a) A function $J : C \longrightarrow \mathbb{R} \cup \{+\infty\}$ is called *convex* if, and only if, for each $(x, y) \in C^2$ and each $0 \leq \alpha \leq 1$:

$$J\big(\alpha\, x + (1-\alpha)\, y\big) \leq \alpha\, J(x) + (1-\alpha)\, J(y). \quad (2.2a)$$

If the inequality in (2.2a) is strict whenever $x \neq y$ and $0 < \alpha < 1$, then J is called *strictly convex*.

(b) A function $S : C \longrightarrow Y$, where Y is another real vector space, is said to *preserve convex combinations* if, and only if, $(x, y) \in C^2$ and each $0 \leq \alpha \leq 1$:

$$S\big(\alpha\, x + (1-\alpha)\, y\big) = \alpha\, S(x) + (1-\alpha)\, S(y). \quad (2.2b)$$

Note that, if S preserves convex combinations, then $S(C)$ is also convex: If $a, b \in S(C)$, then there are $x, y \in C$ such that $S(x) = a$ and $S(y) = b$, and, if $0 \leq \alpha \leq 1$ and S preserve convex combinations, then

$$\alpha a + (1-\alpha) b = \alpha S(x) + (1-\alpha) S(y) = S\big(\alpha x + (1-\alpha) y\big),$$

showing $\alpha a + (1-\alpha) b \in S(C)$.

Remark 2.4. Let X, Y be a real vector spaces, let $C \subseteq X$ be convex. Then $S : C \longrightarrow Y$ preserves convex combinations if, and only if, S is the restriction of an affine map $A : X \longrightarrow Y$ (i.e. if, and only if, there exists a linear map $L : X \longrightarrow Y$ and $a \in X$ such that $A = L + a$ and $S = A\restriction_C$: Suppose S is the restriction of an affine map $A = L + a$ with L and a as above. Then, for each $(x, y) \in C^2$ and each $0 \leq \alpha \leq 1$,

$$\begin{aligned} S\big(\alpha\, x + (1-\alpha)\, y\big) &= \alpha\, Lx + (1-\alpha)\, Ly + a = \alpha\,(Lx + a) + (1-\alpha)\,(Ly + a) \\ &= \alpha\, S(x) + (1-\alpha)\, S(y), \end{aligned} \tag{2.3}$$

showing that S preserves convex combinations. Conversely, assume S preserves convex combinations. Fix $c \in C$. Then there exists a linear map $L : X \longrightarrow Y$ such that $L(x) = S(x + c) - S(c)$ for each $x \in X$ with $x + c \in C$ (as S preserves convex combinations, it canonically extends to the affine hull $\operatorname{aff}(C)$ of C, i.e. $L(x) = S(x + c) - S(c)$ defines a linear map on the linear subspace $V := \operatorname{aff}(C) - c$ of X, which can then be extended to all of X). Thus, letting $a := -L(c) + S(c)$, we obtain, for each $x \in C$: $A(x) = L(x) + a = L(x) - L(c) + S(c) = L(x - c) + S(c) = S(x - c + c) - S(c) + S(c) = S(x)$.

If S preserves convex combinations and $X = \mathbb{R}$, then S is convex, but not strictly convex if C consists of more than one point.

Lemma 2.5. *Let X be a real vector space, let $C \subseteq X$ be convex.*

(a) *If Y is a real vector space, $f : C \longrightarrow Y$ preserves convex combinations and $g : f(C) \longrightarrow \mathbb{R}$ is convex, then $g \circ f$ is convex. If, moreover, f is one-to-one and g is strictly convex, then $g \circ f$ is strictly convex.*

(b) *Suppose $f : C \longrightarrow \mathbb{R}$, let $I \subseteq \mathbb{R}$ be convex such that $f(C) \subseteq I$ and $g : I \longrightarrow \mathbb{R}$. If f and g are both convex and g is increasing, then $g \circ f$ is convex. If, in addition, at least one of the following conditions* (i), (ii) *holds, where*
(i) *f is strictly convex and g is strictly increasing,*
(ii) *f is one-to-one and g is strictly convex,*
then $g \circ f$ is strictly convex.

(c) *If $\lambda \in \mathbb{R}^+$ and $f : C \longrightarrow \mathbb{R}$ is (strictly) convex, then λf is (strictly) convex.*

Proof. Let $(x, y) \in C^2$, $\alpha \in [0, 1]$.

(a): The hypotheses on f and g yield

$$\begin{aligned} (g \circ f)\big(\alpha\, x + (1-\alpha)\, y\big) &= g\Big(f\big(\alpha\, x + (1-\alpha)\, y\big)\Big) = g\big(\alpha\, f(x) + (1-\alpha)\, f(y)\big) \\ &\leq \alpha\, g\big(f(x)\big) + (1-\alpha)\, g\big(f(y)\big) \\ &= \alpha\,(g \circ f)(x) + (1-\alpha)\,(g \circ f)(y), \end{aligned} \tag{2.4}$$

showing that $g \circ f$ is convex. If f is one-to-one and g is strictly convex, then the inequality in (2.4) is strict for $x \neq y$ and $0 < \alpha < 1$, showing that $g \circ f$ is strictly convex.

(b): The convexity of f yields

$$f\big(\alpha\, x + (1-\alpha)\, y\big) \le \alpha\, f(x) + (1-\alpha)\, f(y). \tag{2.5}$$

As g is increasing, one obtains from (2.5)

$$\begin{aligned}(g\circ f)\big(\alpha\, x + (1-\alpha)\, y\big) &= g\Big(f\big(\alpha\, x + (1-\alpha)\, y\big)\Big) \le g\big(\alpha\, f(x) + (1-\alpha)\, f(y)\big)\\ &\le \alpha\, g\big(f(x)\big) + (1-\alpha)\, g\big(f(y)\big)\\ &= \alpha\,(g\circ f)(x) + (1-\alpha)\,(g\circ f)(y),\end{aligned} \tag{2.6}$$

showing the convexity of $g\circ f$. If, in addition, condition (i) is satisfied, then, for $x \ne y$ and $0 < \alpha < 1$, the inequality (2.5) is strict as well as the first inequality in (2.6), proving the strict convexity of $g\circ f$. If condition (ii) is satisfied, then, for $x \ne y$ and $0 < \alpha < 1$, the second inequality in (2.6) is strict, again proving the strict convexity of $g\circ f$.

(c): Here $\lambda > 0$ and the convexity of f imply

$$\lambda f\big(\alpha\, x + (1-\alpha)\, y\big) \le \alpha\, \lambda\, f(x) + (1-\alpha)\, \lambda\, f(y). \tag{2.7}$$

If f is strictly convex, then the inequality in (2.7) is strict whenever $x \ne y$ and $0 < \alpha < 1$. Thus, λf is (strictly) convex given that f is (strictly) convex. ■

Example 2.6. (a) The function $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(x) := |x|$, is convex, but not strictly convex. More generally, if C is a convex subset of a normed vector space X (see Def. 4.1), then $N : C \longrightarrow \mathbb{R}$, $N(x) := \|x\|$ is convex, however not strictly convex if C contains a segment S of a one-dimensional subspace, i.e. $S = \{\lambda x_0 : a \le \lambda \le b\}$ for suitable $x_0 \in X \setminus \{0\}$ and real numbers $0 \le a < b$.

(b) For each $p \in]1, \infty[$, the function $f_p : \mathbb{R} \longrightarrow \mathbb{R}$, $f_p(x) := |x|^p$, is strictly convex.

(c) If C is a convex subset of a normed vector space X and $p \in]1, \infty[$, then $N_p : C \longrightarrow \mathbb{R}$, $N_p(x) := \|x\|^p$, is convex, which follows from (a) and (b) and Lem. 2.5(b) since f_p from (b) is increasing on $[0, \infty[$. The question, whether N_p is strictly convex, is more subtle, cf. Ex. 2.9 below.

Definition 2.7. (a) Let C be a convex subset of a real vector space X. Then p is called an *extreme point* of C if, and only if, $p = \alpha\, x + (1-\alpha)\, y$ with $x, y \in C$ and $0 < \alpha < 1$ implies $p = x = y$ or, equivalently, if, and only if, $p \pm x \in C$ with $x \in X$ implies $x = 0$. The set of all extreme points of C is denoted by $\operatorname{ex}(C)$.

(b) A normed vector space X is called *strictly convex* if, and only if, the set of extreme points of its closed unit ball $\overline{B_1(0)}$ is its entire unit sphere $S_1(0)$, i.e. $\operatorname{ex}\big(\overline{B_1(0)}\big) = S_1(0)$ (see Not. 4.6).

Example 2.8. (a) Every Hilbert space X (see Def. 4.12) is strictly convex, as, for each $p \in S_1(0)$ and $x \in X$ with $p \pm x \in \overline{B_1(0)}$: $1 \ge \|p \pm x\|^2 = \|p\|^2 \pm 2\langle p, x\rangle + \|x\|^2$, i.e. $\|x\|^2 \le \mp 2\langle p, x\rangle$, i.e. $x = 0$.

(b) Clearly, for $n > 1$, $(\mathbb{R}^n, \|\cdot\|_1)$ and $(\mathbb{R}^n, \|\cdot\|_{\max})$ are not strictly convex. The space $L^1[0,1]$ is an example, where even $\operatorname{ex}\big(\overline{B_1(0)}\big) = \emptyset$.

Example 2.9. Let $p \in]1, \infty[$.

(a) If C is a convex subset of a strictly convex normed vector space X, then $N_p : C \longrightarrow \mathbb{R}$, $N_p(x) := \|x\|^p$, is strictly convex: Suppose $x, y \in C$ and $0 < \alpha < 1$. Then

$$\big\|\alpha\, x + (1-\alpha)\, y\big\|^p = \alpha\, \|x\|^p + (1-\alpha)\, \|y\|^p \tag{2.8}$$

implies $\|x\| = \|y\|$ by Ex. 2.6(b). If $\|x\| = 0$, then $x = y = 0$. Otherwise, let $\tilde{x} := x/\|x\|$ and $\tilde{y} := y/\|x\|$ such that $\tilde{x}, \tilde{y} \in S_1(0)$. Then

$$\big\|\alpha\, \tilde{x} + (1-\alpha)\, \tilde{y}\big\|^p = \|x\|^{-p}\, \big\|\alpha\, x + (1-\alpha)\, y\big\|^p \overset{(2.8)}{=} \alpha\, \|\tilde{x}\|^p + (1-\alpha)\, \|\tilde{y}\|^p = 1, \tag{2.9}$$

showing $\alpha\, \tilde{x} + (1-\alpha)\, \tilde{y} \in S_1(0)$ and, thus, $\tilde{x} = \tilde{y}$ as well as $x = y$ by the assumed strict convexity of X.

(b) If X is a normed vector space that is not strictly convex, then there are $x, y \in S_1(0)$, $x \neq y$, and $0 < \alpha < 1$ such that $z := \alpha x + (1-\alpha)y \in S_1(0)$. Then $1 = \|z\|^p = \alpha\|x\|^p + (1-\alpha)\|y\|^p$, showing that $N_p : X \longrightarrow \mathbb{R}$, $N_p(x) := \|x\|^p$, is not strictly convex.

Lemma 2.10. *Let X, Y be real vector spaces, let $C_1 \subseteq X$, $C_2 \subseteq Y$ be convex, and consider $f_1 : C_1 \longrightarrow \mathbb{R}$, $f_2 : C_2 \longrightarrow \mathbb{R}$.*

(a) *If f_1 and f_2 are (strictly) convex, then*

$$(f_1 + f_2) : C_1 \times C_2 \longrightarrow \mathbb{R}, \quad (f_1 + f_2)(y, u) := f_1(y) + f_2(u), \tag{2.10}$$

is (strictly) convex.

(b) *If f_1 and f_2 are convex, and $S : C_2 \longrightarrow C_1$ preserves convex combinations, then*

$$f : C_2 \longrightarrow \mathbb{R}, \quad f(u) := f_1(S(u)) + f_2(u), \tag{2.11}$$

is convex. If at least one of the following additional hypotheses (i) *or* (ii) *is satisfied, then f is strictly convex:*

(i) *f_1 is strictly convex and S is one-to-one.*

(ii) *f_2 is strictly convex.*

Proof. (a): According to Lem. 2.2, $C_1 \times C_2$ is a convex subset of $X \times Y$. Let $(y_1, u_1) \in C_1 \times C_2$, $(y_2, u_2) \in C_1 \times C_2$, and $\alpha \in [0,1]$. Then

$$\begin{aligned}
&(f_1 + f_2)\Big(\alpha\,(y_1, u_1) + (1-\alpha)\,(y_2, u_2)\Big) \\
&= (f_1 + f_2)\Big(\alpha y_1 + (1-\alpha)y_2,\ \alpha u_1 + (1-\alpha)u_2\Big) \\
&= f_1\big(\alpha y_1 + (1-\alpha)y_2\big) + f_2\big(\alpha u_1 + (1-\alpha)u_2\big) \\
&\leq \alpha f_1(y_1) + (1-\alpha) f_1(y_2) + \alpha f_2(u_1) + (1-\alpha) f_2(u_2) \\
&= \alpha(f_1 + f_2)(y_1, u_1) + (1-\alpha)(f_1 + f_2)(y_2, u_2).
\end{aligned} \tag{2.12}$$

If f_1 and f_2 are strictly convex, then equality in (2.12) can only hold for $\alpha \in \{0,1\}$ or $(y_1,u_1)=(y_2,u_2)$, showing the strict convexity of f_1+f_2.

(b): Let $u_1,u_2 \in U$ and $\alpha \in [0,1]$. Then

$$\begin{aligned} f\big(\alpha u_1+(1-\alpha)u_2\big) &= f_1\big(\alpha S(u_1)+(1-\alpha)S(u_2)\big)+f_2\big(\alpha u_1+(1-\alpha)u_2\big) \\ &\le \alpha f_1\big(S(u_1)\big)+(1-\alpha)f_1\big(S(u_2)\big)+\alpha f_2(u_1)+(1-\alpha)f_2(u_2) \\ &= \alpha f(u_1)+(1-\alpha)f(u_2), \end{aligned} \tag{2.13}$$

verifying the convexity of f. If at least one of the additional hypotheses (i) or (ii) is satisfied, then equality in (2.13) can only occur for $\alpha \in \{0,1\}$ or $u_1=u_2$, showing that f is strictly convex in that case. ■

Lemma 2.11. *Let X, Y be normed vector spaces and let $C \subseteq Y$, $U \subseteq X$ be convex. Given $\lambda \in \mathbb{R}_0^+$ and $y_0 \in Y$, consider the functional*

$$J : C \times U \longrightarrow \mathbb{R}, \quad J(y,u) := \frac{1}{2}\|y-y_0\|^2+\frac{\lambda}{2}\|u\|^2. \tag{2.14}$$

(a) *J is convex.*

(b) *If X and Y are strictly convex and $\lambda > 0$, then J is strictly convex.*

(c) *If $S : U \longrightarrow C$ preserves convex combinations, then*

$$f : U \longrightarrow \mathbb{R}, \quad f(u) := J(Su,u) \tag{2.15}$$

is convex. If at least one of the following additional hypotheses (i) *or* (ii) *is satisfied, then f is strictly convex:*

(i) *Y is strictly convex and S is one-to-one.*

(ii) *X is strictly convex and $\lambda > 0$.*

Proof. (a) and (b): According to Lem. 2.2, $C \times U$ is a convex subset of $Y \times X$. Defining

$$f_1 : C \longrightarrow \mathbb{R}, \qquad f_1(y) := \frac{1}{2}\|y-y_0\|^2, \tag{2.16a}$$

$$f_2 : U \longrightarrow \mathbb{R}, \qquad f_2(u) := \frac{\lambda}{2}\|u\|^2, \tag{2.16b}$$

and employing Lem. 2.10(a), it merely remains to show that f_1 is convex (strictly convex if Y is strictly convex), and f_2 is convex (strictly convex if X is strictly convex and $\lambda > 0$).

f_1: The map $y \mapsto \|y-y_0\|^2$ is convex (strictly convex if Y is strictly convex) according to Lem. 2.5(a), as it constitutes a composition of the one-to-one affine map $y \mapsto y-y_0$ (which preserves convex combinations due to Rem. 2.4) with the map $\|\cdot\|^p$, which is always convex according to Ex. 2.6(c) and strictly convex by Ex. 2.9(a), provided that Y is strictly convex.

f_2 is trivially convex for $\lambda = 0$. For $\lambda > 0$, its convexity is a combination of Ex. 2.6(c) with Lem. 2.5(c). If $\lambda > 0$ and X is strictly convex, then f_2 is strictly convex according to Ex. 2.9(a) and Lem. 2.5(c).

(c) now follows from the properties of f_1 and f_2 together with an application of Lem. 2.10(b). ■

Lemma 2.12. *Let X, Y be normed vector spaces, and let $C \subseteq Y$, $U \subseteq X$ be convex,*

$$J : C \times U \longrightarrow \mathbb{R}, \quad S : U \longrightarrow C, \quad f : U \longrightarrow \mathbb{R}, \quad f(u) = J(Su, u). \tag{2.17}$$

If S preserves convex combinations and J is (strictly) convex, then f is (strictly) convex.

Proof. Let $(u, v) \in U^2$ and $0 \leq \alpha \leq 1$. Then

$$\begin{aligned} f\big(\alpha u + (1-\alpha)v\big) &= J\Big(\alpha Su + (1-\alpha)Sv,\, \alpha u + (1-\alpha)v\Big) \\ &= J\Big(\alpha(Su, u) + (1-\alpha)(Sv, v)\Big) \\ &\leq \alpha J(Su, u) + (1-\alpha)J(Sv, v) = \alpha f(u) + (1-\alpha)f(v), \end{aligned} \tag{2.18}$$

showing that f is convex. If J is strictly convex, then equality in (2.18) only holds for $\alpha \in \{0, 1\}$ or $(Su, u) = (Sv, v)$, i.e. only for $\alpha \in \{0, 1\}$ or $u = v$, showing that f is strictly convex if J is strictly convex. ■

Caveat 2.13. In Lem. 2.12, it is *not* sufficient to assume that J is convex in both arguments. For example, consider $J : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$, $J(y, u) = (y-1)^2\,(u+1)^2$. Then $J(\cdot, u)$ is convex for each $u \in \mathbb{R}$ and $J(y, \cdot)$ is convex for each $y \in \mathbb{R}$ (by restricting J to $C := [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$, one can even get J to be strictly convex in both arguments). However, J is *not* convex, and, letting S be the identity, one gets $f(u) = (u-1)^2\,(u+1)^2$. Then f is also *not* convex and has two different global minima, namely at $u = -1$ and $u = 1$ (respectively at $u = -\frac{1}{2}$ and $u = \frac{1}{2}$ if J is restricted to C).

Example 2.14. We investigate the convexity properties of the objective functional from (1.5a), i.e. of

$$J(y, u) := \frac{1}{2}\int_\Omega \big(y(x) - y_\Omega(x)\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_\Gamma u(x)^2 \,\mathrm{d}s(x) \tag{2.19}$$

(the other objective functionals from Sec. 1 can be treated analogously). We assume that the corresponding PDE has a solution operator S that preserves convex combinations (for example, any linear S will work). In other words, for each control u, there is a unique solution $y = S(u)$ to the PDE, and the mapping $u \mapsto S(u)$ preserves convex combinations. Then, instead of J, one can consider the reduced objective functional

$$f(u) := J\big(S(u), u\big) = \frac{1}{2}\int_\Omega \big(S(u)(x) - y_\Omega(x)\big)^2 \,\mathrm{d}x + \frac{\lambda}{2}\int_\Gamma u(x)^2 \,\mathrm{d}s(x)\,. \tag{2.20}$$

Let $U \subseteq L^2(\Gamma)$ be a convex set of admissible control functions, and assume $S(U) \subseteq L^2(\Omega)$. If S preserves convex combinations, then $S(U)$ is convex and, since $L^2(\Gamma)$ and

$L^2(\Omega)$ are Hilbert spaces and, thus, strictly convex by Ex. 2.8(a), from Lem. 2.11(b), we know

$$J : S(U) \times L^2(\Gamma) \longrightarrow \mathbb{R}, \quad J(y,u) := \frac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u\|^2_{L^2(\Gamma)}, \tag{2.21}$$

is strictly convex for $\lambda > 0$. Then Lem. 2.11(c)(ii) (also Lem. 2.12) shows f is strictly convex as well.

If the control-to-state operator S does not preserve convex combinations, then f is, in general, *not* convex. While convexity properties of solution operators of nonlinear PDE are not easy to investigate, we will see an example of nonconvexity and nonuniqueness in a finite-dimensional setting in the next section in Ex. 3.6.

2.2 Relation Between Convexity and the Uniqueness of Extrema

As already mentioned, the uniqueness question with regard to solutions of optimal control problems is linked to convexity properties. This link is due to the following simple, but general, results. We start with a preparatory definition.

Definition 2.15. Let $(X, \|\cdot\|)$ be a normed vector space (see Def. 4.1), $A \subseteq X$, and $f : A \longrightarrow \mathbb{R}$.

(a) Given $x \in A$, f has a *(strict) global min* at x if, and only if, $f(x) \leq f(y)$ ($f(x) < f(y)$) for each $y \in A \setminus \{x\}$.

(b) Given $x \in X$, f has a *(strict) local min* at x if, and only if, there exists $r > 0$ such that $f(x) \leq f(y)$ ($f(x) < f(y)$) for each $y \in \{y \in A : \|y - x\| < r\} \setminus \{x\}$.

Theorem 2.16. *Let $(X, \|\cdot\|)$ be a normed vector space (see Def. 4.1), $C \subseteq X$, and $f : C \longrightarrow \mathbb{R}$. Assume C is a convex set, and f is a convex function.*

(a) *If f has a local min at $x_0 \in C$, then f has a global min at x_0.*

(b) *The set of mins of f is convex.*

Proof. (a): Suppose f has a local min at $x_0 \in C$, and consider an arbitrary $x \in C$, $x \neq x_0$. As x_0 is a local min, there is $r > 0$ such that $f(x_0) \leq f(y)$ for each $y \in C_r := \{y \in C : \|y - x_0\| < r\}$. Note that, for each $\alpha \in \mathbb{R}$,

$$x_0 + \alpha\,(x - x_0) = (1 - \alpha)\,x_0 + \alpha\,x. \tag{2.22}$$

Thus, due to the convexity of C, $x_0 + \alpha\,(x - x_0) \in C$ for each $\alpha \in [0,1]$. Moreover, for sufficiently small α, namely for each $\alpha \in R := \left]0, \min\{1, r/\|x_0 - x\|\}\right[$, one has $x_0 + \alpha\,(x - x_0) \in C_r$. As x_0 is a local min and f is convex, for each $\alpha \in R$, one obtains:

$$f(x_0) \leq f\big((1 - \alpha)\,x_0 + \alpha\,x\big) \leq (1 - \alpha)\,f(x_0) + \alpha\,f(x). \tag{2.23}$$

After subtracting $f(x_0)$ and dividing by $\alpha > 0$, (2.23) yields $f(x_0) \leq f(x)$, showing that x_0 is actually a global min as claimed.

(b): Let $x_0 \in C$ be a min of f. From (a), we already know that x_0 must be a global min. So, if $x \in C$ is any min of f, then it follows that $f(x) = f(x_0)$. If $\alpha \in [0,1]$, then the convexity of f implies that

$$f\big((1-\alpha)\,x_0 + \alpha\,x\big) \leq (1-\alpha)\,f(x_0) + \alpha\,f(x) = f(x_0). \tag{2.24}$$

As x_0 is a global min, (2.24) implies that $(1-\alpha)\,x_0 + \alpha\,x$ is also a global min for each $\alpha \in [0,1]$, showing that the set of mins of f is convex as claimed. ■

Theorem 2.17. *Let $(X, \|\cdot\|)$ be a normed vector space (see Def. 4.1), $C \subseteq X$, and $f : C \longrightarrow \mathbb{R}$. Assume C is a convex set, and f is a strictly convex function. If $x \in C$ is a local min of f, then this is the unique local min of f, and, moreover, it is strict.*

Proof. According to Th. 2.16, every local min of f is also a global min of f. Seeking a contradiction, assume there is $y \in C$, $y \neq x$, such that y is also a min of f. As x and y are both global mins, $f(x) = f(y)$ is implied. Define $z := \frac{1}{2}(x+y)$. Then $z \in C$ due to the convexity of C. Moreover, due to the strict convexity of f,

$$f(z) < \frac{1}{2}\big(f(x) + f(y)\big) = f(x) \tag{2.25}$$

in contradiction to x being a global min. Thus, x must be the unique min of f, also implying that the min must be strict. ■

3 Review: Finite-Dimensional Optimization

3.1 A Finite-Dimensional Optimal Control Problem

Consider the minimization of a real-valued J,

$$\min J(y,u),$$

where J is defined on a pair of finite-dimensional real vectors, i.e. $J : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, $(m,n) \in \mathbb{N}^2$. The function J to be optimized is called the *objective function* of the optimization problem.

Simple examples (set $m = n = 1$) show that, in this generality, the problem can have no solution (e.g. $J = J_1(y,u) := y + u$, $J = J_2(y,u) := e^{y+u}$, $J = J_3(y,u) := y - u$, or $J = J_4(y,u) := yu$), a unique solution (e.g. $J = J_5(y,u) := |y| + |u|$), finitely many solutions (e.g. $J = J_6(y,u) := (y^2-1)^2 + (u^2-1)^2$), or infinitely many solutions (e.g. $J = J_7(y,u) := c \in \mathbb{R}$, $J = J_8(y,u) := |\sin y| + |\sin u|$, or $J = J_9(y,u) := (y+1)^2(y-1)^2(u+1)^2(u-1)^2$).

Given linear maps $A : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $B : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ and a subset U_{ad} of $\mathbb{R}^m$ of so-called *admissible* vectors, one can consider the following modified finite-dimensional optimization problem (see [Trö05, (1.1)]):

$$\min J(y, u), \tag{3.1a}$$
$$Ay = Bu, \quad u \in U_{\text{ad}}. \tag{3.1b}$$

In (3.1), the minimization of the objective function J is subject to the *constraints* $Ay = Bu$ and $u \in U_{\text{ad}}$.

In spite of the constraints, letting $n = m = 1$, $A = B = \mathrm{Id}$, $U_{\text{ad}} = \mathbb{R}$ (or $U_{\text{ad}} =]-5, \infty[$), all the previous simple examples for J still work, showing that, in this generality, the problem can still have no solution, a unique solution, finitely many solutions, or infinitely many solutions. Moreover, for $U_{\text{ad}} = [0, \infty]$, we now also have a unique solution for $J = J_1$, $J = J_2$, $J = J_4$, $J = J_6$, or $J = J_9$. It is desirable to find conditions for J, A, B, and U_{ad}, such that one can prove the existence of a unique solution. We will soon encounter such conditions in Sec. 3.2.

Example 3.1. As a recurring standard example, we will consider the quadratic objective function

$$J : \mathbb{R}^n \times U_{\text{ad}} \longrightarrow \mathbb{R}, \quad J(y, u) := \frac{1}{2}|y - y_0|^2 + \frac{\lambda}{2}|u|^2,$$

where $U_{\text{ad}} \subseteq \mathbb{R}^m$, $y_0 \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^+$, and $|\cdot|$ denotes the Euclidian norm. Note that this can be seen as a finite-dimensional version of the objective functionals J considered in Sec. 1; also cf. Ex. 2.14.

—

A case of particular interest is the one where the map A in (3.1) is invertible. In that case, one can define the maps

$$S : U_{\text{ad}} \longrightarrow \mathbb{R}^n, \qquad S := A^{-1} B, \tag{3.2a}$$
$$f : U_{\text{ad}} \longrightarrow \mathbb{R}, \qquad f(u) := J(Su, u), \tag{3.2b}$$

reformulating (3.1) as

$$\min f(u), \tag{3.3a}$$
$$y = Su, \quad u \in U_{\text{ad}}. \tag{3.3b}$$

Thus, in the setting of (3.3), y is completely determined by u, such that u is the only remaining unknown of this so-called *reduced* optimization problem. One calls u the *control*, $y = Su$ the *state* corresponding to the control u, and S the *control-to-state* operator. In later sections, the constraint (3.1b) will be replaced by a partial differential equation (PDE) (also cf. Sec. 1). The map S will then play the role of the solution operator of this PDE.

Constraints provided in the form of an equality relation between the control u and the state y (such as $Ay = Bu$ in (3.1) and $y = Su$ in (3.3)) are called *equation constraints*.

In the present section, the equation constraints are given as a finite-dimensional linear system. Later (as in Sec. 1), they will take the form of PDE. Constraints that involve only the control (such as $u \in U_{\text{ad}}$ in (3.1) and (3.3)) are called *control constraints*. For the time being, we will restrict ourselves to the consideration of equation and control constraints. However, it can also make sense to consider constraints only involving the state (e.g. of the form $y \in Y_{\text{ad}}$). Not surprisingly, such constraints are then called *state constraints*.

3.2 Existence and Uniqueness

Definition 3.2. Within the setting of (3.2) and (3.3), a control $\bar{u} \in U_{\text{ad}}$ is called an *optimal control* for the problem (3.3), if, and only if, $f(\bar{u}) \leq f(u)$ for each $u \in U_{\text{ad}}$. Moreover, $\bar{y} = S\bar{u}$ is called the corresponding *optimal state* and the pair $(\bar{y}, \bar{u})$ is a *solution* to the (reduced) optimal control problem (3.3).

—

One can now easily proof a first existence theorem ([Trö05, Th. 1.1]):

Theorem 3.3. *Consider the reduced optimal control problem* (3.3)*, i.e.* (3.1) *with an invertible map A. If J is continuous on $\mathbb{R}^n \times U_{\text{ad}}$ and U_{ad} is nonvoid, closed, and bounded, then* (3.3) *has at least one optimal control as defined in Def.* 3.2.

Proof. The continuity of J together with the continuity of A^{-1} and B implies the continuity of f, where f is defined in (3.2b). As U_{ad} is assumed to be a closed and bounded subset of the finite-dimensional space $\mathbb{R}^m$, it is compact. Thus, f is a continuous map defined on a nonempty, compact set, which, in turn, implies that there is at least one $\bar{u} \in U_{\text{ad}}$, where f assumes its minimum (i.e. where it satisfies $f(\bar{u}) \leq f(u)$ for each $u \in U_{\text{ad}}$), completing the proof of the theorem. ■

Theorem 3.4. *Consider the reduced optimal control problem* (3.3)*, i.e.* (3.1) *with an invertible map A. If U_{ad} is nonvoid, convex, closed, and bounded; and J is continuous and strictly convex on $\mathbb{R}^n \times U_{\text{ad}}$, then* (3.3) *has a unique optimal control as defined in Def.* 3.2.

Proof. Let S and f be the functions defined in (3.2a) and (3.2b), respectively. Using Lem. 2.12 with $C := \mathbb{R}^n$ and $U := U_{\text{ad}}$, the strict convexity of J and the linearity of S yield the strict convexity of f. Then the existence of an optimal control is provided by Th. 3.3, while its uniqueness is obtained from Th. 2.17. ■

Example 3.5. Let us apply Th. 3.4 to the objective function J introduced in Ex. 3.1, i.e. to

$$J : \mathbb{R}^n \times U_{\text{ad}} \longrightarrow \mathbb{R}, \quad J(y,u) := \frac{1}{2}|y - y_0|^2 + \frac{\lambda}{2}|u|^2,$$

$y_0 \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^+$. If $U_{\text{ad}} \subseteq \mathbb{R}^m$ is convex, then we can apply Lem. 2.11(b) with $U := U_{\text{ad}}$, $X := \mathbb{R}^m$, $Y := \mathbb{R}^n$, showing that J is strictly convex (where we have also used that $\mathbb{R}^m$

and $\mathbb{R}^n$ with the Euclidean norm constitute Hilbert spaces, which are strictly convex according to Ex. 2.8(a)). As J is clearly continuous, if $U_{\text{ad}} \subseteq \mathbb{R}^m$ is nonvoid, convex, closed, and bounded (e.g. a compact interval or ball) and S and f are the functions defined in (3.2a) and (3.2b), respectively, then Th. 3.4 yields the existence of a unique optimal control for (3.3) as defined in Def. 3.2.

Example 3.6. The goal of the present example is to provide a counterexample to uniqueness in the presence of nonlinearities. More precisely, we will see that a nonlinear S combined with the strictly convex J of Ex. 3.1 can lead to a nonconvex f, which, in turn, can lead to multiple (local and also global) minima. This can already be seen in a one-dimensional setting. For the purposes of this example, we will now temporarily leave the linear setting introduced in Sec. 3.1.

Let $m := n := 1$, $y_0 := 0$, and $U_{\text{ad}} := [-2, 2]$. Then J from Ex. 3.1 becomes

$$J : \mathbb{R} \times [-2, 2] \longrightarrow \mathbb{R}, \quad J(y, u) := \frac{1}{2}\, y^2 + \frac{\lambda}{2}\, u^2, \tag{3.4}$$

also recalling that $\lambda > 0$. Moreover, define

$$f : [-2, 2] \longrightarrow \mathbb{R}, \quad f(u) := (u - 1)^2 (u + 1)^2 + 3\,\lambda. \tag{3.5}$$

Note that $f(u) \geq 3\,\lambda$ for all $u \in [-2, 2]$ and $-\lambda\, u^2 \geq -4\,\lambda$ for all $u \in [-2, 2]$. Thus, $2\, f(u) - \lambda\, u^2 \geq 6\,\lambda - 4\,\lambda = 2\,\lambda > 0$ for all $u \in [-2, 2]$, and one can define

$$S : [-2, 2] \longrightarrow \mathbb{R}, \quad S(u) := \sqrt{2\, f(u) - \lambda\, u^2}. \tag{3.6}$$

One computes

$$J(Su, u) = \frac{1}{2}\big(2\, f(u) - \lambda\, u^2\big) + \frac{\lambda}{2}\, u^2 = f(u), \tag{3.7}$$

showing that J, S, and f satisfy (3.2b) with $\mathbb{R}^n$ replaced by $[-2, 2]$.

Moreover, f is continuous, nonconvex, having exactly two (local and global) minima, namely at $u = -1$ and $u = 1$.

3.3 First Order Necessary Optimality Conditions, Variational Inequality

In one-dimensional calculus, when studying differentiable functions $f : \mathbb{R} \longrightarrow \mathbb{R}$, one learns that a vanishing first derivative $f'(\bar{u}) = 0$ is a *necessary condition* for f to have a (local, in particular, global) extremum (max or min) at $\bar{u}$. One also learns that simple examples (e.g. $f(u) = u^3$ at $\bar{u} = 0$) show that $f'(\bar{u}) = 0$ is not sufficient for f to have a (local, in particular, global) extremum at $\bar{u}$.

Similar necessary optimality conditions that are *first order* in the sense that they involve only first derivatives can also be formulated in multiple finite dimensions (as will be recalled in the present section) and even in infinite-dimensional cases such as the optimal control of PDE as we will see subsequently.

Notation 3.7. If A is a matrix, then let $A^\top$ denote the transpose of A.

Notation 3.8. Given a function $f : \mathbb{R}^m \longrightarrow \mathbb{R}$, $(x_1, \ldots, x_m) \mapsto f(x_1, \ldots, x_m)$, the following notation is used:

$$\begin{aligned} \text{Partial Derivatives:} \quad & \partial_1 f, \ldots, \partial_m f, \quad \text{or} \quad \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_m}. \\ \text{Derivative:} \quad & f' := (\partial_1 f, \ldots, \partial_m f) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_m} \right). \\ \text{Gradient:} \quad & \nabla f := (f')^\top. \end{aligned}$$

Notation 3.9. Given vectors $(u, v) \in \mathbb{R}^m \times \mathbb{R}^m$, $m \in \mathbb{N}$, the scalar product of u and v is denoted by

$$\langle u, v \rangle_{\mathbb{R}^m} := u \bullet v := \sum_{i=1}^{m} u_i \, v_i. \tag{3.8}$$

As in [Trö05], for the sake of readability, both forms of denoting the scalar product introduced in (3.8) will be subsequently used, depending on the situation.

—

If the objective function possesses directional derivatives, then they can be used to formulate necessary conditions for an optimal control $\bar{u}$:

Theorem 3.10. *Let $U_{\text{ad}} \subseteq \mathbb{R}^m$, and assume that $\bar{u} \in U_{\text{ad}}$ minimizes the function $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ (not necessarily given by* (3.2b)*), i.e.*

$$f(\bar{u}) \leq f(u) \quad \textit{for each } u \in U_{\text{ad}}. \tag{3.9}$$

Consider $u \in U_{\text{ad}}$. If $\bar{u} + t\,(u - \bar{u}) \in U_{\text{ad}}$ for each sufficiently small $t > 0$, and, moreover, the directional derivative

$$\delta f(\bar{u}, u - \bar{u}) := \lim_{t \downarrow 0} \frac{1}{t} \Big(f\big(\bar{u} + t\,(u - \bar{u})\big) - f(\bar{u}) \Big) \tag{3.10}$$

exists, then $\bar{u}$ satisfies the variational inequality

$$\delta f(\bar{u}, u - \bar{u}) \;\geq\; 0. \tag{3.11}$$

Proof. Since $\bar{u} + t\,(u - \bar{u}) \in U_{\text{ad}}$ for each sufficiently small $t > 0$, there exists $\varepsilon > 0$ such that

$$\bar{u} + t(u - \bar{u}) = (1 - t)\bar{u} + tu \in U_{\text{ad}}, \quad \text{for each} \quad t \in]0, \varepsilon]. \tag{3.12}$$

By hypothesis, $\bar{u}$ satisfies (3.9), implying, for each $t \in]0, \varepsilon]$:

$$\frac{1}{t} \Big(f\big(\bar{u} + t(u - \bar{u})\big) - f(\bar{u}) \Big) \overset{(3.9)}{\geq} 0. \tag{3.13}$$

Thus, taking the limit for $t \to 0$, (3.13) implies (3.11). ■

In Th. 3.10, we avoided imposing any a priori conditions on the set $U_{\rm ad}$ – in consequence, if $U_{\rm ad}$ is very irregular, there might be few (or no) $u \in U_{\rm ad}$ such that the directional derivative $\delta f(\bar{u}, u - \bar{u})$ exists. We would now like to formulate a corollary for the case that f is *differentiable* on $U_{\rm ad}$. A difficulty arises from the fact that the standard definition for differentiability requires $U_{\rm ad}$ to be open. On the other hand, in many interesting cases, the optimal point $\bar{u}$ lies in the boundary of $U_{\rm ad}$, so it is desirable to allow sets $U_{\rm ad}$ that contain boundary points. This is the reason for dealing with differentiable extensions of f in the sense of the following Def. 3.11.

Definition 3.11. Let $U_{\rm ad} \subseteq O \subseteq \mathbb{R}^m$, where O is open. A function $F : O \longrightarrow \mathbb{R}$ is called a *differentiable extension* of a function $f : U_{\rm ad} \longrightarrow \mathbb{R}$ if, and only if, F is differentiable and $F\restriction_{U_{\rm ad}} = f$.

Corollary 3.12. *Let $U_{\rm ad} \subseteq O \subseteq \mathbb{R}^m$, where $U_{\rm ad}$ is convex and O is open. If $F : O \longrightarrow \mathbb{R}$ is a differentiable extension of $f : U_{\rm ad} \longrightarrow \mathbb{R}$ (not necessarily given by* (3.2b)*), then each minimizer $\bar{u} \in U_{\rm ad}$ of f satisfies*

$$F'(\bar{u})\,(u - \bar{u}) \;\geq\; 0 \quad \textit{for each} \quad u \in U_{\rm ad}. \tag{3.14}$$

Proof. Let $u \in U_{\rm ad}$ be arbitrary. The differentiability of F implies that the directional derivative $\delta F(\bar{u}, u - \bar{u})$ exists and that $\delta F(\bar{u}, u - \bar{u}) = F'(\bar{u})(u - \bar{u})$. On the other hand, the convexity of $U_{\rm ad}$ yields that $\bar{u} + t\,(u - \bar{u}) \in U_{\rm ad}$ for each $t \in [0, 1]$. Therefore, using that F is an extension of f, yields that $\delta f(\bar{u}, u - \bar{u})$ exists and equals $\delta F(\bar{u}, u - \bar{u}) = F'(\bar{u})(u - \bar{u})$. The assertion (3.14) then follows from (3.11) in Th. 3.10. ■

A further easy conclusion from Cor. 3.12 is that $f'(\bar{u}) = 0$ if $\bar{u}$ lies in the *interior* of $U_{\rm ad}$ (Cor. 3.13). In particular, $f'(\bar{u}) = 0$ if $U_{\rm ad} = \mathbb{R}^m$ (no control constraints) or, more generally, if $U_{\rm ad}$ is open. The latter is often assumed when treating extrema in advanced calculus text books. In general, the variational inequality can be strict as will be seen in Ex. 3.14.

Corollary 3.13. *Let $U_{\rm ad} \subseteq \mathbb{R}^m$, and assume that $\bar{u} \in U_{\rm ad}$ lies in the interior of $U_{\rm ad}$ and minimizes the function $f : U_{\rm ad} \longrightarrow \mathbb{R}$ (not necessarily given by* (3.2b)*), assumed differentiable in the interior of $U_{\rm ad}$. Then $f'(\bar{u}) = 0$. Special cases include $U_{\rm ad} = \mathbb{R}^m$ (no control constraints) and any other case, where $U_{\rm ad}$ is open.*

Proof. If $\bar{u}$ lies in the interior of $U_{\rm ad}$, then there is a (convex) open ball B with center $\bar{u}$ such that $B \subseteq U_{\rm ad}$. Then Cor. 3.12 yields that $f'(\bar{u})\,(u - \bar{u}) \;\geq\; 0$ for each $u \in B$. Let e_i denote the i-th standard unit vector of $\mathbb{R}^m$. If $\epsilon > 0$ is sufficiently small, then $\bar{u} \pm \epsilon e_i \in B$ for each $i \in \{1, \ldots, m\}$, implying $f'(\bar{u})\,(\bar{u} \pm \epsilon e_i - \bar{u}) = f'(\bar{u})\,(\pm\epsilon e_i) \geq 0$. Thus, $f'(\bar{u}) = 0$ as claimed. ■

Example 3.14. Let $m = 1$, $U_{\rm ad} = [0, 1]$, $f : [0, 1] \longrightarrow \mathbb{R}$, $f(u) = u$. Then f is minimal at $\bar{u} = 0$, $f'(\bar{u}) = (1) = \mathrm{Id}$, and $f'(\bar{u})(u - \bar{u}) = u > 0$ for each $u \in]0, 1]$, showing that the variational inequality can be strict.

Example 3.15. As already mentioned in the first paragraph of this section, the example $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(u) = u^3$, $\bar{u} = 0$, shows that, in general, the variational inequality is *not* sufficient for $\bar{u}$ to be a min of f ($f'(0) = (0)$, but 0 is not a min of f). As another example, consider $g : [0, 2\pi] \longrightarrow \mathbb{R}$, $g(u) = \sin(u)$. Then, $g'(0) = (1) = \mathrm{Id}$, and $g'(0)(u - 0) = u > 0$ for each $u \in]0, 2\pi]$, but the unique global min of g is at $u = 3\,\pi/2$.

Remark 3.16. Even though the variational inequality is not sufficient for $\bar{u}$ to be a min of f, we will later see in the more general context of minimization and directional derivatives in normed vector spaces (Th. 6.62), that the variational inequality *is* also sufficient, if U_{ad} is convex, f is convex, and $\delta f(\bar{u}, u - \bar{u})$ exists for each $u \in U_{\mathrm{ad}}$.

Theorem 3.17. *In the setting of the reduced optimization problem* (3.3)*, suppose that* $U_{\mathrm{ad}} \subseteq O \subseteq \mathbb{R}^m$*, where* U_{ad} *is convex,* O *is open, and* $J : \mathbb{R}^n \times O \longrightarrow \mathbb{R}$, $(m, n) \in \mathbb{N}^2$*, is of class* C^1*, that means all its partial derivatives* $\partial_1 J, \ldots, \partial_n J, \partial_{n+1} J, \ldots, \partial_{n+m} J$ *exist and are continuous. If* $\bar{u} \in U_{\mathrm{ad}}$ *is an optimal control for* (3.3) *in the sense of Def.* 3.2*, then it satisfies* (3.14) *with* F' *replaced by* f'*, and* f *being defined according to* (3.2b)*, except on* O *instead of* U_{ad}*. Using the chain rule, one can compute* f' *explicitly in terms of* J*,* A*, and* B*, namely, recalling* $S = A^{-1} B$ *from* (3.2)*,*

$$f'(u) = B^\top (A^\top)^{-1} \nabla_y J(Su, u) + \nabla_u J(Su, u) \quad \textit{for each} \quad u \in \mathbb{R}^m, \tag{3.15}$$

using the abbreviations $\nabla_y J = (\partial_1 J, \ldots, \partial_n J)^\top$, $\nabla_u J = (\partial_{n+1} J, \ldots, \partial_{n+m} J)^\top$*. Thus, letting* $\bar{y} := S\bar{u}$*,* (3.11) *can be rewritten in the lengthy form (cf.* [Trö05, (1.6)]*)*

$$\big\langle B^\top (A^\top)^{-1} \nabla_y J(\bar{y}, \bar{u}) + \nabla_u J(\bar{y}, \bar{u}),\, u - \bar{u} \big\rangle_{\mathbb{R}^m} \geq 0 \quad \textit{for each} \quad u \in U_{\mathrm{ad}}. \tag{3.16}$$

Please recall that the vectors $\bar{u}$ *and* u *in* (3.16) *are interpreted as column vectors.*

Proof. According to the definition of S and f in (3.2), one has $S = A^{-1} B$ and $f(u) = J(Su, u)$ for each $u \in O$. Introducing the auxiliary function

$$a : O \longrightarrow \mathbb{R}^n \times \mathbb{R}^m, \quad a(u) := (Su, u), \tag{3.17}$$

it is $f = J \circ a$. As a is linear, $a = a'$. Applying the chain rule yields, for each $u \in O$,

$$f'(u) = J'\big(a(u)\big)\, a'(u) = J'\big(a(u)\big)\, a \tag{3.18}$$

and, thus, for each $(u, v) \in O \times O$, abbreviating $y := Su$,

$$\begin{aligned} f'(u)(v) &= J'(y, u)\,\big(Sv, v\big) \\ &= \big\langle \nabla_y J(y, u),\, A^{-1} B v \big\rangle_{\mathbb{R}^n} + \big\langle \nabla_u J(y, u),\, v \big\rangle_{\mathbb{R}^m} \\ &= \big\langle B^\top (A^\top)^{-1} \nabla_y J(y, u),\, v \big\rangle_{\mathbb{R}^m} + \big\langle \nabla_u J(y, u),\, v \big\rangle_{\mathbb{R}^m}, \end{aligned} \tag{3.19}$$

thereby proving (3.15). Combining (3.15) and (3.14) (with f' instead of F') proves (3.16). ∎

Example 3.18. Let us compute the left-hand side of the variational inequality (3.16) for the function J from Ex. 3.1, i.e. for

$$J : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}, \quad J(y,u) := \frac{1}{2}|y - y_0|^2 + \frac{\lambda}{2}\,|u|^2.$$

Recalling $S = A^{-1}B$, since $\nabla_y J(y,u) = (y - y_0)$ and $\nabla_u J(y,u) = \lambda\, u$, (3.15) implies that

$$f'(u) = B^\top (A^\top)^{-1} (A^{-1}Bu - y_0) + \lambda\, u \quad \text{for each} \quad u \in \mathbb{R}^m. \tag{3.20}$$

Moreover, with $\bar{y} = S\bar{u} = A^{-1}B\bar{u}$, (3.16) becomes

$$\begin{aligned}
&\left\langle B^\top (A^\top)^{-1}\, \nabla_y J(\bar{y},\bar{u}) + \nabla_u J(\bar{y},\bar{u})\,,\, u - \bar{u} \right\rangle_{\mathbb{R}^m} \\
&= \left\langle B^\top (A^\top)^{-1} (\bar{y} - y_0) + \lambda\, \bar{u}\,,\, u - \bar{u} \right\rangle_{\mathbb{R}^m} \\
&= \left\langle B^\top (A^\top)^{-1} (A^{-1}B\bar{u} - y_0) + \lambda\, \bar{u}\,,\, u - \bar{u} \right\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for each} \quad u \in U_{\mathrm{ad}} \subseteq \mathbb{R}^m.
\end{aligned} \tag{3.21}$$

In Ex. 3.20, we will see that the messy-looking condition (3.21) becomes more readable after introducing the so-called adjoint state. Moreover, we will also see that it can be reduced to a linear system in the case of $U_{\mathrm{ad}} = \mathbb{R}^m$ (no control constraints).

3.4 Adjoint Equation, Adjoint State

If the dimension n is large, then, in general, the inverse matrix $(A^\top)^{-1}$ occurring in (3.16) is not at hand (i.e. not easily computable), and it is useful to introduce the quantity

$$\bar{p} := \bar{p}(\bar{y},\bar{u}) := (A^\top)^{-1}\, \nabla_y J(\bar{y},\bar{u}) \tag{3.22}$$

as an additional unknown of the considered problem. The quantity defined in (3.22) is called the *adjoint state* corresponding to $(\bar{y},\bar{u})$. Given $(\bar{y},\bar{u})$, the adjoint state $\bar{p}$ is determined by the equation

$$A^\top \bar{p} = \nabla_y J(\bar{y},\bar{u}), \tag{3.23}$$

which is called the *adjoint equation* of the control problem (3.3).

Corollary 3.19. *As in Th.* 3.17*, consider the setting of the problem* (3.3)*, and suppose that* $U_{\mathrm{ad}} \subseteq O \subseteq \mathbb{R}^m$*, where* U_{ad} *is convex,* O *is open, and* $J : \mathbb{R}^n \times O \longrightarrow \mathbb{R}$*,* $(m,n) \in \mathbb{N}^2$*, is of class* C^1*. If* $\bar{u} \in U_{\mathrm{ad}}$ *is an optimal control for* (3.3) *in the sense of Def.* 3.2 *with corresponding state* $\bar{y} = S\bar{u}$ *and adjoint state* $\bar{p} = (A^\top)^{-1}\, \nabla_y J(\bar{y},\bar{u})$*, then* $(\bar{y},\bar{u},\bar{p})$ *satisfies the following system (cf.* [Trö05, p. 12]*)*

$$A\bar{y} = B\bar{u}, \quad \bar{u} \in U_{\mathrm{ad}}, \tag{3.24a}$$

$$A^\top \bar{p} = \nabla_y J(\bar{y},\bar{u}), \tag{3.24b}$$

$$\left\langle B^\top \bar{p} + \nabla_u J(\bar{y},\bar{u})\,,\, u - \bar{u} \right\rangle_{\mathbb{R}^m} \geq 0 \quad \textit{for each} \quad u \in U_{\mathrm{ad}}, \tag{3.24c}$$

called the system of optimality *for the optimal control problem* (3.3)*.* ■

Example 3.20. As promised at the end of Ex. 3.18, we continue the investigation of the function J from Ex. 3.1 by formulating the resulting system of optimality. According to (3.24):

$$A\bar{y} = B\bar{u}, \quad \bar{u} \in U_{\text{ad}}, \tag{3.25a}$$

$$A^\top \bar{p} = \bar{y} - y_0, \tag{3.25b}$$

$$\left\langle B^\top \bar{p} + \lambda\, \bar{u}\,,\, u - \bar{u} \right\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for each} \quad u \in U_{\text{ad}}. \tag{3.25c}$$

In the case of no control constraints, i.e. $U_{\text{ad}} = \mathbb{R}^m$, one has $f'(\bar{u}) = 0$ by Cor. 3.13. Thus, according to (3.20), $U_{\text{ad}} = \mathbb{R}^m$ implies that (3.25c) can be replaced with

$$B^\top (A^\top)^{-1} (A^{-1} B\, \bar{u} - y_0) + \lambda\, \bar{u} \overset{(3.25a),\,(3.25b)}{=} B^\top \bar{p} + \lambda\, \bar{u} = 0. \tag{3.26}$$

Using the assumption $\lambda > 0$ stated in Ex. 3.1, (3.26) yields

$$\bar{u} = -\frac{B^\top \bar{p}}{\lambda}. \tag{3.27a}$$

By means of (3.25b), we also have an explicit equation for $\bar{y}$, namely

$$\bar{y} = A^\top \bar{p} + y_0. \tag{3.27b}$$

Plugging (3.27) into (3.25a) leads to the following linear system for $\bar{p}$:

$$A(A^\top \bar{p} + y_0) = -\frac{1}{\lambda}\, B\, B^\top \bar{p}, \tag{3.28}$$

or, rearranged,

$$\left(AA^\top + \frac{1}{\lambda}\, B\, B^\top \right) \bar{p} = -A\, y_0. \tag{3.29}$$

Thus, in this case, one can solve the system of optimality by determining the adjoint state $\bar{p}$ from the linear system (3.29). The optimal control $\bar{u}$ and the optimal state $\bar{y}$ are then given by (3.27).

3.5 Lagrange Technique and Karush-Kuhn-Tucker Conditions

3.5.1 Lagrange Function

Introducing an auxiliary function depending on three variables (y, u, p), a so-called *Lagrange function*, one can rewrite the conditions of the system of optimality (3.24) (except for the control constraints $\bar{u} \in U_{\text{ad}}$, at least for the time being) in terms of the gradients of the Lagrange function with respect to the different variables (y, u, p). The Lagrange function will be defined in the statement of the following Cor. 3.21 in (3.30).

Corollary 3.21. *As in Th.* 3.17*, consider the setting of the problem* (3.3)*, and suppose that* $U_{\text{ad}} \subseteq O \subseteq \mathbb{R}^m$*, where* U_{ad} *is convex,* O *is open, and* $J : \mathbb{R}^n \times O \longrightarrow \mathbb{R}$*,* $(m, n) \in \mathbb{N}^2$*, is of class* C^1*. If* $\bar{u} \in U_{\text{ad}}$ *is an optimal control for* (3.3) *in the sense of Def.* 3.2 *with*

corresponding state $\bar{y} = S\bar{u}$ *and adjoint state* $\bar{p} = (A^\top)^{-1} \nabla_y J(\bar{y}, \bar{u})$*, then, introducing the Lagrange function*

$$L : \mathbb{R}^{2n+m} \longrightarrow \mathbb{R}, \quad L(y, u, p) := J(y, u) - \langle Ay - Bu\,,\, p\rangle_{\mathbb{R}^n}, \tag{3.30}$$

$(\bar{y}, \bar{u}, \bar{p})$ *satisfies*

$$\nabla_p L(\bar{y}, \bar{u}, \bar{p}) = 0, \quad \bar{u} \in U_{\text{ad}}, \tag{3.31a}$$
$$\nabla_y L(\bar{y}, \bar{u}, \bar{p}) = 0, \tag{3.31b}$$
$$\big\langle \nabla_u L(\bar{y}, \bar{u}, \bar{p})\,,\, u - \bar{u}\big\rangle_{\mathbb{R}^m} \geq 0 \quad \textit{for each} \quad u \in U_{\text{ad}}. \tag{3.31c}$$

Proof. From Cor. 3.19, we know that $(\bar{y}, \bar{u}, \bar{p})$ satisfies the optimality system (3.24). Moreover, (3.30) implies

$$\nabla_p L(y, u, p) = -Ay + Bu, \tag{3.32a}$$
$$\nabla_y L(y, u, p) = \nabla_y J(y, u) - A^\top p, \tag{3.32b}$$
$$\nabla_u L(y, u, p) = \nabla_u J(y, u) + B^\top p, \tag{3.32c}$$

showing that (3.31a) is the same as (3.24a), (3.31b) is the same as (3.24b), and (3.31c) is the same as (3.24c). ∎

In the present context, the adjoint state $\bar{p}$ is also called *Lagrange multiplier*.

3.5.2 Box Constraints and Karush-Kuhn-Tucker Optimality Conditions

For a special class of control constraints, so-called box constraints defined below, we will deduce another system of first-order necessary optimality conditions (3.37), called Karush-Kuhn-Tucker system. The Karush-Kuhn-Tucker system in a certain sense completes the task started with the formulation of (3.31) in the previous section. Now the control constraints are also formulated in terms of partial gradients of an extended Lagrange function defined in (3.35). The Karush-Kuhn-Tucker system is also structurally simpler than (3.31): Even though, at first glance, the Karush-Kuhn-Tucker system consists of more conditions than (3.31), that is actually deceiving. While the Karush-Kuhn-Tucker system only consists of a finite number of equations and inequalities, the variational inequality (3.31c) actually consists of one condition *for each* $u \in U_{\text{ad}}$, i.e., typically, uncountably many conditions.

Before proceeding to the definition of box constraints, we recall some notation:

Notation 3.22. For $(u, v) \in \mathbb{R}^m \times \mathbb{R}^m$, $m \in \mathbb{N}$, the inequality $u \leq v$ is meant componentwise, i.e. $u \leq v$ if, and only if, $u_i \leq v_i$ for each $i \in \{1, \ldots, m\}$. Analogously, one defines $u \geq v$, $u < v$, and $u > v$.

One speaks of *box constraints* for the control if the control constraints are prescribed via upper and lower bounds. This type of control constraint occurs in numerous applications, for example, in each of the motivating examples of Sec. 1. In the present

finite-dimensional setting, box constraints for the control mean that the admissible set U_{ad} has the form

$$U_{\text{ad}} = \{u \in \mathbb{R}^m : u_a \le u \le u_b\}, \quad \text{where} \quad (u_a, u_b) \in \mathbb{R}^m \times \mathbb{R}^m, \quad u_a \le u_b. \tag{3.33}$$

The bounds u_a and u_b are considered as given and fixed.

Theorem 3.23. *As in Cor.* 3.19*, consider the setting of the problem* (3.3)*, now with the additional assumption of box constraints for the control, i.e.* U_{ad} *is assumed to satisfy* (3.33)*. Still suppose that* $U_{\text{ad}} \subseteq O \subseteq \mathbb{R}^m$*, where* O *is open and* $J : \mathbb{R}^n \times O \longrightarrow \mathbb{R}$*,* $(m, n) \in \mathbb{N}^2$*, is of class* C^1*. If* $\bar{u} \in U_{\text{ad}}$ *is an optimal control for* (3.3) *in the sense of Def.* 3.2 *with corresponding state* $\bar{y} = S\bar{u}$ *and adjoint state* $\bar{p} = (A^\top)^{-1} \nabla_y J(\bar{y}, \bar{u})$*, then* $\bar{u} = (\bar{u}_i)_{i \in \{1,\dots,m\}}$ *satisfies, for each* $i \in \{1, \dots, m\}$*,*

$$\bar{u}_i = \begin{cases} u_{b,i} & \text{where } \big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i < 0, \\ u_{a,i} & \text{where } \big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i > 0. \end{cases} \tag{3.34}$$

Moreover, introducing the extended Lagrange function

$$\begin{aligned} \mathcal{L} &: \mathbb{R}^{2n+3m} \longrightarrow \mathbb{R}, \\ \mathcal{L}(y, u, p, a, b) &:= L(y, u, p) + \langle u_a - u \,,\, a\rangle_{\mathbb{R}^m} + \langle u - u_b \,,\, b\rangle_{\mathbb{R}^m} \\ &= J(y, u) - \langle Ay - Bu \,,\, p\rangle_{\mathbb{R}^n} + \langle u_a - u \,,\, a\rangle_{\mathbb{R}^m} + \langle u - u_b \,,\, b\rangle_{\mathbb{R}^m}, \end{aligned} \tag{3.35}$$

and letting

$$\mu_a := \max\big\{0,\, B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big\}, \quad \mu_b := -\min\big\{0,\, B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big\}, \tag{3.36}$$

the 5*-tuple* $(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b)$ *satisfies*

$$\nabla_p \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0, \tag{3.37a}$$
$$\nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0, \tag{3.37b}$$
$$\nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0, \tag{3.37c}$$
$$\nabla_a \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \le 0, \tag{3.37d}$$
$$\nabla_b \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \le 0, \tag{3.37e}$$
$$\mu_a \ge 0, \quad \mu_b \ge 0, \tag{3.37f}$$
$$(u_{a,i} - \bar{u}_i)\, \mu_{a,i} = (\bar{u}_i - u_{b,i})\, \mu_{b,i} = 0 \quad \text{for each } i \in \{1, \dots, m\}. \tag{3.37g}$$

The system (3.37) *is known as the* Karush-Kuhn-Tucker optimality system*; conditions* (3.37d) – (3.37g) *are called* complementary slackness conditions*.*

Proof. Note that (3.33) implies that U_{ad} is convex. Thus, all hypotheses of Corollaries 3.19 and 3.21 are satisfied, and we know that (3.24) and (3.31) hold. We first show (3.34), followed by the verification of (3.37).

For the convenience of the reader, (3.24c) is restated:

$$\big\langle B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, u - \bar{u}\big\rangle_{\mathbb{R}^m} \ge 0 \quad \text{for each} \quad u \in U_{\text{ad}}. \tag{3.38}$$

Slightly rearranging (3.38) yields

$$\left\langle B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, \bar{u} \right\rangle_{\mathbb{R}^m} \leq \left\langle B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, u \right\rangle_{\mathbb{R}^m} \quad \text{for each} \quad u \in U_{\text{ad}}, \tag{3.39}$$

showing that $\bar{u}$ is a solution to the minimization problem

$$\min_{u \in U_{\text{ad}}} \left\langle B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, u \right\rangle_{\mathbb{R}^m} = \min_{u \in U_{\text{ad}}} \sum_{i=1}^{m} \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i u_i. \tag{3.40}$$

Due to the special form of U_{ad}, the components u_i can be varied completely independently, such that the sum in (3.40) attains its min if, and only if, each summand is minimal. Thus, for each $i \in \{1, \ldots, m\}$,

$$\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i \bar{u}_i = \min_{u_{a,i} \leq u_i \leq u_{b,i}} \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i u_i. \tag{3.41}$$

Now, (3.34) is a direct consequence of (3.41).

As for (3.37), everything except (3.37g) is quite obvious: According to the definition of $\mathcal{L}$ in (3.35), it is $\nabla_p \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = \nabla_p L(\bar{y}, \bar{u}, \bar{p})$ and $\nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = \nabla_y L(\bar{y}, \bar{u}, \bar{p})$ such that (3.37a) is the same as the equation in (3.31a) and (3.24a), and (3.37b) is the same as (3.31b) and (3.24b). Next,

$$\nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = \nabla_u J(\bar{y}, \bar{u}) + B^\top \bar{p} - \mu_a + \mu_b,$$

that means (3.37c) holds because of the way μ_a and μ_b were defined in (3.36). As (3.35) implies $\nabla_a \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = u_a - \bar{u}$ and $\nabla_b \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = \bar{u} - u_b$, (3.37d) and (3.37e) are merely a restatement of the hypothesis $\bar{u} \in U_{\text{ad}}$. The validity of (3.37f) is immediate from (3.36). Finally, (3.37g) follows from (3.34): Let $i \in \{1, \ldots, m\}$. According to (3.36), $\mu_{a,i} \geq 0$. If $\mu_{a,i} = 0$, then $(u_{a,i} - \bar{u}_i)\, \mu_{a,i} = 0$. If $\mu_{a,i} > 0$, then $\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i > 0$ by (3.36), i.e. $\bar{u}_i = u_{a,i}$ by (3.34), again implying $(u_{a,i} - \bar{u}_i)\, \mu_{a,i} = 0$. Analogously, according to (3.36), $\mu_{b,i} \geq 0$. if $\mu_{b,i} = 0$, then $(\bar{u}_i - u_{b,i})\, \mu_{b,i} = 0$. If $\mu_{b,i} > 0$, then $\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i < 0$ by (3.36), i.e. $\bar{u}_i = u_{b,i}$ by (3.34), again implying $(\bar{u}_i - u_{b,i})\, \mu_{b,i} = 0$, thereby concluding the proof of (3.37g) as well as the proof of the theorem. ■

Analogous to $\bar{p}$, the vectors μ_a and μ_b occurring in (3.37) are also referred to as *Lagrange multipliers*. Note that (3.34) does not yield any information on the components $\bar{u}_i$ where $\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \right)_i = 0$.

3.6 A Preview of Optimal Control of PDE

In many respects, and that is the reason for the somewhat detailed review of finite-dimensional optimal control problems in this section, the optimal control theory of PDE can be developed analogously to the finite-dimensional situation. Instead of a finite-dimensional equation, $Ay = Bu$ will represent a PDE, typically with A corresponding to some differential operator and B corresponding to some coefficient or embedding

operator. Guaranteeing the invertibility of A will usually mean restricting its domain to suitable function spaces, e.g. to sets of functions satisfying suitable boundary conditions. Then $S = A^{-1}B$ can be interpreted as the solution operator of the PDE, also called control-to-state operator. The optimality conditions can then be formulated in a form similar to the finite-dimensional case.

4 Review: Functional Analysis Tools

4.1 Normed Vector Spaces

Definition 4.1. Let X be a real vector space. A function $\|\cdot\| : X \longrightarrow \mathbb{R}_0^+$, $x \mapsto \|x\|$, is called a *norm* on X if, and only if, the following conditions (i) – (iii) are satisfied:

(i) For each $x \in X$, one has $\|x\| = 0$ if, and only if, $x = 0$.

(ii) $\|x + y\| \leq \|x\| + \|y\|$ for each $(x, y) \in X^2$.

(iii) $\|\lambda\, x\| = |\lambda|\, \|x\|$ for each $x \in X$, $\lambda \in \mathbb{R}$.

If $\|\cdot\|$ is a norm on X, then $(X, \|\cdot\|)$ is called a *normed vector space*. Frequently, the norm on X is understood and X itself is referred to as a normed vector space.

Remark 4.2. If $\|\cdot\|$ satisfies (ii) and (iii) in Def. 4.1, but not necessarily (i), then $\|\cdot\|$ is called a seminorm on X and $(X, \|\cdot\|)$ is called a seminormed vector space. Seminormed vector spaces where (i) is violated have the significant disadvantage that the corresponding topology does not satisfy the Hausdorff separation axiom.

Definition 4.3. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a real vector space X are called *equivalent* if, and only if, there exist positive constants $(m, M) \in \mathbb{R}^2$, $0 < m \leq M$, such that

$$m\, \|x\|_1 \leq \|x\|_2 \leq M\, \|x\|_1 \quad \text{for each} \quad x \in X. \tag{4.1}$$

Definition 4.4. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in a normed vector space $(X, \|\cdot\|)$.

(a) The sequence is called *convergent* (in X), if, and only if, there exists $x \in X$ such that

$$\lim_{n\to\infty} \|x_n - x\| = 0.$$

If such an $x \in X$ exists, then it is called the *limit* of the sequence. This notion of convergence is sometimes also called *strong* convergence and the limit the *strong* limit of the sequence. This is typically done to avoid confusion with the notion of weak convergence and weak limits that will be introduced in Def. 4.37 below.

(b) The sequence is called a *Cauchy sequence* if, and only if, for each $\epsilon > 0$, there exists some $n_0 \in \mathbb{N}$ such that, for each $(m, n) \in \mathbb{N}^2$,

$$\|x_n - x_m\| \leq \epsilon \quad \text{whenever} \quad n \geq n_0 \text{ and } m \geq n_0.$$

Definition 4.5. A normed vector space $(X, \|\cdot\|)$ is called *complete* or a *Banach space*, if, and only if, every Cauchy sequence in X is convergent in X (i.e. it has a limit $x \in X$).

Notation 4.6. Let $(X, \|\cdot\|)$ be a normed vector space, $x_0 \in X$. Then, for each $r \in \mathbb{R}^+$, $B_r(x_0) := \{x \in X : \|x - x_0\| < r\}$, $\overline{B_r(x_0)} := \{x \in X : \|x - x_0\| \leq r\}$, and $S_r(x_0) := \{x \in X : \|x - x_0\| = r\}$ respectively denote the open ball, the closed ball, and the sphere of radius r with center x_0.

Definition 4.7. A subset B of a normed vector space X is called *bounded* if, and only if, there exists $r \in \mathbb{R}^+$ such that $B \subseteq \overline{B_r(0)}$.

4.2 Bilinear Forms

This section summarizes some elementary properties of bilinear forms on (normed) real vector spaces. Their importance for us is twofold. First, bilinear forms will be encountered frequently as inner products in Hilbert spaces. Second, bilinear forms will be used when studying elliptic linear PDE in Sec. 6.2.

Definition 4.8. Let X be a real vector space. A map $b : X \times X \longrightarrow \mathbb{R}$ is called a *bilinear form* if, and only if,

$$\begin{aligned} b\big(\lambda_1 x_1 + \lambda_2 x_2,\, y\big) &= \lambda_1\, b(x_1, y) + \lambda_2\, b(x_2, y), \\ b\big(x,\, \lambda_1 y_1 + \lambda_2 y_2\big) &= \lambda_1\, b(x, y_1) + \lambda_2\, b(x, y_2) \\ \text{for each} \quad (x, x_1, x_2, y, y_1, y_2) &\in X^6, \quad (\lambda_1, \lambda_2) \in \mathbb{R}^2. \end{aligned} \tag{4.2}$$

Definition 4.9. Let $b : X \times X \longrightarrow \mathbb{R}$ be a bilinear form on a normed vector space X (actually, (ii), (iii), (iv), and (v) use only the linear structure on X).

(i) b is called *bounded* if, and only if, there exists $\alpha_0 \geq 0$ such that

$$|b(x, y)| \leq \alpha_0\, \|x\|\, \|y\| \quad \text{for each} \quad (x, y) \in X^2. \tag{4.3a}$$

(ii) b is called *symmetric* if, and only if,

$$b(x, y) = b(y, x) \quad \text{for each} \quad (x, y) \in X^2. \tag{4.3b}$$

(iii) b is called *skew-symmetric* or *alternating* if, and only if,

$$b(x, y) = -b(y, x) \quad \text{for each} \quad (x, y) \in X^2. \tag{4.3c}$$

(iv) b is called *positive semidefinite* if, and only if,

$$b(x, x) \geq 0 \quad \text{for each} \quad x \in X. \tag{4.3d}$$

(v) b is called *positive definite* if, and only if, b is positive semidefinite and

$$\Big(b(x, x) = 0 \Leftrightarrow x = 0\Big) \qquad \text{for each} \quad x \in X. \tag{4.3e}$$

(vi) b is called *coercive* or *elliptic* if, and only if, there exists $\beta_0 > 0$ such that

$$b(x,x) \geq \beta_0 \|x\|^2 \quad \text{for each} \quad x \in X. \tag{4.3f}$$

Remark: More generally, one defines a function $f : X \longrightarrow \mathbb{R}$ to be coercive if, and only if, $\|x\| \to \infty$ implies $f(x)/\|x\| \to \infty$. Clearly, b is coercive if, and only if, $f_b : X \longrightarrow \mathbb{R}$, $f_b(x) := b(x,x)$, is coercive.

(vii) b is called an *inner product* or *scalar product* on X if, and only if, b is symmetric and positive definite. In that case, it is customary to write $\langle x, y \rangle$ instead of $b(x,y)$.

Remark and Definition 4.10. Let $b : X \times X \longrightarrow \mathbb{R}$ be a bilinear form on a real vector space X. Then there exists a unique decomposition

$$b = \sigma + a \tag{4.4}$$

such that σ is a symmetric bilinear form and a is an alternating bilinear form. Moreover, σ and a can be written explicitly in terms of b as

$$\sigma : X \times X \longrightarrow \mathbb{R}, \qquad \sigma(x,y) := \frac{1}{2}\Big(b(x,y) + b(y,x)\Big), \tag{4.5a}$$

$$a : X \times X \longrightarrow \mathbb{R}, \qquad a(x,y) := \frac{1}{2}\Big(b(x,y) - b(y,x)\Big). \tag{4.5b}$$

The forms σ and a are called the *symmetric* and the *alternating part* of b.

Remark 4.11. If X is a finite-dimensional normed vector space with basis $(e_1, \ldots, e_m)$, then, for a bilinear form $b : X \times X \longrightarrow \mathbb{R}$ with symmetric part σ, the following statements are equivalent:

(i) b is positive definite.

(ii) σ is positive definite.

(iii) b is coercive.

(iv) σ is coercive.

(v) All the eigenvalues of the matrix $S := (\sigma_{ij})$, defined by $\sigma_{ij} := \sigma(e_i, e_j)$, are positive.

4.3 Hilbert Spaces

Definition 4.12. Let X be a real vector space. If $\langle \cdot, \cdot \rangle$ is an inner product on X, then $\big(X, \langle \cdot, \cdot \rangle\big)$ is called an *inner product space* or a *pre-Hilbert space*. An inner product space is called a *Hilbert space* if, and only if, $(X, \|\cdot\|)$ is a Banach space, where the norm $\|\cdot\|$ is defined from the inner product via $\|x\| := \sqrt{\langle x, x \rangle}$. Frequently, the inner product on X is understood and X itself is referred to as an inner product space or Hilbert space.

Lemma 4.13. *The following* Cauchy-Schwarz inequality (4.6) *holds in every inner product space* $\big(X, \langle\cdot,\cdot\rangle\big)$*:*

$$|\langle x, y\rangle| \le \|x\|\,\|y\| \quad \textit{for each} \quad (x, y) \in X^2. \tag{4.6}$$

Proof. See, e.g., [Roy88, p. 245] or [Alt06, Lem. 0.2⟨2⟩]. ■

Definition 4.14. Let $\big(X, \langle\cdot,\cdot\rangle\big)$ be an inner product space and A some (countable or uncountable) index set. A family $(x_\alpha)_{\alpha\in A}$ in X is called an *orthonormal system* if, and only if, $\langle x_\alpha, x_\beta\rangle = 0$ whenever $(\alpha,\beta) \in A^2$, $\alpha \neq \beta$, and $\langle x_\alpha, x_\alpha\rangle = 1$ for each $\alpha \in A$. Given an orthonormal system $\mathcal{O} = (x_\alpha)_{\alpha\in A}$ in X, for each $x \in X$, the numbers $\hat{x}(\alpha) := \langle x, x_\alpha\rangle \in \mathbb{R}$, $\alpha \in A$, are called the *Fourier coefficients* of x with respect to $\mathcal{O}$.

Example 4.15. For each $n \in \mathbb{N}$, define

$$x_n : [0, 2\pi] \longrightarrow \mathbb{R}, \quad x_n(t) := \frac{\sin nt}{\sqrt{\pi}}. \tag{4.7}$$

Then $(x_n)_{n\in\mathbb{N}}$ constitutes an orthonormal system in the Hilbert space $L^2[0, 2\pi]$ endowed with the inner product $\langle x, y\rangle = \int_0^{2\pi} x\,y$ (such spaces will be properly introduced in Sec. 6.1 below, cf. Rem. 6.5): One computes

$$\langle x_n, x_n\rangle = \frac{1}{\pi}\int_0^{2\pi} \sin^2 nt\,\mathrm{d}t = \frac{1}{\pi}\left[\frac{t}{2} - \frac{\sin nt\cos nt}{2n}\right]_0^{2\pi} = 1 \quad \text{for each} \quad n \in \mathbb{N}, \tag{4.8a}$$

and

$$\begin{aligned}\langle x_m, x_n\rangle &= \frac{1}{\pi}\int_0^{2\pi} \sin mt \sin nt\,\mathrm{d}t \\ &= \frac{1}{\pi}\left[\frac{\sin mt\cos nt - \cos mt\sin nt}{2(m-n)} - \frac{\sin mt\cos nt - \cos mt\sin nt}{2(m+n)}\right]_0^{2\pi} \\ &= 0 \quad \text{for each} \quad (m, n) \in \mathbb{N},\ m \neq n.\end{aligned} \tag{4.8b}$$

Bessel Inequality 4.16. *Let X be an inner product space and let $(x_\alpha)_{\alpha\in A}$ be an orthonormal system in X according to Def.* 4.14*. Then, for each $x \in X$, the Bessel inequality*

$$\sum_{\alpha\in A} |\hat{x}(\alpha)| \le \|x\| \tag{4.9}$$

holds. In particular, for each $x \in X$, only countably many of the Fourier coefficients $\hat{x}(\alpha)$ can be nonzero, and, for each sequence $(\alpha_i)_{i\in\mathbb{N}}$ in A,

$$\lim_{i\to\infty} \hat{x}(\alpha_i) = 0. \tag{4.10}$$

Proof. See, e.g., [Rud87, Th. 4.17] or [Alt06, 7.6]. ■

Theorem 4.17. *Let $\big(X, \langle\cdot,\cdot\rangle\big)$ be an inner product space and let $(x_\alpha)_{\alpha\in A}$ be an orthonormal system in X according to Def.* 4.14. *Then the following statements* (i) – (iii) *are equivalent:*

(i) $x = \sum_{\alpha\in A} \hat{x}(\alpha)\, x_\alpha$ *for each* $x \in X$.

(ii) $\langle x, y\rangle = \sum_{\alpha\in A} \hat{x}(\alpha)\, \hat{y}(\alpha)$ *for each* $(x, y) \in X^2$. *This relation is known as* Parseval's identity.

(iii) $\|x\|^2 = \sum_{\alpha\in A} |\hat{x}(\alpha)|^2$ *for each* $x \in X$.

Proof. See, e.g., [Alt06, 7.7] for the case of countable A, and [Rud87, Th. 4.18] for the general case. ■

Definition 4.18. An orthonormal system $\mathcal{O}$ in an inner product space X satisfying the equivalent conditions (i) – (iii) in Th. 4.17 is called a *complete orthonormal system* or an *orthonormal basis*.

Theorem 4.19. *An orthonormal basis $\mathcal{O}$ exists in every Hilbert space $H \neq \{0\}$. Moreover, the cardinality of $\mathcal{O}$ is uniquely determined.*

Proof. See, e.g., [Alt06, Th. 7.8] for the case where H has a countable orthonormal basis, and [Rud87, Th. 4.22] for the general case. ■

4.4 Bounded Linear Operators

Definition 4.20. Let $A : X \longrightarrow Y$ be a function between two normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$.

(a) A is called *continuous* in $x \in X$ if, and only if, for each sequence $(x_n)_{n\in\mathbb{N}}$ in X, $\lim_{n\to\infty} x_n = x$ implies $\lim_{n\to\infty} A(x_n) = A(x)$; A is called *continuous* if, and only if, it is continuous in x for every $x \in X$; A is called *uniformly continuous* if, and only if, for each $\epsilon > 0$, there is $\delta > 0$ such that $(x_1, x_2) \in X^2$ and $\|x_1 - x_2\|_X < \delta$ implies $\big\|A(x_1) - A(x_2)\big\|_Y < \epsilon$; A is called *isometric* if, and only if, $\big\|A(x)\big\|_Y = \|x\|_X$ for each $x \in X$.

(b) A is called bounded if, and only if, there exists a constant $C \geq 0$ such that

$$\|A(x)\|_Y \leq C\|x\|_X \quad \text{for each} \quad x \in X. \tag{4.11}$$

—

Linear functions between normed vector spaces are usually referred to as linear *operators*. Real-valued maps are called *functionals*, in particular, real-valued linear operators are called linear functionals.

Proposition 4.21. *For a linear operator $A : X \longrightarrow Y$ between two normed vector spaces, the following statements are equivalent:*

(a) *A is uniformly continuous.*

(b) *A is continuous.*

(c) *There is $x_0 \in X$ such that A is continuous at x_0.*

(d) *A is bounded.*

Proof. The proof is straightforward; e.g., see [Roy88, Ch. 10, Prop. 2] or [Alt06, Lem. 3.1]. ■

Definition and Remark 4.22. If $A : X \longrightarrow Y$ is a bounded linear operator between normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$, then there exists the minimum of positive constants C satisfying (4.11). This minimum is denoted by $\|A\|$ and is called the *operator norm* of A. Moreover, it holds that

$$\|A\| = \sup\left\{\frac{\|A(x)\|_Y}{\|x\|_X} : x \in X,\ x \neq 0\right\} = \sup\big\{\|A(x)\| : x \in X,\ \|x\|_X = 1\big\}. \quad (4.12)$$

The vector space of all bounded linear operators between X and Y is denoted by $\mathcal{L}(X, Y)$.

Proposition 4.23. *The operator norm constitutes, indeed, a norm on the vector space $\mathcal{L}(X, Y)$ of bounded linear operators between two normed vector spaces X and Y. Moreover, if Y is a Banach space, then $\mathcal{L}(X, Y)$ is also a Banach space.*

Proof. See, e.g., [Roy88, Ch. 10, Prop. 3] or [Alt06, Th. 3.3]. ■

Definition and Remark 4.24. If X is a normed vector space, then the space $\mathcal{L}(X, \mathbb{R})$ of all bounded linear functionals on X is called the *dual* of X. The dual of X is denoted by X^*. According to (4.12), one has, for each $f \in X^*$,

$$\|f\|_{X^*} = \sup\big\{|f(x)| : x \in X,\ \|x\|_X = 1\big\}. \quad (4.13)$$

Moreover, according to Prop. 4.23, the fact that $\mathbb{R}$ is a Banach space implies that X^* is always a Banach space.

Theorem 4.25. *If X is a normed vector space and $0 \neq x \in X$, then there is $f \in X^*$ such that $\|f\|_{X^*} = 1$ and $f(x) = \|x\|_X$.*

Proof. See, e.g., [Yos74, Cor. IV.6.2] or [Alt06, 4.17⟨1⟩]. ■

Remark 4.26. Let $\big(X, \langle\cdot,\cdot\rangle_X\big)$ be an inner product space. For each $y \in X$, the map

$$f_y : X \longrightarrow \mathbb{R}, \quad f_y(x) := \langle x, y\rangle_X \quad (4.14)$$

defines a bounded linear functional on X. Moreover, (4.6) together with $\langle y, y\rangle_X = \|y\|_X^2$ implies

$$\|f_y\|_{X^*} = \sup\big\{|\langle x, y\rangle_X| : x \in X,\ \|x\|_X = 1\big\} = \|y\|_X. \quad (4.15)$$

Riesz Representation Theorem 4.27. *Let $\big(H, \langle\cdot,\cdot\rangle_H\big)$ be a Hilbert space. Then the map*

$$F : H \longrightarrow H^*, \quad F(y) := f_y, \tag{4.16}$$

where f_y is the functional defined in (4.14), constitutes an isometric isomorphism between H and H^*. In particular, for each functional $f \in H^*$, there is a unique $y \in H$ such that $\|y\|_H = \|f\|_{H^*}$ and $f(x) = \langle x, y\rangle_H$ for every $x \in H$.

Proof. See, e.g., [Roy88, Ch. 10, Prop. 28] or [Alt06, Th. 4.1]. ■

Given a Hilbert space H, it is often convenient to use Th. 4.27 to write $H = H^*$, identifying H with its dual H^*.

Definition 4.28. Let X be a normed vector space. The dual of X^* is called the *bidual* of X. One writes $X^{**} := (X^*)^*$.

Theorem 4.29. *Let X be a normed vector space. For each $x \in X$, define a functional φ_x according to*

$$\varphi_x : X^* \longrightarrow \mathbb{R}, \quad \varphi_x(f) := f(x). \tag{4.17}$$

Then φ provides an isometric isomorphism between X and a subspace $\varphi(X)$ of X^{**}.

Proof. See, e.g., [Yos74, Sec. IV.8] or [Alt06, Sec. 6.2]. ■

Definition 4.30. Given a normed vector space X, the map $\varphi : X \longrightarrow X^{**}$ defined in Th. 4.29 is called the *canonical embedding* of X into X^{**}. The space X is called *reflexive* if, and only if, the map φ is surjective, i.e. if, and only if, φ constitutes an isometric isomorphism between X and its bidual.

Caveat 4.31. It can happen that a Banach space X is isometrically isomorphic to its bidual, but *not* reflexive, i.e. the canonical embedding φ is not surjective, but there exists a different isometric isomorphism $\phi : X \cong X^{**}$, $\phi \neq \varphi$. An example of such a Banach space was constructed by R.C. James in 1951 (see [Wer02, Excercise I.4.8] and [Wer02, page 105] for the definition and further references).

Example 4.32. As a consequence of Th. 4.27, every Hilbert space is reflexive. More examples of reflexive spaces are given by the spaces $L^p(E)$, $1 < p < \infty$, defined in Sec. 6.1 below (see Th. 6.6).

4.5 Adjoint Operators

Definition 4.33. Let X, Y be Banach spaces, and let $A : X \longrightarrow Y$ be a bounded linear operator. The map

$$A^* : Y^* \longrightarrow X^*, \quad A^*(f) := f \circ A, \tag{4.18}$$

is called the *adjoint* or *dual operator* of A.

Theorem 4.34. *Let X, Y be Banach spaces, and let $A : X \longrightarrow Y$ be a bounded linear operator. Then the adjoint operator A^* of A according to Def. 4.33 is well-defined, i.e. $f \circ A \in X^*$ for each $f \in Y^*$. Moreover A^* is a bounded linear operator and $\|A^*\| = \|A\|$.*

Proof. See, e.g., [RR96, Th. 7.55]. ■

From the Riesz Representation Th. 4.27, we know that the structure of Hilbert spaces is especially benign in the sense that they are isometrically isomorphic to their duals. In Hilbert spaces, these isomorphisms can be used to pull back the adjoint operators from the dual spaces to the original spaces:

Definition 4.35. Let H_1, H_2 be Hilbert spaces, and let $A : H_1 \longrightarrow H_2$ be a bounded linear operator. Moreover, let $F_1 : H_1 \longrightarrow H_1^*$ and $F_2 : H_2 \longrightarrow H_2^*$ be the isometric isomorphisms given by the Riesz Representation Th. 4.27. Then the *Hilbert adjoint operator* $A^{*,H}$ of A is defined by

$$A^{*,H} : H_2 \longrightarrow H_1, \quad A^{*,H} := F_1^{-1} \circ A^* \circ F_2. \tag{4.19}$$

In the literature, a certain sloppiness is customary, using the same symbol A^* for both the adjoint and the Hilbert adjoint. Outside the present section, we will also adhere to this custom.

Proposition 4.36. *Let H_1, H_2 be Hilbert spaces, and let $A : H_1 \longrightarrow H_2$ and $B : H_2 \longrightarrow H_1$ be bounded linear operators. Then B is the Hilbert adjoint of A if, and only if,*

$$\langle y, Ax\rangle_{H_2} = \langle By, x\rangle_{H_1} \quad \textit{for each} \quad (x, y) \in H_1 \times H_2. \tag{4.20}$$

Proof. Let F_1 and F_2 be as in Def. 4.35.

Suppose $B = A^{*,H} = F_1^{-1} \circ A^* \circ F_2$. Fix $y \in H_2$ and set $x_0 := By$. Then, for each $x \in H_1$,

$$F_1(x_0)(x) = \langle x, x_0\rangle_{H_1} \tag{4.21a}$$

by (4.16) and (4.14). On the other hand

$$F_1(x_0) = F_1(By) = (A^* \circ F_2)(y) = A^*\big(F_2(y)\big) = \big(F_2(y)\big) \circ A,$$

and, thus, for each $x \in H_1$,

$$F_1(x_0)(x) = F_2(y)\big(Ax\big) = \langle Ax, y\rangle_{H_2} \tag{4.21b}$$

again by (4.16) and (4.14). Combining (4.21a) and (4.21b) yields

$$\langle y, Ax\rangle_{H_2} = \langle Ax, y\rangle_{H_2} = F_1(x_0)(x) = \langle x, x_0\rangle_{H_1} = \langle x_0, x\rangle_{H_1} = \langle By, x\rangle_{H_1}, \tag{4.22}$$

showing the validity of (4.20).

Now assume that $B : H_2 \longrightarrow H_1$ is some operator satisfying (4.20). Fix $y \in H_2$. According to the the first part,

$$\langle y, Ax\rangle_{H_2} = \langle A^{*,H}y, x\rangle_{H_1} \quad \text{for each} \quad x \in H_1. \tag{4.23}$$

Hence,

$$\langle By, x\rangle_{H_1} = \langle y, Ax\rangle_{H_2} = \langle A^{*,H}y, x\rangle_{H_1} \quad \text{for each} \quad x \in H_1, \tag{4.24}$$

implying $F_1(By) = F_1(A^{*,H}y)$, which, in turn, implies $By = A^{*,H}y$. Since $y \in H_2$ was arbitrary, $B = A^{*,H}$, concluding the proof of the proposition. ■

4.6 Weak Convergence

Definition 4.37. A sequence $(x_n)_{n\in\mathbb{N}}$ in a normed vector space X is called *weakly convergent* to $x \in X$ (denoted $x_n \rightharpoonup x$) if, and only if,

$$\lim_{n\to\infty} f(x_n) = f(x) \quad \text{for each} \quad f \in X^*. \tag{4.25}$$

Lemma 4.38. *Weak limits in normed vector spaces are unique, i.e., given a sequence $(x_n)_{n\in\mathbb{N}}$ in a normed vector space X, $x_n \rightharpoonup x$, $x_n \rightharpoonup y$, $(x, y) \in X^2$, implies $x = y$.*

Proof. According to Def. 4.37, for each $f \in X^*$,

$$f(x-y) = f(x) - f(y) = \lim_{n\to\infty} f(x_n) - \lim_{n\to\infty} f(x_n) = 0. \tag{4.26}$$

Thus, Th. 4.25 implies that $x - y = 0$, i.e. $x = y$. ■

Remark 4.39. Indeed, in every normed vector space X, strong convergence implies weak convergence, but, in general, weak convergence does not imply strong convergence: If $(x_n)_{n\in\mathbb{N}}$ is a sequence in X, converging strongly to $x \in X$, then, for each $f \in X^*$, $\big|f(x_n) - f(x)\big| \leq \|f\|\,\|x_n - x\| \to 0$, showing $x_n \rightharpoonup x$. For weakly convergent sequences that do not converge strongly, see the following Ex. 4.40.

Example 4.40. Let $\big(H, \langle\cdot,\cdot\rangle\big)$ be a Hilbert space and $\mathcal{O} = (x_n)_{n\in\mathbb{N}}$ an orthonormal system in H. If $f \in H^*$, then, according to the Riesz Representation Th. 4.27, there is $y \in X$ such that $f(x) = \langle x, y\rangle$ for each $x \in X$. Thus,

$$f(x_n) = \langle x_n, y\rangle \overset{\text{Def. 4.14}}{=} \hat{y}(n) \overset{(4.10)}{\to} 0 = f(0), \tag{4.27}$$

showing $x_n \rightharpoonup 0$. On the other hand $(x_n)_{n\in\mathbb{N}}$ does *not* strongly converge to 0 – actually, as

$$\begin{aligned} \|x_m - x_n\| &= \langle x_m - x_n\,,\, x_m - x_n\rangle \\ &= \|x_m\|^2 - 2\langle x_m, x_n\rangle + \|x_n\|^2 = 2 \quad \text{for each} \quad m \neq n, \end{aligned} \tag{4.28}$$

it is not even a Cauchy sequence. According to Th. 4.19, every Hilbert space has an orthonormal basis, and, thus, we see that, in every infinite-dimensional Hilbert space, there are weakly convergent sequences that do not converge strongly. A concrete example is given by the orthonormal sine functions x_n from Ex. 4.15.

Definition 4.41. Consider a subset C of a normed vector space X.

(a) C is called *weakly sequentially closed* if, and only if, for each sequence $(x_n)_{n\in\mathbb{N}}$ in C, $x_n \rightharpoonup x$, $x \in X$, implies that $x \in C$.

(b) The subset of X consisting of all points such that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in C satisfying $x_n \rightharpoonup x$ is called the *weak sequential closure* of C. It is denoted by $\mathrm{cl_w}(C)$.

(c) C is called *relatively weakly sequentially compact* if, and only if, each sequence $(x_n)_{n\in\mathbb{N}}$ in C, has a subsequence that converges weakly to some $x \in X$.

(d) C is called *weakly sequentially compact* if, and only if, each sequence $(x_n)_{n\in\mathbb{N}}$ in C has a subsequence that converges weakly to some $x \in C$.

Lemma 4.42. *Let C be a subset of a normed vector space X. Assume C is relatively weakly sequentially compact. Then C is weakly sequentially compact if, and only if, C is weakly sequentially closed.*

Proof. If $(x_n)_{n\in\mathbb{N}}$ is a sequence in C, then there is a subsequence that converges weakly to $x \in X$. If C is weakly sequentially closed, then $x \in C$, showing that C is weakly sequentially compact. Conversely, if C is weakly sequentially compact and $x_n \rightharpoonup x \in X$, then every subsequence of $(x_n)_{n\in\mathbb{N}}$ also converges weakly to x, showing $x \in C$, i.e. C is weakly sequentially closed. ■

Remark 4.43. Let C be a subset of a normed vector space X. If C is weakly sequentially closed, then it is strongly closed, but, in general, if C is strongly closed that does not imply that C is weakly sequentially closed. Indeed, if $(x_n)_{n\in\mathbb{N}}$ is a sequence in C that converges strongly to $x \in X$, then, by Rem. 4.39, $(x_n)_{n\in\mathbb{N}}$ converges weakly to x. If C is weakly sequentially closed, then $x \in C$, showing that C is strongly closed. For sets that are strongly closed without being weakly sequentially closed, see Ex. 4.45 below.

Caveat 4.44. One has to be especially careful when working with weak sequential closures, as they can have pathological properties: The example of Exercise 9 in [Rud73, Sec. 3] shows that, in general, the weak sequential closure of a set is *not(!)* weakly sequentially closed. However, for *convex* sets C, $\mathrm{cl_w}(C)$ is weakly sequentially closed (see Th. 4.47(c)).

Example 4.45. We come back to the situation considered in Ex. 4.40, i.e. an orthonormal system $\mathcal{O} = (x_n)_{n\in\mathbb{N}}$ in a Hilbert space H. It was shown in Ex. 4.40, that every sequence of orthonormal elements converges weakly to 0. However, it follows from (4.28), that a sequence of orthonormal elements can only converge strongly provided that it is finally constant. Thus, as 0 can never be an element of an orthonormal system, each infinite orthonormal system in a Hilbert space is strongly closed, but not weakly sequentially closed. As in Ex. 4.40, a concrete example is given by the orthonormal system of sine functions x_n from Ex. 4.15.

Lemma 4.46. *Let X be a normed vector space, and let A be a subset of a relatively weakly sequentially compact set $C \subseteq X$. Then:*

(a) *A is always relatively weakly sequentially compact.*

(b) *A is weakly sequentially compact if, and only if, it is weakly sequentially closed.*

Proof. (a): Since each sequence in A is a sequence in C, it must have a subsequence that converges weakly to some $x \in X$, showing that A is relatively weakly sequentially compact.

(b): As A is relatively weakly sequentially compact by (a), A is weakly sequentially compact if, and only if, it is weakly sequentially closed according to Lem. 4.42. ■

Theorem 4.47. *Let C be a convex subset of a normed vector space X.*

(a) *If C is closed, then C is weakly sequentially closed.*

(b) *The weak sequential closure of C is the same as its strong closure, i.e.* $\mathrm{cl_w}(C) = \overline{C}$.

(c) $\mathrm{cl_w}(C)$ *is weakly sequentially closed.*

(d) *Every strongly closed ball $\overline{B_r(0)}$, $r \in \mathbb{R}^+$, is weakly sequentially closed. In particular, if $B \subseteq X$ is bounded, then the weak sequential closure of B is also bounded.*

Proof. (a): See, e.g., [Roy88, Ch. 10, Cor. 23] or [Alt06, Th. 6.13].

(b): $\mathrm{cl_w}(C) \supseteq \overline{C}$ always holds as strong convergence $x_n \to x \in \overline{C}$, $x_n \in C$ for $n \in \mathbb{N}$, implies weak convergence $x_n \rightharpoonup x$. Conversely, let $x \in \mathrm{cl_w}(C)$. Then there is a sequence $(x_n)_{n\in\mathbb{N}}$ such that $x_n \rightharpoonup x$. Since $(x_n)_{n\in\mathbb{N}}$ is also a sequence in $\overline{C}$, and $\overline{C}$ is convex and closed, (a) implies $x \in \overline{C}$.

(c): According to (b), $\mathrm{cl_w}(C) = \overline{C}$. In particular, $\mathrm{cl_w}(C)$ is closed and then (a) implies that $\mathrm{cl_w}(C)$ is weakly sequentially closed.

(d): Since each $\overline{B_r(0)}$, $r \in \mathbb{R}^+$, is closed and convex, it is weakly sequentially closed by (a). ■

Definition 4.48. Consider a function $F : X \longrightarrow Y$ between two normed vector spaces X and Y.

(a) F is called *weakly sequentially continuous* if, and only if, for each sequence $(x_n)_{n\in\mathbb{N}}$ in X, $x_n \rightharpoonup x$, $x \in X$, implies $F(x_n) \rightharpoonup F(x)$.

(b) In the special case $Y = \mathbb{R}$, we call F *weakly sequentially semicontinuous from below* if, and only if, for each sequence $(x_n)_{n\in\mathbb{N}}$ in X, $x_n \rightharpoonup x$, $x \in X$, implies $\liminf\limits_{n\to\infty} F(x_n) \geq F(x)$.

Lemma 4.49. *Every bounded linear operator $A : X \longrightarrow Y$ between two normed vector spaces X and Y is weakly sequentially continuous.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X, $x \in X$, such that $x_n \rightharpoonup x$. One needs to show that $A(x_n) \rightharpoonup A(x)$. To that end, let $f \in Y^*$. Then $f \circ A \in X^*$ and the weak convergence $x_n \rightharpoonup x$ implies that $\lim_{n\to\infty}(f \circ A)(x_n) = (f \circ A)(x)$, establishing $A(x_n) \rightharpoonup A(x)$ as needed. ■

Theorem 4.50. *Let C be a closed and convex subset of a normed vector space X. If $f : C \longrightarrow \mathbb{R}$ is convex and continuous then f is weakly sequentially semicontinuous from below, i.e. for each sequence $(x_n)_{n\in\mathbb{N}}$ in C, $x \in C$, it holds that*

$$x_n \rightharpoonup x \quad \Rightarrow \quad \liminf_{n\to\infty} f(x_n) \geq f(x). \tag{4.29}$$

Proof. The assertion is a consequence of Th. 4.47(a): For each $c \in \mathbb{R}$, the set $A_c := f^{-1}]-\infty, c] = \{x \in C : f(x) \leq c\}$ is closed and convex: It is closed in the relative topology of C as the continuous inverse image of the closed set $]-\infty, c]$. Then, since C is closed in X, A_c is also closed in X. To verify its convexity, let $(x, y) \in A_c^2$ and $\alpha \in [0, 1]$. Then, as f is convex,

$$f\big(\alpha\, x + (1-\alpha)\, y\big) \leq \alpha\, f(x) + (1-\alpha)\, f(y) \ \in\,]-\infty, c] \tag{4.30}$$

since $f(x) \leq c$, $f(y) \leq c$, and $]-\infty, c]$ is convex. Thus, $\alpha\, x + (1-\alpha)\, y \in A_c$, showing that A_c is convex. Seeking a contradiction, suppose $x_n \rightharpoonup x$ in C and $\liminf_{n\to\infty} f(x_n) < f(x)$. Then there is $c \in \mathbb{R}$, $c < f(x)$, and a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that $f(x_{n_k}) \leq c$ (i.e. $x_{n_k} \in A_c$) for each $k \in \mathbb{N}$. Since $(x_{n_k})_{k\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$, one has $x_{n_k} \rightharpoonup x$ such that Th. 4.47(a) implies $x \in A_c$, i.e. $f(x) \leq c$ in contradiction to $c < f(x)$. Thus, the assumption $\liminf_{n\to\infty} f(x_n) < f(x)$ was false, thereby concluding the proof of the theorem. ■

Remark 4.51. The proof of Th. 4.50 shows that its assertion remains true if "convex, continuous functional" is replaced by the weaker hypothesis "convex functional that is semi-continuous from below".

Example 4.52. Let X be a normed vector space and $a \in X$. The distance functional $N_a : X \longrightarrow \mathbb{R}$, $x \mapsto \|x - a\|$ is continuous, convex, and weakly sequentially semicontinuous from below (in particular, letting $a := 0$, the norm itself is continuous, convex, and weakly sequentially semicontinuous from below): Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X such that $\lim_{n\to\infty} x_n = x \in X$. Then $\big|\|x_n - a\| - \|x - a\|\big| \leq \|(x_n - a) - (x - a)\| = \|x_n - x\| \to 0$, proving $\lim_{n\to\infty} \|x_n - a\| = \|x - a\|$ and, thus, the continuity of N_a. Convexity follows since, for each $(x, y) \in X^2$ and $\alpha \in [0, 1]$, the triangle inequality yields $\big\|\alpha\,(x - a) + (1-\alpha)\,(y - a)\big\| \leq \alpha\,\|x - a\| + (1-\alpha)\,\|y - a\|$. Finally, continuity and convexity imply the weak sequential semicontinuity from below via Th. 4.50.

Theorem 4.53. *Every closed ball $\overline{B_r(0)}$, $r \in \mathbb{R}^+$, in a reflexive Banach space X is weakly sequentially compact.*

Proof. See, e.g., [Yos74, Sec. 4 of App. to Ch. V, in particular, proof of "only if" part on p. 143] or [Alt06, Th. 6.10]. ■

Corollary 4.54. *Every bounded subset B of a reflexive Banach space X is relatively weakly sequentially compact.*

Proof. As B is bounded, there is $r \in \mathbb{R}^+$ such that $B \subseteq \overline{B_r(0)}$. According to Th. 4.53, $\overline{B_r(0)}$ is weakly sequentially compact. Then Lem. 4.46(a) implies that B is relatively weakly sequentially compact. ■

Corollary 4.55. *Every closed, bounded, and convex subset B of a reflexive Banach space X is weakly sequentially compact.*

Proof. B is relatively weakly sequentially compact by Cor. 4.54. As it is also weakly sequentially closed by Th. 4.47(a), it is weakly sequentially compact by Lem. 4.42. ■

5 Optimal Control in Reflexive Banach Spaces

5.1 Existence and Uniqueness

The following Th. 5.3 provides an abstract existence and uniqueness result for solutions to optimal control problems in reflexive Banach spaces. Its subsequent application will furnish existence and uniqueness results for the optimal control of PDE. It also has numerous other applications that are of independent interest. A functional analysis application with relevance to the theory of PDE is supplied in Sec. 5.2.1 below, namely the existence of an orthogonal projection. This, in turn, is employed in the usual way to prove the Lax-Milgram theorem in Sec. 5.2.2.

The basic idea of Th. 5.3 is to find conditions on f and U_{ad} such that $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ attains its infimum. If U_{ad} is closed, bounded, and convex, then it is weakly sequentially compact by Cor. 4.55, and it suffices for f to be bounded from below and weakly sequentially semicontinuous from below. However, the condition on U_{ad} to be bounded is often too restrictive, e.g., one might not want to have any control conditions at all for certain applications (i.e. U_{ad} is equal to the entire space). The boundedness condition is also too restrictive for the mentioned functional analysis application in Sec. 5.2.1. On the other hand, one might know that f approaches its infimum inside some bounded set, which turns out to suffice for the purpose of Th. 5.3. Even though this property of f is quite self-explanatory, it seems sufficiently important for us, to be highlighted in a formal definition.

Definition 5.1. Let U_{ad} be a subset of a normed vector space U, $f : U_{\text{ad}} \longrightarrow \mathbb{R}$. If f is bounded from below (i.e. if there exists $m \in \mathbb{R}$ such that $f \geq m$), then it is said to *approach its infimum in a bounded set*, if, and only if, there is $r \in \mathbb{R}^+$ and a sequence $(u_n)_{n\in\mathbb{N}}$ in $U_{\text{ad}} \cap B_r(0)$, where $B_r(0) := \{u \in U : \|u\| < r\}$, such that

$$\inf\{f(u_n) : n \in \mathbb{N}\} = \inf\{f(u) : u \in U_{\text{ad}}\} \in \mathbb{R}. \tag{5.1}$$

Lemma 5.2. *Let U_{ad} be a subset of a normed vector space U, let $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ be bounded from below. Then either one of the following two criteria is sufficient for f to approach its infimum in a bounded set:*

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

(b) *Letting* $j := \inf\{f(u) : u \in U_{\text{ad}}\}$*, there is a bounded set* $B \subseteq U$ *and* $\epsilon \in \mathbb{R}^+$ *such that*

$$u \in U_{\text{ad}} \setminus B \quad \Rightarrow \quad f(u) \geq j + \epsilon. \tag{5.2}$$

Proof. (a) is trivial. Let j and ϵ be as in (b). As B is bounded, there is $r > 0$ such that $B \subseteq B_r(0)$. According to the definition of j, there is a sequence $(u_n)_{n\in\mathbb{N}}$ in U_{ad} such that $\lim_{n\to\infty} f(u_n) = j$. Then (5.2) implies that there is $n_0 \in \mathbb{N}$ such that $u_n \in B \subseteq B_r(0)$ for each $n > n_0$, showing that f approaches its infimum in a bounded set. ■

Theorem 5.3. *Let* U *be a reflexive Banach space (e.g. a Hilbert space), let* U_{ad} *be a nonempty, closed, and convex subset of* U*, and let* $f : U_{\text{ad}} \longrightarrow \mathbb{R}$*. If* f *is bounded from below,* f *approaches its infimum in a bounded set, and* f *is weakly sequentially semicontinuous from below, then the optimal control problem*

$$\min_{u\in U_{\text{ad}}} f(u) \tag{5.3}$$

has at least one solution $\bar{u} \in U_{\text{ad}}$*. In particular,* (5.3) *has at least one solution if* f *is bounded from below, continuous, convex, and approaches its infimum in a bounded set. If, in addition,* f *is* strictly *convex, then* (5.3) *has a* unique *solution.*

Proof. Since f is assumed to be bounded from below, the infimum of its range is a real number, i.e.

$$j := \inf\big\{f(u) : u \in U_{\text{ad}}\big\} > -\infty. \tag{5.4}$$

Thus, we can choose a so-called minimizing sequence for f, that means a sequence $(u_n)_{n\in\mathbb{N}}$ in U_{ad} such that $\lim_{n\to\infty} f(u_n) = j$. Moreover, as f is assumed to approach its infimum inside a bounded set, we can choose a bounded minimizing sequence $(u_n)_{n\in\mathbb{N}}$, i.e. we can choose the sequence such that the u_n remain in some ball $B_r(0)$ for a fixed $r > 0$.

As a closed, bounded, and convex subset of a reflexive Banach space, $U_{\text{ad}} \cap \overline{B_r(0)}$ is weakly sequentially compact by Cor. 4.55. Hence, $(u_n)_{n\in\mathbb{N}}$ has a subsequence $(u_{n_k})_{k\in\mathbb{N}}$ that converges weakly to some $\bar{u} \in U_{\text{ad}} \cap \overline{B_r(0)}$. Finally, since f is weakly sequentially semicontinuous from below by hypothesis,

$$j = \lim_{k\to\infty} f(u_{n_k}) = \liminf_{k\to\infty} f(u_{n_k}) \geq f(\bar{u}). \tag{5.5}$$

On the other hand, $j \leq f(\bar{u})$ by the definition of j, yielding $j = f(\bar{u})$, showing that $\bar{u}$ is a solution to (5.3).

If f is continuous and convex, then it is weakly sequentially semicontinuous from below by Th. 4.50, such that the first part applies if f is also bounded from below and approaches its infimum in a bounded set. If the convexity of f is strict, then the uniqueness of the solution $\bar{u}$ to (5.3) is provided by Th. 2.17. ■

5.2 Applications

5.2.1 Existence of an Orthogonal Projection

We will now apply Th. 5.3 to a class of minimal distance problems, thereby obtaining the existence of orthogonal projections on closed convex sets in reflexive Banach spaces. Only the Hilbert space case will be used subsequently. However, Th. 5.3 yields the more general result without extra difficulty.

Notation 5.4. Let X be a normed vector space, $\emptyset \neq A \subseteq X$, $x \in X$. Define

$$\operatorname{dist}(x, A) := \inf\big\{\|x - y\| : y \in A\big\}. \tag{5.6}$$

Theorem 5.5. *Let C be a nonempty, closed, and convex subset of a reflexive Banach space X (e.g. a Hilbert space). Then, for each $x \in X$, the set*

$$\operatorname{proj}_C(x) := \big\{p \in C : \|x - p\| = \operatorname{dist}(x, C)\big\} \tag{5.7}$$

is nonempty and convex; moreover, if X is a Hilbert space, then $\operatorname{proj}_C(x)$ *contains precisely one element, the so-called* orthogonal projection *of x onto C (cf. Th.* 5.7 *below).*

Proof. Fix $x \in X$. Define

$$f : C \longrightarrow \mathbb{R}, \quad f(y) := \|x - y\|. \tag{5.8}$$

Then $j := \inf\{f(y) : y \in C\} = \operatorname{dist}(x, C) \geq 0$, and, in particular, f is bounded from below. Choose any $\epsilon > 0$. Let $B := B_{j+\epsilon}(x)$ be the ball with center x and radius $j + \epsilon$. If $y \in C \setminus B$, then $f(y) = \|y - x\| \geq j + \epsilon$. Since B is bounded, f satisfies criterion (b) of Lem. 5.2, i.e. f approaches its infimum in a bounded set. According to Ex. 4.52, f is also weakly sequentially semicontinuous from below. Thus, Th. 5.3 applies, showing that the optimal control problem

$$\min_{y \in C} f(y) = \min_{y \in C} \|x - y\| \tag{5.9}$$

has at least one solution $\bar{y} \in \operatorname{proj}_C(x)$, showing $\operatorname{proj}_C(x) \neq \emptyset$. Since the triangle inequality shows that f is convex, $\operatorname{proj}_C(x)$ is convex according to Th. 2.16(b). If X is a Hilbert space, then the parallelogram law holds:

$$\|a + b\|^2 + \|a - b\|^2 = 2\big(\|a\|^2 + \|b\|^2\big) \quad \text{for each } a, b \in X. \tag{5.10}$$

As above, set $j := \operatorname{dist}(x, C)$, assume $p, \tilde{p} \in \operatorname{proj}_C(x)$, and apply (5.10) with $a := x - p$, $b := x - \tilde{p}$ to obtain

$$\begin{aligned} \|p - \tilde{p}\|^2 &= 2\big(\|x - p\|^2 + \|x - \tilde{p}\|^2 - 2\|x - \tfrac{1}{2}(p - \tilde{p})\|^2\big) \\ &\leq 2\big(\|x - p\|^2 + \|x - \tilde{p}\|^2 - 2j^2\big) = 0, \end{aligned} \tag{5.11}$$

showing $p = \tilde{p}$ (the estimate in (5.11) uses the definition of j and the fact that $\frac{1}{2}(p - \tilde{p}) \in C$ as C is convex). ■

Remark 5.6. To see that set $\mathrm{proj}_C(x)$ of (5.7) can consist of more than one point if the norm is not generated by the scalar product of a Hilbert space, one just needs to recall that, for $\mathbb{R}^m$ with the max-norm, balls are actually (hyper)cubes. For example, for $C := [0,1]^2$, $x := (2,0)$, we obtain $\mathrm{proj}_C(x) = \{(1,t) : t \in [0,1]\}$.

—

In Hilbert spaces, the orthogonal projection of Th. 5.5 can be represented via a variational inequality:

Theorem 5.7. *Let C be a nonempty, closed, and convex subset of a Hilbert space X. Then there exists a unique map $p : X \longrightarrow C$ such that, for each $x \in X$,*

$$\|x - p(x)\| = \mathrm{dist}(x, C), \tag{5.12}$$

where $p(x) \in C$ satisfies (5.12) *if, and only if, the variational inequality*

$$\langle x - p(x),\, y - p(x)\rangle_X \leq 0 \tag{5.13}$$

holds for each $y \in C$. Moreover, if C is a linear subspace of X, then $p(x) \in C$ satisfies (5.12) *if, and only if, the variational equality*

$$\langle x - p(x),\, y\rangle_X = 0 \tag{5.14}$$

holds for each $y \in C$. The map p is called the orthogonal projection *from X onto C.*

Proof. The existence and uniqueness of the map p is immediate from the Hilbert space case of Th. 5.5. Suppose $p(x)$ satisfies (5.12). Fix $y \in C$. Due to the convexity of C, for each $\epsilon \in [0,1]$, it is $(1-\epsilon)\,p(x) + \epsilon\, y \in C$. One computes

$$\begin{aligned}
\|x - p(x)\|^2 &= \big(\,\mathrm{dist}(x,C)\big)^2 \leq \big\|x - \big((1-\epsilon)p(x) + \epsilon y\big)\big\|^2 \\
&= \big\|x - p(x) - \epsilon(y - p(x))\big\|^2 \\
&= \big\langle x - p(x) - \epsilon(y - p(x)),\, x - p(x) - \epsilon(y - p(x))\big\rangle_X \\
&= \|x - p(x)\|^2 - 2\,\epsilon\,\langle x - p(x),\, y - p(x)\rangle_X + \epsilon^2\,\|y - p(x)\|^2,
\end{aligned} \tag{5.15}$$

implying, for each $0 < \epsilon \leq 1$,

$$2\,\langle x - p(x),\, y - p(x)\rangle_X \leq \epsilon\,\|y - p(x)\|^2. \tag{5.16}$$

Letting $\epsilon \to 0$ yields (5.13), thereby establishing the case.

Conversely, if $p(x)$ satisfies (5.13) for $y \in C$, then

$$\begin{aligned}
\|x - y\|^2 &= \|x - p(x) + p(x) - y\|^2 \\
&= \|x - p(x)\|^2 + 2\,\langle x - p(x),\, p(x) - y\rangle_X + \|p(x) - y\|^2 \\
&\geq \|x - p(x)\|^2,
\end{aligned} \tag{5.17}$$

showing that $p(x)$ satisfies (5.12) provided that it satisfies (5.13) for each $y \in C$.

Finally, let C be a linear subspace of X. Then, for each $x \in X$, $y \in C$, $\alpha \in \mathbb{R}$, it is $a := (1-\alpha)p(x) + \alpha y \in C$. Furthermore,

$$\begin{aligned}\langle x - p(x),\, a - p(x)\rangle_X &= \langle x - p(x),\, (1-\alpha)p(x) + \alpha y - p(x)\rangle_X \\ &= \alpha\, \langle x - p(x),\, y - p(x)\rangle_X. \end{aligned} \tag{5.18}$$

Hence, if $p(x)$ satisfies (5.13), using (5.13) with y replaced by a, yields

$$0 \geq \langle x - p(x),\, a - p(x)\rangle_X = \alpha\, \langle x - p(x),\, y - p(x)\rangle_X. \tag{5.19}$$

In particular, (5.19) holds for each $\alpha \neq 0$, implying

$$\langle x - p(x),\, y - p(x)\rangle_X = 0 \quad \text{for each} \quad y \in C. \tag{5.20}$$

Since C is a linear subspace, also $-y \in C$. Replacing y by $-y$ in (5.20), we obtain

$$\langle x - p(x),\, -y - p(x)\rangle_X = 0 \quad \text{for each} \quad y \in C. \tag{5.21}$$

Subtraction (5.21) from (5.20) proves (5.14).

Conversely, if $p(x)$ satisfies (5.14) for each $y \in C$, then, as $y - p(x) \in C$ if C is a linear subspace, y can be replaced by $y - p(x)$, immediately implying (5.13) (even with equality). ■

5.2.2 Lax-Milgram Theorem

A strikingly useful tool for establishing the existence of solutions to linear elliptic PDE is provided by the Lax-Milgram Th. 5.8. The basic ingredients to its proof are the Riesz Representation Th. 4.27 and Th. 5.7.

Lax-Milgram Theorem 5.8. *Let X be a Hilbert space, and let $a : X \times X \longrightarrow \mathbb{R}$ be a bilinear form. If a is bounded and coercive (i.e. there are $\alpha_0 > 0$ and $\beta_0 > 0$ such that a satisfies* (4.3a) *and* (4.3f) *with $b = a$), then there exists a unique function $A : X \longrightarrow X$ such that*

$$a(y, x) = \langle y,\, Ax\rangle \quad \textit{for each} \quad (x, y) \in X^2. \tag{5.22}$$

Moreover, this unique function A is linear, bounded, and invertible, where

$$\|A\| \leq \alpha_0, \qquad \|A^{-1}\| \leq \frac{1}{\beta_0}. \tag{5.23}$$

Proof. Given $x \in X$, the map $f_x : X \longrightarrow \mathbb{R}$, $f_x(y) := a(y, x)$ is linear by the bilinearity of a and bounded according to (4.3a). Thus, by the Riesz Representation Theorem 4.27, there is a unique $x_f \in X$ such that $\langle y, x_f\rangle = f_x(y) = a(y, x)$ for each $y \in X$. This shows that $A : X \longrightarrow X$ satisfies (5.22) if, and only if, $Ax := x_f$ for each $x \in X$, proving both existence and uniqueness of A. It remains to verify the claimed properties of A.

A is linear: For each $(y, x_1, x_2) \in X^3$, $(\alpha_1, \alpha_2) \in \mathbb{R}^2$, one computes

$$\begin{aligned}\langle y,\, A(\alpha_1 x_1 + \alpha_2 x_2)\rangle &= a(y,\, \alpha_1 x_1 + \alpha_2 x_2) = \alpha_1\, a(y, x_1) + \alpha_2\, a(y, x_2) \\ &= \alpha_1\, \langle y,\, Ax_1\rangle + \alpha_2\, \langle y,\, Ax_2\rangle = \langle y,\, \alpha_1\, Ax_1 + \alpha_2\, Ax_2\rangle. \end{aligned} \tag{5.24}$$

As (5.24) holds for each $y \in X$, we obtain $A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1\, Ax_1 + \alpha_2\, Ax_2$, establishing the linearity of A.

$\|A\| \leq \alpha_0$: For $x \in X$, (5.22) and (4.3a) imply

$$\|Ax\|^2 = \langle Ax,\, Ax\rangle = a(Ax, x) \leq \alpha_0\, \|Ax\|\, \|x\|, \tag{5.25}$$

hence establishing the case.

For subsequent use, it is noted that (4.3f), (5.22), and (4.6) imply

$$\beta_0\, \|x\|^2 \leq a(x, x) = \langle x,\, Ax\rangle \leq \|x\|\, \|Ax\| \quad \text{for each} \quad x \in X,$$

that means

$$\beta_0\, \|x\| \leq \|Ax\| \quad \text{for each} \quad x \in X. \tag{5.26}$$

The range of A, denoted by $\mathcal{R}(A)$, is a closed subspace of X: Suppose $(x_n)_{n\in\mathbb{N}}$ is a sequence in X, $y \in X$, such that $y = \lim_{n\to\infty} Ax_n$. Then, for each $(m, n) \in \mathbb{N}^2$, (5.26) implies $\beta_0\, \|x_m - x_n\| \leq \|Ax_m - Ax_n\|$, i.e. $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since X is a Hilbert space, there is $x \in X$ such that $x = \lim_{n\to\infty} x_n$. Then the continuity of A yields $Ax = \lim_{n\to\infty} Ax_n = y$, i.e. $y \in \mathcal{R}(A)$, thereby establishing that $\mathcal{R}(A)$ is closed.

A is surjective, i.e. $\mathcal{R}(A) = X$: As $\mathcal{R}(A)$ is closed and convex, Th. 5.7 implies that there is an orthogonal projection from X onto $\mathcal{R}(A)$, i.e., according to (5.14) (since $\mathcal{R}(A)$ is a linear subspace of X), there is a map $p : X \longrightarrow \mathcal{R}(A)$ such that

$$\langle x - p(x),\, y\rangle = 0 \quad \text{for each} \quad x \in X, \quad y \in \mathcal{R}(A). \tag{5.27}$$

Fix $x \in X$ and define $x_0 := x - p(x)$. Then $Ax_0 \in \mathcal{R}(A)$ and

$$a(x_0, x_0) = \langle x_0,\, Ax_0\rangle \overset{(5.27)}{=} 0, \tag{5.28}$$

which, together with (4.3f), implies $x_0 = 0$. Thus $x = p(x) \in \mathcal{R}(A)$, showing that A is surjective.

A is one-to-one, i.e. the kernel of A is $\{0\}$: If $Ax = 0$, then (5.26) yields $x = 0$.

We have shown that A is surjective and one-to-one, i.e. A is invertible. Finally, from (5.26), for each $x \in X$, one has $\beta_0\, \|A^{-1}x\| \leq \|x\|$, proving $\|A^{-1}\| \leq (1/\beta_0)$. ■

Corollary 5.9. *Let X be a Hilbert space, and let $a : X \times X \longrightarrow \mathbb{R}$ be a bilinear form. Suppose there exist positive real constants $\alpha_0 > 0$ and $\beta_0 > 0$ such that a satisfies* (4.3a) *and* (4.3f) *with $b = a$ as in Th.* 5.8. *Then, for each $F \in X^*$, there exists a unique $y \in X$ that satisfies*

$$a(x, y) = F(x) \quad \textit{for each} \quad x \in X. \tag{5.29}$$

Moreover, the unique y that satisfies (5.29) *also satisfies*

$$\|y\| \leq \frac{1}{\beta_0} \|F\|. \tag{5.30}$$

If the bilinear form is an inner product (i.e. symmetric), then the unique y that satisfies (5.29) *is the unique absolute min of the functional*

$$f : X \longrightarrow \mathbb{R}, \quad f(x) := \frac{1}{2}\, a(x,x) - F(x). \tag{5.31}$$

Proof. The Riesz Representation Theorem 4.27 provides a unique $x_F \in X$ such that $\|x_F\| = \|F\|$ and $F(x) = \langle x, x_F \rangle$ for each $x \in X$. Then $a(x,y) = F(x)$ can be rewritten as $a(x,y) = \langle x, x_F \rangle$, such that Th. 5.8 shows that the unique $y \in X$ is given by $y = A^{-1}x_F$, where A is the unique function determined by (5.22). Moreover, (5.23) implies $\|y\| = \|A^{-1}x_F\| \leq (1/\beta_0)\, \|F\|$, i.e. (5.30). Finally, if a is symmetric, then, for each $x \in X$,

$$\begin{aligned} f(x) - f(y) &= \frac{1}{2}\big(a(x,x) - a(y,y)\big) - F(x-y) \\ &= \frac{1}{2}\big(a(x,x) - a(y,y)\big) - a(x-y,\, y) = \frac{1}{2}\big(a(x,x) - 2\,a(y,x) + a(y,y)\big) \\ &= \frac{1}{2}\, a(x-y,\, x-y) \geq \beta_0 \|x-y\|^2, \end{aligned} \tag{5.32}$$

showing that $f(x) - f(y) \geq 0$ with equality if, and only if, $x = y$. ∎

Remark 5.10. If the bilinear form a occurring in the Lax-Milgram Th. 5.8 is also symmetric, then the surjectivity of the map A follows directly from the Riesz Representation Th. 4.27 (i.e. one can avoid using the existence of an orthogonal projection (Th. 5.7)). Indeed, if, in the setting of the Lax-Milgram Th. 5.8, a is also symmetric, then a constitutes a new inner product on X with corresponding norm $\|x\|_a := \sqrt{a(x,x)}$. It follows from a satisfying (4.3a) and (4.3f) that there are $\alpha_0 > 0$ and $\beta_0 > 0$ such that

$$\beta_0 \|x\|^2 \leq a(x,x) \leq \alpha_0 \|x\|^2, \tag{5.33}$$

showing that $\|\cdot\|_a$ constitutes an equivalent norm on the Hilbert space X. If $v \in X$, then, by Th. 4.27, there is $f_v \in X^*$ such that $f_v(x) = \langle x, v \rangle$ for each $x \in X$. Using Th. 4.27 again, this time with respect to the new inner product given by a, there is $y \in X$ such that $f_v(x) = a(x,y)$ for each $x \in X$. Combined,

$$a(x,y) = f_v(x) = \langle x, v \rangle \quad \text{for each} \quad x \in X. \tag{5.34}$$

On the other hand, using the definition of A,

$$a(x,y) = \langle x,\, Ay \rangle \quad \text{for each} \quad x \in X, \tag{5.35}$$

showing $v = Ay$, i.e. A is surjective.

6 Optimal Control of Linear Elliptic PDE

6.1 Sobolev Spaces

6.1.1 L^p-Spaces

Notation 6.1. For each $m \in \mathbb{N}$, let λ_m denote m-dimensional Lebesgue measure.

Definition 6.2. Let E be a measurable subset of $\mathbb{R}^m$, $m \in \mathbb{N}$. For each $p \in [1, \infty[$, let $\mathcal{L}^p(E)$ denote the set of all measurable functions $f : E \longrightarrow \mathbb{R}$ such that

$$\int_E |f(x)|^p \, \mathrm{d}x \, < \infty. \tag{6.1}$$

For each $f \in \mathcal{L}^p(E)$ define

$$\|f\|_p := \left(\int_E |f(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}}. \tag{6.2}$$

Furthermore, define $\mathcal{L}^\infty(E)$ to be the set consisting of all measurable functions $f : E \longrightarrow \mathbb{R}$ such that

$$\|f\|_\infty := \inf \left\{ \sup \left\{ |f(x)| : x \in E \setminus N \right\} : N \text{ measurable and } \lambda_m(N) = 0 \right\} < \infty. \tag{6.3}$$

The number defined in (6.3) is also known as the *essential supremum* of f. In the usual way, we define equivalence relations on the spaces $\mathcal{L}^p(E)$, $p \in [1, \infty]$, by considering functions as equivalent if they only differ on sets of measure zero. The respective sets of equivalence classes are denoted by $L^p(E)$. A certain sloppiness is quite common in that one often does not properly distinguish between elements of $\mathcal{L}^p(E)$ and $L^p(E)$. This sloppiness does not lead to confusion in most cases and it will occasionally be present below.

Riesz-Fischer Theorem 6.3. *For each measurable $E \subseteq \mathbb{R}^m$ and each $p \in [1, \infty]$, the space $\big(L^p(E), \|\cdot\|_p\big)$ is a Banach space (whereas the $\mathcal{L}^p(E)$ are merely seminormed spaces, which is the main reason for working with $L^p(E)$).*

Proof. See, e.g., [Roy88, Ch. 11, Th. 25, Ch. 6, Th. 6] or [Els96, Ch. VI, Th. 2.5]. ■

Hölder Inequality 6.4. *Let $E \subseteq \mathbb{R}^m$ be measurable, $(p, q) \in [1, \infty]^2$ such that $1/p + 1/q = 1$. If $f \in L^p(E)$, $g \in L^q(E)$, then $f\,g \in L^1(E)$ and*

$$\int_E |f\,g| \leq \|f\|_p \, \|g\|_q. \tag{6.4}$$

Proof. See, e.g., [Roy88, Ch. 11, Th. 25, Ch. 6, Th. 4] or [Els96, Ch. VI, Th. 1.5]. ■

Remark 6.5. Observing that, for each measurable $E \subseteq \mathbb{R}^m$, the map $(f,g) \mapsto \int_E f\,g$, defines an inner product on $L^2(E)$ (note that $f\,g \in L^1(E)$ for $f,g \in L^2(E)$ by the Hölder Ineq. 6.4), Th. 6.3 implies that $L^2(E)$ is a Hilbert space.

Riesz Representation Theorem 6.6. *For each measurable $E \subseteq \mathbb{R}^m$ and each $p \in [1,\infty[$, one has $(L^p(E))^* = L^q(E)$ $(1/p + 1/q = 1)$ in the sense that, for each bounded linear functional F on $L^p(E)$, there exists a unique $g \in L^q(E)$ satisfying*

$$F(f) = \int_E f\,g \quad \text{for each} \quad f \in L^p(E), \tag{6.5}$$

and, moreover, $\|F\| = \|g\|_q$. In particular, for $1 < p < \infty$, $L^p(E)$ is reflexive.

Proof. See, e.g., [Roy88, Ch. 11, Th. 29, Ch. 6, Th. 13] or [Els96, Ch. VII, Th. 3.2]. ■

Remark 6.7. As a caveat, it is remarked that, in general, even though $(L^1(E))^* = L^\infty(E)$, one has $(L^\infty(E))^* \neq L^1(E)$, implying that $L^1(E)$ is *not* reflexive. See [Yos74, Ex. IV.9.5] for a representation of $(L^\infty(E))^*$.

Definition 6.8. Given $E \subseteq \mathbb{R}^m$ measurable and a measurable $f : E \longrightarrow \mathbb{R}$, f is called *integrable* if, and only if, $f \in L^1(E)$. Moreover, f is called *locally integrable* if, and only if, for each compact $K \subseteq E$, the restriction of f to K is in $L^1(K)$, i.e. f is integrable over every compact subset of E. The set of all locally integrable functions on E is denoted by $L^1_{\mathrm{loc}}(E)$.

6.1.2 Weak Derivatives

Definition 6.9. The *support* of a function $f : \Omega \longrightarrow \mathbb{R}$ defined on a set $\Omega \subseteq \mathbb{R}^m$, denoted $\operatorname{supp}(f)$, is the closure of the set of points, where f does not vanish, i.e.

$$\operatorname{supp}(f) := \overline{\{x \in \Omega :\ f(x) \neq 0\}} \subseteq \overline{\Omega}. \tag{6.6}$$

Notation 6.10. Let $\Omega \subseteq \mathbb{R}^m$ be open, and let $k \in \mathbb{N}_0$.

(a) Let $C^k(\Omega)$ denote the set of all functions $f : \Omega \longrightarrow \mathbb{R}$ such that f has continuous partial derivatives up to order k (i.e. including those of order k). One also says that functions in $C^k(\Omega)$ are *of class* C^k. Set $C(\Omega) := C^0(\Omega)$ and $C^\infty(\Omega) := \bigcap_{k\in\mathbb{N}_0} C^k(\Omega)$.

(b) Let $C^k(\overline{\Omega})$ denote the set of all functions from $C^k(\Omega)$ such that all their partial derivatives up to order k extend continuously to $\overline{\Omega}$. Set $C(\overline{\Omega}) := C^0(\overline{\Omega})$ and $C^\infty(\overline{\Omega}) := \bigcap_{k\in\mathbb{N}_0} C^k(\overline{\Omega})$.

(c) Let $C^k_0(\Omega)$ denote the set of all functions f from $C^k(\overline{\Omega})$ that have a compact support contained in Ω, i.e. $\operatorname{supp}(f) \subseteq \Omega$. Set $C_0(\Omega) := C^0_0(\Omega)$ and $C^\infty_0(\Omega) := \bigcap_{k\in\mathbb{N}_0} C^k_0(\Omega)$.

Remark 6.11. The elements $\phi \in C_0^\infty(\Omega)$ have numerous particularly benign and useful properties that are subsequently exploited when defining weak derivatives. These properties include the fact that ϕf is integrable for each $f \in L^1_{\mathrm{loc}}(\Omega)$, ϕ satisfies the simplified integration by parts formula (6.11), and ϕ can be differentiated as often as needed. Fortunately, elements of $\phi \in C_0^\infty(\Omega)$ are also very abundant as can be seen from the next theorem.

Theorem 6.12. *If $\Omega \subseteq \mathbb{R}^m$ is open and $p \in [1, \infty[$, then $C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$, i.e. $L^p(\Omega)$ is the closure of $C_0^\infty(\Omega)$ with respect to $\|\cdot\|_p$.*

Proof. See [Alt06, Th. 2.14⟨3⟩]. ■

Fundamental Lemma of Variational Calculus 6.13. *If $\Omega \subseteq \mathbb{R}^m$ is open and $f \in L^1_{\mathrm{loc}}(\Omega)$, then*

$$\int_\Omega f\,\phi = 0 \quad \text{for each} \quad \phi \in C_0^\infty(\Omega) \tag{6.7}$$

implies that $f = 0$ almost everywhere.

Proof. See [Zei90, Prop. 18.36]. ■

Notation 6.14. A *multi-index* α is a finite sequence of nonnegative integers, i.e. $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}_0^m$, $m \in \mathbb{N}$, where $|\alpha| := \alpha_1 + \cdots + \alpha_m$ is called the *order* of α. If $\alpha = (\alpha_1, \ldots, \alpha_m)$ is a multi-index, $\Omega \subseteq \mathbb{R}^m$ is open, and $f \in C^k(\Omega)$, where $k = |\alpha|$, then $D^\alpha f$ denotes a partial derivative of f of order k, namely

$$D^\alpha f := \left(\prod_{i=1}^m \partial_i^{\alpha_i}\right) f = \partial_1^{\alpha_1} \cdots \partial_m^{\alpha_m} f, \tag{6.8}$$

where $\partial_i^0 f := f$. Thus, for example, if $\alpha = (1, 0, 2)$, then $D^\alpha f = \partial_1 \partial_3 \partial_3 f$.

Remark 6.15. Note that the definition in (6.8) is biased with respect to the order of partial derivatives (partial derivatives with respect to the mth variable are always carried out first). Of course, for functions of class C^k, partial derivatives commute and this bias is of no consequence.

Remark 6.16. For $\Omega \subseteq \mathbb{R}^m$ open and bounded, $k \in \mathbb{N}_0$, and $f \in C^k(\overline{\Omega})$, let

$$\|f\|_{C^k(\overline{\Omega})} := \max_{x \in \overline{\Omega}} \sum_{\alpha \in \mathbb{N}_0^m :\, |\alpha| \le k} |D^\alpha f(x)|. \tag{6.9}$$

The numbers $\|f\|_{C^k(\overline{\Omega})}$ are well-defined as each of the continuous functions $|D^\alpha f|$ assumes its maximum on the compact set $\overline{\Omega}$. Moreover, $\|\cdot\|_{C^k(\overline{\Omega})}$ defines a norm on $C^k(\overline{\Omega})$ that makes $C^k(\overline{\Omega})$ into a Banach space.

Remark 6.17. Let $\Omega \subseteq \mathbb{R}^m$ be open. If $f, g \in C^1(\Omega)$ and f or g is in $C_0^1(\Omega)$, then the following *integration by parts* formula holds:

$$\int_\Omega g\,\partial_i f = -\int_\Omega f\,\partial_i g \quad \text{for each} \quad i \in \{1, \ldots, m\}. \tag{6.10}$$

To verify (6.10), one first uses the Fubini theorem to show

$$\phi \in C_0^1(\Omega) \quad \Rightarrow \quad \int_{\mathbb{R}^m} \partial_i \phi = 0 \quad \text{for each} \quad i \in \{1, \dots, m\}. \tag{6.11}$$

Then this implies (6.10) for $\Omega = \mathbb{R}^m$, which, in turn, extends to general open Ω via noticing that fg has compact support *inside* Ω, i.e. extending fg by 0 yields an element of $C_0^1(\mathbb{R}^m)$.

—

One now uses (6.10) to define derivatives (so-called *weak* derivatives) for a larger class of locally integrable functions.

Definition 6.18. Let $\Omega \subseteq \mathbb{R}^m$ be open. If $f \in L^1_{\text{loc}}(\Omega)$ and $\alpha \in \mathbb{N}_0^m$ is a multi-index, then a function $g \in L^1_{\text{loc}}(\Omega)$ is called a *weak derivative* with respect to α if, and only if,

$$\int_\Omega f \, \partial^\alpha \phi = (-1)^{|\alpha|} \int_\Omega g \, \phi \quad \text{for each} \quad \phi \in C_0^\infty(\Omega). \tag{6.12}$$

If (6.12) holds, then we will also write $D^\alpha f$ instead of g. The following Rem. 6.19 provides a justification for this notation.

Remark 6.19. Weak derivatives are unique: If g_1 and g_2 are weak derivatives of f with respect to α, then (6.12) implies $\int_\Omega (g_1 - g_2)\, \phi = 0$ for each $\phi \in C_0^\infty(\Omega)$, yielding $g_1 = g_2$ by Th. 6.13. The uniqueness of weak derivatives together with the integration by parts formula in Rem. 6.17 shows that the weak derivative $D^\alpha f$ coincides with the corresponding classical partial derivative of f provided that $f \in C^k(\Omega)$, where $k := |\alpha|$. As a caveat, it is noted that, if f does not have continuous partials up to order k, then the weak derivatives can *not* be guaranteed to agree with the classical derivatives (see [Rud73, Ex. 6.14] for examples). Here, in such cases, $D^\alpha f$ will always mean the weak derivative of f.

Definition 6.20. Let $\Omega \subseteq \mathbb{R}^m$ be open, $(k, p) \in \mathbb{N}_0 \times [1, \infty]$. By $W^{k,p}(\Omega)$ we denote the subset of $L^1_{\text{loc}}(\Omega)$ consisting of all functions f such that the weak derivatives $D^\alpha f$ exist and lie in $L^p(\Omega)$ for each multi-index α with $|\alpha| \leq k$. The spaces $W^{k,p}(\Omega)$ are referred to as *Sobolev spaces*. The case $p = 2$ is of particular interest, and one sets $H^k(\Omega) := W^{k,2}(\Omega)$. Furthermore, for each $f \in W^{k,p}(\Omega)$, $p < \infty$, define

$$\|f\|_{W^{k,p}(\Omega)} := \left(\sum_{\alpha \in \mathbb{N}_0^m \colon |\alpha| \leq k} \int_\Omega |D^\alpha f|^p \right)^{\frac{1}{p}} = \left(\sum_{\alpha \in \mathbb{N}_0^m \colon |\alpha| \leq k} \|D^\alpha f\|_p^p \right)^{\frac{1}{p}}, \tag{6.13a}$$

and, for each $f \in W^{k,\infty}(\Omega)$,

$$\|f\|_{W^{k,\infty}(\Omega)} := \max \Big\{ \|D^\alpha f\|_\infty : \alpha \in \mathbb{N}_0^m, \ |\alpha| \leq k \Big\}. \tag{6.13b}$$

Theorem 6.21. *Let $\Omega \subseteq \mathbb{R}^m$ be open. For each $(k,p) \in \mathbb{N}_0 \times [1,\infty]$, the Sobolev space $W^{k,p}(\Omega)$ is a Banach space. In particular, each $H^k(\Omega) = W^{k,2}(\Omega)$ is a Hilbert space with respect to the inner product*

$$\langle f, g\rangle_{H^k(\Omega)} = \sum_{\alpha\in\mathbb{N}_0^m\colon |\alpha|\le k} \int_\Omega D^\alpha f\, D^\alpha g = \sum_{\alpha\in\mathbb{N}_0^m\colon |\alpha|\le k} \langle D^\alpha f, D^\alpha g\rangle_{L^2(\Omega)}. \tag{6.14}$$

Proof. See, e.g., [Alt06, p. 64] or [RR96, Th. 6.65]. The case $p < \infty$ is also treated in [Yos74, Prop. I.9.5]. ■

Definition 6.22. Let $\Omega \subseteq \mathbb{R}^m$ be open. For each $(k,p) \in \mathbb{N}_0 \times [1,\infty]$, let $W_0^{k,p}(\Omega)$ denote the closure of the set $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$, i.e. the closure of $C_0^\infty(\Omega)$ with respect to the $\|\cdot\|_{W^{k,p}(\Omega)}$-norm; let $H_0^k(\Omega) := W_0^{k,2}(\Omega)$.

Remark 6.23. As closed subspaces of Banach spaces, the $W_0^{k,p}(\Omega)$ are themselves Banach spaces for each open $\Omega \subseteq \mathbb{R}^m$ and each $(k,p) \in \mathbb{N}_0 \times [1,\infty]$; the $H_0^k(\Omega)$ are Hilbert spaces.

Poincaré-Friedrich Inequality 6.24. *Let $\Omega \subseteq \mathbb{R}^m$ be open and bounded. Then there exists an Ω-dependent constant $c_\Omega > 0$ such that*

$$\|f\|^2_{L^2(\Omega)} = \int_\Omega |f|^2 \le c_\Omega \int_\Omega |\nabla f|^2 = c_\Omega \int_\Omega \sum_{i=1}^m |\partial_i f|^2 \quad \text{for each} \quad f \in H_0^1(\Omega). \tag{6.15}$$

Proof. See, e.g., [Alt06, 4.7] or [RR96, Th. 6.101]. ■

6.1.3 Boundary Issues

Our main goal is the solution and control of PDE. We will see in Sec. 6.2 that solutions of PDE are typically found as elements of Sobolev spaces. On the other hand, as was already seen in the motivating examples of Sec. 1, suitable boundary conditions are an essential part of PDE. Thus, if one is to consider PDE on some open set $\Omega \subseteq \mathbb{R}^m$, then one needs a workable notion of restriction (or trace) of elements of Sobolev spaces (i.e. of weakly differentiable functions) with respect to $\partial\Omega$. A problem arises from the fact that $W^{k,p}(\Omega)$ consists of equivalence classes of functions that only differ on sets of measure zero (with respect to λ_m) and $\lambda_m(\partial\Omega) = 0$. Thus, if $f \in W^{k,p}(\Omega)$, then different functions representing f will, in general, have different restrictions on $\partial\Omega$, and using representatives to define a restriction (or trace) of f is not feasible. On the other hand, continuous functions $f \in C(\overline{\Omega})$ do have a well-defined restriction on $\partial\Omega$, and it turns out that this can be used to define a useful notion of restriction (or trace) of weakly differentiable functions (see Th. 6.35).

In a related issue, one has to address the *regularity* of $\partial\Omega$. For general open sets Ω, the boundary can be extremely pathological. On the other hand, requiring $\partial\Omega$ to be smooth (e.g. of class C^1) seems too restrictive, as surfaces with corners occur in numerous applications. As it turns out, the notion of a Lipschitz boundary (Def. 6.27,

cf. [Alt06, Sec. A6.2], [Trö05, Sec. 2.2.2], [Wlo82, Def. 2.4]) constitutes an effective compromise. Moreover, we will restrict ourselves to the case, where Ω is bounded.

Roughly, a bounded open set $\Omega \subseteq \mathbb{R}^m$ has a Lipschitz boundary if, and only if, the boundary is locally representable as the graph of Lipschitz functions, where Ω lies on only one side of that graph. The technical definition is somewhat tedious and needs some preparation.

Remark 6.25. The case that Ω is an open bounded subset of $\mathbb{R}$ is particularly simple. However, the fact that $\partial\Omega$ is 0-dimensional requires some special treatment and definitions. As in [Trö05], we will avoid this issue by stating subsequent definitions and results only for $\Omega \subseteq \mathbb{R}^m$ with $m \geq 2$. With the appropriate modifications, the case $m = 1$ is usually much simpler to treat. For example, a bounded open set $\Omega \subseteq \mathbb{R}$ has a Lipschitz boundary if, and only if, it consists of finitely many open intervals having positive distance from each other.

Definition 6.26. Let $\Omega \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. A map $f : \Omega \longrightarrow \mathbb{R}$ is called *Lipschitz continuous* if, and only if, there exists $L \in \mathbb{R}_0^+$, such that, for each $(x, y) \in \Omega$, $|f(x) - f(y)| \leq L\,|x - y|$, recalling that we use $|\cdot|$ to denote the Euclidian norm. The set of all Lipschitz continuous maps on Ω is denoted by $C^{0,1}(\Omega)$.

—

We will now state the definition of a set with Lipschitz boundary, followed by some further explanation directly after the definition.

Definition 6.27. Let $\Omega \subseteq \mathbb{R}^m$, $m \geq 2$, be open and bounded. Then Ω is said to have a *Lipschitz boundary* if, and only if, there are finitely many open subsets of $\mathbb{R}^m$, denoted by $U_1, \ldots, U_M$, $M \in \mathbb{N}$, rotations $\rho_j : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, reference points $y^j = (y_1^j, \ldots, y_{m-1}^j) \in \mathbb{R}^{m-1}$, and Lipschitz continuous functions $h_j : \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$, $j \in \{1, \ldots, M\}$, as well as positive real numbers $a > 0$, $b > 0$, such that the following conditions (i) – (iii) are satisfied:

(i) $\partial\Omega \subseteq \bigcup_{j=1}^M U_j$.

(ii) For each $j \in \{1, \ldots, M\}$:

$$U_j = \rho_j\Big(\big\{(x_1, \ldots, x_m) : |y^j - (x_1, \ldots, x_{m-1})| < a,\ |x_m - h_j(y^j)| < b\big\}\Big).$$

(iii) For each $j \in \{1, \ldots, M\}$:

$$U_j \cap \partial\Omega = \rho_j\Big(\big\{(y, h_j(y)) : |y^j - y| < a\big\}\Big), \tag{6.16a}$$

$$U_j \cap \Omega = \rho_j\Big(\big\{(y, x_m) : |y^j - y| < a,\ h_j(y) < x_m < h_j(y) + b\big\}\Big), \tag{6.16b}$$

$$U_j \setminus \overline{\Omega} = \rho_j\Big(\big\{(y, x_m) : |y^j - y| < a,\ h_j(y) - b < x_m < h_j(y)\big\}\Big). \tag{6.16c}$$

—

The rotations ρ_j in Def. 6.27 are needed as, in general, some parts of $\partial\Omega$ are perpendicular to the $(m-1)$-dimensional hyperplane $H := \mathbb{R}^{m-1} \times \{0\} \subseteq \mathbb{R}^m$. However, after rotating H by applying ρ_j, no part of the small patch $U_j \cap \partial\Omega$ is perpendicular to the resulting hyperplane, i.e. to $\rho_j(H)$. Moreover, shifting $U_j \cap \partial\Omega$ for a small distance in the direction perpendicular to $\rho_j(H)$ either results in moving $U_j \cap \partial\Omega$ entirely inside of Ω (cf. (6.16b)) or entirely outside of Ω (cf. (6.16c)), i.e., locally, Ω lies only on one side of $U_j \cap \partial\Omega$. Moreover, a neighborhood N_j of $y^j \in H$ can be deformed into $U_j \cap \partial\Omega$ using h_j in the sense that, after applying the rotation to the deformed piece, the resulting set is identical to $U_j \cap \partial\Omega$ (cf. (6.16a)).

Theorem 6.28. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary. Then the identity map provides an isomorphism between $C^{0,1}(\overline{\Omega})$ and $W^{1,\infty}(\Omega)$:*

$$\mathrm{Id} : C^{0,1}(\overline{\Omega}) \cong W^{1,\infty}(\Omega). \tag{6.17}$$

More precisely, each $f \in C^{0,1}(\overline{\Omega})$ represents an equivalence class that is an element of $W^{1,\infty}(\Omega)$ and each element of $W^{1,\infty}(\Omega)$ contains precisely one representative that lies in $C^{0,1}(\overline{\Omega})$.

Proof. See [Alt06, Th. 8.5⟨2⟩]. The case where $\partial\Omega$ is of class C^1 is also treated in [Eva98, Sec. 5.8, Th. 4–6]. ■

Theorem 6.29. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary. Then, for each $1 \leq p < \infty$ and each $k \in \mathbb{N}_0$, $C^\infty(\overline{\Omega})$ is dense in $W^{k,p}(\Omega)$.*

Proof. See, e.g., [Alt06, Lem. A6.7] or (for the case $p = 2$) [Zei90, Cor. 21.15(a)]. ■

Before formulating a boundary trace theorem for functions in $W^{1,p}(\Omega)$, we need to define the space $L^2(\partial\Omega)$ or, more generally, L^p-spaces of functions living on $\partial\Omega$. In particular, we need to define a suitable measure on $\partial\Omega$, the so-called surface measure. The strategy for that is, given some $x \in \partial\Omega$, to first work in a local neighborhood of x (such as the sets $U_j \cap \partial\Omega$ given by Def. 6.27), and then to patch everything together using a so-called partition of unity. Finally, one needs to show that the resulting definitions do not depend on the choice of local coordinates and the choice of partition of unity. The following Def. 6.31 of a partition of unity is tailored for our purposes. In the literature, one typically finds more general and sometimes slightly modified notions of partitions of unity (see, e.g., [Alt06, 2.19], [Wlo82, Sec. 1.2]).

Definition 6.30. Let $A \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. A finite family $(O_1, \ldots, O_N)$, $N \in \mathbb{N}$, of open sets $O_i \subseteq \mathbb{R}^m$ is called a *finite open cover* of A if, and only if, $A \subseteq \bigcup_{i=1}^N O_i$.

Definition 6.31. Let $A \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. Given a finite open cover $\mathcal{O} = (O_1, \ldots, O_N)$, $N \in \mathbb{N}$, of A, a family of functions $(\eta_1, \ldots, \eta_N)$, $\eta_i \in C_0^\infty(\mathbb{R}^m)$, is called a *partition of unity subordinate to* $\mathcal{O}$ if, and only if, $0 \leq \eta_i \leq 1$, $\mathrm{supp}(\eta_i) \subseteq O_i$ for each $i \in \{1, \ldots, N\}$, and

$$\sum_{i=1}^N \eta_i(x) = 1 \quad \text{for each} \quad x \in A. \tag{6.18}$$

Partition of Unity Theorem 6.32. *Let $A \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. Given a finite open cover $\mathcal{O} = (O_1, \ldots, O_N)$, $N \in \mathbb{N}$, of A, there exists a partition of unity subordinate to $\mathcal{O}$.*

Proof. See [Alt06, 2.19⟨3⟩]. ■

Definition and Remark 6.33. Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary, $m \geq 2$. Let $U_1, \ldots, U_M$, $M \in \mathbb{N}$, $\rho_j : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, $y^j = (y_1^j, \ldots, y_{m-1}^j) \in \mathbb{R}^{m-1}$, $h_j : \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$, $j \in \{1, \ldots, M\}$, and $a > 0$, $b > 0$ be as in Def. 6.27. Let $U_0 := \Omega$. Then, according to Def. 6.27(i), $(U_1, \ldots, U_M)$ forms a finite open cover of $\partial\Omega$ and $\mathcal{O} := (U_0, \ldots, U_M)$ forms a finite open cover of $\overline{\Omega}$. Thus, by Th. 6.32, we can choose a partition of unity $(\eta_0, \ldots, \eta_M)$ subordinate to $\mathcal{O}$.

(a) A function $f : \partial\Omega \longrightarrow \mathbb{R}$ is called λ_{m-1}-*measurable* (respectively λ_{m-1}-*integrable*) if, and only if, for each $j \in \{1, \ldots, m\}$, the function

$$f_j : \mathbb{R}^{m-1} \longrightarrow \mathbb{R}, \quad f_j(y) := \begin{cases} (\eta_j f)\Big(\rho_j\big(y, h_j(y)\big)\Big) & \text{for } y \in B_a(y^j), \\ 0 & \text{for } y \in \mathbb{R}^{m-1} \setminus B_a(y^j) \end{cases} \tag{6.19}$$

is measurable (respectively integrable) in the usual sense. If $f : \partial\Omega \longrightarrow \mathbb{R}$ is integrable or nonnegative and measurable, then define

$$\int_{\partial\Omega} f \, \mathrm{d}\lambda_{m-1} := \sum_{j=1}^{M} \int_{\mathbb{R}^{m-1}} f_j(y) \sqrt{1 + |\nabla h_j(y)|^2} \, \mathrm{d}y \,, \tag{6.20}$$

where it is noted that, as $h_j : \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$ is presumed to be Lipschitz continuous, $h_j \restriction_{B_a(y^j)} \in W^{1,\infty}(B_a(y^j))$ according to (6.17).

Now, given $A \subseteq \partial\Omega$, call A λ_{m-1}-*measurable* if, and only if, the characteristic function $1_A : \partial\Omega \longrightarrow \{0, 1\}$ is λ_{m-1}-measurable. If A is λ_{m-1}-measurable, then define

$$\lambda_{m-1}(A) := \int_{\partial\Omega} 1_A \, \mathrm{d}\lambda_{m-1} \,. \tag{6.21}$$

Then λ_{m-1} defines, indeed, a measure on $\partial\Omega$, called the *surface measure*. Moreover, the integral defined in (6.20) coincides with the integral with respect to the surface measure, thereby justifying the notation $\int_{\partial\Omega} f \, \mathrm{d}\lambda_{m-1}$.

(b) For each $p \in [1, \infty[$, let $\mathcal{L}^p(\partial\Omega)$ denote the set of all λ_{m-1}-measurable functions $f : \partial\Omega \longrightarrow \mathbb{R}$ such that

$$\int_{\partial\Omega} |f|^p \, \mathrm{d}\lambda_{m-1} < \infty. \tag{6.22}$$

For each $f \in \mathcal{L}^p(\partial\Omega)$ define

$$\|f\|_p := \left(\int_{\partial\Omega} |f|^p \, \mathrm{d}\lambda_{m-1} \right)^{\frac{1}{p}}. \tag{6.23}$$

Furthermore, define $\mathcal{L}^\infty(\partial\Omega)$ to be the set consisting of all λ_{m-1}-measurable functions $f : \partial\Omega \longrightarrow \mathbb{R}$ such that

$$\|f\|_\infty := \inf\Big\{ \sup\big\{|f(x)| : x \in (\partial\Omega) \setminus N\big\} : \\ N \ \lambda_{m-1}\text{-measurable and } \lambda_{m-1}(N) = 0\Big\} < \infty. \tag{6.24}$$

Analogous to the λ_m-case, we define equivalence relations on the spaces $\mathcal{L}^p(\partial\Omega)$, $p \in [1,\infty]$, by considering functions as equivalent if they only differ on sets of λ_{m-1}-measure zero. The respective sets of equivalence classes are denoted by $L^p(\partial\Omega)$. Also analogous to the λ_m-case, it is quite common to permit a certain sloppiness, not always properly distinguishing between elements of $\mathcal{L}^p(\partial\Omega)$ and $L^p(\partial\Omega)$.

(c) We define a function

$$\nu : \partial\Omega \longrightarrow \mathbb{R}^m, \tag{6.25}$$

called the *outer unit normal* to Ω, as well as functions

$$\tau_k : \partial\Omega \longrightarrow \mathbb{R}^m \quad \text{for each} \quad k \in \{1,\dots,m-1\}, \tag{6.26}$$

called *unit tangent vectors* to Ω, as follows: Given $x \in \partial\Omega$, there is $j \in \{1,\dots,M\}$ such that $x \in U_j \cap \partial\Omega$. If $x \in U_j \cap \partial\Omega$, then, according to (6.16b), there is $y_x \in B_a(y^j)$ such that

$$x = \rho_j\big(y_x, h_j(y_x)\big). \tag{6.27}$$

Again recalling that $h_j\restriction_{B_a(y^j)} \in W^{1,\infty}(B_a(y^j))$, we define

$$\nu(x) := \big(1 + |\nabla h_j(y_x)|^2\big)^{-\frac{1}{2}} \rho_j\big(\nabla h_j(y_x), -1\big), \tag{6.28}$$

and

$$\tau_k(x) := \Big(1 + \big(\partial_k h_j(y_x)\big)^2\Big)^{-\frac{1}{2}} \rho_j\big(\tau^0(x)\big) \quad \text{for each} \quad k \in \{1,\dots,m-1\}, \tag{6.29}$$

where

$$\tau^0_{k,i}(x) := \begin{cases} 1 & \text{for } i = k, \\ \partial_k h_j(y_x) & \text{for } i = m, \\ 0 & \text{for } i \in \{1,\dots,m\} \setminus \{k,m\}. \end{cases} \tag{6.30}$$

Theorem 6.34. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary, $m \geq 2$. Then the surface measure λ_{m-1} on $\partial\Omega$, the measurability, the integrability, and the integrals with respect to λ_{m-1}, as well as the spaces $\mathcal{L}^p(\partial\Omega)$ and $L^p(\partial\Omega)$, $p \in [1,\infty]$, the normal vector ν and the* tangent space $\operatorname{span}\{\tau_1,\dots,\tau_{m-1}\}$ *to Ω spanned by the tangent vectors, all defined in Def. and Rem.* 6.33*, are independent of the chosen sets $U_1,\dots,U_M$, $M \in \mathbb{N}$, $\rho_j : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, $y^j = (y^j_1,\dots,y^j_{m-1}) \in \mathbb{R}^{m-1}$, $h_j : \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$, $j \in \{1,\dots,M\}$, as well as independent of the chosen partition of unity $(\eta_0,\dots,\eta_M)$. Note that the individual tangent vectors $\tau_1, \dots, \tau_{m-1}$ do depend on the choice of local coordinates h_j and ρ_j.*

Moreover, $L^p(\partial\Omega)$ is a Banach space for each $p \in [1,\infty]$. The outer unit normal $\nu : \partial\Omega \longrightarrow \mathbb{R}^m$ and the unit tangent vectors $\tau_k : \partial\Omega \longrightarrow \mathbb{R}^m$, $k \in \{1,\ldots,m-1\}$, have the following properties: All components are bounded λ_{m-1}-measurable functions, i.e. $\nu_i \in L^\infty(\partial\Omega)$ and $\tau_{k,i} \in L^\infty(\partial\Omega)$ for each $i \in \{1,\ldots,m\}$. Moreover, the vectors are, indeed, unit vectors, i.e. $|\nu| \equiv 1$ and $|\tau_k| \equiv 1$. For λ_{m-1}-almost every $x \in \partial\Omega$ (more precisely, using the notation from Def. and Rem. 6.33(c), for every $x = \rho_j\big(y_x, h_j(y_x)\big) \in \partial\Omega$ such that h_j is differentiable in $y_x \in B_a(y^j)$), $\nu(x)$ is perpendicular to each $\tau_k(x)$, i.e.

$$\big\langle \nu(x),\, \tau_k(x) \big\rangle_{\mathbb{R}^m} = 0 \quad \text{for each} \quad k \in \{1,\ldots,m-1\}; \tag{6.31}$$

and $\nu(x)$ is pointing outwards, *i.e. there is $\epsilon_0 > 0$ such that $x + \epsilon\nu(x) \notin \Omega$ for each $0 \leq \epsilon < \epsilon_0$.*

Proof. See [Alt06, A.6.5⟨1⟩–⟨3⟩]. ■

Theorem 6.35. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary. Then, for each $1 \leq p \leq \infty$, there is a unique bounded linear map $\tau : W^{1,p}(\Omega) \longrightarrow L^p(\partial\Omega)$, called the* trace operator, *such that, for each $f \in C^0(\overline{\Omega}) \cap W^{1,p}(\Omega)$, one has $\tau f = f\restriction_{\partial\Omega}$, where $f\restriction_{\partial\Omega}$ denotes the restriction of f to $\partial\Omega$. The function τf is called the* trace *of f on $\partial\Omega$. Moreover, the trace operator is positive, i.e. if $f \geq 0$ λ_m-almost everywhere on Ω, then $\tau f \geq 0$ λ_{m-1}-almost everywhere on $\partial\Omega$.*

Proof. See, e.g., [Alt06, Lem. A6.6] or (for the case $p = 2$) [Zei90, Th. 21.A(e)]. The positivity of the trace operator follows by noting that, when proving Th. 6.29, for $f \in W^{1,p}(\Omega)$ and $f \geq 0$, one can choose the approximating functions $f_n \in C^\infty(\overline{\Omega})$ to be nonnegative as well (cf. [KS80, Prop. 5.2(ii)]). As $\tau f_n \geq 0$, the continuity of τ then yields $\tau f \geq 0$. ■

Theorem 6.36. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary. Then, for each $1 \leq p \leq \infty$,*

$$W_0^{1,p}(\Omega) = \big\{ f \in W^{1,p}(\Omega) : \tau f = 0 \big\}. \tag{6.32}$$

Proof. The case $p < \infty$ is treated in [Alt06, Lem. A6.10]. The case $p = \infty$ follows from Th. 6.28. ■

6.2 Linear Elliptic PDE

6.2.1 Setting and Basic Definitions, Strong and Weak Formulation

Definition 6.37. An $(m \times m)$-matrix $(a_{ij})_{(i,j)\in\{1,\ldots,m\}^2}$ of measurable functions $a_{ij} : \Omega \longrightarrow \mathbb{R}$ defined on a measurable set $\Omega \subseteq \mathbb{R}^m$ is called *uniformly elliptic* if, and only if, there exists $\alpha_0 \in \mathbb{R}^+$ such that

$$\sum_{i,j=1}^{m} a_{ij}(x)\, \xi_i\, \xi_j \geq \alpha_0\, |\xi|^2 \quad \text{for each} \quad \xi \in \mathbb{R}^m \quad \text{and each} \quad x \in \Omega. \tag{6.33}$$

It is called *almost uniformly elliptic* if, and only if, there exists $\alpha_0 \in \mathbb{R}^+$ such that (6.33) holds for almost every $x \in \Omega$.

Definition 6.38. Let $\Omega \subseteq \mathbb{R}^m$ be bounded and open, $m \geq 2$. Given $h_i \in C^1(\Omega)$, a uniformly elliptic $(m \times m)$-matrix of functions $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$, each $a_{ij} \in C^1(\Omega)$, $g, b \in C^0(\Omega)$, $b \geq 0$, we call the following equations (6.34) for the unknown $y \in C^2(\Omega) \cap C^0(\overline{\Omega})$,

$$-\sum_{i=1}^{m} \partial_i \left(\sum_{j=1}^{m} a_{ij} \partial_j y + h_i \right) + by = g \quad \text{on } \Omega, \tag{6.34a}$$

$$y = 0 \quad \text{on } \partial\Omega, \tag{6.34b}$$

the corresponding *elliptic boundary value problem (BVP)* on Ω in *strong form* with a *homogeneous Dirichlet* condition on the boundary. A function $y \in C^2(\Omega) \cap C^0(\overline{\Omega})$ satisfying (6.34) is called a *strong* or *classical* solution to the problem.

Example 6.39. Let $\Omega \subseteq \mathbb{R}^m$ be bounded and open as before. In the motivating examples of Sec. 1, we considered the heat equation

$$-\operatorname{div}(\kappa \, \nabla y) = g \quad \text{on } \Omega. \tag{6.35}$$

The left-hand side can be written as

$$-\operatorname{div}(\kappa \, \nabla y) = -\sum_{i=1}^{m} \partial_i (\kappa \, \partial_i y), \tag{6.36}$$

such that (6.35) corresponds to (6.34a) with $b = 0$, $h_i = 0$, $a_{ii} = \kappa$, $a_{ij} = 0$ for $i \neq j$, $(i, j) \in \{1, \dots, m\}^2$. For this choice of the a_{ij}, if $\xi \in \mathbb{R}^m$, then

$$\sum_{i,j=1}^{m} a_{ij} \, \xi_i \, \xi_j = \kappa \, |\xi|^2. \tag{6.37}$$

In particular, $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$ is uniformly elliptic if, and only if, the function $\kappa : \Omega \longrightarrow \mathbb{R}$ is measurable and bounded away from 0 by $\alpha_0 > 0$. It is almost uniformly elliptic if, and only if, κ is measurable and $\kappa(x) \geq \alpha_0$ for almost every $x \in \Omega$.

Remark 6.40. If (6.34b) in Def. 6.38 is replaced with "$y = y_0$ on $\partial\Omega$" with some $0 \neq y_0 \in C^0(\partial\Omega)$, then one speaks of a nonhomogeneous Dirichlet condition. Here, for simplicity, we will restrict ourselves to homogeneous Dirichlet conditions. Optimal control of PDE with nonhomogeneous Dirichlet conditions involve some special difficulties (see, e.g., [Lio71, Sec. II.4.2]).

Lemma 6.41. *In the setting of Def.* 6.38, $y \in C^2(\Omega)$ *satisfies* (6.34a) *if, and only if,*

$$\int_\Omega \left(\sum_{i=1}^{m} (\partial_i v) \left(\sum_{j=1}^{m} a_{ij} \partial_j y + h_i \right) + v \, (by - g) \right) \mathrm{d}\lambda_m = 0 \quad \textit{for each} \quad v \in C_0^\infty(\Omega). \tag{6.38}$$

Proof. Suppose $y \in C^2(\Omega)$ satisfies (6.34a). Multiplying (6.34a) by $v \in C_0^\infty(\Omega)$ and integrating the result over Ω yields

$$0 = -\int_\Omega v \left(\sum_{i=1}^m \partial_i \left(\sum_{j=1}^m a_{ij} \partial_j y + h_i \right) + by - g \right) \mathrm{d}\lambda_m \,. \tag{6.39}$$

As $v \in C_0^\infty(\Omega)$ implies $v \in C_0^1(\Omega)$ and the hypotheses on the a_{ij}, h_i, and y imply that $(\sum_{j=1}^m a_{ij}\partial_j y + h_i) \in C^1(\Omega)$, the integration by parts formula (6.10) applies, and one obtains (6.38).

Conversely, if $y \in C^2(\Omega)$ satisfies (6.38) for each $v \in C_0^\infty(\Omega)$, then the integration by parts formula (6.10) yields that y satisfies (6.39) for each $v \in C_0^\infty(\Omega)$. Applying the Fundamental Lemma of Variational Calculus 6.13, one obtains that the equation in (6.34a) holds for almost every $x \in \Omega$. As the hypotheses imply that both sides of (6.34a) are a continuous function on Ω, it follows that the equation in (6.34a) actually holds everywhere in Ω. ■

Using (6.38), we now proceed to formulate the weak form of (6.34) in Def. 6.42 below. The key observation is that (6.38) makes sense if the differentiability and continuity conditions occurring in Def. 6.38 for the functions h_i, a_{ij}, g, and b are replaced by suitable integrability and boundedness conditions (see Def. 6.42), and, in that case, it is no longer necessary to require $y \in C^2(\Omega) \cap C^0(\overline{\Omega})$, but it suffices to demand that y be *weakly* differentiable, i.e. $y \in H^1(\Omega)$. If Ω is a set with Lipschitz boundary (see Def. 6.27), then the homogeneous Dirichlet condition (6.34b) can be incorporated into the weak formulation by requiring $y \in H_0^1(\Omega)$. For our subsequent considerations, the weak formulation will turn out to provide the appropriate setting.

Definition 6.42. Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary, $m \geq 2$. Given $h_i \in L^2(\Omega)$, an almost uniformly elliptic $(m \times m)$-matrix of functions $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$, each $a_{ij} \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, $b \in L^\infty(\Omega)$, $b \geq 0$ almost everywhere, we call the following equation (6.40) for the unknown $y \in H_0^1(\Omega)$,

$$\int_\Omega \left(\sum_{i=1}^m (\partial_i v) \left(\sum_{j=1}^m a_{ij} \partial_j y + h_i \right) + v\,(by - g) \right) \mathrm{d}\lambda_m \; = 0 \quad \text{for each} \quad v \in H_0^1(\Omega), \tag{6.40}$$

the corresponding *elliptic boundary value problem (BVP)* on Ω in *weak form* with a *homogeneous Dirichlet* condition on the boundary. A function $y \in H_0^1(\Omega)$ satisfying (6.40) is called a *weak solution* to the problem.

6.2.2 Existence and Uniqueness of Weak Solutions

Theorem 6.43. *Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary, $m \geq 2$. Given $h_i \in L^2(\Omega)$, $a_{ij} \in L^\infty(\Omega)$, $(i,j) \in \{1,\dots,m\}^2$, $g \in L^2(\Omega)$, $b \in L^\infty(\Omega)$, $b \geq 0$ almost everywhere, the corresponding elliptic BVP on Ω in weak form with a*

homogeneous Dirichlet condition on the boundary has a unique weak solution $y \in H^1_0(\Omega)$. *More precisely, if* $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$ *is almost uniformly elliptic, satisfying* (6.33) *with* $\alpha_0 > 0$, *then there is a unique* $y \in H^1_0(\Omega)$ *satisfying* (6.40), *and, moreover,*

$$\|y\|_{H^1(\Omega)} \leq C \max\Big(\big\{\|g\|_{L^2(\Omega)}\big\} \cup \big\{\|h_i\|_{L^2(\Omega)} : i \in \{1,\dots,m\}\big\}\Big), \tag{6.41}$$

where $C \in \mathbb{R}^+$ *is the constant defined in* (6.52) *below.*

Keeping the h_i, a_{ij}, *and* b *fixed with* $h_i = 0$ *for each* $i \in \{1,\dots,m\}$, *the* solution operator

$$S : L^2(\Omega) \longrightarrow H^1_0(\Omega), \quad g \mapsto y, \tag{6.42}$$

is a bounded linear operator that is one-to-one. Furthermore, S *is also a bounded linear operator when interpreted as a map* $S : L^2(\Omega) \longrightarrow L^2(\Omega)$.

Proof. With the intention of applying Cor. 5.9, we define a bilinear form a and investigate its properties.

Claim 1. Under the hypotheses of Th. 6.43, the map

$$a : H^1_0(\Omega) \times H^1_0(\Omega) \longrightarrow \mathbb{R}, \quad a(u,v) := \sum_{i,j=1}^{m} \int_\Omega (\partial_i u)\, a_{ij} \partial_j v \,\mathrm{d}\lambda_m + \int_\Omega u\, b\, v \,\mathrm{d}\lambda_m\,, \tag{6.43}$$

defines a bounded and coercive bilinear form.

Proof. Note that a is well-defined: If $u, v \in H^1(\Omega)$, then u, v, $\partial_i u$, and $\partial_j v$ all are in $L^2(\Omega)$. Since $a_{ij} \in L^\infty(\Omega)$ and $b \in L^\infty(\Omega)$, one has $a_{ij}\partial_j v \in L^2(\Omega)$ and $bv \in L^2(\Omega)$. Then $(\partial_i u)\, a_{ij} \partial_j v \in L^1(\Omega)$ and $u\, b\, v \in L^1(\Omega)$ such that all the integrals in (6.43) exist as real numbers.

The bilinearity of a is implied by the linearity of the weak derivatives and by the linearity of the integral.

To verify that a is bounded, we estimate for $u, v \in H^1(\Omega)$:

$$\begin{aligned} |a(u,v)| &\leq \sum_{i,j=1}^{m} \|a_{ij}\|_{L^\infty(\Omega)} \|\partial_i u\|_{L^2(\Omega)} \|\partial_j v\|_{L^2(\Omega)} + \|b\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \sum_{i,j=1}^{m} \|a_{ij}\|_{L^\infty(\Omega)} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} + \|b\|_{L^\infty(\Omega)} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\ &= \left(\|b\|_{L^\infty(\Omega)} + \sum_{i,j=1}^{m} \|a_{ij}\|_{L^\infty(\Omega)} \right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}, \end{aligned} \tag{6.44}$$

showing that a is, indeed, bounded.

It remains to check that a is coercive. Since, by hypothesis, the a_{ij} satisfy the ellipticity condition (6.33) and $b \geq 0$, we obtain, for each $u \in H^1(\Omega)$,

$$a(u,u) = \sum_{i,j=1}^{m} \int_\Omega a_{ij}\, (\partial_i u)\, (\partial_j u) \,\mathrm{d}\lambda_m + \int_\Omega b\, u^2 \,\mathrm{d}\lambda_m \geq \alpha_0 \int_\Omega |\nabla u|^2 \,\mathrm{d}\lambda_m\,. \tag{6.45}$$

Finally, using the Poincaré-Friedrich Inequality 6.24, we estimate, for each $u \in H_0^1(\Omega)$,

$$\begin{aligned} \|u\|^2_{H^1(\Omega)} &= \|u\|^2_{L^2(\Omega)} + \sum_{i=1}^m \|\partial_i u\|^2_{L^2(\Omega)} = \|u\|^2_{L^2(\Omega)} + \int_\Omega |\nabla u|^2 \,\mathrm{d}\lambda_m \\ &\overset{(6.15)}{\leq} (c_\Omega + 1) \int_\Omega |\nabla u|^2 \,\mathrm{d}\lambda_m \overset{(6.45)}{\leq} \frac{c_\Omega + 1}{\alpha_0}\, a(u,u), \end{aligned} \tag{6.46}$$

where $c_\Omega > 0$ is the constant from (6.15), showing that a is coercive. ▲

Still with the intention of applying Cor. 5.9, we define an element of $H_0^1(\Omega)^*$.
Claim 2. Under the hypotheses of Th. 6.43, the map

$$F : H_0^1(\Omega) \longrightarrow \mathbb{R}, \quad F(v) := -\int_\Omega \left(\sum_{i=1}^m h_i\, \partial_i v - gv \right) \mathrm{d}\lambda_m\,, \tag{6.47}$$

defines a bounded linear functional.

Proof. Once again, linearity is clear from the linearity of the derivatives and the integral. To see that F is bounded, let

$$C_{hg} := \max\Big(\big\{ \|g\|_{L^2(\Omega)} \big\} \cup \big\{ \|h_i\|_{L^2(\Omega)} : i \in \{1,\dots,m\} \big\} \Big) \in \mathbb{R}_0^+ \tag{6.48}$$

and compute, for each $v \in H_0^1(\Omega)$,

$$\begin{aligned} |F(v)| &\leq \sum_{i=1}^m \|h_i\|_{L^2(\Omega)}\, \|\partial_i v\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}\, \|v\|_{L^2(\Omega)} \\ &\leq C_{hg} \left(\|v\|_{L^2(\Omega)} + \sum_{i=1}^m \|\partial_i v\|_{L^2(\Omega)} \right) \leq C_{hg}\, \sqrt{1+m}\, \sqrt{\|v\|^2_{L^2(\Omega)} + \sum_{i=1}^m \|\partial_i v\|^2_{L^2(\Omega)}} \\ &= C_{hg}\, \sqrt{1+m}\, \|v\|_{H^1(\Omega)}, \end{aligned} \tag{6.49}$$

thereby establishing the case. ▲

We are now in a position to apply Cor. 5.9 with $V = H_0^1(\Omega)$. Since a and F satisfy the hypotheses of Cor. 5.9, there is a unique $y \in H_0^1(\Omega)$ that satisfies (5.29), i.e.

$$\begin{aligned} a(v,y) &= \sum_{i,j=1}^m \int_\Omega (\partial_i v)\, a_{ij} \partial_j y \,\mathrm{d}\lambda_m + \int_\Omega v\, b\, y \,\mathrm{d}\lambda_m \\ &= F(v) = -\int_\Omega \left(\sum_{i=1}^m h_i\, \partial_i v - gv \right) \mathrm{d}\lambda_m \quad \text{for each} \quad v \in H_0^1(\Omega), \end{aligned} \tag{6.50}$$

which is the same as

$$\sum_{i,j=1}^{m} \int_\Omega (\partial_i v)\, a_{ij} \partial_j y \,\mathrm{d}\lambda_m + \int_\Omega v\, b\, y \,\mathrm{d}\lambda_m + \int_\Omega \left(\sum_{i=1}^{m} h_i\, \partial_i v - g v \right) \mathrm{d}\lambda_m = 0$$

$$\text{for each} \quad v \in H_0^1(\Omega). \tag{6.51}$$

As (6.51) is precisely the defining relation (6.40) for a weak solution of the homogeneous elliptic boundary value problem on Ω, the proof of the existence and uniqueness part of the theorem is complete. Also from Cor. 5.9, we know that the norm of y can be estimated according to (5.30) by the norm of F and the constant occurring in the coercivity condition for a. Using (6.46) and (6.49), one obtains the estimate

$$\|y\|_{H^1(\Omega)} \le C\, C_{hg}, \quad C := \frac{(c_\Omega + 1)\sqrt{1+m}}{\alpha_0}, \tag{6.52}$$

thereby verifying (6.41).

Claim 3. The solution operator S is linear.

Proof. Let $g_1, g_2 \in L^2(\Omega)$, $\alpha, \beta \in \mathbb{R}$. By the definition of S, $y_1 := S(g_1)$ and $y_2 := S(g_2)$ are the weak solutions corresponding to g_1 and g_2, respectively. According to (6.40), using $h_i = 0$, one obtains

$$\int_\Omega \left(\sum_{i=1}^{m} (\partial_i v) \sum_{j=1}^{m} a_{ij} \partial_j y_1 + v\,(b y_1 - g_1) \right) \mathrm{d}\lambda_m = 0 \quad \text{for each} \quad v \in H_0^1(\Omega), \tag{6.53a}$$

$$\int_\Omega \left(\sum_{i=1}^{m} (\partial_i v) \sum_{j=1}^{m} a_{ij} \partial_j y_2 + v\,(b y_2 - g_2) \right) \mathrm{d}\lambda_m = 0 \quad \text{for each} \quad v \in H_0^1(\Omega). \tag{6.53b}$$

Adding (6.53a) and (6.53b) after multiplying by α and β, respectively, yields, for each $v \in H_0^1(\Omega)$,

$$\int_\Omega \left(\sum_{i=1}^{m} (\partial_i v) \sum_{j=1}^{m} a_{ij} \partial_j (\alpha\, y_1 + \beta\, y_2) + v \Big(b(\alpha\, y_1 + \beta\, y_2) - \alpha\, g_1 - \beta\, g_2 \Big) \right) \mathrm{d}\lambda_m = 0, \tag{6.54}$$

showing $\alpha y_1 + \beta y_2 = S(\alpha g_1 + \beta g_2)$, which is the unique weak solution corresponding to $\alpha g_1 + \beta g_2$. ▲

Claim 4. The solution operator S is one-to-one.

Proof. Since S is linear, it suffices to show that $Sg = 0$ implies $g = 0$. If $Sg = 0$, that means (6.40) holds with $y = 0$ and $h_i = 0$:

$$\int_\Omega v\, g \,\mathrm{d}\lambda_m = 0 \quad \text{for each} \quad v \in H_0^1(\Omega). \tag{6.55}$$

According to Def. 6.22, we know $C_0^\infty(\Omega) \subseteq H_0^1$, such that (6.55) together with the Fundamental Lemma of Variational Calculus 6.13 implies $g = 0$. ▲

That $S : L^2(\Omega) \longrightarrow H_0^1(\Omega)$ is bounded follows directly from (6.41). That $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ is bounded follows from (6.41) together with $\|S(g)\|_{L^2(\Omega)} \leq \|S(g)\|_{H^1(\Omega)}$. ■

Example 6.44. Letting $\Omega \subseteq \mathbb{R}^m$ be bounded, open, and with Lipschitz boundary as before, we come back to the heat equation

$$-\operatorname{div}(\kappa \, \nabla y) = g \quad \text{on } \Omega, \tag{6.56}$$

previously considered in Ex. 6.39 and Sec. 1. In Ex. 6.39, we noted that its coefficient matrix is almost uniformly elliptic if, and only if, κ is measurable and $\kappa(x) \geq \alpha_0 > 0$ for almost every $x \in \Omega$. Applying Th. 6.43, it follows that, for (6.56), the corresponding BVP in weak form with a homogeneous Dirichlet condition on the boundary, i.e. the following equation (6.57) for the unknown $y \in H_0^1(\Omega)$,

$$\int_\Omega \left(\sum_{i=1}^m \kappa \, (\partial_i v) \, (\partial_i y) - v \, g \right) \mathrm{d}\lambda_m \, = 0 \quad \text{for each} \quad v \in H_0^1(\Omega), \tag{6.57}$$

has a unique weak solution $y \in H_0^1(\Omega)$, provided that $\kappa \in L^\infty(\Omega)$, $\kappa \geq \alpha_0 > 0$ almost everywhere, and $g \in L^2(\Omega)$. Moreover, $\|y\|_{H^1(\Omega)} \leq C \, \|g\|_{L^2(\Omega)}$ and the solution operator S is linear, bounded, and one-to-one.

6.3 Optimal Control Existence and Uniqueness

We will now define a class of (reduced) optimal control problems for linear elliptic PDE as were studied in the previous Sec. 6.2. Combining Th. 6.43 on the existence and uniqueness of weak solutions to linear elliptic PDE with Th. 5.3 on the existence and uniqueness of optimal control in reflexive Banach spaces will provide existence and uniqueness results for the optimal control of linear elliptic PDE.

Definition 6.45. This definition defines what we will call an *elliptic optimal control problem*, an EOCP for short. We start with the general setting. An EOCP always includes the following general assumptions (A-1) – (A-6):

(A-1) The set $\Omega \subseteq \mathbb{R}^m$ is bounded and open with Lipschitz boundary, $m \geq 2$.

(A-2) The functions $a_{ij} \in L^\infty(\Omega)$, $(i,j) \in \{1, \ldots, m\}^2$, are such that $(a_{ij})_{(i,j) \in \{1,\ldots,m\}^2}$ is almost uniformly elliptic (cf. Def. 6.37).

(A-3) $b \in L^\infty(\Omega)$ with $b \geq 0$ almost everywhere.

(A-4) $U_{\mathrm{ad}} \subseteq L^2(\Omega)$.

(A-5) $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ is the solution operator according to Th. 6.43 (cf. (6.42)) for the corresponding elliptic BVP (with $h_i = 0$ for each $i \in \{1, \ldots, m\}$) on Ω in weak form with a homogeneous Dirichlet condition on the boundary.

(A-6) $J : L^2(\Omega) \times U_{\mathrm{ad}} \longrightarrow \mathbb{R}$.

Given Assumptions (A-1) – (A-6), the EOCP is completed by the reduced optimal control problem

$$\min J(y,u), \tag{6.58a}$$
$$u \in U_{\text{ad}}, \tag{6.58b}$$
$$y = Su, \tag{6.58c}$$

where u is the *control*, y is the *state*, (6.58b) are the *control constraints*, (6.58c) are the *equation constraints* (here, more precisely, elliptic PDE constraints), and S is the *control-to-state* operator.

A function $\bar{u} \in L^2(\Omega)$ is called an *optimal control* for the EOCP if, and only if, $u = \bar{u}$ satisfies (6.58).

Theorem 6.46. *There exists an optimal control $\bar{u} \in L^2(\Omega)$ for the elliptic optimal control problem* (EOCP) *of Def.* 6.45*, provided that U_{ad} is nonempty, closed, and convex; J is continuous, convex and such that*

$$f : U_{\text{ad}} \longrightarrow \mathbb{R}, \quad f(u) := J(Su,u), \tag{6.59}$$

is bounded from below and approaches its infimum in a bounded set (cf. Def. 5.1*). If, in addition, J is strictly convex, then the optimal control $\bar{u}$ is unique.*

Proof. As $L^2(\Omega)$ is a Hilbert space and as U_{ad} is assumed nonempty, closed, and convex, U_{ad} satisfies the hypotheses of Th. 5.3. Since f is assumed to be bounded from below and to approach its infimum in a bounded set, for f to satisfy the hypotheses of Th. 5.3, it remains to show that f is continuous and convex. According to Th. 6.43, the solution operator S is a continuous linear operator from $L^2(\Omega)$ into $L^2(\Omega)$. This, together with the assumed continuity of J, implies the continuity of f. Observing that we can apply Lem. 2.12 to the present situation (with $X := Y := C := L^2(\Omega)$ and $U := U_{\text{ad}}$), the linearity of S together with the assumed (strict) convexity of J implies the (strict) convexity of f. Thus, Th. 5.3 applies, providing an optimal control $\bar{u} \in U_{\text{ad}}$ that is unique if J is strictly convex. ∎

Example 6.47. In the setting of Def. 6.45, consider the objective functional

$$J : L^2(\Omega) \times U_{\text{ad}} \longrightarrow \mathbb{R}, \quad J(y,u) := \frac{1}{2}\,\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)} \tag{6.60}$$

with given $y_\Omega \in L^2(\Omega)$, $\lambda \in \mathbb{R}_0^+$. Note that this is precisely the objective functional that was considered for several of the motivational examples of Sec. 1, e.g. in (1.8a).

It follows from Ex. 4.52 that J is continuous. If U_{ad} is convex, then J is convex (and even strictly convex for $\lambda > 0$) as a consequence of Lem. 2.11(a),(b) (with $X := Y := C := L^2(\Omega)$ and $U := U_{\text{ad}}$, also recalling that the Hilbert space $L^2(\Omega)$ is strictly convex). To apply Th. 6.46, we need to consider f according to (6.59), i.e.

$$f : U_{\text{ad}} \longrightarrow \mathbb{R}, \quad f(u) := \frac{1}{2}\,\|Su - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}. \tag{6.61}$$

Then f is always bounded from below by 0. Moreover, if $\lambda > 0$, then f approaches its infimum in a bounded set: Let $j := \inf\{f(u) : u \in U_{\text{ad}}\} \in \mathbb{R}_0^+$ and choose an arbitrary $\epsilon > 0$. For the bounded set

$$B := \left\{ u \in L^2(\Omega) : \|u\|_{L^2(\Omega)} \leq \sqrt{\frac{2(j+\epsilon)}{\lambda}} \right\}, \tag{6.62}$$

one obtains

$$u \in U_{\text{ad}} \setminus B \quad \Rightarrow \quad f(u) = \frac{1}{2}\|Su - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u\|^2_{L^2(\Omega)} \geq \frac{\lambda}{2}\|u\|^2_{L^2(\Omega)} > j + \epsilon, \tag{6.63}$$

such that criterion (b) of Lem. 5.2 implies that f approaches its infimum in a bounded set. Hence, if $\lambda > 0$, then Th. 6.46 yields existence and uniqueness of an optimal control $\bar{u}$ whenever U_{ad} is nonempty, closed, and convex.

For $\lambda = 0$, Th. 6.46 still yields existence under the additional assumption that U_{ad} is bounded. And even though strict convexity of J will generally fail for $\lambda = 0$, we still obtain strict convexity of f via Lem. 2.11(c)(i), since S is one-to-one according to Th. 6.43. Thus, we still have uniqueness of $\bar{u}$ even for $\lambda = 0$.

Notation 6.48. Let E be a measurable subset of $\mathbb{R}^m$, $m \in \mathbb{N}$. Given $p \in [1, \infty]$, $(a, b) \in L^p(E) \times L^p(E)$ with $a \leq b$ almost everywhere, define

$$L^p_{a,b}(E) := \big\{ f \in L^p(E) : a(x) \leq f(x) \leq b(x) \text{ for almost every } x \in E \big\}. \tag{6.64}$$

Lemma 6.49. *Let E be a measurable subset of $\mathbb{R}^m$, $m \in \mathbb{N}$, let $p \in [1, \infty]$, $(a, b) \in L^p(E) \times L^p(E)$ with $a \leq b$ almost everywhere. Then $L^p_{a,b}(E)$ is a convex, closed, and bounded subset of $L^p(E)$ (i.e. closed with respect to $\|\cdot\|_{L^p(E)}$) for each $p \in [1, \infty]$.*

Proof. Let $f, g \in L^p_{a,b}(E)$, $\alpha \in [0, 1]$. Then, for almost every $x \in E$,

$$\alpha\, a(x) \leq \alpha\, f(x) \leq \alpha\, b(x), \tag{6.65a}$$
$$(1-\alpha)\, a(x) \leq (1-\alpha)\, g(x) \leq (1-\alpha)\, b(x). \tag{6.65b}$$

Adding (6.65a) and (6.65b) yields, for almost every $x \in E$,

$$a(x) \leq \big(\alpha f + (1-\alpha)g\big)(x) \leq b(x), \tag{6.66}$$

showing the convexity of $L^p_{a,b}(E)$.

If $a, b \in L^p(E)$, then the pointwise defined function $F := |a| \vee |b| := \max\{|a|, |b|\}$ is also in $L^p(E)$. Moreover, for each $f \in L^p_{a,b}(E)$, one has $|f| \leq F$, and, thus, $\|f\|_{L^p(E)} \leq \|F\|_{L^p(E)}$, showing that f is bounded in $L^p(E)$.

To verify that $L^p_{a,b}(E)$ is closed in $L^p(E)$, consider $f_n \in L^p_{a,b}(E)$, $n \in \mathbb{N}$, and $f \in L^p(E)$ such that

$$\lim_{n\to\infty} \|f_n - f\|_{L^p(E)} = 0. \tag{6.67}$$

Seeking a contradiction, assume $f \notin L^p_{a,b}(E)$. Then there is $\epsilon > 0$ and a measurable set $A \subseteq E$, $\lambda_m(A) > 0$, such that $f(x) \geq b(x) + \epsilon$ or $f(x) \leq a(x) - \epsilon$ for each $x \in A$. If $p < \infty$, then one computes, for each $n \in \mathbb{N}$,

$$\|f_n - f\|^p_{L^p(E)} = \int_E |f_n - f|^p \geq \int_A |f_n - f|^p \geq \int_A \epsilon^p = \lambda_m(A)\, \epsilon^p > 0, \tag{6.68}$$

in contradiction to (6.67). If $p = \infty$, then, for each $n \in \mathbb{N}$,

$$\|f_n - f\|_{L^p(E)} \geq \|f_n - f\|_{L^p(A)} \geq \epsilon > 0, \tag{6.69}$$

again in contradiction to (6.67). ■

Example 6.50. In the setting of Def. 6.45, we consider special forms of the set of admissible controls U_{ad}.

(a) If $U_{\text{ad}} = L^2_{a,b}(\Omega)$ with $a, b \in L^2(\Omega)$, $a \leq b$, then one speaks of *box constraints* on the control. Please note that this constitutes the infinite-dimensional version of the finite-dimensional box constraints considered in Sec. 3.5.2 (in particular, cf. (3.33)). According to Lem. 6.49, U_{ad} is a bounded, closed, and convex subset of $L^2(\Omega)$. Thus, if J is defined according to (6.60), then we know from Ex. 6.47 that the EOCP of Def. 6.45 has a unique solution $\bar{u} \in U_{\text{ad}}$ for each $\lambda \geq 0$.

(b) Consider the case of no control constraints, i.e. $U_{\text{ad}} = L^2(\Omega)$. Again, considering J defined according to (6.60), we know from Ex. 6.47 that the EOCP of Def. 6.45 has a unique solution $\bar{u} \in U_{\text{ad}}$ for each $\lambda > 0$. The case $\lambda = 0$ is now excluded due to the unboundedness of U_{ad} (cf. Ex. 6.47).

6.4 First Order Necessary Optimality Conditions, Variational Inequality

6.4.1 Differentiability in Normed Vector Spaces

In the following Def. 6.51, we start by generalizing the notion of directional derivative defined in Th. 3.10 to functions having their domain and range in general normed vector spaces. In the same setting, we then proceed to definitions of increasingly strong notions of derivatives.

Definition 6.51. Let U and V be normed vector spaces, $U_{\text{ad}} \subseteq U$, $f : U_{\text{ad}} \longrightarrow V$, and consider some $u \in U_{\text{ad}}$.

(a) Let $h \in U$. If $u + t\,h \in U_{\text{ad}}$ for each sufficiently small $t > 0$, and, moreover, the limit

$$\delta f(u, h) := \lim_{t \downarrow 0} \frac{1}{t} \big(f\big(u + t\,h\big) - f(u) \big) \tag{6.70}$$

exists in V, then $\delta f(u, h)$ is called the *directional derivative* of f at u in the direction h. This definition actually still makes sense even if U is merely a real vector space; it does not make use of any norm or topology on U.

(b) Let u be in the interior of U_{ad}. If the directional derivative $\delta f(u,h)$ exists for each $h \in U$, then the map

$$\delta f(u,\cdot) : U \longrightarrow V, \quad h \mapsto \delta f(u,h), \tag{6.71}$$

is called the *first variation* of f at u (note that, if u is in the interior of U_{ad}, then, for each $h \in U$, $u + t\,h \in U_{\mathrm{ad}}$ for each sufficiently small $t > 0$).

If, in addition, the first variation constitutes a bounded linear operator, then it is called the *Gâteaux derivative* of f at u. In that case, f is called *Gâteaux differentiable* at u. If the first variation is a Gâteaux derivative, then one highlights this by writing $f'_{\mathrm{G}}(u)$ instead of $\delta f(u,\cdot)$.

(c) As in (b), assume u to be in the interior of U_{ad}. Then f is called *Fréchet differentiable* at u if, and only if, there exists a bounded linear operator $A : U \longrightarrow V$ such that

$$h \neq 0, \quad \|h\|_U \to 0 \quad \Rightarrow \quad \frac{\big\|f(u+h) - f(u) - Ah\big\|_V}{\|h\|_U} \to 0. \tag{6.72}$$

In that case, one writes $f'_{\mathrm{F}}(u)$ instead of A and calls $f'_{\mathrm{F}}(u)$ the *Fréchet derivative* of f at u.

The function f is called Gâteaux (resp. Fréchet) differentiable if, and only if, it is Gâteaux (resp. Fréchet) differentiable at each $u \in U_{\mathrm{ad}}$, U_{ad} open.

Lemma 6.52. *Let U and V be normed vector spaces, $U_{\mathrm{ad}} \subseteq U$ open, $f : U_{\mathrm{ad}} \longrightarrow V$. If f is Fréchet differentiable at $u \in U_{\mathrm{ad}}$, then the Fréchet derivative $f'_{\mathrm{F}}(u)$ is unique and equal to the Gâteaux derivative, i.e. $f'_{\mathrm{F}}(u) = f'_{\mathrm{G}}(u)$. Moreover, f is continuous at u.*

Proof. If f is Fréchet differentiable, then there is a bounded linear operator $A : U \longrightarrow V$ satisfying (6.72). For fixed $0 \neq h \in U$, letting $t \downarrow 0$, yields $\|t\,h\|_U \to 0$ such that (6.72) implies

$$0 = \lim_{t \downarrow 0} \frac{\big\|f(u+t\,h) - f(u) - A(t\,h)\big\|_V}{\|t\,h\|_U} = \lim_{t \downarrow 0} \left\| \frac{1}{t}\big(f\big(u+t\,h\big) - f(u)\big) - Ah \right\|_V, \tag{6.73}$$

thereby identifying Ah as $\delta f(u,h) = f'_{\mathrm{G}}(u)(h)$. In particular, as the value $\delta f(u,h)$ is unique for each $(u,h) \in U_{\mathrm{ad}} \times U$ according to its definition, the Fréchet derivative is unique. To see that f is continuous at u, consider a sequence $(u_n)_{n\in\mathbb{N}}$ in $U_{\mathrm{ad}} \setminus \{u\}$ such that $u_n \to u$. Letting $h_n := u_n - u$, one obtains $\|h_n\|_U \to 0$. Applying (6.72) once again, yields

$$0 = \lim_{n\to\infty} \frac{\big\|f(u+h_n) - f(u) - Ah_n\big\|_V}{\|h_n\|_U}. \tag{6.74}$$

Thus, there exists $n_0 \in \mathbb{N}$ such that, for each $n \geq n_0$, $\|f(u+h_n) - f(u) - Ah_n\|_V < \|h_n\|_U$. In consequence, for each $n \geq n_0$,

$$\begin{aligned} \big\|f(u_n) - f(u)\big\|_V &\leq \big\|f(u_n) - f(u) - Ah_n\big\|_V + \big\|Ah_n\big\|_V \\ &= \big\|f(u+h_n) - f(u)\big\|_V + \big\|Ah_n\big\|_V \\ &< \|h_n\|_U + \|A\|\,\|h_n\|_U \to 0, \end{aligned} \tag{6.75}$$

thereby showing $f(u_n) \to f(u)$, proving the continuity of f. ■

Remark 6.53. If U is a normed vector space, $U_{\text{ad}} \subseteq U$ open, $u \in U_{\text{ad}}$, and $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ is Gâteaux differentiable at u, then, according to Def. 6.51(b), $f'_{\text{G}}(u) \in U^*$.

Remark 6.54. Let U and V be normed vector spaces, $U_{\text{ad}} \subseteq U$, $f : U_{\text{ad}} \longrightarrow V$, $u \in U_{\text{ad}}$, $h \in U$. If $\delta f(u,h)$ exists, then it is immediate from (6.70) that the function

$$f\restriction_{D_h}, \quad D_h := U_{\text{ad}} \cap \{u + th : t \in \mathbb{R}_0^+\} \tag{6.76}$$

is continuous in u. In particular, if f has a first variation $\delta f(u,\cdot)$ for each $u \in U$, then f is continuous in every straight direction. However, the following Ex. 6.55 shows that this does not imply that f has to be continuous.

Example 6.55. In (a), we first construct a function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ that is Gâteaux differentiable in $(0,0)$, but not continuous in $(0,0)$. Then, in (b), we show that by working a little harder, we can even get f to be everywhere Gâteaux differentiable, but still not continuous in $(0,0)$.

(a) Define

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := \begin{cases} 1 & \text{if } x > 0 \text{ and } y = x^2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.77}$$

Since $\lim_{n\to\infty}(n^{-1}, n^{-2}) = (0,0)$ and $\lim_{n\to\infty} f(n^{-1}, n^{-2}) = 1 \neq 0 = f(0,0)$, f is not continuous in $(0,0)$. On the other hand, we show that

$$\delta f\big((0,0),h\big) = 0 \quad \text{for each } h = (h_1,h_2) \in \mathbb{R}^2, \tag{6.78}$$

that means, in particular, f is Gâteaux differentiable at $(0,0)$ with $f'_{\text{G}}(0,0) \equiv 0$. If $h_1 \leq 0$ or $h_2 \leq 0$, then, for each $t > 0$, we have $f((0,0) + th) = f(th_1, th_2) = 0$, showing $\delta f\big((0,0),h\big) = 0$. It remains the case, where $h_1 > 0$ and $h_2 > 0$. In that case, we obtain $t^2 h_1^2 < t h_2$ for each $0 < t < \frac{h_2}{h_1^2}$. Thus, for such t, $f(th_1, th_2) = 0$, once again proving $\delta f\big((0,0),h\big) = 0$.

(b) For each $a \in \mathbb{R}^2$ and each $\epsilon > 0$, let $\phi_{a,\epsilon} \in C_0^\infty(\mathbb{R}^2)$ satisfying $\phi_{a,\epsilon}(a) = 1$ and $\operatorname{supp}(\phi_{a,\epsilon}) \subseteq B_\epsilon(a)$, where we consider $\mathbb{R}^2$ endowed with the max-norm. Let $C := \{(x,y) \in \mathbb{R}^2 : y = x^2\}$. For each $n \in \mathbb{N} \setminus \{1\}$, define

$$a_n := (n^{-1}, n^{-3}), \tag{6.79a}$$

$$\delta_n := \operatorname{dist}\big(a_n, C\big). \tag{6.79b}$$

As $a_n \notin C$ and C is closed, we know $\delta_n > 0$ for each $n \in \mathbb{N} \setminus \{1\}$. Next, for each $n \in \mathbb{N}$, define

$$\epsilon_n := \min\left\{\frac{\operatorname{dist}(a_n, a_{n+1})}{2}, \frac{1}{n^3}, \delta_n\right\} > 0, \tag{6.79c}$$

$$B := \bigcup_{n\in\mathbb{N}\setminus\{1\}} B_{\epsilon_n}(a_n). \tag{6.79d}$$

Note that the B_n are all disjoint since $\operatorname{dist}(a_n, a_{n+1}) < \operatorname{dist}(a_n, a_{n-1})$. Thus, we can define

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := \begin{cases} \phi_{a_n,\epsilon_n}(x,y) & \text{if } (x,y) \in B_{\epsilon_n}(a_n), \\ 0 & \text{if } (x,y) \in \mathbb{R}^2 \setminus B. \end{cases} \tag{6.79e}$$

Due to the definition of ϵ_n, we have $(0,0) \notin B$. Thus, as $f(a_n) = \phi_{a_n,\epsilon_n}(a_n) = 1$ for each $n \in \mathbb{N}$, f not being continuous in $(0,0)$ follows precisely as in (a).

To show $f \in C^\infty\big(\mathbb{R}^2 \setminus \{(0,0)\}\big)$, we first compute the closure $\overline{B}$ of B:

Claim 1. $\overline{B} = \{(0,0)\} \cup \bigcup_{n\in\mathbb{N}\setminus\{1\}} \overline{B}_{\epsilon_n}(a_n)$.

Proof. For the inclusion "$\supseteq$", merely observe that, for each point p in the right-hand side, we can clearly find a sequence in B converging to p. For "$\subseteq$", we show that the right-hand side must contain all cluster points of sequences in B: Let $(b_k)_{k\in\mathbb{N}}$ be a sequence in B that does not have a cluster point in any $\overline{B}_{\epsilon_n}(a_n)$. Thus, for each $n \in \mathbb{N}$, the set $\{k \in \mathbb{N} : b_k \in B_{\epsilon_n}(a_n)\}$ must be finite. In other words, for each $\epsilon > 0$, $B_\epsilon(0,0)$ must contain all but finitely many of the b_k, showing $\lim_{k\to\infty} = (0,0)$. ▲

Claim 2. $f \in C^\infty\big(\mathbb{R}^2 \setminus \{(0,0)\}\big)$.

Proof. Let $(x,y) \in \mathbb{R}^2 \setminus \{(0,0)\}$. We consider several cases:

$(x,y) \in B$ (i.e. $(x,y) \in B_{\epsilon_n}(a_n)$ for some $n \in \mathbb{N} \setminus \{1\}$: f is C^∞ in (x,y), since f coincides with the C^∞ function ϕ_{a_n,ϵ_n} on the open neighborhood $B_{\epsilon_n}(a_n)$ of (x,y).

$(x,y) \in \partial B_{\epsilon_n}(a_n)$ for some $n \in \mathbb{N} \setminus \{1\}$: Let

$$\eta_0 := \begin{cases} \min\Big\{ \operatorname{dist}\big(\overline{B}_{\epsilon_n}(a_n), \overline{B}_{\epsilon_{n+1}}(a_{n+1})\big), \operatorname{dist}\big(\overline{B}_{\epsilon_n}(a_n), \overline{B}_{\epsilon_{n-1}}(a_{n-1})\big)\Big\} & \text{for } n > 2, \\ \operatorname{dist}\big(\overline{B}_{\epsilon_n}(a_n), \overline{B}_{\epsilon_{n+1}}(a_{n+1})\big) & \text{for } n = 2. \end{cases} \tag{6.80}$$

Due to the choice of the ϵ_n, one has $\eta_0 > 0$ and, in consequence,

$$\eta_1 := \min\Big\{\epsilon_0, \operatorname{dist}\big((x,y), \operatorname{supp}(\phi_{a_n,\epsilon_n})\big)\Big\} > 0. \tag{6.81}$$

We now obtain that f is C^∞ in (x,y), since $f \equiv 0$ on the open neighborhood $B_{\eta_1}(x,y)$ of (x,y).

$(x,y) \in \mathbb{R}^2 \setminus \overline{B}$: Since $\mathbb{R}^2 \setminus \overline{B}$ is open, there is $\epsilon > 0$ such that $B_\epsilon(x,y) \subseteq R^2 \setminus \overline{B}$. Thus, f is C^∞ in (x,y), since $f \equiv 0$ on the open neighborhood $B_\epsilon(x,y)$ of (x,y).

Due to Cl. 1, we have covered every possible case. ▲

As in (a), we show the validity of (6.78), that means, in particular, f is Gâteaux differentiable at $(0,0)$ with $f'_{\mathrm{G}}(0,0) \equiv 0$: Due to the choice of the ϵ_n, we have

$$B \subseteq \{(x,y) \in \mathbb{R}^2 : 1 > x > 0 \text{ and } x^2 > y > 0\}. \tag{6.82}$$

Thus, as $f(x,y) = 0$ for each $(x,y) \notin B$, we can argue as in (a): If $h_1 \leq 0$ or $h_2 \leq 0$, then, for each $t > 0$, we have $f((0,0) + th) = f(th_1, th_2) = 0$ as $(th_1, th_2) \notin B$, showing $\delta f\big((0,0), h\big) = 0$. It remains the case, where $h_1 > 0$ and $h_2 > 0$, and, in that case, we obtain $t^2 h_1^2 < th_2$ for each $0 < t < \frac{h_2}{h_1^2}$. Thus, for such t, we again have $(th_1, th_2) \notin B$ and $f(th_1, th_2) = 0$, proving $\delta f\big((0,0), h\big) = 0$ also in this case.

Remark 6.56. In [IT79, p. 24], one can find an example of a function f where the first variation $\delta f(0, \cdot)$ is nonlinear, such that f is not Gâteaux differentiable at 0.

Example 6.57. (a) Each constant map $f : U \longrightarrow V$ between normed vector spaces U and V is Fréchet differentiable with $f'_{\mathrm{F}}(u) = 0$ for every $u \in U$: Just notice that, for constant f and $A = 0$, the numerator in the conclusion of (6.72) vanishes identically.

(b) A linear map $f : U \longrightarrow V$ between normed vector spaces U and V always has a first variation at each $u \in U$ and $\delta f(u,h) = f(h)$ for each $h \in U$. Moreover, if f is linear, then the following statements are equivalent:

(i) f is continuous.

(ii) f is Gâteaux differentiable. In that case, $f'_{\mathrm{G}}(u) = f$ for each $u \in U$.

(iii) f is Fréchet differentiable. In that case, $f'_{\mathrm{F}}(u) = f$ for each $u \in U$.

First, note that, for linear f, the quantity in the limit of (6.70) equals $f(h)$ for every $(u,h) \in U^2$ and each $t > 0$, showing $\delta f(u,h) = f(h)$. Next, if f and $A : U \longrightarrow V$ are linear, then, for each $u \in U$ and each $0 \neq h \in U$,

$$\frac{\big\|f(u+h) - f(u) - Ah\big\|_V}{\|h\|_U} = \frac{\big\|f(h) - Ah\big\|_V}{\|h\|_U}. \tag{6.83}$$

Thus, if f is continuous, then one can choose $A := f$, causing the expression in (6.83) to vanish identically, showing that (6.72) is satisfied. In particular, f is Fréchet differentiable with $f'_{\mathrm{F}}(u) = f$. This shows that (i) implies (iii). According to Lem. 6.52, (iii) implies (ii). If f is Gâteaux differentiable, then $\delta f(u, \cdot) = f$ is a continuous linear operator, establishing that (ii) implies (i).

(c) We consider the one-dimensional case, i.e. $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$, where $U_{\mathrm{ad}} \subseteq \mathbb{R}$. Consider $u \in U_{\mathrm{ad}}$ and $h > 0$. If there is $\epsilon > 0$ such that $[u, u+\epsilon] \subseteq U_{\mathrm{ad}}$, then $\delta f(u,h)$ exists if, and only if, f is differentiable from the right at u in the classical sense. In that case $\delta f(u,h) = h\, f'_+(u)$, where $f'_+(u)$ denotes the derivative from the right of f at u. Analogously, if there is $\epsilon > 0$ such that $[u-\epsilon, u] \subseteq U_{\mathrm{ad}}$, then $\delta f(u,-h)$ exists if, and only if, f is differentiable from the left at u. In that case $\delta f(u,-h) = -h\, f'_-(u)$, where $f'_-(u)$ denotes the derivative from the left of f at u.

Claim 1. The following statements are equivalent:

(i) u is in the interior of U_{ad} and $f'_+(u) = f'_-(u)$, i.e. the classical derivative $f'(u)$ of f at u exists.

(ii) f is Gâteaux differentiable at u. In that case, $f'_{\mathrm{G}}(u)(h) = h\,f'(u)$.

(iii) f is Fréchet differentiable at u. In that case, $f'_{\mathrm{F}}(u)(h) = h\,f'(u)$.

Proof. (iii) implies (ii) according to Lem. 6.52. If f is Gâteaux differentiable at u, then $\delta f(u,h)$ exists for all $h \in \mathbb{R}$, i.e. u must be an interior point of U_{ad}. Moreover, for $h > 0$, the linearity of $\delta f(u,\cdot)$ yields

$$f'_+(u) = \frac{\delta f(u,h)}{h} = \frac{-\delta f(u,h)}{-h} = \frac{\delta f(u,-h)}{-h} = f'_-(u), \tag{6.84}$$

showing that (ii) implies (i). Finally, assuming (i), one concludes, for $h \to 0$, $h \neq 0$,

$$\frac{f(u+h) - f(u) - h\,f'(u)}{h} = \frac{f(u+h)-f(u)}{h} - f'(u) \to f'(u) - f'(u) = 0, \tag{6.85}$$

i.e. (i) implies (iii), thereby establishing the case. ▲

(d) Let $(a,b) \in \mathbb{R}$, $a < b$. We let $U := C[a,b]$ endowed with the $\|\cdot\|_\infty$-norm, and consider point functionals f_x, i.e., for $x \in [a,b]$,

$$f_x : C[a,b] \longrightarrow \mathbb{R}, \quad f_x(u) := u(x). \tag{6.86}$$

Since, for each $u \in C[a,b]$, $|f_x(u)| \leq \|u\|_\infty$, f_x is a bounded linear functional. According to (b), f_x is Fréchet differentiable, and, for each $u \in U$,

$$(f_x)'_{\mathrm{F}}(u)(h) = f_x(h) = h(x). \tag{6.87}$$

(e) Let H be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. Consider

$$f : H \longrightarrow \mathbb{R}, \quad f(u) := \|u\|^2. \tag{6.88}$$

Claim 1. The map f is Fréchet differentiable, and, for each $(u,h) \in H^2$,

$$f'_{\mathrm{F}}(u)(h) = \langle 2u, h\rangle. \tag{6.89}$$

Proof. Fixing $u \in H$ and letting

$$A : H \longrightarrow \mathbb{R}, \quad Ah := \langle 2u, h\rangle, \tag{6.90}$$

the Cauchy-Schwarz inequality shows that A is a bounded linear operator. To verify the Fréchet differentiability of f, we compute, for each $h \in U$,

$$\mathcal{F}(h) := f(u+h) - f(u) - Ah = \langle u+h, u+h\rangle - \|u\|^2 - \langle 2u, h\rangle = \|h\|^2. \tag{6.91}$$

If $h \to 0$, $h \neq 0$, then $\mathcal{F}(h)/\|h\| = \|h\| \to 0$, establishing the Fréchet differentiability of f as well as $A = f'_{\mathrm{F}}(u)$. ▲

(f) We apply (e) to $H = L^2(E)$ for some measurable set $E \subseteq \mathbb{R}^m$. Then

$$f : L^2(E) \longrightarrow \mathbb{R}, \quad f(u) := \|u\|^2 = \int_E u^2 \, \mathrm{d}\lambda_m \,. \tag{6.92}$$

From (e), we know that f is Fréchet differentiable with

$$f'_{\mathrm{F}}(u) : L^2(E) \longrightarrow \mathbb{R}, \quad f'_{\mathrm{F}}(u)(h) = \langle 2u, h \rangle = 2 \int_E u\, h \, \mathrm{d}\lambda_m \,. \tag{6.93}$$

(g) Fix $(a, b) \in \mathbb{R}^2$ such that $a < b$, and $\epsilon \in]0, 1[$. Given a measurable function $u : [a, b] \longrightarrow \mathbb{R}$ such that $u(x) \geq \epsilon$ for almost every $x \in [0, 1]$, let

$$A := \big\{ x \in [a, b] : u(x) \leq 1 \big\}, \quad B := \big\{ x \in [a, b] : u(x) > 1 \big\}. \tag{6.94}$$

Thus, as $\ln(x) < x$ for each $x \in \mathbb{R}^+$,

$$\int_a^b \big| \ln\big(u(x)\big) \big| \, \mathrm{d}x \, \leq |\ln(\epsilon)| \, \lambda_1(A) + \int_B u(x) \, \mathrm{d}x \,. \tag{6.95}$$

Letting, for $p \in [1, \infty]$,

$$L^p_\epsilon := \big\{ u \in L^p[a, b] : u \geq \epsilon \text{ a.e. on } [a, b] \big\}, \tag{6.96}$$

(6.95) shows that

$$f_p : L^p_\epsilon \longrightarrow \mathbb{R}, \quad f_p(u) := \int_a^b \ln\big(u(x)\big) \, \mathrm{d}x \,, \tag{6.97}$$

is well-defined for each $p \in [1, \infty]$. Consider $u_0 \equiv 1$. Then $u_0 \in L^p_\epsilon$ since $\epsilon < 1$.

Claim 1. For each $1 \leq p < \infty$, f_p does not have a first variation at $u_0 \equiv 1$.

Proof. For each $1 \leq p < \infty$, we have $L^p[a, b] \not\subseteq L^\infty[a, b]$, i.e. there is $h_p \in L^p[a, b]$ such that h_p is unbounded from below. Then, for each $t > 0$, $t h_p$ is still unbounded from below such that $u_0 + t\, h_p \notin L^p_\epsilon$ for every $t > 0$, showing, in particular, that f_p does not have a first variation at u_0. ▲

Claim 2. For each $p \in [1, \infty]$ and each $h \in L^\infty[a, b]$, f_p has a directional derivative at $u_0 \equiv 1$ in direction h. Moreover, f_∞ is Fréchet differentiable at u_0 with

$$(f_\infty)'_{\mathrm{F}}(u_0) : L^\infty[a, b] \longrightarrow \mathbb{R}, \quad (f_\infty)'_{\mathrm{F}}(u_0)(h) = \int_a^b h(x) \, \mathrm{d}x \,. \tag{6.98}$$

Proof. For $h = 0$, there is nothing to show. Thus, let $0 \neq h \in L^\infty[a, b]$. First, note that, for each $0 < t < (1 - \epsilon)/\|h\|_\infty$, one has, for almost every $x \in [a, b]$,

$$|t\, h(x)| \leq \frac{(1 - \epsilon)\, |h(x)|}{\|h\|_\infty} \leq 1 - \epsilon, \tag{6.99}$$

such that $u_0 + th \in L^p_\epsilon$. Moreover, noting $f_p(u_0) = 0$, for each such t,

$$\delta(t) := \frac{1}{t}\big(f_p(u_0 + th) - f_p(u_0)\big) = \frac{1}{t}\int_a^b \ln\big(1 + t\,h(x)\big)\,\mathrm{d}x\,. \tag{6.100}$$

As (6.99) implies $|t\,h(x)| \leq 1 - \epsilon < 1$, the Taylor expansion formula yields, for almost every $x \in [a,b]$, the existence of $-t\,|h(x)| < \xi(t,x) < t\,|h(x)|$ such that

$$\ln\big(1 + t\,h(x)\big) = t\,h(x) - \frac{t^2\,\big(h(x)\big)^2}{2\,\big(1 + \xi(t,x)\big)^2}. \tag{6.101}$$

Thus,

$$\left|\delta(t) - \int_a^b h(x)\,\mathrm{d}x\right| = \int_a^b \frac{t\,\big(h(x)\big)^2}{2\,\big(1 + \xi(t,x)\big)^2}\,\mathrm{d}x \;\leq \frac{t\,(b-a)\,\|h\|_\infty^2}{2\,\epsilon^2}, \tag{6.102}$$

showing that

$$\delta f_p(u_0, h) = \lim_{t \downarrow 0} \delta(t) = \int_a^b h(x)\,\mathrm{d}x\,. \tag{6.103}$$

Since $\delta f_p(u_0,h)$ is linear in $h \in L^\infty[a,b]$, and since $f_\infty(u_0,\cdot)$ is bounded due to $|\delta f_p(u_0,h)| \leq \|h\|_\infty\,(b-a)$, f_∞ is Gâteaux differentiable at u_0.

It remains to show that f_∞ also has a Fréchet derivative at $u_0 \equiv 1$, i.e.

$$h \neq 0, \quad \|h\|_\infty \to 0 \quad \Rightarrow \quad \frac{\left|\int_a^b \ln\big(1 + h(x)\big)\,\mathrm{d}x \;- \int_a^b h(x)\,\mathrm{d}x\right|}{\|h\|_\infty} \to 0. \tag{6.104}$$

Again, we apply the Taylor expansion formula. For $\|h\|_\infty < 1$, one finds that, for almost every $x \in [a,b]$, there is $-|h(x)| < \xi(x) < |h(x)|$ such that

$$\left|\int_a^b \ln\big(1 + h(x)\big)\,\mathrm{d}x \;- \int_a^b h(x)\,\mathrm{d}x\right| = \int_a^b \frac{\big(h(x)\big)^2}{2\,\big(1 + \xi(x)\big)^2}\,\mathrm{d}x \;\leq \frac{(b-a)\,\|h\|_\infty^2}{2\,(1 - \|h\|_\infty)^2}, \tag{6.105}$$

thereby proving the validity of (6.104) and the Fréchet differentiability of f_∞. ▲

Lemma 6.58. *Let U, V be normed vector spaces.*

(a) *Let $U_{\mathrm{ad}} \subseteq U$ be open, $u \in U_{\mathrm{ad}}$. If $\alpha \in \mathbb{R}$ and $f : U_{\mathrm{ad}} \longrightarrow V$ has a directional derivative in direction $h \in U$ (resp. a first variation, a Gâteaux derivative, or a Fréchet derivative) at u, then αf has a directional derivative in direction h (resp. a first variation, a Gâteaux derivative, or a Fréchet derivative) at u, and it holds that*

$$\delta(\alpha f)(u,h) = \alpha\,\delta f(u,h), \tag{6.106a}$$
$$(\alpha f)'_{\mathrm{G}}(u) = \alpha\,f'_{\mathrm{G}}(u), \tag{6.106b}$$
$$(\alpha f)'_{\mathrm{F}}(u) = \alpha\,f'_{\mathrm{F}}(u), \tag{6.106c}$$

respectively.

(b) *Let $U_{\mathrm{ad}} \subseteq U$ be open, $u \in U_{\mathrm{ad}}$. If $f : U_{\mathrm{ad}} \longrightarrow V$ and $g : U_{\mathrm{ad}} \longrightarrow V$ both have a directional derivative in direction $h \in U$ (resp. a first variation, a Gâteaux derivative, or a Fréchet derivative) at u, then $f + g$ has a directional derivative in direction h (resp. a first variation, a Gâteaux derivative, or a Fréchet derivative) at u, and it holds that*

$$\delta(f+g)(u,h) = \delta f(u,h) + \delta g(u,h), \tag{6.107a}$$
$$(f+g)'_{\mathrm{G}}(u) = f'_{\mathrm{G}}(u) + g'_{\mathrm{G}}(u), \tag{6.107b}$$
$$(f+g)'_{\mathrm{F}}(u) = f'_{\mathrm{F}}(u) + g'_{\mathrm{F}}(u), \tag{6.107c}$$

respectively.

(c) *The* chain rule *holds for Fréchet differentiable maps: Let U, V, Z be normed vector spaces, let $U_{\mathrm{ad}} \subseteq U$ and $V_{\mathrm{ad}} \subseteq V$ be open, $f : U_{\mathrm{ad}} \longrightarrow V$, $g : V_{\mathrm{ad}} \longrightarrow Z$, $f(U_{\mathrm{ad}}) \subseteq V_{\mathrm{ad}}$. If f is Fréchet differentiable at $u \in U_{\mathrm{ad}}$, and g is Fréchet differentiable at $f(u)$, then $g \circ f$ is Fréchet differentiable at u and*

$$(g \circ f)'_{\mathrm{F}}(u) = g'_{\mathrm{F}}\big(f(u)\big) \circ f'_{\mathrm{F}}(u). \tag{6.108}$$

Proof. (a): If the directional derivative $\delta f(u,h)$ exists, then

$$\begin{aligned}\delta(\alpha f)(u,h) &= \lim_{t \downarrow 0} \frac{1}{t} \big(\alpha\, f\big(u + t\,h\big) - \alpha\, f(u)\big) \\ &= \alpha \lim_{t \downarrow 0} \frac{1}{t} \big(f\big(u + t\,h\big) - f(u)\big) = \alpha\, \delta f(u,h),\end{aligned} \tag{6.109}$$

proving (6.106a) and (6.106b). Moreover, if f is Fréchet differentiable, then, for $h \neq 0$, $\|h\|_U \to 0$,

$$\frac{\big\|\alpha f(u+h) - \alpha f(u) - \alpha f'_{\mathrm{F}}(u)(h)\big\|_V}{\|h\|_U} = |\alpha| \, \frac{\big\|f(u+h) - f(u) - f'_{\mathrm{F}}(u)(h)\big\|_V}{\|h\|_U} \to 0, \tag{6.110}$$

showing that the Fréchet differentiability of f implies that of αf as well as (6.106c).

(b): If the directional derivatives $\delta f(u,h)$ and $\delta g(u,h)$ exist, then

$$\begin{aligned}\delta(f+g)(u,h) &= \lim_{t \downarrow 0} \frac{1}{t} \big((f+g)\big(u + t\,h\big) - (f+g)(u)\big) \\ &= \lim_{t \downarrow 0} \frac{1}{t} \big(f\big(u + t\,h\big) - f(u)\big) + \lim_{t \downarrow 0} \frac{1}{t} \big(g\big(u + t\,h\big) - g(u)\big) \\ &= \delta(f+g)(u,h),\end{aligned} \tag{6.111}$$

proving (6.107a) and (6.107b). Moreover, if f is Fréchet differentiable, then, for $h \neq 0$, $\|h\|_U \to 0$,

$$\begin{aligned}&\frac{\big\|(f+g)(u+h) - (f+g)(u) - (f'_{\mathrm{F}}(u) + g'_{\mathrm{F}}(u))(h)\big\|_V}{\|h\|_U} \\ &\leq \frac{\big\|f(u+h) - f(u) - f'_{\mathrm{F}}(u)(h)\big\|_V}{\|h\|_U} + \frac{\big\|g(u+h) - g(u) - g'_{\mathrm{F}}(u)(h)\big\|_V}{\|h\|_U} \to 0,\end{aligned} \tag{6.112}$$

showing that the Fréchet differentiability of f and g implies that of $f+g$ as well as (6.107c).

(c): Define

$$r_f(h) := f(u+h) - f(u) - f'_{\mathrm{F}}(u)(h), \tag{6.113a}$$

$$r_g(h) := g\big(f(u)+h\big) - g\big(f(u)\big) - g'_{\mathrm{F}}\big(f(u)\big)(h), \tag{6.113b}$$

$$r_{g\circ f}(h) := (g\circ f)(u+h) - (g\circ f)(u) - \Big(g'_{\mathrm{F}}\big(f(u)\big)\circ f'_{\mathrm{F}}(u)\Big)(h), \tag{6.113c}$$

where $r_f(h)$ and $r_{g\circ f}(h)$ are defined for each h such that $u+h \in U_{\mathrm{ad}}$, and $r_g(h)$ is defined for each h such that $f(u)+h \in V_{\mathrm{ad}}$. Then the Fréchet differentiability of f at u and of g at $f(u)$ imply

$$h \neq 0, \quad \|h\|_U \to 0 \quad \Rightarrow \quad \|r_f(h)\|_V/\|h\|_U \to 0, \tag{6.114a}$$

$$h \neq 0, \quad \|h\|_V \to 0 \quad \Rightarrow \quad \|r_g(h)\|_Z/\|h\|_V \to 0, \tag{6.114b}$$

whereas, one needs to show

$$h \neq 0, \quad \|h\|_U \to 0 \quad \Rightarrow \quad \|r_{g\circ f}(h)\|_V/\|h\|_U \to 0. \tag{6.114c}$$

Since $f'_{\mathrm{F}}(u)$ is bounded and $h \to 0$ implies $r_f(h) \to 0$, for each sufficiently small $h \in U$, one has $f(u) + f'_{\mathrm{F}}(u)(h) + r_f(h) \in V_{\mathrm{ad}}$. In the following, we only consider sufficiently small $h \in U$ such that $f(u) + f'_{\mathrm{F}}(u)(h) + r_f(h) \in V_{\mathrm{ad}}$ and, simultaneously, $u+h \in U_{\mathrm{ad}}$. Then

$$\begin{aligned}(g\circ f)(u+h) &= g\Big(f(u) + f'_{\mathrm{F}}(u)(h) + r_f(h)\Big)\\ &= g\big(f(u)\big) + g'_{\mathrm{F}}\big(f(u)\big)\Big(f'_{\mathrm{F}}(u)(h) + r_f(h)\Big) + r_g\Big(f'_{\mathrm{F}}(u)(h) + r_f(h)\Big),\end{aligned} \tag{6.115}$$

implying

$$r_{g\circ f}(h) = g'_{\mathrm{F}}\big(f(u)\big)\Big(r_f(h)\Big) + r_g\Big(f'_{\mathrm{F}}(u)(h) + r_f(h)\Big). \tag{6.116}$$

Next, one notes that, for $h \neq 0$, $\|h\|_U \to 0$,

$$\frac{\Big\|g'_{\mathrm{F}}\big(f(u)\big)\Big(r_f(h)\Big)\Big\|_Z}{\|h\|_U} \leq \frac{\big\|g'_{\mathrm{F}}\big(f(u)\big)\big\|\,\big\|r_f(h)\big\|_V}{\|h\|_U} \to 0 \quad \text{as} \quad \frac{\big\|r_f(h)\big\|_V}{\|h\|_U} \to 0. \tag{6.117}$$

Thus, it merely remains to show that, for $h \neq 0$, $\|h\|_U \to 0$,

$$\frac{\Big\|r_g\Big(f'_{\mathrm{F}}(u)(h) + r_f(h)\Big)\Big\|_Z}{\|h\|_U} \to 0. \tag{6.118}$$

To that end, for $0 \neq \eta \in V$ such that $f(u)+\eta \in V$, define

$$s(\eta) := r_g(\eta)/\|\eta\|_V. \tag{6.119}$$

This allows to rewrite the left-hand side (6.118) as

$$\frac{\Big\|r_g\Big(f'_{\mathrm{F}}(u)(h)+r_f(h)\Big)\Big\|_Z}{\|h\|_U} = \frac{\big\|f'_{\mathrm{F}}(u)(h)+r_f(h)\big\|_V \Big\|s\Big(f'_{\mathrm{F}}(u)(h)+r_f(h)\Big)\Big\|_Z}{\|h\|_U}. \quad (6.120)$$

Now, if $h \neq 0$, $\|h\|_U \to 0$, then $y := \Big(f'_{\mathrm{F}}(u)(h)+r_f(h)\Big) \to 0$ such that (6.119) and (6.114b) imply $\|s(y)\|_Z \to 0$. Finally, note that

$$\big\|f'_{\mathrm{F}}(u)(h)+r_f(h)\big\|_V/\|h\|_U \leq \big\|f'_{\mathrm{F}}(u)\big\| + \big\|r_f(h)\big\|_V/\|h\|_U, \quad (6.121)$$

which remains bounded for $\|h\|_U \to 0$, showing that both sides of (6.120) converge to 0 for $\|h\|_U \to 0$, thereby proving (6.118) and part (c) of the lemma. ■

Example 6.59. The following examples shows that, for the chain rule of Lem. 6.58(c) to hold, it does *not* suffice for f to be Fréchet differentiable at u and g merely Gâteaux differentiable at $f(u)$.

(a) Let

$$f : \mathbb{R} \longrightarrow \mathbb{R}^2, \quad f(x) := (x, x^2). \quad (6.122a)$$

Then f is clearly Fréchet differentiable (even C^∞) on $\mathbb{R}$. If g is taken to be the function f from Ex. 6.55(a), then g is Gâteaux differentiable at $(0,0) = f(0)$, however,

$$g \circ f : \mathbb{R} \longrightarrow \mathbb{R}, \quad (g \circ f)(x) = g(x, x^2) = \begin{cases} 0 & \text{for } x \leq 0, \\ 1 & \text{for } x > 0, \end{cases} \quad (6.122b)$$

is *not* Gâteaux differentiable at 0.

(b) By modifying g in (a), we can get $g \circ f$ to be an arbitrary function $h : \mathbb{R} \longrightarrow \mathbb{R}$: Define

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad g(x,y) := \begin{cases} h(x) & \text{if } y = x^2, \\ h(0) & \text{otherwise.} \end{cases} \quad (6.123)$$

With the obvious minor modifications, the argument from Ex. 6.55(a) still shows that g is Gâteaux differentiable at $(0,0)$ with $g'_{\mathrm{G}}(0,0) \equiv 0$. Of course, h might or might not be Gâteaux differentiable at 0 (it can be as irregular as one wants it to be, e.g. nonmeasurable).

(c) By modifying f to

$$f : \mathbb{R} \longrightarrow \mathbb{R}^2, \quad f(x) := (x, x^3), \quad (6.124)$$

and setting g to f from Ex. 6.55(b), we can even get g to be everywhere Gâteaux differentiable, while $g \circ f$ is still not Gâteaux differentiable at 0 (still not even continuous at 0).

Example 6.60. (a) Let $(a,b) \in \mathbb{R}^2$, $a < b$. We consider $U := C[a,b]$ endowed with the $\|\cdot\|_\infty$-norm, and investigate point functionals composed with a differentiable map g. More precisely, let $x \in [a,b]$, let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be differentiable (in the classical sense), and consider

$$g_x : C[a,b] \longrightarrow \mathbb{R}, \quad g_x(u) := g\big(u(x)\big). \tag{6.125}$$

We know, according to Ex. 6.57(c), that g is Fréchet differentiable with $g'_{\mathrm{F}}\big(u\big)(h) = h\, g'\big(u\big)$. We also know, according to Ex. 6.57(d), that $f_x : C[a,b] \longrightarrow \mathbb{R}$, $f_x(u) := u(x)$, is Fréchet differentiable with $(f_x)'_{\mathrm{F}}(u)(h) = h(x)$. Since $g_x = g \circ f_x$, we can apply the chain rule of Lem. 6.58(c) to conclude that g_x is Fréchet differentiable and, for each $u \in C[a,b]$,

$$(g_x)'_{\mathrm{F}}(u) : C[a,b] \longrightarrow \mathbb{R}, \quad (g_x)'_{\mathrm{F}}(u)(h) = g'\big(u(x)\big)\, h(x). \tag{6.126}$$

(b) We would now like to apply the chain rule to differentiate our favorite objective functional. We reformulate it in a general Hilbert space setting: Let U and H be Hilbert spaces, let $S : U \longrightarrow H$ be a bounded linear operator, $y_0 \in H$, $\lambda \in \mathbb{R}_0^+$, and consider

$$f : U \longrightarrow \mathbb{R}, \quad f(u) := \frac{1}{2}\,\|Su - y_0\|_H^2 + \frac{\lambda}{2}\,\|u\|_U^2. \tag{6.127}$$

Note that $f = f_1 + f_2$, where

$$f_1 : U \longrightarrow \mathbb{R}, \quad f_1(u) := \frac{1}{2}\,\|Su - y_0\|_H^2, \tag{6.128a}$$

$$f_2 : U \longrightarrow \mathbb{R}, \quad f_2(u) := \frac{\lambda}{2}\,\|u\|_U^2. \tag{6.128b}$$

We decompose f_1 even further according to $f_1 = f_{12} \circ f_{11}$, where

$$f_{11} : U \longrightarrow H, \quad f_{11}(u) := Su - y_0, \tag{6.129a}$$

$$f_{12} : H \longrightarrow \mathbb{R}, \quad f_{12}(y) := \|y\|_H^2/2. \tag{6.129b}$$

First, observe that, according to Ex. 6.57(a),(b) and Lem. 6.58(b), f_{11} is Fréchet differentiable and, for each $u \in U$,

$$(f_{11})'_{\mathrm{F}}(u) : U \longrightarrow H, \quad (f_{11})'_{\mathrm{F}}(u)(h) = Sh. \tag{6.130a}$$

Second, according to Ex. 6.57(e) and Lem. 6.58(a), f_{12} is Fréchet differentiable and, for each $y \in H$,

$$(f_{12})'_{\mathrm{F}}(y) : H \longrightarrow \mathbb{R}, \quad (f_{12})'_{\mathrm{F}}(y)(h) = \langle y, h\rangle_H. \tag{6.130b}$$

Third, applying the chain rule of Lem. 6.58(c) yields that f_1 is Fréchet differentiable and, for each $u \in U$,

$$\begin{aligned}(f_1)'_{\mathrm{F}}(u) : U \longrightarrow \mathbb{R}, \quad (f_1)'_{\mathrm{F}}(u)(h) &= \langle Su - y_0,\, Sh\rangle_H = \langle S^*(Su - y_0),\, h\rangle_U \\ &= \langle Su, Sh\rangle_H - \langle y_0, Sh\rangle_H,\end{aligned} \tag{6.130c}$$

where S^* is the (Hilbert) adjoint operator of S (see Sec. 4.5). Fourth, according to Ex. 6.57(e) and Lem. 6.58(a), f_2 is Fréchet differentiable and, for each $u \in U$,

$$(f_2)'_{\mathrm{F}}(u) : U \longrightarrow \mathbb{R}, \quad (f_2)'_{\mathrm{F}}(u)(h) = \lambda \langle u, h \rangle_U. \tag{6.130d}$$

Fifth and last, according to Lem. 6.58(b), f is Fréchet differentiable and, for each $u \in U$,

$$\begin{aligned} f'_{\mathrm{F}}(u) : U \longrightarrow \mathbb{R}, \quad f'_{\mathrm{F}}(u)(h) &= \langle Su - y_0,\, Sh \rangle_H + \lambda \langle u, h \rangle_U \\ &= \langle S^*(Su - y_0) + \lambda\, u,\, h \rangle_U \\ &= \langle Su, Sh \rangle_H - \langle y_0, Sh \rangle_H + \lambda \langle u, h \rangle_U. \end{aligned} \tag{6.130e}$$

6.4.2 Variational Inequality, Adjoint State

The key observation is that the proof of Th. 3.10 still works, basically without change, in the infinite-dimensional case. Thus, as in the finite-dimensional case, we get a variational inequality involving the derivative of f at $\bar{u}$ as a necessary condition for f to be minimal at $\bar{u}$. Moreover, if f is convex, then the variational inequality also turns out to be sufficient. All this is precisely stated and proved in Th. 6.62 below. For the sufficiency part, we will use the following Lem. 6.61.

Lemma 6.61. *Let U be a normed vector space, $U_{\mathrm{ad}} \subseteq U$, $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$. Consider $(\bar{u}, u) \in U_{\mathrm{ad}}^2$ such that*

$$[\bar{u}, u] := \mathrm{conv}\{\bar{u}, u\} := \big\{\alpha\, \bar{u} + (1 - \alpha)\, u : \alpha \in [0, 1]\big\} \subseteq U_{\mathrm{ad}}. \tag{6.131}$$

If f is convex on $[\bar{u}, u]$ and the directional derivative of f at $\bar{u}$ in the direction $u - \bar{u}$, i.e. $\delta f(\bar{u}, u - \bar{u})$, exists, then

$$\delta f(\bar{u}, u - \bar{u}) \leq f(u) - f(\bar{u}). \tag{6.132}$$

In particular, if U_{ad} is convex and f is convex on U_{ad}, then (6.132) *holds whenever $\delta f(\bar{u}, u - \bar{u})$ exists.*

Proof. Let $t \in]0, 1]$. Then the convexity of f on $[\bar{u}, u]$ yields

$$\begin{aligned} f\big(\bar{u} + t(u - \bar{u})\big) - f(\bar{u}) &= f\big(tu + (1 - t)\bar{u}\big) - f(\bar{u}) \\ &\leq t f(u) + (1 - t) f(\bar{u}) - f(\bar{u}) = t\,\big(f(u) - f(\bar{u})\big). \end{aligned} \tag{6.133}$$

Dividing by t in (6.133) and taking the limit $t \downarrow 0$ yields the claimed relation (6.132). ■

Theorem 6.62. *Let U be a normed vector space, $U_{\mathrm{ad}} \subseteq U$, and assume that $\bar{u} \in U_{\mathrm{ad}}$ minimizes the function $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$, i.e.*

$$f(\bar{u}) \leq f(u) \quad \textit{for each } u \in U_{\mathrm{ad}}. \tag{6.134}$$

Consider $u \in U_{\mathrm{ad}}$. If $\bar{u}+t\,(u-\bar{u}) \in U_{\mathrm{ad}}$ for each sufficiently small $t>0$, and, moreover, the directional derivative

$$\delta f(\bar{u}, u-\bar{u}) = \lim_{t \downarrow 0} \frac{1}{t}\Big(f\big(\bar{u}+t\,(u-\bar{u})\big) - f(\bar{u})\Big) \tag{6.135}$$

exists in $\mathbb{R}$, then $\bar{u}$ satisfies the variational inequality

$$\delta f(\bar{u}, u-\bar{u}) \;\geq\; 0. \tag{6.136}$$

If, in addition, U_{ad} is convex, f is convex, and $\delta f(\bar{u}, u-\bar{u})$ exists for each $u \in U_{\mathrm{ad}}$, then the validity of (6.136) for each $u \in U_{\mathrm{ad}}$ is both necessary and sufficient for $\bar{u}$ to minimize f.

Proof. Since $\bar{u}+t\,(u-\bar{u}) \in U_{\mathrm{ad}}$ for each sufficiently small $t>0$, there exists $\varepsilon>0$ such that

$$\bar{u}+t(u-\bar{u}) = (1-t)\bar{u}+tu \in U_{\mathrm{ad}}, \quad \text{for each} \quad t \in]0,\varepsilon]. \tag{6.137}$$

By hypothesis, $\bar{u}$ satisfies (6.134), implying, for each $t \in]0,\varepsilon]$:

$$\frac{1}{t}\Big(f\big(\bar{u}+t(u-\bar{u})\big) - f(\bar{u})\Big) \overset{(6.134)}{\geq} 0. \tag{6.138}$$

Thus, taking the limit for $t \downarrow 0$, (6.138) implies (6.136).

If $\delta f(\bar{u}, u-\bar{u})$ exists for each $u \in U_{\mathrm{ad}}$, then the above considerations show that (6.134) implies (6.136) for each $u \in U_{\mathrm{ad}}$. Conversely, if $\delta f(\bar{u}, u-\bar{u})$ exists for each $u \in U_{\mathrm{ad}}$, U_{ad} and f are convex, and (6.136) holds for each $u \in U_{\mathrm{ad}}$, then

$$0 \overset{(6.136)}{\leq} \delta f(\bar{u}, u-\bar{u}) \overset{(6.132)}{\leq} f(u) - f(\bar{u}), \tag{6.139}$$

thereby establishing the validity of (6.134). ∎

As in the finite-dimensional case (cf. Cor. 3.13), at interior points $\bar{u}$, we can strengthen the variational inequality to a variational equality:

Corollary 6.63. *Let U be a normed vector space and assume that $\bar{u} \in U_{\mathrm{ad}}$ lies in the interior of $U_{\mathrm{ad}} \subseteq U$. If $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$ is Gâteaux differentiable at $\bar{u}$ and f is minimal at $\bar{u}$, then $f'_{\mathrm{G}}(\bar{u}) = 0$. Special cases include $U_{\mathrm{ad}} = U$ (no control constraints) and any other case, where U_{ad} is open.*

Proof. The proof is essentially the same as that of Cor. 3.13: If $\bar{u}$ lies in the interior of U_{ad}, then there is a (convex) ball B with center $\bar{u}$ such that $B \subseteq U_{\mathrm{ad}}$. Then (6.136) implies

$$f'_{\mathrm{G}}(\bar{u})(u-\bar{u}) = \delta f(\bar{u}, u-\bar{u}) \;\geq\; 0 \quad \text{for each } u \in B. \tag{6.140}$$

Let $v \in U$ be arbitrary. If $\epsilon>0$ is sufficiently small, then $\bar{u} \pm \epsilon v \in B$, implying $f'_{\mathrm{G}}(\bar{u})\,(\bar{u} \pm \epsilon v - \bar{u}) = f'_{\mathrm{G}}(\bar{u})\,(\pm \epsilon v) \geq 0$. Thus, using the linearity of $f'_{\mathrm{G}}(\bar{u})$, $f'_{\mathrm{G}}(\bar{u}) = 0$ as claimed. ∎

Example 6.64. (a) We reconsider the situation from Ex. 6.60(b), i.e. Hilbert spaces U and H, $S : U \longrightarrow H$ linear and bounded, $y_0 \in H$, $\lambda \in \mathbb{R}_0^+$, and

$$f : U \longrightarrow \mathbb{R}, \quad f(u) := \frac{1}{2} \|Su - y_0\|_H^2 + \frac{\lambda}{2} \|u\|_U^2. \tag{6.141}$$

In Ex. 6.47, we remarked that f is convex, and in Ex. 6.60(b), we determined that f is Fréchet differentiable with Fréchet derivative according to (6.130e). Thus, if U_{ad} is a convex subset of U, then we obtain from Th. 6.62 together with (6.130e) that $\bar{u} \in U_{\mathrm{ad}}$ minimizes f in U_{ad} if, and only if, f satisfies the corresponding variational inequality. More precisely, the following statements (6.142a) – (6.142c) are equivalent:

$$f(\bar{u}) \leq f(u) \quad \text{for each } u \in U_{\mathrm{ad}}, \tag{6.142a}$$

$$\langle S\bar{u} - y_0,\, S(u - \bar{u})\rangle_H + \lambda \langle \bar{u}, u - \bar{u}\rangle_U \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}}, \tag{6.142b}$$

$$\langle S^*(S\bar{u} - y_0) + \lambda\, \bar{u},\, u - \bar{u}\rangle_U \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}}. \tag{6.142c}$$

(b) Let us now specialize (a) to the EOCP of Def. 6.45. In particular, $H = U = L^2(\Omega)$, and $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ is the solution operator of the corresponding elliptic BVP on Ω with homogeneous Dirichlet condition on the boundary. We consider the EOCP with the additional assumption that $U_{\mathrm{ad}} \subseteq L^2(\Omega)$ is convex and that $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$, $f(u) = J(Su, u)$, coincides with f as given by (6.141), where, as before, we choose to write y_Ω instead of y_0 in the context of the EOCP. Then $\bar{u} \in U_{\mathrm{ad}}$ is an optimal control for the EOCP if, and only if, there exists $(\bar{y}, \bar{p}) \in H_0^1(\Omega) \times H_0^1(\Omega)$ such that the triple $(\bar{u}, \bar{y}, \bar{p}) \in U_{\mathrm{ad}} \times H_0^1(\Omega) \times H_0^1(\Omega)$ satisfies the following *system of optimality* (6.143):

$$\bar{u} \in U_{\mathrm{ad}}, \tag{6.143a}$$

$$\bar{y} = S\bar{u}, \tag{6.143b}$$

$$\bar{p} = S^*(\bar{y} - y_\Omega), \tag{6.143c}$$

$$\int_\Omega (\bar{p} + \lambda\, \bar{u})\, (u - \bar{u}) \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}}, \tag{6.143d}$$

where (6.143c) and (6.143d) can be combined into the following equivalent, albeit S^*-free, form:

$$\int_\Omega (\bar{y} - y_\Omega)(Su - \bar{y}) + \lambda \int_\Omega \bar{u}\, (u - \bar{u}) \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}}. \tag{6.144}$$

The function $\bar{p}$ determined by (6.143c) is called the *adjoint state* corresponding to $\bar{u}$ and $\bar{y}$. To make proper use of (6.143c) in this context, we need to determine S^* more explicitly, which is related to the topic of the following section.

6.4.3 Adjoint Equation

As it turns out, the adjoint S^* of the solution operator $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ of Th. 6.43 can itself be interpreted as the solution operator of a linear elliptic BVP with

homogeneous Dirichlet boundary conditions. Moreover, to obtain the BVP for S^*, one only needs to transpose the matrix (a_{ij}). In particular, S is *self-adjoint*, i.e. $S = S^*$, if (a_{ij}) is symmetric. This is the contents of the following Prop. 6.65. The situation of Th. 6.43 (i.e. the class of PDE under consideration) is still particularly simple. In general, the relation between S and S^* can be much more complicated, and, even for symmetric (a_{ij}), one can not expect the solution operator of a PDE to be self-adjoint.

Proposition 6.65. *Let $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ be the solution operator of Th. 6.43. The precise setting is as in Def. 6.42 (with $h_i = 0$): Let the bounded and open set $\Omega \subseteq \mathbb{R}^m$ be a set with Lipschitz boundary, $m \geq 2$, let $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$, $a_{ij} \in L^\infty(\Omega)$, be almost uniformly elliptic, and $b \in L^\infty(\Omega)$, $b \geq 0$ almost everywhere. The adjoint operator $S^* : L^2(\Omega) \longrightarrow L^2(\Omega)$ to S is the solution operator for the elliptic BVP with homogeneous Dirichlet conditions on the boundary, where $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$ has been replaced by its transpose, i.e. by $(a_{ji})_{(i,j)\in\{1,\dots,m\}^2}$. In particular, if $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$ is symmetric, then S is self-adjoint, i.e. $S = S^*$.*

Proof. Let $B : L^2(\Omega) \longrightarrow L^2(\Omega)$ denote the solution operator for the elliptic BVP with homogeneous Dirichlet conditions on the boundary, where $(a_{ij})_{(i,j)\in\{1,\dots,m\}^2}$ has been replaced by its transpose. We need to show $B = S^*$, i.e. that B is the adjoint of S. Here, when we speak of the adjoint operator of S, more precisely, we mean the Hilbert adjoint in the sense of Sec. 4.5. Thus, following Prop. 4.36 and letting $\langle \cdot, \cdot \rangle$ denote the inner product in $L^2(\Omega)$, we need to show that

$$\langle y, Su \rangle = \langle By, u \rangle \quad \text{for each} \quad (u, y) \in L^2(\Omega) \times L^2(\Omega). \tag{6.145}$$

According to the definition of S in Th. 6.43, given $(u, y) \in L^2(\Omega) \times L^2(\Omega)$, Su and By are the unique respective solutions to the following elliptic BVP in weak form with a homogeneous Dirichlet condition on the boundary, namely (6.146) and (6.147). Here, as is customary, for the sake of better readability, we write both BVP in strong form, which, here, is nothing more than notational candy symbolizing the weak form:

$$-\sum_{i=1}^{m}\sum_{j=1}^{m} \partial_i\big(a_{ij}\partial_j(Su)\big) + b\,Su = u \quad \text{on } \Omega, \tag{6.146a}$$

$$Su = 0 \quad \text{on } \partial\Omega, \tag{6.146b}$$

$$-\sum_{i=1}^{m}\sum_{j=1}^{m} \partial_i\big(a_{ji}\partial_j(By)\big) + b\,By = y \quad \text{on } \Omega, \tag{6.147a}$$

$$By = 0 \quad \text{on } \partial\Omega. \tag{6.147b}$$

If $Su \in H_0^1(\Omega)$ is the weak solution to (6.146), then (6.40) (with $h_i = 0$ y replaced by Su, and g replaced by u) holds for each $v \in H_0^1(\Omega)$. Choosing $v = By$ yields

$$\int_\Omega \left(\sum_{i=1}^{m} (\partial_i(By)) \sum_{j=1}^{m} a_{ij}\partial_j(Su) + By\,(b\,Su - u) \right) \mathrm{d}\lambda_m = 0. \tag{6.148a}$$

Analogously, if $By \in H_0^1(\Omega)$ is the weak solution to (6.147), then (6.40) (with $h_i = 0$, y replaced by By, and g replaced by y) holds for each $v \in H_0^1(\Omega)$. Choosing $v = Su$ yields

$$\int_\Omega \left(\sum_{i=1}^m (\partial_i (Su)) \sum_{j=1}^m a_{ji} \partial_j (By) + Su\,(b\,By - y) \right) \mathrm{d}\lambda_m = 0. \tag{6.148b}$$

Subtracting (6.148b) from (6.148a), one obtains

$$\int_\Omega y\,Su\,\mathrm{d}\lambda_m = \int_\Omega u\,By\,\mathrm{d}\lambda_m\,, \tag{6.149}$$

which is exactly (6.145), i.e. $B = S^*$ is verified. ■

Example 6.66. As in Ex. 6.64(b), we consider the EOCP of Def. 6.45 with U_{ad} convex and $f : U_{\mathrm{ad}} \longrightarrow \mathbb{R}$ given by (6.141). We can now use Prop. 6.65 to reformulate the system of optimality (6.143), explicitly stating the PDE which determine $\bar{y}$ and $\bar{p}$: Under the stated hypotheses, $\bar{u} \in U_{\mathrm{ad}}$ is an optimal control for the EOCP if, and only if, there exists $(\bar{y}, \bar{p}) \in H_0^1(\Omega) \times H_0^1(\Omega)$ such that $(\bar{u}, \bar{y}, \bar{p}) \in U_{\mathrm{ad}} \times H_0^1(\Omega) \times H_0^1(\Omega)$ satisfies

$$\bar{u} \in U_{\mathrm{ad}}, \tag{6.150a}$$

$$-\sum_{i=1}^m \sum_{j=1}^m \partial_i \big(a_{ij} \partial_j \bar{y}\big) + b\,\bar{y} = \bar{u} \quad \text{on } \Omega, \tag{6.150b}$$

$$\bar{y} = 0 \quad \text{on } \partial\Omega, \tag{6.150c}$$

$$-\sum_{i=1}^m \sum_{j=1}^m \partial_i \big(a_{ji} \partial_j \bar{p}\big) + b\,\bar{p} = \bar{y} - y_\Omega \quad \text{on } \Omega, \tag{6.150d}$$

$$\bar{p} = 0 \quad \text{on } \partial\Omega, \tag{6.150e}$$

$$\int_\Omega (\bar{p} + \lambda\,\bar{u})\,(u - \bar{u}) \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}}, \tag{6.150f}$$

where, again, (6.150b) – (6.150e) are supposed to mean that $\bar{y}$ and $\bar{p}$ are *weak* solutions of the respective BVP. The BVP for the adjoint state $\bar{p}$, consisting of (6.150d) and (6.150e) is called the *adjoint equation* for the problem under consideration.

6.4.4 Pointwise Formulations of the Variational Inequality

For a while, we will now focus on the EOCP of Def. 6.45 with the objective functional

$$J :\, L^2(\Omega) \times U_{\mathrm{ad}} \longrightarrow \mathbb{R}, \quad J(y,u) = \frac{1}{2}\,\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}, \tag{6.151}$$

i.e.

$$f :\, U_{\mathrm{ad}} \longrightarrow \mathbb{R}, \quad f(u) = J(Su,u) = \frac{1}{2}\,\|Su - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}, \tag{6.152}$$

where $y_\Omega \in L^2(\Omega)$, $\lambda \in \mathbb{R}_0^+$. Moreover, for simplicity, we will only consider the situation where $U_{\rm ad} = L^2_{a,b}(\Omega)$ with $a, b \in L^2(\Omega)$, $a \leq b$ (see Not. 6.48, box constraints on the control), and, alternatively, $U_{\rm ad} = L^2(\Omega)$ (no control constraints). The main goal of the present section is to formulate equivalent pointwise versions of the variational inequality (6.143d) (also see (6.150f)). The validity of the equivalences actually has nothing to do with the EOCP and the functions J and f above. Therefore, in Th. 6.68, the equivalences are stated and proved independently of the specific context of the EOCP. The EOCP is then considered as an application in Ex. 6.69.

Notation 6.67. Given $(a, b) \in \mathbb{R}^2$, $a \leq b$, the projection from $\mathbb{R}$ onto the interval $[a, b]$ is denoted by $P_{[a,b]}$, i.e.

$$P_{[a,b]} : \mathbb{R} \longrightarrow [a,b], \quad P_{[a,b]}(\alpha) := \min\big\{b, \max\{a, \alpha\}\big\} = \begin{cases} a & \text{for } \alpha < a, \\ \alpha & \text{for } \alpha \in [a,b], \\ b & \text{for } b < \alpha. \end{cases} \tag{6.153}$$

Theorem 6.68. *Let $\Omega \subseteq \mathbb{R}^m$ be measurable and bounded, and let $U_{\rm ad} \subseteq L^2(\Omega)$. Furthermore, let $g \in L^2(\Omega)$, $\bar{u} \in U_{\rm ad}$, and consider the following variational inequality:*

$$\int_\Omega g\,(u - \bar{u}) \geq 0 \quad \textit{for each } u \in U_{\rm ad}. \tag{6.154}$$

(a) *If $U_{\rm ad} = L^2_{a,b}(\Omega)$ with $a, b \in L^2(\Omega)$, $a \leq b$, then* (6.154) *is equivalent to each of the following* pointwise *conditions* (6.155a) – (6.155d) *holding for almost every $x \in \Omega$:*

$$\bar{u}(x) = \begin{cases} a(x) & \textit{for } g(x) > 0, \\ \in [a(x), b(x)] & \textit{for } g(x) = 0, \\ b(x) & \textit{for } g(x) < 0, \end{cases} \tag{6.155a}$$

$$g(x)\,\big(\xi - \bar{u}(x)\big) \geq 0 \qquad \textit{for each } \xi \in [a(x), b(x)], \tag{6.155b}$$

$$g(x)\,\bar{u}(x) \leq g(x)\,\xi \qquad \textit{for each } \xi \in [a(x), b(x)], \tag{6.155c}$$

$$\min_{\xi \in [a(x), b(x)]} g(x)\,\xi = g(x)\,\bar{u}(x). \tag{6.155d}$$

Here, (6.155b) *and* (6.155c) *constitute* pointwise *variational inequalities in $\mathbb{R}$, whereas the form* (6.155d) *is known as the* weak minimum principle. *Moreover, for each $\lambda \geq 0$, if one defines $\bar{p} := g - \lambda\,\bar{u}$ (i.e. $g = \bar{p} + \lambda\,\bar{u}$), then all the previous conditions are also equivalent to the* minimum principle

$$\min_{\xi \in [a(x), b(x)]} \left(\bar{p}(x)\xi + \frac{\lambda\,\xi^2}{2} \right) = \bar{p}(x)\bar{u}(x) + \frac{\lambda\,\bar{u}(x)^2}{2}. \tag{6.155e}$$

If $\lambda > 0$, then all the previous conditions are also equivalent to the projection formula

$$\bar{u}(x) = P_{[a(x), b(x)]}\left(-\frac{\bar{p}(x)}{\lambda} \right), \tag{6.155f}$$

where $P_{[a(x),b(x)]}$ is according to Not. 6.67.

(b) *If* $U_{\mathrm{ad}} = L^2(\Omega)$, *then* (6.154) *is equivalent to*

$$g(x) = 0 \quad \textit{for almost every } x \in \Omega. \tag{6.156}$$

Proof. (a): We start by showing the following implications: "(6.154) $\Rightarrow$ (6.155a) $\Rightarrow$ (6.155b) $\Rightarrow$ (6.154)", "(6.155b) $\Leftrightarrow$ (6.155c)", and "(6.155c) $\Leftrightarrow$ (6.155d)". Then (6.155e) and (6.155f) are considered subsequently.

"(6.154) $\Rightarrow$ (6.155a)": Proceeding by contraposition, assume there is a measurable set $E \subseteq \Omega$ such that $\lambda_m(E) > 0$, $g > 0$ on E, and $\bar{u} > a$ on E. Then there is also $\epsilon > 0$ and a measurable set $E_\epsilon \subseteq E$, $\lambda_m(E_\epsilon) > 0$, such that $g(x) \geq \epsilon$ and $\bar{u}(x) \geq a(x) + \epsilon$ for every $x \in E_\epsilon$. In that case, one can define

$$u := \begin{cases} a & \text{on } E_\epsilon, \\ \bar{u} & \text{on } \Omega \setminus E_\epsilon. \end{cases} \tag{6.157}$$

Since a and $\bar{u}$ are in $L^2_{a,b}(\Omega)$, so is u. Furthermore,

$$\int_\Omega g\,(u - \bar{u}) = \int_{E_\epsilon} g\,(a - \bar{u}) \leq -\epsilon^2\,\lambda_m(E_\epsilon) < 0, \tag{6.158}$$

showing that (6.154) fails. Similarly, if there is a measurable set $E \subseteq \Omega$ such that $\lambda_m(E) > 0$, $g < 0$ on E, and $\bar{u} < b$ on E, then there is $\epsilon > 0$ and a measurable set $E_\epsilon \subseteq E$, $\lambda_m(E_\epsilon) > 0$, such that $g(x) \leq -\epsilon$ and $\bar{u}(x) \leq b(x) - \epsilon$ for every $x \in E_\epsilon$. In that case, one can define

$$u := \begin{cases} b & \text{on } E_\epsilon, \\ \bar{u} & \text{on } \Omega \setminus E_\epsilon, \end{cases} \tag{6.159}$$

yielding

$$\int_\Omega g\,(u - \bar{u}) = \int_{E_\epsilon} g\,(b - \bar{u}) \leq -\epsilon^2\,\lambda_m(E_\epsilon) < 0, \tag{6.160}$$

again proving failure of (6.154).

"(6.155a) $\Rightarrow$ (6.155b)": Let $x \in \Omega$ such that (6.155a) holds and fix $\xi \in [a(x), b(x)]$. If $g(x) > 0$, then $g(x)\big(\xi - \bar{u}(x)\big) = g(x)\big(\xi - a(x)\big) \geq 0$ as $\xi \geq a(x)$. If $g(x) = 0$, then $g(x)\big(\xi - \bar{u}(x)\big) = 0$. If $g(x) < 0$, then $g(x)\big(\xi - \bar{u}(x)\big) = g(x)\big(\xi - b(x)\big) \geq 0$ as $\xi \leq b(x)$. Thus, (6.155b) holds in each case.

"(6.155b) $\Rightarrow$ (6.154)": If $u \in L^2_{a,b}(\Omega)$, then $u(x) \in [a(x), b(x)]$ for almost every $x \in \Omega$. If (6.155b) holds for almost every $x \in \Omega$ and $u \in L^2_{a,b}(\Omega)$, then, in consequence, $g(x)\big(u(x) - \bar{u}(x)\big) \geq 0$ for almost every $x \in \Omega$, which, in turn, implies (6.154).

(6.155b) and (6.155c) are trivially equivalent as they are merely simple algebraic rearrangements of each other.

The equivalence of (6.155c) and (6.155d) is also immediate due to the definition of the minimum.

We now let $\lambda \geq 0$ and consider $\bar{p} := g - \lambda\,\bar{u}$, i.e. $g = \bar{p} + \lambda\,\bar{u}$. We need to show that the previous conditions are equivalent to the minimum principle (6.155e). If $\lambda = 0$, then

(6.155e) and (6.155d) are identical and there is nothing to show. Thus, let $\lambda > 0$. Then, for each $x \in \Omega$ such that $a(x) \leq b(x)$, one can apply Th. 6.62 to the minimization of the convex function $\gamma : [a(x), b(x)] \longrightarrow \mathbb{R}$, $\gamma(\xi) := \bar{p}(x)\,\xi + \lambda\,\xi^2/2$. According to Th. 6.62, γ is minimal at $\bar{\xi} \in [a(x), b(x)]$ if, and only if,

$$(\xi - \bar{\xi})\,\gamma'(\bar{\xi}) = (\xi - \bar{\xi})\,\big(\bar{p}(x) + \lambda\,\bar{\xi}\big) \geq 0 \quad \text{for each } \xi \in [a(x), b(x)]. \tag{6.161}$$

Thus, if (6.155e) holds, then (6.161) is valid with $\bar{\xi} = \bar{u}(x)$, i.e.

$$\bar{u}(x)\big(\bar{p}(x) + \lambda\,\bar{u}(x)\big) \leq \xi\,\big(\bar{p}(x) + \lambda\,\bar{u}(x)\big) \quad \text{for each } \xi \in [a(x), b(x)], \tag{6.162}$$

which is the same as (6.155c). Conversely, if (6.155c) holds, then (6.161) is valid with $\bar{\xi} = \bar{u}(x)$, implying (6.155e) and establishing the case.

For (a), it only remains to verify that, in the case $g = \bar{p} + \lambda\,\bar{u}$, $\lambda > 0$, all the previous conditions are also equivalent to (6.155f). We choose to carry out this verification by showing "(6.155a) $\Leftrightarrow$ (6.155f)": To that end, note that $g = \bar{p} + \lambda\,\bar{u}$, $\lambda > 0$, implies

$$g(x) > 0 \qquad \Leftrightarrow \qquad -\frac{\bar{p}(x)}{\lambda} < \bar{u}(x), \tag{6.163a}$$

$$g(x) = 0 \qquad \Leftrightarrow \qquad -\frac{\bar{p}(x)}{\lambda} = \bar{u}(x), \tag{6.163b}$$

$$g(x) < 0 \qquad \Leftrightarrow \qquad -\frac{\bar{p}(x)}{\lambda} > \bar{u}(x). \tag{6.163c}$$

Also, using (6.153), (6.155f) can be reformulated as

$$\bar{u}(x) = \begin{cases} a(x) & \text{for } -\frac{\bar{p}(x)}{\lambda} < a(x), \\ -\frac{\bar{p}(x)}{\lambda} & \text{for } -\frac{\bar{p}(x)}{\lambda} \in [a(x), b(x)], \\ b(x) & \text{for } b(x) < -\frac{\bar{p}(x)}{\lambda}. \end{cases} \tag{6.164}$$

According to (6.163), (6.164) is clearly the same as (6.155a), finishing the proof of (a).

(b): As in (a) for "(6.154) $\Rightarrow$ (6.155a)", proceeding by contraposition, assume there is a measurable set $E \subseteq \Omega$ such that $\lambda_m(E) > 0$ and $g \neq 0$ on E. Then there is also $\epsilon > 0$ such that at least one of the following two cases holds: (i) there is a measurable set $E_\epsilon \subseteq E$, $\lambda_m(E_\epsilon) > 0$, such that $g(x) \geq \epsilon$ for every $x \in E_\epsilon$, or (ii) there is a measurable set $E_\epsilon \subseteq E$, $\lambda_m(E_\epsilon) > 0$, such that $g(x) \leq -\epsilon$ for every $x \in E_\epsilon$. In case (i), define

$$u := \begin{cases} \bar{u} - 1 & \text{on } E_\epsilon, \\ \bar{u} & \text{on } \Omega \setminus E_\epsilon. \end{cases} \tag{6.165}$$

Since Ω is bounded, it is $u \in L^2(\Omega)$. Furthermore,

$$\int_\Omega g\,(u - \bar{u}) = -\int_{E_\epsilon} g \leq -\epsilon\,\lambda_m(E_\epsilon) < 0, \tag{6.166}$$

showing that (6.154) fails. Likewise, in case (ii), define

$$u := \begin{cases} \bar{u} + 1 & \text{on } E_\epsilon, \\ \bar{u} & \text{on } \Omega \setminus E_\epsilon. \end{cases} \tag{6.167}$$

A computation analogous to (6.166) verifies failure of (6.154) also in this case. ■

Example 6.69. As in Ex. 6.66 and Ex. 6.64(b), we consider the EOCP of Def. 6.45 with U_{ad} convex and $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ given by (6.141). For the convenience of the reader, we restate f and the system of optimality (6.143): If f is given by

$$f : U_{\text{ad}} \longrightarrow \mathbb{R}, \quad f(u) := \frac{1}{2} \int_\Omega (Su - y_\Omega)^2 + \frac{\lambda}{2} \int_\Omega u^2, \tag{6.168}$$

then $\bar{u} \in U_{\text{ad}}$ is an optimal control for the EOCP if, and only if, $(\bar{u}, \bar{y}, \bar{p}) \in U_{\text{ad}} \times H_0^1(\Omega) \times H_0^1(\Omega)$ satisfies the system of optimality (6.169):

$$\bar{u} \in U_{\text{ad}}, \tag{6.169a}$$
$$\bar{y} = S\bar{u}, \tag{6.169b}$$
$$\bar{p} = S^*(\bar{y} - y_\Omega), \tag{6.169c}$$
$$\int_\Omega (\bar{p} + \lambda\, \bar{u})\, (u - \bar{u}) \geq 0 \quad \text{for each } u \in U_{\text{ad}}. \tag{6.169d}$$

Depending on the forms of U_{ad} and λ, we can apply Th. 6.68 to replace (6.169d) with equivalent pointwise statements, which can help to gain more insight into the problem at hand.

(a) Consider $\lambda = 0$ and $U_{\text{ad}} = L^2_{a,b}(\Omega)$ with $a, b \in L^2(\Omega)$, $a \leq b$. Then Th. 6.68(a) applies with $g = \bar{p}$, i.e. one obtains that (6.169d) is equivalent to each of the following pointwise conditions (6.170a) – (6.170d) holding for almost every $x \in \Omega$:

$$\bar{u}(x) = \begin{cases} a(x) & \text{for } \bar{p}(x) > 0, \\ \in [a(x), b(x)] & \text{for } \bar{p}(x) = 0, \\ b(x) & \text{for } \bar{p}(x) < 0, \end{cases} \tag{6.170a}$$
$$\bar{p}(x)\, \big(\xi - \bar{u}(x)\big) \geq 0 \qquad \text{for each } \xi \in [a(x), b(x)], \tag{6.170b}$$
$$\bar{p}(x)\, \bar{u}(x) \leq \bar{p}(x)\, \xi \qquad \text{for each } \xi \in [a(x), b(x)], \tag{6.170c}$$
$$\min_{\xi \in [a(x), b(x)]} \bar{p}(x)\, \xi = \bar{p}(x)\, \bar{u}(x). \tag{6.170d}$$

In cases, where $\bar{p} \neq 0$ almost everywhere on Ω, (6.170a) shows that, almost everywhere, the optimal control $\bar{u}$ coincides either with the upper bound b or with the lower bound a. In this case, one says that the optimal control is *bang-bang*.

(b) Consider $\lambda = 0$ and $U_{\text{ad}} = L^2(\Omega)$. Then Th. 6.68(b) yields $\bar{p} = 0$ a.e. on Ω, which, in turn, implies $S\bar{u} = \bar{y} = y_\Omega$ a.e. In particular, in this case, the EOCP has an optimal control $\bar{u}$ if, and only if, y_Ω is in the range of S. In particular, the EOCP has *no solution* if $y_\Omega \in L^2(\Omega) \setminus H_0^1(\Omega)$.

(c) Consider $\lambda > 0$ and $U_{\text{ad}} = L^2_{a,b}(\Omega)$ with $a, b \in L^2(\Omega)$, $a \leq b$. Then Th. 6.68(a) applies with $g = \bar{p} + \lambda\, \bar{u}$, i.e. (6.169d) is equivalent to each of the following pointwise

conditions (6.171a) – (6.171f) holding for almost every $x \in \Omega$:

$$\bar{u}(x) = \begin{cases} a(x) & \text{for } \bar{p}(x) + \lambda\,\bar{u}(x) > 0, \\ \in [a(x), b(x)] & \text{for } \bar{p}(x) + \lambda\,\bar{u}(x) = 0, \\ b(x) & \text{for } \bar{p}(x) + \lambda\,\bar{u}(x) < 0, \end{cases} \tag{6.171a}$$

$$\big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\big(\xi - \bar{u}(x)\big) \geq 0 \qquad \text{for each } \xi \in [a(x), b(x)], \tag{6.171b}$$

$$\big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\bar{u}(x) \leq \big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\xi \qquad \text{for each } \xi \in [a(x), b(x)], \tag{6.171c}$$

$$\min_{\xi \in [a(x), b(x)]} \big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\xi = \big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\bar{u}(x), \tag{6.171d}$$

$$\min_{\xi \in [a(x), b(x)]} \left(\bar{p}(x)\xi + \frac{\lambda\,\xi^2}{2}\right) = \bar{p}(x)\bar{u}(x) + \frac{\lambda\,\bar{u}(x)^2}{2}, \tag{6.171e}$$

$$\bar{u}(x) = P_{[a(x), b(x)]}\left(-\frac{\bar{p}(x)}{\lambda}\right), \tag{6.171f}$$

where $P_{[a(x),b(x)]}$ is according to Not. 6.67.

(d) Consider $\lambda > 0$ and $U_{\text{ad}} = L^2(\Omega)$. Then Th. 6.68(b) yields $\bar{p} + \lambda\,\bar{u} = 0$ a.e. on Ω, i.e.

$$\bar{u} = -\frac{\bar{p}}{\lambda}. \tag{6.172}$$

Thus, in this case, the system of (6.169) can be reformulated as

$$\bar{y} = -S(\bar{p}/\lambda), \tag{6.173a}$$

$$\bar{p} = S^*(\bar{y} - y_\Omega), \tag{6.173b}$$

or, using the more explicit form (6.150), as

$$-\sum_{i=1}^{m}\sum_{j=1}^{m} \partial_i\big(a_{ij}\partial_j \bar{y}\big) + b\,\bar{y} = -\bar{p}/\lambda \quad \text{on } \Omega, \tag{6.174a}$$

$$\bar{y} = 0 \quad \text{on } \partial\Omega, \tag{6.174b}$$

$$-\sum_{i=1}^{m}\sum_{j=1}^{m} \partial_i\big(a_{ji}\partial_j \bar{p}\big) + b\,\bar{p} = \bar{y} - y_\Omega \quad \text{on } \Omega, \tag{6.174c}$$

$$\bar{p} = 0 \quad \text{on } \partial\Omega. \tag{6.174d}$$

In particular, the optimal state $\bar{y}$ and the adjoint state $\bar{p}$ are the solution of the system of coupled PDE (6.174). Given $\bar{y}$ and $\bar{p}$, the optimal control is provided by (6.172).

6.4.5 Lagrange Multipliers and Karush-Kuhn-Tucker Optimality Conditions

In Sec. 3.5.2, we used Lagrange multipliers to transform the variational inequality (3.31c) of the finite-dimensional optimization problem (3.3) with box constraints on

the control into a finite number of equations and inequalities formulated within the Karush-Kuhn-Tucker optimality system (3.37). Thereby, we achieved a structural simplification of the optimality system, as the variational inequality (3.31c) typically consists of uncountably many conditions (one for each $u \in U_{\mathrm{ad}}$). In the present section, we will proceed in an analogous fashion for the infinite-dimensional optimal control problem (6.169).

Remark 6.70. For each real-valued function $f : X \longrightarrow \mathbb{R}$, one can define its *positive part* $f^+ : X \longrightarrow \mathbb{R}$ and its *negative part* $f^- : X \longrightarrow \mathbb{R}$ by letting

$$f^+ := \frac{1}{2}\big(f + |f|\big) = \max\{0,\, f\}, \tag{6.175a}$$

$$f^- := \frac{1}{2}\big(|f| - f\big) = -\min\{0,\, f\}. \tag{6.175b}$$

Then (6.175) immediately implies

$$f^+ \geq 0, \quad f^- \geq 0, \tag{6.176a}$$

$$f = f^+ - f^-. \tag{6.176b}$$

Theorem 6.71. *Let $\Omega \subseteq \mathbb{R}^m$ be measurable and bounded, and $U_{\mathrm{ad}} = L^2_{a,b}(\Omega)$, $(a, b) \in L^2(\Omega) \times L^2(\Omega)$, $a \leq b$. Furthermore, let $\bar{p} \in L^2(\Omega)$, $\bar{u} \in U_{\mathrm{ad}}$, and $\lambda \geq 0$. Then the following statements* (i) – (iii) *are equivalent:*

(i) $\int_\Omega (\bar{p} + \lambda\, \bar{u})\, (u - \bar{u}) \geq 0$ *for each $u \in U_{\mathrm{ad}}$.*

(ii) *For almost every $x \in \Omega$, the following* complementary slackness conditions *hold:*

$$\big(\bar{p}(x) + \lambda\, \bar{u}(x)\big)^+ \big(a(x) - \bar{u}(x)\big) = \big(\bar{p}(x) + \lambda\, \bar{u}(x)\big)^- \big(\bar{u}(x) - b(x)\big) = 0. \tag{6.177}$$

(iii) *There exist $\mu_a \in L^2(\Omega)$ and $\mu_b \in L^2(\Omega)$ satisfying*

$$\mu_a \geq 0, \quad \mu_b \geq 0, \tag{6.178a}$$

$$\bar{p} + \lambda \bar{u} - \mu_a + \mu_b = 0, \tag{6.178b}$$

and the complementary slackness conditions

$$\mu_a(x) \big(a(x) - \bar{u}(x)\big) = \mu_b(x) \big(\bar{u}(x) - b(x)\big) = 0 \tag{6.179}$$

for almost every $x \in \Omega$. In this context, μ_a and μ_b are called Lagrange multipliers.

Proof. "(i) $\Rightarrow$ (ii)": We apply Th. 6.68(a) with $g := \bar{p} + \lambda \bar{u}$, obtaining that (i) implies (6.155a). We need to show, for almost every $x \in \Omega$,

$$g^+(x) \big(a(x) - \bar{u}(x)\big) = g^-(x) \big(\bar{u}(x) - b(x)\big) = 0. \tag{6.180}$$

If $g(x) > 0$, then $g^-(x) = 0$ and, by (6.155a), $\bar{u}(x) = a(x)$, such that (6.180) holds. If $g(x) = 0$, then $g^+(x) = g^-(x) = 0$ and (6.180) holds. If $g(x) < 0$, then $g^+(x) = 0$ and, by (6.155a), $\bar{u}(x) = b(x)$, such that (6.180) also holds.

"(ii) ⇒ (iii)": Defining $\mu_a := (\bar{p} + \lambda\,\bar{u})^+$, $\mu_b := (\bar{p} + \lambda\,\bar{u})^-$, one observes that (6.178) is identical to (6.176), and (6.179) is identical to (6.177).

"(iii) ⇒ (i)": Let $u \in U_{\text{ad}}$, i.e., in particular, $a(x) \leq u(x) \leq b(x)$ for almost every $x \in \Omega$. We show that every $x \in \Omega$ satisfying (6.179) and $a(x) \leq u(x) \leq b(x)$ also satisfies

$$\big(\bar{p}(x) + \lambda\,\bar{u}(x)\big)\,\big(u(x) - \bar{u}(x)\big) \geq 0. \tag{6.181}$$

Thus, assume $x \in \Omega$ satisfies (6.179). If $a(x) = b(x)$, then $u(x) = \bar{u}(x)$ and (6.181) holds. If $a(x) < \bar{u}(x) < b(x)$, then (6.179) implies $\mu_a(x) = \mu_b(x) = 0$, which, in turn implies $\bar{p}(x) + \lambda\,\bar{u}(x) = 0$ according to (6.178b). Hence, (6.181) holds. Next, if $b(x) > a(x) = \bar{u}(x)$, then $u(x) - \bar{u}(x) \geq 0$ and (6.179) implies $\mu_b(x) = 0$. Using (6.178) yields

$$\bar{p}(x) + \lambda\bar{u}(x) \overset{(6.178b)}{=} \mu_a(x) \overset{(6.178a)}{\geq} 0, \tag{6.182}$$

again showing that (6.181) holds. Finally, if $a(x) < b(x) = \bar{u}(x)$, then $u(x) - \bar{u}(x) \leq 0$ and (6.179) implies $\mu_a(x) = 0$. Using (6.178) yields

$$\bar{p}(x) + \lambda\bar{u}(x) \overset{(6.178b)}{=} -\mu_b(x) \overset{(6.178a)}{\leq} 0, \tag{6.183}$$

such that (6.181) holds in every case possible. Clearly, (6.181) implies (i). ■

Example 6.72. If $U_{\text{ad}} = L^2_{a,b}(\Omega)$, then Th. 6.71 allows to rewrite the system of optimality (6.169) as the Karush-Kuhn-Tucker optimality system

$$\begin{aligned}
a \leq\ \bar{u} &\leq b, && \text{(6.184a)}\\
\bar{y} &= S\bar{u}, && \text{(6.184b)}\\
\bar{p} &= S^*(\bar{y} - y_\Omega), && \text{(6.184c)}\\
\mu_a \geq 0, &\quad \mu_b \geq 0, && \text{(6.184d)}\\
\bar{p} + \lambda\bar{u} - \mu_a + \mu_b &= 0, && \text{(6.184e)}\\
\mu_a\,(a - \bar{u}) = \mu_b\,(\bar{u} - b) &= 0. && \text{(6.184f)}
\end{aligned}$$

More precisely, if $U_{\text{ad}} = L^2_{a,b}(\Omega)$, then $\bar{u} \in U_{\text{ad}}$ is an optimal control for the EOCP of Def. 6.45 with $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ given by (6.168) if, and only if, $(\bar{u}, \bar{y}, \bar{p}, \mu_a, \mu_b) \in L^2(\Omega) \times H^1_0(\Omega) \times H^1_0(\Omega) \times L^2(\Omega) \times L^2(\Omega)$ satisfies the Karush-Kuhn-Tucker optimality system (6.184).

7 Introduction to Numerical Methods

This section follows [Trö05, Sec. 2.12]. It is merely meant as a short introduction to the topic of numerical methods for the optimal control of PDE. An extensive literature on this topic is available, see, e.g., [Bet01, GS80, Kel99] as well as references therein.

7.1 Conditional Gradient Method

7.1.1 Abstract Case: Hilbert Space

Let U be a Hilbert space, and consider a Gâteaux differentiable objective functional $f : U \longrightarrow \mathbb{R}$. Moreover, let U_{ad} be a nonempty, closed, bounded, and convex subset of U, and consider the optimal control problem

$$\min_{u \in U_{\mathrm{ad}}} f(u). \tag{7.1}$$

The idea is to approximate a solution $\bar{u} \in U_{\mathrm{ad}}$ to (7.1) by a sequence $(u_n)_{n \in \mathbb{N}}$ in U_{ad}. Of course, without further conditions on f, we do not know, in general, if such a solution $\bar{u}$ exists. However, from Th. 6.62, we know that

$$f'_{\mathrm{G}}(\bar{u})(u - \bar{u}) \geq 0 \quad \text{for each } u \in U_{\mathrm{ad}} \tag{7.2}$$

is a necessary condition for $\bar{u}$ to be a solution to (7.1).

Given the iterative solution u_n, the next iteration u_{n+1} is found by determining a *direction of descent*, i.e. $v_n \in U$ such that f, in some local neighborhood of u_n, is decreasing in the direction $v_n - u_n$. Once such a direction $v_n - u_n$ is found, one still needs to determine a *step size* s_n such that $u_{n+1} = u_n + s_n(v_n - u_n)$ is a suitable next iteration in the approximating sequence.

Direction Search: The new direction $v_n - u_n$ is determined by the solution v_n to the auxiliary optimal control problem

$$\min_{v \in U_{\mathrm{ad}}} f'_{\mathrm{G}}(u_n)\,(v). \tag{7.3}$$

The function $g_n : U \longrightarrow \mathbb{R}$, $g_n(v) := f'_{\mathrm{G}}(u_n)\,(v)$ is continuous and linear (in particular, convex on U_{ad}). Thus, by Th. 5.3, (7.3) has a solution $v_n \in U_{\mathrm{ad}}$. In Sec. 7.1.2 below, we will see how to determine such a solution v_n in a concrete example. If $f'_{\mathrm{G}}(u_n)(v_n - u_n) \geq 0$, then (7.3) implies, for each $u \in U_{\mathrm{ad}}$,

$$f'_{\mathrm{G}}(u_n)(u - u_n) \geq f'_{\mathrm{G}}(u_n)(v_n - u_n) \geq 0, \tag{7.4}$$

i.e. $\bar{u} := u_n$ satisfies (7.2), i.e., if f is convex on U_{ad}, then Th. 6.62 implies that $\bar{u} := u_n$ is a solution to (7.1). Otherwise,

$$f'_{\mathrm{G}}(u_n)(v_n - u_n) < 0. \tag{7.5}$$

Since

$$f'_{\mathrm{G}}(u_n)(v_n - u_n) = \lim_{t \downarrow 0} \frac{1}{t} \big(f\big(u_n + t\,(v_n - u_n)\big) - f(u_n) \big), \tag{7.6}$$

(7.5) implies that $f\big(u_n + t\,(v_n - u_n)\big) < f(u_n)$ for each sufficiently small $t > 0$. Thus, in the direction $v_n - u_n$, f is decreasing in some local neighborhood of u_n. Note that, for convex U_{ad}, the $u_n + t\,(v_n - u_n)$ are elements of U_{ad} for each $t \in [0, 1]$.

Step Size Search: We now assume that, given $u_n \in U_{\text{ad}}$, we have already found $v_n \in U_{\text{ad}}$ as a solution to (7.3). As noted above, if $f'_{\text{G}}(u_n)(v_n - u_n) \geq 0$ and f is convex on U_{ad}, then u_n is already optimal, i.e. we can stop the iteration. However, one can also skip the check if $f'_{\text{G}}(u_n)(v_n - u_n) \geq 0$ and perform the step size search anyway. If u_n was already optimal, then one will find $s_n = 0$ (see below), if f is convex, then $s_n = 0$ will guarantee optimality of u_n. The new step size $s_n \in [0,1]$ is determined as the solution to another auxiliary minimization problem, namely, the one-dimensional problem

$$\min_{s\in[0,1]} f\big(u_n + s\,(v_n - u_n)\big). \tag{7.7}$$

Even though Gâteaux differentiability of f does, generally, not imply continuity of f, it does imply continuity of $f\restriction_{[u_n,v_n]}$, and, thus, of $g : [0,1] \longrightarrow \mathbb{R}$, $g(s) := f\big(u_n + s\,(v_n - u_n)\big)$. As g must attain its min on the compact set $[0,1]$, (7.7) must have a solution s_n. If (7.5) is satisfied, then $s_n > 0$. Thus, $s_n = 0$ implies $f'_{\text{G}}(u_n)(v_n - u_n) \geq 0$, and the converse holds if f is convex. If f is not convex, then $s_n > 0$ can occur even if $f'_{\text{G}}(u_n)(v_n - u_n) \geq 0$. If f is locally convex at u_n, then this means one has luckily espaced a local min of f at u_n. Again, we will solve (7.7) for a concrete situation in Sec. 7.1.2 below.

Once both v_n and s_n are determined as described, one sets $u_{n+1} := u_n + s_n\,(v_n - u_n)$.

The described algorithm for the construction of the u_n is known as the *conditional gradient method*. If f is continuous and convex on U_{ad}, then we know from Th. 5.3, that (7.1) has a solution $\bar{u} \in U_{\text{ad}}$. In that case, one can also guarantee that the conditional gradient method produces a monotonically decreasing sequence $\big(f(u_n)\big)_{n\in\mathbb{N}}$ (stricly decreasing unless $u_n = \bar{u}$). One can show that the u_n actually converge to a solution $\bar{u}$ of (7.1); a disadvantage is the generally rather slow convergence rate [GS80].

7.1.2 Application: Elliptic Optimal Control Problem

We apply the conditional gradient method to the EOCP of Def. 6.45 with $f : U_{\text{ad}} \longrightarrow \mathbb{R}$ given by (6.168), $U_{\text{ad}} = L^2_{a,b}(\Omega)$, $(a,b) \in L^2(\Omega) \times L^2(\Omega)$, $a \leq b$. According to Examples 6.47 and 6.60(b), f is convex and Fréchet differentiable and, for each $u \in U = L^2(\Omega)$, $f'_{\text{G}}(u) = f'_{\text{F}}(u)$ is given by (6.130e) as

$$\begin{aligned} f'_{\text{F}}(u) :\, L^2(\Omega) \longrightarrow \mathbb{R}, \quad f'_{\text{F}}(u)(h) &= \int_\Omega \big(S^*(Su - y_\Omega) + \lambda\, u\big)\, h \\ &= \int_\Omega \big(p + \lambda\, u\big)\, h, \end{aligned} \tag{7.8}$$

where, as usual, we have used the adjoint state p, i.e.

$$y = Su, \tag{7.9a}$$

$$p = S^*(y - y_\Omega). \tag{7.9b}$$

Each iteration of the conditional gradient method can be divided into five steps **S1** – **S5**. We first describe these five steps consecutively, subsequently providing further comments and details.

S1 Given the control $u_n \in U_{\text{ad}}$, determine the corresponding state $y_n = Su_n$ as the solution of the state equation.

S2 Given the state y_n, determine the corresponding adjoint state $p_n = S^*(y_n - y_\Omega)$ as the solution of the adjoint equation.

S3 For the new direction $v_n - u_n$, determine v_n as a solution to (7.3), i.e. as a solution to

$$\min_{v \in U_{\text{ad}}} \int_\Omega \big(p_n + \lambda\, u_n\big)\, v. \tag{7.10}$$

S4 Determine the new step size $s_n \in [0,1]$ as a solution to (7.7), i.e. as a solution to

$$\min_{s \in [0,1]} f\big(u_n + s\,(v_n - u_n)\big). \tag{7.11}$$

If $s_n = 0$, then u_n is an optimal control for the EOCP and the iteration is halted (see comment below).

S5 Put $u_{n+1} := u_n + s_n\,(v_n - u_n)$ and, with n replaced by $n+1$, proceed to **S1** for the next iteration step.

S1 and **S2** both require the solution of a PDE, and, typically, will have to be carried out numerically. Thus, any convergence and error analysis of the method will have to take into account the convergence and error analysis of the numerical method used for solving the PDE.

According to Sec. 7.1.1, one should check in **S3** if $\big(p_n + \lambda\, u_n\big)\,(v_n - u_n) \geq 0$, concluding optimality and halting if this is the case (as f is convex), proceeding to **S4** only if $\big(p_n + \lambda\, u_n\big)\,(v_n - u_n) < 0$. However, as also remarked in Sec. 7.1.1, checking if $s_n = 0$ in **S4** is an equivalent (and, in practice, easier) way of checking for optimality on a convex f: If $\big(p_n + \lambda\, u_n\big)\,(v_n - u_n) \geq 0$, then u_n is optimal, and **S4** will yield $s_n = 0$. Otherwise, i.e. if $\big(p_n + \lambda\, u_n\big)\,(v_n - u_n) < 0$, then necessarily $s_n > 0$ in **S4**.

For **S3**, one needs to determine a solution v_n to (7.10). As U_{ad} is convex, and the function to be minimized is linear (in particular, convex), according to Th. 6.62, $v_n \in U_{\text{ad}}$ is a solution to (7.10) if, and only if,

$$\int_\Omega \big(p_n + \lambda\, u_n\big)\,(v - v_n) \geq 0 \quad \text{for each } v \in U_{\text{ad}}. \tag{7.12}$$

From Th. 6.68(a), we see that

$$v_n(x) := \begin{cases} a(x) & \text{for } p_n(x) + \lambda\, u_n(x) > 0, \\ \big(a(x) + b(x)\big)/2 & \text{for } p_n(x) + \lambda\, u_n(x) = 0, \\ b(x) & \text{for } p_n(x) + \lambda\, u_n(x) < 0, \end{cases} \tag{7.13}$$

is a possible choice for v_n, where one should note that the middle case will hardly ever occur during numerical calculations.

Performing **S4** is also easy for the f under consideration. We have to minimize $g : [0,1] \longrightarrow \mathbb{R}$, where, using the abbreviations $y_n := Su_n$, $w_n := Sv_n$,

$$\begin{aligned} g(s) &= f\big(u_n + s\,(v_n - u_n)\big) \\ &= \frac{1}{2}\,\big\|S\big(u_n + s\,(v_n - u_n)\big) - y_\Omega\big\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\big\|\big(u_n + s\,(v_n - u_n)\big)\big\|^2_{L^2(\Omega)} \\ &= \frac{1}{2}\,\big\|y_n + s\,(w_n - y_n) - y_\Omega\big\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\big\|\big(u_n + s\,(v_n - u_n)\big)\big\|^2_{L^2(\Omega)} \\ &= \frac{1}{2}\,\big\|y_n - y_\Omega\big\|^2_{L^2(\Omega)} + s\,\langle y_n - y_\Omega,\, w_n - y_n\rangle_{L^2(\Omega)} + \frac{s^2}{2}\,\big\|w_n - y_n\big\|^2_{L^2(\Omega)} \\ &\quad + \frac{\lambda}{2}\,\|u_n\|^2_{L^2(\Omega)} + \lambda\, s\,\langle u_n,\, v_n - u_n\rangle_{L^2(\Omega)} + s^2\,\frac{\lambda}{2}\,\|v_n - u_n\|^2_{L^2(\Omega)} \\ &= g_0 + g_1\, s + g_2\, s^2, \end{aligned} \tag{7.14}$$

with

$$g_0 := \frac{1}{2}\,\big\|y_n - y_\Omega\big\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u_n\|^2_{L^2(\Omega)}, \tag{7.15a}$$

$$g_1 := \langle y_n - y_\Omega,\, w_n - y_n\rangle_{L^2(\Omega)} + \lambda\,\langle u_n,\, v_n - u_n\rangle_{L^2(\Omega)}, \tag{7.15b}$$

$$g_2 := \frac{1}{2}\,\big\|w_n - y_n\big\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|v_n - u_n\|^2_{L^2(\Omega)}. \tag{7.15c}$$

Thus, g is a quadratic function, with $g_2 \geq 0$ according to (7.15c). If $g_2 > 0$, then the min of g on $[0,1]$ is given by the projection of the zero of its derivative, i.e. by

$$s_n := P_{[0,1]}\left(-\frac{g_1}{2\,g_2}\right). \tag{7.16}$$

If $g_2 = 0$ and $g_1 > 0$, then $s_n = 0$. If $g_2 = 0$ and $g_1 < 0$, then $s_n = 1$. If $g_2 = g_1 = 0$, then g is constant and any $s_n \in [0,1]$ will do.

7.2 Projected Gradient Method

The *projected gradient method* is similar to the conditional gradient method of the previous section. For the projected gradient method, one always uses the antigradient, i.e $-(p_n + \lambda\, u_n)$, for the new direction $v_n - u_n$. Thus, in **S3** of Sec. 7.1.2, one sets

$$v_n := u_n - (p_n + \lambda\, u_n). \tag{7.17}$$

However, one now has the additional problem that, in general, even if $u_n \in U_{\rm ad}$, the quantity $u_n - s\,(p_n + \lambda\, u_n)$ can be nonadmissible (i.e. $u_n - s\,(p_n + \lambda\, u_n) \notin U_{\rm ad}$) for every $s > 0$. This is remedied by projecting $u_n - s\,(p_n + \lambda\, u_n)$ back onto $U_{\rm ad}$. However, now, instead of (7.7), for the new step size s_n, one wants to solve the more difficult auxiliary minimization problem

$$\min_{s\in[0,1]} f\big(P_{[a,b]}\big(u_n + s\,(v_n - u_n)\big)\big), \tag{7.18}$$

at least approximatively. This usually requires evaluations of f, and, thus, solving the PDE, which can be numerically costly. Possible strategies mentioned in [Trö05, Sec. 2.12.2] are the *bisection method* and *Armijo's method*.

Even though each step of the projected gradient method is typically more difficult than the corresponding step of the conditional gradient method, in many situations, this is more than compensated by a faster convergence rate. For more information on the projected gradient method and its convergence theory, see, e.g., [GS80, Kel99].

7.3 Transformation into Finite-Dimensional Problems

Here, the general strategy is to use discretization techniques to transform the infinite-dimensional optimal control problem of minimizing an objective functional with PDE constraints into, possibly large, but finite-dimensional optimization problems, approximating the original problem. We will restrict ourselves to illustrating the technique in a rather simple situation.

7.3.1 Finite-Dimensional Formulation in Nonreduced Form

For our illustrating example, we consider the particularly simple space domain

$$\Omega :=]0,1[\times]0,1[,$$

and we choose the optimal control problem

$$\text{minimize} \quad J(y,u) := \frac{1}{2}\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|u\|^2_{L^2(\Omega)},$$

subject to the PDE constraints

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

and control constraints

$$a \le u \le b \quad \text{on } \Omega,$$

where $\lambda > 0$ and $y_\Omega, a, b \in L^2(\Omega)$, $a \le b$. The reason for choosing Ω to be the unit square is its admitting simple discretizations into n^2 ($n \in \mathbb{N}$) small squares

$$\overline{\Omega} = \bigcup_{i,j=1}^{n} \overline{\Omega}_{ij}, \tag{7.19a}$$

$$\Omega_{ij} := \left]\frac{i-1}{n}, \frac{i}{n}\right[\times \left]\frac{j-1}{n}, \frac{j}{n}\right[. \tag{7.19b}$$

More complicated domains Ω are, in general, more difficult to discretize, but the principal stretegy remains the same.

To proceed further, we introduce the $(n+1)^2$ points

$$x_{ij} := (hi, hj), \quad h := \frac{1}{n}, \quad i,j = 0,\dots,n,$$

with neighborhoods

$$\tilde{\Omega}_{ij} := B_{\frac{h}{2},\|\cdot\|_{\max}}(x_{ij}).$$

Thus, the x_{ij} are precisely the vertices of the squares Ω_{ij} as well as the centers of the $\tilde{\Omega}_{ij}$. The goal is to determine values y_{ij} as solutions to a suitable finite-dimensional optimization problem such that $y_{ij} \approx y(x_{ij})$, i.e. such that the y_{ij} are approximations to the value of y at x_{ij} (or on $\tilde{\Omega}_{ij}$).

Using difference quotients of difference quotients, one obtains the classical approximation

$$-\Delta y(x_{ij}) = -\partial_1\partial_1 y(x_{ij}) - \partial_2\partial_2 y(x_{ij}) \approx \frac{4y_{ij} - (y_{i-1,j} + y_{i,j-1} + y_{i+1,j} + y_{i,j+1})}{h^2}. \quad (7.20)$$

At the boundary of Ω, we use

$$y_{0j} := y_{i0} := y_{nj} := y_{in} := 0 \quad \text{for each } i,j = 0,\dots,n.$$

Analogous to the y_{ij}, we introduce approximations

$$u_{ij} \approx u(x_{ij}), \quad y_{\Omega,ij} \approx y_\Omega(x_{ij}), \quad a_{ij} \approx a(x_{ij}), \quad b_{ij} \approx b(x_{ij}),$$

where $y_{\Omega,ij}$, a_{ij}, b_{ij} have to be computed from the given functions y_Ω, a, b, respectively, whereas the u_{ij} constitute additional unknowns to be determined from the finite-dimensional optimization problem below. By employing an enumeration of the index set $\{(i,j) : i,j = 1,\dots,n-1\}$, we organize the y_{ij}, u_{ij}, $y_{\Omega,ij}$, a_{ij}, and b_{ij} into vectors

$$\vec{y} = (y_1,\dots,y_{(n-1)^2}), \qquad \vec{u} = (u_1,\dots,u_{(n-1)^2}), \qquad \vec{y}_\Omega = (y_{\Omega,1},\dots,y_{(n-1)^2}),$$
$$\vec{a} = (a_1,\dots,a_{(n-1)^2}), \qquad \vec{b} = (b_1,\dots,b_{(n-1)^2}),$$

respectively. Approximating the functions y, u, and y_Ω as being constant on each $\tilde{\Omega}_{ij}$ with values according to the above approximations, one obtains

$$\begin{aligned} J(y,u) &= \frac{1}{2}\int_\Omega (y-y_\Omega)^2 + \frac{\lambda}{2}\int_\Omega u^2 = \frac{1}{2}\sum_{i,j=0}^{n}\int_{\tilde{\Omega}_{ij}} \big((y-y_\Omega)^2 + \lambda u^2\big) \\ &\approx \frac{1}{2}\sum_{i,j=1}^{n-1}\int_{\tilde{\Omega}_{ij}} \big((y_{ij}-y_{\Omega,ij})^2 + \lambda u_{ij}^2\big) = \frac{h^2}{2}\sum_{i,j=1}^{n-1}\big((y_{ij}-y_{\Omega,ij})^2 + \lambda u_{ij}^2\big). \end{aligned}$$

Thus, our original optimal control problem is translated into the finite-dimensional optimization problem

$$\text{minimize} \quad \frac{1}{2}\sum_{k=1}^{(n-1)^2}\big((y_k - y_{\Omega,k})^2 + \lambda u_k^2\big),$$

subject to the equation constraints

$$A_h\,\vec{y}^{\top} = \vec{u}^{\top},$$

and control constraints

$$\vec{a} \le \vec{u} \le \vec{b}$$

for the unknown components of $\vec{y}$ and $\vec{u}$, where the entries of the matrix A_h are given according to (7.20). Discrete optimization problems of this form can, for example, be solved by the `MATLAB` function `quadprog`.

That the size of the discrete problem is typically large, is a drawback of the approach considered in the present section. It is often useful to obtain smaller problems by discretizing the reduced optimal control problem, which is discussed in the next section.

7.3.2 Finite-Dimensional Formulation in Reduced Form

The strategy of the previous section consists of discretizing the nonreduced optimal control problem directly, i.e. the variable y is not eliminated by writing it as a function of u. In contrast, in the following, we will employ the PDE's solution operator S to eliminate the variable y. We start by discretizing u. Here, the idea is to approximate u as a finite linear combination of M basis functions $e_1, \dots, e_M$, $M \in \mathbb{N}$,

$$u(x) \approx \sum_{k=1}^{M} u_k\, e_k(x), \quad u_k \in \mathbb{R}.$$

Continueing to employ the notation introduced in the previous section, for the e_k, we choose the $M = n^2$ characteristic functions

$$1_{\Omega_{ij}}(x) := \begin{cases} 1 & \text{for } x \in \Omega_{ij}, \\ 0 & \text{for } x \in \Omega \setminus \Omega_{ij}. \end{cases}$$

Next, we introduce the functions

$$y_k := S(e_k), \quad k = 1, \dots, M,$$

where $S : L^2(\Omega) \longrightarrow L^2(\Omega)$ is the PDE's solution operator according to Th. 6.43.

In practice, the computation of the y_k means the numerical solution of n^2 PDE. And as a fine discretization of Ω can be necessary to achieve an acceptable accuracy for the approximation of u, this can be computationally costly. However, in Sec. 7.3.3 below, we will see a trick to avoid the computation of the y_k.

Plugging the approximations for u into $J(y,u) = J(Su,u)$ yields the finite-dimensional optimization problem

$$\text{minimize} \quad f_h(\vec{u}) = f_h(u_1, \dots, u_M) := \frac{1}{2} \left\| \sum_{k=1}^{M} u_k\, y_k - y_\Omega \right\|^2_{L^2(\Omega)} + \frac{\lambda}{2} \left\| \sum_{k=1}^{M} u_k\, e_k \right\|^2_{L^2(\Omega)}, \tag{7.21a}$$

subject to the control constraints

$$\vec{a} \le \vec{u} \le \vec{b} \tag{7.21b}$$

for the unknown components of $\vec{u}$. In order to solve (7.21), we rewrite $f_h(\vec{u})$ by making use of the inner product in $L^2(\Omega)$:

$$\begin{aligned} f_h(\vec{u}) &= \frac{1}{2}\,\|y_\Omega\|^2_{L^2(\Omega)} - \left\langle y_\Omega,\, \sum_{k=1}^{M} u_k\, y_k \right\rangle + \frac{1}{2}\left\langle \sum_{k=1}^{M} u_k\, y_k,\, \sum_{l=1}^{M} u_l\, y_l \right\rangle \\ &\quad + \frac{\lambda}{2}\left\langle \sum_{k=1}^{M} u_k\, e_k,\, \sum_{l=1}^{M} u_l\, e_l \right\rangle \\ &= \frac{1}{2}\,\|y_\Omega\|^2_{L^2(\Omega)} - \sum_{k=1}^{M} u_k\,\langle y_\Omega,\, y_k\rangle + \frac{1}{2}\sum_{k,l=1}^{M} u_k u_l\,\langle y_k,\, y_l\rangle + \frac{\lambda}{2}\sum_{k=1}^{M} u_k u_l\,\langle e_k,\, e_l\rangle. \end{aligned}$$

Since $\frac{1}{2}\,\|y_\Omega\|^2_{L^2(\Omega)}$ is constant, (7.21) is equivalent to

$$\text{minimize} \quad \tilde{f}_h(\vec{u}) := \vec{v}\,\vec{u}^\top + \frac{1}{2}\,\vec{u}C\vec{u}^\top + \frac{\lambda}{2}\,\vec{u}D\vec{u}^\top, \tag{7.22a}$$

subject to the control constraints

$$\vec{a} \le \vec{u} \le \vec{b} \tag{7.22b}$$

for the unknown components of $\vec{u}$, where

$$\begin{aligned} \vec{v} &:= (v_k)_{k=1}^M, & v_k &:= -\langle y_\Omega, y_k\rangle, \\ C &:= (c_{kl})_{k,l=1}^M, & c_{kl} &:= \langle y_k,\, y_l\rangle, \\ D &:= (d_{kl})_{k,l=1}^M, & d_{kl} &:= \langle e_k,\, e_l\rangle. \end{aligned}$$

Note that due to the choice of the e_k as the characteristic functions of the disjoint sets Ω_{ij}, the matrix D is diagonal:

$$\langle e_k,\, e_l\rangle = \begin{cases} \|e_k\|^2_{L^2(\Omega)} & \text{for } k = l, \\ 0 & \text{for } k \ne l. \end{cases}$$

Once again, (7.22) can, for example, be solved by the `MATLAB` function `quadprog`.

7.3.3 Trick to Solve the Reduced Form Without Formulating it First

The numerical solution of the finite-dimensional optimization problem (7.22), requires computations of expressions of the form $D\vec{x}$ and $C\vec{x}$ with $x \in \mathbb{R}^M$, where for large M one should only store the nonzero elements of D and C. While expressions of the form $D\vec{x}$ are easy to compute, $C\vec{x}$ is typically more involved, since C depends on the y_k, where y_k resulted from solving the PDE with e_k on the right-hand side.

However, using the following trick, one can evaluate $C\vec{x}$ without having to compute the y_k first: For each row vector $\vec{x} \in \mathbb{R}^M$, we have

$$(C\vec{x})_k = \sum_{l=1}^{M} c_{kl} x_l = \sum_{l=1}^{M} x_l \langle y_k, y_l \rangle = \sum_{l=1}^{M} x_l \langle Se_k, Se_l \rangle = \left\langle S^*S \sum_{l=1}^{M} x_l e_l, e_k \right\rangle$$
$$= \langle S^*Sx_h, e_k \rangle,$$

where $x_h := \sum_{l=1}^{M} x_l e_l$.

Thus, instead of solving $M = n^2$ PDE once, one now has to solve 2 PDE for every new application of C. In general, it will depend on n and on the number of applications of C needed during the solution of (7.22), which strategy is preferable.

7.4 Active Set Methods

For simplicity, we remain in the setting established in Sec. 7.3.1.

Active set methods are motivated by the observation that, if $\bar{u}$ is an optimal control, then

$$\bar{u}(x) = \begin{cases} a(x) & \text{for } -\frac{\bar{p}(x)}{\lambda} < a(x), \\ -\frac{\bar{p}(x)}{\lambda} & \text{for } -\frac{\bar{p}(x)}{\lambda} \in [a(x), b(x)], \\ b(x) & \text{for } b(x) < -\frac{\bar{p}(x)}{\lambda} \end{cases} \tag{7.23}$$

where $\bar{p} = S^*(S\bar{u} - y_\Omega)$ is the adjoint state (this is due to Ex. 6.69(c), where we have written (6.170a) in the equivalent form from (6.164)). The relation (7.23) suggests that the quantity $-\frac{\bar{p}(x)}{\lambda}$ can be considered as a measure for the activity of the control constraints.

The continuous version of the active set algorithm reads as follows:

S0 Choose arbitrary initial functions $u_0, p_0 \in L^2(\Omega)$.

S1 Given the control u_n and the adjoint state p_n, determine the new *active sets* A^+_{n+1} and A^-_{n+1} as well as the *inactive set* I_{n+1}:

$$A^+_{n+1} := \left\{ x \in \Omega : -\frac{p_n(x)}{\lambda} > b(x) \right\},$$
$$A^-_{n+1} := \left\{ x \in \Omega : -\frac{p_n(x)}{\lambda} < a(x) \right\},$$
$$I_{n+1} := \Omega \setminus (A^+_{n+1} \cup A^-_{n+1}).$$

S2 Determine y_{n+1} and p_{n+1} from the coupled system of PDE

$$-\Delta y_{n+1} = \begin{cases} a & \text{on } A^-_{n+1}, \\ -\frac{p_{n+1}}{\lambda} & \text{on } I_{n+1}, \\ b & \text{on } A^+_{n+1}, \end{cases}$$
$$-\Delta p_{n+1} = y_{n+1} - y_\Omega.$$

S3 Set

$$u_{n+1} := \begin{cases} a & \text{on } A_{n+1}^-, \\ -\frac{p_{n+1}}{\lambda} & \text{on } I_{n+1}, \\ b & \text{on } A_{n+1}^+. \end{cases}$$

Proceed to **S1** for the next iteration step.

Active set methods can be interpreted as Newton methods and, thus, show rapid convergence rates [BIK99, KR02].

It remains to formulate the discrete version of the above algorithm. According to the discretization approach from Sec. 7.3.2, we seek a solution $\vec{u}$ to (7.21). For our present purposes, it will be useful to introduce the operator

$$S_h : \mathbb{R}^M \longrightarrow L^2(\Omega), \quad S_h(\vec{u}) := \sum_{k=1}^{M} u_k \, y_k,$$

which occurs on the right-hand side of (7.21a) ($y_k = S(e_k)$ as before). Note that the adjoint operator of S_h is a map $S_h^* : L^2(\Omega) \longrightarrow \mathbb{R}^M$.

Steps **D2** and **D3** of the algorithm below require the computation of expressions of the form $S_h^*(S_h\vec{u}_n - y_\Omega)$. As we have used e_k to denote the basis functions, let ϵ_k denote the standard unit vectors in $\mathbb{R}^M$, i.e.

$$\epsilon_{kl} := \delta_{kl} := \begin{cases} 1 & \text{for } k = l, \\ 0 & \text{for } k \neq l. \end{cases}$$

Then, for $u \in L^2(\Omega)$, the kth component of $S_h^* u \in \mathbb{R}^M$ is

$$\begin{aligned} (S_h^* u)_k &= \langle S_h^* u, \, \epsilon_k \rangle_{\mathbb{R}^M} = \langle u, \, S_h \epsilon_k \rangle_{L^2(\Omega)} = \left\langle u, \, \sum_{l=1}^{M} \delta_{kl} \, y_l \right\rangle_{L^2(\Omega)} = \langle u, \, y_k \rangle_{L^2(\Omega)} \\ &= \langle u, \, S e_k \rangle_{L^2(\Omega)} = \langle S^* u, \, e_k \rangle_{L^2(\Omega)}. \end{aligned} \tag{7.24}$$

Thus, to compute $S_h^* y_\Omega$, it suffices to solve precisely one PDE, namely the one corresponding to $S^* y_\Omega$. Given $\vec{u} \in \mathbb{R}^M$, by setting $u := S_h \vec{u}$ in (7.24), one obtains

$$(S_h^* S_h \vec{u})_k = \left\langle \sum_{l=1}^{M} u_l y_l, \, y_k \right\rangle_{L^2(\Omega)} = \left\langle S^* S \left(\sum_{l=1}^{M} u_l e_l \right), \, e_k \right\rangle_{L^2(\Omega)}. \tag{7.25}$$

In consequence, as in Sections 7.3.2 and 7.3.3, one has the choice of either computing the y_k by solving $M = n^2$ PDE once, or solving two PDE whenever an expression of the form $S_h^*(S_h\vec{u}_n - y_\Omega)$ needs to be calculated. In general, it will depend on the situation (e.g. size of M, desired accuracy, etc.), which strategy is faster.

The discrete active set algorithm can be formulated as follows:

D0 Choose arbitrary initial vectors $\vec{u}_0, \vec{p}_0 \in \mathbb{R}^M$.

D1 Given $\vec{u}_n$ and $\vec{p}_n$, determine the new *finite active sets* A^+_{n+1} and A^-_{n+1} as well as the *finite inactive set* I_{n+1}:

$$\begin{aligned} A^+_{n+1} &:= \left\{ k \in \{1, \dots, M\} : -\frac{p_{n,k}}{\lambda} > b_k \right\}, \\ A^-_{n+1} &:= \left\{ k \in \{1, \dots, M\} : -\frac{p_{n,k}}{\lambda} < a_k \right\}, \\ I_{n+1} &:= \{1, \dots, M\} \setminus (A^+_{n+1} \cup A^-_{n+1}). \end{aligned}$$

D2 Determine $\vec{u}_{n+1}$ from the linear system

$$u_{n+1,k} = \begin{cases} a_k & \text{for } k \in A^-_{n+1}, \\ -\lambda^{-1}\big(S_h^*(S_h\vec{u}_n - y_\Omega)\big)_k, & \\ b_k & \text{for } k \in A^+_{n+1}. \end{cases}$$

D3 Set $\vec{p}_{n+1} := S_h^*(S_h\vec{u}_n - y_\Omega)$.

Proceed to **D1** for the next iteration step.

The iteration is halted once $A^+_{n+1} = A^+_n$ and $A^-_{n+1} = A^-_n$. One can show that the active sets must become stationary after finitely many steps and that, at this stage, one can set $\vec{u} := \vec{u}_n$ to obtain a solution to (7.21) (see [BIK99, KR02]).

References

[Alt06] Hans Wilhelm Alt. *Lineare Funktionalanalysis*, 5th ed. Springer-Verlag, Berlin, 2006 (German).

[Bet01] J.T. Betts. *Practical Methods for Optimal Control Using Nonlinear Programming*. SIAM, Philadelphia, USA, 2001.

[BIK99] M. Bergounioux, K. Ito, and K. Kunisch. *Primal-dual strategy for constrained optimal control problems*. SIAM J. Control and Optimization **37** (1999), 1176–1194.

[Els96] J. Elstrodt. *Maß- und Integrationstheorie*. Grundwissen Mathematik, Springer-Verlag, Berlin, 1996 (German).

[Eva98] Lawrence C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics, Vol. 19, American Mathematical Society, Providence, Rhode Island, 1998, reprinted with corrections in 2002.

[GS80] W.A. Gruver and E.W. Sachs. *Algorithmic Methods in Optimal Control*. Pitman, London, 1980.

[IT79] A.D. Ioffe and V.M. Tihomirov. *Theory of Extremal Problems*. Studies in Mathematics and its Applications, Vol. 6, North-Holland Publishing Company, Amsterdam, 1979.

[Kel99] C.T. Kelley. *Iterative Methods for Optimization*. SIAM, Philadelphia, USA, 1999.

[KR02] K. Kunisch and A. Rösch. *Primal-dual active set strategy for a general class of constrained optimal control problems*. SIAM J. on Optimization **13** (2002), 321–334.

[KS80] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and their Applications*. Pure and Applied Mathematics, Vol. 88, Academic Press, New York, 1980.

[Lio71] J.L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, 1st ed. Grundlehren der mathematischen Wissenschaften, Vol. 170, Springer-Verlag, Berlin, 1971.

[Roy88] H.L. Royden. *Real Analysis*, 3rd ed. Macmillian Publ., New York, 1988.

[RR96] M. Renardy and R.C. Rogers. *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics, Vol. 13, Springer-Verlag, New York, 1996, corrected 2nd printing.

[Rud73] W. Rudin. *Functional Analyis*. McGraw-Hill Book Company, New York, 1973.

[Rud87] W. Rudin. *Real and Complex Analysis*, 3rd ed. McGraw-Hill Book Company, New York, 1987.

[Trö05] Fredi Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen*, 1st ed. Vieweg Verlag, Wiesbaden, Germany, 2005 (German).

[Wer02] D. Werner. *Funktionalanalysis*, 4th ed. Springer-Verlag, Berlin, 2002 (German).

[Wlo82] J. Wloka. *Partielle Differentialgleichungen*. B.G. Teubner, Stuttgart, Germany, 1982 (German).

[Yos74] K. Yosida. *Functional Analysis*, 4th ed. Grundlehren der mathematischen Wissenschaften, Vol. 123, Springer-Verlag, Berlin, 1974.

[Zei90] E. Zeidler. *Nonlinear Functional Analysis and its Applications*. Vol. II/A, Springer-Verlag, New York, 1990.