

Introduction and examples

1.1. What is optimal control?

The mathematical theory of optimal control has in the past few decades rapidly developed into an important and separate field of applied mathematics. One area of application of this theory lies in aviation and space technology: aspects of optimization come into play whenever the motion of an aircraft or a space vessel (which can be modeled by ordinary differential equations) has to follow a trajectory that is “optimal” in a sense to be specified.

Let us explain this by a simple example: a vehicle that at time $t = 0$ is at the space point A moves along a straight line and stops at time $T > 0$ at another point B on that line. Suppose that the vehicle can be accelerated along the line in either direction by a variable force whose maximal strength is the same in both directions. For example, this situation might represent a jet engine that can be switched between forward and backward thrust. What is the minimal time $T > 0$ needed for the travel, provided that the available thrust $u(t)$ at time t is subject to the constraint $-1 \leq u(t) \leq 1$? Here, $u(t) = +1$ (respectively, $u(t) = -1$) corresponds to maximal forward (respectively, backward) acceleration.

To model this situation, let $y(t)$ denote the position of the vehicle at time t , m the mass of the vehicle (which is assumed to remain constant during the process), and $y_0, y_T \in \mathbb{R}$ the points corresponding to the positions A and B . The mathematical problem then reads as follows:

Minimize $T > 0$, subject to the constraints

$$\begin{aligned} m y''(t) &= u(t) \\ y(0) &= y_0 \\ y'(0) &= 0, \\ \\ y(T) &= y_T \\ y'(T) &= 0 \\ |u(t)| &\leq 1 \quad \forall t \in [0, T]. \end{aligned}$$

The above problem, which is referred to as *the rocket car* in the textbook by Macki and Strauss [MS82], exhibits all the essential features of an *optimal control problem*:

- a *cost functional* to be minimized (here, the time $T > 0$ needed for the travel),
- an initial value problem for a differential equation (here, $m y'' = u$, $y(0) = y_0$, $y'(0) = 0$) describing the motion, in order to determine the *state* y ,
- a *control function* u , and
- various constraints (here, $y(T) = y_T$, $y'(T) = 0$, $|u| \leq 1$) that have to be obeyed.

The control u may be freely chosen within the given constraints (e.g., for the rocket car, by stepping on the gas or the brake pedal), while the state is uniquely determined by the differential equation and the initial conditions. We have to choose u in such a way that the cost function is minimized. Such controls are called *optimal*. In the case of the rocket car, intuition immediately tells us what the optimal choice should be. For this reason, this example is often used to test theoretical results.

The optimal control of ordinary differential equations is of interest not only for aviation and space technology. In fact, it is also important in fields such as robotics, movement sequences in sports, and the control of chemical processes and power plants, to name just a few of the various applications. In many cases, however, the processes to be optimized can no longer be adequately modeled by *ordinary* differential equations; instead, *partial* differential equations have to be employed for their description. For instance, heat conduction, diffusion, electromagnetic waves, fluid flows, freezing processes, and many other physical phenomena can be modeled by partial differential equations.

In these fields, there are numerous interesting problems in which a given cost functional has to be minimized subject to a differential equation and certain constraints being satisfied. The difference from the above problem

“merely” consists of the fact that a partial differential equation has to be dealt with in place of an ordinary one. In this textbook, we will discuss, through examples in the form of mathematically simplified case studies, the optimal control of heating processes, two-phase problems, and fluid flows.

There are many types of partial differential equations. Here, we focus on linear and semilinear elliptic and parabolic partial differential equations, since a satisfactory regularity theory is available for the solutions to such equations. This is not the case for hyperbolic equations. Also, the treatment of quasilinear partial differential equations is considerably more difficult, and the theory of their optimal control is still an open field in many respects.

We begin our study with problems involving linear equations and quadratic cost functionals. To this end, we introduce simple model problems in the next section. In the following chapters, they will repeatedly serve as illustrations of theoretical results. This analysis will be facilitated by the fact that the Hilbert space setting suffices as a functional analytic framework in the case of linear-quadratic theory. The later chapters deal with semilinear equations. Here, the examples under study will be less academic. Owing to the presence of nonlinearities, the mathematical analysis will have to be more delicate.

1.2. Examples of convex problems

1.2.1. Optimal stationary heating.

Optimal boundary heating. Let us consider a body that is to be heated or cooled and which occupies the spatial domain $\Omega \subset \mathbb{R}^3$. We apply to its boundary Γ a heat source u (the *control*), which is constant in time but depends on the location x on the boundary, that is, $u = u(x)$. Our aim is to choose the control in such a way that the corresponding temperature distribution $y = y(x)$ in Ω (the *state*) is the best possible approximation to a desired stationary temperature distribution $y_\Omega = y_\Omega(x)$ in Ω . We can model this in the following way:

$$\min J(y, u) := \frac{1}{2} \int_{\Omega} |y(x) - y_\Omega(x)|^2 dx + \frac{\lambda}{2} \int_{\Gamma} |u(x)|^2 ds(x),$$

subject to the *state equation*

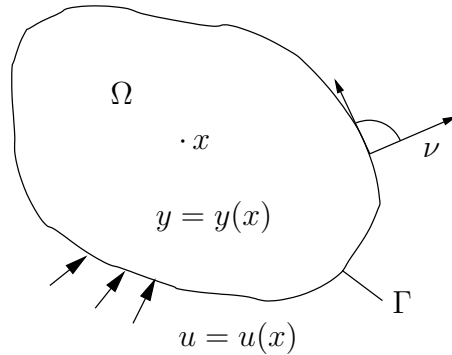
$$\boxed{\begin{array}{ll} -\Delta y & = 0 \quad \text{in } \Omega \\ \frac{\partial y}{\partial \nu} & = \alpha(u - y) \quad \text{on } \Gamma \end{array}}$$

and the *pointwise control constraints*

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{on } \Gamma.$$

Such pointwise bounds for the control are quite natural, since the available capacities for heating or cooling are usually restricted. The constant $\lambda \geq 0$ can be viewed as a measure of the energy costs needed to implement the control u . From the mathematical viewpoint, this term also serves as a regularization parameter; it has the effect that possible optimal controls show improved regularity properties.

Throughout this textbook, we will denote the element of surface area by ds and the outward unit normal to Γ at $x \in \Gamma$ by $\nu(x)$. The function α represents the heat transmission coefficient from Ω to the surrounding medium. The functional J to be minimized is called the *cost functional*. The factor $1/2$ appearing in it has no influence on the solution of the problem. It is introduced just for the sake of convenience: it will later cancel out a factor 2 arising from differentiation. We seek an optimal control $u = u(x)$ together with the associated state $y = y(x)$. The minus sign in front of the Laplacian Δ appears to be unmotivated at first glance. It is introduced because Δ is not a coercive operator, while $-\Delta$ is.



Boundary control.

Observe that in the above problem the cost functional is quadratic, the state is governed by a linear elliptic partial differential equation, and the control acts on the boundary of the domain. We thus have a *linear-quadratic elliptic boundary control problem*.

Remark. The problem is strongly simplified. Indeed, in a realistic model Laplace's equation $\Delta y = 0$ has to be replaced by the stationary heat conduction equation $\operatorname{div}(a \operatorname{grad} y) = 0$, where the coefficient a can depend on x or even on y . If $a = a(y)$ or $a = a(x, y)$, then the partial differential equation is quasilinear. In addition, it will in many cases be more natural to describe the process by a time-dependent partial differential equation.

Optimal heat source. In a similar way, the control can act as a *heat source in the domain* Ω . Problems of this kind arise if the body Ω is heated by electromagnetic induction or by microwaves. Assuming at first that the boundary temperature vanishes, we obtain the following problem:

$$\min J(y, u) := \frac{1}{2} \int_{\Omega} |y(x) - y_{\Omega}(x)|^2 dx + \frac{\lambda}{2} \int_{\Omega} |u(x)|^2 dx,$$

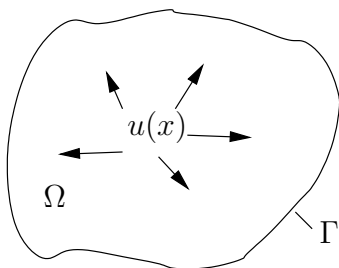
subject to

$$\begin{cases} -\Delta y = \beta u & \text{in } \Omega \\ y = 0 & \text{on } \Gamma \end{cases}$$

and

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{in } \Omega.$$

Here, the coefficient $\beta = \beta(x)$ is prescribed.



Distributed control.

Observe that by the special choice $\beta = \chi_{\Omega_c}$ (where χ_E denotes the characteristic function of a set E), it can be achieved that u acts only in a subdomain $\Omega_c \subset \Omega$. This problem is a *linear-quadratic elliptic control problem with distributed control*. It can be more realistic to prescribe an exterior temperature y_a rather than assume that the boundary temperature vanishes. Then a better model

is given by the state equation

$$\begin{cases} -\Delta y = \beta u & \text{in } \Omega \\ \frac{\partial y}{\partial \nu} = \alpha (y_a - y) & \text{on } \Gamma. \end{cases}$$

1.2.2. Optimal nonstationary boundary control. Let $\Omega \subset \mathbb{R}^3$ represent a potato that is to be roasted over a fire for some period of time $T > 0$. We denote its temperature by $y = y(x, t)$, with $x \in \Omega$, $t \in [0, T]$. Initially, the potato has temperature $y_0 = y_0(x)$, and we want to serve it at a pleasant palatable temperature y_Ω at the final time T . We now introduce notation that will be used throughout this book: we write $Q := \Omega \times (0, T)$ and $\Sigma := \Gamma \times (0, T)$. The problem then reads as follows:

$$\min J(y, u) := \frac{1}{2} \int_{\Omega} |y(x, T) - y_\Omega(x)|^2 dx + \frac{\lambda}{2} \int_0^T \int_{\Gamma} |u(x, t)|^2 ds(x) dt,$$

subject to

$$\begin{cases} y_t - \Delta y = 0 & \text{in } Q \\ \frac{\partial y}{\partial \nu} = \alpha (u - y) & \text{on } \Sigma \\ y(x, 0) = y_0(x) & \text{in } \Omega \end{cases}$$

and

$$u_a(x, t) \leq u(x, t) \leq u_b(x, t) \quad \text{on } \Sigma.$$

By continued turning of the spit, we produce $u(x, t)$. The heating process has to be described by the *nonstationary heat equation*, which is a parabolic differential equation. We thus have to deal with a *linear-quadratic parabolic boundary control problem*. Here and throughout this textbook, y_t denotes the partial derivative of y with respect to t .

1.2.3. Optimal vibrations. Suppose that a group of pedestrians crosses a bridge, trying to excite oscillations in it. This can be modeled (strongly abstracted) as follows: let $\Omega \subset \mathbb{R}^2$ denote the domain of the bridge, $y = y(x, t)$ its transversal displacement, $u = u(x, t)$ the force density acting in the vertical direction, and $y_d = y_d(x, t)$ a desired evolution of the transversal vibrations. We then obtain the optimal control problem

$$\min J(y, u) := \frac{1}{2} \int_0^T \int_{\Omega} |y(x, t) - y_d(x, t)|^2 dx dt + \frac{\lambda}{2} \int_0^T \int_{\Omega} |u(x, t)|^2 dx dt,$$

subject to

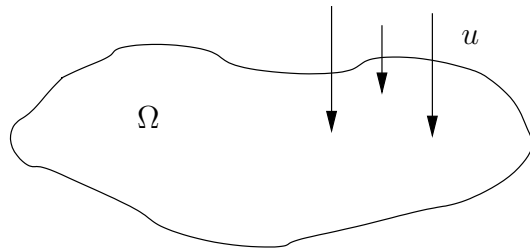
$$\begin{aligned} y_{tt} - \Delta y &= u && \text{in } Q \\ y(0) &= y_0 && \text{in } \Omega \\ y_t(0) &= y_1 && \text{in } \Omega \\ y &= 0 && \text{on } \Sigma \end{aligned}$$

and

$$u_a(x, t) \leq u(x, t) \leq u_b(x, t) \quad \text{in } Q.$$

This is a *linear-quadratic hyperbolic control problem with distributed control*.

Since hyperbolic problems are not the subject of this textbook, we refer the interested reader to the standard monograph by Lions [Lio71], as well as to Ahmed and Teo [AT81]. Interesting control problems for oscillating elastic networks have been treated by Lagnese et al. [LLS94]. An elementary introduction to the controllability of oscillations can be found in the textbook by Krabs [Kra95].



Excitation of vibrations.

In the linear-quadratic case, the theory of hyperbolic problems has many similarities to the parabolic theory studied in this textbook. However, the treatment of semilinear hyperbolic problems is much more difficult, since the smoothing properties of the associated solution operators are weaker.

As a consequence, many of the techniques presented in this book fail in the hyperbolic case.

1.3. Examples of nonconvex problems

So far, we have only considered linear differential equations. However, linear models do not suffice for many real-world phenomena. Instead, one often needs quasilinear or, much simpler, semilinear equations. Recall that a second-order equation is called *semilinear* if the main parts (that is, the expressions involving highest-order derivatives) of the differential operators considered in the domain and on the boundary are linear with respect to the desired solution. For such equations, the theory of optimal control is well developed.

Optimal control problems with semilinear state equations are, as a rule, nonconvex, even if the cost functional is convex. In the following section, we will discuss examples of semilinear state equations. Associated optimal control problems can be obtained by prescribing a cost functional and suitable constraints.

1.3.1. Problems involving semilinear elliptic equations.

Heating with radiation boundary condition. If the heat radiation of the heated body is taken into account, then we obtain a problem with a nonlinear Stefan–Boltzmann boundary condition. In this case, the control u is given by the temperature of the surrounding medium:

$$\begin{aligned} -\Delta y &= 0 && \text{in } \Omega \\ \frac{\partial y}{\partial \nu} &= \alpha(u^4 - y^4) && \text{on } \Gamma. \end{aligned}$$

In this example, the nonlinearity y^4 occurs in the boundary condition, while the heat conduction equation itself is linear.

Simplified superconductivity. The following simplified (Ginzburg–Landau) model for superconductivity was considered by Ito and Kunisch [IK96] to test numerical methods for optimal control problems:

$$\begin{aligned} -\Delta y - y + y^3 &= u && \text{in } \Omega \\ y|_{\Gamma} &= 0 && \text{on } \Gamma. \end{aligned}$$

For analytic reasons, we will later discuss the simpler equation $-\Delta y + y + y^3 = u$, which is also of interest in the theory of superconductivity; see [IK96].

Control of stationary flows. Stationary flows of incompressible media in two- or three-dimensional spatial domains Ω are described by the stationary

Navier–Stokes equations

$$\begin{aligned} -\frac{1}{Re} \Delta u + (u \cdot \nabla) u + \nabla p &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma \\ \operatorname{div} u &= 0 \quad \text{in } \Omega; \end{aligned}$$

see Temam [Tem79] and Galdi [Gal94]. Here, in contrast to the notation used so far, $u = u(x) \in \mathbb{R}^3$ denotes the velocity vector of the particle located at the space point x ; moreover, $p = p(x)$ and $f = f(x)$ represent the pressure and the density of the volume force, respectively. The constant Re is called the *Reynolds number*. In this example, f is the control, and the nonlinearity arises from the first-order differential operator $(u \cdot \nabla)$ being applied to u , which results in (with D_i denoting the partial derivative with respect to x_i)

$$(u \cdot \nabla) u = u_1 D_1 u + u_2 D_2 u + u_3 D_3 u = \sum_{i=1}^3 u_i \begin{bmatrix} D_i u_1 \\ D_i u_2 \\ D_i u_3 \end{bmatrix}.$$

The above mathematical model is of particular interest in relation to electrically conducting fluids that can be influenced by magnetic fields. A possible target for the optimization could be the realization of a desired stationary flow pattern.

1.3.2. Problems involving semilinear parabolic equations.

The examples from Section 1.3.1. Both of the examples involving semilinear elliptic equations discussed in Section 1.3.1 can be formulated in nonstationary form. The first example leads to a parabolic initial-boundary value problem with Stefan–Boltzmann condition for the temperature $y(x, t)$:

$$\begin{aligned} y_t - \Delta y &= 0 && \text{in } Q \\ \frac{\partial y}{\partial \nu} &= \alpha (u^4 - y^4) && \text{on } \Sigma \\ y(\cdot, 0) &= 0 && \text{in } \Omega. \end{aligned}$$

An optimal control problem for a system of this type was initially investigated by Sachs [Sac78]; see also Schmidt [Sch89]. Similarly, a nonstationary analogue of the simplified model for superconductivity can be studied:

$$\begin{aligned} y_t - \Delta y - y + y^3 &= u && \text{in } Q \\ y|_{\Gamma} &= 0 && \text{on } \Sigma \\ y(\cdot, 0) &= 0 && \text{in } \Omega. \end{aligned}$$

A phase field model. Many phase change phenomena (e.g., melting or solidification) can be modeled by systems of *phase field equations* of the

following type:

$$\begin{aligned}
 u_t + \frac{\ell}{2}\varphi_t &= \kappa \Delta u + f && \text{in } Q \\
 \tau\varphi_t &= \xi^2 \Delta \varphi + g(\varphi) + 2u && \text{in } Q \\
 \frac{\partial u}{\partial \nu} &= 0, \quad \frac{\partial \varphi}{\partial \nu} = 0 && \text{on } \Sigma \\
 u(\cdot, 0) &= u_0, \quad \varphi(\cdot, 0) = \varphi_0 && \text{in } \Omega.
 \end{aligned}$$

In a liquid-solid transition, the quantity $u = u(x, t)$ represents a temperature, and the so-called *phase function* $\varphi = \varphi(x, t) \in [-1, 1]$ describes the degree of solidification, where $\{\varphi = 1\}$ and $\{\varphi = -1\}$ correspond to the liquid and solid phases, respectively. The function f represents a controllable heat source, and $-g$ is the derivative of a so-called “double well” potential G . One standard form for G is $G(z) = \frac{1}{8}(z^2 - 1)^2$. In many applications g has the form $g(z) = az + bz^2 - cz^3$, with bounded coefficient functions a, b , and $c > 0$. For the precise physical meaning of the quantities κ, ℓ, τ , and ξ we refer the interested reader to Section 4.4 in the monograph by Brokate and Sprekels [BS96].

In this example, the target of optimization could be the approximation of a desired evolution of the melting/solidification process. First results for related control problems have been published by Chen and Hoffmann [CH91] and by Hoffmann and Jiang [HJ92].

Control of nonstationary flows. Nonstationary flows of incompressible fluids are described by the *nonstationary Navier–Stokes equations*

$$\begin{aligned}
 u_t - \frac{1}{Re} \Delta u + (u \cdot \nabla) u + \nabla p &= f && \text{in } Q \\
 \operatorname{div} u &= 0 && \text{in } Q \\
 u &= 0 && \text{on } \Sigma \\
 u(\cdot, 0) &= u_0 && \text{in } \Omega.
 \end{aligned}$$

Here, a volume force f acts on the fluid, whose velocity is initially equal to u_0 and is zero at the boundary (“no-slip condition”). Depending on the particular circumstances, other boundary conditions may also be of interest. One of the first contributions to the mathematical theory of optimal control of fluid flows is due to Abergel and Temam [AT90].

1.4. Basic concepts for the finite-dimensional case

Some fundamental concepts of optimal control theory can easily be explained by considering optimization problems in Euclidean space with finitely many equality constraints. A little detour into finite-dimensional optimization has

the advantage that the basic ideas will not be complicated by technical details from partial differential equations or functional analysis.

1.4.1. Finite-dimensional optimal control problems. Suppose that $J = J(y, u)$, $J : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, denotes a cost functional to be minimized, and that an $n \times n$ matrix A , an $n \times m$ matrix B , and a nonempty set $U_{ad} \subset \mathbb{R}^m$ are given (where “ad” stands for “admissible”). We consider the *optimization problem*

$$(1.1) \quad \boxed{\begin{array}{l} \min J(y, u) \\ Ay = Bu, \quad u \in U_{ad}. \end{array}}$$

We seek vectors y and u minimizing the cost functional J subject to the constraints $Ay = Bu$ and $u \in U_{ad}$. In this connection, we introduce the following convention: Unless specified otherwise, throughout this book vectors will always be regarded as *column* vectors.

Example. Often quadratic cost functionals are used, for instance

$$J(y, u) = |y - y_d|^2 + \lambda |u|^2,$$

where $|\cdot|$ denotes the Euclidean norm. ◇

As it stands, (1.1) is a standard optimization problem in which the unknowns y and u play similar roles. But this situation changes if we make the additional assumption that the matrix A has an inverse A^{-1} . Indeed, we can then solve for y in (1.1), obtaining

$$(1.2) \quad y = A^{-1}Bu,$$

and for any $u \in \mathbb{R}^m$ there is a uniquely determined solution $y \in \mathbb{R}^n$; that is, we may choose (i.e. “control”) u in an arbitrary way to produce the associated y as a dependent quantity. We therefore call u the control vector or, for short, the *control*, and y the associated state vector or *state*. In this way, (1.1) becomes a finite-dimensional optimal control problem.

Next, we introduce the *solution matrix* of our control system

$$S : \mathbb{R}^m \rightarrow \mathbb{R}^n, \quad S = A^{-1}B.$$

Then $y = Su$, and, owing to (1.2), we can eliminate y from J to obtain the *reduced cost functional* f ,

$$J(y, u) = J(Su, u) =: f(u).$$

For instance, for the quadratic function in the above example we get $f(u) = |Su - y_d|^2 + \lambda |u|^2$. The problem (1.1) then becomes the nonlinear optimization

problem

$$(1.3) \quad \min f(u), \quad u \in U_{ad}.$$

In this *reduced problem* only the control u appears as an unknown.

In the following sections, we will discuss some basic ideas that will be repeatedly encountered in similar forms in the optimal control of partial differential equations.

1.4.2. Existence of optimal controls.

Definition. A vector $\bar{u} \in U_{ad}$ is called an optimal control for problem (1.1) if $f(\bar{u}) \leq f(u)$ for all $u \in U_{ad}$; then $\bar{y} := S\bar{u}$ is called the optimal state associated with \bar{u} .

Optimal or locally optimal quantities will be indicated by overlining, as in \bar{u} .

Theorem 1.1. Suppose that J is continuous on $\mathbb{R}^n \times U_{ad}$ and that the set U_{ad} is nonempty, bounded, and closed. If the matrix A is invertible, then (1.1) has at least one solution.

Proof: Obviously, the continuity of J implies that f is also continuous on U_{ad} . Moreover, as a bounded and closed set in a finite-dimensional space, U_{ad} is compact. By the well-known Weierstrass theorem, f attains its minimum in U_{ad} . Hence, there is some $\bar{u} \in U_{ad}$ such that $f(\bar{u}) = \min_{u \in U_{ad}} f(u)$. \square

This proof becomes more complicated in the case of optimal control problems for partial differential equations, since bounded and closed sets need not be compact in (infinite-dimensional) function spaces.

1.4.3. First-order necessary optimality conditions. In this section, we investigate what conditions the optimal vectors \bar{u} and \bar{y} must satisfy. We do this in the hope that we will be able to extract enough information from these conditions to determine \bar{u} and \bar{y} . Usually, this will have to be done using numerical methods.

Notation. We use the following notation for the derivatives of functions $f : \mathbb{R}^m \rightarrow \mathbb{R}$:

$$\begin{aligned} D_i &= \frac{\partial}{\partial x_i}, & D_x &= \frac{\partial}{\partial x} && \text{(partial derivatives)} \\ f'(x) &= (D_1 f(x), \dots, D_m f(x)) && \text{(derivative)} \\ \nabla f(u) &= f'(u)^\top && \text{(gradient)} \end{aligned}$$

where $^\top$ stands for transposition. For functions $f = f(x, y) : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$, we denote by $D_x f$ the row vector of partial derivatives of f with respect to

x_1, \dots, x_m , and by $\nabla_x f$ the corresponding column vector. The expressions $D_y f$ and $\nabla_y f$ are defined in a similar way. Moreover,

$$(u, v)_{\mathbb{R}^m} = u \cdot v = \sum_{i=1}^m u_i v_i$$

denotes the standard Euclidean scalar product in \mathbb{R}^m . For the sake of convenience, we will use both kinds of notation for the scalar product between vectors. The application of $f'(u)$ to a column vector $h \in \mathbb{R}^m$, denoted by $f'(u)h$, coincides with the directional derivative of f in the direction h ,

$$f'(u)h = (\nabla f(u), h)_{\mathbb{R}^m} = \nabla f(u) \cdot h.$$

We now make the additional assumption that the cost functional J is continuously differentiable with respect to y and u ; that is, the partial derivatives $D_y J(y, u)$ and $D_u J(y, u)$ with respect to y and u are continuous in (y, u) . Then, by virtue of the chain rule, $f(u) = J(Su, u)$ is continuously differentiable.

Example. Suppose that $f(u) = \frac{1}{2}|Su - y_d|^2 + \frac{\lambda}{2}|u|^2$. Then it follows that

$$\begin{aligned} \nabla f(u) &= S^\top(Su - y_d) + \lambda u, & f'(u) &= (S^\top(Su - y_d) + \lambda u)^\top, \\ f'(u)h &= (S^\top(Su - y_d) + \lambda u, h)_{\mathbb{R}^m}. & & \diamond \end{aligned}$$

Theorem 1.2. *Let U_{ad} be convex. Then any optimal control \bar{u} for (1.1) satisfies the variational inequality*

$$(1.4) \quad f'(\bar{u})(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad}.$$

This simple yet fundamental result is a special case of Lemma 2.21 on page 63. It reflects the observation that f cannot decrease in any direction at a minimum point.

Invoking the chain rule and the rules for total differentials, we can determine the derivative f' in (1.4), which is given by $f' = D_y J S + D_u J$. We find that

$$\begin{aligned} f'(\bar{u})h &= D_y J(S\bar{u}, \bar{u})Sh + D_u J(S\bar{u}, \bar{u})h \\ &= (\nabla_y J(\bar{y}, \bar{u}), A^{-1}Bh)_{\mathbb{R}^n} + (\nabla_u J(\bar{y}, \bar{u}), h)_{\mathbb{R}^m} \\ (1.5) \quad &= (B^\top(A^\top)^{-1}\nabla_y J(\bar{y}, \bar{u}) + \nabla_u J(\bar{y}, \bar{u}), h)_{\mathbb{R}^m}. \end{aligned}$$

Hence, the variational inequality (1.4) takes the somewhat clumsy form

$$(1.6) \quad (B^\top(A^\top)^{-1}\nabla_y J(\bar{y}, \bar{u}) + \nabla_u J(\bar{y}, \bar{u}), u - \bar{u})_{\mathbb{R}^m} \geq 0 \quad \forall u \in U_{ad}.$$

It can be considerably simplified by introducing the adjoint state, a simple trick that is of utmost importance in optimal control theory.

1.4.4. Adjoint state and reduced gradient. As motivation, let us assume that the use of the inverse matrix A^{-1} is too costly for numerical calculations. This is usually the case for realistic optimal control problems. Then, a numerical method that avoids the explicit calculation of A^{-1} (e.g., the conjugate gradient method) must be used for the solution of the linear system $Ay = b$. The same applies for A^\top . We therefore replace the term $(A^\top)^{-1}\nabla_y J(\bar{y}, \bar{u})$ in (1.6) by a new variable \bar{p} ,

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

The quantity \bar{p} corresponding to the pair (\bar{y}, \bar{u}) can be determined by solving the linear system

$$(1.7) \quad A^\top \bar{p} = \nabla_y J(\bar{y}, \bar{u}).$$

Definition. *The equation (1.7) is called the adjoint equation, and its solution \bar{p} is called the adjoint state associated with (\bar{y}, \bar{u}) .*

Example. In the case of the quadratic function $J(y, u) = \frac{1}{2}|y - y_d|^2 + \frac{\lambda}{2}|u|^2$, we obtain the adjoint equation

$$A^\top \bar{p} = \bar{y} - y_d,$$

since $\nabla_y J(y, u) = y - y_d$. ◇

The introduction of the adjoint state has two advantages: the first-order necessary optimality conditions simplify, and the use of the inverse matrix $(A^\top)^{-1}$ is avoided. Also, the form of the gradient of f simplifies. Indeed, with $\bar{y} = S\bar{u}$, it follows from (1.5) that

$$\nabla f(\bar{u}) = B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}).$$

The vector $\nabla f(\bar{u})$ is referred to as the *reduced gradient*. Moreover, since $\bar{y} = S\bar{u}$, the directional derivative $f'(\bar{u})h$ at an arbitrary point \bar{u} is given by

$$f'(\bar{u})h = (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}), h)_{\mathbb{R}^m}.$$

The two expressions above involving the adjoint state \bar{p} do not depend on whether \bar{u} is optimal or not. We will encounter them repeatedly in control problems for partial differential equations. Moreover, the use of the adjoint state \bar{p} also simplifies Theorem 1.2:

Theorem 1.3. *Suppose that the matrix A is invertible, and let \bar{u} be an optimal control for (1.1) with associated state \bar{y} . Then the adjoint equation (1.7) has a unique solution \bar{p} such that*

$$(1.8) \quad (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}), u - \bar{u})_{\mathbb{R}^m} \geq 0 \quad \forall u \in U_{ad}.$$

The assertion follows directly from the variational inequality (1.6) and the definition of \bar{p} . In summary, we have derived the following *optimality system* for the unknown vectors \bar{y} , \bar{u} , and \bar{p} , which can be used to determine the optimal control:

$$(1.9) \quad \boxed{\begin{aligned} Ay &= Bu, \quad u \in U_{ad} \\ A^\top p &= \nabla_y J(y, u) \\ (B^\top p + \nabla_u J(y, u), v - u)_{\mathbb{R}^m} &\geq 0 \quad \forall v \in U_{ad}. \end{aligned}}$$

Every solution (\bar{y}, \bar{u}) to the optimal control problem (1.1) must, together with \bar{p} , satisfy this system.

No restrictions on u . In this case, $U_{ad} = \mathbb{R}^m$. Then $u - \bar{u}$ may attain any value $h \in \mathbb{R}^m$, and thus the variational inequality (1.8) reduces to the equation

$$B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) = 0.$$

Example. Suppose that

$$J(y, u) = \frac{1}{2} |Cy - y_d|^2 + \frac{\lambda}{2} |u|^2,$$

with a given $n \times n$ matrix C . Then, obviously,

$$\nabla_y J(y, u) = C^\top (Cy - y_d), \quad \nabla_u J(y, u) = \lambda u.$$

The optimality system becomes

$$\begin{aligned} Ay &= Bu, \quad u \in U_{ad} \\ A^\top p &= C^\top (Cy - y_d) \\ (B^\top p + \lambda u, v - u)_{\mathbb{R}^m} &\geq 0 \quad \forall v \in U_{ad}. \end{aligned}$$

If $U_{ad} = \mathbb{R}^m$, then $B^\top \bar{p} + \lambda \bar{u} = 0$. In the case where $\lambda > 0$, we can solve for \bar{u} to obtain

$$(1.10) \quad \bar{u} = -\frac{1}{\lambda} B^\top \bar{p}.$$

Substitution in the two other relations yields the optimality system

$$\boxed{\begin{aligned} Ay &= -\frac{1}{\lambda} B B^\top p \\ A^\top p &= C^\top (Cy - y_d), \end{aligned}}$$

which is a linear system for the unknowns \bar{y} and \bar{p} . Once \bar{y} and \bar{p} have been recovered from it, the optimal control \bar{u} can be determined from (1.10). \diamond

Remark. We have chosen a linear equation in (1.1) for the sake of simplicity. The fully nonlinear problem

$$(1.11) \quad \min J(y, u), \quad T(y, u) = 0, \quad u \in U_{ad}$$

will be discussed in Exercise 2.1 on page 116.

1.4.5. Lagrangians. By using the Lagrangian function from basic calculus, the optimality system can also be formulated as a *Lagrange multiplier rule*.

Definition. *The function*

$$L : \mathbb{R}^{2n+m} \rightarrow \mathbb{R}, \quad L(y, u, p) := J(y, u) - (Ay - Bu, p)_{\mathbb{R}^n},$$

is called the Lagrangian function or Lagrangian.

Using L , we can formally eliminate the equality constraints from (1.1), while retaining the seemingly simpler restriction $u \in U_{ad}$ in explicit form. Upon comparison, we find that the second and third conditions in the optimality system are equivalent to

$$\begin{aligned} \nabla_y L(\bar{y}, \bar{u}, \bar{p}) &= 0 \\ (\nabla_u L(\bar{y}, \bar{u}, \bar{p}), u - \bar{u})_{\mathbb{R}^m} &\geq 0 \quad \forall u \in U_{ad}. \end{aligned}$$

Conclusion. *The adjoint equation (1.7) is equivalent to $\nabla_y L(\bar{y}, \bar{u}, \bar{p}) = 0$ and thus can be recovered by differentiating the Lagrangian with respect to y . Similarly, the variational inequality follows from differentiation of L with respect to u .*

Consequently, (\bar{y}, \bar{u}) is a solution to the necessary optimality conditions of the following minimization problem without equality constraints:

$$(1.12) \quad \min_{y, u} L(y, u, p), \quad u \in U_{ad}, \quad y \in \mathbb{R}^n.$$

By the way, this does not imply that (\bar{y}, \bar{u}) can always be determined numerically as a solution to (1.12). In fact, the “right” \bar{p} is usually not known, and (1.12) may not be solvable or could even lead to wrong solutions. The vector $\bar{p} \in \mathbb{R}^n$ also plays the role of a *Lagrange multiplier*. It corresponds to the equation $Ay - Bu = 0$.

We remark that the above conclusion remains valid for the fully nonlinear problem (1.11), provided that the Lagrangian is defined by $L(y, u, p) := J(y, u) - (T(y, u), p)_{\mathbb{R}^n}$.

1.4.6. Discussion of the variational inequality. In later chapters the admissible set U_{ad} will be defined by upper and lower bounds, so-called *box constraints*. We assume this here too, i.e.,

$$(1.13) \quad U_{ad} = \{u \in \mathbb{R}^m : u_a \leq u \leq u_b\}.$$

Here, $u_a \leq u_b$ are given vectors in \mathbb{R}^m , where the inequalities are to be understood componentwise, that is, $u_{a,i} \leq u_i \leq u_{b,i}$ for $i = 1, \dots, m$. Rewriting the variational inequality (1.8) as

$$(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}), \bar{u})_{\mathbb{R}^m} \leq (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}), u)_{\mathbb{R}^m} \quad \forall u \in U_{ad},$$

we find that \bar{u} solves the linear optimization problem

$$\min_{u \in U_{ad}} (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}), u)_{\mathbb{R}^m} = \min_{u \in U_{ad}} \sum_{i=1}^m (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i u_i.$$

If U_{ad} is given as in (1.13), then it follows from the fact that the u_i are independent from each other that

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

for $i = 1, \dots, m$. Hence, we must have

$$(1.14) \quad \bar{u}_i = \begin{cases} u_{b,i} & \text{if } (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i < 0 \\ u_{a,i} & \text{if } (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i > 0. \end{cases}$$

No direct information can be recovered from the variational inequality for the components that satisfy $(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i = 0$. However, in many cases useful information can still be extracted simply from the fact that this equation holds.

1.4.7. Formulation as a Karush–Kuhn–Tucker system. Up to now, the Lagrangian L has only been used to eliminate the conditions in equation form. The same can be done with the inequality constraints induced by U_{ad} . To this end, we introduce the quantities

$$(1.15) \quad \begin{aligned} \mu_a &:= (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_+ \\ \mu_b &:= (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_- \end{aligned}$$

We have $\mu_{a,i} = (B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i$ if the right-hand side is positive, and $\mu_{a,i} = 0$ otherwise; likewise, $\mu_{b,i} = |(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i|$ for a negative right-hand side, and $\mu_{b,i} = 0$ otherwise. Invoking (1.14), we deduce the relations

$$\begin{aligned} \mu_a &\geq 0, & u_a - \bar{u} &\leq 0, & (u_a - \bar{u}, \mu_a)_{\mathbb{R}^m} &= 0, \\ \mu_b &\geq 0, & \bar{u} - u_b &\leq 0, & (\bar{u} - u_b, \mu_b)_{\mathbb{R}^m} &= 0. \end{aligned}$$

In optimization theory, these are usually referred to as *complementary slackness conditions* or *complementarity conditions*.

The inequalities hold trivially, so that only the equations have to be verified. We confine ourselves to showing the first orthogonality condition: in view of (1.14), the strict inequality $u_{a,i} < \bar{u}_i$ can only be valid if $(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i \leq 0$. By definition, this implies that $\mu_{a,i} = 0$, hence $(u_{a,i} - \bar{u}_i) \mu_{a,i} = 0$. If $\mu_{a,i} > 0$, then, owing to the definition of μ_a , also $(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i > 0$, and from (1.14) we conclude that $u_{a,i} = \bar{u}_i$. Again, it follows that $(u_{a,i} - \bar{u}_i) \mu_{a,i} = 0$. Summation over i then yields $(u_a - \bar{u}, \mu_a)_{\mathbb{R}^m} = 0$.

Note that (1.15) implies that $\mu_a - \mu_b = \nabla_u J(\bar{y}, \bar{u}) + B^\top \bar{p}$, so that

$$(1.16) \quad \nabla_u J(\bar{y}, \bar{u}) + B^\top \bar{p} - \mu_a + \mu_b = 0.$$

We now introduce an extended Lagrangian \mathcal{L} by adding the inequality constraints in the following way:

$$\begin{aligned} \mathcal{L}(y, u, p, \mu_a, \mu_b) &:= J(y, u) - (Ay - Bu, p)_{\mathbb{R}^n} + (u_a - u, \mu_a)_{\mathbb{R}^m} \\ &\quad + (u - u_b, \mu_b)_{\mathbb{R}^m}. \end{aligned}$$

Then (1.16) can be expressed in the form

$$\nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0.$$

Moreover, the adjoint equation is equivalent to the equation

$$\nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0,$$

since $\nabla_y L = \nabla_y \mathcal{L}$. Hence, μ_a and μ_b are the Lagrange multipliers corresponding to the inequality constraints $u_a - u \leq 0$ and $u - u_b \leq 0$. The optimality conditions can therefore be rewritten in the following alternative form.

Theorem 1.4. *Suppose that A is invertible, U_{ad} is given by (1.13), and \bar{u} is an optimal control for (1.1) with associated state \bar{y} . Then there exist Lagrange multipliers $\bar{p} \in \mathbb{R}^n$ and $\mu_i \in \mathbb{R}^m$, $i = 1, 2$, such that the following conditions hold:*

$\begin{aligned} \nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) &= 0 \\ \nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) &= 0 \\ \mu_a &\geq 0, \quad \mu_b \geq 0 \\ (u_a - \bar{u}, \mu_a)_{\mathbb{R}^m} &= (\bar{u} - u_b, \mu_b)_{\mathbb{R}^m} = 0. \end{aligned}$
--

The above optimality system, which combines the conditions of Theorem 1.4 with the constraints

$$Ay - Bu = 0, \quad u_a \leq u \leq u_b,$$

constitutes the famous *Karush–Kuhn–Tucker conditions*.

In order to be able to compare later with the results in Section 4.10, we are now going to state the second-order sufficient optimality conditions; see, e.g., [GT97b], [GMW81], or [Lue84]. To this end, we introduce index sets corresponding to the *active* inequality constraints, $I(\bar{u}) = I_a(\bar{u}) \cup I_b(\bar{u})$, and to the *strongly active* inequality constraints, $A(\bar{u}) \subset I(\bar{u})$. We have

$$\begin{aligned} I_a(\bar{u}) &= \{i : \bar{u}_i = u_{a,i}\}, & I_b(\bar{u}) &= \{i : \bar{u}_i = u_{b,i}\}, \\ A(\bar{u}) &= \{i : \mu_{a,i} > 0 \text{ or } \mu_{b,i} > 0\}. \end{aligned}$$

Moreover, let $C(\bar{u})$ denote the *critical cone* consisting of all $h \in \mathbb{R}^m$ with the properties

$$\begin{aligned} h_i &= 0 & \text{for } i \in A(\bar{u}) \\ h_i &\geq 0 & \text{for } i \in I_a(\bar{u}) \setminus A(\bar{u}) \\ h_i &\leq 0 & \text{for } i \in I_b(\bar{u}) \setminus A(\bar{u}). \end{aligned}$$

By definition of μ_a and μ_b , we have $i \in A(\bar{u}) \Leftrightarrow |(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}))_i| > 0$.

Hence, an active constraint for u is strongly active if and only if the corresponding component of the gradient of f does not vanish.

Theorem 1.5. *Suppose that U_{ad} is given by (1.13), and let $(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b)$ satisfy the Karush–Kuhn–Tucker conditions. If*

$$\begin{bmatrix} y \\ u \end{bmatrix}^\top \begin{bmatrix} \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) & \mathcal{L}_{yu}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \\ \mathcal{L}_{uy}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) & \mathcal{L}_{uu}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} > 0$$

for all $(y, u) \neq (0, 0)$ with $Ay = Bu$ and $u \in C(\bar{u})$, then (\bar{y}, \bar{u}) is locally optimal for (1.1).

In the above theorem, \mathcal{L}_{yy} , \mathcal{L}_{yu} , and \mathcal{L}_{uu} denote the second-order partial derivatives $D_y^2 \mathcal{L}$, $D_u D_y \mathcal{L}$, and $D_u^2 \mathcal{L}$, respectively. Owing to a standard compactness argument, the definiteness condition of the theorem is equivalent to the existence of some $\delta > 0$ such that

$$\begin{bmatrix} y \\ u \end{bmatrix}^\top \begin{bmatrix} \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) & \mathcal{L}_{yu}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \\ \mathcal{L}_{uy}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) & \mathcal{L}_{uu}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} \geq \delta (|y|^2 + |u|^2)$$

for all corresponding (y, u) . If A is invertible, it even suffices to postulate that the above quadratic form is greater than or equal to $\delta |u|^2$.

Generalization to partial differential equations. In optimal control problems for partial differential equations, the argumentation follows similar lines to that above. In this case, the equation $Ay = Bu$ stands for an elliptic or parabolic boundary value problem, with A being a differential operator and B representing some coefficient or embedding operator. The solution matrix $S = A^{-1}B$ corresponds to the part of the solution operator associated with the differential equation that occurs in the cost functional. The associated optimality conditions will be of the same form as those established above.

Lagrangians are also powerful tools in the control theory of partial differential equations. In the formal Lagrange method, they are used as convenient means to formally derive optimality conditions that can easily be memorized. Their application in the rigorous proof of optimality conditions is not so straightforward; in fact, it is based on the Karush–Kuhn–Tucker theory of optimization problems in Banach spaces, which will be discussed in Chapter 6.

Linear-quadratic elliptic control problems

2.1. Normed spaces

In the first few sections, we present some basic notions from functional analysis. We are guided by the principle of covering only the material that is absolutely necessary for a proper understanding of the subsequent section. The proofs will not be given; in this regard, the interested reader is referred to standard textbooks on functional analysis such as those by Alt [Alt99], Kantorovich and Akilov [KA64], Kreyszig [Kre78], Lusternik and Sobolev [LS74], Wouk [Wou79], or Yosida [Yos80].

We assume that the reader is already familiar with the concept of a linear space over the field \mathbb{R} of real numbers. Standard examples include the n -dimensional Euclidean space \mathbb{R}^n and the space of continuous real-valued functions defined on an interval $[a, b] \subset \mathbb{R}$. Their elements are vectors $x = (x_1, \dots, x_n)^\top$ or functions $x : [a, b] \rightarrow \mathbb{R}$, respectively. In both spaces, operations of addition “+” of two elements and multiplication by real numbers are defined that obey the familiar rules in linear spaces.

Definition. *Let X be a linear space over \mathbb{R} . A mapping $\|\cdot\| : X \rightarrow \mathbb{R}$ is called a norm on X if the following hold for all $x, y \in X$ and $\lambda \in \mathbb{R}$:*

- (i) $\|x\| \geq 0$, and $\|x\| = 0 \Leftrightarrow x = 0$
- (ii) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)
- (iii) $\|\lambda x\| = |\lambda| \|x\|$ (homogeneity)

If $\|\cdot\|$ is a norm on X , then $\{X, \|\cdot\|\}$ is called a (real) normed space.

The space \mathbb{R}^n is a normed space when equipped with the *Euclidean norm*

$$|x| = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}.$$

The space of continuous real-valued mappings $x : [a, b] \rightarrow \mathbb{R}$ is a normed space, denoted by $C[a, b]$, with respect to the *maximum norm* of $x(\cdot)$,

$$\|x\|_{C[a,b]} = \max_{t \in [a,b]} |x(t)|.$$

Another normed space, denoted by $C_{L^2}[a, b]$, is obtained if we endow the space of continuous real-valued functions with the L^2 norm,

$$\|x\|_{C_{L^2}[a,b]} = \left(\int_a^b |x(t)|^2 dt \right)^{1/2}.$$

The reader will be asked in Exercise 2.2 to verify that the norm axioms (i)–(iii) are satisfied for the latter two examples.

Remark. By definition, the space X and the associated norm together define a normed space. The introduction of another norm on the same space leads to a different normed space. However, it is usually clear which norm is under consideration; in such a situation, we will simply refer to the normed space X without making any reference to the particular norm.

Definition. Let $\{X, \|\cdot\|\}$ be a normed space, and let $\{x_n\}_{n=1}^\infty \subset X$ be a sequence.

- (i) The sequence is said to be *convergent* if there is some $x \in X$ such that $\lim_{n \rightarrow \infty} \|x_n - x\| = 0$.
- (ii) We call x the *limit* of the sequence, written as $\lim_{n \rightarrow \infty} x_n = x$.
- (iii) The sequence is called a *Cauchy sequence* if for any $\varepsilon > 0$ there is some $n_0 = n_0(\varepsilon) \in \mathbb{N}$ such that $\|x_n - x_m\| \leq \varepsilon$ for all $n > n_0(\varepsilon)$ and $m > n_0(\varepsilon)$.

Any convergent sequence in a normed space is also a Cauchy sequence, but the converse is in general false, as the following example shows.

Example. Consider the sequence of functions in the space $C_{L^2}[0, 2]$ defined by $x_n(t) = \min\{1, t^n\}$ for $t \in [0, 2]$, $n \in \mathbb{N}$. Then

$$\begin{aligned} \|x_n - x_m\|_{C_{L^2}[0,2]}^2 &= \int_0^1 (t^n - t^m)^2 dt = \int_0^1 (t^{2n} - 2t^{n+m} + t^{2m}) dt \\ &= \frac{1}{2n+1} - \frac{2}{n+m+1} + \frac{1}{2m+1} \leq \frac{2}{2m+1} \end{aligned}$$

for $m \leq n$. Hence, we have a Cauchy sequence. However, its pointwise limit

$$x(t) = \lim_{n \rightarrow \infty} x_n(t) = \begin{cases} 0 & 0 \leq t < 1 \\ 1 & 1 \leq t \leq 2 \end{cases}$$

is not continuous on $[0, 2]$ and thus not an element of $C_{L^2}[0, 2]$. \diamond

Definition. A normed space $\{X, \|\cdot\|\}$ is said to be complete if every Cauchy sequence in X converges, i.e., has a limit in X . A complete normed space is called a Banach space.

The spaces \mathbb{R}^n and $C[a, b]$ are Banach spaces with respect to their natural norms $|\cdot|$ and $\|\cdot\|_{C[a,b]}$, while $\{C_{L^2}[a, b], \|\cdot\|_{C_{L^2}[a,b]}\}$ is not complete and hence not a Banach space.

In a Banach space there does not necessarily exist an equivalent to the scalar product of two vectors in \mathbb{R}^n , which is fundamental for the concept of orthogonality.

Definition. Let H be a real linear space. A mapping $(\cdot, \cdot) : H \rightarrow \mathbb{R}$ is called a scalar product on H if the following conditions are satisfied for all $u, v, u_1, u_2 \in H$ and $\lambda \in \mathbb{R}$:

- (i) $(u, u) \geq 0$, and $(u, u) = 0 \Leftrightarrow u = 0$
- (ii) $(u, v) = (v, u)$
- (iii) $(u_1 + u_2, v) = (u_1, v) + (u_2, v)$
- (iv) $(\lambda u, v) = \lambda (u, v)$.

If (\cdot, \cdot) is a scalar product on H , then $\{H, (\cdot, \cdot)\}$ is called a pre-Hilbert space.

Remark. Again, we speak of the pre-Hilbert space H instead of $\{H, (\cdot, \cdot)\}$ if it is clear which scalar product is being considered on H .

The space \mathbb{R}^n is a pre-Hilbert space with respect to the scalar product $(u, v) := u^\top v$, and $C_{L^2}[a, b]$ is a pre-Hilbert space when equipped with the scalar product

$$(u, v) = \int_a^b u(t)v(t) dt.$$

Every pre-Hilbert space $\{H, (\cdot, \cdot)\}$ is a normed space with respect to its natural norm (see Exercise 2.3)

$$\|u\| := \sqrt{(u, u)}.$$

We then have the *Cauchy–Schwarz inequality*:

$$|(u, v)| \leq \|u\| \|v\| \quad \forall u, v \in H.$$

Definition. A pre-Hilbert space $\{H, (\cdot, \cdot)\}$ is called a Hilbert space if it is complete with respect to the norm

$$\|u\| := \sqrt{(u, u)}.$$

The Euclidean space \mathbb{R}^n is a Hilbert space with respect to the standard scalar product, while $C_{L^2}[a, b]$ is not complete and hence not a Hilbert space.

2.2. Sobolev spaces

In this section, we will recall basic notions from the theory of L^p spaces and Sobolev spaces, which are indispensable prerequisites for the next chapters. In the following, $E \subset \mathbb{R}^N$ denotes a nonempty, bounded, and Lebesgue measurable set having the N -dimensional Lebesgue measure $|E|$.

2.2.1. L^p spaces.

Definition. We denote by $L^p(E)$, $1 \leq p < \infty$, the linear space of all (equivalence classes of) Lebesgue measurable functions y that satisfy

$$\int_E |y(x)|^p dx < \infty.$$

In this connection, functions that differ only on a set of zero measure are identified with each other and considered to belong to the same equivalence class. Endowed with the norm

$$\|y\|_{L^p(E)} = \left(\int_E |y(x)|^p dx \right)^{1/p},$$

$L^p(E)$, with $1 < p < \infty$, becomes a Banach space which is reflexive (this notion will be defined in Section 2.4).

Definition. We denote by $L^\infty(E)$ the Banach space of all (equivalence classes of) Lebesgue measurable and essentially bounded functions, equipped with the norm

$$\|y\|_{L^\infty(E)} = \operatorname{ess\,sup}_{x \in E} |y(x)| := \inf_{|F|=0} \left(\sup_{x \in E \setminus F} |y(x)| \right).$$

By “ess sup” we mean the *essential* maximum or supremum of a function. This excludes any maxima that change upon the removal of single points that are isolated in a certain sense and thus not essential. For instance, the function $y : [0, 1] \rightarrow \mathbb{R}$ which attains the values zero on $(0, 1]$ and one at $x = 0$ has maximum 1 but essential supremum 0.

In the following, $\Omega \subset \mathbb{R}^N$ is a *domain*, i.e., an open and connected set, whose boundary is generally denoted by Γ . Moreover, $v : \Omega \rightarrow \mathbb{R}$ is a function defined in Ω , and the closure of a set E will be denoted by \bar{E} .

Definition.

(i) Let $k \in \mathbb{N}$. We denote by $C^k(\Omega)$ the linear space of all real-valued functions on Ω that, together with their partial derivatives up to order k , are continuous in Ω .

(ii) The set $\text{supp } v = \overline{\{x \in \Omega : v(x) \neq 0\}}$ is called the support of v . It is the smallest closed set outside of which v vanishes identically.

(iii) $C_0^k(\Omega)$, $k \in \mathbb{N} \cup \{0, \infty\}$, denotes the set of all k -times continuously differentiable functions with compact support in Ω .

The case of $k = \infty$, i.e., the set $C_0^\infty(\Omega)$ of so-called *test functions*, is of special interest to us. Test functions vanish on the boundary Γ and thus yield zero boundary integrals upon integration by parts; on the other hand, they can be differentiated up to arbitrary order. Both of these properties will be exploited in the definition of Sobolev spaces. We remark that, since the topology of $C_0^\infty(\Omega)$ will not be needed here, we have used the notion of *set* instead of *space* for $C_0^\infty(\Omega)$.

Next, we recall the notion of *multi-indices*, i.e., vectors $\alpha = (\alpha_1, \dots, \alpha_N)$ having nonnegative integer components. The number $|\alpha| = \alpha_1 + \dots + \alpha_N$ is called the *length* of the multi-index. The components α_i are used to indicate how often a function has to be differentiated with respect to x_i . For example, the multi-index $\alpha = (1, 0, 2)$ means that we have to differentiate once with respect to x_1 and twice with respect to x_3 , but not with respect to x_2 , that is,

$$D^{(1,0,2)}y = \frac{\partial^3 y}{\partial x_1 \partial x_3^2}.$$

Hence, $D^\alpha y(x)$ is shorthand for $D_1^{\alpha_1} \dots D_N^{\alpha_N} y(x)$, and the length $|\alpha|$ represents the total order of differentiation. We put $D^{(0)}y := y$.

Definition. Let $\Omega \subset \mathbb{R}^N$ be bounded. For any $k \in \mathbb{N} \cup \{0\}$, we denote by $C^k(\bar{\Omega})$ the linear space of all elements of $C^k(\Omega)$ that together with their partial derivatives up to order k can be continuously extended to $\bar{\Omega}$. In the $k = 0$ case, we write simply $C(\bar{\Omega})$ instead of $C^0(\bar{\Omega})$.

The spaces $C^k(\bar{\Omega})$ are Banach spaces with respect to the following norms:

$$\|y\|_{C(\bar{\Omega})} = \max_{x \in \bar{\Omega}} |y(x)|, \quad \|y\|_{C^k(\bar{\Omega})} = \sum_{|\alpha| \leq k} \|D^\alpha y\|_{C(\bar{\Omega})}, \quad \text{for } k \in \mathbb{N}.$$

2.2.2. Regular domains. The theory of partial differential equations requires the spatial domains Ω to have sufficiently smooth boundary. The following definition is given in the books by Nečas [Nec67], Ladyzhenskaya et al. [LSU68], Gajewski et al. [GGZ74], and Adams [Ada78]. Comprehensive treatment of Lipschitz domains can be found in the monographs by Alt [Alt99], Grisvard [Gri85], and Wloka [Wlo87].

Definition. Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$, be a bounded domain with boundary Γ . We say that Ω , or Γ , belongs to the class $C^{k,1}$, $k \in \mathbb{N} \cup \{0\}$, if there exist finitely many local coordinate systems S_1, \dots, S_M , functions h_1, \dots, h_M , and numbers $a > 0$ and $b > 0$ that have the following properties:

- (i) The functions h_i , $1 \leq i \leq M$, are k -times differentiable on the closed $(N - 1)$ -dimensional cube

$$\bar{Q}_{N-1} = \{y = (y_1, \dots, y_{N-1}) : |y_i| \leq a, i = 1 \dots N - 1\},$$

and the partial derivatives of order k are Lipschitz continuous on \bar{Q}_{N-1} .

- (ii) For any $P \in \Gamma$ there is some $i \in \{1, \dots, M\}$ such that in the coordinate system S_i there is some $y \in Q_{N-1}$ with $P = (y, h_i(y))$.
- (iii) In the local coordinate system S_i we have

$$(y, y_N) \in \Omega \Leftrightarrow y \in \bar{Q}_{N-1}, h_i(y) < y_N < h_i(y) + b;$$

$$(y, y_N) \notin \Omega \Leftrightarrow y \in \bar{Q}_{N-1}, h_i(y) - b < y_N < h_i(y).$$

The geometrical meaning of condition (iii) is that the domain lies locally on one side of the boundary. Domains and boundaries of class $C^{0,1}$ are called *Lipschitz domains* (or *regular domains*) and *Lipschitz boundaries*, respectively. Boundaries of class $C^{k,1}$ are referred to as $C^{k,1}$ boundaries.

Using the local coordinate systems S_i , we can introduce a Lebesgue measure on Γ in a natural way. To this end, suppose that the set $E \subset \Gamma$ can be completely represented by the coordinate system S_i , that is, for every $P \in E$ there is some $y \in Q_{N-1}$ such that $P = (y, h_i(y))$. Moreover, let $D = (h_i)^{-1}(E) \subset \bar{Q}_{N-1}$ denote the counter-image of E . Then the set E is called *measurable* if D is measurable with respect to the $(N - 1)$ -dimensional Lebesgue measure. The measure of E is then defined by

$$|E| = \int_D \sqrt{1 + |\nabla h_i(y_1, \dots, y_{N-1})|^2} dy_1 \dots dy_{N-1};$$

see [Ada78] or [GGZ74]. For a set E whose representation requires several different local coordinate systems, the measure will be put together appropriately by using a suitable partition of unity. We also use the fact that the Lipschitz function h_i is almost everywhere differentiable by Rademacher's

theorem (see [Alt99] or [Cas92]). Having defined the surface measure, we can proceed in the usual way to introduce the notions of measurable and integrable functions on Γ . We denote the surface measure by $ds(x)$ or ds .

2.2.3. Weak derivatives and Sobolev spaces. In bounded Lipschitz domains Ω , Gauss's theorem is valid. In particular, for $y, v \in C^1(\bar{\Omega})$ we have the *integration by parts formula*

$$\int_{\Omega} v(x) D_i y(x) dx = \int_{\Gamma} v(x) y(x) \nu_i(x) ds(x) - \int_{\Omega} y(x) D_i v(x) dx.$$

Here, $\nu_i(x)$ denotes the i th component of the outward unit normal $\nu(x)$ to Γ at $x \in \Gamma$, and ds is the Lebesgue surface measure on Γ . If, in addition, $v = 0$ on Γ , then it follows that

$$\int_{\Omega} y(x) D_i v(x) dx = - \int_{\Omega} v(x) D_i y(x) dx.$$

More generally, if $y \in C^k(\bar{\Omega})$, $v \in C_0^k(\Omega)$, and some multi-index α of length $|\alpha| \leq k$ are given, then repeated integration by parts yields

$$\int_{\Omega} y(x) D^\alpha v(x) dx = (-1)^{|\alpha|} \int_{\Omega} v(x) D^\alpha y(x) dx.$$

This relation motivates a generalization of the classical notion of derivatives that will be explained now. To this end, we denote by $L_{loc}^1(\Omega)$ the set of all *locally integrable* functions in Ω , that is, the set of all functions that are Lebesgue integrable on every compact subset of Ω .

Definition. Let $y \in L_{loc}^1(\Omega)$ and some multi-index α be given. If a function $w \in L_{loc}^1(\Omega)$ satisfies

$$(2.1) \quad \int_{\Omega} y(x) D^\alpha v(x) dx = (-1)^{|\alpha|} \int_{\Omega} w(x) v(x) dx \quad \forall v \in C_0^\infty(\Omega),$$

then w is called the *weak derivative of y (associated with α)*.

In other words, w is the weak derivative of y if it satisfies the formula of integration by parts in the same manner as the (strong) derivative $D^\alpha y$ would if y belonged to $C^k(\bar{\Omega})$. This observation and the easily proven fact that y can have at most one weak derivative justify our henceforth denoting the weak derivative by the same symbol as the strong one, that is, we write $w = D^\alpha y$.

Example. Consider the function $y(x) = |x|$ in $\Omega = (-1, 1)$. We can easily check that the first-order weak derivative is given by

$$y'(x) := w(x) = \begin{cases} -1, & x \in (-1, 0) \\ +1, & x \in [0, 1). \end{cases}$$

Indeed, we obtain for each $v \in C_0^\infty(-1, 1)$ that

$$\begin{aligned} \int_{-1}^1 |x| v'(x) dx &= \int_{-1}^0 (-x) v'(x) dx + \int_0^1 x v'(x) dx \\ &= -x v(x) \Big|_{-1}^0 - \int_{-1}^0 (-1) v(x) dx + x v(x) \Big|_0^1 - \int_0^1 (+1) v(x) dx \\ &= - \int_{-1}^1 w(x) v(x) dx. \end{aligned}$$

Note that the value of y' at $x = 0$ is immaterial, since an isolated point has zero measure. \diamond

Weak derivatives do not necessarily exist. However, if they do, then they may belong to “better” spaces than merely $L_{loc}^1(\Omega)$, e.g., to the space $L^p(\Omega)$. This gives rise to the following notion:

Definition. Let $1 \leq p < \infty$ and $k \in \mathbb{N}$. We denote by $W^{k,p}(\Omega)$ the linear space of all functions $y \in L^p(\Omega)$ having weak derivatives $D^\alpha y$ in $L^p(\Omega)$ for all multi-indices α of length $|\alpha| \leq k$, endowed with the norm

$$\|y\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \int_{\Omega} |D^\alpha y(x)|^p dx \right)^{1/p}.$$

Analogously, for $p = \infty$, $W^{k,\infty}(\Omega)$ is defined, equipped with the norm

$$\|y\|_{W^{k,\infty}(\Omega)} = \max_{|\alpha| \leq k} \|D^\alpha y\|_{L^\infty(\Omega)}.$$

The spaces $W^{k,p}(\Omega)$ are Banach spaces (see, e.g., [Ada78], [Wlo87]). They are referred to as *Sobolev spaces*. For the particularly interesting case of $p = 2$, we write

$$H^k(\Omega) := W^{k,2}(\Omega).$$

Since $H^1(\Omega)$ is of special importance for our purposes, we repeat the definition given above more explicitly for this space. We have

$$H^1(\Omega) = \{y \in L^2(\Omega) : D_i y \in L^2(\Omega), i = 1, \dots, N\},$$

and the norm is given by

$$\|y\|_{H^1(\Omega)} = \left(\int_{\Omega} (y^2 + |\nabla y|^2) dx \right)^{1/2},$$

where $|\nabla y|^2 = (D_1 y)^2 + \dots + (D_N y)^2$. With the scalar product

$$(u, v)_{H^1(\Omega)} = \int_{\Omega} u v dx + \int_{\Omega} \nabla u \cdot \nabla v dx,$$

$H^1(\Omega)$ becomes a Hilbert space.

A hidden difficulty arises when one wants to assign boundary values to functions from Sobolev spaces. For instance, how do we interpret the statement that a function $y \in W^{k,p}(\Omega)$ vanishes on Γ ? After all, since Γ , as a subset of \mathbb{R}^N , has zero measure, the values of any function $y \in L^p(\Omega)$ can be changed arbitrarily on Γ without affecting y as an element of $L^p(\Omega)$; indeed, functions that have equal values except on a set of zero measure are regarded as equal in the sense of $L^p(\Omega)$.

We now recall the notion of the *closure* of a set $E \subset X$ in a normed space $\{X, \|\cdot\|\}$, which is by definition the set

$$\bar{E} = \{x \in X : x \text{ is the limit of some sequence } \{x_n\}_{n=1}^\infty \subset E\}.$$

We say that a set $E \subset X$ is *dense* in X if $\bar{E} = X$. With this notion, we can define another class of Sobolev spaces.

Definition. *The closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. Moreover, we put $H_0^k(\Omega) := W_0^{k,2}(\Omega)$.*

Obviously, $W_0^{k,p}(\Omega)$, endowed with the norm $\|\cdot\|_{W^{k,p}(\Omega)}$, is a normed space and, as a closed subspace of $W^{k,p}(\Omega)$, also a Banach space. Also note that by definition $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$.

The elements of $W_0^{k,p}(\Omega)$ can be regarded as functions for which all derivatives up to order $k-1$ vanish at the boundary. This is a consequence of the following result, which answers the question of in what sense functions from $W^{k,p}(\Omega)$ have boundary values.

Theorem 2.1 (Trace theorem). *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain and let $1 \leq p \leq \infty$. Then there exists a linear and continuous mapping $\tau : W^{1,p}(\Omega) \rightarrow L^p(\Gamma)$ such that for all $y \in W^{1,p}(\Omega) \cap C(\bar{\Omega})$ we have $(\tau y)(x) = y(x)$ for all $x \in \Gamma$.*

In particular, for $p = 2$ it follows that $\tau : H^1(\Omega) \rightarrow L^2(\Gamma)$. In the case of continuous functions, τy coincides with the restriction $y|_\Gamma$ of y to Γ .

The proof of the trace theorem can be found, e.g., in the monographs of Adams [Ada78], Evans [Eva98], Nečas [Nec67], and Wloka [Wlo87]. We note that it follows from the embedding result Theorem 7.1 on page 355 that for $p > N$, the elements of $W^{1,p}(\Omega)$ can be identified with elements of $C(\bar{\Omega})$. In this case, τ defines a continuous mapping from $W^{1,p}(\Omega)$ into $C(\Gamma)$.

Definition. *The element τy is called the trace of y on Γ , and the mapping τ is called the trace operator.*

Remark. In the following we will, for the sake of simplicity, use the notation $y|_\Gamma$ in place of τy . In this sense, $y|_{\Gamma_0}$ is, for measurable subsets $\Gamma_0 \subset \Gamma$, defined as the restriction of τy to Γ_0 .

Since the trace operator is continuous and thus bounded, there exists some constant $c_\tau = c_\tau(\Omega, p)$ such that

$$\|y|_\Gamma\|_{L^p(\Gamma)} \leq c_\tau \|y\|_{W^{1,p}(\Omega)} \quad \forall y \in W^{1,p}(\Omega).$$

Moreover, for bounded Lipschitz domains Ω it follows that

$$H_0^1(\Omega) = \{y \in H^1(\Omega) : y|_\Gamma = 0\};$$

see, e.g., [Ada78] or [Wlo87]. Finally, we note that in $H_0^1(\Omega)$ a norm can be defined by

$$\|y\|_{H_0^1(\Omega)}^2 := \int_\Omega |\nabla y|^2 dx,$$

which turns out to be equivalent to the norm in $H^1(\Omega)$. Consequently, there are suitable positive constants c_1 and c_2 such that

$$c_1 \|y\|_{H_0^1(\Omega)} \leq \|y\|_{H^1(\Omega)} \leq c_2 \|y\|_{H_0^1(\Omega)} \quad \forall y \in H_0^1(\Omega);$$

cf. the estimate (2.10) on page 33 and the remark following it.

2.3. Weak solutions to elliptic equations

In order to keep the exposition to a reasonable length, we shall not give a comprehensive treatment of elliptic boundary value problems. Instead, we confine ourselves to a few types of elliptic equations, for which basic concepts of optimal control theory will be developed later in this book; in particular, we shall focus on equations containing the Laplacian or, more generally, differential operators in divergence form. In this section, we generally assume that $\Omega \subset \mathbb{R}^N$, $N \geq 2$, is a bounded Lipschitz domain with boundary Γ .

2.3.1. Poisson's equation. We begin our study with the elliptic boundary value problem

$$(2.2) \quad \boxed{\begin{array}{ll} -\Delta y = f & \text{in } \Omega \\ y = 0 & \text{on } \Gamma, \end{array}}$$

where $f \in L^2(\Omega)$ is given. Such functions f may be very irregular. For example, imagine that the open unit square $\Omega \subset \mathbb{R}^2$ is divided into square subdomains in the form of a chessboard, and that f equals unity on the black squares and zero on the others. Since the boundaries between the subdomains have zero Lebesgue measure, and since functions belonging to

$L^2(\Omega)$ cannot be distinguished on sets of zero measure, we do not have to specify the values of f on the interior boundaries.

Obviously, Poisson's equation $-\Delta y = f$ cannot have a classical solution $y \in C^2(\Omega) \cap C^1(\bar{\Omega})$ for such an f . Instead, we seek a *weak solution* y in the space $H_0^1(\Omega)$. Its definition is based on a *variational* formulation of (2.2).

To this end, we assume for the time being that f is sufficiently smooth and that $y \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is a classical solution to (2.2). The domain Ω is generally assumed to be bounded. Multiplying Poisson's equation by an arbitrary test function $v \in C_0^\infty(\Omega)$ and integrating over Ω , we obtain

$$-\int_{\Omega} v \Delta y \, dx = \int_{\Omega} f v \, dx,$$

whence, upon using integration by parts,

$$-\int_{\Gamma} v \partial_{\nu} y \, ds + \int_{\Omega} \nabla y \cdot \nabla v \, dx = \int_{\Omega} f v \, dx.$$

Here, $\partial_{\nu} y$ denotes the normal derivative of y , i.e., the directional derivative of v in the direction of the outward unit normal ν to Γ . Recall that $\partial_{\nu} y = \nabla y \cdot \nu$. Since v vanishes on Γ , it follows that

$$\int_{\Omega} \nabla y \cdot \nabla v \, dx = \int_{\Omega} f v \, dx.$$

Note that this equation holds for *any* $v \in C_0^\infty(\Omega)$. Recalling that $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, and observing that for fixed y all expressions in the equation depend continuously on $v \in H_0^1(\Omega)$, we conclude its validity for all $v \in H_0^1(\Omega)$. Conversely, one can show that any sufficiently smooth $y \in H_0^1(\Omega)$ satisfying the above equation for each $v \in C_0^\infty(\Omega)$ is a classical solution to Poisson's equation $-\Delta y = f$. In summary, the following definition is justified:

Definition. We call $y \in H_0^1(\Omega)$ a *weak solution to the boundary value problem (2.2)* if it satisfies the so-called *weak or variational formulation*

$$(2.3) \quad \int_{\Omega} \nabla y \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega).$$

Equation (2.3) is also referred to as a *variational equality*. The boundary condition $y|_{\Gamma} = 0$ is encoded in the definition of the solution space $H_0^1(\Omega)$. It is remarkable that only (weak) first-order derivatives are needed for a second-order equation.

In order to be able to treat equations more general than Poisson's with a unified approach, we write $V = H_0^1(\Omega)$ and define the *bilinear form*

$a : V \times V \rightarrow \mathbb{R}$,

$$(2.4) \quad a[y, v] := \int_{\Omega} \nabla y \cdot \nabla v \, dx.$$

Then the weak formulation (2.3) can be rewritten in the abstract form

$$a[y, v] = (f, v)_{L^2(\Omega)} \quad \forall v \in V.$$

Next, we define the linear and continuous *functional* (for this notion, see Section 2.4) $F : V \rightarrow \mathbb{R}$,

$$F(v) := (f, v)_{L^2(\Omega)}.$$

Then (2.3) attains the general form

$$(2.5) \quad \boxed{a[y, v] = F(v) \quad \forall v \in V.}$$

We denote by V^* the dual space of V , the space of all linear and continuous functionals on V (see page 42); hence, $F \in V^*$. The following result is of fundamental importance to the existence theory for linear elliptic equations. It forms the basis for the proof of the existence and uniqueness of a weak solution to (2.2), as well as to the other linear elliptic boundary value problems investigated in this book.

Lemma 2.2 (Lax and Milgram). *Let V be a real Hilbert space, and let $a : V \times V \rightarrow \mathbb{R}$ denote a bilinear form. Moreover, suppose that there exist positive constants α_0 and β_0 such that the following conditions are satisfied for all $v, y \in V$:*

$$(2.6) \quad |a[y, v]| \leq \alpha_0 \|y\|_V \|v\|_V \quad (\text{boundedness})$$

$$(2.7) \quad a[y, y] \geq \beta_0 \|y\|_V^2 \quad (V\text{-ellipticity}).$$

Then for every $F \in V^$ the variational equation (2.5) admits a unique solution $y \in V$. Moreover, there is some constant $c_a > 0$, which does not depend on F , such that*

$$(2.8) \quad \|y\|_V \leq c_a \|F\|_{V^*}.$$

The application of the Lax–Milgram lemma to the case of homogeneous Dirichlet boundary conditions $y|_{\Gamma} = 0$ requires the following estimate.

Lemma 2.3 (Friedrichs inequality). For any bounded Lipschitz domain Ω there is a constant $c(\Omega) > 0$, which depends only on the domain Ω , such that

$$\int_{\Omega} |y|^2 dx \leq c(\Omega) \int_{\Omega} |\nabla y|^2 dx \quad \forall y \in H_0^1(\Omega).$$

The proof of this lemma can be found, e.g., in Alt [Alt99], Casas [Cas92], Nečas [Nec67], and Wloka [Wlo87]. Observe that the validity of the Friedrichs inequality is restricted to functions with zero boundary values, i.e. those in $H_0^1(\Omega)$; it cannot hold for general functions in $H^1(\Omega)$, as the counterexample $y(x) \equiv 1$ shows.

Theorem 2.4. *If Ω is a bounded Lipschitz domain, then for every $f \in L^2(\Omega)$ problem (2.2) has a unique weak solution $y \in H_0^1(\Omega)$. Moreover, there is a constant $c_P > 0$, which does not depend on f , such that*

$$(2.9) \quad \|y\|_{H^1(\Omega)} \leq c_P \|f\|_{L^2(\Omega)}.$$

Proof: We apply the Lax–Milgram lemma in $V = H_0^1(\Omega)$. To this end, we verify that the bilinear form (2.4) satisfies the conditions (2.6) and (2.7). Since $H_0^1(\Omega)$ is a subspace of $H^1(\Omega)$, we use the standard H^1 norm; see, however, Remark (i) following this proof. The boundedness condition (2.6) for a follows from the Cauchy–Schwarz inequality:

$$\begin{aligned} \left| \int_{\Omega} \nabla y \cdot \nabla v dx \right| &\leq \left(\int_{\Omega} |\nabla y|^2 dx \right)^{1/2} \left(\int_{\Omega} |\nabla v|^2 dx \right)^{1/2} \\ &\leq \left(\int_{\Omega} (|y|^2 + |\nabla y|^2) dx \right)^{1/2} \left(\int_{\Omega} (|v|^2 + |\nabla v|^2) dx \right)^{1/2} \\ &\leq \|y\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \end{aligned}$$

To show the V -ellipticity, we estimate, using the Friedrichs inequality,

$$\begin{aligned} a[y, y] = \int_{\Omega} |\nabla y|^2 dx &= \frac{1}{2} \int_{\Omega} |\nabla y|^2 dx + \frac{1}{2} \int_{\Omega} |\nabla y|^2 dx \\ &\geq \frac{1}{2} \int_{\Omega} |\nabla y|^2 dx + \frac{1}{2c(\Omega)} \int_{\Omega} |y|^2 dx \\ (2.10) \quad &\geq \frac{1}{2} \min \{1, c(\Omega)^{-1}\} \|y\|_{H^1(\Omega)}^2. \end{aligned}$$

Hence, the assumptions of Lemma 2.2 are satisfied in $V = H_0^1(\Omega)$. The boundedness of the functional F is again a consequence of the Cauchy–Schwarz inequality. Indeed, we have

$$|F(v)| = |(f, v)_{L^2(\Omega)}| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)},$$

so that $\|F\|_{V^*} \leq \|f\|_{L^2(\Omega)}$. Lemma 2.2 yields the existence of a unique solution y to (2.2). Inserting the above estimate for F in (2.8), we conclude that $\|y\|_{H^1(\Omega)} \leq c_a \|F\|_{V^*} \leq c_a \|f\|_{L^2(\Omega)}$, which proves (2.9). \square

Remarks.

(i) Inequality (2.10) shows that

$$\|y\|_{H_0^1(\Omega)} := \left(\int_{\Omega} |\nabla y|^2 dx \right)^{1/2}$$

defines a norm in $H_0^1(\Omega)$ that is equivalent to the standard norm of $H^1(\Omega)$. If $V = H_0^1(\Omega)$ is endowed with this norm a priori, then the assumptions of Lemma 2.2 are directly fulfilled. This is one reason why $\|y\|_{H_0^1(\Omega)}$ is frequently used.

(ii) The Lax–Milgram lemma is also valid for functionals $F \in V^*$ that are not generated by some $f \in L^2(\Omega)$. This fact will be used in the next section.

2.3.2. Boundary conditions of the third kind. In a similar way, we can treat the boundary value problem

$$(2.11) \quad \boxed{\begin{array}{ll} -\Delta y + c_0 y & = f \quad \text{in } \Omega \\ \partial_\nu y + \alpha y & = g \quad \text{on } \Gamma. \end{array}}$$

Here, the functions $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$, as well as the nonnegative coefficient functions $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^\infty(\Gamma)$, are prescribed. The boundary condition in (2.11) is usually referred to as a *boundary condition of the third kind* or a *Robin boundary condition*. Again, ∂_ν denotes the directional derivative in the direction of the outward unit normal ν to Γ .

The above problem is treated in a similar way as (2.2). We multiply the partial differential equation by an arbitrary $v \in C^1(\bar{\Omega})$. Under the same assumptions as in Section 2.3.1, integration by parts leads to

$$- \int_{\Gamma} v \partial_\nu y ds + \int_{\Omega} \nabla y \cdot \nabla v dx + \int_{\Omega} c_0 y v dx = \int_{\Omega} f v dx.$$

Substitution of the boundary condition $\partial_\nu y = g - \alpha y$ then yields that

$$(2.12) \quad \int_{\Omega} \nabla y \cdot \nabla v dx + \int_{\Omega} c_0 y v dx + \int_{\Gamma} \alpha y v ds = \int_{\Omega} f v dx + \int_{\Gamma} g v ds$$

for all $v \in C^1(\bar{\Omega})$. Using the fact that $C^1(\bar{\Omega})$ is for Lipschitz domains Ω a dense subset of $H^1(\Omega)$, and assuming that $y \in H^1(\Omega)$, we finally arrive at the following definition:

Definition. A function $y \in H^1(\Omega)$ is called a weak solution to the boundary value problem (2.11) if the variational equality (2.12) holds for all $v \in H^1(\Omega)$.

The boundary condition in (2.11) does not need to be accounted for in the solution space. As a so-called *natural boundary condition*, it follows automatically for sufficiently smooth solutions. In order to apply the Lax–Milgram lemma to the present situation, we put $V := H^1(\Omega)$ and define the functional F and the bilinear form a , respectively, by

$$(2.13) \quad \begin{aligned} F(v) &:= \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds, \\ a[y, v] &:= \int_{\Omega} \nabla y \cdot \nabla v \, dx + \int_{\Omega} c_0 y v \, dx + \int_{\Gamma} \alpha y v \, ds. \end{aligned}$$

Observe that in this case F can no longer be identified with a function $f \in L^2(\Omega)$; F has a more complicated structure and can only be interpreted as an element of V^* . The variational formulation (2.12) is again of the form (2.5). To prove the V -ellipticity of a this time, we need the following inequality.

Lemma 2.5. *Let $\Omega \subset \mathbb{R}^N$ denote a bounded Lipschitz domain, and let $\Gamma_1 \subset \Gamma$ be a measurable set such that $|\Gamma_1| > 0$. Then there exists a constant $c(\Gamma_1) > 0$, which is independent of $y \in H^1(\Omega)$, such that*

$$(2.14) \quad \|y\|_{H^1(\Omega)}^2 \leq c(\Gamma_1) \left(\int_{\Omega} |\nabla y|^2 \, dx + \left(\int_{\Gamma_1} y \, ds \right)^2 \right)$$

for all $y \in H^1(\Omega)$.

The proof of this generalization of the Friedrichs inequality can be found, e.g., in [Cas92] or [Wlo87]. The Friedrichs inequality obviously arises as a special case with $\Gamma_1 := \Gamma$ and functions $y \in H_0^1(\Omega)$. An analogous inequality holds for subsets of Ω : for any set $E \subset \Omega$ having positive measure there exists some constant $c(E) > 0$, which is independent of $y \in H^1(\Omega)$, such that the *generalized Poincaré inequality*

$$(2.15) \quad \|y\|_{H^1(\Omega)}^2 \leq c(E) \left(\int_{\Omega} |\nabla y|^2 \, dx + \left(\int_E y \, dx \right)^2 \right)$$

holds for all $y \in H^1(\Omega)$; see [Cas92] or [GGZ74]. In the case where $E := \Omega$, Poincaré's inequality results.

We are now in a position to show the existence of a weak solution.

Theorem 2.6. *Let $\Omega \subset \mathbb{R}^N$ be a Lipschitz domain, and suppose that almost-everywhere nonnegative functions $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^\infty(\Gamma)$ are given such that*

$$\int_{\Omega} (c_0(x))^2 \, dx + \int_{\Gamma} (\alpha(x))^2 \, ds(x) > 0.$$

Then for every given pair $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$, the boundary value problem (2.11) has a unique weak solution $y \in H^1(\Omega)$. Moreover, there is some constant $c_R > 0$, independent of f and g , such that

$$(2.16) \quad \|y\|_{H^1(\Omega)} \leq c_R (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}).$$

Proof: We apply the Lax–Milgram lemma in $V = H^1(\Omega)$. To this end, we have to verify that the bilinear form (2.13) is bounded and V -elliptic. In this proof, as throughout this textbook, $c > 0$ denotes a generic constant that depends only on the data of the problem. First, one easily derives

$$|a[y, v]| = \left| \int_{\Omega} \nabla y \cdot \nabla v \, dx + \int_{\Omega} c_0 y v \, dx + \int_{\Gamma} \alpha y v \, ds \right| \leq \alpha_0 \|y\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)},$$

i.e., the boundedness of a . Indeed, this is an immediate consequence of the estimates

$$\begin{aligned} \left| \int_{\Omega} c_0 y v \, dx \right| &\leq \|c_0\|_{L^\infty(\Omega)} \|y\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \|c_0\|_{L^\infty(\Omega)} \|y\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}, \\ \left| \int_{\Gamma} \alpha y v \, ds \right| &\leq \|\alpha\|_{L^\infty(\Gamma)} \|y\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)} \\ &\leq \|\alpha\|_{L^\infty(\Gamma)} c \|y\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}, \end{aligned}$$

where the trace theorem has been used for the latter estimate.

To show the V -ellipticity, we argue as follows. In view of the assumptions, we have $c_0 \neq 0$ in $L^\infty(\Omega)$ or $\alpha \neq 0$ in $L^\infty(\Gamma)$. If $c_0 \neq 0$, then there exist a measurable set $E \subset \Omega$ with $|E| > 0$ and some $\delta > 0$ such that $c_0(x) \geq \delta$ for all $x \in E$. Hence, invoking (2.15) and the inequality $(\int_E y \, dx)^2 \leq |E| \int_E y^2 \, dx$, we find that

$$\begin{aligned} a[y, y] &= \int_{\Omega} (|\nabla y|^2 + c_0 |y|^2) \, dx + \int_{\Gamma} \alpha |y|^2 \, ds \geq \int_{\Omega} |\nabla y|^2 \, dx + \delta \int_E |y|^2 \, dx \\ &\geq \min \{1, \delta\} \left(\int_{\Omega} |\nabla y|^2 \, dx + \int_E |y|^2 \, dx \right) \\ &\geq \frac{\min \{1, \delta\}}{c(E) \max \{1, |E|\}} \|y\|_{H^1(\Omega)}^2. \end{aligned}$$

In the case where $\alpha \neq 0$ there exist a measurable set $\Gamma_1 \subset \Gamma$ with $|\Gamma_1| > 0$ and some $\delta > 0$ such that $\alpha(x) \geq \delta$ for all $x \in \Gamma_1$. In view of (2.14), similar reasoning yields that in this case,

$$(2.17) \quad a[y, y] \geq \int_{\Omega} |\nabla y|^2 \, dx + \delta \int_{\Gamma_1} |y|^2 \, ds \geq \frac{\min \{1, \delta\}}{c(\Gamma_1) \max \{1, |\Gamma_1|\}} \|y\|_{H^1(\Omega)}^2.$$

Consequently, the assumptions of Lemma 2.2 are satisfied. In addition, employing the trace theorem once more, we can conclude as follows:

$$\begin{aligned}
 |F(v)| &\leq \int_{\Omega} |f v| \, dx + \int_{\Gamma} |g v| \, ds \\
 &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)} \\
 &\leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} + c \|g\|_{L^2(\Gamma)} \|v\|_{H^1(\Omega)} \\
 &\leq \tilde{c} (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}) \|v\|_{H^1(\Omega)}.
 \end{aligned}$$

But this means that $\|F\|_{V^*} \leq \tilde{c} (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)})$, and the asserted estimate for $\|y\|_{H^1(\Omega)}$ then follows from the Lax–Milgram lemma. This concludes the proof. \square

2.3.3. Differential operators in divergence form. The boundary value problems investigated in Sections 2.3.1 and 2.3.2 are special cases of the problem

$$(2.18) \quad \boxed{
 \begin{array}{ll}
 \mathcal{A}y + c_0 y = f & \text{in } \Omega \\
 \partial_{\nu_{\mathcal{A}}} y + \alpha y = g & \text{on } \Gamma_1 \\
 y = 0 & \text{on } \Gamma_0.
 \end{array}
 }$$

Here, \mathcal{A} is an elliptic differential operator of the form

$$(2.19) \quad \mathcal{A}y(x) = - \sum_{i,j=1}^N D_i (a_{ij}(x) D_j y(x)), \quad x \in \Omega.$$

The coefficient functions a_{ij} of \mathcal{A} are assumed to belong to $L^\infty(\Omega)$ and to satisfy the symmetry condition $a_{ij}(x) = a_{ji}(x)$ for all $i, j \in \{1, \dots, N\}$ and $x \in \Omega$. Moreover, they are assumed to satisfy with some $\gamma_0 > 0$ the *condition of uniform ellipticity*, that is,

$$(2.20) \quad \sum_{i,j=1}^N a_{ij}(x) \xi_i \xi_j \geq \gamma_0 |\xi|^2 \quad \forall \xi \in \mathbb{R}^N$$

for almost all $x \in \Omega$. In this more general case we denote by $\partial_{\nu_{\mathcal{A}}}$ the directional derivative in the direction of the *conormal* vector $\nu_{\mathcal{A}}$ whose components are given by

$$(2.21) \quad (\nu_{\mathcal{A}})_i(x) = \sum_{j=1}^N a_{ij}(x) \nu_j(x), \quad 1 \leq i \leq N.$$

Observe that with the $N \times N$ matrix function $A = (a_{ij})$ we have $\nu_{\mathcal{A}} = A \nu$.

The boundary $\Gamma = \Gamma_0 \cup \Gamma_1$ is split into two disjoint measurable subsets Γ_0 and Γ_1 , one of which may be empty. Moreover, almost-everywhere non-negative functions $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^2(\Gamma_1)$ are given, as well as functions $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_1)$.

The appropriate solution space for problem (2.18) is

$$V := \{y \in H^1(\Omega) : y|_{\Gamma_0} = 0\}.$$

We thus have $\tau y = 0$ almost everywhere in Γ_0 . The associated bilinear form a is given by

$$(2.22) \quad a[y, v] := \int_{\Omega} \sum_{i,j=1}^N a_{ij} D_i y D_j v \, dx + \int_{\Omega} c_0 y v \, dx + \int_{\Gamma_1} \alpha y v \, ds,$$

and the weak solution $y \in V$ is defined as the solution to the variational equality

$$a[y, v] = (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\Gamma_1)} \quad \forall v \in V.$$

We have the following well-posedness result.

Theorem 2.7. *Suppose that $\Omega \subset \mathbb{R}^N$ is a bounded Lipschitz domain, and suppose that the assumptions from above are satisfied. Moreover, assume that $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^\infty(\Gamma_1)$ satisfy $c_0(x) \geq 0$ and $\alpha(x) \geq 0$ almost everywhere in Ω and in Γ_1 , respectively. If one of the conditions*

- (i) $|\Gamma_0| > 0$
- (ii) $\Gamma_1 = \Gamma$ and $\int_{\Omega} (c_0(x))^2 \, dx + \int_{\Gamma} (\alpha(x))^2 \, ds(x) > 0$

is satisfied, then for all pairs $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_1)$ problem (2.18) has a unique weak solution $y \in V$. Moreover, there is a constant $c_A > 0$, which depends on neither f nor g , such that

$$(2.23) \quad \|y\|_{H^1(\Omega)} \leq c_A (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_1)}) \quad \forall f \in L^2(\Omega), \forall g \in L^2(\Gamma_1).$$

The proof proceeds along the same lines as that of Theorem 2.6, using Lemma 2.2; see Exercise 2.4. Compare this also with the treatment of equations of the form (2.18) in [Cas92], [Lio71], or [Wlo87].

Remarks.

(i) Assumption (ii) above is equivalent to saying that at least one of the following conditions is satisfied: there is a set $E \subset \Omega$ with $|E| > 0$ such that $c_0(x) > 0$ for almost all $x \in E$; or, there is a set $D \subset \Gamma$ with $|D| > 0$ such that $\alpha(x) > 0$ for almost all $x \in D$.

(ii) In all three cases studied above, the Dirichlet boundary conditions that occurred were merely homogeneous. There are good reasons for this. First, a nonhomogeneous boundary condition of the form $y|_{\Gamma} = g$ automatically entails that g has the regularity $g \in H^{1/2}(\Gamma)$, provided that $y \in H^1(\Omega)$ (fractional-order Sobolev spaces will be defined in Section 2.14.2). If, as in later sections, g were a control, then it would have to be chosen from $H^{1/2}(\Gamma)$. This does not make sense in many practical applications.

Moreover, the standard variational formulation does not work in the case of inhomogeneous Dirichlet boundary conditions. A possible way out is a reduction to homogeneous boundary conditions by using a function that satisfies the inhomogeneous Dirichlet conditions. In Lions [Lio71], inhomogeneous Dirichlet problems for elliptic and parabolic equations were treated using the so-called *transposition method*. For the parabolic case, we also refer to Bensoussan et al. [BDPDM92, BDPDM93], where semigroups and the variation of constants formula were employed. Recent results on boundary control involving boundary conditions of Dirichlet type can be found in, e.g., [CR06], [KV07], and [Vex07].

(iii) The estimates (2.9), (2.16), and (2.23), of the type $\|y\| \leq c(\|f\| + \|g\|)$, are equivalent to saying that the mappings $f \mapsto y$ and $(f, g) \mapsto y$ are continuous between the respective spaces.

Data belonging to L^p spaces with $p < 2$. We reconsider problem (2.18) from page 37 in the form

$$\begin{aligned} \mathcal{A}y + c_0 y &= f && \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y + \alpha y &= g && \text{on } \Gamma, \end{aligned}$$

where the assumptions of Theorem 2.7 condition (ii) are assumed to hold. Till now, it has been assumed that $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$. We are now going to demonstrate that a unique solution $y \in H^1(\Omega)$ exists also if $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$, for suitably chosen $1 < r, s < 2$. For this purpose, we interpret f and g as functionals on $H^1(\Omega)^*$ and define

$$F_1(v) = \int_{\Omega} f(x)v(x) dx, \quad F_2(v) = \int_{\Gamma} g(x)v(x) ds(x).$$

From Sobolev's embedding result, Theorem 7.1 on page 355, we infer that the embedding $H^1(\Omega) \hookrightarrow L^p(\Omega)$ is continuous for all $p < \infty$ if $N = \dim \Omega = 2$, and continuous for all $p \leq 2N/(N-2)$ if $N = \dim \Omega > 2$. Owing to Hölder's inequality, we have

$$|F_1(v)| \leq \|f\|_{L^r(\Omega)} \|v\|_{L^p(\Omega)},$$

where $1/r + 1/p = 1$. In the case of $N = 2$, an arbitrarily large p may be chosen, that is, r may be arbitrarily close to unity. Hence, for $N = 2$ we have $F_1 \in H^1(\Omega)^*$ if $f \in L^r(\Omega)$ merely for some $r > 1$. In the $N > 2$ case, the smallest possible r is given by

$$\frac{1}{r} + \frac{N-2}{2N} = 1 \quad \Rightarrow \quad r = \frac{2N}{N+2}.$$

Thus, $F_1 \in H^1(\Omega)^*$ for $N > 2$ if $f \in L^r(\Omega)$ for some $r \geq 2N/(N+2)$.

In a similar way, we can study F_2 , invoking Theorem 7.2 on page 355. The results can be summarized as follows: if $N = 2$, then the trace τy belongs to $L^p(\Gamma)$ for all $p < \infty$, while in the case $N > 2$ we have $\tau y \in L^p(\Gamma)$ only if $p \leq 2(N-1)/(N-2)$. In summary, $F_2 \in H^1(\Omega)^*$ provided that $g \in L^s(\Gamma)$, where $s > 1$ if $N = 2$ and $s \geq 2 - 2/N$ if $N > 2$.

In any of these cases, the Lax–Milgram theorem ensures the existence of a uniquely determined solution $y \in H^1(\Omega)$ to the above problem. Moreover, we have, with a suitable generic constant $c > 0$, the estimate

$$\|y\|_{H^1(\Omega)} \leq c (\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\Gamma)}).$$

2.4. Linear mappings

2.4.1. Continuous linear operators and functionals. The results of this section are listed without proof. They can be found in most standard textbooks on functional analysis, e.g., Alt [Alt99], Kantorovich and Akilov [KA64], Kreyszig [Kre78], Lusternik and Sobolev [LS74], Wouk [Wou79], and Yosida [Yos80].

In the following, $\{U, \|\cdot\|_U\}$ and $\{V, \|\cdot\|_V\}$ denote normed spaces over \mathbb{R} .

Definition. We say that a mapping $A : U \rightarrow V$ is linear or a linear operator if $A(u+v) = Au + Av$ and $A(\lambda v) = \lambda Av$ for all $u, v \in U$ and $\lambda \in \mathbb{R}$. A linear mapping $f : U \rightarrow \mathbb{R}$ is called a linear functional.

More generally, real- or complex-valued mappings are referred to as *functionals*.

Definition. We call a mapping $A : U \rightarrow V$ continuous on U if for any sequence $\{u_n\}_{n=1}^\infty \subset U$ with $\lim_{n \rightarrow \infty} \|u_n - u\|_U = 0$ we have $\lim_{n \rightarrow \infty} \|Au_n - Au\|_V = 0$.

Definition. A linear operator $A : U \rightarrow V$ is said to be bounded if there is a constant $c(A) > 0$ such that

$$\|Au\|_V \leq c(A) \|u\|_U \quad \forall u \in U.$$

Theorem 2.8. A linear operator is bounded if and only if it is continuous.

Example. We take $U = V = C[0, 1]$ and consider the integral operator A defined by

$$(Au)(t) = \int_0^1 e^{t-s} u(s) ds, \quad t \in [0, 1].$$

Obviously, A is a linear mapping from U into itself. To prove that A is continuous, we show its boundedness and employ the above theorem. We have

$$\begin{aligned} |(Au)(t)| &\leq e^t \int_0^1 e^{-s} |u(s)| ds \leq e^t (1 - e^{-1}) \max_{t \in [0,1]} |u(t)| \\ &\leq (e - 1) \|u\|_{C[0,1]} \end{aligned}$$

and, therefore,

$$\|Au\|_U = \max_{t \in [0,1]} |(Au)(t)| \leq (e - 1) \|u\|_U.$$

Consequently, A is bounded with $c(A) = e - 1$. ◇

Definition. If $A : U \rightarrow V$ is a linear and continuous operator, then

$$\|A\|_{\mathcal{L}(U,V)} = \sup_{\|u\|_U=1} \|Au\|_V < +\infty.$$

The finite number $\|A\|_{\mathcal{L}(U,V)}$ is called the (operator) norm of A . The shorter notation $\|A\|$ is also commonly used.

Since for linear operators continuity is equivalent to boundedness, there is some $c > 0$ such that $\|Au\|_V \leq c \|u\|_U$ for all $u \in U$. Obviously, $c = \|A\|$ is the smallest such constant. Also, the term *norm* is justified, because $\|A\|$ is in fact a norm on the linear space of all linear and continuous mappings from U into V ; the reader will be asked to verify this in Exercise 2.5.

Definition. $\mathcal{L}(U, V)$ denotes the normed space of all linear and continuous mappings from U into V , endowed with the operator norm $\|\cdot\|_{\mathcal{L}(U,V)}$. If $U = V$, then we write $\mathcal{L}(U, V) =: \mathcal{L}(U)$.

The space $\mathcal{L}(U, V)$ is complete (and thus a Banach space) if V is complete.

Example: multiplication operator. Let $U = V = L^\infty(\Omega)$, and let a fixed function $a \in L^\infty(\Omega)$ be given. We consider the operator $A : U \rightarrow V$ given by

$$(Au)(x) = a(x)u(x) \quad \text{for almost every } x \in \Omega.$$

A is bounded, since

$$\|Au\|_V = \|a(\cdot)u(\cdot)\|_{L^\infty(\Omega)} \leq \|a\|_{L^\infty(\Omega)} \|u\|_{L^\infty(\Omega)},$$

where obviously the latter estimate cannot be improved. In conclusion, $A \in \mathcal{L}(L^\infty(\Omega))$, and $\|A\|_{\mathcal{L}(L^\infty(\Omega))} = \|a\|_{L^\infty(\Omega)}$.

As an illustration, consider the operator $A : L^\infty(0, 1) \rightarrow L^\infty(0, 1)$,

$$(Au)(x) = x^2 u(x).$$

We have $\|A\| = 1$, since the function $a(x) = x^2$ belongs to the unit sphere in $L^\infty(0, 1)$. \diamond

Definition. *The space of all continuous linear functionals on $\{U, \|\cdot\|_U\}$, denoted by U^* , is called the dual space of U .*

Observe that $U^* = \mathcal{L}(U, \mathbb{R})$. The associated norm is given by

$$\|f\|_{U^*} = \sup_{\|u\|_U=1} |f(u)|.$$

Moreover, since \mathbb{R} is a complete space, the dual space U^* is always a Banach space.

Example. We consider the linear functional $f(u) = u(\frac{1}{2})$ on $U = C[0, 1]$. Since

$$|f(u)| = |u(1/2)| \leq \max_{t \in [0, 1]} |u(t)| = 1 \cdot \|u\|_{C[0, 1]} \quad \forall u \in C[0, 1],$$

we see that f is bounded with $\|f\|_{U^*} \leq 1$. Moreover, for $u(t) \equiv 1$ it follows that $|f(u)| = 1 = \|u\|$, and thus $\|f\|_{U^*} \geq 1$. In summary, $\|f\|_{U^*} = 1$. \diamond

In the following, we are concerned with the explicit representation of continuous linear functionals, aiming at a characterization of dual spaces. Note that there can be many different ways to represent the same continuous linear functional; for instance, the expressions

$$(2.24) \quad F(v) = \int_0^1 \ln(\exp(3v(x) - 5)) dx + 5, \quad G(v) = 3v,$$

while looking quite different, represent the same functional on \mathbb{R} . The following result, which settles the representation problem for Hilbert spaces in terms of the scalar product, is of fundamental importance.

Theorem 2.9 (Riesz representation theorem). *Let $\{H, (\cdot, \cdot)_H\}$ be a real Hilbert space. Then for any continuous linear functional $F \in H^*$ there exists a uniquely determined $f \in H$ such that $\|F\|_{H^*} = \|f\|_H$ and*

$$F(v) = (f, v)_H \quad \forall v \in H.$$

By virtue of this result, we can identify H^* with H , writing $H = H^*$. For example, in the case of the functional on $H = \mathbb{R}$ for which different representations were given in (2.24), the canonical form referred to in the theorem is that of G , with $f = 3 \in \mathbb{R}$.

Next, we introduce the fundamental notion of *reflexivity*. To this end, let U denote a real Banach space with associated dual space U^* . We fix an

arbitrary $u \in U$, let f vary over U^* , and consider the mapping $F_u : U^* \rightarrow \mathbb{R}$ induced by u ,

$$F_u : f \mapsto f(u).$$

Clearly, F_u is linear, and its continuity is a consequence of the simple estimate

$$|F_u(f)| = |f(u)| \leq \|u\|_U \|f\|_{U^*}.$$

Hence, the functional F_u induced by u belongs to the dual space $(U^*)^* =: U^{**}$ of U^* . Since the mapping $u \mapsto F_u$ turns out to be injective, we may identify u with F_u , thereby interpreting $u \in U$ as an element of U^{**} .

The space U^{**} is called the *bidual space* of U . In light of the above identification, it is always true that $U \subset U^{**}$. The mapping $u \mapsto F_u$ from U into U^{**} is called the *canonical embedding* or *canonical mapping*. If this mapping is surjective, i.e., if $U = U^{**}$, then U is called a *reflexive* space. In the case of reflexive spaces, taking the dual twice leads back to the original space. In particular, we infer from the Riesz representation theorem that Hilbert spaces are always reflexive.

Example. The spaces $L^p(E)$ introduced in Section 2.2.1 are also reflexive if $1 < p < \infty$. In fact, it can be shown that the dual space $L^p(E)^*$ can be identified with $L^q(E)$, where the *conjugate exponent* q of p is given by the relation $\frac{1}{p} + \frac{1}{q} = 1$. More precisely, to every continuous linear functional $F \in L^p(E)^*$ there corresponds a uniquely determined function $f \in L^q(E)$ such that

$$F(u) = \int_E f(x) u(x) dx \quad \forall u \in L^p(E).$$

Repeating this argument, we arrive at the conclusion that the bidual space $L^p(E)^{**}$ can be identified with $L^p(E)$, which proves the reflexivity. Observe that the continuity of the above functional F is a consequence of *Hölder's inequality for integrals*,

$$(2.25) \quad \int_E |f(x)| |u(x)| dx \leq \left(\int_E |f(x)|^q dx \right)^{\frac{1}{q}} \left(\int_E |u(x)|^p dx \right)^{\frac{1}{p}}.$$

◇

Remark. Note that the spaces $L^\infty(E)$ and $L^1(E)$ are *not* reflexive. Indeed, while $L^1(E)^*$ can be identified with $L^\infty(E)$, the dual space of $L^\infty(E)$ cannot be identified with $L^1(E)$.

2.4.2. Weak convergence. The contents of this subsection are of importance mainly for proving the existence of optimal controls; thus, they may for the time being be skipped by readers who are more interested in actually *finding* the solution to optimal control problems. In the following, the

underlying spaces will always be Banach spaces, even though not all of the results require the completeness property.

Definition. Let U be a real Banach space. We say that a sequence $\{u_n\}_{n=1}^\infty \subset U$ converges weakly to some $u \in U$ if

$$\lim_{n \rightarrow \infty} f(u_n) = f(u) \quad \forall f \in U^*.$$

We denote weak convergence by the symbol \rightharpoonup , that is, we write $u_n \rightharpoonup u$ as $n \rightarrow \infty$.

The limit u is uniquely determined and is called the *weak limit* of the sequence. Moreover, it follows from the Banach–Steinhaus theorem, which is a consequence of the *principle of uniform boundedness*, that $\{\|u_n\|\}_{n=1}^\infty \subset \mathbb{R}$ is bounded for any weakly convergent sequence $\{u_n\}_{n=1}^\infty \subset U$.

Examples.

(i) If a sequence $\{u_n\}_{n=1}^\infty \subset U$ converges *strongly* (that is, with respect to the norm of U) to $u \in U$, then it also converges weakly to u , i.e.,

$$u_n \rightarrow u \quad \Rightarrow \quad u_n \rightharpoonup u \quad \text{as } n \rightarrow \infty.$$

(ii) By virtue of the Riesz representation theorem, weak convergence in a Hilbert space $\{H, (\cdot, \cdot)\}$ is equivalent to

$$\lim_{n \rightarrow \infty} (v, u_n) \rightarrow (v, u) \quad \forall v \in H.$$

Moreover, if $u_n \rightharpoonup u$ and $v_n \rightarrow v$ (strong convergence), then $(v_n, u_n) \rightarrow (v, u)$ as $n \rightarrow \infty$; see Exercise 2.8. In other words, the scalar products of the terms of a weakly convergent sequence and a strongly convergent one tend to the scalar product of the associated limits.

(iii) We consider in the Hilbert space $H = L^2(0, 2\pi)$ the sequence of functions

$$u_n(x) = \frac{1}{\sqrt{\pi}} \sin(nx), \quad x \in (0, 2\pi).$$

Moreover, let $f \in L^2(0, 2\pi)$ be arbitrary. Then

$$(f, u_n) = \int_0^{2\pi} f(x) \frac{1}{\sqrt{\pi}} \sin(nx) dx$$

defines the n th Fourier coefficient associated with f with respect to the orthonormal system consisting of the functions $\sin(nx)/\sqrt{\pi}$, $n \in \mathbb{N}$, in $L^2(0, 2\pi)$. Owing to the well-known *Bessel inequality*, the sequence of coefficients tends to zero as $n \rightarrow \infty$, that is,

$$(f, u_n) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Now observe that $0 = (f, 0)$ for all $f \in H$. Consequently, the sequence $\{u_n\}_{n=1}^\infty$ converges weakly to the zero function:

$$u_n = \frac{1}{\sqrt{\pi}} \sin(n \cdot) \rightharpoonup 0 \quad \text{as } n \rightarrow \infty.$$

On the other hand, we have

$$\|u_n\|^2 = \frac{1}{\pi} \int_0^{2\pi} \sin^2(nx) dx = 1 \quad \forall n \in \mathbb{N}.$$

◇

Conclusion. *There exist sequences that converge weakly to the zero function even though all their terms belong to the unit sphere.*

The above sequence of sine functions, while converging weakly to the zero function, oscillates ever more strongly as n increases. This example shows that little (if any) information about the actual pointwise convergence behavior can be extracted from the mere fact that a sequence is weakly convergent. Therefore, the notion of weak convergence is not of major importance from the numerical point of view. However, in the context of proving existence results it plays a fundamental role. We are now going to provide some results that form the conceptual basis for the application of the notion of weak convergence.

Definition. *Let U and V denote real Banach spaces. A mapping $F : U \rightarrow V$ is said to be weakly sequentially continuous if the following holds: whenever a sequence $\{u_n\}_{n=1}^\infty \subset U$ converges weakly in U to some $u \in U$, its image $\{F(u_n)\}_{n=1}^\infty \subset V$ converges weakly to $F(u)$ in V ; that is,*

$$u_n \rightharpoonup u \quad \Rightarrow \quad F(u_n) \rightharpoonup F(u) \quad \text{as } n \rightarrow \infty.$$

Examples.

(i) *Every continuous linear operator $A : U \rightarrow V$ is weakly sequentially continuous.*

The proof of this statement is easy: suppose that $u_n \rightharpoonup u$. We have to show that then $Au_n \rightharpoonup Au$, i.e., that $f(Au_n) \rightarrow f(Au)$ for all $f \in V^*$. Now if $f \in V^*$ is fixed, then the functional $F(u) := f(Au)$ is obviously linear and continuous on U , and thus belongs to U^* . Hence, we must have $F(u_n) \rightarrow F(u)$ or, in view of the definition of F , $f(Au_n) \rightarrow f(Au)$. Since f was arbitrarily chosen, we can conclude that $Au_n \rightharpoonup Au$.

(ii) The functional $f(u) = \|u\|$ is not weakly sequentially continuous in the Hilbert space $H = L^2(0, 2\pi)$. The sequence of sine functions $u_n(x) =$

$\sin(nx)/\sqrt{\pi}$ from above serves as a counterexample. Indeed, we know that $u_n \rightarrow 0$ as $n \rightarrow \infty$ but

$$\lim_{n \rightarrow \infty} f(u_n) = \lim_{n \rightarrow \infty} \|u_n\| = 1 \neq \|0\| = f(0).$$

The fact that the norm in the Hilbert space $H = L^2(0, 2\pi)$ is not weakly sequentially continuous presents a problem that will have to be attended to when dealing with infinite-dimensional Banach spaces. It is one reason for the introduction of the concept of *weak lower semicontinuity*; cf. the example following Theorem 2.12. \diamond

Definition. *Let M be a subset of a real Banach space U . We say that M is weakly sequentially closed if the limit of every weakly convergent sequence $\{u_n\}_{n=1}^{\infty} \subset M$ lies in M . We say that M is weakly sequentially relatively compact if every sequence $\{u_n\}_{n=1}^{\infty} \subset M$ contains a weakly convergent subsequence; if, in addition, M is weakly sequentially closed, then M is said to be weakly sequentially compact.*

The reader will be asked to verify in Exercise 2.7 that every strongly convergent sequence also converges weakly. As the above example involving sine functions shows, the contrary is false in general; that is to say, in general there are more weakly convergent sequences than strongly convergent ones.

Conclusion. *Any weakly sequentially closed set is also (strongly) closed; however, not every (strongly) closed set must be weakly sequentially closed.*

For instance, the unit sphere in the space $H = L^2(0, 2\pi)$ is closed but not weakly sequentially closed: the sequence of sine functions $\{\sin(nx)/\sqrt{\pi}\}$ belongs to the unit sphere while its weak limit, the zero function, does not.

The next two results can be found in, e.g., [Kre78], [Wou79], and [Yos80].

Theorem 2.10. *Every bounded subset of a reflexive Banach space is weakly sequentially relatively compact.*

The above result is the main reason why the concept of weak convergence is of such fundamental importance: it says that the notion of weak sequential relative compactness can in a certain sense take over the role of relative compactness. It follows from a theorem of Eberlein and Shmulian that this property even characterizes reflexive Banach spaces; see [Yos80].

Definition.

- (i) A subset C of a real Banach space U is said to be convex if for any pair $u, v \in C$ and any $\lambda \in [0, 1]$ the convex combination $\lambda u + (1 - \lambda)v$ lies in C .
- (ii) A functional $f : C \rightarrow \mathbb{R}$ is said to be convex if

$$f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v)$$

for all $\lambda \in [0, 1]$ and all $u, v \in C$. The functional is said to be strictly convex if the above inequality holds with $<$ in place of \leq whenever $u \neq v$ and $\lambda \in (0, 1)$.

Theorem 2.11. *Every convex and closed subset of a Banach space is weakly sequentially closed. If the space is reflexive and the set is in addition bounded, then it is weakly sequentially compact.*

The first assertion of the theorem is an easy consequence of Mazur's theorem, which states that the weak limit of a weakly convergent sequence is at the same time the strong limit of a sequence consisting of suitable convex combinations of the terms of the sequence. This part of the assertion is already true in normed spaces; see [BP78] and [Wer97]. The second assertion follows from Theorem 2.10.

Theorem 2.12. *Every continuous and convex functional $f : U \rightarrow \mathbb{R}$ on a Banach space U is weakly lower semicontinuous; that is, for any sequence $\{u_n\}_{n=1}^{\infty} \subset U$ such that $u_n \rightharpoonup u$ as $n \rightarrow \infty$ we have*

$$\liminf_{n \rightarrow \infty} f(u_n) \geq f(u).$$

For a proof of this result, we refer the interested reader to [BP78], [Wer97], or [Wou79]. Note that the preceding two theorems underline the key importance of the concept of convexity for the treatment of optimization problems in function spaces.

Example. The functional $f(u) = \|u\|$ is obviously continuous on any Banach space. It is also convex, since it follows from the triangle inequality and homogeneity that

$$\|\lambda u + (1 - \lambda)v\| \leq \lambda \|u\| + (1 - \lambda)\|v\| \quad \forall \lambda \in [0, 1], \quad \forall u, v \in U.$$

Owing to the above theorem, the norm functional is thus weakly lower semicontinuous on U . \diamond

Remark. In the literature, the notions of weak compactness and weak closedness in the sense of the weak topology are often used in place of weak sequential compactness and weak sequential closedness, respectively. This may lead to confusion and sometimes renders the study of the relevant literature a bit difficult. It should be noted, however, that in reflexive Banach spaces the two concepts are equivalent; see [Alt99], Section 6.7, or [Con90].

2.5. Existence of optimal controls

In this chapter, we are concerned with optimal control problems for linear elliptic differential equations. In the course of our study, we will discuss the following fundamental questions: Does a solution to the problem (i.e. an optimal control with associated optimal state) exist? What optimality conditions must possible solutions necessarily satisfy? How can their solutions be determined numerically? We first investigate the problem of existence, beginning with the simplest of the examples presented in Section 2.3, namely the boundary value problem for Poisson's equation.

We remark generally that if for a given problem existence cannot be shown by standard techniques, this is often due to mistakes made during the process of modeling; such mistakes are also likely to lead to numerical difficulties.

In this section, we make the following general assumptions on the data that characterize the problems under study. In this connection, E denotes a set whose actual meaning varies from case to case and will become clear from the context.

Assumption 2.13. $\Omega \subset \mathbb{R}^N$ denotes a bounded Lipschitz domain with boundary Γ , and we assume that we are given $\lambda \geq 0$, $y_\Omega \in L^2(\Omega)$, $y_\Gamma \in L^2(\Gamma)$, $\beta \in L^\infty(\Omega)$, and $\alpha \in L^\infty(\Gamma)$ with $\alpha(x) \geq 0$ for almost every $x \in \Gamma$, as well as functions $u_a, u_b, v_a, v_b \in L^2(E)$ having the property that $u_a(x) \leq u_b(x)$ and $v_a(x) \leq v_b(x)$ for almost every $x \in E$.

In this connection, y_Ω and y_Γ represent desired functions (i.e., *targets* to be approximated), α and β are coefficient functions, and the functions u_a, u_b, v_a , and v_b will define the sets of admissible controls acting on $E = \Omega$ or on $E = \Gamma$.

In most cases to follow, the control function will be denoted by u . This commonly used notation goes back to the Russian word “**u**pravlenie” for control. If, however, both a distributed control and a boundary control occur in a problem, then u will denote the boundary control and v the distributed control.

2.5.1. Optimal stationary heat sources. As the first case study, we investigate the problem of finding an optimal heat source under homogeneous Dirichlet boundary conditions, which can be written in the form

$$(2.26) \quad \min J(y, u) := \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2,$$

subject to the constraints

$$(2.27) \quad \boxed{\begin{array}{l} -\Delta y = \beta u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}}$$

and

$$(2.28) \quad u_a(x) \leq u(x) \leq u_b(x) \quad \text{for almost every } x \in \Omega.$$

First, we have to decide from which class of functions the control u should be selected. Continuous functions are not eligible, since the set of all continuous functions u such that $u_a \leq u \leq u_b$ does not, as a rule, have the compactness properties needed to prove existence; for instance, this applies to the case of continuous bounds satisfying $u_a(x) < u_b(x)$ on Ω . Moreover, it will turn out that optimal controls may have jump discontinuities if $\lambda = 0$. With these considerations, a natural choice for the control space is given by the Hilbert space $L^2(\Omega)$. We thus define the *set of admissible controls* by

$$U_{ad} = \{u \in L^2(\Omega) : u_a(x) \leq u(x) \leq u_b(x) \text{ for almost every } x \in \Omega\}.$$

U_{ad} is a nonempty, closed, and convex subset of $L^2(\Omega)$; see Exercise 2.9. Its elements are called *admissible controls*.

Owing to Theorem 2.4 on page 33, to every $u \in U_{ad}$ there corresponds a unique weak solution $y \in H_0^1(\Omega)$ to the boundary value problem (2.27), called *the state associated with u* . The space

$$Y := H_0^1(\Omega)$$

is referred to as the *state space*. The dependence of y on u is expressed by the notation $y = y(u)$. The context will always ensure that this expression cannot be confused with the value $y(x)$ of y at $x \in \bar{\Omega}$.

Definition. We call a control $\bar{u} \in U_{ad}$ optimal and $\bar{y} = y(\bar{u})$ the associated optimal state if

$$J(\bar{y}, \bar{u}) \leq J(y(u), u) \quad \forall u \in U_{ad}.$$

For the treatment of the existence question, we now rewrite the optimal control problem as an optimization problem in terms of u .

Definition. The mapping $G : L^2(\Omega) \rightarrow H_0^1(\Omega)$, $u \mapsto y(u)$, defined by Theorem 2.4 on page 33 is called the control-to-state operator.

Obviously, G is a linear mapping and, by virtue of the estimate (2.9), also continuous.

In view of the obvious estimate $\|y\|_{L^2(\Omega)} \leq \|y\|_{H^1(\Omega)}$, the space $H^1(\Omega)$ and its subspace $H_0^1(\Omega)$ are linearly and continuously embedded in $L^2(\Omega)$. Therefore, G may also be viewed as a continuous linear operator with range in $L^2(\Omega)$, which we will do henceforth. In other words, we consider the operator $E_Y G$ instead of G , where $E_Y : H^1(\Omega) \rightarrow L^2(\Omega)$ denotes the embedding operator that assigns to each function $y \in Y = H^1(\Omega)$ the same function in $L^2(\Omega)$. More precisely, we have to interpret E_Y first as an operator acting between $H_0^1(\Omega)$ and $L^2(\Omega)$. However, $H_0^1(\Omega)$ is a subspace of $H^1(\Omega)$, and the norms of the two spaces are equivalent, so we avoid in this way the use of two different embedding operators. Note that E_Y is a linear and continuous operator. The operator thus defined is denoted by S , that is,

$$S = E_Y G.$$

In the following, S will always represent that part of the state y that actually occurs in the quadratic cost functional. This can be either y itself or its trace $y|_\Gamma$. In the problem of stationary heat sources, we thus have

$$S : L^2(\Omega) \rightarrow L^2(\Omega), \quad u \mapsto y(u).$$

The use of S has the advantage that the adjoint operator S^* (see Section 2.7 for the definition of this notion) also acts in the space $L^2(\Omega)$. Moreover, the optimal control problem (2.26)–(2.28) reduces to the following quadratic optimization problem in the Hilbert space $L^2(\Omega)$:

$$(2.29) \quad \boxed{\min_{u \in U_{ad}} f(u) := \frac{1}{2} \|S u - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2.}$$

The functional f just defined is referred to as the *reduced functional*. The following existence result for problem (2.29) will be applied repeatedly during the course of this textbook.

Theorem 2.14. Let $\{U, \|\cdot\|_U\}$ and $\{H, \|\cdot\|_H\}$ denote real Hilbert spaces, and let a nonempty, closed, bounded, and convex set $U_{ad} \subset U$, as well as some $y_d \in H$ and constant $\lambda \geq 0$ be given. Moreover, let $S : U \rightarrow H$ be a continuous linear operator. Then the quadratic Hilbert space optimization

problem

$$(2.30) \quad \min_{u \in U_{ad}} f(u) := \frac{1}{2} \|Su - y_d\|_H^2 + \frac{\lambda}{2} \|u\|_U^2$$

admits an optimal solution \bar{u} . If $\lambda > 0$ or S is injective, then the solution is uniquely determined.

Proof: Since $f(u) \geq 0$, there exists the infimum

$$j := \inf_{u \in U_{ad}} f(u),$$

and there is a sequence $\{u_n\}_{n=1}^\infty \subset U_{ad}$ such that $f(u_n) \rightarrow j$ as $n \rightarrow \infty$. U_{ad} is bounded and closed but—in contrast to the existence result of Theorem 1.1 for the finite-dimensional case—not necessarily compact. However, as a Hilbert space, H is reflexive; hence, by virtue of Theorem 2.11, its bounded, closed, and convex subset U_{ad} is weakly sequentially compact. Consequently, some subsequence $\{u_{n_k}\}_{k=1}^\infty$ converges weakly to some $\bar{u} \in U_{ad}$, that is,

$$u_{n_k} \rightharpoonup \bar{u} \quad \text{as } k \rightarrow \infty.$$

Since S is continuous, f is also continuous. At this point it would be a mistake to conclude that this implies $f(u_{n_k}) \rightarrow f(\bar{u})$. Instead, we have to invoke the convexity of f , which together with the continuity ensures that f is weakly lower semicontinuous. Consequently,

$$f(\bar{u}) \leq \liminf_{k \rightarrow \infty} f(u_{n_k}) = j.$$

Since $\bar{u} \in U_{ad}$, we must have $f(\bar{u}) = j$, and \bar{u} is therefore an optimal control.

The asserted uniqueness follows from the *strict convexity* of f . If $\lambda > 0$, this follows immediately from the second summand of f , while in the case of $\lambda = 0$ the strict convexity is a consequence of the injectivity of S ; see Exercise 2.10. \square

Remark. The proof only made use of the fact that f is continuous and convex. The existence result thus holds for *any* functional $f : U \rightarrow \mathbb{R}$ having these properties in a Hilbert space U . By virtue of Theorem 2.11, the whole assertion remains true also for reflexive Banach spaces U .

As a consequence of the above theorem, we obtain an existence and uniqueness result for the elliptic optimal control problem (2.26)–(2.28):

Theorem 2.15. *Suppose that the conditions of Assumption 2.13 are fulfilled. Then the problem (2.26)–(2.28) has at least one optimal control \bar{u} . If, in addition, $\lambda > 0$ or $\beta \neq 0$ almost everywhere in Ω , then the solution is unique.*

Proof: We apply the previous theorem with $U = H = L^2(\Omega)$, $y_d = y_\Omega$, and $S = E_Y G$. The set $U_{ad} = \{u \in L^2(\Omega) : u_a \leq u \leq u_b \text{ a.e. in } \Omega\}$ is bounded, closed, and convex. Hence, it follows from Theorem 2.14 that the corresponding problem (2.30) admits at least one solution \bar{u} , which is unique if $\lambda > 0$. In the $\lambda = 0$ case, we have $\beta \neq 0$ almost everywhere in Ω , which implies that the operator S is injective. Indeed, if $Su = 0$, then $y = 0$, and inserting this into the differential equation yields $\beta u = 0$ and thus $u = 0$ almost everywhere in Ω . In conclusion, S is injective, that is, we have uniqueness also for this case. This concludes the proof of the assertion. \square

Remark. In the proof of Theorem 2.14, \bar{u} is obtained as the limit of a weakly convergent sequence $\{u_{n_k}\}$. Since the control-to-state operator $G : L^2(\Omega) \rightarrow H_0^1(\Omega)$ is a continuous linear operator, it is also weakly continuous. This implies that the sequence of states $\{y_{n_k}\}$ converges weakly in $H_0^1(\Omega)$ to $\bar{y} = G\bar{u}$.

We now allow for one or both of the inequality constraints defining U_{ad} to be absent. Formally, this can be expressed by putting $u_a = -\infty$ and/or $u_b = +\infty$. Then U_{ad} is no longer bounded and hence not weakly sequentially compact. However, we still have existence and uniqueness if $\lambda > 0$, as the following result shows.

Theorem 2.16. *Suppose that U_{ad} is nonempty, closed, and convex. If $\lambda > 0$, then problem (2.30) has a unique optimal solution.*

Proof: By assumption, there exists some $u_0 \in U_{ad}$. Now observe that if $\|u\|_U^2 > 2\lambda^{-1}f(u_0)$, then

$$f(u) = \frac{1}{2} \|Su - y_d\|_H^2 + \frac{\lambda}{2} \|u\|_U^2 \geq \frac{\lambda}{2} \|u\|_U^2 > f(u_0).$$

Therefore, the search for an optimum can be restricted to the closed, convex, and bounded set $U_{ad} \cap \{u \in U : \|u\|_U^2 \leq 2\lambda^{-1}f(u_0)\}$. The remainder of the proof now proceeds along the same lines as that of the preceding theorem. \square

As an immediate consequence, we obtain the following result.

Theorem 2.17. *Suppose that $u_a = -\infty$ and/or $u_b = +\infty$. If $\lambda > 0$, then under the given conditions the problem (2.26)–(2.28) of finding the optimal stationary heat source has a uniquely determined optimal control.*

Optimal stationary heat source with prescribed outside temperature. We now recall another variant of the problem of finding an optimal

stationary heat source, in which a Robin boundary condition was given instead of a homogeneous boundary condition of Dirichlet type. The state equation is in this case given by

$$\boxed{\begin{aligned} -\Delta y &= \beta u && \text{in } \Omega \\ \partial_\nu y &= \alpha (y_a - y) && \text{on } \Gamma, \end{aligned}}$$

where the outside temperature $y_a \in L^2(\Gamma)$ and an almost-everywhere non-negative function $\alpha \in L^\infty(\Gamma)$ with $\int_\Gamma (\alpha(x))^2 ds > 0$ are prescribed.

This problem can be treated similarly to the case with homogeneous Dirichlet boundary condition, where this time the state space is given by $Y = H^1(\Omega)$. Owing to Theorem 2.6, for each pair of functions $u \in L^2(\Omega)$ and $y_a \in L^2(\Gamma)$ there is a unique weak solution $y \in Y$ to the above boundary value problem. By the superposition principle, we may decompose y in the form

$$y = y(u) + y_0,$$

where $y(u) \in H^1(\Omega)$ is the solution to the boundary value problem for Poisson's equation with homogeneous boundary condition corresponding to the pair $(\beta u, y_a = 0)$, while $y_0 \in H^1(\Omega)$ solves the boundary value problem for Laplace's equation with inhomogeneous boundary condition associated with the pair $(\beta u = 0, y_a)$. Clearly, $G : u \mapsto y(u)$ is linear and maps $L^2(\Omega)$ continuously into $H^1(\Omega)$. Again, we interpret G as an operator with range in $L^2(\Omega)$, that is, $S : L^2(\Omega) \rightarrow L^2(\Omega)$, $S = E_Y G$, so that

$$y = S u + y_0.$$

The problem then attains the form

$$(2.31) \quad \min_{u \in U_{ad}} f(u) := \frac{1}{2} \|S u - (y_\Omega - y_0)\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2.$$

Invoking Theorem 2.14 and Theorem 2.16, we immediately find that the existence results established in Theorem 2.15 and in Theorem 2.17, respectively, remain valid under the above hypotheses. In particular, there exists a unique optimal control if $\lambda > 0$. If $\lambda = 0$, existence still follows if the threshold functions are bounded; we have uniqueness in this case if $\beta \neq 0$ almost everywhere in Ω .

2.5.2. Optimal stationary boundary temperature. In a similar way, we can treat the problem of finding the optimal stationary boundary temperature. It has the form

$$(2.32) \quad \min J(y, u) := \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Gamma)}^2,$$

subject to the constraints

$$(2.33) \quad \boxed{\begin{array}{ll} -\Delta y = 0 & \text{in } \Omega \\ \partial_\nu y = \alpha(u - y) & \text{on } \Gamma \end{array}}$$

and

$$(2.34) \quad u_a(x) \leq u(x) \leq u_b(x) \quad \text{for almost every } x \in \Gamma.$$

To guarantee existence and uniqueness of a solution to the above elliptic boundary value problem, we additionally require that

$$(2.35) \quad \int_{\Gamma} (\alpha(x))^2 ds(x) > 0.$$

We seek the control u in $L^2(\Gamma)$ and the corresponding state y in the state space $Y = H^1(\Omega)$. The set of admissible controls is

$$U_{ad} = \{u \in L^2(\Gamma) : u_a(x) \leq u(x) \leq u_b(x) \quad \text{for almost every } x \in \Gamma\}.$$

By virtue of Theorem 2.6, for any $u \in L^2(\Gamma)$ the elliptic boundary value problem (2.33) has a unique weak solution $y = y(u) \in H^1(\Omega)$. The operator $G : L^2(\Gamma) \rightarrow H^1(\Omega)$, $u \mapsto y(u)$, is continuous. We interpret G as a continuous linear operator mapping $L^2(\Gamma)$ into $L^2(\Omega)$, that is, we take $S = E_Y G$ and $S : L^2(\Gamma) \rightarrow L^2(\Omega)$. We have the following result.

Theorem 2.18. *Suppose that the conditions of Assumption 2.13 on page 48 and (2.35) are satisfied. Then problem (2.32)–(2.34) has an optimal control, which is unique if $\lambda > 0$.*

This result is also a consequence of Theorem 2.14. By virtue of Theorem 2.16, it carries over to the case of unbounded admissible sets U_{ad} .

2.5.3. General elliptic equations and cost functionals * In this section, we study the general problem

$$(2.36) \quad \min J(y, u, v) := \frac{\lambda_\Omega}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda_\Gamma}{2} \|y - y_\Gamma\|_{L^2(\Gamma)}^2 \\ + \frac{\lambda_v}{2} \|v\|_{L^2(\Omega)}^2 + \frac{\lambda_u}{2} \|u\|_{L^2(\Gamma_1)}^2,$$

subject to the constraints

$$(2.37) \quad \boxed{\begin{array}{ll} \mathcal{A}y + c_0 y = \beta_\Omega v & \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y + \alpha y = \beta_\Gamma u & \text{on } \Gamma_1 \\ y = 0 & \text{on } \Gamma_0 \end{array}}$$

and

$$(2.38) \quad \begin{aligned} v_a(x) &\leq v(x) \leq v_b(x) && \text{for a.e. } x \in \Omega \\ u_a(x) &\leq u(x) \leq u_b(x) && \text{for a.e. } x \in \Gamma_1. \end{aligned}$$

Here, the uniformly elliptic differential operator \mathcal{A} and the sets Γ_0 and Γ_1 are defined as in Section 2.3.3 on page 37.

Assumption 2.19. *Suppose that Assumption 2.13 on page 48 holds. In addition, let $c_0 \in L^\infty(\Omega)$, $\beta_\Omega \in L^\infty(\Omega)$, $\beta_\Gamma \in L^\infty(\Gamma_1)$ as well as constants $\lambda_\Omega \geq 0$, $\lambda_\Gamma \geq 0$, $\lambda_v \geq 0$, and $\lambda_u \geq 0$ be prescribed. Moreover, suppose that the functions c_0 and α satisfy one of the assumptions (i) or (ii) from Theorem 2.7 on page 38.*

We begin our analysis by recalling that the appropriate state space in this case is

$$V = \{y \in H^1(\Omega) : y|_{\Gamma_0} = 0\}.$$

Under the above assumptions, the control-to-state mapping $G : (u, v) \mapsto y$ is linear and maps $L^2(\Gamma_1) \times L^2(\Omega)$ continuously into V . Again, we use $S = E_Y G$, $S : L^2(\Gamma_1) \times L^2(\Omega) \rightarrow L^2(\Omega)$. The *boundary observation operator* $S_\Gamma = \tau \circ G$, $(u, v) \mapsto y|_\Gamma$, maps $L^2(\Gamma_1) \times L^2(\Omega)$ continuously into $L^2(\Gamma)$. The sets of admissible controls are given by

$$\begin{aligned} V_{ad} &= \{v \in L^2(\Omega) : v_a(x) \leq v(x) \leq v_b(x) \quad \text{for a.e. } x \in \Omega\}, \\ U_{ad} &= \{u \in L^2(\Gamma_1) : u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Gamma_1\}. \end{aligned}$$

Finally, after elimination of y the cost functional J attains the reduced form

$$\begin{aligned} J(y, u, v) = f(u, v) &= \frac{\lambda_\Omega}{2} \|S(u, v) - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda_\Gamma}{2} \|S_\Gamma(u, v) - y_\Gamma\|_{L^2(\Gamma)}^2 \\ &\quad + \frac{\lambda_v}{2} \|v\|_{L^2(\Omega)}^2 + \frac{\lambda_u}{2} \|u\|_{L^2(\Gamma_1)}^2. \end{aligned}$$

In this example, both a *distributed control* v and a *boundary control* u occur. In addition, the cost functional contains terms of y that act in the domain as well as terms that act on the boundary (*distributed observation* and *boundary observation*). Also, this functional is convex and continuous with respect to (v, u) , so that Theorem 2.14 on page 50 applies. Observe that the second summand of the cost functional (2.36) has an impact only on Γ_1 , since y is prescribed on Γ_0 .

By virtue of Theorem 2.14 on page 50, there exists an optimal pair $(\bar{u}, \bar{v}) \in U_{ad} \times V_{ad}$, which is unique if $\lambda_u > 0$ and $\lambda_v > 0$. We note that for unbounded U_{ad} , existence follows as in Theorem 2.16 if $\lambda_u > 0$ and $\lambda_v > 0$.

2.6. Differentiability in Banach spaces

Gâteaux derivatives. For the derivation of necessary optimality conditions in the later sections of this book, we will need a generalization of the notion of derivatives. We begin here with first-order derivatives; higher-order derivatives will be encountered later in this book. We caution the reader not to confuse the meaning that the Banach spaces $\{U, \|\cdot\|_U\}$ and $\{V, \|\cdot\|_V\}$ have in this section with their later meaning in optimal control problems. In the following, \mathcal{U} will always denote a nonempty and open subset of U , while F will always denote a mapping from \mathcal{U} into V .

Definition. Let $u \in \mathcal{U}$ and $h \in U$ be given. If the limit

$$\delta F(u, h) := \lim_{t \downarrow 0} \frac{1}{t} (F(u + th) - F(u))$$

exists in V , then it is called the directional derivative of F at u in the direction h . If this limit exists for all $h \in U$, then the mapping $h \mapsto \delta F(u, h)$ is termed the first variation of F at u .

Observe that the openness of \mathcal{U} implies that $u + th$ belongs to \mathcal{U} , and therefore to the domain of F , provided that $t > 0$ is sufficiently small. Hence, the above definition is meaningful.

The first variation is not necessarily a linear mapping, as is demonstrated by the following example from [IT79]: the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, which in terms of polar coordinates is given by $f(x) = r \cos(\varphi)$, has a first variation at the origin that is nonlinear with respect to h , namely $\delta f(0, h) = f(h)$.

Definition. Suppose that the first variation $\delta F(u, h)$ at $u \in \mathcal{U}$ exists, and suppose there exists a continuous linear operator $A : U \rightarrow V$ such that

$$\delta F(u, h) = Ah \quad \forall h \in U.$$

Then F is said to be Gâteaux differentiable at u , and A is referred to as the Gâteaux derivative of F at u . We write $A = F'(u)$.

It follows from the definition that Gâteaux derivatives can be determined as directional derivatives (which we will do below). Note also that in the case where $V = \mathbb{R}$, that is, if a functional $f : \mathcal{U} \rightarrow \mathbb{R}$ is Gâteaux differentiable at a point $u \in \mathcal{U}$, then $f'(u)$ is an element of the dual space U^* .

Sometimes the Gâteaux derivative is not denoted by $F'(u)$ but rather, for example, by $F'_G(u)$. This is done in order to avoid confusion with the Fréchet derivative $F'(u)$ to be introduced below. Note that if the Fréchet derivative exists, then so does the Gâteaux derivative, and we have $F'(u) = F'_G(u)$. The converse is false, in general. However, since in all examples and exercises to be considered in this book the Gâteaux derivatives that occur will also be Fréchet derivatives, we simply use for the sake of convenience the common notation $F'(u)$.

Examples.

(i) *Evaluation of a function at a point.*

Let $U = \mathcal{U} = C[0, 1]$. We define $f : \mathcal{U} \rightarrow \mathbb{R}$ by

$$f(u(\cdot)) = \sin(u(1)).$$

Suppose that $h = h(x)$ is another element of $C[0, 1]$. We calculate the directional derivative of f at $u(\cdot)$ in the direction $h(\cdot)$. We have

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{1}{t} (f(u + th) - f(u)) &= \lim_{t \rightarrow 0} \frac{1}{t} (\sin(u(1) + th(1)) - \sin(u(1))) \\ &= \left. \frac{d}{dt} \sin(u(1) + th(1)) \right|_{t=0} \\ &= \cos(u(1) + th(1)) h(1) \Big|_{t=0} = \cos(u(1)) h(1). \end{aligned}$$

Hence, $\delta f(u, h) = \cos(u(1)) h(1)$. The mapping $h(\cdot) \mapsto \cos(u(1)) h(1)$ is linear and continuous with respect to $h \in C[0, 1]$, and therefore the Gâteaux derivative $f'(u)$ exists at any point $u \in U$ and satisfies

$$f'(u) h = \cos(u(1)) h(1).$$

Remark. In this example, it is impossible to express $f'(u)$ *without* reference to the increment h . We therefore have to use the evaluation rule for the functional $f'(u) \in U^*$.

(ii) *Square of the norm in Hilbert spaces.*

Let $\{H, (\cdot, \cdot)_H\}$ be a real Hilbert space equipped with the standard norm $\|\cdot\|_H$. We determine the Gâteaux derivative of the functional $f : H = \mathcal{U} \rightarrow \mathbb{R}$,

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

We have

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{1}{t} (f(u + th) - f(u)) &= \lim_{t \rightarrow 0} \frac{1}{t} \left(\|u + th\|_H^2 - \|u\|_H^2 \right) \\ &= \lim_{t \rightarrow 0} \frac{2t (u, h)_H + t^2 \|h\|_H^2}{t} \\ &= 2 (u, h)_H, \end{aligned}$$

and therefore

$$f'(u) h = (2u, h)_H.$$

If we identify H with its dual space H^* in the sense of the Riesz representation theorem, then we obtain for $f(u) = \|u\|_H^2$ the simple formula

$$f'(u) = 2u.$$

The expression, which results from identifying $f'(u)$ with an element of H , is called the *gradient* of f . We thus distinguish between the derivative, which is given by the rule $f'(u) h = (2u, h)_H$, and the gradient $f'(u) = 2u$.

(iii) *Application to the norm in $L^2(\Omega)$.*

By virtue of (ii), the Gâteaux derivative of the functional

$$f(u) := \|u(\cdot)\|_{L^2(\Omega)}^2 = \int_{\Omega} |u(x)|^2 dx$$

is given by

$$f'(u) h = \int_{\Omega} 2u(x) h(x) dx \quad \forall h \in L^2(\Omega).$$

Identification of $L^2(\Omega)^*$ and $L^2(\Omega)$ yields the gradient $(f'(u))(x) = 2u(x)$, for almost every $x \in \Omega$. \diamond

All the mappings considered in the above examples have even better differentiability properties. In fact, they are actually Fréchet differentiable.

Fréchet derivatives.

As before, let $\{U, \|\cdot\|_U\}$ and $\{V, \|\cdot\|_V\}$ denote real Banach spaces and \mathcal{U} an open subset of U .

Definition. A mapping $F : \mathcal{U} \subset U \rightarrow V$ is said to be Fréchet differentiable at $u \in \mathcal{U}$ if there exist an operator $A \in \mathcal{L}(U, V)$ and a mapping $r(u, \cdot) : U \rightarrow V$ with the following properties: for all $h \in U$ such that $u + h \in \mathcal{U}$, we have

$$F(u + h) = F(u) + Ah + r(u, h),$$

where the so-called remainder r satisfies the condition

$$\frac{\|r(u, h)\|_V}{\|h\|_U} \rightarrow 0 \quad \text{as } \|h\|_U \rightarrow 0.$$

The operator A is then called the Fréchet derivative of F at u , and we write $A = F'(u)$. If A is Fréchet differentiable at every point $u \in \mathcal{U}$, then A is said to be Fréchet differentiable in \mathcal{U} .

Since \mathcal{U} is an open set, we have $u + h \in \mathcal{U}$ for all $h \in U$ with sufficiently small norm. Hence, the relation to be satisfied by the remainder $r(u, h)$ is meaningful at least for all $h \in U$ from a small ball about the origin. We also remark that it is often more convenient to prove Fréchet differentiability by showing that

$$(2.39) \quad \frac{\|F(u + h) - F(u) - Ah\|_V}{\|h\|_U} \rightarrow 0 \quad \text{as } \|h\|_U \rightarrow 0,$$

which is obviously equivalent to postulating that $F(u + h) - F(u) - Ah = r(u, h)$, where $\|r(u, h)\|_V/\|h\|_U \rightarrow 0$ as $\|h\|_U \rightarrow 0$.

Examples.

(iv) The following function taken from [IT79] is a standard example illustrating the fact that Gâteaux differentiability is not sufficient to guarantee Fréchet differentiability: we consider the mapping $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x, y) = \begin{cases} 1 & \text{if } y = x^2 \text{ and } x \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

This function is Gâteaux differentiable at the origin. It is, however, not even continuous at the origin, let alone Fréchet differentiable.

(v) The functional $f(u) = \sin(u(1))$ is Fréchet differentiable at every $u \in C[0, 1]$.

(vi) The mapping $f(u) = \|u\|_H^2$ is Fréchet differentiable on every Hilbert space H ; see Exercise 2.11.

(vii) Every continuous linear operator A is Fréchet differentiable. Indeed, $A(u + h) = Au + Ah + r(u, h)$ holds with $r(u, h) = 0$, and we conclude: “The derivative of a continuous linear operator is given by the operator itself.” \diamond

Calculation of Fréchet derivatives. Evidently, every Fréchet differentiable mapping F is also Gâteaux differentiable, and the two derivatives coincide (i.e., $F'_G(u) = F'(u)$; see also the remarks following the definition of the Gâteaux derivative). Hence, the explicit form of a Fréchet derivative can be determined through the Gâteaux derivative, which ultimately amounts to

calculating the directional derivative. This has already been demonstrated on pages 57 and 58.

Theorem 2.20 (Chain rule). *Suppose that Banach spaces U , V , and Z are given, and let $\mathcal{U} \subset U$ and $\mathcal{V} \subset V$ denote open sets. Let $F : \mathcal{U} \rightarrow \mathcal{V}$ and $G : \mathcal{V} \rightarrow Z$ be Fréchet differentiable at $u \in \mathcal{U}$ and at $F(u) \in \mathcal{V}$, respectively. Then the composition $E = G \circ F : \mathcal{U} \rightarrow Z$, defined by $E(u) = G(F(u))$, is Fréchet differentiable at u , and*

$$E'(u) = G'(F(u)) F'(u).$$

Example. Let two real Hilbert spaces $\{U, (\cdot, \cdot)_U\}$ and $\{H, (\cdot, \cdot)_H\}$ be given, and let $z \in H$ be fixed. For some $S \in \mathcal{L}(U, H)$ we consider the functional $E : U \rightarrow \mathbb{R}$,

$$E(u) = \|Su - z\|_H^2.$$

In this case, E can be expressed in the form $E(u) = G(F(u))$, where $G(v) = \|v\|_H^2$ and $F(u) = Su - z$. We know already from examples (ii) and (vi) that

$$G'(v)h = (2v, h)_H, \quad F'(u)h = Sh.$$

The chain rule thus yields

$$\begin{aligned} E'(u)h &= G'(F(u))F'(u)h = (2v, F'(u)h)_H \\ (2.40) \quad &= 2(Su - z, Sh)_H \\ &= 2(S^*(Su - z), h)_U. \end{aligned}$$

Here, $S^* \in \mathcal{L}(H^*, U^*)$ denotes the so-called *adjoint* of S , which will be defined in Section 2.7. \diamond

Remark. The above results and further information concerning the differentiability of operators and functionals can be found, e.g., in [Car67], [IT79], [Jah94], and [KA64].

2.7. Adjoint operators

If A is an $m \times n$ matrix and A^\top its transpose, then

$$(Au, v)_{\mathbb{R}^m} = (u, A^\top v)_{\mathbb{R}^n} \quad \text{for all } u \in \mathbb{R}^n \text{ and } v \in \mathbb{R}^m.$$

In a similar way, for real Hilbert spaces $\{U, (\cdot, \cdot)_U\}$ and $\{V, (\cdot, \cdot)_V\}$ one can assign to any linear and continuous operator $A : U \rightarrow V$ a so-called adjoint operator A^* , which allows the transformation $(Au, v)_V = (u, A^*v)_U$ for all $u \in U$ and $v \in V$.

The corresponding definition in Banach spaces is more general. To this end, let two real Banach spaces U and V , a continuous linear operator $A :$

$U \rightarrow V$, and a functional $f \in V^*$ be given. We can then define the functional $g = f \circ A : U \rightarrow \mathbb{R}$,

$$g(u) = f(Au).$$

The mapping g is obviously linear; its continuity follows from the estimate

$$|g(u)| \leq \|f\|_{V^*} \|A\|_{\mathcal{L}(U,V)} \|u\|_U.$$

Hence, g belongs to the dual space U^* , and we have the estimate

$$(2.41) \quad \|g\|_{U^*} \leq \|A\|_{\mathcal{L}(U,V)} \|f\|_{V^*}.$$

Definition. *The mapping $A^* : V^* \rightarrow U^*$ defined by $f \mapsto g = f \circ A$ is called the adjoint operator or dual operator of A .*

It follows from the above arguments that

$$(A^* f)(u) = f(Au) \quad \forall u \in U,$$

$$\|A^* f\|_{U^*} \leq \|A\|_{\mathcal{L}(U,V)} \|f\|_{V^*} \quad \forall f \in V^*.$$

Remark. In many texts the notation A' for the adjoint or dual operator is used. We have chosen to write A^* in order to avoid any possible confusion with derivatives. We also remark that the notion of *adjoint operator* is often reserved for Hilbert spaces. Below we will therefore—but only for a moment—write A^* ; note the typographical difference between A^* and A^* . For the definition of the adjoint or dual operator, we follow Alt [Alt99], Kreyszig [Kre78], and Wouk [Wou79].

An immediate consequence of estimate (2.41) is that A^* is continuous, so that $A^* \in \mathcal{L}(V^*, U^*)$. More precisely, we have $\|A^*\|_{\mathcal{L}(V^*, U^*)} \leq \|A\|_{\mathcal{L}(U, V)}$. We even have equality of these norms; see, e.g., [LS74], [Wou79].

For better readability, in the following we will make use of the so-called *duality pairing*, which resembles a scalar product: if a functional $f \in V^*$ is evaluated at $v \in V$, then we write

$$f(v) = \langle f, v \rangle_{V^*, V}.$$

This notation makes the definition of the operator A^* more transparent; indeed, we have

$$\langle f, Au \rangle_{V^*, V} = \langle A^* f, u \rangle_{U^*, U} =: \langle u, A^* f \rangle_{U, U^*} \quad \forall f \in V^*, \forall u \in U.$$

This form, while easily memorized, can lead to the misconception that A^* is already explicitly determined by it (for instance, in terms of a matrix representation or via an integral operator). This is, however, not to be expected, since a functional $f \in V^*$ may admit several completely different representations; see (2.24) on page 42. Explicit expressions for adjoint operators can be derived if results like the Riesz representation theorem are available that

provide a characterization in concrete form of continuous linear functionals. We therefore confine ourselves from now on to adjoint operators in Hilbert spaces.

Definition. Let real Hilbert spaces $\{U, (\cdot, \cdot)_U\}$ and $\{V, (\cdot, \cdot)_V\}$ as well as an operator $A \in \mathcal{L}(U, V)$ be given. An operator A^* is called the Hilbert space adjoint or adjoint of A if

$$(2.42) \quad (v, Au)_V = (A^*v, u)_U \quad \forall u \in U, \quad \forall v \in V.$$

Remark. The terms *dual*, *adjoint*, and *Hilbert space adjoint* are not used consistently in the literature. We will use *adjoint operator* in both Banach and Hilbert spaces, since the definition will always become clear from the context. Moreover, dual spaces and adjoint operators will generally be marked by $*$.

Examples.

(i) Let $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ denote a linear operator, which is represented by an $m \times n$ matrix also denoted by A . Since $(v, Au)_{\mathbb{R}^m} = (A^\top v, u)_{\mathbb{R}^n}$ for all $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, the (Hilbert space) adjoint A^* can be identified with the transposed matrix A^\top .

(ii) We consider in the Hilbert space $L^2(0, 1)$ the integral operator

$$(Au)(t) = \int_0^t e^{(t-s)} u(s) ds.$$

It is easily seen that A is linear and continuous on $L^2(0, 1)$; see Exercise 2.12. Its adjoint A^* can be calculated as follows:

$$\begin{aligned} (v, Au)_{L^2(0,1)} &= \int_0^1 v(t) \left(\int_0^t e^{(t-s)} u(s) ds \right) dt \\ &= \int_0^1 \int_0^t v(t) e^{(t-s)} u(s) ds dt \\ &= \int_0^1 \int_s^1 v(t) e^{(t-s)} u(s) dt ds && \text{(Fubini's theorem)} \\ &= \int_0^1 u(s) \left(\int_s^1 e^{(t-s)} v(t) dt \right) ds \\ &= \int_0^1 \left(\int_t^1 e^{(s-t)} v(s) ds \right) u(t) dt && \text{(exchange of variables)} \\ &= (A^*v, u)_{L^2(0,1)}, \end{aligned}$$

where the adjoint operator has the representation

$$(A^*v)(t) = \int_t^1 v(s)e^{(s-t)} ds. \quad \diamond$$

2.8. First-order necessary optimality conditions

In Section 2.5, the existence and uniqueness of optimal controls was demonstrated for selected types of elliptic optimal control problems. In this section, we will invoke the first derivative of the cost functional to derive conditions that optimal solutions have to satisfy. These necessary conditions allow for far-reaching conclusions concerning the form of optimal controls and the verification that numerically determined controls are actually optimal. In addition, they form the theoretical basis for the development of numerical methods.

2.8.1. Quadratic optimization in Hilbert spaces. For proving the existence of optimal controls, we transformed the control problems under investigation into a reduced quadratic optimization problem in terms of u , namely

$$(2.43) \quad \min_{u \in U_{ad}} f(u) := \frac{1}{2} \|Su - y_d\|_H^2 + \frac{\lambda}{2} \|u\|_U^2.$$

To this minimization problem, the following fundamental result can be applied. It is the key to the derivation of first-order necessary optimality conditions in the presence of control constraints.

Lemma 2.21. *Let C denote a nonempty and convex subset of a real Banach space U , and let the real-valued mapping f be Gâteaux differentiable in an open subset of U containing C . If $\bar{u} \in C$ is a solution to the problem*

$$\min_{u \in C} f(u),$$

then it solves the variational inequality

$$(2.44) \quad f'(\bar{u})(u - \bar{u}) \geq 0 \quad \forall u \in C.$$

Conversely, if $\bar{u} \in C$ solves the variational inequality (2.44) and f is convex, then \bar{u} is a solution to the minimization problem $\min_{u \in C} f(u)$.

Proof: Let $u \in C$ be arbitrary. Since C is convex, $\bar{u} + t(u - \bar{u}) \in C$ for any $t \in (0, 1]$. Since \bar{u} is optimal, $f(\bar{u} + t(u - \bar{u})) \geq f(\bar{u})$ and hence also

$$\frac{1}{t} (f(\bar{u} + t(u - \bar{u})) - f(\bar{u})) \geq 0 \quad \text{for } t \in (0, 1].$$

Letting $t \downarrow 0$, we arrive at $f'(\bar{u})(u - \bar{u}) \geq 0$, which proves the validity of (2.44).

Now suppose that \bar{u} solves the variational inequality. Since f is convex, it follows from a standard argument that

$$f(u) - f(\bar{u}) \geq f'(\bar{u})(u - \bar{u}) \quad \forall u \in C.$$

By (2.44), the right-hand side of this inequality is nonnegative, whence $f(u) \geq f(\bar{u})$ follows. This concludes the proof of the assertion. \square

Lemma 2.21 yields a *necessary*, and in the case of convexity also *sufficient*, so-called *first-order optimality condition*. It is apparent that the result remains valid if merely the existence of all directional derivatives of f is postulated. It can even make sense to consider only the directional derivatives with respect to all directions from a dense subspace of U , as the following example shows.

Example. Let $\varepsilon \in (0, 1)$ be fixed, and let $C_\varepsilon = \{u \in L^2(a, b) : u(x) \geq \varepsilon \text{ for a.e. } x \in (a, b)\}$. The functional

$$f(u) = \int_a^b \ln(u(x)) \, dx,$$

which is well defined on C_ε , is not Gâteaux differentiable at $\bar{u} \in C$, $\bar{u}(x) \equiv 1$, in the sense of $L^2(a, b)$. However, directional derivatives exist in any direction $h \in L^\infty(a, b)$. In fact, we have

$$\delta f(\bar{u}, h) = \int_a^b \frac{h(x)}{\bar{u}(x)} \, dx = \int_a^b h(x) \, dx.$$

Functionals of this type occur in the study of interior-point methods for the solution of optimization problems in function spaces. \diamond

We are now going to apply Lemma 2.21 to the quadratic optimization problem (2.43).

Theorem 2.22. *Suppose that real Hilbert spaces U and H , a nonempty and convex set $U_{ad} \subset U$, some $y_d \in H$, and a constant $\lambda \geq 0$ are given. Moreover, let $S : U \rightarrow H$ denote a continuous linear operator. Then $\bar{u} \in U_{ad}$ is a solution to the minimization problem (2.43) if and only if \bar{u} solves the variational inequality*

$$(2.45) \quad (S^*(S\bar{u} - y_d) + \lambda\bar{u}, u - \bar{u})_U \geq 0 \quad \forall u \in U_{ad}.$$

Proof: In view of (2.40), the gradient of the functional f defined in (2.43) is of the form

$$(2.46) \quad f'(\bar{u}) = S^*(S\bar{u} - y_d) + \lambda\bar{u}.$$

The assertion is thus a direct consequence of Lemma 2.21. \square

In many instances it is advantageous to write the variational inequality (2.45) in the equivalent form

$$(2.47) \quad (S\bar{u} - y_d, Su - S\bar{u})_H + \lambda (\bar{u}, u - \bar{u})_U \geq 0 \quad \forall u \in U_{ad},$$

which avoids the adjoint operator S^* .

Below, we apply the variational inequality to our various optimal control problems, following the scheme indicated in Section 1.4.

2.8.2. Optimal stationary heat source. The problem (2.26)–(2.28) defined on page 49 reads

$$\min J(y, u) := \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2,$$

subject to

$$\boxed{\begin{array}{l} -\Delta y = \beta u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}}$$

and

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Omega.$$

As above, we denote the solution operator of the boundary value problem by S , viewed as a mapping in $L^2(\Omega)$. In view of (2.45), any optimal control \bar{u} must obey the variational inequality

$$(2.48) \quad (S^*(S\bar{u} - y_\Omega) + \lambda\bar{u}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall u \in U_{ad},$$

where the adjoint operator S^* is yet to be determined. For this purpose, we prove the following preparatory result.

Lemma 2.23. *Let functions $z, u \in L^2(\Omega)$ and $c_0, \beta \in L^\infty(\Omega)$ with $c_0 \geq 0$ a.e. in Ω be given, and let y and p denote, respectively, the weak solutions to the elliptic boundary value problems*

$$\begin{array}{ll} -\Delta y + c_0 y = \beta u & -\Delta p + c_0 p = z \quad \text{in } \Omega \\ y = 0 & p = 0 \quad \text{on } \Gamma. \end{array}$$

Then

$$(2.49) \quad \int_{\Omega} z y \, dx = \int_{\Omega} \beta p u \, dx.$$

Proof: We invoke the variational formulations of the above boundary value problems. For y , insertion of the test function $p \in H_0^1(\Omega)$ yields

$$\int_{\Omega} (\nabla y \cdot \nabla p + c_0 y p) \, dx = \int_{\Omega} \beta p u \, dx,$$

while for p we obtain with the test function $y \in H_0^1(\Omega)$ that

$$\int_{\Omega} \left(\nabla p \cdot \nabla y + c_0 p y \right) dx = \int_{\Omega} z y dx.$$

Since the left-hand sides are equal, the assertion immediately follows. \square

Lemma 2.24. *For the boundary value problem (2.27), the adjoint operator $S^* : L^2(\Omega) \rightarrow L^2(\Omega)$ is given by*

$$S^* z := \beta p,$$

where $p \in H_0^1(\Omega)$ is the weak solution to the boundary value problem

$$\begin{aligned} -\Delta p &= z \quad \text{in } \Omega \\ p &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Proof: According to (2.42) on page 62, the operator S^* is given by the relation

$$(z, Su)_{L^2(\Omega)} = (S^* z, u)_{L^2(\Omega)} \quad \forall z \in L^2(\Omega), \quad \forall u \in L^2(\Omega).$$

Invoking Lemma 2.23 with $c_0 = 0$ and $y = Su$, we find that

$$(z, Su)_{L^2(\Omega)} = (z, y)_{L^2(\Omega)} = (\beta p, u)_{L^2(\Omega)}.$$

Owing to Theorem 2.4 on page 33, the mapping $z \mapsto \beta p$ is linear and continuous from $L^2(\Omega)$ into itself. Since z and u can be chosen arbitrarily and S^* is uniquely determined, we conclude that $S^* z = \beta p$. \square

The construction of S^* in the above proof, which is based on Lemma 2.23, is not easy to understand intuitively. In Section 2.10, we will get acquainted with the *formal Lagrange method*, which is an effective tool for finding the form of the partial differential equation from which S^* can be determined.

Remark. As we know, $S = E_Y G$ has range in the space $H_0^1(\Omega)$. However, if we had considered the operator $G : L^2(\Omega) \rightarrow H_0^1(\Omega)$ instead of S , then (after identifying $L^2(\Omega)^*$ with $L^2(\Omega)$) the adjoint operator $G^* : H_0^1(\Omega)^* \rightarrow L^2(\Omega)$ would have occurred. We have avoided the space $H_0^1(\Omega)^*$ by choosing $S : L^2(\Omega) \rightarrow L^2(\Omega)$. This choice restricts the applicability of the above theory to a certain extent; it is, however, simpler and suffices for the time being. In Section 2.13, we will briefly explain how to work in $H_0^1(\Omega)^*$. There are good reasons not to identify $H_0^1(\Omega)^*$ with the Hilbert space $H_0^1(\Omega)$ in this approach.

Adjoint state and optimality system. The variational inequality (2.48) can be easily transformed if S^* is known.

Definition. The weak solution $p \in H_0^1(\Omega)$ to the adjoint equation

$$(2.50) \quad \begin{aligned} -\Delta p &= \bar{y} - y_\Omega && \text{in } \Omega \\ p &= 0 && \text{on } \Gamma \end{aligned}$$

is called the adjoint state associated with \bar{y} .

The right-hand side of the adjoint equation belongs to $L^2(\Omega)$, since $y_\Omega \in L^2(\Omega)$ by assumption and $\bar{y} \in Y = H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. From Theorem 2.4 on page 33, we infer that (2.50) admits a unique solution $p \in H_0^1(\Omega)$. Putting $z = \bar{y} - y_\Omega$, we conclude from Lemma 2.24 that

$$S^*(S\bar{u} - y_\Omega) = S^*(\bar{y} - y_\Omega) = \beta p,$$

whence, upon invoking (2.48),

$$(\beta p + \lambda \bar{u}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall u \in U_{ad}.$$

Thus, it follows directly from the variational inequality (2.44) that the following result holds.

Theorem 2.25. Suppose that \bar{u} is an optimal control for the problem (2.26)–(2.28) of optimal stationary heat sources from page 49, and let \bar{y} denote the associated state. Then the adjoint equation (2.50) has a unique weak solution p that satisfies the variational inequality

$$(2.51) \quad \int_{\Omega} (\beta(x)p(x) + \lambda \bar{u}(x))(u(x) - \bar{u}(x)) \, dx \geq 0 \quad \forall u \in U_{ad}.$$

Conversely, any control $\bar{u} \in U_{ad}$ which, together with its associated state $\bar{y} = y(\bar{u})$ and the solution p to (2.50), satisfies the variational inequality (2.51) is optimal.

The sufficiency part of the statement follows from the convexity of f . Hence, a control u , together with the optimal state y and the adjoint state p , is optimal for problem (2.26)–(2.28) if and only if the triple (u, y, p) satisfies the following *optimality system*:

$$(2.52) \quad \boxed{\begin{aligned} -\Delta y &= \beta u & -\Delta p &= y - y_\Omega \\ y|_\Gamma &= 0 & p|_\Gamma &= 0 \\ & & u &\in U_{ad} \\ (\beta p + \lambda u, v - u)_{L^2(\Omega)} &\geq 0 & \forall v &\in U_{ad}. \end{aligned}}$$

Discussion of pointwise optimality conditions. In this section, we perform a detailed analysis of the variational inequality (2.51). We begin

our investigation by rewriting it in the form

$$\int_{\Omega} (\beta p + \lambda \bar{u}) \bar{u} \, dx \leq \int_{\Omega} (\beta p + \lambda \bar{u}) u \, dx \quad \forall u \in U_{ad}$$

hence

$$(2.53) \quad \int_{\Omega} (\beta p + \lambda \bar{u}) \bar{u} \, dx = \min_{u \in U_{ad}} \int_{\Omega} (\beta p + \lambda \bar{u}) u \, dx.$$

Conclusion. *Under the assumption that the expression inside the bracket in (2.53) is known, we obtain \bar{u} as the solution to a linear optimization problem in a function space.*

This simple observation forms the basis of the *conditioned gradient method*; see Section 2.12.1.

It is intuitively clear that the variational inequality can also be formulated in *pointwise* form. The following lemma provides insight in this direction.

Lemma 2.26. *A necessary and sufficient condition for the variational inequality (2.51) to be satisfied is that for almost every $x \in \Omega$,*

$$(2.54) \quad \bar{u}(x) = \begin{cases} u_a(x) & \text{if } \beta(x)p(x) + \lambda \bar{u}(x) > 0 \\ \in [u_a(x), u_b(x)] & \text{if } \beta(x)p(x) + \lambda \bar{u}(x) = 0 \\ u_b(x) & \text{if } \beta(x)p(x) + \lambda \bar{u}(x) < 0. \end{cases}$$

An equivalent condition is given by the pointwise variational inequality in \mathbb{R} ,

$$(2.55) \quad (\beta(x)p(x) + \lambda \bar{u}(x))(v - \bar{u}(x)) \geq 0 \quad \forall v \in [u_a(x), u_b(x)], \text{ for a.e. } x \in \Omega.$$

Proof: (i) First, we show that (2.51) implies (2.54). To this end, let \bar{u} , u_a , and u_b be arbitrary but fixed representatives of the corresponding equivalence classes in the sense of L^∞ . Suppose that (2.54) is false. We consider the measurable sets

$$\begin{aligned} A_+(\bar{u}) &= \{x \in \Omega : \beta(x)p(x) + \lambda \bar{u}(x) > 0\}, \\ A_-(\bar{u}) &= \{x \in \Omega : \beta(x)p(x) + \lambda \bar{u}(x) < 0\}. \end{aligned}$$

By our assumption, there is a set $E_+ \subset A_+(\bar{u})$ having positive measure such that $\bar{u}(x) > u_a(x)$ for all $x \in E_+$, or there is a set $E_- \subset A_-(\bar{u})$ having positive measure such that $\bar{u}(x) < u_b(x)$ for all $x \in E_-$. In the first case,

we define the function $u \in U_{ad}$,

$$u(x) = \begin{cases} u_a(x) & \text{for } x \in E_+ \\ \bar{u}(x) & \text{for } x \in \Omega \setminus E_+. \end{cases}$$

Then

$$\begin{aligned} & \int_{\Omega} (\beta(x)p(x) + \lambda \bar{u}(x))(u(x) - \bar{u}(x)) dx \\ &= \int_{E_+} (\beta(x)p(x) + \lambda \bar{u}(x))(u_a(x) - \bar{u}(x)) dx < 0, \end{aligned}$$

since the first factor is positive on E_+ while the second is negative. This evidently contradicts (2.51). The other case can be handled in a similar way by putting $u(x) = u_b(x)$ on E_- and $u(x) = \bar{u}(x)$ otherwise.

(ii) Next, we show that (2.54) implies (2.55). We have for almost every $x \in A_+(\bar{u})$ that $\bar{u}(x) = u_a(x)$, and thus $v - \bar{u}(x) \geq 0$ for all $v \in [u_a(x), u_b(x)]$. Hence, by the definition of $A_+(\bar{u})$,

$$(\beta(x)p(x) + \lambda \bar{u}(x))(v - \bar{u}(x)) \geq 0 \quad \text{for almost every } x \in A_+(\bar{u}).$$

Similar reasoning shows that this inequality also holds almost everywhere in $A_-(\bar{u})$. Since this is trivially the case whenever $\beta(x)p(x) + \lambda \bar{u}(x) = 0$, (2.55) holds almost everywhere in Ω .

(iii) Finally, we show that (2.55) implies (2.51). To this end, let $u \in U_{ad}$ be arbitrarily chosen. Since $\bar{u}(x) \in [u_a(x), u_b(x)]$ for almost every $x \in \Omega$, we may put $v := u(x)$ in (2.55) to find that

$$(\beta(x)p(x) + \lambda \bar{u}(x))(u(x) - \bar{u}(x)) \geq 0 \quad \text{for a.e. } x \in \Omega.$$

Integration yields that (2.51) holds. \square

Next we observe that by a simple rearrangement of terms the pointwise variational inequality (2.55) can be rewritten in the form

$$(2.56) \quad (\beta(x)p(x) + \lambda \bar{u}(x)) \bar{u}(x) \leq (\beta(x)p(x) + \lambda \bar{u}(x)) v \quad \forall v \in [u_a(x), u_b(x)],$$

for almost every $x \in \Omega$. Here, as in (2.55), v is a real number, not a function.

Theorem 2.27. *A control $\bar{u} \in U_{ad}$ is optimal for (2.26)–(2.28) if and only if it satisfies, together with the adjoint state p from (2.50), one of the following two minimum conditions for almost all $x \in \Omega$:*

the weak minimum principle

$$\min_{v \in [u_a(x), u_b(x)]} \left\{ (\beta(x)p(x) + \lambda \bar{u}(x)) v \right\} = (\beta(x)p(x) + \lambda \bar{u}(x)) \bar{u}(x)$$

or the minimum principle

$$\min_{v \in [u_a(x), u_b(x)]} \left\{ \beta(x) p(x) v + \frac{\lambda}{2} v^2 \right\} = \beta(x) p(x) \bar{u}(x) + \frac{\lambda}{2} \bar{u}(x)^2.$$

Proof: The weak minimum principle is evidently nothing but a reformulation of (2.56). The minimum principle is also easily verified: a real number \bar{v} solves for fixed x the (convex) quadratic optimization problem in \mathbb{R} ,

$$\min_{v \in [u_a(x), u_b(x)]} g(v) := \beta(x) p(x) v + \frac{\lambda}{2} v^2,$$

if and only if the variational inequality

$$g'(\bar{v})(v - \bar{v}) \geq 0 \quad \forall v \in [u_a(x), u_b(x)]$$

is satisfied, that is, if

$$(\beta(x) p(x) + \lambda \bar{v})(v - \bar{v}) \geq 0 \quad \forall v \in [u_a(x), u_b(x)].$$

The minimum condition follows from taking $\bar{v} = \bar{u}(x)$. \square

The derived pointwise conditions can be further evaluated in order to extract additional information. Depending on the choice of λ , different consequences result.

Case 1: $\lambda = 0$. In this case, it follows from (2.54) that almost everywhere,

$$(2.57) \quad \bar{u}(x) = \begin{cases} u_a(x) & \text{if } \beta(x) p(x) > 0 \\ u_b(x) & \text{if } \beta(x) p(x) < 0. \end{cases}$$

At points $x \in \Omega$ where $\beta(x) p(x) = 0$, no information concerning $\bar{u}(x)$ can be extracted. If $\beta(x) p(x) \neq 0$ almost everywhere in Ω , then \bar{u} is a so-called *bang-bang control*, that is, the values $\bar{u}(x)$ coincide almost everywhere with one of the threshold values $u_a(x)$ or $u_b(x)$.

Case 2: $\lambda > 0$. We interpret the second relation in (2.54) as saying that “ \bar{u} is undetermined if $\lambda \bar{u} + \beta p = 0$ ”. This is not really true, since the equation $\lambda \bar{u} + \beta p = 0$ yields $\bar{u}(x) = -\lambda^{-1} \beta(x) p(x)$ and therefore provides a hint towards a complete understanding of the minimum condition.

Theorem 2.28. *If $\lambda > 0$, then \bar{u} is an optimal control to the problem (2.26)–(2.28) if and only if it satisfies, together with the associated adjoint state p , the projection formula*

$$(2.58) \quad \bar{u}(x) = \mathbb{P}_{[u_a(x), u_b(x)]} \left\{ -\frac{1}{\lambda} \beta(x) p(x) \right\} \quad \text{for almost every } x \in \Omega.$$

Here, for real numbers $a \leq b$ $\mathbb{P}_{[a,b]}$ denotes the projection of \mathbb{R} onto $[a, b]$,

$$\mathbb{P}_{[a,b]}(u) := \min \{b, \max\{a, u\}\}.$$

Proof: The assertion is a direct consequence of Theorem 2.27: indeed, the solution to the quadratic optimization problem in \mathbb{R} formulated in terms of the minimum principle

$$\min_{v \in [u_a(x), u_b(x)]} \left\{ \beta(x)p(x)v + \frac{\lambda}{2} v^2 \right\}$$

is given by the projection formula (2.58). The reader will be asked to verify this claim in Exercise 2.13. \square

Case 2a: $\lambda > 0$ and $U_{ad} = L^2(\Omega)$. In this case, the control is unconstrained, and we can infer from (2.58) (or directly from (2.55)) that

$$(2.59) \quad \bar{u} = -\frac{1}{\lambda} \beta p.$$

Putting this in the state equation leads to the optimality system

$-\Delta y = -\lambda^{-1} \beta^2 p$	$-\Delta p = y - y_\Omega$
$y _\Gamma = 0$	$p _\Gamma = 0.$

This is a coupled system of two elliptic boundary value problems for the determination of $y = \bar{y}$ and p . Once p has been found, the optimal control \bar{u} is obtained from (2.59).

Formulation as a Karush–Kuhn–Tucker system. By the introduction of Lagrange multipliers, the variational inequality (2.51) in the optimality system can be reformulated in terms of additional equations. The associated technique was explained in Section 1.4.7.

Theorem 2.29. *The variational inequality (2.51) is equivalent to the existence of almost-everywhere nonnegative functions $\mu_a, \mu_b \in L^2(\Omega)$ that satisfy the equation*

$$(2.60) \quad \beta p + \lambda \bar{u} - \mu_a + \mu_b = 0$$

as well as the complementarity conditions

$$(2.61) \quad \mu_a(x) (u_a(x) - \bar{u}(x)) = \mu_b(x) (\bar{u}(x) - u_b(x)) = 0 \quad \text{for a.e. } x \in \Omega.$$

Proof. (i) We first show that (2.60) and (2.61) are consequences of the variational inequality (2.51). To this end, we follow the treatment in Section

1.4.7 and define the functions

$$(2.62) \quad \begin{aligned} \mu_a(x) &:= (\beta(x)p(x) + \lambda\bar{u}(x))_+, \\ \mu_b(x) &:= (\beta(x)p(x) + \lambda\bar{u}(x))_-. \end{aligned}$$

Here, we use the usual definitions of s_+ and s_- for $s \in \mathbb{R}$, namely

$$s_+ = \frac{1}{2}(s + |s|), \quad s_- = \frac{1}{2}(|s| - s).$$

Then, by definition, $\mu_a \geq 0$, $\mu_b \geq 0$, and $\beta p + \lambda\bar{u} = \mu_a - \mu_b$, which shows (2.60). Moreover, in view of (2.54), the following implications are valid for almost every $x \in \Omega$:

$$\begin{aligned} (\beta p + \lambda\bar{u})(x) > 0 &\Rightarrow \bar{u}(x) = u_a(x) \\ (\beta p + \lambda\bar{u})(x) < 0 &\Rightarrow \bar{u}(x) = u_b(x) \\ u_a(x) < \bar{u}(x) < u_b(x) &\Rightarrow (\beta p + \lambda\bar{u})(x) = 0. \end{aligned}$$

From these implications, we can conclude the validity of (2.61), since in both products at least one of the factors vanishes for almost all $x \in \Omega$. Indeed, suppose that $\mu_a(x) > 0$. Then obviously $\mu_b(x) = 0$; in addition, $(\beta p + \lambda\bar{u})(x) = \mu_a(x) > 0$, which implies that $\bar{u}(x) - u_a(x) = 0$. Next, suppose that $\mu_a(x) = 0$. We have to show that the second product also vanishes. In fact, if $\mu_b(x) > 0$, then $(\beta p + \lambda\bar{u})(x) < 0$, and thus $\bar{u}(x) - u_b(x) = 0$.

(ii) Conversely, assume that $\bar{u} \in U_{ad}$ satisfies (2.60) and (2.61), and let $u \in U_{ad}$ be given. We have to discuss three different cases.

First, for almost all x with $u_a(x) < \bar{u}(x) < u_b(x)$, it follows from the complementarity conditions (2.61) that $\mu_a(x) = \mu_b(x) = 0$, whence, upon invoking (2.60),

$$(\beta p + \lambda\bar{u})(x) = 0.$$

In conclusion, we have

$$(2.63) \quad (\beta(x)p(x) + \lambda\bar{u}(x)) (u(x) - \bar{u}(x)) \geq 0.$$

In the case where $u_a(x) = \bar{u}(x)$, we find, from $u \in U_{ad}$, that $u(x) - \bar{u}(x) \geq 0$. Moreover, equation (2.61) immediately yields that $\mu_b(x) = 0$. Therefore, we can infer from equation (2.60) that

$$\beta(x)p(x) + \lambda\bar{u}(x) = \mu_a(x) \geq 0,$$

whence inequality (2.63) again follows. The third case $\bar{u}(x) = u_b(x)$ is treated similarly. In summary, (2.63) holds for almost every $x \in \Omega$, and integration over Ω yields the validity of the variational inequality (2.51). This concludes the proof of the theorem. \square

By virtue of the above theorem, we can replace the optimality system (2.52), which contains the variational inequality, by the following *Karush–Kuhn–Tucker system*:

$$(2.64) \quad \boxed{\begin{aligned} -\Delta y &= \beta u & -\Delta p &= y - y_\Omega \\ y|_\Gamma &= 0 & p|_\Gamma &= 0 \\ \beta p + \lambda u - \mu_a + \mu_b &= 0 \\ u_a \leq u \leq u_b, \quad \mu_a \geq 0, \quad \mu_b \geq 0 \\ \mu_a (u_a - u) &= \mu_b (u - u_b) = 0. \end{aligned}}$$

Here, the relations in the last three lines hold for almost every $x \in \Omega$.

Definition. The functions $\mu_a, \mu_b \in L^2(\Omega)$ defined in Theorem 2.29 are called Lagrange multipliers associated with the inequality constraints $u_a \leq u$ and $u \leq u_b$, respectively.

Remark. The system (2.64) can be derived directly by using a Lagrangian function provided that the existence of multipliers $\mu_a, \mu_b \in L^2(\Omega)$ is assumed; see Section 6.1. However, the existence of such multipliers cannot directly be concluded from the Karush–Kuhn–Tucker theory in Banach spaces, since the set of almost-everywhere nonnegative functions in $L^2(\Omega)$ has empty interior. By explicitly defining the multipliers μ_a and μ_b , we have circumvented this difficulty here. A detailed analysis of this problem will be given in Section 6.1.

The reduced gradient of the cost functional. The calculation of the reduced gradient, that is, the gradient of $f(u) = J(y(u), u)$, is also simplified by invoking the adjoint state. The representation of $f'(u)$ given in the following lemma will apply to almost all optimal control problems to be studied in this book.

Lemma 2.30. *The gradient of the functional*

$$f(u) = J(y(u), u) = \frac{1}{2} \|y - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2$$

is given by

$$f'(u) = \beta p + \lambda u,$$

where $p \in H_0^1(\Omega)$ denotes the weak solution to the adjoint equation

$$(2.65) \quad \begin{aligned} -\Delta p &= y - y_\Omega & \text{in } \Omega \\ p &= 0 & \text{on } \Gamma \end{aligned}$$

and $y = y(u)$ is the state associated with u .

Proof: Invoking equation (2.46) on page 64, we conclude from Lemma 2.24 that

$$f'(u)h = (S^*(Su - y_\Omega) + \lambda u, h)_{L^2(\Omega)} = (\beta p + \lambda u, h)_{L^2(\Omega)}.$$

By virtue of the Riesz representation theorem, $f'(u)$ is identified with $\beta p + \lambda u$. \square

We conclude this section by reformulating the variational inequality (2.48) on page 65. Owing to the definition of the adjoint S^* , the variational inequality is equivalent to

$$(2.66) \quad (S\bar{u} - y_\Omega, Su - S\bar{u})_{L^2(\Omega)} + \lambda(\bar{u}, u - \bar{u})_{L^2(\Omega)} \geq 0 \quad \forall u \in U_{ad}.$$

With $\bar{y} = S\bar{u}$ and $y = Su$, it follows that

$$(2.67) \quad f'(\bar{u})(u - \bar{u}) = (\bar{y} - y_\Omega, y - \bar{y})_{L^2(\Omega)} + \lambda(\bar{u}, u - \bar{u})_{L^2(\Omega)} \geq 0.$$

The form (2.67) of the variational inequality makes it possible to apply the next result, Lemma 2.31, to determine S^* . Even though the operator S^* will not appear explicitly, it will stand behind the construction. We prefer to use this approach in what follows.

2.8.3. Stationary heat sources and boundary conditions of the third kind. In this section, we will treat the optimal control problem (2.69)–(2.71) defined below. We begin our analysis by proving an analogue of Lemma 2.23 that can be applied directly to determine the adjoint equation.

Lemma 2.31. *Let functions $a_\Omega, v \in L^2(\Omega)$, $a_\Gamma, u \in L^2(\Gamma)$, $c_0, \beta_\Omega \in L^\infty(\Omega)$, and $\alpha, \beta_\Gamma \in L^\infty(\Gamma)$ be given, where $\alpha \geq 0$ and $c_0 \geq 0$ almost everywhere. Moreover, let y and p denote the weak solutions to the elliptic boundary value problems*

$$\begin{aligned} -\Delta y + c_0 y &= \beta_\Omega v & -\Delta p + c_0 p &= a_\Omega \\ \partial_\nu y + \alpha y &= \beta_\Gamma u & \partial_\nu p + \alpha p &= a_\Gamma. \end{aligned}$$

Then

$$(2.68) \quad \int_\Omega a_\Omega y \, dx + \int_\Gamma a_\Gamma y \, ds = \int_\Omega \beta_\Omega p v \, dx + \int_\Gamma \beta_\Gamma p u \, ds.$$

Proof: We use the variational formulations of the above two boundary value problems. Inserting $p \in H^1(\Omega)$ in the equation for y , we find that

$$\int_\Omega (\nabla y \cdot \nabla p + c_0 y p) \, dx + \int_\Gamma \alpha y p \, ds = \int_\Omega \beta_\Omega p v \, dx + \int_\Gamma \beta_\Gamma p u \, ds,$$

and insertion of $y \in H^1(\Omega)$ in the equation for p yields

$$\int_{\Omega} (\nabla p \cdot \nabla y + c_0 p y) dx + \int_{\Gamma} \alpha p y ds = \int_{\Omega} a_{\Omega} y dx + \int_{\Gamma} a_{\Gamma} y ds.$$

From this, the assertion immediately follows. \square

With this result in hand, it is now easy to treat the problem of finding the optimal stationary heat source for a Robin boundary condition; for the sake of simplicity, we assume the latter to be homogeneous. We also include a boundary term in the cost functional. The problem then reads:

$$(2.69) \quad \min J(y, u) := \frac{\lambda_{\Omega}}{2} \|y - y_{\Omega}\|_{L^2(\Omega)}^2 + \frac{\lambda_{\Gamma}}{2} \|y - y_{\Gamma}\|_{L^2(\Gamma)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Omega)}^2,$$

subject to

$$(2.70) \quad \boxed{\begin{array}{ll} -\Delta y = \beta u & \text{in } \Omega \\ \partial_{\nu} y + \alpha y = 0 & \text{on } \Gamma \end{array}}$$

and

$$(2.71) \quad u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. in } \Omega.$$

We postulate that $\lambda \geq 0$, $\lambda_{\Omega} \geq 0$, $\lambda_{\Gamma} \geq 0$ and $\alpha \in L^{\infty}(\Gamma)$, where $\alpha \geq 0$ almost everywhere and $\|\alpha\|_{L^{\infty}(\Gamma)} \neq 0$, and also that $y_{\Omega} \in L^2(\Omega)$ and $y_{\Gamma} \in L^2(\Gamma)$. The optimal quantities \bar{u} , \bar{y} , and p then satisfy the optimality condition

$$\int_{\Omega} (\beta p + \lambda \bar{u})(u - \bar{u}) dx \geq 0 \quad \forall u \in U_{ad},$$

where the adjoint state p solves the boundary value problem

$$\boxed{\begin{array}{ll} -\Delta p = \lambda_{\Omega} (\bar{y} - y_{\Omega}) & \text{in } \Omega \\ \partial_{\nu} p + \alpha p = \lambda_{\Gamma} (\bar{y} - y_{\Gamma}) & \text{on } \Gamma. \end{array}}$$

The above relations are derived as in the next subsection, by invoking Lemma 2.31; see Exercise 2.14.

2.8.4. Optimal stationary boundary temperature. Let us recall the boundary control problem (2.32)–(2.34) from page 53:

$$\min J(y, u) := \frac{1}{2} \|y - y_{\Omega}\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Gamma)}^2,$$

subject to

$$\boxed{\begin{array}{ll} -\Delta y = 0 & \text{in } \Omega \\ \partial_\nu y + \alpha y = \alpha u & \text{on } \Gamma \end{array}}$$

and

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Gamma.$$

Owing to Theorem 2.6, the control-to-state operator $G : u \mapsto y(u)$ is a continuous linear mapping from $L^2(\Gamma)$ into $H^1(\Omega)$. However, we again consider G as an operator with range in $L^2(\Omega)$, that is, $S = E_Y G : L^2(\Gamma) \rightarrow L^2(\Omega)$, with the embedding operator $E_Y : H^1(\Omega) \rightarrow L^2(\Omega)$. The cost functional then attains the form

$$J(y, u) = f(u) = \frac{1}{2} \|Su - y_\Omega\|_{L^2(\Omega)}^2 + \frac{\lambda}{2} \|u\|_{L^2(\Gamma)}^2.$$

We now proceed similarly as in Section 2.8.2. A simpler method for the construction of the adjoint equation will be given later by the Lagrange method. To begin with, let $\bar{u} \in U_{ad}$ and \bar{y} denote the optimal control and its associated state, respectively. We employ Theorem 2.22 on page 64 and rearrange the resulting variational inequality as in (2.67) to get

$$(2.72) \quad f'(\bar{u})(u - \bar{u}) = (\bar{y} - y_\Omega, y - \bar{y})_{L^2(\Omega)} + \lambda (\bar{u}, u - \bar{u})_{L^2(\Gamma)} \geq 0.$$

We intend to apply Lemma 2.31. Comparison of the boundary value problems satisfied by y indicates that the choices $\beta_\Omega = 0$, $\beta_\Gamma = \alpha$, and $c_0 = 0$ have to be made. With this, we see that the expression $(\bar{y} - y_\Omega, y - \bar{y})_{L^2(\Omega)}$ attains the form of the left-hand side of equation (2.68), provided we replace $y - \bar{y}$ by y and make the choices $a_\Omega = \bar{y} - y_\Omega$ and $a_\Gamma = 0$. Our plan is to express $y - \bar{y}$ in terms of $u - \bar{u}$ in order to calculate $f'(u)$ from (2.72).

In view of the above considerations, we are motivated to define p as the solution to the following adjoint equation:

$$(2.73) \quad \boxed{\begin{array}{ll} -\Delta p = \bar{y} - y_\Omega & \text{in } \Omega \\ \partial_\nu p + \alpha p = 0 & \text{on } \Gamma. \end{array}}$$

The right-hand side of the differential equation belongs to $L^2(\Omega)$, since $y_\Omega \in L^2(\Omega)$ by assumption and $\bar{y} \in Y = H^1(\Omega) \hookrightarrow L^2(\Omega)$. Owing to Theorem 2.6, problem (2.73) admits a unique weak solution $p \in H^1(\Omega)$ that

satisfies

$$(2.74) \quad \int_{\Omega} \nabla p \cdot \nabla v \, dx + \int_{\Gamma} \alpha p v \, ds = \int_{\Omega} (\bar{y} - y_{\Omega}) v \, dx \quad \forall v \in H^1(\Omega).$$

The optimal state $\bar{y} = S\bar{u}$ is the weak solution to the state equation associated with \bar{u} , while $y = Su$ corresponds to u . Hence, by the linearity of the state equation, we have $y - \bar{y} = S(u - \bar{u})$. Lemma 2.31 applied with $y = y - \bar{y}$ and $v = u - \bar{u}$ yields that

$$\int_{\Omega} (\bar{y} - y_{\Omega})(y - \bar{y}) \, dx = \int_{\Gamma} \alpha p (u - \bar{u}) \, ds.$$

With this, (2.72) becomes

$$f'(\bar{u})(u - \bar{u}) = \int_{\Gamma} (\lambda \bar{u} + \alpha p)(u - \bar{u}) \, ds \geq 0 \quad \forall u \in U_{ad}.$$

The form of the derivative $f'(\bar{u})$ does not depend on the fact that \bar{u} is optimal. Hence, we obtain as a side result that the reduced gradient $f'(u)$ at an arbitrary u is of the form

$$(2.75) \quad f'(u) = \alpha p|_{\Gamma} + \lambda u,$$

where p solves the associated adjoint equation

$$\begin{aligned} -\Delta p &= y(u) - y_{\Omega} && \text{in } \Omega \\ \partial_{\nu} p + \alpha p &= 0 && \text{on } \Gamma. \end{aligned}$$

In accordance with the Riesz representation theorem, we have expressed the derivative $f'(u)$ as an element of $L^2(\Gamma)$, namely the gradient.

Summarizing the above considerations, we have proved the following result.

Theorem 2.32. *Let \bar{u} denote an optimal control to the problem (2.32)–(2.34) on page 53, and let \bar{y} denote the associated state. Then the adjoint equation (2.73) has a unique solution p such that the variational inequality*

$$(2.76) \quad \int_{\Gamma} (\alpha(x)p(x) + \lambda \bar{u}(x))(u(x) - \bar{u}(x)) \, ds(x) \geq 0 \quad \forall u \in U_{ad}$$

is satisfied. Conversely, every control $\bar{u} \in U_{ad}$ that, together with $\bar{y} := y(\bar{u})$ and the solution p to (2.73), solves the variational inequality (2.76) is optimal.

Further discussion of the variational inequality (2.76) follows the same lines as in the case of Poisson's equation. In this case, we obtain that

$$(2.77) \quad \bar{u}(x) = \begin{cases} u_a(x) & \text{if } \alpha(x)p(x) + \lambda \bar{u}(x) > 0 \\ \in [u_a(x), u_b(x)] & \text{if } \alpha(x)p(x) + \lambda \bar{u}(x) = 0 \\ u_b(x) & \text{if } \alpha(x)p(x) + \lambda \bar{u}(x) < 0, \end{cases}$$

and the *weak minimum principle* becomes

$$\min_{u_a(x) \leq v \leq u_b(x)} \left\{ (\alpha(x)p(x) + \lambda \bar{u}(x)) v \right\} = (\alpha(x)p(x) + \lambda \bar{u}(x)) \bar{u}(x)$$

for almost every $x \in \Gamma$.

In addition, we have the following result.

Theorem 2.33 (Minimum principle). *Suppose that \bar{u} is an optimal control for the problem (2.32)–(2.34) on page 53, and let p denote the associated adjoint state. Then, for almost every $x \in \Gamma$, the minimum*

$$\min_{u_a(x) \leq v \leq u_b(x)} \left\{ \alpha(x)p(x)v + \frac{\lambda}{2}v^2 \right\}$$

is attained at $v = \bar{u}(x)$. Hence, for $\lambda > 0$ we have for almost every $x \in \Gamma$ the projection formula

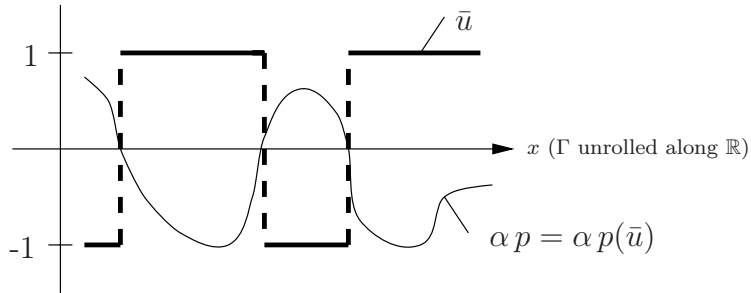
$$(2.78) \quad \bar{u}(x) = \mathbb{P}_{[u_a(x), u_b(x)]} \left\{ -\frac{1}{\lambda} \alpha(x)p(x) \right\}.$$

Conversely, a control $\bar{u} \in U_{ad}$ is optimal if it satisfies, together with the associated adjoint state p , the projection formula (2.78).

The proof of this result is identical to that for the problem of finding the optimal stationary heat source. In the unconstrained case where $u_a = -\infty$ and $u_b = \infty$, one obtains that

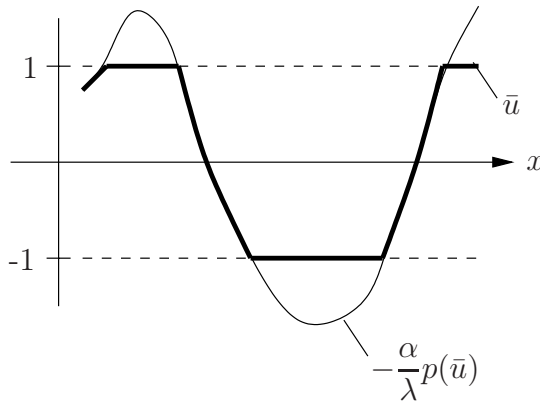
$$\bar{u}(x) = -\frac{1}{\lambda} \alpha(x)p(x).$$

In the special case of $\lambda = 0$, we have to distinguish between different cases as in (2.57) on page 70. As an illustration, we choose a two-dimensional domain Ω and imagine that its boundary Γ is unrolled onto part of the real axis. As bounds, we prescribe $u_a = -1$ and $u_b = +1$.



Optimal control for $\lambda = 0$.

For $\lambda > 0$, we obtain \bar{u} as the projection of the function $-\lambda^{-1}\alpha p$ onto $[-1, 1]$.



Optimal control for $\lambda > 0$.

2.8.5. A linear optimal control problem. Let us consider the linear problem with distributed control v and boundary control u :

$$\min J(y, u, v) := \int_{\Omega} (a_{\Omega} y + \lambda_{\Omega} v) dx + \int_{\Gamma} (a_{\Gamma} y + \lambda_{\Gamma} u) ds,$$

subject to

$\begin{aligned} -\Delta y &= \beta_{\Omega} v && \text{in } \Omega \\ \partial_{\nu} y + \alpha y &= \beta_{\Gamma} u && \text{on } \Gamma, \end{aligned}$

and

$$v_a(x) \leq v(x) \leq v_b(x) \quad \text{a.e. in } \Omega, \quad u_a(x) \leq u(x) \leq u_b(x) \quad \text{a.e. on } \Gamma.$$

We impose the following conditions on the data of this problem: the functions a_Ω , λ_Ω and a_Γ , λ_Γ are square integrable on their domains Ω and Γ , respectively, β_Ω and β_Γ are bounded and measurable on Ω and Γ , respectively, and the bounds v_a , v_b , u_a , and u_b are square integrable as well. In addition, α is nonnegative almost everywhere and does not vanish almost everywhere.

Then, the optimality conditions for an optimal triple $(\bar{y}, \bar{v}, \bar{u})$ read

$$\int_{\Omega} (\beta_{\Omega} p + \lambda_{\Omega})(v - \bar{v}) dx + \int_{\Gamma} (\beta_{\Gamma} p + \lambda_{\Gamma})(u - \bar{u}) ds \geq 0 \quad \forall v \in V_{ad}, \forall u \in U_{ad},$$

where the adjoint state p is given by

$$\begin{aligned} -\Delta p &= a_{\Omega} && \text{in } \Omega \\ \partial_{\nu} p + \alpha p &= a_{\Gamma} && \text{on } \Gamma. \end{aligned}$$

The reader will be asked to derive these relations in Exercise 2.15.

Linear control problems arise, for instance, if nonlinear optimal control problems are linearized at optimal points. By linearization and application of the necessary conditions to the linearized problem, optimality conditions for nonlinear problems can be derived. This is one possible way to treat nonlinear problems.

2.9. Construction of test examples

To validate numerical methods for the solution of optimal control problems, test examples are needed for which the exact solutions are known explicitly. By means of such test examples it can be checked whether a numerical method yields correct results. Invoking the necessary optimality conditions proved above, it is not hard to construct such examples. However, partial differential equations require a different approach than ordinary ones.

Indeed, in the optimal control theory of ordinary differential equations it is possible, at least for specifically chosen examples, to solve the state equation in closed form if an analytic expression is prescribed for the control. In the case of partial differential equations, this is much more difficult: even in the simplest cases the best we can hope for is to obtain a series expansion of the state y for a given u . Therefore, we take the opposite approach: we simply prescribe the desired solution triple (\bar{u}, \bar{y}, p) , and then adjust the state equation and the cost functional in such a way that \bar{u} , \bar{y} , and p satisfy the necessary optimality conditions.

2.9.1. Bang-bang control. By *bang-bang controls* we mean control functions whose values almost all lie on the boundary of the admissible set. Such controls occur in certain situations if the regularization parameter λ in front