Optimal Control of Partial Differential Equations

Theory, Methods and Applications

Fredi Tröltzsch

Graduate Studies in Mathematics

Volume 112

American Mathematical Society

Optimal Control of Partial Differential Equations

Theory, Methods and Applications

Optimal Control of Partial Differential Equations

Theory, Methods and Applications

Fredi Tröltzsch

Translated by Jürgen Sprekels

Graduate Studies
in Mathematics

Volume 112

American Mathematical Society
Providence, Rhode Island

Originally published in German by Friedr. Vieweg & Sohn Verlag, 65189 Wiesbaden, Germany, under the title: "Fredi Tröltzsch: Optimale Steuerung partieller Differentialgleichungen." 1. Auflage (1st edition). © Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH, Wiesbaden, 2005

Translated by Jürgen Sprekels

2000 *Mathematics Subject Classification*. Primary 49–01, 49K20, 35J65, 35K60, 90C48, 35B37.

For additional information and updates on this book, visit
www.ams.org/bookpages/gsm-112

Library of Congress Cataloging-in-Publication Data

Tröltzsch, Fredi, 1951–
[Optimale Steuerung partieller Differentialgleichungen. English]
Optimal control of partial differential equations : theory, methods and applications / Fredi Tröltzsch.
p. cm. — (Graduate studies in mathematics : v. 112)
Includes bibliographical references and index.
ISBN 978-0-8218-4904-0 (alk. paper)
1. Control theory. 2. Differential equations, Partial. 3. Mathematical optimization. I. Title.

QA402.3.T71913 2010
515′.642—dc22

2009037756

10 9 8 7 6 5 4 3 2 1 15 14 13 12 11 10

To my wife Silvia

Contents

Preface to the English edition

In addition to correcting some misprints and inaccuracies in the German edition, some parts of this book were revised and expanded. The sections dealing with gradient methods were shortened in order to make space for primal-dual active set strategies; the exposition of the latter now leads to the systems of linear equations to be solved. Following the suggestions of several readers, a derivation of the associated Green's functions is provided, using Fourier's method. Moreover, some references are discussed in greater detail, and some recent references on the numerical analysis of state-constrained problems have been added.

The sections marked with an asterisk may be skipped; their contents are not needed to understand the subsequent sections. Within the text, the reader will find formulas in framed boxes. Such formulas contain either results of special importance or the partial differential equations being studied in that section.

I am indebted to all readers who have pointed out misprints and supplied me with suggestions for improvements—in particular, Roland Herzog, Markus Müller, Hans Josef Pesch, Lothar v. Wolfersdorf, and Arnd Rösch. Thanks are also due to Uwe Prüfert for his assistance with the LaTeX typesetting. In the revision of the results on partial differential equations, I was supported by Eduardo Casas and Jens Griepentrog; I am very grateful for their cooperation. Special thanks are due to Jürgen Sprekels for his careful and competent translation of this textbook into English. His suggestions have left their mark in many places. Finally, I have to thank Mrs. Jutta Lohse for her careful proofreading of the English translation.

Berlin, July 2009

F. Tröltzsch

Preface to the German edition

The mathematical optimization of processes governed by partial differential equations has seen considerable progress in the past decade. Ever faster computational facilities and newly developed numerical techniques have opened the door to important practical applications in fields such as fluid flow, microelectronics, crystal growth, vascular surgery, and cardiac medicine, to name just a few. As a consequence, the communities of numerical analysts and optimizers have taken a growing interest in applying their methods to optimal control problems involving partial differential equations; at the same time, the demand from students for this expertise has increased, and there is a growing need for textbooks that provide an introduction to the fundamental concepts of the corresponding mathematical theory.

There are a number of monographs devoted to various aspects of the optimal control of partial differential equations. In particular, the comprehensive text by J. L. Lions [**Lio71**] covers much of the theory of linear equations and convex cost functionals. However, the interest in the class notes of my lectures held at the technical universities in Chemnitz and Berlin revealed a clear demand for an introductory textbook that also includes aspects of nonlinear optimization in function spaces.

The present book is intended to meet this demand. We focus on basic concepts and notions such as:

- Existence theory for linear and semilinear partial differential equations
- Existence of optimal controls

- Necessary optimality conditions and adjoint equations
- Second-order sufficient optimality conditions
- Foundation of numerical methods

In this connection, we will always impose constraints on the control functions, and sometimes also on the state of the system under study. In order to keep the exposition to a reasonable length, we will not address further important subjects such as controllability, Riccati equations, discretization, error estimates, and Hamilton–Jacobi–Bellman theory.

The first part of the textbook deals with convex problems involving quadratic cost functionals and linear elliptic or parabolic equations. While these results are rather standard and have been treated comprehensively in [**Lio71**], they are well suited to facilitating the transition to problems involving semilinear equations. In order to make the theory more accessible to readers having only minor knowledge of these fields, some basic notions from functional analysis and the theory of linear elliptic and parabolic partial differential equations will also be provided.

The focus of the exposition is on nonconvex problems involving semilinear equations. Their treatment requires new techniques from analysis, optimization, and numerical analysis, which to a large extent can presently be found only in original papers. In particular, fundamental results due to E. Casas and J.-P. Raymond concerning the boundedness and continuity of solutions to semilinear equations will be needed.

This textbook is mainly devoted to the analysis of the problems, although numerical techniques will also be addressed. Numerical methods could easily fill another book. Our exposition is confined to brief introductions to the basic ideas, in order to give the reader an impression of how the theory can be realized numerically. Much attention will be paid to revealing hidden mathematical difficulties that, as experience shows, are likely to be overlooked.

The material covered in this textbook will not fit within a one-term course, so the lecturer will have to select certain parts. One possible strategy is to confine oneself to elliptic theory (linear-quadratic and nonlinear), while neglecting the chapters on parabolic equations. This would amount to concentrating on Sections 1.2–1.4, 2.3–2.10, and 2.12 for linear-quadratic theory, and on Sections 4.1–4.6 and 4.8–4.10 for nonlinear theory. The chapters devoted to elliptic problems do not require results from parabolic theory as a prerequisite.

Alternatively, one could select the linear-quadratic elliptic theory and add Sections 3.3–3.7 on linear-quadratic parabolic theory. Further topics

can also be covered, provided that the students have a sufficient working knowledge of functional analysis and partial differential equations.

The sections marked with an asterisk may be skipped; their contents are not needed to understand the subsequent sections. Within the text, the reader will find formulas in framed boxes. Such formulas contain either results of special importance or the partial differential equations being studied in that section.

During the process of writing this book, I received much support from many colleagues. M. Hinze, P. Maaß, and L. v. Wolfersdorf read various chapters, in parts jointly with their students. W. Alt helped me with the typographical aspects of the exposition, and the first impetus to writing this textbook came from T. Grund, who put my class notes into a first LaTeX version. My colleagues C. Meyer, U. Prüfert, T. Slawig, and D. Wachsmuth in Berlin, and my students I. Neitzel and I. Yousept, proofread the final version. I am indebted to all of them. I also thank Mrs. U. Schmickler-Hirzebruch and Mrs. P. Rußkamp of Vieweg-Verlag for their very constructive cooperation during the preparation and implementation of this book project.

Berlin, April 2005

F. Tröltzsch

Chapter 1

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

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

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, $y(T) = y_T$, $y'(T) = 0$, $|u| \le 1$) 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:

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

subject to the *state equation*

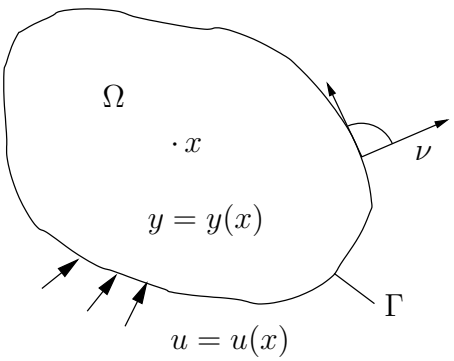
$$\boxed{\begin{array}{rcll} -\Delta y & = & 0 & \text{in } \Omega \\ \dfrac{\partial y}{\partial \nu} & = & \alpha\,(u-y) & \text{on } \Gamma \end{array}}$$

and the *pointwise control constraints*

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

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:

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

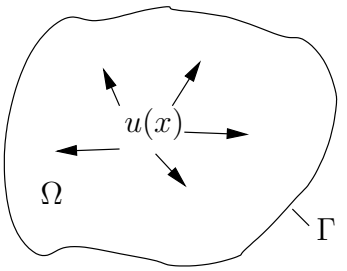
subject to

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

and

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

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

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

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:

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

subject to

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

and

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

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

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

subject to

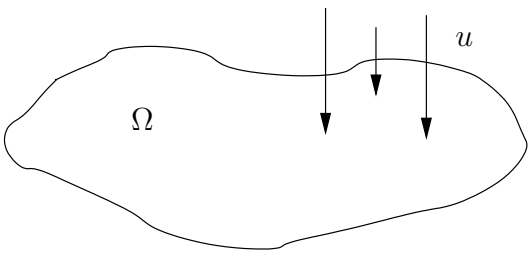
$$\begin{aligned} y_{tt} - \Delta y &= u && \text{in } Q \\ y(0) &= y_0 && \text{in } \Omega \\ y_t(0) &= y_1 && \text{in } \Omega \\ y &= 0 && \text{on } \Sigma \end{aligned}$$

and

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

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)

$$(u\cdot\nabla)\,u = u_1\,D_1u + u_2D_2u + u_3D_3u = \sum_{i=1}^{3} u_i \begin{bmatrix} D_iu_1 \\ D_iu_2 \\ D_iu_3 \end{bmatrix}.$$

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 $J = J(y,u)$, $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, 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*

$$\boxed{\begin{array}{ll} \min J(y,u) & \\ A\,y = B\,u, & u \in U_{ad}. \end{array}} \tag{1.1}$$

We seek vectors y and u minimizing the cost functional J subject to the constraints $A\,y = B\,u$ 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

$$y = A^{-1}B\,u, \tag{1.2}$$

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

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

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

$$J(y,u) = J(S\,u,u) =: f(u).$$

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

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

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 $f(\bar{u}) \le f(u)$ 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 $f(\bar{u}) = \min\limits_{u \in U_{ad}} f(u)$. □

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 $f : \mathbb{R}^m \to \mathbb{R}$:

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

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

$x_1, ..., x_m$, 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,

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

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,

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

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

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

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

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

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

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

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

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

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 $A\,y = b$. 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

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

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

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

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*

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

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:

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

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

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

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

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

The optimality system becomes

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

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

$$\bar{u} = -\frac{1}{\lambda}\,B^\top \bar{p}. \tag{1.10}$$

Substitution in the two other relations yields the optimality system

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

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). ◇

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

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

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*

$$L : \mathbb{R}^{2n+m} \to \mathbb{R}, \quad L(y,u,p) := J(y,u) - \big(A\,y - B\,u\,,\,p\big)_{\mathbb{R}^n},$$

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

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

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:

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

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 $L(y,u,p) := J(y,u) - \big(T(y,u)\,,\,p\big)_{\mathbb{R}^n}$.

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.,

$$U_{ad} = \big\{u \in \mathbb{R}^m \,:\, u_a \le u \le u_b\big\}. \tag{1.13}$$

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

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

we find that $\bar u$ solves the linear optimization problem

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

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

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

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

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

No direct information can be recovered from the variational inequality for the components that satisfy $\big(B^\top \bar p + \nabla_u J(\bar y, \bar u)\big)_i = 0$. 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

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

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

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

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 $\big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i \leq 0$. 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 $\big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i > 0$, 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

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

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

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

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 by 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:*

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

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

$$A\,y - B\,u = 0, \qquad u_a \le u \le u_b,$$

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

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

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

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

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 $(y, u) \ne (0, 0)$ with $A\,y = B\,u$ 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

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

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 $A\,y = B\,u$ 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.

Chapter 2

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 $x = (x_1, \ldots, x_n)^\top$ or functions $x : [a, b] \to \mathbb{R}$, 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 $\|\cdot\| : X \to \mathbb{R}$ 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)	$\|\lambda\, x\| = \lvert\lambda\rvert\, \|x\|$	(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*

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

The space of continuous real-valued mappings $x : [a,b] \to \mathbb{R}$ 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,

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

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

(ii) *We call x the* limit *of the sequence, written as* $\lim_{n\to\infty} x_n = x$.

(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* $\|x_n - x_m\| \le \varepsilon$ 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

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

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

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

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

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 $\left\{C_{L^2}[a,b], \|\cdot\|_{C_{L^2}[a,b]}\right\}$ 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* $(\cdot\,,\cdot) : H \to \mathbb{R}$ *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) $(u\,,u) \ge 0$, *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* $\left\{H, (\cdot\,,\cdot)\right\}$ *is called a* pre-Hilbert space.

Remark. Again, we speak of the pre-Hilbert space H instead of $\left\{H, (\cdot\,,\cdot)\right\}$ 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 $\left\{H, (\cdot\,,\cdot)\right\}$ 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*:

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

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* $L^p(E)$, $1 \le p < \infty$, *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

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

$L^p(E)$, 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*

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

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 $y : [0,1] \to \mathbb{R}$ 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, $v : \Omega \to \mathbb{R}$ 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* 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 $\alpha = (\alpha_1, \ldots, \alpha_N)$ having nonnegative integer components. The number $|\alpha| = \alpha_1 + \ldots + \alpha_N$ 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 $D_1^{\alpha_1} \ldots D_N^{\alpha_N}\, y(x)$, 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:

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

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 $S_1, \ldots, S_M$, functions $h_1, \ldots, h_M$, 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*
$$\bar{Q}_{N-1} = \big\{ y = (y_1, ..., y_{N-1}) \,:\, |y_i| \leq a,\ i = 1 \ldots N-1 \big\},$$
and the partial derivatives of order k are Lipschitz continuous on $\bar{Q}_{N-1}$.

(ii) *For any $P \in \Gamma$ there is some $i \in \{1, \ldots, M\}$ 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
$$|E| = \int_D \sqrt{1 + |\nabla h_i(y_1, \ldots, y_{N-1})|^2}\ dy_1 \ldots dy_{N-1};$$
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 ith 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 $|\alpha| \le k$ 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 $L^1_{loc}(\Omega)$ 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 $y \in L^1_{loc}(\Omega)$ and some multi-index α be given. If a function $w \in L^1_{loc}(\Omega)$ satisfies*

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

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. ◇

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

Definition. *Let $1 \le p < \infty$ 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 $|\alpha| \le k$, endowed with the norm*

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

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

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

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

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

and the norm is given by

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

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

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

$H^1(\Omega)$ 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 $\big\{X, \|\cdot\|\big\}$, which is by definition the set

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

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 $1 \le p \le \infty$. Then there exists a linear and continuous mapping $\tau : W^{1,p}(\Omega) \to L^p(\Gamma)$ such that for all $y \in W^{1,p}(\Omega) \cap C(\bar{\Omega})$ we have $\big(\tau\, y\big)(x) = y(x)$ for all $x \in \Gamma$.*

In particular, for $p = 2$ it follows that $\tau : H^1(\Omega) \to L^2(\Gamma)$. In the case of continuous functions, $\tau\, y$ 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 $\tau\, y$ 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 $\tau\, y$. In this sense, $y|_{\Gamma_0}$ is, for measurable subsets $\Gamma_0 \subset \Gamma$, defined as the restriction of $\tau\, y$ to Γ_0.

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

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

Moreover, for bounded Lipschitz domains Ω it follows that

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

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

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

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

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

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 $\Omega \subset \mathbb{R}^N,\ N \ge 2$, is a bounded Lipschitz domain with boundary Γ.

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

$$\boxed{\begin{array}{rcll} -\Delta y & = & f & \text{in } \Omega \\ y & = & 0 & \text{on } \Gamma, \end{array}} \tag{2.2}$$

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 $y \in C^2(\Omega) \cap C(\bar{\Omega})$ 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

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

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*

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

$$a[y,v] := \int_\Omega \nabla y \cdot \nabla v \; dx. \tag{2.4}$$

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

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

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

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

Then (2.3) attains the general form

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

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 $a : V \times V \to \mathbb{R}$ 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, $y \in V$:*

$$\big|a[y,v]\big| \;\le\; \alpha_0 \,\|y\|_V \|v\|_V \quad (\textit{boundedness}) \tag{2.6}$$

$$a[y,y] \;\ge\; \beta_0 \,\|y\|_V^2 \quad (\textit{V-ellipticity}). \tag{2.7}$$

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*

$$\|y\|_V \le c_a \,\|F\|_{V^*}. \tag{2.8}$$

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

$$\int_\Omega |y|^2\,dx \le c(\Omega) \int_\Omega |\nabla y|^2\,dx \quad \forall\, y \in H_0^1(\Omega).$$

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*

$$\|y\|_{H^1(\Omega)} \le c_P\, \|f\|_{L^2(\Omega)}. \tag{2.9}$$

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:

$$\begin{aligned}
\Big|\int_\Omega \nabla y \cdot \nabla v\; dx\Big| &\le \Big(\int_\Omega |\nabla y|^2\,dx\Big)^{1/2} \Big(\int_\Omega |\nabla v|^2\,dx\Big)^{1/2} \\
&\le \Big(\int_\Omega \big(|y|^2 + |\nabla y|^2\big)\; dx\Big)^{1/2} \Big(\int_\Omega \big(|v|^2 + |\nabla v|^2\big)\; dx\Big)^{1/2} \\
&\le \|y\|_{H^1(\Omega)}\; \|v\|_{H^1(\Omega)}.
\end{aligned}$$

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

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

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

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

so that $\|F\|_{V^*} \le \|f\|_{L^2(\Omega)}$. 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 $\|y\|_{H^1(\Omega)} \le c_a \|F\|_{V^*} \le c_a \|f\|_{L^2(\Omega)}$, which proves (2.9). □

Remarks.
(i) Inequality (2.10) shows that

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

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

$$\boxed{\begin{aligned} -\Delta y + c_0\, y &= f && \text{in } \Omega \\ \partial_\nu y + \alpha\, y &= g && \text{on } \Gamma. \end{aligned}} \tag{2.11}$$

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

$$\int_\Omega \nabla y \cdot \nabla v\; dx + \int_\Omega c_0\, y\, v\; dx + \int_\Gamma \alpha\, y\, v\, ds = \int_\Omega f\, v\; dx + \int_\Gamma g\, v\; ds \tag{2.12}$$

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

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

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*

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

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 $y \in H^1_0(\Omega)$. 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*

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

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*

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

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

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

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

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

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 $c_0(x) \ge \delta$ for all $x \in E$. Hence, invoking (2.15) and the inequality $\big(\int_E y \, dx\big)^2 \le |E| \int_E y^2 \, dx$, we find that

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

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 $\alpha(x) \ge \delta$ for all $x \in \Gamma_1$. In view of (2.14), similar reasoning yields that in this case,

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

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

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

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

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

$$\boxed{\begin{aligned} \mathcal{A}\,y + c_0\,y &= f && \text{in } \Omega \\ \partial_{\nu_\mathcal{A}} y + \alpha\,y &= g && \text{on } \Gamma_1 \\ y &= 0 && \text{on } \Gamma_0. \end{aligned}} \tag{2.18}$$

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

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

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,

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

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

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

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

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

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

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

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

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

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* $\displaystyle\int_\Omega (c_0(x))^2\, dx + \int_\Gamma (\alpha(x))^2\, ds(x) > 0$

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 $c_{\mathcal{A}} > 0$, which depends on neither f nor g, such that

$$\|y\|_{H^1(\Omega)} \leq c_{\mathcal{A}} \big(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_1)} \big) \qquad \forall f \in L^2(\Omega),\ \forall g \in L^2(\Gamma_1). \tag{2.23}$$

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 $\|y\| \le c\,(\|f\| + \|g\|)$, 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

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

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 $p \le 2N/(N-2)$ if $N = \dim\,\Omega > 2$. Owing to Hölder's inequality, we have

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

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 $\tau\, y$ 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

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

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 $A : U \to V$ is* linear *or a* linear operator *if $A(u+v) = A\,u + A\,v$ and $A\,(\lambda\, v) = \lambda\, A\, v$ for all u, $v \in U$ and $\lambda \in \mathbb{R}$. A linear mapping $f : U \to \mathbb{R}$ is called a* linear functional.

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

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

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

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

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

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

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

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

and, therefore,

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

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

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

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

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 $\|A\,u\|_V \leq c\,\|u\|_U$ 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 $A : U \to V$ given by

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

A is bounded, since

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

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 $A : L^\infty(0,1) \to L^\infty(0,1)$,

$$\big(A\,u\big)(x) = x^2\,u(x).$$

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

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 $f(u) = u\big(\frac{1}{2}\big)$ on $U = C[0,1]$. Since

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

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

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

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

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*

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

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 $F_u : U^* \to \mathbb{R}$ induced by u,

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

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

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

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*,

$$\int_E |f(x)|\, |u(x)|\, dx \le \Big(\int_E |f(x)|^q\, dx \Big)^{\frac{1}{q}} \Big(\int_E |u(x)|^p\, dx \Big)^{\frac{1}{p}}. \tag{2.25}$$

◇

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*

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

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

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.,

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

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

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

Moreover, if $u_n \rightharpoonup u$ and $v_n \to v$ (strong convergence), then $\big(v_n\,,\,u_n\big) \to \big(v\,,\,u\big)$ as $n \to \infty$; 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

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

defines the nth 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 $n \to \infty$, that is,

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

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

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

On the other hand, we have

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

$\diamond$

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 $F : U \to V$ 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,*

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

Examples.

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

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

(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 $u_n(x) =$

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

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

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. ◇

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 $\{\sin(n\,x)/\sqrt{\pi}\}$ 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 $f : C \to \mathbb{R}$ 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 $f : U \to \mathbb{R}$ 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 $n \to \infty$ we have*

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

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

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

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

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

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

subject to the constraints

$$\boxed{\begin{array}{rcll} -\Delta y & = & \beta\, u & \text{in } \Omega \\ y & = & 0 & \text{on } \Gamma \end{array}} \tag{2.27}$$

and

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

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 $u_a \le u \le u_b$ 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

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

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 $y \in H^1_0(\Omega)$ to the boundary value problem (2.27), called *the state associated with* u. The space

$$Y := H^1_0(\Omega)$$

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*

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

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 $G : L^2(\Omega) \to H_0^1(\Omega)$, $u \mapsto y(u)$, 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 $\|y\|_{L^2(\Omega)} \le \|y\|_{H^1(\Omega)}$, 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 $E_Y : H^1(\Omega) \to L^2(\Omega)$ 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

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

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 $L^2(\Omega)$:

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

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 $\lambda \ge 0$ be given. Moreover, let $S : U \to H$ be a continuous linear operator. Then the quadratic Hilbert space optimization*

problem

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

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 $f(u_n) \to j$ as $n \to \infty$. 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,

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

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

$$f(\bar{u}) \leq \lim_{k\to\infty} \inf f(u_{n_k}) = j.$$

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. □

Remark. The proof only made use of the fact that f is continuous and convex. The existence result thus holds for *any* functional $f : U \to \mathbb{R}$ 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 $U_{ad} = \{u \in L^2(\Omega) : u_a \le u \le u_b \text{ a.e. in } \Omega\}$ 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. □

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 $G : L^2(\Omega) \to H_0^1(\Omega)$ 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

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

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

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 nonnegative 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, $S : L^2(\Omega) \to L^2(\Omega)$, $S = E_Y G$, so that

$$y = S\, u + y_0.$$

The problem then attains the form

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

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

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

subject to the constraints

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

and

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

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

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

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

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

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 $G : L^2(\Gamma) \to H^1(\Omega)$, $u \mapsto y(u)$, 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 $S : L^2(\Gamma) \to L^2(\Omega)$. 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

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

subject to the constraints

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

and

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

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

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

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 $S = E_Y G$, $S : L^2(\Gamma_1) \times L^2(\Omega) \to L^2(\Omega)$. 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

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

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

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

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*

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

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 $f : \mathbb{R}^2 \to \mathbb{R}$, 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 $A : U \to V$ such that*

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

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 $f : \mathcal{U} \to \mathbb{R}$ is Gâteaux diffentiable 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 $f : \mathcal{U} \to \mathbb{R}$ by

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

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

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

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

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

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

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

We have

$$\begin{aligned}\lim_{t\to 0}\frac{1}{t}\big(f(u+t\,h)-f(u)\big) &= \lim_{t\to 0}\frac{1}{t}\Big(\|u+t\,h\|_H^2-\|u\|_H^2\Big)\\ &= \lim_{t\to 0}\frac{2\,t\,\big(u\,,\,h\big)_H+t^2\,\|h\|_H^2}{t}\\ &= 2\,\big(u\,,\,h\big)_H,\end{aligned}$$

and therefore

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

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

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

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 $f'(u)\,h=\big(2\,u\,,\,h\big)_H$, and the gradient $f'(u)=2\,u$.

(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

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

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

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

Definition. *A mapping $F:\mathcal{U}\subset U\to V$ 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 $r(u,\cdot):U\to V$ with the following properties: for all $h\in U$ such that $u+h\in\mathcal{U}$, we have*

$$F(u+h)=F(u)+A\,h+r(u,h),$$

where the so-called remainder r *satisfies the condition*

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

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

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

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

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 $f : \mathbb{R}^2 \to \mathbb{R}$,

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

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 $F : \mathcal{U} \to \mathcal{V}$ and $G : \mathcal{V} \to Z$ be Fréchet differentiable at $u \in \mathcal{U}$ and at $F(u) \in \mathcal{V}$, respectively. Then the composition $E = G \circ F : \mathcal{U} \to Z$, defined by $E(u) = G\big(F(u)\big)$, is Fréchet differentiable at u, and*

$$E'(u) = G'\big(F(u)\big)\, F'(u).$$

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

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

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

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

The chain rule thus yields

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

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

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

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

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

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

$$g(u) = f(A\,u).$$

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

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

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

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

Definition. *The mapping $A^* : V^* \to U^*$ 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

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

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

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 $A^\star$; note the typographical difference between $A^\star$ 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 $\|A^*\|_{\mathcal{L}(V^*,U^*)} \le \|A\|_{\mathcal{L}(U,V)}$. 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

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

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* $A^\star$ *is called the* Hilbert space adjoint *or* adjoint *of* A *if*

$$\big(v\,,\,A\,u\big)_V = \big(A^\star v\,,\,u\big)_U \quad \forall u \in U, \quad \forall v \in V. \tag{2.42}$$

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 $A : \mathbb{R}^n \to \mathbb{R}^m$ denote a linear operator, which is represented by an $m \times n$ matrix also denoted by A. Since $(v\,,\,A\,u)_{\mathbb{R}^m} = (A^\top v\,,\,u)_{\mathbb{R}^n}$ 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

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

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:

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

where the adjoint operator has the representation

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

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

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

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

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

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, $f(u + t\,(u-\bar u)) \ge f(\bar u)$ and hence also

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

Letting $t \downarrow 0$, we arrive at $f'(\bar u)(u-\bar u) \ge 0$, 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. □

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 $C_\varepsilon = \big\{u \in L^2(a,b) : u(x) \geq \varepsilon \text{ for a.e. } x \in (a,b)\big\}$. The functional

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

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. ◇

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 $S : U \to H$ 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*

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

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

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

The assertion is thus a direct consequence of Lemma 2.21. □

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

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

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

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

subject to

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

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

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

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*

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

Then

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

Proof: We invoke the variational formulations of the above boundary value problems. For y, insertion of the test function $p \in H^1_0(\Omega)$ yields

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

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

$$\int_\Omega \Big(\nabla p \cdot \nabla y + c_0\, p\, y \Big)\, dx = \int_\Omega z\, y\, dx.$$

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

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

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

Invoking Lemma 2.23 with $c_0 = 0$ and $y = S\,u$, we find that

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

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 $S^*\, z = \beta\, p$. □

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 $G : L^2(\Omega) \to H_0^1(\Omega)$ instead of S, then (after identifying $L^2(\Omega)^*$ with $L^2(\Omega)$) the adjoint operator $G^* : H_0^1(\Omega)^* \to L^2(\Omega)$ would have occurred. We have avoided the space $H_0^1(\Omega)^*$ by choosing $S :\, L^2(\Omega) \to L^2(\Omega)$. 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

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

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),

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

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*

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

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*:

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

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

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

hence

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

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* $x \in \Omega$,

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

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

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

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

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

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 $A_+(\bar{u})$,

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

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

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

Integration yields that (2.51) holds. □

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

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

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

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

or the minimum principle

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

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

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

is satisfied, that is, if

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

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

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

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

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: $\boldsymbol{\lambda > 0}$. 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

$$\bar u(x) = \mathbb{P}_{[u_a(x),u_b(x)]}\Big\{-\frac{1}{\lambda}\beta(x)\,p(x)\Big\} \quad \textit{for almost every } x \in \Omega. \tag{2.58}$$

Here, for real numbers $a \le b$ $\mathbb{P}_{[a,b]}$ denotes the projection of $\mathbb{R}$ onto $[a,b]$,

$$\mathbb{P}_{[a,b]}(u) := \min\big\{b\,,\,\max\{a\,,\,u\}\big\}.$$

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. □

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

$$\bar{u} = -\frac{1}{\lambda}\,\beta\,p. \tag{2.59}$$

Putting this in the state equation leads to the optimality system

$$\boxed{\begin{array}{rclcrcl} -\Delta y &=& -\lambda^{-1}\,\beta^2\,p & \qquad & -\Delta p &=& y - y_\Omega \\ y|_\Gamma &=& 0 & & p|_\Gamma &=& 0. \end{array}}$$

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*

$$\beta\,p + \lambda\bar{u} - \mu_a + \mu_b = 0 \tag{2.60}$$

as well as the complementarity conditions

$$\mu_a(x)\,\big(u_a(x) - \bar{u}(x)\big) = \mu_b(x)\,\big(\bar{u}(x) - u_b(x)\big) = 0 \quad \text{for a.e. } x \in \Omega. \tag{2.61}$$

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

$$\begin{aligned} \mu_a(x) &:= \big(\beta(x)\,p(x) + \lambda\bar{u}(x)\big)_+, \\ \mu_b(x) &:= \big(\beta(x)\,p(x) + \lambda\bar{u}(x)\big)_-. \end{aligned} \tag{2.62}$$

Here, we use the usual definitions of s_+ and s_- for $s \in \mathbb{R}$, namely

$$s_+ = \frac{1}{2}\big(s + |s|\big), \qquad s_- = \frac{1}{2}\big(|s| - s\big).$$

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 $x \in \Omega$:

$$\begin{aligned} (\beta\,p + \lambda\,\bar{u})(x) > 0 \quad &\Rightarrow \quad \bar{u}(x) = u_a(x) \\ (\beta\,p + \lambda\,\bar{u})(x) < 0 \quad &\Rightarrow \quad \bar{u}(x) = u_b(x) \\ u_a(x) < \bar{u}(x) < u_b(x) \quad &\Rightarrow \quad (\beta\,p + \lambda\,\bar{u})(x) = 0. \end{aligned}$$

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

$$\big(\beta(x)\,p(x) + \lambda\,\bar{u}(x)\big)\,\big(u(x) - \bar{u}(x)\big) \geq 0. \tag{2.63}$$

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. □

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*:

$$
\boxed{
\begin{array}{rclcrcl}
-\Delta y & = & \beta\, u & & -\Delta p & = & y - y_\Omega \\
y|_\Gamma & = & 0 & & p|_\Gamma & = & 0 \\
\multicolumn{7}{c}{\beta\, p + \lambda u - \mu_a + \mu_b = 0} \\
\multicolumn{7}{c}{u_a \le u \le u_b, \quad \mu_a \ge 0, \quad \mu_b \ge 0} \\
\multicolumn{7}{c}{\mu_a \big(u_a - u\big) = \mu_b \big(u - u_b\big) = 0.}
\end{array}
} \tag{2.64}
$$

Here, the relations in the last three lines hold for almost every $x \in \Omega$.

Definition. *The functions μ_a, $\mu_b \in L^2(\Omega)$ defined in Theorem 2.29 are called* Lagrange multipliers *associated with the inequality constraints $u_a \le u$ and $u \le u_b$, respectively.*

Remark. The system (2.64) can be derived directly by using a Lagrangian function provided that the existence of multipliers μ_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 $f(u) = J\big(y(u), u\big)$, 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*

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

is given by

$$f'(u) = \beta\, p + \lambda\, u,$$

where $p \in H^1_0(\Omega)$ denotes the weak solution to the adjoint equation

$$
\begin{array}{rcll}
-\Delta p & = & y - y_\Omega & \text{in} \quad \Omega \\
p & = & 0 & \text{on} \quad \Gamma
\end{array} \tag{2.65}
$$

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

$$f'(u)\,h = \big(S^*(S\,u - y_\Omega) + \lambda\,u\,,\,h\big)_{L^2(\Omega)} = \big(\beta\,p + \lambda\,u\,,\,h\big)_{L^2(\Omega)}.$$

By virtue of the Riesz representation theorem, $f'(u)$ is identified with $\beta\,p + \lambda\,u$. □

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

$$\big(S\,\bar u - y_\Omega\,,\,Su - S\bar u\big)_{L^2(\Omega)} + \lambda\,\big(\bar u\,,\,u - \bar u\big)_{L^2(\Omega)} \geq 0 \quad \forall u \in U_{ad}. \tag{2.66}$$

With $\bar y = S\bar u$ and $y = Su$, it follows that

$$f'(\bar u)(u - \bar u) = \big(\bar y - y_\Omega\,,\,y - \bar y\big)_{L^2(\Omega)} + \lambda\,\big(\bar u\,,\,u - \bar u\big)_{L^2(\Omega)} \geq 0. \tag{2.67}$$

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*

$$\begin{aligned} -\Delta y + c_0\,y &= \beta_\Omega\,v & \qquad -\Delta p + c_0\,p &= a_\Omega \\ \partial_\nu\,y + \alpha\,y &= \beta_\Gamma\,u & \qquad \partial_\nu\,p + \alpha\,p &= a_\Gamma. \end{aligned}$$

Then

$$\int_\Omega a_\Omega\,y\,dx + \int_\Gamma a_\Gamma\,y\,ds = \int_\Omega \beta_\Omega\,p\,v\,dx + \int_\Gamma \beta_\Gamma\,p\,u\,ds. \tag{2.68}$$

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

$$\int_\Omega \big(\nabla y \cdot \nabla p + c_0\,y\,p\big)\,dx + \int_\Gamma \alpha\,y\,p\,ds = \int_\Omega \beta_\Omega\,p\,v\,dx + \int_\Gamma \beta_\Gamma\,p\,u\,ds,$$

and insertion of $y \in H^1(\Omega)$ in the equation for p yields

$$\int_\Omega \big(\nabla p \cdot \nabla y + c_0\, p\, y\big)\, dx + \int_\Gamma \alpha\, p\, y\, ds = \int_\Omega a_\Omega\, y\, dx + \int_\Gamma a_\Gamma\, y\, ds.$$

From this, the assertion immediately follows. □

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:

$$\min\ J(y,u) := \frac{\lambda_\Omega}{2}\,\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda_\Gamma}{2}\,\|y - y_\Gamma\|^2_{L^2(\Gamma)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}, \tag{2.69}$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y &=& \beta\, u & \text{in } \Omega \\ \partial_\nu y + \alpha\, y &=& 0 & \text{on } \Gamma \end{array}} \tag{2.70}$$

and

$$u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. in } \Omega. \tag{2.71}$$

We postulate that $\lambda \ge 0$, $\lambda_\Omega \ge 0$, $\lambda_\Gamma \ge 0$ and $\alpha \in L^\infty(\Gamma)$, where $\alpha \ge 0$ almost everywhere and $\|\alpha\|_{L^\infty(\Gamma)} \ne 0$, 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

$$\int_\Omega (\beta\, p + \lambda\, \bar u)(u - \bar u)\, dx \ge 0 \qquad \forall u \in U_{ad},$$

where the adjoint state p solves the boundary value problem

$$\boxed{\begin{array}{rcll} -\Delta p &=& \lambda_\Omega\,(\bar y - y_\Omega) & \text{in } \Omega \\ \partial_\nu p + \alpha\, p &=& \lambda_\Gamma\,(\bar y - y_\Gamma) & \text{on } \Gamma. \end{array}}$$

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:

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

subject to

$$\boxed{\begin{aligned} -\Delta y &= 0 && \text{in } \Omega \\ \partial_\nu y + \alpha\, y &= \alpha\, u && \text{on } \Gamma \end{aligned}}$$

and

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

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, $S = E_Y G : L^2(\Gamma) \to L^2(\Omega)$, with the embedding operator $E_Y : H^1(\Omega) \to L^2(\Omega)$. The cost functional then attains the form

$$J(y,u) = f(u) = \frac{1}{2}\,\|Su - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Gamma)}.$$

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

$$f'(\bar u)(u - \bar u) = \big(\bar y - y_\Omega\,,\, y - \bar y\big)_{L^2(\Omega)} + \lambda\,\big(\bar u\,,\, u - \bar u\big)_{L^2(\Gamma)} \ge 0. \tag{2.72}$$

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:

$$\boxed{\begin{aligned} -\Delta p &= \bar y - y_\Omega && \text{in } \Omega \\ \partial_\nu p + \alpha\, p &= 0 && \text{on } \Gamma. \end{aligned}} \tag{2.73}$$

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

$$\int_\Omega \nabla p \cdot \nabla v \, dx + \int_\Gamma \alpha \, p \, v \, ds = \int_\Omega (\bar{y} - y_\Omega) \, v \, dx \quad \forall \, v \in H^1(\Omega). \tag{2.74}$$

The optimal state $\bar{y} = S\,\bar{u}$ is the weak solution to the state equation associated with $\bar{u}$, while $y = S\,u$ 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

$$f'(\bar{u})(u - \bar{u}) = \int_\Gamma (\lambda \, \bar{u} + \alpha \, p)(u - \bar{u}) \, ds \geq 0 \qquad \forall \, u \in U_{ad}.$$

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

$$f'(u) = \alpha \, p|_\Gamma + \lambda \, u, \tag{2.75}$$

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*

$$\int_\Gamma \big(\alpha(x) \, p(x) + \lambda \, \bar{u}(x)\big)\big(u(x) - \bar{u}(x)\big) \, ds(x) \geq 0 \qquad \forall \, u \in U_{ad} \tag{2.76}$$

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

$$\bar{u}(x) = \begin{cases} u_a(x) & \text{if} \quad \alpha(x)\,p(x) + \lambda\,\bar{u}(x) > 0 \\ \in [u_a(x), u_b(x)] & \text{if} \quad \alpha(x)\,p(x) + \lambda\,\bar{u}(x) = 0 \\ u_b(x) & \text{if} \quad \alpha(x)\,p(x) + \lambda\,\bar{u}(x) < 0, \end{cases} \tag{2.77}$$

and the *weak minimum principle* becomes

$$\min_{u_a(x) \le v \le u_b(x)} \Big\{ \big(\alpha(x)\,p(x) + \lambda\,\bar{u}(x)\big)\,v \Big\} = \big(\alpha(x)\,p(x) + \lambda\,\bar{u}(x)\big)\,\bar{u}(x)$$

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*

$$\min_{u_a(x) \le v \le u_b(x)} \left\{ \alpha(x)\,p(x)\,v + \frac{\lambda}{2} v^2 \right\}$$

is attained at $v = \bar{u}(x)$. Hence, for $\lambda > 0$ we have for almost every $x \in \Gamma$ the projection formula

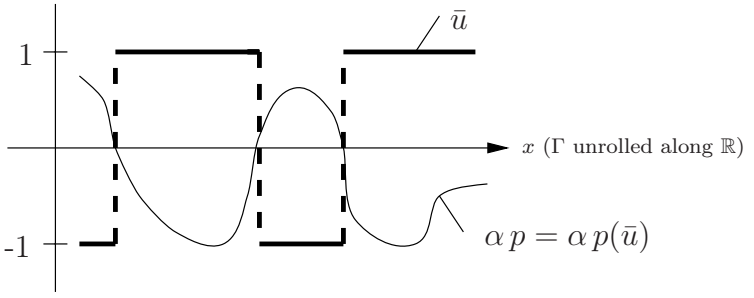
$$\bar{u}(x) = I\!P_{[u_a(x), u_b(x)]} \left\{ -\frac{1}{\lambda} \alpha(x)\,p(x) \right\}. \tag{2.78}$$

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

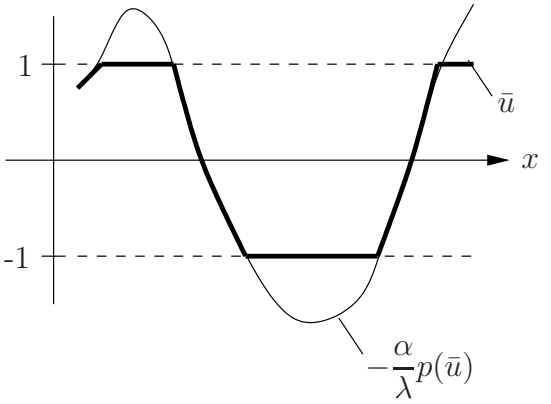
$$\bar{u}(x) = -\frac{1}{\lambda}\,\alpha(x)\,p(x).$$

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

$$\boxed{\begin{array}{rcll} -\Delta y & = & \beta_\Omega\, v & \text{in } \Omega \\ \partial_\nu y + \alpha\, y & = & \beta_\Gamma\, u & \text{on } \Gamma, \end{array}}$$

and

$$v_a(x) \le v(x) \le v_b(x) \quad \text{a.e. in } \Omega, \qquad u_a(x) \le u(x) \le u_b(x) \quad \text{a.e. on } \Gamma.$$

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

of $\|u\|^2$ vanishes. In the following, we will construct a case in which such a control appears for the problem

$$\min \int_\Omega |y - y_\Omega|^2 \, dx,$$

subject to

$$\begin{aligned} -\Delta y &= u + e_\Omega \\ y|_\Gamma &= 0 \end{aligned}$$

and

$$-1 \le u(x) \le 1,$$

where e_Ω is yet to be defined.

This problem differs from that of the optimal stationary heat source investigated in Section 2.8.2 only by the term e_Ω in the state equation. It is, however, easily seen that this term influences neither the adjoint equation (2.50) nor the variational inequality (2.51); see Exercise 2.16.

As the domain, we choose again the unit square $\Omega = (0,1)^2$. We look for a *chessboard function* $\bar{u}$ as optimal control. To this end, we subdivide the unit square like a chessboard into $8 \times 8 = 64$ congruent subsquares. Within these subsquares, the optimal control $\bar{u}$ shall alternately attain the values $+1$ or -1.

In view of the necessary optimality condition (2.57), and since $u_a = -1$ and $u_b = 1$, $\bar{u}$ is optimal if and only if

$$\bar{u}(x) = -\operatorname{sign} p(x) \qquad \text{for a.e. } x \in \Omega,$$

where we have put $\operatorname{sign}(0) := [-1,1]$ to express that $\bar{u}$ can vary here arbitrarily in $[-1,1]$. An adjoint state that fits the chessboard pattern is given by

$$p(x) = p(x_1, x_2) = -\frac{1}{128\,\pi^2} \sin(8\,\pi\,x_1) \sin(8\,\pi\,x_2).$$

The factor $1/(128\,\pi^2)$ simplifies other expressions. The associated control $\bar{u}$ has value $+1$ in the lower left subsquare of Ω and changes sign according to the chessboard pattern. Next, we choose as optimal state the function

$$\bar{y}(x) = \sin(\pi\,x_1) \sin(\pi\,x_2).$$

Note that $\bar{y}$ vanishes on Γ; moreover, it solves the Poisson equation

$$-\Delta \bar{y} = 2\pi^2\,\bar{y} = 2\pi^2 \sin(\pi\,x_1) \sin(\pi\,x_2).$$

For the state equation to be solved by $\bar{y}$, we must have $-\Delta\bar{y} = \bar{u} + e_\Omega$, that is,

$$e_\Omega = -\Delta\bar{y} - \bar{u} = 2\pi^2 \sin(\pi\,x_1) \sin(\pi\,x_2) + \operatorname{sign}\big(-\sin(8\,\pi\,x_1) \sin(8\,\pi\,x_2)\big).$$

Correspondingly, the adjoint state satisfies

$$\Delta p(x) = 2\,(8\,\pi)^2\,\sin(8\,\pi\,x_1)\sin(8\,\pi\,x_2)\frac{1}{128\,\pi^2} = \sin(8\,\pi\,x_1)\sin(8\,\pi\,x_2),$$

and it has to be a solution to the boundary value problem

$$\begin{aligned} -\Delta p &= \bar{y} - y_\Omega \\ p|_\Gamma &= 0. \end{aligned}$$

Therefore, we choose $y_\Omega = \bar{y} + \Delta p$, that is,

$$y_\Omega(x) = \sin(\pi\,x_1)\sin(\pi\,x_2) + \sin(8\,\pi\,x_1)\sin(8\,\pi\,x_2).$$

2.9.2. Distributed control and Neumann boundary condition. In this section, we consider a problem with homogeneous *Neumann* boundary condition $\partial_\nu y = 0$, namely:

$$\min\, J(y,u) := \frac{1}{2}\int_\Omega |y - y_\Omega|^2\,dx + \int_\Gamma e_\Gamma\, y\,ds + \frac{1}{2}\int_\Omega |u|^2\,dx, \tag{2.79}$$

subject to

$$\boxed{\begin{aligned} -\Delta y + y &= u + e_\Omega \\ \partial_\nu y &= 0 \end{aligned}}$$

and the control constraints

$$0 \le u(x) \le 1.$$

For the sake of convenience, we again choose as domain the unit square $\Omega = (0,1)^2$, which has center $\hat{x} = (0.5, 0.5)^\top$. The functions y_Ω, e_Ω, and e_Γ will again be fitted in such a way that a desired solution results. To this end, we put $r = |x - \hat{x}| = \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2}$. The desired optimal control is the function depicted below, which is given by

$$\bar{u}(x) = \begin{cases} 1 & \text{if} \quad r > \frac{1}{3} \\ 12\,r^2 - \frac{1}{3} & \text{if} \quad r \in [\frac{1}{6}, \frac{1}{3}] \\ 0 & \text{if} \quad r < \frac{1}{6} \end{cases}$$

or, equivalently, by

$$\bar{u}(x) = \min\big\{1\,,\, \max\{0\,,\, 12\,r^2 - 1/3\}\big\}. \tag{2.80}$$

As the reader will be asked to show in Exercise 2.17, $\bar{u}$ can be recovered from the projection formula

$$\bar{u}(x) = \mathbb{P}_{[0,1]}\big\{-p(x)\big\} \quad \text{for a.e. } x \in \Omega,$$

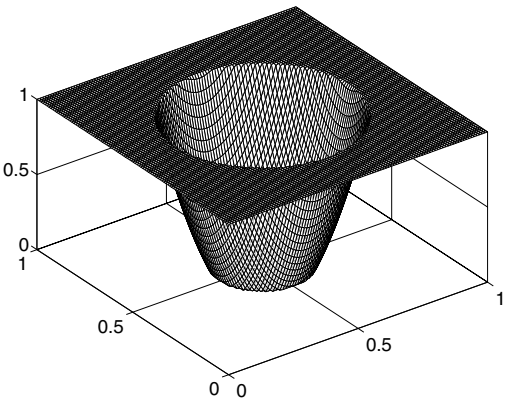
where the adjoint state p is the weak solution to the boundary value problem

$$\begin{aligned} -\Delta p + p &= \bar{y} - y_\Omega \\ \partial_\nu p &= e_\Gamma. \end{aligned} \tag{2.81}$$

We thus fix the adjoint state by putting

$$p(x) = -12\,|x - \hat{x}|^2 + \frac{1}{3} = -12\,r^2 + \frac{1}{3}\,.$$

The graph of $-p$ is a paraboloid, which is cut by the planes $\{p = 0\}$ and $\{p = 1\}$ in such a way that the above control $\bar{u}$ results. Having defined the adjoint state and control, we now choose the associated state $\bar{y}$. Since its normal derivative must vanish on Γ, we simply take $\bar{y}(x) \equiv 1$. For this function to satisfy the state equation, the function $e_\Omega = e_\Omega(x)$ is used as compensation. Clearly, since $\Delta\bar{y} = 0$, we must choose $e_\Omega = 1 - \bar{u}$. Hence, after substituting the expression (2.80) for $\bar{u}$,



Constructed control.

$$e_\Omega = 1 - \min\left\{1, \max\{0,\, 12\,r^2 - 1/3\}\right\}.$$

The functions y_Ω and e_Γ can still be chosen in order to fit equation (2.81) for the adjoint state. Application of the Laplacian to p yields

$$\Delta p = D_1^2 p + D_2^2 p = -12\,\{2 + 2\} = -48,$$

whence we conclude that

$$y_\Omega(x) = (\bar{y} + \Delta p - p)(x) = 1 - 48 - \frac{1}{3} + 12\,|x - \hat{x}|^2 = -\frac{142}{3} + 12\,r^2.$$

We still have to satisfy the boundary condition for p. To this end, we have included the boundary integral in the cost functional. Indeed, e_Γ needs to satisfy the relation

$$\partial_\nu p = e_\Gamma.$$

Obviously,

$$D_1 p = -24\,(x_1 - 0.5), \qquad D_2 p = -24\,(x_2 - 0.5),$$

so that on the part of the boundary given by $\{(x_1, x_2) \in \bar{\Omega} : x_1 = 0\}$ we have

$$\partial_\nu p = -D_1 p\,|_{x_1 = 0} = 24\,(0 - 0.5) = -12.$$

On the other parts of Γ, the same value results, so we may choose $e_\Gamma(x) = \partial_\nu p(x) \equiv -12$.

2.10. The formal Lagrange method

In the preceding section, we determined the actual form of the adjoint equation in a more or less intuitive way. However, this equation can easily be derived by means of a Lagrangian function. In this connection, we recall that in the finite-dimensional case treated in Section 1.4, the adjoint equation resulted from the derivative $D_y L$ of the Lagrangian with respect to y. Given the right formalism, this ought to be possible here as well.

Necessary optimality conditions in function spaces can be directly deduced from the Karush–Kuhn–Tucker theory for optimization problems in Banach spaces. This method, which might be called the *exact Lagrange method*, will be investigated in Chapter 6. In many cases, its application is difficult, requiring a lot of experience in matching the operators, functionals, and spaces involved. Indeed, the given quantities have to be differentiable in the chosen Banach spaces, adjoint operators have to be determined, and the Lagrange multipliers must exist in the right spaces. We will discuss some examples of the application of this method in Sections 2.13, 6.1.3, and 6.2.

Up to now, we have taken a different approach: we first expressed the state y by means of the control-to-state mapping G in the form $y = G(u)$. From this, we derived a variational inequality that was simplified by the introduction of the adjoint state p. The adjoint state is the Lagrange multiplier associated with the boundary value problem for the partial differential equation if this is defined as in Sections 2.13 or 6.1.3. This approach is equivalent to application of the general Karush–Kuhn–Tucker theory, since in this way one actually proves a Lagrange multiplier rule. However, because of the spaces involved, the application of Karush–Kuhn–Tucker theorems can be very complicated.

The above remarks apply to the *proof* of optimality conditions. It is, however, a completely different task to *derive* them (e.g., to determine the adjoint equation for complex problems), as well as to find a form for the conditions that is easy to memorize. For these purposes, the *formal Lagrange method* to be introduced here is particularly well suited. Basically, it is just the (exact) Lagrange principle described in, e.g., Ioffe and Tihomirov [**IT79**] and Luenberger [**Lue69**]. This principle will be discussed in Chapter 6, in particular for a problem involving a semilinear elliptic equation in Section 6.1.3.

The formal Lagrange method differs from the exact method in the following way: differential operators such as $-\Delta$ or ∂_ν are written formally, and all multipliers are regarded as functions, without specifying the underlying

spaces. One simply assumes that the state y and the multipliers that occur, as well as their derivatives, are all square integrable. In this way, L^2 scalar products can be used, and one avoids functionals from more general dual spaces.

This approach, while being justified in a certain sense, is not mathematically rigorous. It is not meant to be used as a tool for rigorous proofs, but primarily as a convenient means to derive and formulate the correct optimality conditions. Once these have been established, it does not matter anymore how they were found. This line of argument is particularly helpful for complex problems involving nonlinear systems of partial differential equations.

While the adjoint equation is too easy to guess for the problem of finding the optimal stationary heat source, it is quite instructive to demonstrate the basic ideas of the technique by means of the problem of finding the optimal stationary boundary temperature, defined in (2.32)–(2.34):

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

subject to

$$\boxed{\begin{array}{rcll} -\Delta y & = & 0 & \text{in } \Omega \\ \partial_\nu y + \alpha y & = & \alpha\, u & \text{on } \Gamma \end{array}}$$

and

$$u_a(x) \le u(x) \le u_b(x) \quad \text{a.e. on } \Gamma.$$

Here, three constraints have to be obeyed, two difficult ones (the boundary value problem) and a harmless one (the pointwise constraint $u \in U_{ad}$). We now apply the Lagrangian principle, eliminating only the equations by means of Lagrange multipliers p_1 and p_2. To this end, we define, still somewhat formally, the *Lagrangian function*

$$\mathcal{L} = \mathcal{L}(y,u,p) = J(y,u) - \int_\Omega (-\Delta y)\, p_1 \; dx - \int_\Gamma \big(\partial_\nu y - \alpha(u-y)\big)\, p_2 \; ds.$$

Here, the *Lagrange multipliers* p_1 and p_2 are functions defined on Ω and Γ, respectively, which in $\mathcal{L}$ are expressed as the vector $p := (p_1, p_2)$.

The definition of $\mathcal{L}$ is not rigorous for three reasons: we only know that $y \in H^1(\Omega)$, so neither Δy nor $\partial_\nu y$ need to be functions. Indeed, without further knowledge concerning the higher regularity of y, we can only claim that $\Delta y \in H^1(\Omega)^*$ and $\partial_\nu y \in H^{-1/2}(\Gamma)$ (for the definition of $H^{-1/2}(\Gamma)$, see, e.g., Lions and Magenes [**LM72**]). This has the unpleasant consequence that

the integrals which occur might be meaningless. In addition, it has to be clarified what regularity p_1 and p_2 have.

Nevertheless, we continue our approach, simply taking sufficient smoothness of p_1 and p_2 for granted, and integrate by parts using the second Green's formula. For the sake of brevity, we omit the differentials in the integrals. We obtain

$$\mathcal{L}(y,u,p) = J(y,u) + \int_\Gamma p_1 \partial_\nu y - \int_\Gamma y\, \partial_\nu p_1 + \int_\Omega y\, \Delta p_1 - \int_\Gamma \big(\partial_\nu y - \alpha\,(u-y)\big)\, p_2.$$

Recalling the Lagrange principle, we expect the pair $(\bar y, \bar u)$, together with the Lagrange multipliers p_1 and p_2, to satisfy the optimality conditions associated with the problem

$$\min \mathcal{L}(y,u,p), \qquad y \text{ unconstrained}, \ u \in U_{ad}.$$

Since y is now formally unconstrained, the derivative of $\mathcal{L}$ with respect to y has to vanish at the optimal point, that is,

$$\begin{aligned} D_y\mathcal{L}(\bar y,\bar u,p)\,h \;&=\; \int_\Omega \big((\bar y - y_\Omega) + \Delta p_1\big)\, h\, dx + \int_\Gamma \big(p_1 - p_2\big)\, \partial_\nu h\, ds \\ &\quad - \int_\Gamma \big(\partial_\nu p_1 + \alpha\, p_2\big)\, h\, ds = 0 \quad \forall\, h \in Y = H^1(\Omega). \end{aligned} \tag{2.82}$$

Here, we have used the fact that the derivative of a linear mapping is that mapping itself. Moreover, from the box constraints for u we deduce the variational inequality

$$D_u\mathcal{L}(\bar y,\bar u,p)(u-\bar u) \ge 0 \qquad \forall\, u \in U_{ad}.$$

Let us take a closer look at equation (2.82). First, we choose some $h \in C_0^\infty(\Omega)$, so that $h = \partial_\nu h = 0$ on Γ. We obtain that

$$\int_\Omega \big((\bar y - y_\Omega) + \Delta p_1\big)\, h\; dx = 0 \qquad \forall\, h \in C_0^\infty(\Omega),$$

and the density of $C_0^\infty(\Omega)$ in $L^2(\Omega)$ implies that

$$-\Delta p_1 = \bar y - y_\Omega \quad \text{in } \Omega.$$

This is already the first half of the adjoint system that had only been guessed at before, and we see that the first integral in (2.82) vanishes. Next, we only postulate $h|_\Gamma = 0$ and let $\partial_\nu h$ vary (see the remark (iii) below). For all such h we evidently have

$$\int_\Gamma (p_1 - p_2)\, \partial_\nu h\; ds = 0,$$

which is only possible if $p_1 = p_2$ on Γ. Finally, we vary h on Γ and consider the only remaining term in (2.82),

$$0 = -\int_\Gamma \big(\partial_\nu p_1 + \alpha\, p_2\big) h\; ds = -\int_\Gamma \big(\partial_\nu p_1 + \alpha\, p_1\big)\; h\; ds.$$

Since h is arbitrary on Γ (see remark (iii) below), we can infer that

$$\partial_\nu p_1 + \alpha p_1 = 0 \text{ on } \Gamma.$$

If we now put $p := p_1$ and $p_2 := p|_\Gamma$, then we recover the adjoint equation for the problem of the optimal stationary boundary temperature, which had been introduced intuitively before.

Observe that the variational inequality is also easily derived. We have

$$D_u\mathcal{L}(\bar{y},\bar{u},p)\,(u-\bar{u}) = \int_\Gamma \lambda\,\bar{u}\,(u-\bar{u})\;ds + \int_\Gamma \alpha\,p\,(u-\bar{u})\;ds \geq 0.$$

Next, we introduce the Lagrangian function in such a way that all the occurring terms are meaningful.

Definition. *The Lagrangian function* $\mathcal{L}: \; H^1(\Omega)\times L^2(\Gamma)\times H^1(\Omega)\to\mathbb{R}$ *for problem* (2.32)–(2.34) *is defined by*

$$\mathcal{L}(y,u,p) := J(y,u) - \int_\Omega \nabla y\cdot\nabla p\;dx + \int_\Gamma \alpha\,(u-y)\,p\;ds. \tag{2.83}$$

This form arises from the earlier, only formally correct form of $\mathcal{L}$, upon integrating once by parts. The terms containing $\partial_\nu y$ cancel each other out. We readily convince ourselves of the following equivalences:

$$\begin{aligned} &D_y\mathcal{L}(\bar{y},\bar{u},p)\,h = 0 \quad \forall\, h\in H^1(\Omega) && \Leftrightarrow \quad \text{weak formulation of (2.73);}\\ &D_u\mathcal{L}(\bar{y},\bar{u},p)\,(u-\bar{u}) \geq 0 \quad \forall\, u\in U_{ad} && \Leftrightarrow \quad \text{variational inequality (2.76).}\end{aligned}$$

Conclusion. *The Lagrangian function just introduced yields the optimality conditions derived in Section 2.8.4. The adjoint equation and the variational inequality are obtained by taking the derivative with respect to the state y and the control u, respectively.*

Although $\mathcal{L}$ is now properly defined by (2.83), one question remains unanswered: How do we know that such a $p\in H^1(\Omega)$ exists? In the preceding section, we avoided this problem by *defining* p directly as the solution to the adjoint equation; note, however, that the Lagrange method presented in this section has the sole purpose of determining the correct form of the adjoint equation. Nevertheless, the existence of p as Lagrange multiplier can be concluded directly from the Karush–Kuhn–Tucker theory in Banach spaces; see Section 2.13.

Remarks.
(i) If we assume right from the beginning that the Lagrange multipliers p_1 and p_2

coincide on the boundary, then the second integration by parts leading to the term $-\Delta p_1$ is superfluous. In fact, one integration by parts suffices, and the variational form of the adjoint equation is immediately established (cf. the argument in the next section). Unfortunately, this simplified approach does not always succeed. This is the case, e.g., if boundary controls of Dirichlet type are considered; see Exercise 2.19 on page 118.

(ii) The gradient of $f(u) = J(y(u), u)$ can be obtained from

$$f'(u) = D_u \mathcal{L}(y, u, p)$$

if $y = y(u)$ and $p = p(u)$ are inserted. We thus can memorize as a rule that this reduced gradient is calculated by differentiating the Lagrangian with respect to the control.

(iii) Above, we argued that $\partial_\nu h$ is essentially arbitrary on Γ within the set of all h with $h|_\Gamma = 0$. This is a consequence of the fact that the mapping $h \mapsto (\tau\, h, \partial_\nu h)$ is surjective from $H^2(\Omega)$ onto $H^{3/2}(\Gamma) \times H^{1/2}(\Gamma)$ (see, e.g., **[Ada78]**, Thm. 7.53). A similar conclusion can be drawn for h, since the trace operator $\tau : H^1(\Omega) \to H^{1/2}(\Gamma)$ is also a surjective mapping.

We have demonstrated the application of the Lagrangian function for problem (2.32)–(2.34) only. Other problems can be treated similarly. We also note that it is possible to eliminate the box constraints for the control by incorporating them into the Lagrangian by means of additional Lagrange multipliers μ_a and μ_b. This was demonstrated earlier in Section 1.4.7. The Lagrangian (2.83) then has to be extended in the following way:

$$\mathcal{L}(y, u, p, \mu_a, \mu_b) := J(y, u) - \int_\Omega \nabla y \cdot \nabla p \, dx + \int_\Gamma \alpha\,(u - y)\, p \, ds + \int_\Gamma \big(\mu_a(u_a - u) + \mu_b(u - u_b)\big)\, dx. \tag{2.84}$$

The formal Lagrange method described in this section will repeatedly serve us well when adjoint equations are to be determined. Lagrangian functions like (2.83) or (2.84) are the proper tool for formulating the optimality conditions in a both elegant and rigorous way. Later on, we will also employ them to represent second-order sufficient optimality conditions rigorously in a form that is commonly used in both finite-dimensional optimization and the foundation of numerical methods.

2.11. Further examples *

2.11.1. Differential operators in divergence form. In the same way, the general problem (2.36)–(2.38) on page 54 can be treated using the Lagrange method:

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

subject to

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

and

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

The state y vanishes on Γ_0; therefore, only its restriction to Γ_1 contributes to the boundary term in the cost functional.

Suppose that Assumption 2.19 on page 55 holds. The natural state space is evidently

$$Y = \left\{ y \in H^1(\Omega) :\, y|_{\Gamma_0} = 0 \right\}.$$

The optimality conditions can be derived as in Section 2.8.4. First, we construct the adjoint equation. The Lagrangian can be defined by

$$\begin{aligned}\mathcal{L}(y,u,v,p) \;=\; & J(y,u,v) - \int_\Omega \left(\mathcal{A}\,y \,+\, c_0\,y \,-\, \beta_\Omega\,v\right) p\,dx \\ & - \int_{\Gamma_1} \left(\partial_{\nu_\mathcal{A}} y \,+\, \alpha\,y \,-\, \beta_\Gamma\,u\right) p\,ds,\end{aligned}$$

since the boundary condition $y|_{\Gamma_0} = 0$ is already accounted for in the space Y. As a simplification, we tacitly assume that the multiplier p occurring in the boundary integral coincides with the trace of the multiplier p appearing in the integral over Ω. This follows as in Section 2.10. Again, we could have chosen different multipliers p_1 and p_2. After integration by parts and with the bilinear form a defined in (2.22) on page 38 it follows that for all $h \in Y$ we have

$$D_y\mathcal{L}(\bar y,\bar u,\bar v,p)\,h = \int_\Omega \lambda_\Omega\,(\bar y - y_\Omega)\,h\,dx + \int_\Gamma \lambda_\Gamma\,(\bar y - y_\Gamma)\,h\,ds - a[h,p] = 0.$$

It is not necessary to integrate by parts once more in order to transfer the second-order differential operator to p. Indeed, the above relation is just the variational equation corresponding to the weak solution to the adjoint equation

$$\boxed{\begin{array}{rcll} \mathcal{A}\,p + c_0\,p &=& \lambda_\Omega\,(\bar{y} - y_\Omega) & \text{in } \Omega \\ \partial_{\nu_\mathcal{A}} p + \alpha\,p &=& \lambda_\Gamma\,(\bar{y} - y_\Gamma) & \text{on } \Gamma_1 \\ p &=& 0 & \text{on } \Gamma_0. \end{array}} \tag{2.85}$$

Remark. The symmetry condition postulated for the operator $\mathcal{A}$ is a prerequisite neither for the well-posedness of the state equation nor for the optimality conditions. If this condition is dropped, then $\mathcal{A}$ has to be replaced in the adjoint equation by its (formally) associated adjoint operator, which is defined by

$$\mathcal{A}^*\,p(x) = -\sum_{i,j=1}^{N} D_j\,\big(a_{ij}(x)\,D_i\,p(x)\big).$$

We have generally postulated the symmetry condition, because it will be needed later in this textbook for certain regularity results.

The formal Lagrange method gave us a hint to introduce the adjoint equation in exactly this way. Owing to Theorem 2.7 on page 38, there exists a unique solution $p \in Y$ to this equation, and hence one and only one adjoint state associated with $\bar{y}$. Invoking a generalization of Lemma 2.31 on page 74 for the construction of S^*, one can derive the first-order necessary optimality conditions. The interested reader will be asked to do this in Exercise 2.18. Instead, we employ here the Lagrange method again. We obtain the variational inequalities

$$\begin{aligned} D_v\mathcal{L}(\bar{y},\bar{u},\bar{v},p)(v-\bar{v}) &= \int_\Omega (\lambda_v\,\bar{v} + \beta_\Omega\,p)(v-\bar{v})\;dx \geq 0 \quad \forall\, v \in V_{ad}, \\ D_u\mathcal{L}(\bar{y},\bar{u},\bar{v},p)(u-\bar{u}) &= \int_{\Gamma_1} (\lambda_u\,\bar{u} + \beta_\Gamma\,p)(u-\bar{u})\;ds \geq 0 \quad \forall\, u \in U_{ad}. \end{aligned}$$

In summary, the following necessary and sufficient first-order optimality conditions hold.

Theorem 2.34. *Suppose that Assumption 2.19 holds. Then the pair of controls $(\bar{v},\bar{u}) \in V_{ad} \times U_{ad}$, together with the associated state $\bar{y} \in Y$ and the corresponding adjoint state $p \in Y$ defined by* (2.85)*, is optimal if and only if the following variational inequalities are satisfied:*

$$\begin{aligned} (\lambda_v\,\bar{v} + \beta_\Omega\,p\,,\,v-\bar{v})_{L^2(\Omega)} &\geq 0 \quad \forall\, v \in V_{ad}, \\ (\lambda_u\,\bar{u} + \beta_\Gamma\,p\,,\,u-\bar{u})_{L^2(\Gamma_1)} &\geq 0 \quad \forall\, u \in U_{ad}. \end{aligned}$$

2.11.2. Optimal stationary heat source with given outside temperature. This problem, which was mentioned on page 4, is a special case of the above general problem with the choices $\mathcal{A} := -\Delta$, $c := 0$, $\beta_\Gamma := \alpha$, $\beta_\Omega := \beta$, and $u_a = u_b = y_a$. The boundary component of the control is fixed, since y_a is prescribed. The cost functional is even more general; by putting $\lambda_\Gamma = \lambda_u = 0$ and $\lambda_\Omega = 1$, it can be brought into the form discussed in Section 1.2.1. Here, the control v plays the role of u. The necessary optimality conditions follow immediately from Theorem 2.34.

2.12. Numerical methods

In this section, we are going to sketch some fundamental concepts for the numerical solution of linear-quadratic elliptic problems. This field of mathematics has been well established for many years, and there are quite a few methods available to tackle even complex problems successfully. Here, and in the other sections dealing with numerical methods, we can merely give the reader a feel for how to approach such problems numerically. It would be completely beyond the scope of this book if we attempted to present an even half-way comprehensive treatment of these methods. We therefore refer the interested reader to the relevant literature—in particular, to the monographs by Betts [**Bet01**], Gruver and Sachs [**GS80**], Hinze et al. [**HPUU09**], Kelley [**Kel99**], and Ito and Kunisch [**IK08**].

In most cases, we ignore the fact that the problems have to be discretized, tacitly assuming that the partial differential equations that arise can be solved explicitly. To be sure, a serious numerical analysis has to include the discretization of the equations by, e.g., finite differences or the finite element method. However, without these technical details it is much easier to illustrate the special features that originate from optimization theory. Nevertheless, finite difference techniques and the finite element method (FEM) will be briefly touched upon in Sections 2.12.3 and 2.12.4, respectively.

We begin our analysis with gradient methods. Historically, these were among the first techniques by which control problems for partial differential equations were solved. These methods are slow but easily implemented, and therefore well suited for exercises and first numerical tests. Moreover, for very complex and highly nonlinear problems they are still often the first choice.

After that, we study the direct transformation into a finite-dimensional optimization problem, using the finite difference method; we also derive the reduced optimal control problem, which proves to be advantageous in certain situations. Finally, we explain the basic ideas of the so-called *primal-dual*

active set strategy, which is one of the most efficient and commonly used numerical methods nowadays.

2.12.1. The conditioned gradient method. We discuss the conditioned gradient method mainly for didactic reasons, since it provides much insight into how the necessary optimality conditions can be exploited for the construction of numerical methods. In addition, it is easy to implement for exercises and testing. However, this method is comparatively slow, being only linearly convergent. In general, the so-called *projected gradient method* converges faster. For a detailed discussion of the conditioned gradient method, we refer the reader to Gruver and Sachs [**GS80**].

The conditioned gradient method in Hilbert spaces. We first formulate the conditioned gradient method for an optimization problem in the Hilbert space U:

$$\min_{u\in U_{ad}} f(u),$$

where $f : U \to \mathbb{R}$ denotes a Gâteaux differentiable functional and $U_{ad} \subset U$ is a nonempty, bounded, closed, and convex set. Suppose that the iterates $u_1, \dots, u_n$ have already been determined, so that u_n is the current approximation to the solution. Then the following steps have to be taken:

S1 (*Determination of a new descent direction*) We determine a direction v_n by solving the Hilbert space optimization problem

$$f'(u_n)\, v_n = \min_{v\in U_{ad}} f'(u_n)\, v.$$

This problem is linear with respect to the cost functional and, in view of the assumptions on U_{ad}, solvable. If $f'(u_n)(v_n - u_n) \geq 0$, then u_n solves the variational inequality (why?) and is a solution. The algorithm terminates. Otherwise, if $f'(u_n)(v_n - u_n) < 0$, then $v_n - u_n$ is a descent direction.

S2 (*Line search and step size control*) Find $s_n \in (0, 1]$ from solving the one-dimensional optimization problem

$$f\big(u_n + s_n\,(v_n - u_n)\big) = \min_{s\in(0,1]} f\big(u_n + s\,(v_n - u_n)\big).$$

Then put $u_{n+1} := u_n + s_n\,(v_n - u_n)$, $n := n + 1$, and go to S1.

In the convex case, the sequence $\big\{f(u_n)\big\}_{n=1}^{\infty}$ converges in a strictly decreasing fashion to the optimal value (descent method). Note that, since $u_n, v_n \in U_{ad}$, the convex combination $u_n + s_n\,(v_n - u_n)$ also belongs to U_{ad}. This is an essential feature of this method.

Application to elliptic control problems. We now apply the conditioned gradient method to the problem of finding the optimal stationary heat source:

$$\min J(y,u) := \frac{1}{2}\,\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}, \tag{2.86}$$

subject to

$$\boxed{\begin{aligned} -\Delta y &= \beta\, u \\ y|_\Gamma &= 0 \end{aligned}} \tag{2.87}$$

and

$$u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Omega. \tag{2.88}$$

The associated control space is $U = L^2(\Omega)$, U_{ad} is defined as before through the given box constraints (2.88) for u, and the reduced cost functional f is given by $f(u) = J\big(y(u), u\big)$. Owing to Lemma 2.30 on page 73, we have

$$f'(u_n)\,v = \int_\Omega (\beta\, p_n + \lambda\, u_n)\, v \; dx,$$

where p_n is the solution to the adjoint equation

$$\boxed{\begin{aligned} -\Delta p_n &= y_n - y_\Omega \\ (p_n)|_\Gamma &= 0. \end{aligned}} \tag{2.89}$$

Suppose that $u_1, \ldots, u_n$ are already known. The method then proceeds as follows:

S1 Determine the state y_n corresponding to u_n by solving the state equation (2.87).

S2 Determine the adjoint state p_n by solving the adjoint equation (2.89).

S3 (*Direction search*) Find v_n by solving the linear optimization problem

$$\min_{v \in U_{ad}} \int_\Omega (\beta\, p_n + \lambda u_n)\, v \, dx.$$

If the v_n obtained from this linear problem has the same value as u_n, then v_n is optimal, and the algorithm terminates. Since this is unlikely to happen in practice, it makes sense to incorporate a stopping criterion. For instance, one terminates the algorithm if the value for v_n is not smaller at least by $\varepsilon > 0$ than that for u_n.

S4 (*Step size control*) Determine the step size s_n by solving

$$f\big(u_n + s_n\,(v_n - u_n)\big) = \min_{s\,\in(0,1]} f\big(u_n + s\,(v_n - u_n)\big).$$

S5 Set $u_{n+1} := u_n + s_n\,(v_n - u_n)$, $n := n+1$, and go to S1.

Remarks on the practical performance. The steps S3 and S4 can be carried out analytically in the following way:

For S3: A meaningful direction v_n is evidently

$$v_n(x) := \begin{cases} u_a(x) & \text{if} \quad \lambda\,u_n(x) + \beta(x)\,p_n(x) > 0 \\ \dfrac{1}{2}(u_a + u_b)(x) & \text{if} \quad \lambda\,u_n(x) + \beta(x)\,p_n(x) = 0 \\ u_b(x) & \text{if} \quad \lambda\,u_n(x) + \beta(x)\,p_n(x) < 0. \end{cases}$$

Note that the second case is unlikely to occur in practice.

For S4: We exploit the fact that f is a quadratic cost functional. We have

$$\begin{aligned} f\big(u_n + s\,(v_n - u_n)\big) \;&=\; \frac{1}{2}\,\big\|y_n + s\,(w_n - y_n) - y_\Omega\big\|^2_{L^2(\Omega)} \\ &\quad + \frac{\lambda}{2}\,\big\|u_n + s\,(v_n - u_n)\big\|^2_{L^2(\Omega)}, \end{aligned}$$

where $y_n = y(u_n)$ and $w_n = y(v_n)$ denote the states associated with u_n and v_n, respectively. Since here only dependence on the step size s matters, we put $g(s) := f\big(u_n + s\,(v_n - u_n)\big)$. Straightforward computation yields that

$$\begin{aligned} g(s) \;&=\; \frac{1}{2}\,\big\|y_n - y_\Omega\big\|^2_{L^2(\Omega)} + s\,\big(y_n - y_\Omega\,,\,w_n - y_n\big)_{L^2(\Omega)} \\ &\quad + \frac{s^2}{2}\,\big\|w_n - y_n\big\|^2_{L^2(\Omega)} \;+\; \frac{\lambda}{2}\,\big\|u_n\big\|^2_{L^2(\Omega)} \;+\; \lambda\,s\,\big(u_n\,,\,v_n - u_n\big)_{L^2(\Omega)} \\ &\quad + s^2\,\frac{\lambda}{2}\,\big\|v_n - u_n\big\|^2_{L^2(\Omega)}. \end{aligned}$$

The function g is a parabola of the form $g(s) = g_0 + g_1\,s + g_2\,s^2$, where the constants g_i can be determined beforehand. Hence, the problem

$$s_n := \;\arg\,\min_{s\in(0,1]} g(s)$$

can be solved by hand. Its solution, which is the projection of the zero of $g'(s)$ onto $[0,1]$, is called the *exact step size*.

Initially, the conditioned gradient method exhibits fast convergence, but it then becomes slow. This behavior is characteristic of gradient methods.

Every iteration step requires the solution of two elliptic boundary value problems. An advantage is the possibility of carrying out the steps S3 and S4 analytically. In contrast to this, for the projected gradient method described below the determination of the step size is more complicated; it may require the solution of additional differential equations.

As mentioned initially, the conditioned gradient method can only be carried out in connection with a suitable discretization. Both the control u and the state y have to be replaced by discrete approximations, for instance by piecewise constant or piecewise linear functions. It should be clear how the steps of the algorithm described above have to be performed for the discretized quantities. In fact, the method just requires the solution of the elliptic boundary value problems involved and a prescription of routines for calculation of the integrals that occur. Note that this also applies to the $\lambda = 0$ case.

2.12.2. The projected gradient method. A better gradient method is the so-called *projected gradient method*. Since this method will be discussed in some detail in Chapter 3 for parabolic problems, here we only briefly explain its differences from the conditioned gradient method: in step S3, we choose as descent direction the negative gradient

$$v_n := -(\beta\, p_n + \lambda\, u_n).$$

To guarantee admissibility, the step size s_n is determined by solving the one-dimensional minimization problem

$$f\big(\mathbb{P}_{[u_a,u_b]}(u_n + s_n\, v_n)\big) = \min_{s>0}\ f\big(\mathbb{P}_{[u_a,u_b]}(u_n + s\, v_n)\big).$$

If there are no control constraints, then the optimal step size is obtained from solving

$$f\big(u_n + s_n\, v_n\big) = \min_{s>0} f(u_n + s\, v_n).$$

In the linear-quadratic case studied here, s_n can easily be determined as the exact step size (see page 94). The new approximation for the optimal control is obtained by setting

$$u_{n+1} := \mathbb{P}_{[u_a,u_b]}(u_n + s_n\, v_n),$$

regardless of how s_n has been determined.

In the presence of restrictions, the determination of the step size is a nontrivial task. The exact step size can usually no longer be calculated, so an acceptable step size has to be constructed numerically. For instance, one can employ the method of *bisection*: starting from a small initial step size s_0, e.g., the step size used in the previous iteration step, one takes consecutively $s = \frac{s_0}{2}, \frac{s_0}{4}, \frac{s_0}{8}$, and so on, until an s is found such that $f\big(\mathbb{P}_{[u_a,u_b]}(u_n + s\, v_n)\big)$ is sufficiently smaller than the previous value $f(u_n)$. Otherwise, the algorithm

is terminated after a prescribed number of bisections of the step size. Another meaningful method for the determination of the step size is *Armijo's rule* from the theory of nonlinear optimization; see, e.g., Nocedal and Wright [**NW99**] and Polak [**Pol97**].

For each new step size, the evaluation of the cost functional requires the solution of a partial differential equation, which is costly. Therefore, one sometimes chooses to work with a fixed, sufficiently small step size as long as a sufficiently large descent can be maintained.

Very nice expositions of the projected gradient method and its convergence properties in the finite-dimensional case can be found in Gruver and Sachs [**GS80**], Kelley [**Kel99**], and (with geometric illustration) Nocedal and Wright [**NW99**]; the Hilbert space case is treated in, e.g., Hinze et al. [**HPUU09**].

2.12.3. Transformation into a finite-dimensional quadratic optimization problem.

Derivation of a discretized problem. For the numerical treatment of problem (2.86)–(2.88), we have to discretize the elliptic boundary value problem, the state y, and the control u. If the boundary of the domain consists of pieces that are parallel to the axes, then *finite difference methods* offer a very simple means to this end. Here, we will pursue this approach for the sake of simplicity, even though it is unsatisfactory from the viewpoint of numerical analysis: optimal controls are, as a rule, not smooth enough to guarantee the convergence of the finite difference method; also, the related error estimates do not apply. Therefore, the finite element method is almost exclusively applied in the relevant literature. We will briefly discuss this method in Section 2.12.4. On the other hand, the finite difference method is easy to implement and thus very suitable for the purpose of testing.

Once again, we consider the problem of finding the optimal stationary heat source, (2.86)–(2.88). As the domain, we choose the two-dimensional unit square $\Omega = (0,1)^2$, which we subdivide into n^2 congruent subsquares:

$$\bar{\Omega} = \bigcup_{i,j=1}^{n} \bar{\Omega}_{ij}, \quad \Omega_{ij} = \Big(\frac{i-1}{n}, \frac{i}{n}\Big) \times \Big(\frac{j-1}{n}, \frac{j}{n}\Big), \quad i,j = 1,\dots,n. \tag{2.90}$$

The corners of these subsquares,

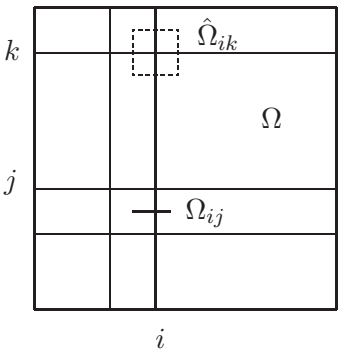
$$x_{ij} = h \begin{bmatrix} i \\ j \end{bmatrix}, \qquad i,\, j = 0,\dots,n,$$

form an equidistant grid of step size $h = 1/n$. The finite difference method determines approximations y_{ij} for the values $y(x_{ij})$ of the state at the grid points. Using the standard five-point stencil, we approximate $-\Delta y(x_{ij})$ at

the inner grid points by

$$-\Delta y(x_{ij}) \sim \frac{4\,y_{ij} - [y_{(i-1)\,j} + y_{i\,(j-1)} + y_{(i+1)\,j} + y_{i\,(j+1)}]}{h^2}, \tag{2.91}$$

$i,\, j = 1 \ldots n-1$. In addition, we define open squares $\hat{\Omega}_{ij}$ of side length h that are parallel to the axes and have midpoints x_{ij} (see the figure). Within these subsquares, the control u is assumed to be constant, i.e., we put $u(x) \equiv u_{ij}$ on $\hat{\Omega}_{ij}$ for $i, j = 1, \ldots, n-1$; within the remaining part of Ω in the vicinity of the boundary, we simply put $u(x) = 0$.



Subdivison of Ω.

Next, we number the quantities x_{ij}, y_{ij}, and u_{ij} lexicographically, e.g., from the southwest corner to the northeast corner of Ω. In this way, we obtain vectors $\vec{x} = (x_1, \ldots, x_{(n-1)^2})^\top$, $\vec{y} = (y_1, \ldots, y_{(n-1)^2})^\top$, and $\vec{u} = (u_1, \ldots, u_{(n-1)^2})^\top \in \mathbb{R}^{(n-1)^2}$. Moreover, we define the vectors $\vec{y}_\Omega$, $\vec{u}_a$, and $\vec{u}_b$ by putting $y_{\Omega,i} = y_\Omega(x_i)$, $u_{a,i} = u_a(x_i)$, and $u_{b,i} = u_b(x_i)$.

Finally, we replace the integrals occurring in the cost functional by midpoint rules, for example, $\int_{\hat{\Omega}_{ij}} y(x)\,dx \approx h^2 y_{ij}$. Then, dividing the cost functional by h^2, we infer from (2.91) a problem of the form

$$\min \frac{1}{2} \sum_{i=1}^{(n-1)^2} \left[(y_i - y_{\Omega,i})^2 + \lambda\, u_i^2\right]$$

$$A_h\, \vec{y} = B_h\, \vec{u}, \qquad \vec{u}_a \le \vec{u} \le \vec{u}_b.$$

Here, A_h is an $(n-1)^2 \times (n-1)^2$ matrix, and B_h is a diagonal matrix with the entries $b_{ii} = \beta(x_i)$.

Remark. A discretized version of the problem can also be set up using the finite element method; the matrices and the discretized cost functional then look different. We will discuss this approach in Section 2.12.4 in connection with the primal-dual active set strategy.

The above quadratic optimization problem with linear equality and inequality constraints can be implemented in existing numerical routines. For instance, the code quadprog in MATLAB can be employed. In addition, other programs for large optimization problems are available; a selection can be found on the website of NEOS (*NEOS Server for Optimization*). Elliptic

problems in two-dimensional domains do not present too many difficulties, as the results of Maurer and Mittelmann [**MM00, MM01**] for semilinear elliptic problems with control and state constraints show.

In this method, which is often referred to as *first discretize, then optimize*, $\vec{y}$ and $\vec{u}$ play the role of independent variables; the fact that they are state and control, respectively, is not exploited. In the next section, we will briefly demonstrate how to express the problem in terms of the control u alone. The dimensionality of this problem is smaller, which reduces the storage problems arising during the solution of quadratic optimization problems. This method works as long as the computer is able to handle the partial differential equation.

Formulation of a reduced optimization problem. If the control can be expressed as a linear combination of a relatively small number of basis functions, then it makes sense to reduce the problem to a quadratic optimization problem in terms of u alone. The basis functions may either result from a discretization (as described below) or be prescribed by the application under investigation.

Thus, let the control function u be of the form

$$u(x) = \sum_{i=1}^{m} u_i \, e_i(x), \tag{2.92}$$

with finitely many given functions $e_i : \Omega \to \mathbb{R}$ and real variables u_i. Moreover, suppose that the control constraints are either equivalent to

$$u_a \le u_i \le u_b, \quad i = 1, \ldots, m,$$

with given real numbers $u_a < u_b$, or given in this form from the beginning.

Example. Suppose that the basis functions $\epsilon_{ij}(x)$ are defined, with respect to the subdivision (2.90), by

$$\epsilon_{ij}(x) = \begin{cases} 1 & x \in \Omega_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

Their values on the boundaries between the subsquares are immaterial, since these sets have zero measure. We number the above functions from 1 to $m = n^2$, that is, $e_1 = \epsilon_{11}, \ldots, e_n = \epsilon_{1n}, e_{n+1} = \epsilon_{21}, \ldots, e_m = \epsilon_{nn}$. The resulting u is obviously a step function. Note that here the control is assumed to be constant on the subsquares Ω_{ij}, whereas in the last section the subsquares $\hat{\Omega}_{ij}$ were used (which was due to the finite difference method used for the solution of the differential equation).

For simplicity, we again assume that the state y associated with a given control u can be determined exactly as the solution to the state equation.

The main workload in formulating the reduced problem then consists of the determination of the functions

$$y_i(x) := \big(Se_i\big)(x), \quad i = 1, \ldots, m,$$

which are the solutions to the boundary value problems

$$\begin{aligned} -\Delta y &= \beta\, e_i \\ y|_\Gamma &= 0. \end{aligned}$$

In other words, m partial differential equations have to be solved in this approach. In the case of the subdivision into subsquares of side length $h = 0.01$ described above, this already amounts to 10^4 equations. Therefore, setting up a reduced problem is worthwhile only if either m is comparatively small or the same problem has to be solved several times with different data.

Once we have determined the functions y_i, we obtain the state $y = Su$ by superposition:

$$y = \sum_{i=1}^{m} u_i\, y_i. \tag{2.93}$$

Inserting the expressions (2.92) and (2.93) in the cost function yields the finite-dimensional cost functional

$$f_m(u_1, \ldots, u_m) = \frac{1}{2} \left\| \sum_{i=1}^{m} u_i\, y_i - y_\Omega \right\|^2_{L^2(\Omega)} + \frac{\lambda}{2} \left\| \sum_{i=1}^{m} u_i\, e_i \right\|^2_{L^2(\Omega)}.$$

The finite-dimensional approximation of the original problem then consists of finding $\vec{u} = (u_1, \ldots, u_m)^\top$ as the solution to

$$(P_m) \qquad \min_{u_a \le u_i \le u_b} f_m(\vec{u}).$$

A slight rearrangement of the cost functional yields

$$\begin{aligned} f_m(\vec{u}) &= \frac{1}{2} \big\|y_\Omega\big\|^2_{L^2(\Omega)} - \Big(y_\Omega\,,\, \sum_{i=1}^{m} u_i\, y_i\Big)_{L^2(\Omega)} \\ &\quad + \frac{1}{2}\Big(\sum_{i=1}^{m} u_i\, y_i\,,\, \sum_{j=1}^{m} u_j\, y_j\Big)_{L^2(\Omega)} + \frac{\lambda}{2}\Big(\sum_{i=1}^{m} u_i\, e_i\,,\, \sum_{j=1}^{m} u_j\, e_j\Big)_{L^2(\Omega)} \end{aligned}$$

$$= \frac{1}{2}\|y_\Omega\|^2_{L^2(\Omega)} - \sum_{i=1}^m u_i\,(y_\Omega\,,\,y_i)_{L^2(\Omega)} + \frac{1}{2}\sum_{i,j=1}^m u_i\,u_j\,(y_i\,,\,y_j)_{L^2(\Omega)}$$
$$+ \frac{\lambda}{2}\sum_{i,j=1}^m u_i\,u_j\,(e_i\,,\,e_j)_{L^2(\Omega)}.$$

Therefore, up to the constant $\frac{1}{2}\|y_\Omega\|^2_{L^2(\Omega)}$, (P_m) is equivalent to the finite-dimensional *quadratic optimization problem*

$$\boxed{\begin{array}{l}\min\left\{\frac{1}{2}\vec{u}^\top\big(C+\lambda D\big)\,\vec{u} - \vec{a}^\top\vec{u}\right\}\\[2ex] \vec{u}_a \le \vec{u} \le \vec{u}_b,\end{array}} \tag{2.94}$$

where $\vec{u}_a = (u_a,\dots,u_a)^\top$ and $\vec{u}_b = (u_b,\dots,u_b)^\top$. Evidently, the following quantities have to be calculated beforehand:

$$\begin{aligned}\vec{a} &= (a_i),\ a_i = (y_\Omega\,,\,y_i)_{L^2(\Omega)}\\ C &= (c_{ij}),\ c_{ij} = (y_i\,,\,y_j)_{L^2(\Omega)}\\ D &= (d_{ij}),\ d_{ij} = (e_i\,,\,e_j)_{L^2(\Omega)}.\end{aligned}$$

In the case of step functions, we have $(e_i\,,\,e_j)_{L^2(\Omega)} = \delta_{ij}\|e_i\|^2_{L^2(\Omega)}$, so that D is a diagonal matrix, $D = \operatorname{diag}\Big(\|e_i\|^2_{L^2(\Omega)}\Big)$.

For the solution of such problems, the aforementioned MATLAB code quadprog can be used. Numerous other codes can be found on the internet website of NEOS (NEOS Server for Optimization).

2.12.4. The primal-dual active set strategy.

The infinite-dimensional case. For control problems involving partial differential equations, this method goes back to Ito and Kunisch [**IK00**]. Here, it will be explained for the optimal stationary heat source problem,

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

subject to

$$u \in U_{ad} = \big\{u \in L^2(\Omega)\,:\,u_a(x) \le u(x) \le u_b(x)\ \text{ for a.e. } x \in \Omega\big\},$$

where $S : L^2(\Omega) \to L^2(\Omega)$ denotes the solution operator of the boundary value problem

$$\begin{aligned} -\Delta y &= u \\ y|_\Gamma &= 0 \end{aligned}$$

(for better readability, we assume $\beta \equiv 1$).

According to the projection relation (2.58) on page 70, the optimal control has to satisfy

$$\bar{u}(x) = \mathbb{P}_{[u_a(x),u_b(x)]} \left\{ -\lambda^{-1} p(x) \right\},$$

where the adjoint state p is the solution to problem (2.50) on page 67. Therefore, with

$$\mu = -\left(\lambda^{-1} p + \bar{u}\right) = -\lambda^{-1} f'(\bar{u}),$$

we have

(2.95)
$$\bar{u}(x) = \begin{cases} u_a(x) & \text{if} \quad -\lambda^{-1} p(x) < u_a(x) & (\Leftrightarrow \mu(x) < 0) \\ -\lambda^{-1} p(x) & \text{if} \quad -\lambda^{-1} p(x) \in [u_a(x), u_b(x)] & (\Leftrightarrow \mu(x) = 0) \\ u_b(x) & \text{if} \quad -\lambda^{-1} p(x) > u_b(x) & (\Leftrightarrow \mu(x) > 0). \end{cases}$$

In the first case, $\mu(x) < 0$ and thus $\bar{u}(x) + \mu(x) < u_a(x)$, by the definition of μ and the fact that $\bar{u} = u_a$. Similarly, $\bar{u}(x) + \mu(x) > u_b(x)$ in the third case. In the second case, we have $\mu(x) = 0$, hence $\bar{u}(x) + \mu(x) = -\lambda^{-1} p(x) \in [u_a(x), u_b(x)]$. In summary, $u = \bar{u}$ satisfies the relations

(2.96)
$$u(x) = \begin{cases} u_a(x) & \text{if} \quad u(x) + \mu(x) < u_a(x) \\ -\lambda^{-1} p(x) & \text{if} \quad u(x) + \mu(x) \in [u_a(x), u_b(x)] \\ u_b(x) & \text{if} \quad u(x) + \mu(x) > u_b(x). \end{cases}$$

Conversely, if $u \in U_{ad}$ satisfies (2.96), then u satisfies the projection condition and is therefore optimal. To see this, we discuss the first relation: since $u = u_a$, it immediately follows that $\mu(x) < 0$; hence

$$0 > -\lambda^{-1} p - u = -\lambda^{-1} p - u_a$$

so that $-\lambda^{-1} p < u_a$, and thus $u(x) = \mathbb{P}_{[u_a(x),u_b(x)]} \left\{ -\lambda^{-1} p(x) \right\}$. The other cases are treated similarly.

Summarizing, we conclude that the quantity $u + \mu$ indicates whether the inequality restrictions are active or not. These considerations motivate the *primal-dual active set strategy* described next.

As initial guesses, two arbitrary functions u_0, $\mu_0 \in L^2(\Omega)$ are chosen, where u_0 is not necessarily admissible. Suppose that the iterates u_{n-1} and μ_{n-1} have already been found. To determine u_n, the following steps are carried out:

S1 (*New active, respectively, inactive set*)
Put

$$\begin{aligned} A_n^b &= \{x : u_{n-1}(x) + \mu_{n-1}(x) > u_b(x)\} \\ A_n^a &= \{x : u_{n-1}(x) + \mu_{n-1}(x) < u_a(x)\} \\ I_n &= \Omega \setminus (A_n^b \cup A_n^a). \end{aligned}$$

If $A_n^a = A_{n-1}^a$ and $A_n^b = A_{n-1}^b$, then we have optimality and terminate the algorithm. Otherwise, we continue with the next step.

S2 (*New control*)
Determine the solution y, $p \in H_0^1(\Omega)$ to the linear system

$$\begin{aligned} -\Delta y &= u \\ -\Delta p &= y - y_\Omega \end{aligned} \qquad u = \begin{cases} u_a & \text{on} \quad A_n^a \\ -\lambda^{-1} p & \text{on} \quad I_n \\ u_b & \text{on} \quad A_n^b. \end{cases}$$

Observe that this system constitutes the first-order necessary optimality condition for a solvable linear-quadratic optimal control problem (namely the one that results if, in the initially given problem, the control u is fixed by taking $u = u_a$ on A_n^a and $u = u_b$ on A_n^b); hence, it admits a unique solution.

One then puts

$$u_n := u, \quad p_n := p, \quad \mu_n := -(\lambda^{-1} p_n + u_n), \quad n := n + 1,$$

and continues with step **S1**.

It is convenient to rewrite the linear system to be solved in step S2 in a slightly different form. To this end, let χ_n^a and χ_n^b denote the characteristic functions of the sets A_n^a and A_n^b, respectively. Then, evidently,

$$u + (1 - \chi_n^a - \chi_n^b)\, \lambda^{-1} p = \chi_n^a\, u_a + \chi_n^b\, u_b,$$

and in step S2 the following system has to be solved, where y and p vanish on the boundary:

$$\begin{aligned} -\Delta y \quad -u &= 0 \\ -\Delta p \quad -y &= -y_\Omega \\ (1 - \chi_n^a - \chi_n^b)\, \lambda^{-1} p \quad +u &= \chi_n^a\, u_a + \chi_n^b\, u_b. \end{aligned} \tag{2.97}$$

We also observe that the case distinction in (2.95) remains unchanged if the function μ is replaced by $c\,\mu$ with some constant $c > 0$. We may therefore work with $c\,\mu_{n-1}$ in place of μ_{n-1} in step S1, which can be of benefit in numerical computations. Sometimes this is also used for the convergence analysis.

For a detailed analysis of the method and of more general variants in which, e.g., admissibility is enforced by a multiplier shift, we refer the interested reader to the book by Ito and Kunisch [**IK08**], as well as to Ito and Kunisch [**IK00**], Bergounioux et al. [**BIK99**], and Kunisch and Rösch [**KR02**]. The method can be interpreted as a semismooth Newton method, which explains the fact that it usually converges superlinearly; see [**IK08**] or [**HPUU09**].

Numerical realization using a finite element method. The numerical realization of the primal-dual active set strategy requires the setting up of a discretized analogue. In contrast to the discretization of Poisson's equation by finite differences described on page 96, this time we employ linear finite elements. To fix things, suppose that $\Omega \in \mathbb{R}^2$ is a polygonal domain, which is subdivided by a regular triangulation into finitely many triangles with pairwise disjoint interiors. Associated with the triangulation is a finite set of continuous piecewise linear basis functions $\{\Phi_1, ..., \Phi_\ell\} \subset H_0^1(\Omega)$. Neither the regular triangulation nor the basis functions are described in greater detail; for more information, the reader is referred to the monographs by Braess [**Bra07**], Brenner and Scott [**BS94**], Ciarlet [**Cia78**], and Grossmann and Roos [**GR05**].

The control is assumed to be piecewise constant. More precisely, u is assumed to be constant on each triangle of the triangulation. We denote by e_i, $i = 1, ..., m$, the associated set of unit step functions, i.e., e_i equals unity on the subtriangle with index i and zero elsewhere.

In summary, we use the ansatz

$$y(x) = \sum_{i=1}^{\ell} y_i \Phi_i(x), \qquad u(x) = \sum_{i=1}^{m} u_i e_i(x),$$

with the real unknowns y_i and u_j, for $i = 1, \ldots, \ell$, $j = 1, \ldots, m$. Inserting these expressions into the weak form of the boundary value problem and choosing Φ_j as the test function, we find that

$$\int_\Omega \sum_{i=1}^{\ell} y_i \nabla\Phi_i \cdot \nabla\Phi_j \, dx = \int_\Omega \sum_{i=1}^{m} u_i \, e_i \Phi_j \, dx.$$

In terms of the unknown vectors $\vec{y} = (y_1, \dots, y_\ell)^\top$ and $\vec{u} = (u_1, \dots, u_m)^\top$, this is a system of linear equations of the form

$$K_h \vec{y} = B_h \, \vec{u}$$

where the elements of the *stiffness matrix* K_h and the matrix B_h are given by

$$k_{h,ij} = \int_\Omega \nabla\Phi_i \cdot \nabla\Phi_j \, dx, \qquad b_{h,ij} = \int_\Omega \Phi_i \, e_j \, dx.$$

The positive number h, called the *mesh size* of the triangulation, is a measure of the refinement of the mesh. For the cost functional, we find after a straightforward computation the identity

$$\frac{1}{2} \left\| y - y_\Omega \right\|^2_{L^2(\Omega)} + \frac{\lambda}{2} \left\| u \right\|^2_{L^2(\Omega)} = \frac{1}{2} \vec{y}^\top M_h \vec{y} - \vec{a_h}^\top \vec{y} + \frac{\lambda}{2} \vec{u}^\top D_h \vec{u} + \frac{1}{2} \left\| y_\Omega \right\|^2_{L^2(\Omega)},$$

where the entries of the *mass matrices* M_h and D_h and of the vector $\vec{a}_h$ are given by

$$m_{h,ij} = \int_\Omega \Phi_i \, \Phi_j \, dx, \quad d_{h,ij} = \int_\Omega e_i \, e_j \, dx, \quad a_{h,i} = \int_\Omega \Phi_i \, y_\Omega \, dx.$$

The constant term $\frac{1}{2}\|y_\Omega\|^2_{L^2(\Omega)}$ does not influence the minimization. Since the interiors of the triangles are pairwise disjoint, D_h is a diagonal matrix.

Summarizing, we have the following discretized analogue of the initial optimal control problem:

$$\boxed{\begin{aligned} \min \left\{ \frac{1}{2} \vec{y}^\top M_h \vec{y} - \vec{a_h}^\top \vec{y} + \frac{\lambda}{2} \vec{u}^\top D_h \vec{u} \right\} \\ K_h \, \vec{y} = B_h \vec{u} \\ \vec{u}_a \, \le \, \vec{u} \, \le \, \vec{u}_b, \end{aligned}} \tag{2.98}$$

where $\vec{u}_a = (u_a, \dots, u_a)^\top$ and $\vec{u}_b = (u_b, \dots, u_b)^\top$. The associated optimality system reads

$$\begin{aligned} & K_h \, \vec{y} = B_h \vec{u}, \qquad \vec{u}_a \le \vec{u} \le \vec{u}_b \\ & K_h \, \vec{p} = M_h \vec{y} - \vec{a}_h \\ & (\lambda \, D_h \vec{u} + B_h^\top \vec{p})^\top (\vec{v} - \vec{u}) \ge 0 \quad \forall \, \vec{u}_a \le \vec{v} \le \vec{u}_b. \end{aligned}$$

Since D_h is a diagonal matrix, we can argue as in the infinite-dimensional case. Putting

$$\vec{\mu} = - \left((\lambda D_h)^{-1} B_h^\top \vec{p} + \vec{u} \right),$$

for the components of the optimal vector $\vec{u}$ we obtain

$$u_i = \begin{cases} u_a & \text{if } u_i + \mu_i < u_a \\ \mu_i & \text{if } u_i + \mu_i \in [u_a, u_b] \\ u_b & \text{if } u_i + \mu_i > u_b \end{cases} \qquad 1 \le i \le m.$$

Comparison with the infinite-dimensional case thus motivates the following primal-dual active set strategy:

First, we choose initial guesses $\vec{u_0}$ and $\vec{\mu_0}$. In the nth step of the algorithm, we define the sets of active and inactive restrictions according to

$$\begin{aligned} A_n^b &= \big\{ i \in \{1, \dots, m\} \,:\, u_{n-1,i} + \mu_{n-1,i} > u_b \big\} \\ A_n^a &= \big\{ i \in \{1, \dots, m\} \,:\, u_{n-1,i} + \mu_{n-1,i} < u_a \big\} \\ I_n &= \{1, \dots, m\} \setminus (A_n^b \cup A_n^a). \end{aligned}$$

Having defined the new active sets A_n^a and A_n^b, we determine, in analogy to the characteristic functions χ_n^a and χ_n^b introduced on page 102, the diagonal matrices X_n^a and X_n^b with the diagonal elements

$$X_{n,ii}^a = \begin{cases} 1 & \text{if } i \in A_n^a \\ 0 & \text{otherwise} \end{cases}, \qquad X_{n,ii}^b = \begin{cases} 1 & \text{if } i \in A_n^b \\ 0 & \text{otherwise} \end{cases}.$$

Next, we put $E_h := (\lambda D_h)^{-1}(I - X_n^a - X_n^b)$. The diagonal elements $e_{h,ii}$ of E_h vanish if and only if $i \in A_n^a \cup A_n^b$. We then have to solve the following system of linear equations for $\vec{p}$, $\vec{y}$, $\vec{u}$:

$$\begin{bmatrix} 0 & K_h & -B_h \\ K_h & -M_h & 0 \\ E_h B_h^\top & 0 & I \end{bmatrix} \begin{bmatrix} \vec{p} \\ \vec{y} \\ \vec{u} \end{bmatrix} = \begin{bmatrix} 0 \\ -\vec{a}_h \\ X_n^a \, \vec{u}_a + X_n^b \, \vec{u}_b \end{bmatrix}. \tag{2.99}$$

Once this has been done, we put $\vec{u}_n := \vec{u}$ and $\mu_n := -\big((\lambda D_h)^{-1} B_h^\top \vec{p}_n + \vec{u}_n\big)$.

It can be shown that this algorithm terminates at the optimal solution after finitely many steps. This happens when $A_n^a = A_{n-1}^a$ and $A_n^b = A_{n-1}^b$ for the first time, because then $\vec{u}_n$ satisfies all the restrictions. In this sense, all iterates generated by the primal-dual active set strategy are inadmissible, except for the last one, when the optimum is achieved. For the proofs, we refer the reader to [**BIK99**] and [**KR02**]. It is possible to achieve the admissibility of all iterates by a shift of the multipliers; see [**IK00**]. Moreover, as in the infinite-dimensional case, $c\,\mu_n$ may be used instead of μ_n.

Often, the piecewise linear ansatz $u = \sum_{i=1}^n u_i \Phi_i$ is used in place of the piecewise constant one. Then D_h is no longer a diagonal matrix, and the above discussion of the variational inequality fails. In this situation,

the active set strategy for the continuous case described in Section 2.12.4 is applied pointwise at the mesh points of the triangulation; that is to say, in the case distinctions the points $x \in \Omega$ are replaced by the mesh points x_i.

Remark. An important task is of course to estimate the error of the method, i.e., the amount by which the exact optimal control and the optimal control of the discretized problem differ. For elliptic problems we refer to [**ACT02**], [**CM02b**], [**CMT05**], [**Hin05**], [**HPUU09**], and [**MR04**], and for parabolic problems to [**Mal81**], [**Rös04**], and [**TT96**].

Other active set strategies. The strategy described above generates inadmissible controls until the optimum is reached. With projected Newton methods, there exist schemes that always generate admissible controls and exhibit a convergence behavior comparable to that of the primal-dual active set strategy described above. Such methods have been successfully applied to parabolic problems, in particular. We refer to Bertsekas [**Ber82**] and Kelley [**Kel99**] regarding the foundation of these methods, and to Kelley and Sachs [**KS94, KS95**] with respect to their application to the solution of parabolic optimal control problems.

Direct solution of the optimality system. Another technique that can be recommended is the direct numerical solution of the nonsmooth optimality system

$$\begin{aligned} -\Delta y &= \beta\, \mathbb{P}_{[u_a,u_b]}\left\{-\lambda^{-1}\beta\, p\right\} & \qquad -\Delta p &= y - y_\Omega \\ y|_\Gamma &= 0 & p|_\Gamma &= 0. \end{aligned}$$

The application of this method to parabolic problems will be discussed in some detail on page 170.

2.13. The adjoint state as a Lagrange multiplier *

2.13.1. Elliptic equations with data from V^*. Using the theory of weak solutions, all of the elliptic boundary value problems investigated in this chapter have been transformed into the general form

$$a[y, v] = F(v), \tag{2.100}$$

with a continuous bilinear form $a : V \times V \to \mathbb{R}$ and a functional $F \in V^*$. We now consider the bilinear form from a different point of view. To this end, observe that for any fixed $y \in V$ the linear mapping $a_y : V \to \mathbb{R}$, $v \mapsto a[y, v]$, is continuous on V and thus an element of V^*. The linear mapping $A : V \to V^*$, $y \mapsto a_y$, is continuous, since it satisfies, with the constant α_0 from the inequality (2.6) on page 32 (continuity of the bilinear

form a),

$$\begin{aligned} \|A\,y\|_{V^*} &= \sup_{\|v\|_V=1} \big|a_y(v)\big| = \sup_{\|v\|_V=1} \big|a[y,v]\big| \\ &\le \sup_{\|v\|_V=1} \alpha_0\,\|y\|_V\,\|v\|_V = \alpha_0\,\|y\|_V. \end{aligned}$$

Clearly, this implies that $\|A\| \le \alpha_0$. Hence, A is bounded and therefore continuous. Moreover, we have

$$a[y,v] = a_y\,(v) \quad \forall\, y,\, v \in V, \tag{2.101}$$

so that the variational equality $a[y,v] = F(v)$ can be regarded as an equality in V^*, namely,

$$A\,y = F.$$

By virtue of the Lax–Milgram lemma, for any functional $F \in V^*$ this equation has under the corresponding assumptions a unique solution $y \in V$, and $\|y\|_V \le c_a\,\|F\|_{V^*}$. Consequently, the inverse operator $A^{-1} : V^* \to V$, $F \mapsto y$, exists and is continuous. Observe that the continuity of A^{-1} can also be concluded from the well-known open mapping theorem, since A is surjective. In summary, we have shown the following result.

Lemma 2.35. *Every V-elliptic and bounded bilinear form $a = a[y,v]$ generates via*

$$\langle A\,y\,,\,v\rangle_{V^*,V} = a[y,v] \qquad \forall y,\, v \in V$$

a continuous and bijective linear operator $A : V \to V^*$. The inverse operator $A^{-1} : V^* \to V$ is continuous as well.

The application of Lemma 2.35 offers several advantages. In particular, one can work with the operator A in a similar way as with the matrix A in Section 1.4. The representation becomes symmetric, since it holds for the adjoint operator $A^* : (V^*)^* \to V^*$ and therefore $A^* : V \to V^*$, provided that V is reflexive. In this textbook, this will always be the case.

2.13.2. Application to the proof of optimality conditions. In this section, we aim to demonstrate the advantages of Lemma 2.35 by applying it to the problem of finding the optimal stationary boundary temperature. Here, $y \in H^1(\Omega)$ is the weak solution to the state equation

$$\begin{aligned} -\Delta y &= 0 \quad \text{in} \quad \Omega \\ \partial_\nu y + \alpha\,y &= \alpha\,u \quad \text{on} \quad \Gamma, \end{aligned}$$

where we postulate that $\alpha \geq 0$ almost everywhere on Γ and $\|\alpha\|_{L^\infty(\Gamma)} > 0$. We choose $V = H^1(\Omega)$ and define a and F by

$$\begin{aligned} a[y,v] &= \int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Gamma \alpha \, y \, (\tau \, v) \, ds \\ F(v) &= \int_\Gamma \alpha \, u \, (\tau \, v) \, ds, \end{aligned}$$

where $\tau : V \to L^2(\Gamma)$ denotes the trace operator. Now let $A : V \to V^*$ be the operator generated by the bilinear form a. Then the given problem can be rewritten in the form

$$\begin{aligned} \min \, J(y,u) &:= \frac{1}{2} \, \|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2} \, \|u\|^2_{L^2(\Gamma)} \\ A \, y &= B \, u, \quad u \in U_{ad}. \end{aligned} \tag{2.102}$$

Here, the linear operator $B : L^2(\Gamma) \to V^* = H^1(\Omega)^*$ is defined by

$$\langle B \, u \, , \, v \rangle_{V^*,V} = \int_\Gamma \alpha \, u \, (\tau \, v) \, ds \quad \forall v \in V.$$

In the following, we identify $L^2(\Omega)^*$ with $L^2(\Omega)$, but not V^* with V since this would, for example, necessitate using the H^1 scalar product at places where this is not appropriate.

Next, we define the solution operator $G := A^{-1} : V^* \to V$, $u \mapsto y = GBu$. Again, we consider G as an operator S with range in $L^2(\Omega)$, which is meaningful since $V \hookrightarrow L^2(\Omega)$. We thus put $S = E_V G$, with the embedding operator $E_V : V \to L^2(\Omega)$. Then $S : V^* \to L^2(\Omega)$ and

$$S = E_V \, A^{-1}.$$

The adjoint operator S^* maps $L^2(\Omega)$ into V and thus into $L^2(\Omega)$. Putting $y = E_V G \, B \, u = S \, B \, u$ in the cost functional $J(y,u)$, we find that problem (2.102) can be rewritten as

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

Theorem 2.14 on page 50 yields the existence of an optimal control $\bar{u}$, and from (2.45) in Theorem 2.22 on page 64 we deduce the associated variational inequality

$$\big(B^* S^* (\bar{y} - y_\Omega) + \lambda \, \bar{u} \, , \, u - \bar{u}\big)_{L^2(\Gamma)} \geq 0 \quad \forall u \in U_{ad}. \tag{2.103}$$

Next, we define the adjoint state p by

$$p := S^*(\bar{y} - y_\Omega) = \big(A^{-1}\big)^* \, E_V^* \, (\bar{y} - y_\Omega).$$

Then p solves the adjoint equation

$$A^*p = E_V^*\,(\bar{y} - y_\Omega). \tag{2.104}$$

The operator $E_V^* : L^2(\Omega) \to V^*$ assigns the function $\bar{y} - y_\Omega$ to itself, but we consider it as a functional on V, defined by the equation $\big(E_V^*\,(\bar{y} - y_\Omega)\big)(v) := (\bar{y} - y_\Omega\,,\, v)_{L^2(\Omega)}$. It remains to determine the explicit form of A^*. It follows from the symmetry of a that

$$\langle A\,y\,,\, v\rangle_{V^*,V} = a[y,v] = a[v,y] = \langle A\,v\,,\, y\rangle_{V^*,V} = \langle y\,,\, A\,v\rangle_{V,V^*} \quad \forall y,\, v \in V,$$

which shows that $A = A^*$. Therefore, the adjoint equation (2.104) is equivalent to the boundary value problem

$$\begin{aligned} -\Delta p &= \bar{y} - y_\Omega \\ \partial_\nu p + \alpha\, p &= 0. \end{aligned} \tag{2.105}$$

Finally, the adjoint operator $B^* : V \to L^2(\Gamma)$ is given by $B^*p = \alpha\,(\tau\, p)$, so that the variational inequality (2.103) takes the form

$$(\alpha\,(\tau\, p) + \lambda\,\bar{u}\,,\, u - \bar{u})_{L^2(\Gamma)} \geq 0 \quad \forall u \in U_{ad}. \tag{2.106}$$

All these results have already been derived above, using different techniques. However, the method employed here is more general. It is applied, in particular, in the monograph by [**Lio71**], which we have followed here. For instance, it allows for the use of control functions from L^r spaces with $r < 2$ or even more general functionals from V^* (cf. page 40). In addition, with this method the adjoint state p can be defined as a Lagrange multiplier in a natural way, as will be explained in the next section.

2.13.3. The adjoint state as a multiplier. The result just proved can also be deduced from the Lagrange multiplier rule for optimization problems in Banach spaces. In anticipation of the contents of Chapter 6, we demonstrate this for problem (2.102). This is an optimization problem in a Banach space of the type (6.1) on page 324, with the equality constraint $A\,y - B\,v = 0$ for the unknown $u := (y, v) \in U := Y \times L^2(\Gamma)$ and the range space $Z := Y^*$ for the equation. To ensure compatibility with the notation of Chapter 6, we denote the control by v (instead of u), and put $Y := H^1(\Omega)$ and $V_{ad} := U_{ad}$. We thus consider the problem

$$\min J(y,v) := \frac{1}{2}\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{1}{2}\|v\|^2_{L^2(\Gamma)}, \quad A\,y - B\,v = 0, \quad v \in V_{ad}.$$

The corresponding Lagrangian function $L : Y \times U \times Y^{**} \to \mathbb{R}$ reads, in view of (6.2) on page 325,

$$\begin{aligned} L(y,v,z^*) &= J(y,v) + \big\langle z^*\,,\, A\,y - B\,v\big\rangle_{Y^{**},Y^*} \\ &= \frac{1}{2}\|E_V y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\|v\|^2_{L^2(\Gamma)} + \big\langle z^*\,,\, A\,y - B\,v\big\rangle_{Y^{**},Y^*}. \end{aligned}$$

Since Y is reflexive, we can identify Y^{**} with Y and hence z^* with an element of Y. Evidently, A is a surjective operator. Therefore, the regularity condition (6.11) for equality constraints by Zowe and Kurcyusz is fulfilled. Hence, owing to Theorem 6.3 on page 330, there exists a Lagrange multiplier $z^* \in Y$ such that the variational inequality (6.13) holds, i.e., in the present case,

$$D_{(y,v)}L(\bar y, \bar v, z^*)(y - \bar y, v - \bar v) \geq 0 \quad \forall (y,v) \in Y \times V_{ad}.$$

Since $y \in Y$ may be chosen arbitrarily, it follows that

$$D_y L(\bar y, \bar v, z^*) = 0,$$

so that

$$E_V^*(E_V\,\bar y - y_\Omega) + A^* z^* = 0.$$

Putting $p := -z^*$ and recalling that $E_V\,\bar y = \bar y$, we finally arrive at equation (2.104),

$$A^*\,p = E_V^*(\bar y - y_\Omega).$$

Its unique solution is the weak solution to problem (2.105). We have thus shown the following result.

Lemma 2.36. *The adjoint state p associated with the optimal control $\bar v$ of the optimal stationary heat source problem is the (uniquely determined) Lagrange multiplier corresponding to the state equation $A\,y - B\,v = 0$.*

For the sake of completeness, we mention that the variational inequality

$$D_v L(\bar y, \bar v, z^*)(v - \bar v) \geq 0 \quad \text{for all } v \in V_{ad}$$

implies, as is to be expected, the variational inequality (2.106), here formulated with v in place of u.

In the same way, all the other problems for elliptic equations can be treated. A similar approach is also appropriate for parabolic equations. Difficulties arise for certain classes of nonlinear equations (see Chapter 4), since right-hand sides $v \in Y^*$ only lead to $y \in H^1(\Omega)$. We will, however, need boundedness of the state y, which is not guaranteed in $H^1(\Omega)$ if $N \geq 2$. In the study of problems with state constraints, we will later even need the regularity $y \in C(\bar\Omega)$.

2.14. Higher regularity for elliptic problems

2.14.1. Limited applicability of the state space $H^1(\Omega)$. In the present chapter, the state was generally chosen from the space $H^1(\Omega)$. While this is appropriate for standard linear-quadratic problems, simple changes already lead to difficulties that cannot be resolved in $H^1(\Omega)$, as the following two examples show.

Evaluation at a point. In this example, we replace the quadratic integral functional used so far by the value of the state at a fixed point $x_0 \in \Omega$. We thus consider the following problem:

$$\min\, y(x_0),$$

subject to

$$\begin{aligned} -\Delta y + y &= 0 \quad \text{in} \quad \Omega \\ \partial_\nu\, y + \alpha\, y &= u \quad \text{on} \quad \Gamma \end{aligned}$$

and constraints on the control $u \in L^2(\Gamma)$,

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

This problem is not well posed in $H^1(\Omega)$ unless Ω is one-dimensional. Indeed, y must be continuous for $y(x_0)$ to be defined, and functions in $H^1(\Omega)$ need not be continuous if $\dim \Omega \ge 2$. We will treat this problem in Section 6.2.1, using $Y = H^1(\Omega) \cap C(\bar\Omega)$ as the state space.

Best approximation with respect to the maximum norm. A similar situation as in the above example arises for the problem

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

subject to the same restrictions as in the case of the point-evaluation functional. Here, the target y_Ω has to be approximated by y uniformly, not in the quadratic mean.

Also in this case, $H^1(\Omega)$ is not the appropriate state space: the state y needs to be continuous for J to be defined. Moreover, while J is convex, it is not differentiable. In Section 6.2.1, this problem will be treated by transforming it into a problem with differentiable cost functional and pointwise state constraints.

2.14.2. Sobolev–Slobodetskii spaces. In this section, we briefly discuss Sobolev spaces for noninteger orders of differentiation. These spaces play an important role in the theory of partial differential equations. In this book, they will be used in Section 2.15 to prove that the optimal control for the problem of finding the optimal stationary boundary temperature belongs to

$H^1(\Gamma)$. For the next definition, we follow the lines of Thm. 7.48 in Adams [**Ada78**].

Definition. *Let $\Omega \subset \mathbb{R}^N$ be a bounded domain, and let $s > 0$ be noninteger, with $\lambda = s - [s] > 0$ being its noninteger part. Then we denote by $H^s(\Omega)$ the normed space of all functions $v \in H^{[s]}(\Omega)$ such that*

$$\sum_{|\alpha|=[s]} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{N+2\lambda}}\, dx\, dy < \infty,$$

endowed with the norm

$$\|v\|^2_{H^s(\Omega)} = \|v\|^2_{H^{[s]}(\Omega)} + \sum_{|\alpha|=[s]} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{N+2\lambda}}\, dx\, dy.$$

To define the spaces $H^s(\Gamma)$, we need the representations $y_N = h_i(y)$, for $y \in Q_{N-1}$, with respect to all the local coordinate systems S_i, which were used to introduce the notion of $C^{k,1}$ domains in Section 2.2.2. A function v belongs to $H^s(\Gamma)$ if and only if all the functions $v_i(y) = v(y, h_i(y))$ belong to the space $H^s(Q_{N-1})$.

A thorough treatment of the mathematical foundations is given in Wloka [**Wlo87**], Chap. 1, §4; we also refer the reader to Adams [**Ada78**] and Alt [**Alt99**]. In the same way, one can define the spaces $H^k(\Gamma)$ for integer k in $C^{k-1,1}$ domains, using the Sobolev space $H^k(Q_{N-1})$; see Gajewski et al. [**GGZ74**].

These so-called *Sobolev–Slobodetskii spaces* turn out to be Hilbert spaces if equipped with the corresponding scalar product. The spaces $W^{s,p}(\Omega)$, for $p \neq 2$, are introduced in a similar way; see [**Ada78**]. In particular, one has $H^s(\Omega) = W^{s,2}(\Omega)$.

2.14.3. Higher regularity of solutions. The examples discussed in Section 2.14.1 show that H^1 regularity of solutions to elliptic boundary value problems does not suffice for important classes of optimal control problems. One therefore tries to find additional conditions relating to the smoothness of the boundary and/or the prescribed data which guarantee better regularity properties. In this section, we collect some standard results from the relevant literature.

Boundary value problems in $C^{1,1}$ domains. We consider the Dirichlet problem

$$\begin{aligned} \mathcal{A}y + \lambda y &= f \quad \text{in } \Omega \\ y &= g \quad \text{on } \Gamma \end{aligned} \tag{2.107}$$

and the Neumann problem

$$\begin{aligned} \mathcal{A}y + \lambda y &= f \quad \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y &= g \quad \text{on } \Gamma, \end{aligned} \tag{2.108}$$

with the elliptic operator $\mathcal{A}$ defined in (2.19) on page 37. We assume that the coefficient functions a_{ij} obey both the symmetry condition and the ellipticity condition (2.20). In addition, we assume that $a_{ij} \in C^{0,1}(\bar{\Omega})$ for all $i, j \in \{1, \ldots, N\}$, and we postulate that Ω is a bounded $C^{1,1}$ domain. Moreover, $\lambda \in \mathbb{R}$ is prescribed.

Homogeneous boundary data. In this case, we can deduce the following result from Thms. 2.2.2.3 and 2.2.2.5 in Grisvard [**Gri85**]:

If $g = 0$, $\lambda \geq 0$, and $f \in L^2(\Omega)$, then the weak solution to the Dirichlet problem (2.107) *belongs to $H^2(\Omega)$. If, in addition, $\lambda > 0$, then the same result holds for the Neumann problem* (2.108)*.*

Inhomogeneous boundary data. The following result is a consequence of Thms. 2.4.2.5 and 2.4.2.7 in [**Gri85**]:

Let $1 < p < \infty$. If $g \in W^{2-1/p,p}(\Gamma)$, $\lambda \geq 0$, and $f \in L^p(\Omega)$, then the weak solution y to the Dirichlet problem (2.107) *belongs to $W^{2,p}(\Omega)$. If $\lambda > 0$ and $g \in W^{1-1/p,p}(\Gamma)$, then the same result holds for the Neumann problem* (2.108)*.*

The results cited above were proved in [**Gri85**] for more general boundary operators that include, in particular, the case of Robin boundary conditions. Note that the results for homogeneous boundary conditions follow as special cases from the $W^{2,p}$ results, since $g = 0$ is smooth.

Lipschitz domains. For Lipschitz domains, the following interesting result due to Jerison and Kenig [**JK81**] holds:

Let Ω be a Lipschitz domain. If $\mathcal{A} = -\Delta$, $\lambda > 0$, $f = 0$, and $g \in L^2(\Gamma)$, then the weak solution y to the Neumann problem belongs to $H^{3/2}(\Omega)$.

Convex domains. The following result shows that additional regularity can be expected for convex domains:

If Ω is a bounded and convex domain, then the results stated for $C^{1,1}$ domains and homogeneous boundary data remain valid; that is, if $g = 0$, $f \in L^2(\Omega)$, and $\lambda \geq 0$ or $\lambda > 0$, respectively, for the Dirichlet or Neumann problems, then $y \in H^2(\Omega)$.

This result follows from Thms. 3.2.1.2 and 3.2.1.3 in **[Gri85]**. Further regularity results for C^∞ boundaries can be found in Triebel **[Tri95]**. Moreover, the *Stampacchia technique* can be employed to prove boundedness or continuity of the solution y under slightly weaker conditions. We will come back to this in Sections 4.2 and 7.2.2.

2.15. Regularity of optimal controls

In the problems studied above, the controls were chosen from the Hilbert spaces $L^2(\Gamma)$ or $L^2(\Omega)$; hence, no better regularity than L^2 can at first glance be expected for the optimal controls. It turns out, however, that for $\lambda > 0$ the mere fact of optimality entails additional regularity.

Optimal stationary heat sources. We treat this problem for zero boundary temperature; the case of a boundary condition of the third kind with prescribed outside temperature can be handled analogously. The optimality system associated with the corresponding problem (2.26)–(2.28) (see page 49) reads

$$\begin{aligned} -\Delta y &= \beta\, u & \qquad -\Delta p &= y - y_\Omega \\ y|_\Gamma &= 0 & \qquad p|_\Gamma &= 0 \\ & u = \mathbb{P}_{[u_a,u_b]}\left\{-\lambda^{-1}\,\beta\, p\right\}. \end{aligned}$$

We have the following regularity result.

Theorem 2.37. *Suppose that β is Lipschitz continuous on $\bar\Omega$, and assume that u_a, $u_b \in H^1(\Omega)$ and $y_\Omega \in L^2(\Omega)$. Then the optimal control for the optimal stationary heat source problem* (2.26)–(2.28) *on page 49 belongs to $H^1(\Omega)$.*

Proof. Since $y - y_\Omega$ belongs to $L^2(\Omega)$, the solution to the adjoint equation satisfies $p \in H^1_0(\Omega)$. This holds for the product $\beta\, p$ only if β is sufficiently smooth. To ensure this, we have postulated that β is Lipschitz continuous on $\bar\Omega$. Hence, we have $\beta\, p \in H^1_0(\Omega)$; see Grisvard **[Gri85]**, Thm. 1.4.1.2. Moreover, u_a, $u_b \in H^1(\Omega)$.

Finally, we claim that the projection operator $\mathbb{P}_{[u_a,u_b]}$ maps $H^1(\Omega)$ into itself. Indeed, this is a consequence of a result due to Stampacchia and

Kinderlehrer [**KS80**], which states that the mapping $u(\cdot) \mapsto |u(\cdot)|$ maps $H^1(\Omega)$ continuously into itself. In conclusion, the optimal control $u = \bar{u}$ belongs to $H^1(\Omega)$. □

Optimal stationary boundary temperature. The optimality system for this problem reads

$$\begin{aligned} -\Delta y &= 0 & \qquad -\Delta p &= y - y_\Omega \\ \partial_\nu y + \alpha\, y &= \alpha\, u & \qquad \partial_\nu p + \alpha\, p &= 0 \end{aligned}$$

$$u = \mathbb{P}_{[u_a, u_b]} \left\{ -\lambda^{-1}\, \alpha\, p|_\Gamma \right\}.$$

In this case, the above method fails to yield the expected regularity. Again, the adjoint equation yields $p \in H^1(\Omega)$. However, the projection relation for u involves the trace of p, which by Theorem 7.3 merely belongs to $H^{1/2}(\Gamma)$. We can thus expect $u \in H^{1/2}(\Gamma)$ at best. Nevertheless, additional regularity can be recovered under natural conditions.

Theorem 2.38. *Let Ω be a bounded $C^{1,1}$ domain, let α be Lipschitz continuous, and let u_a, $u_b \in H^1(\Gamma)$. Then the solution $\bar{u}$ to the problem of the optimal stationary boundary control belongs to $H^1(\Gamma)$.*

Proof: We write the adjoint equation in the form

$$\begin{aligned} -\Delta p + p &= y - y_\Omega + p \\ \partial_\nu p &= -\alpha\, p. \end{aligned}$$

The solution $p = p_1 + p_2$ is composed of two summands p_1 and p_2. Here, p_1 is the solution to the boundary value problem $-\Delta p_1 + p_1 = f$, $\partial_\nu p_1 = 0$, with $f := y - y_\Omega + p$. The summand p_2 is the solution to $-\Delta p_2 + p_2 = 0$, $\partial_\nu p_2 = g$, with $g := -\alpha\, p$.

Since Ω is a domain of class $C^{1,1}$, we have $p_1 \in H^2(\Omega)$ owing to the regularity results for the homogeneous Neumann problem from [**Gri85**], which are collected in Section 2.14.3.

Also, p_2 belongs to $H^2(\Omega)$: indeed, we have $p \in H^1(\Omega)$, so that $p|_\Gamma \in H^{1/2}(\Gamma)$ by virtue of Theorem 7.3 on page 356. Then, owing to Thm. 1.4.1.2 in [**Gri85**], the product $g = -\alpha\, p|_\Gamma$ also belongs to $H^{1/2}(\Gamma)$. The claim now follows from the regularity result for the nonhomogeneous Neumann problem in $C^{1,1}$ domains stated in Section 2.14.3: we can infer that the mapping $g \mapsto p_2$ is continuous from $H^{1/2}(\Gamma) = W^{1-1/2,2}(\Gamma)$ into $W^{2,2}(\Omega) = H^2(\Omega)$.

In summary, $p \in H^2(\Omega)$. Hence, by virtue of Theorem 7.3, the trace of p belongs at least to $H^1(\Gamma)$. The assertion now follows from the projection

formula for $\bar{u}$ and the continuity of the mapping $u(\cdot) \mapsto |u(\cdot)|$ in $H^1(\Gamma)$, since the product of the Lipschitz function α and the $H^1(\Gamma)$ function $p|_\Gamma$ belongs to $H^1(\Gamma)$; see **[Gri85]**, Thm. 1.4.1.2. □

2.16. Exercises

2.1 We sketch the treatment of the nonlinear problem

$$\begin{aligned} &\min\ J(y,u) \\ &T(y,u) = 0, \qquad u \in U_{ad}. \end{aligned}$$

Here, in addition to the quantities J and U_{ad} defined on page 10, a continuously differentiable mapping $T : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is given. Suppose that the Jacobian matrix $D_y T(\bar{y}, \bar{u})$ is nonsingular at the optimal point $(\bar{y}, \bar{u})$. Then the solution z to the equation linearized at $(\bar{y}, \bar{u})$,

$$D_y T(\bar{y}, \bar{u})(z - \bar{y}) + D_u T(\bar{y}, \bar{u})(u - \bar{u}) = 0,$$

behaves, in a suitable neighborhood of $(\bar{y}, \bar{u})$ and up to an error of higher order than $\|u - \bar{u}\|$, like the solution y to the equation $T(y, u) = 0$. In comparison with the linear state equation (1.1) on page 10, $D_y T(\bar{y}, \bar{u})$ takes over the role of A and $D_u T(\bar{y}, \bar{u})$ that of B. It is therefore plausible to conjecture that the pair $(\bar{y}, \bar{u})$ satisfies the following optimality system in place of (1.9) on page 14:

$$\begin{aligned} &T(y,u) = 0, \qquad u \in U_{ad} \\ &D_y T(y,u)^\top p = \nabla_y J(y,u) \\ &\big(D_u T(y,u)^\top p + \nabla_u J(y,u)\,,\, v - u\big)_{\mathbb{R}^m} \geq 0 \quad \forall\, v \in U_{ad}. \end{aligned}$$

Use the implicit function theorem to prove this conjecture.

2.2 Show that the expressions $\|x\|_{C[a,b]}$ (maximum norm) and $\|x\|_{C_{L^2}[a,b]}$ introduced in Section 2.1 satisfy the norm axioms.

2.3 Let $\big\{H, (\cdot\,,\,\cdot)\big\}$ be a pre-Hilbert space. Show that $\|u\| := \sqrt{(u\,,\,u)}$ defines a norm on H.

2.4 Prove Theorem 2.7 on page 38.

2.5 Show that $\|A\|_{\mathcal{L}(U,V)} = \sup\limits_{\|u\|_U = 1} \|A\,u\|_V$ defines a norm in $\mathcal{L}(U, V)$.

2.6 Determine the operator norm of the integral operator $A : C[0,1] \to C[0,1]$,

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

2.7 Show that every strongly convergent sequence in a normed space converges weakly.

2.8 Prove that in Hilbert spaces $\big\{H, (\cdot\,,\,\cdot)\big\}$ the following important result holds: if $u_n \rightharpoonup u$ and $v_n \to v$, then $(u_n\,,\,v_n) \to (u\,,\,v)$ as $n \to \infty$.

2.9 Suppose that Assumption 2.13 on page 48 is fulfilled. Show that the set

$$\big\{u \in L^2(\Omega)\ :\ u_a(x) \leq u(x) \leq u_b(x)\ \text{ for a.e. } x \in \Omega\big\}$$

is convex and closed. Use the following well-known result: if $\|u_n\|_{L^2(\Omega)} \to 0$, then there exists a subsequence of $\{u_n\}_{n=1}^{\infty}$ that converges to zero almost everywhere in Ω.

2.10 Suppose that $\{Y, \|\cdot\|_Y\}$ and $\{U, \|\cdot\|_U\}$ are Hilbert spaces, and let $y_d \in Y$, $\lambda \geq 0$, and an operator $S \in \mathcal{L}(U, Y)$ be given. Show that the functional

$$f(u) = \|S\,u - y_d\|_Y^2 + \lambda\,\|u\|_U^2$$

is strictly convex if λ is positive or S is injective.

2.11 Show that the following functionals are continuously Fréchet differentiable:

a) $f(u) = \sin\big(u(1)\big)$, in $C[0,1]$;

b) $f(u) = \|u\|_H^2$, in any Hilbert space $\{H, (\cdot\,,\,\cdot)\}$.

2.12 Show that the linear integral operator

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

is well defined in $H = L^2(0,1)$ and maps H continuously into itself.

2.13 Let real numbers u_a, u_b, β, p and some $\lambda > 0$ be given. Solve the quadratic optimization problem in $\mathbb{R}$,

$$\min_{v \in [u_a, u_b]} \left\{ \beta\,p\,v + \frac{\lambda}{2}\,v^2 \right\},$$

by deriving a projection formula of the type (2.58) on page 70.

2.14 Prove the necessary optimality conditions for problem (2.69)–(2.71) on page 74.

2.15 Derive the necessary optimality conditions for the linear optimal control problem on page 79. *Hint:* Use the fact that the value of the cost functional at an arbitrary triple cannot be smaller than that at an optimal triple, and write down the differential equation satisfied by the difference between an arbitrary and an optimal triple.

2.16 Let a bounded Lipschitz domain $\Omega \subset \mathbb{R}^N$ and functions $y_\Omega \in L^2(\Omega)$, $e_\Omega \in L^2(\Omega)$, and $e_\Gamma \in L^2(\Gamma)$ be given, where e_Γ is assumed to be the trace of a function $y \in H^2(\Omega)$. Derive the necessary optimality conditions for the problem

$$\min \int_\Omega |y - y_\Omega|^2\,dx,$$

subject to $-\Delta y = u + e_\Omega$, $y|_\Gamma = e_\Gamma$, and the box constraints $-1 \leq u(x) \leq 1$.

2.17 Derive the necessary optimality conditions for the following problem (cf. page 82):

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

subject to $-\Delta y + y = u + e_\Omega$, $\partial_\nu y = e_\Gamma$, and the box constraints $0 \leq u(x) \leq 1$.

2.18 Prove Theorem 2.34 on page 90, that is, the first-order necessary optimality conditions for the problem (2.36)–(2.38) on page 54, which were derived only formally in Section 2.11.1 by using the formal Lagrange method.

2.19 Apply the formal Lagrange method to the problem with boundary control of Dirichlet type,

$$\min \int_\Omega |y - y_\Omega|^2 \, dx + \lambda \int_\Gamma |u|^2 \, ds,$$

subject to

$$\begin{aligned} -\Delta y &= 0 \\ y|_\Gamma &= u \end{aligned}$$

and $-1 \le u(x) \le 1$, and state the necessary optimality conditions that are to be expected. *Hint:* Use different multipliers p_1 and p_2 for the Laplace equation and the boundary condition, respectively.

Chapter 3

Linear-quadratic parabolic control problems

3.1. Introduction

Preliminary remarks.

Elliptic differential equations model stationary physical processes such as heat conduction processes with equilibrium temperature distributions. If the process under investigation is nonstationary, then *time* comes into play as an additional physical parameter. As an archetypical case, let us consider the problem of finding the optimal nonstationary boundary temperature discussed in Section 1.2.2: our task is to control the temperature in a spatial domain Ω, which is initially given by $y_0(x)$, $x \in \Omega$, in such a way that a desired final temperature distribution $y_\Omega(x)$, $x \in \Omega$, is achieved within a finite period of time $T > 0$. A simplified mathematical model for this problem reads as follows:

$$\text{(3.1)}\quad \min\ J(y,u) := \frac{1}{2}\int_\Omega \left|y(x,T) - y_\Omega(x)\right|^2 dx + \frac{\lambda}{2}\int_0^T\int_\Gamma |u(x,t)|^2\, ds(x)dt,$$

subject to

$$\text{(3.2)}\quad \boxed{\begin{array}{rcll} y_t - \Delta y &=& 0 & \text{in } Q := \Omega \times (0,T) \\ \partial_\nu y + \alpha\, y &=& \beta\, u & \text{on } \Sigma := \Gamma \times (0,T) \\ y(x,0) &=& y_0(x) & \text{in } \Omega \end{array}}$$

and

$$u_a(x,t) \le u(x,t) \le u_b(x,t) \quad \text{for a.e. } (x,t) \in \Sigma. \tag{3.3}$$

In contrast to elliptic problems, the process evolves within the space-time cylinder $Q := \Omega \times (0,T)$. The control function $u = u(x,t)$ acts on the spatial boundary Γ; it is therefore defined on the set $\Sigma := \Gamma \times (0,T)$.

In this chapter, we will pursue a similar strategy to that in the elliptic case. First, we will show that the initial-boundary value problem (3.2) admits for any given control $u = u(x,t)$ a unique solution $y = y(x,t)$ in a suitable function space. Then we will investigate the solvability of the optimal control problem, that is, whether there exists an optimal control $\bar{u}$ with associated optimal state $\bar{y}$. This will again follow from the continuity of the solution operator $G : u \mapsto y$. Finally, necessary optimality conditions will be derived.

From the viewpoint of optimization, this approach is in principle the same as in the elliptic case. However, the theory of weak solutions is slightly more involved for parabolic equations: in addition to second-order spatial derivatives, a first-order derivative of y with respect to the time variable also occurs. This makes the use of another solution space necessary, and eventually leads to the conclusion that the associated adjoint equation has to be taken as an equation that runs backwards in time.

Formal derivation of the optimality conditions. In order to get, from the very beginning, an idea of what sort of optimality conditions can be expected, we once more apply the formal Lagrange technique. It will turn out later that this approach actually leads to the correct result. To this end, we introduce the Lagrangian function $\mathcal{L}$ associated with the problem (3.1)–(3.3),

$$\mathcal{L}(y,u,p) = J(y,u) - \iint_Q (y_t - \Delta y)\, p_1 \, dx\, dt - \iint_\Sigma \big(\partial_\nu y + \alpha\, y - \beta\, u\big)\, p_2 \, ds\, dt,$$

in which we account for the "difficult" constraints involving derivatives. The initial condition and the inequality constraints for u are not eliminated by introducing corresponding Lagrange multipliers. We note that in this case it would suffice to choose the same function p in both integrals; however, since this leads to difficulties in other situations, we prefer to work with two different functions p_1 and p_2, yet eventually arrive at the conclusion that $p_1 = p_2$ on Σ. It is advisable to take such an approach whenever there is doubt.

In the following, we will often write $y(\cdot,t)$ or, for short, $y(t)$, and we will regard y as a function with values in a Banach space (for the explanation of this notion, see page 141). The first form $y(\cdot,t)$ stresses the dependence on x.

The *set of admissible controls* is defined by

$$U_{ad} = \big\{ u \in L^2(\Sigma) \,:\, u_a(x,t) \le u(x,t) \le u_b(x,t) \quad \text{for a.e. } (x,t) \in \Sigma \big\}.$$

The initial condition $y(\cdot,0) = y_0$ for y has to be taken into account. Therefore, the Lagrange method at first yields the variational inequality

$$D_y \mathcal{L}(\bar{y}, \bar{u}, p)(y - \bar{y}) \ge 0$$

for all sufficiently smooth functions y with $y(\cdot,0) = y_0$. Substituting $y := y - \bar{y}$, we find that $D_y \mathcal{L}(\bar{y}, \bar{u}, p) y \ge 0$ for all y satisfying $y(\cdot,0) = 0$. Since $-y$ also belongs to this class of functions, we finally obtain that $D_y \mathcal{L}(\bar{y}, \bar{u}, p) y = 0$. With respect to u, the known variational inequality follows. In conclusion, we expect the following necessary optimality conditions:

$$\boxed{\begin{array}{rcll} D_y \mathcal{L}(\bar{y}, \bar{u}, p)\, y & = & 0 & \text{for all } y \text{ with } y(0) = 0 \\ D_u \mathcal{L}(\bar{y}, \bar{u}, p)\, (u - \bar{u}) & \ge & 0 & \text{for all } u \in U_{ad}. \end{array}}$$

The first equation leads to the adjoint equation. Since the derivative of the linear and continuous mapping $y \mapsto y(\cdot, T)$ coincides with the mapping itself, we find that

$$\begin{aligned} D_y \mathcal{L}(\bar{y}, \bar{u}, p)\, y \;=\; & \int_\Omega \big(\bar{y}(T) - y_\Omega\big)\, y(T)\, dx - \iint_Q \big(y_t - \Delta y\big)\, p_1 \, dx\, dt \\ & - \iint_\Sigma \big(\partial_\nu y + \alpha y\big)\, p_2 \, ds\, dt. \end{aligned}$$

Upon integrating by parts (with respect to t in y_t and with respect to x in Δy), invoking Green's formula, and finally regrouping the terms, we obtain that any sufficiently smooth y with $y(0) = 0$ satisfies

$$\begin{aligned} 0 \;=\; & \int_\Omega \big(\bar{y}(T) - y_\Omega\big)\, y(T)\, dx - \int_\Omega y(T)\, p_1(T)\, dx + \iint_Q y\, p_{1,t}\, dx\, dt \\ & + \iint_\Sigma p_1\, \partial_\nu y \, ds\, dt - \iint_\Sigma y\, \partial_\nu p_1 \, ds\, dt + \iint_Q y\, \Delta p_1 \, dx\, dt \\ & - \iint_\Sigma p_2\, \partial_\nu y \, ds\, dt - \iint_\Sigma \alpha\, y\, p_2 \, ds\, dt \\ = \; & \int_\Omega \big(\bar{y}(T) - y_\Omega - p_1(T)\big)\, y(T)\, dx + \iint_Q \big(p_{1,t} + \Delta p_1\big)\, y \, dx\, dt \\ & - \iint_\Sigma \big(\partial_\nu p_1 + \alpha\, p_2\big)\, y \, ds\, dt + \iint_\Sigma \big(p_1 - p_2\big)\, \partial_\nu y \, ds\, dt. \end{aligned}$$

Note that the term containing $y(0)$ vanishes, since $y(0) = 0$.

First, we note that for all $y \in C_0^\infty(Q)$ the expressions $y(T)$, $y(0)$ and y, $\partial_\nu y$ vanish on Ω and Σ, respectively. Therefore,

$$\iint_Q \big(p_{1,t} + \Delta p_1\big)\, y \; dx\, dt = 0 \qquad \forall\, y \in C_0^\infty(Q).$$

Now observe that $C_0^\infty(Q)$ is dense in $L^2(Q)$. We therefore must have

$$p_{1,t} + \Delta p_1 = 0 \quad \text{in } Q.$$

In particular, the integral over Q vanishes in the above equation. Next, we no longer require that $y(T) = 0$ and consider the set of all functions $y \in C^1(\bar{Q})$ such that $y|_\Sigma = 0$. For such functions, it follows that

$$\int_\Omega \big(\bar{y}(T) - y_\Omega - p_1(T)\big)\, y(T)\, dx = 0.$$

The possible values $y(T)$ form a dense subset of $L^2(\Omega)$. We do not discuss the validity of this claim here, since we are arguing formally; similarly, we will claim below that the boundary values and normal derivatives of smooth functions y form dense subsets in $L^2(\Sigma)$. If, however, the claim is true, then it follows that

$$p_1(T) = \bar{y}(T) - y_\Omega \quad \text{in } \Omega.$$

Next, we no longer require $y|_\Sigma = 0$ and vary over all functions $y \in C^1(\bar{Q})$. We obtain

$$\iint_\Sigma (\partial_\nu p_1 + \alpha\, p_2)\, y\, ds\, dt = 0.$$

Claiming that the set $\big\{y|_\Sigma :\ y \in C^1(\bar{Q})\big\}$ is dense in $L^2(\Sigma)$, we conclude that

$$\partial_\nu p_1 + \alpha\, p_2 = 0 \quad \text{in } \Sigma.$$

In summary, we have

$$\iint_\Sigma \big(p_1 - p_2\big)\, \partial_\nu y\, ds\, dt = 0$$

for all sufficiently smooth y. Claiming that the set of normal derivatives $\partial_\nu y$ is dense in $L^2(\Sigma)$, we finally find that $p_2 = p_1$ on Σ. We thus put $p := p_1$ to obtain that $p_2 = p$ on Σ. In conclusion, our formal argument yields the following system as the *adjoint equation*:

(3.4)

$$\begin{array}{rcll} -p_t & = & \Delta p & \text{in } Q \\ \partial_\nu p + \alpha\, p & = & 0 & \text{on } \Sigma \\ p(T) & = & \bar{y}(T) - y_\Omega & \text{in } \Omega. \end{array}$$

Evaluation of the variational inequality for $D_u \mathcal{L}$ yields

$$\begin{aligned} D_u\mathcal{L}(\bar{y},\bar{u},p)(u-\bar{u}) &= \lambda \iint_\Sigma \bar{u}\,(u-\bar{u})\,ds\,dt + \iint_\Sigma \beta\,p\,(u-\bar{u})\,ds\,dt \\ &= \iint_\Sigma (\lambda\,\bar{u} + \beta\,p)(u-\bar{u})\,ds\,dt \geq 0. \end{aligned}$$

Hence, the following *variational inequality* has to be satisfied:

$$\boxed{\iint_\Sigma (\lambda\,\bar{u} + \beta\,p)(u-\bar{u})\;ds\,dt \geq 0 \qquad \forall\, u \in U_{ad}.} \tag{3.5}$$

The above derivation was only formal, with not much attention paid to mathematical rigor. For instance, in calculations we treated the time derivatives y_t and p_t as if they were ordinary functions; also, we have not specified the function spaces to which y and p, as well as their derivatives, are supposed to belong. Moreover, the careless use of the initial and final values of y and p was rather bold. Finally, the claimed density of the boundary values of y and of $\partial_\nu y$ does not always hold, unless precise smoothness assumptions are imposed on the boundary of Ω. Consequently, we can for the time being only guess that our result is the correct one. (Note, however, that later in this chapter a mathematically rigorous derivation of the optimality conditions will still be given.) Nevertheless, a careful application of the Lagrangian function yields, in any case, a convenient formulation of these conditions that is also easy to memorize.

Recommended course of study of the upcoming sections. In this chapter, we will introduce the notion of weak solutions to linear parabolic equations, prove existence and uniqueness of such solutions, and then investigate the questions originating from optimization theory.

At first, however, the spatially one-dimensional parabolic case will be treated using the Fourier technique. This method does not need the theory of weak solutions to parabolic equations as a prerequisite; it therefore benefits those readers who might prefer to defer the study of this theory until later. The Fourier method is also interesting in itself, since it is equivalent to the theory of strongly continuous semigroups. In addition, in this comparatively elementary way one can deduce the well-known bang-bang principle for optimal controls in the case of a pure final-value functional.

Readers who want to acquaint themselves immediately with the theory of weak solutions may omit Section 3.2 for the time being and continue with Section 3.3, for which Section 3.2 is not required.

3.2. Fourier's method in the spatially one-dimensional case

3.2.1. One-dimensional model problems.

A boundary control problem. To provide some physical background, we once more interpret the following control problems as heating problems. We consider the problem of heating the one-dimensional spatial domain $\Omega = (0,1)$ in an optimal way by means of a control $u = u(t)$ that acts at the right boundary point $x = 1$:

$$\min\ J(y,u) := \frac{1}{2}\int_0^1 \big|y(x,T) - y_\Omega(x)\big|^2\,dx + \frac{\lambda}{2}\int_0^T |u(t)|^2\,dt, \tag{3.6}$$

subject to

$$\boxed{\begin{array}{rcll} y_t(x,t) &=& y_{xx}(x,t) & \text{in } (0,1)\times(0,T)\\ y_x(0,t) &=& 0 & \text{in } (0,T)\\ y_x(1,t) &=& \beta\,u(t) - \alpha\,y(1,t) & \text{in } (0,T)\\ y(x,0) &=& 0 & \text{in } (0,1), \end{array}} \tag{3.7}$$

and the control constraints

$$u_a(t) \le u(t) \le u_b(t) \quad \text{for a.e. } t \in (0,T). \tag{3.8}$$

Assumption 3.1. *Assume that we are given real numbers $T > 0$ (the period of heating), $\alpha \ge 0$ (heat transmission coefficient), and $\lambda \ge 0$, as well as functions $\beta \in L^\infty(0,T)$ with $\beta(t) \ge 0$ for almost every $t \in (0,T)$, $y_\Omega \in L^2(0,1)$ (the desired final temperature distribution), and u_a, $u_b \in L^2(0,T)$ such that $u_a(t) \le u_b(t)$ for almost every $t \in (0,T)$.*

The control u that we seek is supposed to belong to $L^2(0,T)$. Note that any admissible control u is automatically essentially bounded if u_a and u_b are bounded and measurable.

From the physical viewpoint, we ought to have $\beta = \alpha$, since the physically correct form of the right-hand boundary condition is given by $y_x(1,t) = \alpha\big(u(t) - y(1,t)\big)$ (the heat flux $y_x(1,t)$ at the boundary is proportional to the difference between the outside temperature $u(t)$ and the boundary temperature $y(1,t)$). However, the above form of the boundary condition allows for admitting a Neumann boundary condition (by choosing $\alpha = 0$). It also seems appropriate for purely mathematical reasons to decouple α and β.

Remarks. The above problem is merely academic from the physical viewpoint. Indeed, it is rather unrealistic to consider a one-dimensional spatial domain Ω, i.e.,

a real interval. One could imagine the heating of a very thin beam of unit length which, except for the right endpoint $x = 1$, is completely isolated. The boundary condition $y_x = 0$ at the left endpoint $x = 0$ models isolation; it may also model a symmetry condition if the beam has length 2 (with left endpoint at $x = -1$ and right endpoint at $x = 1$) and the heating occurs with the same outside temperature $u(t)$ at both endpoints.

A somewhat more realistic interpretation is the heating of an infinitely long plate of unit thickness (with both plane surfaces assumed to be orthogonal to the x-axis) whose right surface is heated by $u(t)$ while the left surface is isolated. Analogously, we can think of a plate of thickness 2, where both left and right outside temperatures are given by $u(t)$.

Recall, however, that our main concern in this book is not the description of heat conduction processes; instead, we intend to explain the basic ideas governing optimal control problems. The easiest way to do this is to consider simple toy models. Also, the assumption of a homogeneous boundary temperature is due only to methodological considerations.

A problem involving a controllable heat source. Another physically meaningful situation is the control of a heat source distributed in the spatial domain. A typical example is the heating of metals by electromagnetic induction. For a change, we consider a different cost functional. Here, the aim is to follow a desired nonstationary temperature evolution $y_Q(x,t)$ in $Q = (0,1) \times (0,T)$, which is modeled in the following way:

(3.9)
$$\min J(y,u) := \frac{1}{2}\int_0^T \int_0^1 \left|y(x,t) - y_Q(x,t)\right|^2 dx\,dt + \frac{\lambda}{2}\int_0^T \int_0^1 |u(x,t)|^2\,dx\,dt,$$

subject to

(3.10)
$$\boxed{\begin{array}{rcll} y_t(x,t) &=& y_{xx}(x,t) + u(x,t) & \text{in } (0,1)\times(0,T) \\ y_x(0,t) &=& 0 & \text{in } (0,T) \\ y_x(1,t) + \alpha\, y(1,t) &=& 0 & \text{in } (0,T) \\ y(x,0) &=& 0 & \text{in } (0,1) \end{array}}$$

and

(3.11)
$$u_a(x,t) \le u(x,t) \le u_b(x,t) \quad \text{for a.e. } (x,t) \in Q.$$

Here, homogeneous Neumann boundary conditions are prescribed (to model isolation). Analogously, a fixed outside temperature could have been prescribed in the form of a boundary condition of the third kind. The given functions y_Q and $u_a \le u_b$ are assumed to belong to $L^2(Q)$.

3.2.2. Integral representation of solutions—Green's function. Under suitable assumptions, the solutions to linear parabolic initial-boundary value problems can be expressed by means of Fourier series, which are obtained by separation of variables. This theory is comparatively simple, since it does not need the theory of weak solutions. Unfortunately, its applicability is limited to geometrically simple spatial domains (e.g., in the case of $N > 1$ to cubes or balls; see, however, Glashoff and Weck [**GW76**]).

To fix things, consider the initial-boundary value problem for the one-dimensional heat equation

(3.12)
$$\boxed{\begin{aligned} y_t(x,t) - y_{xx}(x,t) &= f(x,t) \\ y_x(0,t) &= 0 \\ y_x(1,t) + \alpha\, y(1,t) &= u(t) \\ y(x,0) &= y_0(x) \end{aligned}}$$

in $Q = (0,1) \times (0,T)$, where $f \in L^2(Q)$, $y_0 \in L^2(0,1)$, $u \in L^2(0,T)$, and the constant $\alpha \geq 0$ are given. If the data f, y_0, and u are sufficiently smooth, then there exists a unique *classical* solution y, which can be expressed by means of a *Green's function* $G = G(x,\xi,t)$ in the form

(3.13)
$$\boxed{\begin{aligned} y(x,t) = &\int_0^1 G(x,\xi,t)\, y_0(\xi)\, d\xi + \int_0^t \int_0^1 G(x,\xi,t-s)\, f(\xi,s)\, d\xi\, ds \\ &+ \int_0^t G(x,1,t-s)\, u(s)\, ds. \end{aligned}}$$

In the physically interesting cases for the constant α, G has the form of a Fourier series:

(3.14)
$$G(x,\xi,t) = \begin{cases} 1 + 2\displaystyle\sum_{n=1}^{\infty} \cos(n\pi x)\cos(n\pi\xi)\exp(-n^2\pi^2 t) & \text{for } \alpha = 0 \\ \displaystyle\sum_{n=1}^{\infty} \frac{1}{N_n}\cos(\mu_n x)\cos(\mu_n \xi)\exp(-\mu_n^2 t) & \text{for } \alpha > 0. \end{cases}$$

Here, the $\mu_n \geq 0$ denote the solutions to the equation $\mu \tan \mu = \alpha$, ordered according to increasing magnitude, while $N_n = 1/2 + \sin(2\mu_n)/(4\mu_n)$ are normalizing factors. The numbers $n\,\pi$ and μ_n are the *eigenvalues* of the differential operator $\partial^2/\partial x^2$ combined with the homogeneous version of

the boundary conditions formulated in (3.12); the functions $\cos(n\pi x)$ and $\cos(\mu_n x)$ are the corresponding *eigenfunctions*. The above series expansions will be derived in Section 3.8. We also refer to Tychonov and Samarski [**TS64**].

The Green's function is nonnegative and symmetric with respect to the variables x and ξ. For $x = \xi$, it becomes singular at $t = 0$; more precisely, G has a so-called *weak singularity* at $(x = \xi, t = 0)$. Moreover, we have $y \in L^2(Q)$ if $f \in L^2(Q)$, $y_0 \in L^2(0,1)$, and $u \in L^2(0,T)$. The corresponding estimation can be found in [**Trö84b**], where estimates for the Green's function from Friedman [**Fri64**] are employed.

Definition. *We call the function y given by* (3.13) *with square integrable functions f, u, and y_0 the* generalized solution *to* (3.12).

Each of the three summands in (3.13) defines a linear operator. The following two cases are interesting for the discussion of the initial-boundary value problem (3.12):

(i) $u := \beta u$, $f = y_0 = 0$ (boundary control):
In this case, only the final distribution $y(x,T)$ occurs in the cost functional (3.9), and (3.13) simplifies to

$$y(x,T) = \int_0^T G(x,1,T-s)\,\beta(s)\,u(s)\,ds =: \big(Su\big)(x). \tag{3.15}$$

The integral operator S represents the part of the state y that appears in the cost functional, here the final value $y(T)$. The operator S maps $L^2(0,T)$ continuously into $L^2(0,1)$; see [**Trö84b**]. We therefore consider S as a mapping between these spaces,

$$S : L^2(0,T) \to L^2(0,1).$$

The above property also follows from the equivalence with weak solutions and their properties; see Theorem 3.13 on page 150.

(ii) $u = y_0 = 0$ (distributed control):
This case arises from the control of the heat source. From (3.13), we deduce the representation

$$y(x,t) = \int_0^t \int_0^1 G(x,\xi,t-s)\,f(\xi,s)\,d\xi\,ds =: \big(\mathbf{S}\,f\big)(x,t). \tag{3.16}$$

$\mathbf{S}$ is a linear and continuous mapping from the space $L^2(Q) = L^2((0,1)\times(0,T))$ into itself. In the following, we consider $\mathbf{S}$ as an operator in $L^2(Q)$, even though it is actually continuous as a linear mapping from $L^2(Q)$ into

$C([0,T],L^2(0,1))$ (see **[Trö84b]**). The latter is a space of functions with values in a Banach space, which will be introduced in Section 3.4.1 on page 142. It expresses a certain continuity with respect to the variable t, which, in particular, ensures the convergence $y(x,t) \to 0$ as $t \downarrow 0$. We conclude that

$$\mathbf{S} : L^2(Q) \to L^2(Q).$$

These properties may also be deduced from Theorem 3.13 on page 150 concerning weak solutions.

3.2.3. Necessary optimality conditions. With the aid of the integral representations derived above, the spatially one-dimensional parabolic optimal control problems in Section 3.2.1 can be readily treated theoretically.

Boundary control problem. We first investigate problem (3.6)–(3.8), which can be rewritten in the form

$$\min J(y,u) := \frac{1}{2}\,\|y(T) - y_\Omega\|^2_{L^2(0,1)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(0,T)},$$

subject to $u \in U_{ad}$ and

$$\begin{aligned} y_t(x,t) &= y_{xx}(x,t) \\ y_x(0,t) &= 0 \\ y_x(1,t) + \alpha\, y(1,t) &= \beta(t)\, u(t) \\ y(x,0) &= 0, \end{aligned}$$

where $U_{ad} = \big\{u \in L^2(0,T) \,:\, u_a(t) \le u(t) \le u_b(t) \ \text{ for a.e. } t \in (0,T)\big\}$.

Substituting the integral representation (3.15) for $y(x,T)$, i.e., $S\,u = y(\cdot,T)$, in the cost functional, we obtain a quadratic optimization problem in a Hilbert space, namely

$$\min_{u\in U_{ad}} f(u) := \frac{1}{2}\,\|S\,u - y_\Omega\|^2_{L^2(0,1)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(0,T)}.$$

Since $S : L^2(0,T) \to L^2(0,1)$ is continuous, we infer from Theorem 2.14 on page 50 the existence of an optimal control $\bar u$ for the boundary control problem (3.6)–(3.8), which is unique if $\lambda > 0$. Let $\bar y$ denote the associated generalized solution. In view of Theorem 2.22 on page 64, the necessary and sufficient optimality condition is given by the variational inequality (2.45) on page 64, which in the present case takes the form

$$\big(S^*(S\,\bar u - y_\Omega) + \lambda\,\bar u\,,\, u - \bar u\big)_{L^2(0,T)} \ge 0 \quad \forall u \in U_{ad}. \tag{3.17}$$

The variational inequality contains the adjoint operator S^*. From our experience with elliptic problems, we expect it to be closely connected to an

adjoint differential equation. We now determine the form of the adjoint S^* of the integral operator S, defined through the relation

$$\big(v\,,\,S\,u\big)_{L^2(0,1)} = \big(S^*v\,,\,u\big)_{L^2(0,T)} \qquad \forall\, u \in L^2(0,T),\ \ \forall\, v \in L^2(0,1).$$

Fixing arbitrary elements $u \in L^2(0,T)$ and $v \in L^2(0,1)$ and invoking Fubini's theorem, we find that

$$\begin{aligned}
\big(v\,,\,Su\big)_{L^2(0,1)} &= \int_0^1 v(x)\,\Big(\int_0^T G(x,1,T-s)\,\beta(s)\,u(s)\,ds\Big)\,dx \\
&= \int_0^1 \int_0^T u(s)\,\beta(s)\,G(x,1,T-s)\,v(x)\,ds\,dx \\
&= \int_0^T u(s)\,\Big(\beta(s)\,\int_0^1 G(x,1,T-s)\,v(x)\,dx\Big)ds \\
&= \big(u\,,\,S^*v\big)_{L^2(0,T)} = \big(S^*v\,,\,u\big)_{L^2(0,T)}.
\end{aligned}$$

We thus have

$$\big(S^*v\big)(t) = \beta(t)\,\int_0^1 G(\xi,1,T-t)\,v(\xi)\,d\xi. \tag{3.18}$$

Lemma 3.2. *$\big(S^*v\big)(t) = \beta(t)\,p(1,t)$, where p is the generalized solution to the parabolic final value problem*

$$\begin{aligned}
-p_t(x,t) &= p_{xx}(x,t) \\
p_x(0,t) &= 0 \\
p_x(1,t) + \alpha\,p(1,t) &= 0 \\
p(x,T) &= v(x).
\end{aligned}$$

Proof: Owing to the symmetry of the Green's function, it follows from equation (3.18) that

$$\big(S^*v\big)(t) = \beta(t)\,\int_0^1 G(1,\xi,T-t)\,v(\xi)\,d\xi.$$

We now introduce the functions

$$p(x,t) = \int_0^1 G(x,\xi,T-t)\,v(\xi)\,d\xi$$

and, with the time transformation $\tau = T - t$,

$$\tilde{p}(x,\tau) = p(x,t) = p(x,T-\tau) = \int_0^1 G(x,\xi,\tau)\,v(\xi)\,d\xi.$$

Recalling (3.13), we see that $\tilde{p}$ solves the initial-boundary value problem

$$\begin{aligned} \tilde{p}_\tau(x,t) &= \tilde{p}_{xx}(x,t) \\ \tilde{p}_x(0,\tau) &= 0 \\ \tilde{p}_x(1,\tau) + \alpha\,\tilde{p}(1,\tau) &= 0 \\ \tilde{p}(x,0) &= v(x). \end{aligned}$$

The assertion now follows from the substitution $\tilde{p}(x,\tau) = p(x,t)$, using the fact that

$$D_\tau\,\tilde{p}(x,\tau) = D_\tau p(x,T-\tau) = -D_t\,p(x,t).$$

□

Remark. The equation for p runs backwards in time. It is, however, well posed, since a final condition is prescribed and not an initial condition; if an initial condition were prescribed instead, we would have the typical case of an ill-posed backwards parabolic equation known from the theory of inverse problems.

We are now in a position to state the necessary (and, owing to the convexity, also sufficient) optimality conditions.

Theorem 3.3. *A control $\bar{u} \in U_{ad}$ with associated state $\bar{y}$ is optimal for the one-dimensional boundary control problem* (3.6)–(3.8) *if and only if it satisfies the variational inequality*

$$\int_0^T \big(\beta(t)\,p(1,t) + \lambda\,\bar{u}(t)\big)\big(u(t) - \bar{u}(t)\big)\,dt \geq 0 \quad \forall u \in U_{ad}, \tag{3.19}$$

where $p \in L^2(Q)$ is the generalized solution to the adjoint equation

$$\begin{aligned} -p_t(x,t) &= p_{xx}(x,t) \\ p_x(0,t) &= 0 \\ p_x(1,t) + \alpha\,p(1,t) &= 0 \\ p(x,T) &= \bar{y}(x,T) - y_\Omega(x). \end{aligned} \tag{3.20}$$

Proof: The assertion follows directly from inserting the representation of S^* from Lemma 3.2, with $v := \bar{y}(T) - y_\Omega$, into the variational inequality (3.17). □

The function p is called the *adjoint state* associated with $(\bar{u}, \bar{y})$. In analogy to Lemma 2.26 on page 69, we have the following result.

Lemma 3.4. *The variational inequality* (3.19) *holds if and only if for almost every* $t \in [0,T]$ *the following variational inequality in* $\mathbb{R}$ *is satisfied:*

$$\big(\beta(t)\,p(1,t) + \lambda\,\bar{u}(t)\big)\big(v - \bar{u}(t)\big) \geq 0 \qquad \forall\, v \in [u_a(t), u_b(t)]. \tag{3.21}$$

Conclusion. *If* $\lambda > 0$*, then for almost every* $t \in (0,T)$ *the optimal control satisfies the projection relation*

$$\bar{u}(t) = \mathbb{P}_{[u_a(t),u_b(t)]}\Big\{-\frac{\beta(t)}{\lambda}p(1,t)\Big\}.$$

If $\lambda = 0$*, then for all points* t *such that* $\beta(t)\,p(1,t) \neq 0$*,* $\bar{u}(t)$ *is determined by*

$$\bar{u}(t) = \begin{cases} u_a(t) & \text{if} \quad \beta(t)\,p(1,t) > 0 \\ u_b(t) & \text{if} \quad \beta(t)\,p(1,t) < 0. \end{cases}$$

The proof of this result proceeds similarly to that in the elliptic case; see page 70.

Distributed control. The problem (3.9)–(3.11) on page 125 can be treated similarly. The existence of an optimal control $\bar{u}$ follows from Theorem 2.14 on page 50. Since in this case S is injective, we have uniqueness. For application of the necessary optimality condition (2.45) on page 64, we again have to determine the adjoint operator $\mathbf{S}^*$ in $L^2(Q)$. Proceeding as in the case of the boundary control, we find that

$$\begin{aligned} \big(v\,,\,\mathbf{S}u\big)_{L^2(Q)} &= \int_0^T \int_0^1 \Big(\int_0^t \int_0^1 u(\xi,s)\,G(x,\xi,t-s)\,v(x,t)\,d\xi\,ds \Big)\,dx\,dt \\ &= \int_0^T \int_0^1 u(x,t)\,\Big(\int_t^T \int_0^1 G(\xi,x,s-t)\,v(\xi,s)\,d\xi\,ds \Big)\,dx\,dt \\ &= \big(u\,,\,\mathbf{S}^*v\big)_{L^2(Q)}, \end{aligned}$$

so that

$$\big(\mathbf{S}^*v\big)(x,t) = \int_t^T \int_0^1 G(\xi,x,s-t)\,v(\xi,s)\,d\xi\,ds.$$

Lemma 3.5. *The function*

$$p(x,t) = \int_t^T \int_0^1 G(\xi,x,s-t)\,v(\xi,s)\,d\xi\,ds$$

is the generalized solution to

$$\begin{aligned} -p_t(x,t) &= p_{xx}(x,t) + v(x,t) \\ p_x(0,t) &= 0 \\ p_x(1,t) + \alpha\, p(1,t) &= 0 \\ p(x,T) &= 0. \end{aligned}$$

Proof: We use the Green's function (3.14) and substitute $\tau = T - t$ and $\sigma = T - s$. Then the integration variable σ runs from $T - t = \tau$ to 0. In addition, $d\sigma = -ds$. We thus obtain

$$\begin{aligned} p(x, T-\tau) &= -\int_\tau^0 \int_0^1 G(\xi, x, \tau - \sigma)\, v(\xi, T-\sigma)\, d\xi\, d\sigma \\ &= \int_0^\tau \int_0^1 G(x, \xi, \tau - \sigma)\, v(\xi, T-\sigma)\, d\xi\, d\sigma =: \tilde{p}(x,\tau). \end{aligned}$$

In view of equation (3.13), $\tilde{p}$ is the generalized solution to the (forward) initial value problem $\tilde{p}_\tau = \tilde{p}_{xx} + v(x, T-\tau)$, $\tilde{p}(x,0) = 0$, with homogeneous boundary conditions. Since $\tilde{p}_\tau = -p_t(x, T-\tau) = -p_t(x,t)$, the assertion of the lemma follows. □

Summarizing, we have proved the following necessary optimality conditions:

Theorem 3.6. *A control $\bar{u} \in U_{ad}$ with associated state $\bar{y}$ is optimal for the optimal heat source problem* (3.9)–(3.11) *on page 125 if and only if with the solution $p \in L^2(Q)$ to the adjoint equation*

$$\begin{aligned} -p_t(x,t) &= p_{xx}(x,t) + \bar{y}(x,t) - y_Q(x,t) \\ p_x(0,t) &= 0 \\ p_x(1,t) + \alpha\, p(1,t) &= 0 \\ p(x,T) &= 0 \end{aligned}$$

we have the variational inequality

$$\iint_Q (p + \lambda\, \bar{u})\, (u - \bar{u})\, dx\, dt \geq 0 \quad \forall u \in U_{ad}$$

or, equivalently, the pointwise relation

$$\big(p(x,t) + \lambda\, \bar{u}(x,t)\big)\big(v - \bar{u}(x,t)\big) \geq 0 \qquad \forall\, v \in [u_a(x,t), u_b(x,t)]$$

for almost every $(x,t) \in Q$.

As in the elliptic case, explicit expressions for $\bar{u}$ can be derived from the above theorem: if $\lambda > 0$, then the optimal control has to obey the projection

formula

$$\bar{u}(x,t) = \mathbb{P}_{[u_a(x,t),u_b(x,t)]}\Big\{-\frac{1}{\lambda}p(x,t)\Big\} \quad \text{for a.e. } (x,t)\in Q,$$

while for $\lambda = 0$ we have, almost everywhere in Q,

$$\bar{u}(x,t) = \begin{cases} u_a(x,t) & \text{if} \quad p(x,t) > 0 \\ u_b(x,t) & \text{if} \quad p(x,t) < 0. \end{cases} \tag{3.22}$$

3.2.4. The bang-bang principle. We now reconsider the optimal boundary control problem (3.6)–(3.8), but this time without regularization, that is, for $\lambda = 0$. It turns out that the missing regularizing term is reflected in a lower regularity of the optimal control. For simplicity, we choose $u_a = -1$, $u_b = 1$, and $\beta(t) \equiv 1$; in addition, let $\alpha \geq 0$ and $y_\Omega \in L(0,1)$ be given. We consider the problem

$$\min \ \frac{1}{2}\int_0^1 \big|y(x,T) - y_\Omega(x)\big|^2\,dx,$$

subject to

$$\begin{array}{rcll} y_t(x,t) & = & y_{xx}(x,t) & \text{in} \quad (0,1)\times(0,T) \\ y_x(0,t) & = & 0 & \text{in} \quad (0,T) \\ y_x(1,t) & = & u(t) - \alpha\, y(1,t) & \text{in} \quad (0,T) \\ y(x,0) & = & 0 & \text{in} \quad (0,1) \end{array}$$

and

$$|u(t)| \leq 1 \qquad \text{in } (0,T).$$

Owing to (3.22), we have

$$\bar{u}(t) = \begin{cases} 1 & \text{if } p(1,t) < 0 \\ -1 & \text{if } p(1,t) > 0. \end{cases}$$

This relation does not yield any information for points t where $p(1,t) = 0$. However, it turns out that for problems of the above type this can only occur at isolated points if the optimal value is positive.

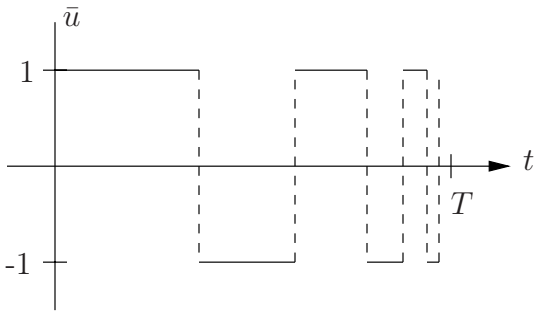
Theorem 3.7 (Bang-bang principle)**.** *Let $\bar{u}$ be an optimal control of the boundary control problem* (3.6)–(3.8) *on page 124, where $\lambda = 0$, $u_a = -1$, $u_b = 1$, and $\beta = 1$. If*

$$\|\bar{y}(\cdot,T) - y_\Omega\|^2_{L^2(0,1)} > 0,$$

then the function $p(1,t)$ has at most countably many zeros that can accumulate only at $t = T$. We thus have

$$|\bar{u}(t)| = 1 \quad \textit{for a.e. } t\in(0,T),$$

and $\bar{u}$ is piecewise equal to ± 1, with at most countably many switching points at the zeros of $p(1,t)$.



Bang-bang control.

Proof: We assume $\alpha > 0$, so that the second expression for G in (3.14) applies. The case where $\alpha = 0$ can be treated by analogous reasoning.

Let $d := \bar{y}(\cdot, T) - y_\Omega$. Then, by assumption, $\|d\|_{L^2(0,1)} \neq 0$. In view of (3.18) and Theorem 3.3, for $x = 1$ and $0 \leq t < T$ the adjoint state $\bar{p}$ has the form

$$\begin{aligned} p(1,t) &= \big(S^* d\big)(1,t) = \int_0^1 G(\xi, 1, T-t)\, d(\xi)\, d\xi \\ &= \sum_{n=1}^{\infty} \tfrac{1}{\sqrt{N_n}} \cos(\mu_n)\, \exp(-\mu_n^2 (T-t)) \underbrace{\int_0^1 \frac{1}{\sqrt{N_n}} \cos(\mu_n \xi)\, d(\xi)\, d\xi}_{=d_n}. \end{aligned}$$

Here, d_n denotes the nth Fourier coefficient of d with respect to the orthonormal system formed by the eigenfunctions $\frac{1}{\sqrt{N_n}} \cos(\mu_n x)$; see page 127. Owing to Bessel's inequality, the sequence $\{d_n\}_{n=1}^{\infty}$ is square summable. Moreover, the μ_n behave asymptotically like $(n-1)\pi$ for $n \to \infty$. Since for $0 \leq t < T$ the terms $\exp(-\mu_n^2(T-t))$ and all their derivatives with respect to t decay exponentially fast as $n \to \infty$, we can infer that the above series is infinitely differentiable with respect to t in $[0, T)$; hence, the complex extension of the series,

$$\varphi(z) := \sum_{n=1}^{\infty} \frac{1}{\sqrt{N_n}} \cos(\mu_n)\, \exp\big(-\mu_n^2 (T-z)\big)\, d_n,$$

defines an analytic function of the complex variable z in the half plane $\{\mathrm{Re}(z) < T\}$. We have to distinguish two different cases:

(i) $\varphi(z) \not\equiv 0$: By the identity theorem for analytic functions, φ can have at most finitely many zeros in any compact subset of the half plane $\{\mathrm{Re}(z) < T\}$ and, in particular, in any real interval $[0, T-\varepsilon]$ with $0 < \varepsilon < T$.

(ii) $\varphi(z) \equiv 0$: In this case $p(1,t) \equiv 0$ in $(-\infty, T)$, that is,

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{N_n}} \cos(\mu_n) \exp\big(-\mu_n^2(T-t)\big)\, d_n = 0 \quad \forall\, t < T.$$

Multiplying this equation by $\exp(\mu_1^2(T-t))$, we see that

$$\frac{1}{\sqrt{N_1}} \cos(\mu_1)\, d_1 + \sum_{n=2}^{\infty} \frac{1}{\sqrt{N_n}} \cos(\mu_n) \exp\big(-(\mu_n^2 - \mu_1^2)(T-t)\big)\, d_n = 0.$$

Letting $t \to -\infty$, we find that $d_1 = 0$. Repeating this process inductively, we deduce that $d_n = 0$ for all $n \in \mathbb{N}$. In conclusion, we must have $d = 0$, which contradicts the assumption. Therefore, case (ii) cannot occur, and the assertion is proved. □

Conclusion. *Under the assumptions of the bang-bang principle ($\lambda = 0$, positive optimal value) the optimal control is uniquely determined.*

Proof: Let $\bar{u}_1$ and $\bar{u}_2$ be two different optimal controls with associated optimal states $\bar{y}_1$ and $\bar{y}_2$, and let j denote the optimal value for $\|y(T) - y_\Omega\|$. We claim that then $u := (\bar{u}_1 + \bar{u}_2)/2$ is also optimal. Indeed, u is admissible since U_{ad} is convex, and the corresponding state is evidently $y = (\bar{y}_1 + \bar{y}_2)/2$. On the other hand, the triangle inequality yields that

$$\begin{aligned} \|y(T) - y_\Omega\|_{L^2(0,1)} &= \big\|\frac{1}{2}\big(\bar{y}_1(T) + \bar{y}_2(T)\big) - y_\Omega\big\|_{L^2(0,1)} \\ &\le \frac{1}{2}\big\|\bar{y}_1(T) - y_\Omega\|_{L^2(0,1)} + \frac{1}{2}\big\|\bar{y}_2(T) - y_\Omega\big\|_{L^2(0,1)} = j, \end{aligned}$$

which obviously implies that

$$\big\|y(T) - y_\Omega\big\|_{L^2(0,1)} = j.$$

Now, owing to the last theorem, $\bar{u}_1$ and $\bar{u}_2$ are bang-bang solutions. But then, $(\bar{u}_1 + \bar{u}_2)/2$ cannot be a bang-bang solution, which contradicts the above theorem. □

Example. Consider the following problem due to Schittkowski **[Sch79]**:

$$\min \frac{1}{2} \int_0^1 \big|y(x,T) - y_\Omega(x)\big|^2\, dx,$$

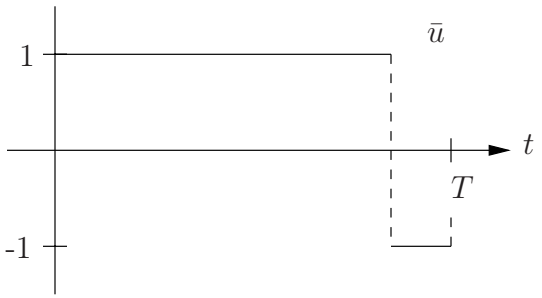
subject to

$$\begin{aligned} y_t(x,t) &= y_{xx}(x,t) \\ y_x(0,t) &= 0 \\ y_x(1,t) &= u(t) - y(1,t) \\ y(x,0) &= 0 \end{aligned}$$

and

$$-1 \le u(t) \le 1.$$

With the choices $y_\Omega(x) = \frac{1}{2}(1 - x^2)$ and $T = 1.58$, one obtains a numerical solution that has a switching point at $t = 1.329$; see the figure below.



Computed bang-bang control.

By combining numerical calculations with careful estimates for the Fourier series, it can be shown that the optimal value is positive and that the optimal control has exactly one switching point in $[1.329, 1.3294]$; see **[DT09]**.

References. Further results concerning the bang-bang principle in the control theory of parabolic equations are collected in **[GS80]**; see also **[GS77]**, **[GW76]**, **[Kno77]**, **[Sac78]**, **[Sch80]**, **[Sch89]**, and **[Trö84b]**, to name just a few of the numerous contributions. In **[Kar77]**, it was proved that if the maximum norm is used in place of the L^2 norm, then the optimal control can have only finitely many switching points for a positive optimal value. Numerical applications, using the calculation of switching points, have been reported in, e.g., **[ET86]**, **[Mac83b]** and, for the case of mixed control-state constraints, **[Trö84a]**.

Green's functions may also be used in higher dimensions. We do not pursue this here, and refer the interested reader to **[GW76]** or, for the application of the integral equation to semilinear equations, to **[Trö84b]**. A generalization of this concept is the approach via strongly continuous semigroups; detailed expositions of their use in control theory have been given in **[BDPDM92]**, **[BDPDM93]**, and **[Fat99]**.

3.3. Weak solutions in $W_2^{1,0}(Q)$

We consider as a typical case the parabolic analogue to problem (2.11) on page 34,

$$
\boxed{\begin{array}{rcll} y_t - \Delta y + c_0\, y &=& f & \text{in } Q = \Omega \times (0,T) \\ \partial_\nu y + \alpha\, y &=& g & \text{on } \Sigma = \Gamma \times (0,T) \\ y(\cdot,0) &=& y_0(\cdot) & \text{in } \Omega, \end{array}} \tag{3.23}
$$

confining ourselves from the beginning to boundary conditions of the third kind. It does not present any problem to split the boundary Γ into disjoint pieces Γ_0 and Γ_1 and prescribe homogeneous Dirichlet data on Γ_0; see (2.11) on page 34. The case of *inhomogeneous* nonsmooth Dirichlet data is more difficult to handle. We do not delve into this case here; boundary value problems of this type are treated, e.g., in [**Lio71**] or, using the theory of strongly continuous semigroups, in [**BDPDM92**], [**BDPDM93**], and [**Fat99**]. An approximation approach via weak solutions with Robin boundary conditions can be found in [**AEFR00**] and [**BBEFR03**].

We make the following assumptions:

Assumption 3.8. *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain with boundary Γ, and let $T > 0$ denote a fixed final time. Moreover, assume that functions $c_0 \in L^\infty(Q)$ and $\alpha \in L^\infty(\Sigma)$, where $\alpha(x,t) \geq 0$ for almost every $(x,t) \in \Sigma$, are prescribed.*

The functions c_0, α, f, and g all depend on (x,t); we assume that $f \in L^2(Q)$, $g \in L^2(\Sigma)$, and $y_0 \in L^2(\Omega)$.

Let us briefly discuss the difficulties we have to face when trying to define an appropriate notion of a solution to the parabolic problem (3.23). From a classical solution $y = y(x,t)$, one would expect the existence of all derivatives that appear, as well as their continuity in the interior of the space-time cylinder $Q = \Omega \times (0,T)$, i.e., $y \in C^{2,1}(Q)$. This requirement is much too restrictive for optimal control problems in which the controls belong to L^2 spaces. Recall that in the elliptic case we merely required the existence of the weak first-order spatial partial derivatives $D_i = \partial/\partial_{x_i}$; the boundary value problem was brought into a variational form in which half of the derivatives were shifted to the test function $v = v(x)$.

For parabolic problems, we follow a similar approach at first. Again, we introduce a variational formulation that requires the existence of the weak first-order partial derivatives $D_i = \partial/\partial_{x_i}$ only; the remaining half of the spatial derivatives will be shifted to the test function $v = v(x,t)$.

In addition, we have to account for the time t. With respect to t, also only the weak derivative comes into question. There are two possibilities: either we postulate the existence of the weak time derivative of y (which then is not needed for the test function v) or we do not (in which case the weak

time derivative is shifted to v). In most situations, the requirement that y_t belong to a space of functions, for instance $y_t \in L^2(Q)$, is too restrictive. Hence, initially, only the second possibility remains. This, however, is a source of asymmetry in the treatment of y and of v that renders the control theory more difficult. Eventually, we will obtain the existence of y_t, but as a functional and not as a function.

To begin with, we introduce two function spaces that are commonly used for the treatment of parabolic initial-boundary value problems (cf. the standard textbook [**LSU68**]).

Definition. *We denote by $W_2^{1,0}(Q)$ the normed space of all (equivalence classes of) functions $y \in L^2(Q)$ having weak first-order partial derivatives with respect to $x_1, ..., x_N$ in $L^2(Q)$, endowed with the norm*

$$\|y\|_{W_2^{1,0}(Q)} = \Big(\int_0^T \int_\Omega \big(|y(x,t)|^2 + |\nabla y(x,t)|^2 \big) \; dx\, dt \Big)^{1/2}.$$

Here, ∇ stands for the gradient with respect to x, i.e., $\nabla := \nabla_x$. We have

$$W_2^{1,0}(Q) = \big\{ y \in L^2(Q) \;\; : \; D_i\, y \in L^2(Q) \;\; \forall\, i = 1, \ldots, N \big\}.$$

The space $W_2^{1,0}(Q)$ is often referred to in the literature as $H^{1,0}(Q)$; it coincides with the space $L^2\big(0,T; H^1(\Omega)\big)$ that will be introduced in Section 3.4.1. One should note that in the notation $W_2^{1,0}(Q)$, the left and right upper indices indicate the order of the derivatives with respect to x and t, respectively, while the lower index indicates the order of integrability. In contrast to this, the second upper index in $W^{k,p}(\Omega)$ reflects the order of integrability. However, context and the different domains Q and Ω ought to prevent any confusion.

The elements of $W_2^{1,0}(Q)$ possess all first-order spatial derivatives in weak form, that is, there are functions $w_i \in L^2(Q)$ such that

$$\iint\limits_Q y(x,t)\, D_i\, v(x,t)\, dx\, dt = - \iint\limits_Q w_i(x,t)\, v(x,t)\, dx\, dt \quad \forall v \in C_0^\infty(Q).$$

Then, we put $D_i y(x,t) := w_i(x,t)$. Note that $W_2^{1,0}(Q)$ becomes a Hilbert space when endowed with the natural scalar product; see Ladyzhenskaya et al. [**LSU68**].

Definition. *The space $W_2^{1,1}(Q)$, defined by*

$$W_2^{1,1}(Q) = \big\{ y \in L^2(Q) \;\; : \;\; y_t \in L^2(Q) \text{ and } D_i\, y \in L^2(Q) \;\forall\, i = 1, \ldots, N \big\},$$

is a normed space with the norm

$$\|y\|_{W_2^{1,1}(Q)} = \Big(\int_0^T \int_\Omega \big(|y(x,t)|^2 + |\nabla y(x,t)|^2 + |y_t(x,t)|^2 \big) \; dx\, dt \Big)^{1/2}.$$

Here, $\nabla := \nabla_x$ is again the gradient with respect to x.

$W_2^{1,1}(Q)$ also becomes a Hilbert space when endowed with the natural scalar product. Note that this space coincides with $H^1(Q)$. Its elements have, apart from the weak partial derivatives with respect to the x_i, also a weak partial derivative with respect to t; that is, there is a function $w \in L^2(Q)$, denoted by $w = y_t$, such that

$$\iint\limits_Q y(x,t)\, v_t(x,t)\, dx\, dt = - \iint\limits_Q w(x,t)\, v(x,t)\, dx\, dt \quad \forall\, v \in C_0^\infty(Q).$$

We now transform problem (3.23) into a variational formulation by multiplying the parabolic partial differential equation by a test function $v \in C^1(\bar{Q})$ and then integrating over Q. Since we do not yet know what regularity can be expected from a "solution" y to (3.23), we argue *formally*, assuming that y is a classical solution for which all the integrals appearing below exist; in particular, y is assumed to be continuous on $\bar{Q}$. However, eventually the variational formulation should also be meaningful if we merely have $y \in W_2^{1,0}(Q)$, the space to which y should belong as a weak solution. Upon integrating over Q and by parts, we obtain for all $v \in C^1(\bar{Q})$ that

$$\begin{aligned} (3.24) \qquad & \int_0^T \int_\Omega y_t\, v\, dx\, dt - \int_0^T \int_\Omega v\, \Delta y\, dx\, dt + \int_0^T \int_\Omega c_0\, y\, v\, dx\, dt \\ &= \int_\Omega y(x,t)\, v(x,t)\, dx \Big|_0^T - \iint_Q \big(y\, v_t - \nabla y \cdot \nabla v - c_0\, y\, v \big)\, dx\, dt \\ &\quad - \iint_\Sigma v\, \partial_\nu\, y\, ds\, dt \\ &= \iint_Q f\, v\, dx\, dt. \end{aligned}$$

In this formulation, $y(x,0)$ and $y(x,T)$ occur. These values are not necessarily defined, since functions $y \in W_2^{1,0}(Q)$ need not be continuous in t. While the given initial value $y_0(x)$ can be inserted for $y(x,0)$, the final value $y(x,T)$ cannot be eliminated so easily. The test function $v = v(x,t)$ is smoother; it belongs to $C^1(\bar{Q})$. But the same operations as above can be performed even if merely $v \in W_2^{1,1}(Q)$; in particular, the functions $v(\cdot,0)$ and $v(\cdot,T)$ are for all $v \in W_2^{1,1}(Q)$ well defined as traces in $L^2(\Omega)$; see **[LSU68]**. It therefore makes sense to require that $v(x,T) = 0$ in order to get rid of the

term $y(x,T)$. Hence, substituting the boundary condition $\partial_\nu y = g - \alpha\, y$, we deduce that for all $v \in W_2^{1,1}(Q)$ with $v(x,T) = 0$,

$$\boxed{\begin{aligned} &\iint_Q \big(- y\, v_t + \nabla y \cdot \nabla v + c_0\, y\, v \big)\, dx\, dt + \iint_\Sigma \alpha\, y\, v\, ds\, dt \\ &\quad = \iint_Q f\, v\, dx\, dt + \iint_\Sigma g\, v\, ds\, dt + \int_\Omega y_0\, v(\cdot,0)\, dx. \end{aligned}} \tag{3.25}$$

We thus arrive at the following definition.

Definition. *We call a function $y \in W_2^{1,0}(Q)$ a* weak solution *to the initial-boundary value problem* (3.23) *if the variational equality* (3.25) *is satisfied for all $v \in W_2^{1,1}(Q)$ such that $v(\cdot,T) = 0$.*

Remark. Evidently, all terms occurring in (3.25) are well defined if $y \in W_2^{1,0}(Q)$ and $v \in W_2^{1,1}(Q)$. However, so far the regularity of y does not permit us to conclude the existence of the initial value $y(\cdot,0)$.

Theorem 3.9. *Suppose that Assumption 3.8 holds. Then the parabolic initial-boundary value problem* (3.23) *has a unique weak solution in $W_2^{1,0}(Q)$. Moreover, there is a constant $c_p > 0$, which is independent of f, g, and y_0, such that*

$$\max_{t\in[0,T]} \|y(\cdot,t)\|_{L^2(\Omega)} + \|y\|_{W_2^{1,0}(Q)} \le c_p \big(\|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)} \big) \tag{3.26}$$

for all $f \in L^2(Q)$, $g \in L^2(\Sigma)$, and $y_0 \in L^2(\Omega)$.

The above result is a special case of the more general Theorem 7.9 in Section 7.3.1. It ensures, in particular, that y is a continuous mapping from $[0,T]$ into $L^2(\Omega)$, that is, $y \in C\big([0,T], L^2(\Omega)\big)$ (the latter space will be introduced in the next section). Therefore, the norm $\max\limits_{t\in[0,T]} \|y(\cdot,t)\|_{L^2(\Omega)}$ and the initial and final values $y(\cdot,0)$ and $y(\cdot,T)$ are defined, and the initial condition $y(\cdot,0) = y_0$ is satisfied.

Conclusion. *The linear mapping $(f,g,y_0) \mapsto y$ is a continuous operator from $L^2(Q) \times L^2(\Sigma) \times L^2(\Omega)$ into $W_2^{1,0}(Q)$ and into $L^2\big(0,T; H^1(\Omega)\big) \cap C\big([0,T], L^2(\Omega)\big)$.*

The variational formulation (3.25) has a severe disadvantage: the test function v has to belong to the space $W_2^{1,1}(Q)$ and satisfy $v(\cdot,T)=0$, and eventually we should insert the adjoint state p in place of v. As a rule, however, p neither belongs to $W_2^{1,1}(Q)$ nor needs to satisfy $p(\cdot,T)=0$. Hence, the asymmetry between the requirements on y and the requirements on the test function v is a disturbing factor in the theory of optimal control, and a different approach is needed. Summarizing, we conclude that the space $W_2^{1,0}(Q)$ is not well suited for the study of optimal control problems.

3.4. Weak solutions in $W(0,T)$

3.4.1. Functions with values in Banach spaces. The concept of functions with values in Banach spaces is a fundamental tool for the treatment of nonstationary equations (evolution equations). Here, we consider only functions that are defined on a compact interval $[a,b]\subset\mathbb{R}$. We will call such functions *vector-valued functions* in this textbook; the term *abstract functions* is also commonly used in the relevant literature.

Definition. *Any mapping from $[a,b]\subset\mathbb{R}$ into a Banach space X is called a* vector-valued function.

Examples. Depending on the choice of the space X, we have the following special cases:

(i) $\boldsymbol{X=\mathbb{R}}$. Then a vector-valued function $y:[a,b]\to\mathbb{R}$ is just a real-valued function of one variable.

(ii) $\boldsymbol{X=\mathbb{R}^N}$. Then $y:[a,b]\to\mathbb{R}^N$ is a function that assigns to the variable t a vector

$$y(t)=\begin{bmatrix} y_1(t)\\ \vdots\\ y_N(t)\end{bmatrix}\in\mathbb{R}^N.$$

(iii) $\boldsymbol{X=H^1(\Omega)}$. In this case, for any $t\in[a,b]$ the value $y(t)$ of a function $y:[a,b]\to H^1(\Omega)$ is an element of $H^1(\Omega)$ and hence a function itself that acts on the spatial variable $x\in\Omega$. In other words, $y(t)=y(\cdot,t)\in H^1(\Omega)$ for any fixed $t\in[a,b]$, that is, the function $x\mapsto y(x,t)$ belongs to $H^1(\Omega)$.

The solution $y\in W_2^{1,0}(Q)$ to our parabolic initial-boundary value problem is of this type. ◇

We are now going to introduce some important spaces of vector-valued functions.

Definition. *Let $\{X, \|\cdot\|_X\}$ be a real Banach space. We say that a vector-valued function $y : [a,b] \to X$ is* continuous at the point $t \in [a,b]$ *if we have* $\lim_{\tau \to t} \|y(\tau) - y(t)\|_X = 0$. *We denote the space of all vector-valued functions that are continuous at every $t \in [a,b]$ by $C\big([a,b], X\big)$. The space $C\big([a,b], X\big)$ is a Banach space with respect to the norm*

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

Example. The space $C\big([0,T], L^2(\Omega)\big)$ consists of all real-valued functions $y = y(x,t)$ that are measurable in $\Omega \times [0,T]$, square integrable with respect to $x \in \Omega$ for every $t \in [0,T]$, and continuous in t with respect to the norm of $L^2(\Omega)$. For any $t \in [0,T]$ we have

$$\int_\Omega |y(x,t)|^2 \, dx < \infty,$$

and for $\tau \to t$ we have

$$\|y(\tau) - y(t)\|_{L^2(\Omega)} = \Big(\int_\Omega \big|y(x,\tau) - y(x,t)\big|^2 \, dx \Big)^{1/2} \to 0.$$

The norm of y is given by

$$\|y\|_{C([0,T],L^2(\Omega))} = \max_{t \in [0,T]} \|y(t)\|_{L^2(\Omega)} = \max_{t \in [0,T]} \Big(\int_\Omega |y(x,t)|^2 \, dx \Big)^{1/2}. \qquad \diamond$$

For functions $y \in C\big([a,b], X\big)$, the *Riemann integral* $I = \int_a^b y(t)\, dt$ can be defined in much the same way as in the cases where $X = \mathbb{R}$ or $X = \mathbb{R}^N$: we define the integral as the limit of Riemann sums

$$\sum_{i=1}^{k} y(\xi_i) \ (t_i - t_{i-1}),$$

with arbitrary intermediate points $\xi_i \in [t_{i-1}, t_i]$, as the step size of the subdivision $a = t_0 < t_1 < \ldots < t_k = b$ tends to zero. Note that the integral I is an element of the space X; see Hille and Phillips [**HP57**].

Definition. *A vector-valued function $y : [a,b] \to X$ is called a* step function *if there are finitely many $y_i \in X$ and Lebesgue measurable, pairwise disjoint sets $M_i \subset [a,b]$, $1 \le i \le m$, such that $[a,b] = \bigcup_{i=1}^m M_i$ and $y(t) = y_i$ for every $t \in M_i$, $1 \le i \le m$.*

Definition. *A vector-valued function $y : [a,b] \to X$ is said to be* measurable *if there exists a sequence $\{y_k\}_{k=1}^\infty$ of step functions $y_k : [a,b] \to X$ such that $y(t) = \lim_{k \to \infty} y_k(t)$ for almost every $t \in [a,b]$.*

Now we are ready to introduce L^p spaces of vector-valued functions.

Definition.

(i) *We denote by $L^p(a,b;X)$, $1 \le p < \infty$, the linear space of all (equivalence classes of) measurable vector-valued functions $y : [a,b] \to X$ having the property that*

$$\int_a^b \|y(t)\|_X^p \, dt < \infty.$$

The space $L^p(a,b;X)$ is a Banach space with respect to the norm

$$\|y\|_{L^p(a,b;X)} := \Big(\int_a^b \|y(t)\|_X^p \, dt \Big)^{1/p}.$$

(ii) *We denote by $L^\infty(a,b;X)$ the Banach space of all (equivalence classes of) measurable vector-valued functions $y : [a,b] \to X$ having the property that*

$$\|y\|_{L^\infty(a,b;X)} := \operatorname{ess\ sup}_{[a,b]} \|y(t)\|_X < \infty.$$

In the spaces just defined, functions that differ at most on a subset of $[a,b]$ of zero Lebesgue measure belong to the same equivalence class and thus are regarded as equal. Observe that obviously $C\big([a,b],X\big) \subset L^p(a,b;X) \subset L^q(a,b;X)$ for $1 \le q \le p \le \infty$.

Example. For any $T > 0$ the function $y(x,t) = \dfrac{e^t}{\sqrt{x}}$ is an element of $C\big([0,T],L^1(0,1)\big)$ and thus also of $L^p\big(0,T;L^1(0,1)\big)$ whenever $1 \le p \le \infty$. ◇

In $L^1(a,b;X)$, and hence also in the spaces $L^p(a,b;X)$ with $1 \le p \le \infty$ and $C\big([a,b],X\big)$, the *Bochner integral* can be defined for any vector-valued function. For step functions y it is defined by

$$\int_a^b y(t) \, dt := \sum_{i=1}^m y_i \, |M_i|,$$

where $y_i \in X$ is the value of y on M_i and $|M_i|$ denotes the Lebesgue measure of the set M_i, $1 \le i \le m$. It is apparent that the integral is an element of X. For arbitrary $y \in L^1(a,b;X)$, we can argue as follows: since y is measurable, there exists a sequence $\{y_k\}_{k=1}^\infty$ of step functions converging almost everywhere on $[a,b]$ to y. Then the Bochner integral of y is defined by

$$\int_a^b y(t) \, dt = \lim_{k\to\infty} \int_a^b y_k(t) \, dt.$$

This limit is independent of the choice of the sequence $\{y_k\}_{k=1}^{\infty}$; see Hille and Phillips [**HP57**] or Pazy [**Paz83**]. The Bochner integral is the analogue of the Lebesgue integral for $X = \mathbb{R}$.

Example. The vector-valued elements of $L^2\big(0,T;H^1(\Omega)\big)$ can be viewed as real-valued functions of the variables x and t, i.e., $y = y(x,t)$ with $x \in \Omega$ and $t \in [0,T]$. For each t, $y(\cdot,t)$ belongs to $H^1(\Omega)$ with respect to x. The norm, given by

$$\begin{aligned} \|y\|_{L^2(0,T;H^1(\Omega))} &= \Big(\int_0^T \|y(t)\|^2_{H^1(\Omega)}\,dt\Big)^{1/2} \\ &= \Big(\int_0^T\!\!\int_\Omega \big(|y(x,t)|^2 + |\nabla_x y(x,t)|^2\big)\,dx\,dt\Big)^{1/2}, \end{aligned}$$

is obviously the same as that in $W_2^{1,0}(Q)$. This suggests that the two spaces might coincide. Indeed, it turns out that they are isometric and isomorphic to each other, that is,

$$W_2^{1,0}(Q) \cong L^2\big(0,T;H^1(\Omega)\big).$$

More precisely, it can be shown that any $y \in W_2^{1,0}(Q)$ coincides—modulo a modification on a set of zero measure—with a function in $L^2\big(0,T;H^1(\Omega)\big)$ and vice versa. For a proof, we refer the reader to [**HP57**]. ◇

Additional information about vector-valued functions can be found, e.g., in Hille and Phillips [**HP57**], Pazy [**Paz83**], Tanabe [**Tan79**], and Wloka [**Wlo87**].

3.4.2. Vector-valued functions and parabolic problems. Once again, we derive a variational formulation of the parabolic problem (3.23) on page 137. We proceed similarly as in the derivation of (3.25). However, this time we keep $t \in (0,T)$ fixed, multiplying (3.23) by a function $v \in H^1(\Omega) = V$ of x alone and integrating over Ω only. It follows (only formally, since it is still unclear how to define the time derivative y_t) that

$$\begin{aligned} \int_\Omega y_t(t)\,v\,dx &= -\int_\Omega \nabla y(t)\cdot\nabla v\,dx + \int_\Omega \big(f(t) - c_0(t)\,y(t)\big)\,v\,dx \\ &\quad + \int_\Gamma \big(g(t) - \alpha(t)\,y(t)\big)\,v\,ds \quad \forall\, t \in [0,T], \end{aligned} \tag{3.27}$$

where we have suppressed the dependence on x. Since $L^2(Q) \cong L^2\big(0,T;L^2(\Omega)\big)$ and $L^2(\Sigma) \cong L^2\big(0,T;L^2(\Gamma)\big)$, we may regard f, g, and y as vector-valued functions in t that coincide almost everywhere with the corresponding real-valued functions. For example, by Fubini's theorem the function $x \mapsto f(t,x)$ belongs to $L^2(\Omega)$ for almost all $t \in (0,T)$; modifying f for the

remaining t to, e.g., zero, we obtain the vector-valued function $t \mapsto f(t,\cdot)$ belonging to $L^2\big(0,T;L^2(\Omega)\big)$.

For almost every $t \in (0,T)$, the expression on the right-hand side of (3.27) is linear and continuous in v and assigns to each $v \in H^1(\Omega)$ a real number. Hence, it defines for almost every t a linear and continuous functional $F = F(t) \in H^1(\Omega)^*$, and we have

$$\int_\Omega y_t(t)\, v\, dx = F(t)\, v \quad \forall\, v \in H^1(\Omega), \qquad \text{for a.e. } t \in [0,T]. \tag{3.28}$$

Since the right-hand side of (3.27) defines a functional $F = F(t) \in H^1(\Omega)^*$, so should the left-hand side; consequently, $y_t(t)$ ought to be a continuous linear functional on $H^1(\Omega)$, i.e., $y_t(t) \in H^1(\Omega)^*$ for almost every $t \in (0,T)$.

Now, every weak solution $y \in W_2^{1,0}(Q)$ can—after an appropriate modification on a null set—be understood as an element of $L^2\big(0,T;H^1(\Omega)\big)$. In the proof of Theorem 3.12 it will be shown that then

$$\int_0^T \|F(t)\|^2_{H^1(\Omega)^*}\, dt < \infty;$$

in other words, $F \in L^2\big(0,T;H^1(\Omega)^*\big)$. Comparing this with the left-hand side of (3.28), we conclude that

$$y_t \in L^2\big(0,T,H^1(\Omega)^*\big). \tag{3.29}$$

This observation gives us an important hint as to in which space the derivative y_t should be looked for. However, we still do not know how this derivative has to be understood. To this end, we need the notion of *vector-valued distributions*.

3.4.3. Vector-valued distributions. In the following, let V be a Banach space, where we have $V = H^1(\Omega)$ in mind. For given $y \in Y$, we define a *vector-valued distribution* $\mathcal{T} : C_0^\infty(0,T) \to V$ by setting

$$\mathcal{T}\varphi := \int_0^T y(t)\, \varphi(t)\, dt \quad \forall \varphi \in C_0^\infty(0,T).$$

As usual, we identify the generating function y with the distribution $\mathcal{T}$, which is then investigated in place of y. In order to stress the dependence on y, the notation $\mathcal{T}_y$ is also commonly used in the literature; but we avoid this notation for the sake of clarity.

We now introduce the derivative $\mathcal{T}'$ as a vector-valued distribution, in terms of the Bochner integral:

$$\mathcal{T}'\varphi := -\int_0^T y(t)\, \varphi'(t)\, dt.$$

Likewise, one can define $\mathcal{T}''$ by putting $\mathcal{T}''\varphi := \int_0^T y(t)\,\varphi''(t)\,dt$, etc. Observe that y is a vector-valued function, while φ is real-valued. If a vector-valued function $w = w(t) \in L^1(0,T;V)$ exists such that

$$\mathcal{T}'\varphi = -\int_0^T y(t)\,\varphi'(t)\;dt = \int_0^T w(t)\,\varphi(t)\;dt \qquad \forall\,\varphi \in C_0^\infty(0,T),$$

then $\mathcal{T}'$ can be identified with w, since it is induced by w. Since we have identified $\mathcal{T}$ with the generating function y and $\mathcal{T}'$ with w, we do the same with the generating functions and define

$$y'(t) := w(t).$$

In this sense, we have $y' \in L^1(0,T;V)$ here. Hidden behind this is the formula of integration by parts, which for continuously differentiable $y : [0,T] \to V$ and $\varphi \in C_0^\infty(0,T)$ yields that

$$\int_0^T y(t)\,\varphi'(t)\,dt = y(t)\,\varphi(t)\big|_0^T - \int_0^T y'(t)\,\varphi(t)\,dt = -\int_0^T y'(t)\,\varphi(t)\,dt.$$

Remark. The weak derivatives defined in (2.1) on page 27 can be introduced in the same way: to this end, one considers the real-valued distribution $\mathcal{T} : C_0^\infty(\Omega) \to \mathbb{R}$,

$$\mathcal{T}\varphi := \int_\Omega y(x)\,\varphi(x)\,dx \qquad \forall \varphi \in C_0^\infty(\Omega),$$

and regards y as a distribution. In this way, $D^\alpha y$ is initially defined as a distributional derivative. Those distributional derivatives that are generated by locally integrable functions are then weak derivatives. This approach is employed in many texts dealing with Sobolev spaces. Obviously, the above function w is an analogue of the weak derivative $w = D^\alpha y$ introduced in (2.1).

If we even have $w \in L^2(0,T;V)$, then y would belong to the class of functions in $L^2(0,T;V)$ having a regular derivative in $L^2(0,T;V)$. However, this class of functions is too small for our purposes. Indeed, relation (3.29) indicates that the derivative $y' = y_t$ in the parabolic equation (3.23) should belong to the larger space $L^2(0,T;V^*)$.

In the following definition, we consider y as a vector-valued function in $L^2(0,T;V^*)$, so that $\mathcal{T}$ maps into V^*. This is possible since V is continuously embedded in V^*, as we shall see below.

Definition. *We denote by $W(0,T)$ the linear space of all $y \in L^2(0,T;V)$ having a (distributional) derivative $y' \in L^2(0,T;V^*)$, equipped with the norm*

$$\|y\|_{W(0,T)} = \Big(\int_0^T \big(\|y(t)\|_V^2 + \|y'(t)\|_{V^*}^2\big)\;dt\Big)^{1/2}.$$

The space $W(0,T) = \{y \in L^2(0,T;V) \,:\, y' \in L^2(0,T;V^*)\}$ is a Hilbert space with the scalar product

$$\big(u\,,\,v\big)_{W(0,T)} = \int_0^T \big(u(t)\,,\,v(t)\big)_V \,dt + \int_0^T \big(u'(t)\,,\,v'(t)\big)_{V^*} \,dt.$$

Here, we have used the abbreviation $\big(F\,,\,G\big)_{V^*} := \big(J\,F\,,\,J\,G\big)_V$, where $J : V^* \to V$ is the duality mapping from the Riesz representation theorem that assigns to each functional $F \in V^*$ the corresponding $f \in V$.

To facilitate comprehension of the arguments that follow, we now introduce the notion of a *Gelfand triple*. In our case, we have the following situation: the space $V = H^1(\Omega)$ is continuously and densely embedded in the Hilbert space $H = L^2(\Omega)$. By the Riesz representation theorem, we identify the dual H^* with H. Also, V is a Hilbert space that could be identified with its dual V^*; however, we avoid doing this for obvious reasons—for instance, in integrations by parts we tacitly use the scalar product of H and not that of V, which would have to be used if V were identified with V^*.

Any $f \in H$ can via

$$v \mapsto (f\,,\,v)_H \in \mathbb{R} \qquad \forall\, v \in V$$

be regarded as an element of V^*. In this sense,

$$V \subset H = H^* \subset V^*.$$

Now, it can be shown that the embedding $H \subset V^*$ is also dense and continuous; see Wloka [**Wlo87**], page 253 and following. The chain of dense and continuous embeddings

$$V \subset H \subset V^*$$

is called a *Gelfand triple*. Owing to the Riesz representation theorem, functionals $F \in V^*$ can be continuously extended to the larger space H if and only if they are of the form

$$F(v) = (f\,,\,v)_H,$$

with a fixed $f \in H$. In the case of $V = H^1(\Omega)$ and $H = L^2(\Omega)$, this means that

$$H^1(\Omega) \subset L^2(\Omega) \subset H^1(\Omega)^*,$$

and $F \in H^1(\Omega)^*$ can be continuously extended to $L^2(\Omega)$ if and only if there is some $f \in L^2(\Omega)$ such that

$$F(v) = \int_\Omega f(x)\,v(x)\,dx = \big(f\,,\,v\big)_{L^2(\Omega)} \quad \forall v \in H^1(\Omega).$$

The following results, which can be found in Wloka [**Wlo87**] or Zeidler [**Zei90a**], hold for any Gelfand triple. They are of fundamental importance for our purposes.

Theorem 3.10. *Every $y \in W(0,T)$ coincides—possibly after a suitable modification on a set of zero measure—with an element of $C\big([0,T],H\big)$. In this sense, we have the continuous embedding $W(0,T) \hookrightarrow C\big([0,T],H\big)$.*

In particular, it follows that for any $y \in W(0,T)$, the values $y(0)$ and $y(T)$ belong to H. Moreover, there exists a constant $c_E > 0$ such that

$$\|y\|_{C([0,T],H)} \le c_E \, \|y\|_{W(0,T)} \qquad \forall\, y \in W(0,T).$$

Theorem 3.11. *For all y, $p \in W(0,T)$ the* formula of integration by parts holds:

$$\begin{aligned} \int_0^T \big(y'(t)\,,\,p(t)\big)_{V^*,V}\,dt &= \big(y(T)\,,\,p(T)\big)_H - \big(y(0)\,,\,p(0)\big)_H \\ &\quad - \int_0^T \big(p'(t)\,,\,y(t)\big)_{V^*,V}\,dt. \end{aligned}$$

Here, $(F\,,\,v)_{V^*,V} := F(v)$ for $F \in V^*$ and $v \in V$. This notation, which resembles a scalar product, is commonly used in the literature.

Conclusion. Taking $p = y$ in the formula of integration by parts, we find that any $y \in W(0,T)$ satisfies the useful identity

$$\int_0^T \big(y'(t)\,,\,y(t)\big)_{V^*,V}\,dt = \frac{1}{2}\|y(T)\|_H^2 - \frac{1}{2}\|y(0)\|_H^2. \tag{3.30}$$

Hence, we can formally write

$$\int_0^T \big(y'(t)\,,\,y(t)\big)_{V^*,V}\,dt = \int_0^T \frac{1}{2}\frac{d}{dt}\|y(t)\|_H^2\,dt = \frac{1}{2}\|y(T)\|_H^2 - \frac{1}{2}\|y(0)\|_H^2.$$

3.4.4. Weak solutions from $W_2^{1,0}(Q)$ belong to $W(0,T)$. In this section, we will show that weak solutions to our parabolic initial-boundary value problems belong to $W(0,T)$. To this end, we again consider the problem (3.23) on page 137,

$$\boxed{\begin{aligned} y_t - \Delta y + c_0\, y &= f && \text{in } Q \\ \partial_\nu y + \alpha\, y &= g && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega, \end{aligned}}$$

and require that Assumption 3.8 be satisfied. We have the following result.

Theorem 3.12. *Let $y \in W_2^{1,0}(Q)$ be the weak solution to problem (3.23), which exists according to Theorem 3.9. Then y belongs—possibly after a modification on a set of zero measure—to $W(0,T)$.*

Proof: It follows from (3.25) that, for all $v \in W_2^{1,1}(Q)$ with $v(T) = 0$,

$$
\begin{aligned}
-\iint_Q y\, v_t \, dx\, dt \;&=\; -\iint_Q \nabla y \cdot \nabla v \, dx\, dt - \iint_Q c_0\, y\, v \, dx\, dt \\
&\quad - \iint_\Sigma \alpha\, y\, v \, ds\, dt + \int_\Omega y_0\, v(0) \, dx + \iint_Q f\, v \, dx\, dt + \iint_\Sigma g\, v \, ds\, dt.
\end{aligned}
$$

In particular, we may insert any function of the form $v(x,t) := v(x)\varphi(t)$, where $\varphi \in C_0^\infty(0,T)$ and $v \in V = H^1(\Omega)$. Setting $H = L^2(\Omega)$ and $H^N = H \times H \times \ldots \times H$ (N times), we find that

$$
\begin{aligned}
-\int_0^T \big(y(t)\,\varphi'(t)\,,\, v\big)_H \, dt \;&=\; -\int_0^T \big(\nabla y(t)\,,\, \nabla v\big)_{H^N} \varphi(t)\, dt \\
&\quad - \int_0^T \big(c_0\, y(t)\,,\, v\big)_H \varphi(t)\, dt - \int_0^T \big(\alpha(t)\, y(t)\,,\, v\big)_{L^2(\Gamma)} \varphi(t)\, dt \\
&\quad + \int_0^T \big(f(t)\,,\, v\big)_H \varphi(t) \, dt + \int_0^T \big(g(t)\,,\, v\big)_{L^2(\Gamma)} \varphi(t)\, dt.
\end{aligned}
$$

The initial condition vanishes, since $\varphi(0) = 0$. Now $y \in L^2(Q)$, by the definition of $W_2^{1,0}(Q)$. Hence, by Fubini's theorem, $y(\cdot,t) \in L^2(\Omega)$ for almost every $t \in (0,T)$. Moreover, $D_i y \in L^2(Q)$ for $i = 1, \ldots, N$, and thus $\nabla y(\cdot,t) \in (L^2(\Omega))^N = H^N$ for almost every $t \in (0,T)$. Finally, $y(\cdot,t) \in H^1(\Omega)$, and thus $y(\cdot,t) \in L^2(\Gamma)$ for almost every $t \in (0,T)$.

On the set of measure zero in $[0,T]$ where one of the above statements possibly does not hold, we put $y(t) = 0$, which does not change the vector-valued function y in the sense of L^2 spaces. Hence, we see that for any fixed t, the expressions in the integrals on the right-hand side define linear functionals $F_i(t) : H^1(\Omega) \to \mathbb{R}$, namely

$$
\begin{aligned}
F_1(t) :&\; v \;\mapsto\; \big(\nabla y(t)\,,\, \nabla v\big)_{H^N} \\
F_2(t) :&\; v \;\mapsto\; \big(c_0(t)\, y(t)\,,\, v\big)_H \\
F_3(t) :&\; v \;\mapsto\; \big(\alpha(t)\, y(t)\,,\, v\big)_{L^2(\Gamma)} \\
F_4(t) :&\; v \;\mapsto\; \big(f(t)\,,\, v\big)_H \\
F_5(t) :&\; v \;\mapsto\; \big(g(t)\,,\, v\big)_{L^2(\Gamma)},
\end{aligned}
\tag{3.31}
$$

in that order. We claim that the functionals $F_i(t)$, $1 \le i \le 5$, are bounded and thus continuous on V for every t. We verify this claim only for $F_1(t)$ and $F_3(t)$, leaving the other cases as an easy exercise for the reader. First, we have

$$|F_1(t)v| \le \int_\Omega |\nabla y(t)|\,|\nabla v|\,dx \le \|y(t)\|_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)} \quad \forall\, v \in H^1(\Omega).$$

Note that the function $t \mapsto \|y(t)\|_{H^1(\Omega)}$ belongs to $L^2(0,T)$, and $\|y(t)\|_{H^1(\Omega)}$ is by construction everywhere finite. $F_3(t)$ can be estimated similarly:

$$\begin{aligned} |F_3(t)v| &\le \int_\Gamma |\alpha(t)|\,|y(t)|\,|v|\,ds \le \|\alpha\|_{L^\infty(\Sigma)}\,\|y(t)\|_{L^2(\Gamma)}\,\|v\|_{L^2(\Gamma)} \\ &\le \tilde{c}\,\|\alpha\|_{L^\infty(\Sigma)}\,\|y(t)\|_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)}. \end{aligned}$$

Hence, $F_3(t)$ is bounded, and $\|F_3(t)\|_{H^1(\Omega)^*} \le \tilde{c}\,\|\alpha\|_{L^\infty(\Sigma)}\,\|y(t)\|_{H^1(\Omega)}$.

In this way, we find that $F_i(t) \in V^* = H^1(\Omega)^*$ for every t, and there is some constant $c > 0$ such that

$$\sum_{i=1}^5 \|F_i(t)\|_{V^*} \le c\,\big(\|y(t)\|_{H^1(\Omega)} + \|f(t)\|_{L^2(\Omega)} + \|g(t)\|_{L^2(\Gamma)}\big). \tag{3.32}$$

Since the expression on the right-hand side belongs to $L^2(0,T)$, so does the expression on the left-hand side, showing that $F_i \in L^2(0,T;V^*)$ for $1 \le i \le 5$. But then the functional F on the right-hand side of the variational formulation, being just the sum of the F_i, also belongs to $L^2(0,T;V^*)$.

Rewriting the variational formulation in terms of F, we obtain that for all $v \in V$ we have the chain of equalities

$$\begin{aligned} \Big(-\int_0^T y(t)\,\varphi'(t)\,dt\,,\,v\Big)_{L^2(\Omega)} &= -\int_0^T \big(y(t)\,\varphi'(t)\,,\,v\big)_{L^2(\Omega)}\,dt \\ = \int_0^T \big(F(t)\,\varphi(t)\,,\,v\big)_{V^*,V}\,dt &= \Big(\int_0^T F(t)\,\varphi(t)\,dt\,,\,v\Big)_{V^*,V} \end{aligned}$$

and therefore, as an equation in the space V^*,

$$-\int_0^T y(t)\,\varphi'(t)\,dt = \int_0^T F(t)\,\varphi(t)\,dt \qquad \forall\,\varphi \in C_0^\infty(0,T).$$

But this means that $y' = F$ in the sense of vector-valued distributions; hence, $y' \in L^2(0,T;V^*)$. In conclusion, $y \in W(0,T)$, and the assertion is proved. □

Theorem 3.13. *The weak solution y to the problem* (3.23) *satisfies an estimate of the form*

$$\|y\|_{W(0,T)} \le c_w\,\big(\|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)}\big),$$

with some constant $c_w > 0$ that does not depend on (f, g, y_0). In other words, the mapping $(f, g, y_0) \mapsto y$ defines a continuous linear operator from $L^2(Q)\times L^2(\Sigma)\times L^2(\Omega)$ into $W(0,T)$ and, in particular, into $C\big([0,T], L^2(\Omega)\big)$.

Proof: We estimate the norm $\|y\|_{W(0,T)}$ or, more precisely, its square

$$\|y\|^2_{W(0,T)} = \|y\|^2_{L^2(0,T;H^1(\Omega))} + \|y'\|^2_{L^2(0,T;H^1(\Omega)^*)}.$$

For the first summand, we obtain from Theorem 3.9 on page 140, with a generic constant $c > 0$, the estimate

$$\|y\|^2_{L^2(0,T;H^1(\Omega))} = \|y\|^2_{W_2^{1,0}(Q)} \le c\,\big(\|f\|^2_{L^2(Q)} + \|g\|^2_{L^2(\Sigma)} + \|y_0\|^2_{L^2(\Omega)}\big). \tag{3.33}$$

The second summand requires only a little more effort. Indeed, with the functionals F_i defined in (3.31), we have

$$\|y'\|_{L^2(0,T;H^1(\Omega)^*)} = \Big\| \sum_{i=1}^5 F_i \Big\|_{L^2(0,T;H^1(\Omega)^*)} \le \sum_{i=1}^5 \|F_i\|_{L^2(0,T;H^1(\Omega)^*)}.$$

We estimate only the norm of F_1, leaving the others to the reader. Using the above estimates, in particular (3.32) and (3.33), we find, with a generic constant $c > 0$, that

$$\begin{aligned} \|F_1\|^2_{L^2(0,T;V^*)} &= \int_0^T \|F_1(t)\|^2_{V^*}\,dt \;\le\; \int_0^T c\,\|y(t)\|^2_{H^1(\Omega)}\,dt \\ &\le c\,\|y\|^2_{W_2^{1,0}(Q)} \;\le\; c\,\big(\|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)}\big)^2. \end{aligned}$$

The norms of $F_2, \ldots, F_5$ can be estimated similarly. The assertion is thus proved. □

Now that the existence of the derivative $y_t := y'$ has been shown, we are in a position to reformulate the variational equality (3.25). Assuming that $y \in W(0,T)$ in (3.25) on page 140, we can keep the final value $y(T)$, which is now well defined. It follows from (3.24) on page 139 that for all $v \in W_2^{1,1}(Q)$ we have

$$\begin{aligned} &-\iint_Q y\,v_t + \iint_Q \nabla y\cdot\nabla v + \iint_Q c_0\,y\,v + \iint_\Sigma \alpha\,y\,v \\ &= \iint_Q f\,v + \iint_\Sigma g\,v + \int_\Omega y_0\,v(\cdot,0) - \int_\Omega y(\cdot,T)\,v(\cdot,T), \end{aligned}$$

where the differentials in the integrals have been omitted for the sake of brevity. We now use the facts that y, $v \in W(0,T)$ and $y(0) = y_0$. Invoking

the formula of integration by parts on page 148, we find that for all $v \in W(0,T)$,

$$\begin{aligned} (3.34) \qquad & \int_0^T \big(y_t\,,\,v\big)_{V^*,V} + \iint_Q \nabla y \cdot \nabla v + \iint_Q c_0\, y\, v + \iint_\Sigma \alpha\, y\, v \\ & = \iint_Q f\, v + \iint_\Sigma g\, v, \\ & y(0) = y_0, \end{aligned}$$

where y_t is a vector-valued function in $L^2(0,T;V^*)$. The extension from $v \in W^{1,1}(Q)$ to $v \in W(0,T)$ follows from a density argument; indeed, the integrals appearing in the above equation are continuous with respect to v in $W(0,T)$, and the functions in $W^{1,1}(Q)$, regarded as elements of $W(0,T)$, form a dense subset of $W(0,T)$. The above variational formulation is valid even for $v \in L^2(0,T;V)$, since all of the expressions involved are continuous also in this space. Therefore, (3.34) can be rewritten in the following equivalent form:

(3.35)

$$\boxed{\begin{aligned} & \int_0^T \big(y_t\,,\,v\big)_{V^*,V}\,dt + \iint_Q \big(\nabla y \cdot \nabla v + c_0\, y\, v\big)\,dx\,dt + \iint_\Sigma \alpha\, y\, v\,ds\,dt \\ & \quad = \iint_Q f\, v\,dx\,dt + \iint_\Sigma g\, v\,ds\,dt \qquad \forall v \in L^2(0,T;V), \\ & \quad y(0) = y_0. \end{aligned}}$$

The solution mapping $(f,g,y_0) \mapsto y$ corresponding to the initial-boundary value problem (3.23),

$$\begin{aligned} y_t - \Delta y + c_0\, y &= f \qquad \text{in } Q = \Omega \times (0,T) \\ \partial_\nu y + \alpha\, y &= g \qquad \text{on } \Sigma = \Gamma \times (0,T) \\ y(\cdot,0) &= y_0 \qquad \text{in } \Omega, \end{aligned}$$

has the structure

$$(3.36) \qquad y = G_Q\, f + G_\Sigma\, g + G_0\, y_0,$$

with continuous linear operators $G_Q : L^2(Q) \to W(0,T)$, $G_\Sigma : L^2(\Sigma) \to W(0,T)$, and $G_0 : L^2(\Omega) \to W(0,T)$ which are defined by

$$\begin{aligned} G_Q &: f \mapsto y \quad \text{for} \quad g = 0,\ y_0 = 0 \\ G_\Sigma &: g \mapsto y \quad \text{for} \quad f = 0,\ y_0 = 0 \\ G_0 &: y_0 \mapsto y \quad \text{for} \quad f = 0,\ g = 0. \end{aligned}$$

The basis for this representation is Theorem 3.13.

3.5. Parabolic optimal control problems

As in the elliptic case, we begin our analysis by transforming selected linear-quadratic parabolic optimal control problems into quadratic optimization problems in Hilbert spaces and proving their solvability.

To begin with, we fix some general assumptions on the given quantities. These concern the spatial domain Ω and its boundary Γ, the final time $T > 0$, target functions y_Ω, y_Q, y_Σ that are to be approximated, the initial distribution y_0, coefficients α and β, as well as bounds u_a, u_b, v_a, v_b that—depending on the particular problem—have to be obeyed on either $E = Q = \Omega \times (0,T)$ or $E = \Sigma = \Gamma \times (0,T)$. The actual meaning of the set E can be discerned from the context.

Assumption 3.14. *Let $\Omega \subset \mathbb{R}^N$ be a domain with Lipschitz boundary Γ, and let $\lambda \geq 0$ be a fixed constant. Assume that we are given functions $y_\Omega \in L^2(\Omega)$, $y_Q \in L^2(Q)$, $y_\Sigma \in L^2(\Sigma)$, α, $\beta \in L^\infty(E)$, and u_a, u_b, v_a, $v_b \in L^2(E)$ with $u_a(x,t) \leq u_b(x,t)$ and $v_a(x,t) \leq v_b(x,t)$ for almost every $(x,t) \in E$. Here, depending on the specific problem under study, $E = Q$ or $E = \Sigma$.*

3.5.1. Optimal nonstationary boundary temperature. We consider the problem (3.1)–(3.3) from page 119:

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

subject to

$$\begin{array}{rcll} y_t - \Delta y & = & 0 & \text{in } Q \\ \partial_\nu y + \alpha\, y & = & \beta\, u & \text{on } \Sigma \\ y(0) & = & 0 & \text{in } \Omega \end{array}$$

and

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

Owing to Theorem 3.13 on page 150 and Theorem 3.12 on page 149, the initial-boundary value problem (3.2) on page 119 has for any given control $u \in L^2(\Sigma)$ a unique weak solution $y \in W(0,T)$, represented by (cf. equation (3.36))

$$y = G_\Sigma(\beta\, u).$$

For evaluation of the cost functional the full information about y is not needed, only the final value $y(T)$. The *observation operator* $E_T : y \mapsto y(T)$ is a continuous and linear mapping from $W(0,T)$ into $L^2(\Omega)$, since the embedding $W(0,T) \hookrightarrow C\big([0,T], L^2(\Omega)\big)$ has these properties. Hence, for some constant $c > 0$, $\|y(T)\|_{L^2(\Omega)} \le \|y\|_{C([0,T],L^2(\Omega))} \le c\,\|y\|_{W(0,T)}$, and we have

$$y(T) = E_T\, G_\Sigma(\beta\, u) =: S\, u.$$

Again, S represents the part of the state that appears in the cost functional. In summary, the composition $u \mapsto y \mapsto y(T)$ is a continuous linear mapping

$$S : u \mapsto y(T)$$

from the control space $L^2(\Sigma)$ into the space $L^2(\Omega)$ that contains $y(T)$. Replacing the expression $y(T)$ in the cost functional (3.1) by $S\,u$, we eliminate the initial-boundary value problem (3.2) (of course, only theoretically). Then the optimal control problem (3.1)–(3.3) becomes a quadratic optimization problem in the Hilbert space $U = L^2(\Sigma)$:

$$\min_{u \in U_{ad}} f(u) := \frac{1}{2}\,\|Su - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Sigma)}, \tag{3.37}$$

where

$$U_{ad} = \big\{u \in L^2(\Sigma) \,:\, u_a(x,t) \le u(x,t) \le u_b(x,t) \quad \text{for a.e. } (x,t) \in \Sigma\big\}.$$

Obviously, the functional f is convex and continuous, and the admissible set U_{ad} is a nonempty, closed, bounded, and convex subset of the Hilbert space $L^2(\Sigma)$. Hence, we can infer from Theorem 2.14 on page 50 the following existence result.

Theorem 3.15. *Suppose that Assumption 3.14 on page 153 holds with* $E := \Sigma$. *Then the optimization problem* (3.37), *and hence the optimal nonstationary boundary temperature problem* (3.1)–(3.3) *on page 119, has at least one optimal control* $\bar u \in U_{ad}$. *If* $\lambda > 0$, *then* $\bar u$ *is uniquely determined.*

The problem just discussed is a *parabolic boundary control problem* with final-value cost functional. Next, we will use a similar approach to study a problem with distributed heat source control. For a change, we assume isolation at the boundary and minimize a functional that involves the evolution of the boundary temperature (*boundary observation*).

3.5.2. Optimal nonstationary heat source. This section deals with the optimal control problem

(3.38)
$$\min J(y,u) := \frac{1}{2}\iint_{\Sigma} \bigl|y(x,t) - y_\Sigma(x,t)\bigr|^2 \, ds(x)\,dt + \frac{\lambda}{2}\iint_{Q} |u(x,t)|^2 \, dx\,dt,$$

subject to

(3.39)
$$\boxed{\begin{aligned} y_t - \Delta y &= \beta\, u && \text{in } Q \\ \partial_\nu y &= 0 && \text{on } \Sigma \\ y(0) &= 0 && \text{in } \Omega \end{aligned}}$$

and

(3.40)
$$u_a(x,t) \le u(x,t) \le u_b(x,t) \qquad \text{for a.e. } (x,t) \in Q.$$

We have to find the *optimal heat source* u, aiming at the best possible approximation of a desired *evolution of the boundary temperatue* y_Σ, where the costs due to the control action are accounted for by the term $\frac{1}{2}\,\lambda\|u\|^2$. This can also be seen as an *inverse problem*: an unknown heat source u distributed within the body Ω has to be recovered from measurements of the temperature evolution $y|_\Sigma$ on the surface Γ. In this context, and also in the interpretation as an optimal control problem, $\lambda > 0$ plays the role of a *regularization parameter*.

By virtue of Theorem 3.12 on page 149 and Theorem 3.13 on page 150 applied with $f := \beta\, u$, $g = 0$, and $y_0 = 0$, there exists for any $u \in L^2(Q)$ a unique weak solution $y \in W(0,T)$ to the parabolic initial-boundary value problem (3.39). The mapping $u \mapsto y$ defines a linear (since $y(x,0) = 0$) and continuous operator from $L^2(Q)$ into $W(0,T)$ and, by the definition of $W(0,T)$, into $L^2\bigl(0,T;H^1(\Omega)\bigr)$ as well. With the control-to-state operator $G_Q : L^2(Q) \to W(0,T)$ defined in (3.36), it has the representation

$$y = G_Q(\beta\, u).$$

The evaluation of the cost functional only requires knowledge of the boundary values $y(x,t)|_\Sigma$. Since the trace operator $y \mapsto y|_\Gamma$ maps $H^1(\Omega)$ continuously into $L^2(\Gamma)$, the mapping $E_\Sigma : y \mapsto y|_\Sigma$ defines a continuous linear operator from $L^2\bigl(0,T;H^1(\Omega)\bigr)$ into $L^2\bigl(0,T;L^2(\Gamma)\bigr)$. Consequently, the mapping $u \mapsto y \mapsto y|_\Sigma$, i.e., the operator

$$S : u \mapsto y|_\Sigma,$$

maps the control space $L^2(Q)$ continuously into the space $L^2\big(0,T;L^2(\Gamma)\big) \cong L^2\big((0,T)\times\Gamma\big) = L^2(\Sigma)$ to which $y|_\Sigma$ belongs. With the operators thus defined, S takes the form

$$S\,u = E_\Sigma\, G_Q(\beta\, u). \tag{3.41}$$

Substituting $y = S\,u$ in the cost functional $J(y,u)$, we eliminate the parabolic initial-boundary value problem to arrive at the following quadratic minimization problem in the Hilbert space $U = L^2(Q)$:

$$\min_{u\in U_{ad}} f(u) := \frac{1}{2}\,\|Su - y_\Sigma\|^2_{L^2(\Sigma)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(Q)}. \tag{3.42}$$

Invoking Theorem 2.14, we conclude the existence of an optimal control, i.e., the solvability of the problem. We have thus shown the following result.

Theorem 3.16. *Suppose that Assumption 3.14 on page 153 holds with $E = Q$. Then the optimal nonstationary heat source problem* (3.38)–(3.40) *has at least one optimal control $\bar u \in U_{ad}$. If $\lambda > 0$, then $\bar u$ is uniquely determined.*

3.6. Necessary optimality conditions

In this section, we will derive the first-order necessary optimality conditions for the problems stated in Sections 3.5.1 and 3.5.2. First, a variational inequality will be derived that still involves the state y; then, y will be eliminated by means of the adjoint state to deduce a variational inequality for the control.

3.6.1. An auxiliary result for adjoint operators. Consider the parabolic problem

$$\begin{aligned} -p_t - \Delta p + c_0\, p &= a_Q \\ \partial_\nu p + \alpha\, p &= a_\Sigma \\ p(\cdot,T) &= a_\Omega, \end{aligned} \tag{3.43}$$

with bounded and measurable coefficient functions c_0 and α and prescribed functions $a_Q \in L^2(Q)$, $a_\Sigma \in L^2(\Sigma)$, and $a_\Omega \in L^2(\Omega)$. We define the bilinear form

$$a[t;y,v] := \int_\Omega \big(\nabla y\cdot\nabla v + c_0(\cdot,t)\,y\,v\big)\,dx + \int_\Gamma \alpha(\cdot,t)\,y\,v\,ds.$$

We have the following well-posedness result.

Lemma 3.17. *The parabolic problem* (3.43) *has a unique weak solution* $p \in W_2^{1,0}(Q)$*, which is the solution to the variational problem*

$$\begin{aligned}&\iint_Q p\, v_t \, dx\, dt \;+ \int_0^T a[t; p, v]\, dt\\ &= \int_\Omega a_\Omega\, v(T)\, dx \;+ \iint_Q a_Q\, v\, dx\, dt + \iint_\Sigma a_\Sigma\, v\, ds\, dt\\ &\forall\, v \in W_2^{1,1}(Q) \;\; \textit{with} \;\; v(\cdot, 0) = 0.\end{aligned}$$

We have $p \in W(0,T)$*, and there is a constant* $c_a > 0$*, which does not depend on the given functions, such that*

$$\|p\|_{W(0,T)} \le c_a \left(\|a_Q\|_{L^2(Q)} + \|a_\Sigma\|_{L^2(\Sigma)} + \|a_\Omega\|_{L^2(\Omega)}\right).$$

Proof: Let $\tau \in [0,T]$, and put $\tilde{p}(\tau) := p(T-\tau)$ and $\tilde{v}(\tau) := v(T-\tau)$. Then $\tilde{p}(0) = p(T)$, $\tilde{p}(T) = p(0)$, $\tilde{v}(0) = v(T)$, $\tilde{v}(T) = v(0)$, $\tilde{a}_Q(\cdot, t) := a_Q(\cdot, T-\tau)$, etc., and also

$$\iint_Q p\, v_t \, dx\, dt = -\iint_Q \tilde{p}\, \tilde{v}_\tau \, dx\, d\tau,$$

and so on. Consequently, the asserted variational formulation is equivalent to the definition of the weak solution to the (forward) parabolic initial-boundary value problem

$$\begin{aligned}\tilde{p}_\tau - \Delta\tilde{p} + c_0\, \tilde{p} &= \tilde{a}_Q\\ \partial_\nu \tilde{p} + \alpha\, \tilde{p} &= \tilde{a}_\Sigma\\ \tilde{p}(0) &= a_\Omega.\end{aligned}$$

By Theorem 3.9 on page 140, there is a unique weak solution $\tilde{p}$, which by Theorem 3.12 on page 149 belongs to $W(0,T)$. The assertion now follows from reversing the time transformation. □

Since $p \in W(0,T)$, we can, in analogy to equation (3.35) on page 152, rewrite after an integration by parts the variational formulation of the adjoint equation in the following shorter form:

(3.44)

$$\boxed{\begin{aligned}\int_0^T \left\{-\big(p_t\,,\, v\big)_{V^*,V} + a[t; p, v]\right\} dt &= \iint_Q a_Q\, v\, dx\, dt + \iint_\Sigma a_\Sigma\, v\, ds\, dt\\ &\qquad \forall\, v \in L^2(0,T;V)\\ p(T) &= a_\Omega.\end{aligned}}$$

Like in the elliptic case, for the derivation of the adjoint system we need the following somewhat technical result.

Theorem 3.18. *Let $y \in W(0,T)$ be the solution to the parabolic problem*

$$\begin{aligned} y_t - \Delta y + c_0\, y &= b_Q\, v \\ \partial_\nu y + \alpha\, y &= b_\Sigma\, u \\ y(0) &= b_\Omega\, w, \end{aligned}$$

with coefficient functions $c_0, b_Q \in L^\infty(Q)$, $\alpha, b_\Sigma \in L^\infty(\Sigma)$, and $b_\Omega \in L^\infty(\Omega)$ and controls $v \in L^2(Q)$, $u \in L^2(\Sigma)$, and $w \in L^2(\Omega)$. Moreover, let square integrable functions a_Ω, a_Q, and a_Σ be given, and let $p \in W(0,T)$ be the weak solution to (3.43)*. Then we have*

$$\begin{aligned} &\int_\Omega a_\Omega\, y(\cdot,T)\, dx + \iint_Q a_Q\, y\, dx\, dt + \iint_\Sigma a_\Sigma\, y\, ds\, dt \\ &= \iint_\Sigma b_\Sigma\, p\, u\, ds\, dt + \iint_Q b_Q\, p\, v\, dx\, dt + \int_\Omega b_\Omega\, p(\cdot,0)\, w\, dx. \end{aligned}$$

Proof: We use the variational formulations for y and p. For y, using the test function p, we have

(3.45)
$$\int_0^T \Big\{ \big(y_t\,,\, p\big)_{V^*,V} + a[t;y,p] \Big\}\, dt = \iint_Q b_Q\, p\, v\, dx\, dt + \iint_\Sigma b_\Sigma\, p\, u\, ds\, dt,$$

with the initial condition $y(0) = b_\Omega\, w$. Analogously, taking y as test function in the equation for p, we find that

$$\int_0^T \Big\{ -\big(p_t\,,\, y\big)_{V^*,V} + a[t;p,y] \Big\}\, dt = \iint_Q a_Q\, y\, dx\, dt + \iint_\Sigma a_\Sigma\, y\, ds\, dt, \tag{3.46}$$

with the final condition $p(T) = a_\Omega$. Integrating by parts in (3.45), we obtain

$$\begin{aligned} \int_0^T \Big\{ -\big(p_t\,,\, y\big)_{V^*,V} + a[t;y,p] \Big\}\, dt \;&=\; -\big(y(T)\,,\, a_\Omega\big)_{L^2(\Omega)} \\ &+ \big(b_\Omega\, w\,,\, p(0)\big)_{L^2(\Omega)} + \iint_Q b_Q\, p\, v\, dx\, dt + \iint_\Sigma b_\Sigma\, p\, u\, ds\, dt. \end{aligned} \tag{3.47}$$

Since the left-hand sides of this equation and equation (3.46) coincide, the right-hand sides of (3.46) and (3.47) must also be equal, hence the assertion follows. □

3.6.2. Optimal nonstationary boundary temperature. In this section, we determine the necessary optimality conditions for the problem (3.1)–(3.3) on page 119:

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

subject to

$$\boxed{\begin{aligned} y_t - \Delta y &= 0 \\ \partial_\nu y + \alpha\, y &= \beta\, u \\ y(0) &= y_0 \end{aligned}}$$

and

$$u_a \le u \le u_b.$$

In this problem, the initial state $y_0 \in L^2(\Omega)$ may differ from zero, a situation we have avoided so far for the sake of simpler exposition. From the previous section, we know that in the case of $y_0 = 0$ the problem has an optimal control $\bar u$ with associated state $\bar y$. The reader is invited to show the corresponding result for $y_0 \neq 0$ in Exercise 3.2 on page 178.

We now determine the form of the adjoint problem. This could easily be done by means of the formal Lagrange method. However, we are sufficiently experienced by now to be able to deduce the correct form directly: each of the terms occurring in the derivative of the cost functional J with respect to y appears as a right-hand side in the adjoint system. The domains of definition of these terms have to coincide with the domains in which the corresponding condition in the adjoint system is valid. In the present case, the derivative of the cost functional with respect to y is given by the function $\bar y(T) - y_\Omega$, which is defined in Ω. Hence, the derivative must appear in the condition of the adjoint system that has to be satisfied in Ω. Keeping in mind that the adjoint system ought to be a parabolic final value problem backwards in time, in which only the final value $p(T)$ has Ω as its domain, we conclude that the *adjoint state* p associated with $\bar y$ must solve the following *adjoint system*:

$$\boxed{\begin{aligned} -p_t - \Delta p &= 0 && \text{in } Q \\ \partial_\nu p + \alpha\, p &= 0 && \text{on } \Sigma \\ p(T) &= \bar y(T) - y_\Omega && \text{in } \Omega. \end{aligned}} \tag{3.48}$$

Theorem 3.19. *Let $\bar u \in U_{ad}$ be a control with associated state $\bar y$, and let $p \in W(0,T)$ be the corresponding adjoint state that solves* (3.48). *Then $\bar u$*

is an optimal control for the optimal nonstationary boundary temperature problem (3.1)–(3.3) *on page 119 if and only if the variational inequality*

$$\iint_\Sigma \big(\beta(x,t)\,p(x,t) + \lambda\,\bar u(x,t)\big)\,\big(u(x,t) - \bar u(x,t)\big)\,ds(x)dt \geq 0 \tag{3.49}$$

holds for all $u \in U_{ad}$.

Proof: Let $S : L^2(\Sigma) \to L^2(\Omega)$ be the continuous linear operator that, for the homogeneous initial condition $y_0 = 0$, assigns to each control u the final value $y(T)$ of the weak solution y to the state equation. Moreover, let $\hat y = G_0\,y_0$ denote the weak solution corresponding to $y_0 \neq 0$ and $u = 0$. Then it follows from the superposition principle for linear equations that

$$y(T) - y_\Omega = S\,u + (G_0\,y_0)(T) - y_\Omega = S\,u - z,$$

where $z := y_\Omega - \big(G_0\,y_0\big)(T)$. The above control problem then takes the form

$$\min_{u\in U_{ad}} f(u) := \frac{1}{2}\,\|Su - z\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Sigma)}.$$

The general variational inequality (2.47) on page 65 yields, for all $u \in U_{ad}$,

$$\begin{aligned} 0 &\leq \big(S\bar u - z\,,\,S(u-\bar u)\big)_{L^2(\Omega)} + \lambda\,\big(\bar u\,,\,u-\bar u\big)_{L^2(\Sigma)} \\ &= \int_\Omega \big(\bar y(T) - y_\Omega\big)\big(y(T) - \bar y(T)\big)\,dx + \lambda \iint_\Sigma \bar u\,(u-\bar u)\,ds\,dt. \end{aligned} \tag{3.50}$$

Here, we have used the identity

$$Su - S\bar u = Su + \big(G_0\,y_0\big)(T) - \big(G_0\,y_0\big)(T) - S\bar u = y(T) - \bar y(T)$$

and, once more, that $z = y_\Omega - \big(G_0\,y_0\big)(T)$. Now put $\tilde y := y - \bar y$ and apply Theorem 3.18 with the specifications $a_\Omega = \bar y(T) - y_\Omega$, $a_Q = 0$, $a_\Sigma = 0$, $b_\Sigma = \beta$, $v = 0$, $w = 0$, $y := \tilde y$, and $\tilde u := u - \bar u$. Note that $w = 0$ in this situation since, by definition, $Su(0) = 0$; the part originating from y_0 is incorporated into z. It follows that

$$\big(\bar y(T) - y_\Omega\,,\,\tilde y(T)\big)_{L^2(\Omega)} = \iint_\Sigma \beta\,p\,\tilde u\,ds\,dt.$$

Substituting this result in inequality (3.50), we find that

$$\begin{aligned} 0 &\leq \int_\Omega \big(\bar y(T) - y_\Omega\big)\big(y(T) - \bar y(T)\big)\,dx + \lambda \iint_\Sigma \bar u\,(u-\bar u)\,ds\,dt \\ &= \iint_\Sigma \beta\,p\,(u-\bar u)\,ds\,dt + \lambda \iint_\Sigma \bar u\,(u-\bar u)\,ds\,dt \\ &= \iint_\Sigma (\beta\,p + \lambda\,\bar u)\,(u-\bar u)\,ds\,dt, \end{aligned}$$

which concludes the proof of the assertion. □

Next, we employ the method described on page 69 for elliptic problems to derive a number of results concerning the possible form of optimal controls.

Theorem 3.20. *A control $\bar u \in U_{ad}$ with associated state $\bar y$ is optimal for the problem* (3.1)–(3.3) *on page 119 if and only if it satisfies, together with the adjoint state p from* (3.48)*, the following conditions for almost all $(x,t) \in \Sigma$: the* weak minimum principle

$$\big(\beta(x,t)\,p(x,t) + \lambda\,\bar u(x,t)\big)\big(v - \bar u(x,t)\big) \ge 0 \quad \forall v \in \big[u_a(x,t), u_b(x,t)\big], \tag{3.51}$$

the minimum principle

$$\beta(x,t)\,p(x,t)\,\bar u(x,t) + \frac{\lambda}{2}\,\bar u(x,t)^2 = \min_{v\in[u_a(x,t),u_b(x,t)]} \Big\{\beta(x,t)\,p(x,t)\,v + \frac{\lambda}{2}v^2\Big\}, \tag{3.52}$$

and, in the case of $\lambda > 0$, the projection formula

$$\bar u(x,t) = \mathbb{P}_{[u_a(x,t),u_b(x,t)]}\left\{-\frac{1}{\lambda}\beta(x,t)\,p(x,t)\right\}. \tag{3.53}$$

Proof: For the proof of this assertion one takes, starting from the variational inequality (3.49), the same series of steps that led from Theorem 2.25 on page 67 to Theorem 2.27 on page 69 in the elliptic case. □

Conclusion. In the $\lambda > 0$ case, the triple $(\bar u, \bar y, p)$ satisfies the *optimality system*

$$\boxed{\begin{array}{rcl@{\qquad}rcl} y_t - \Delta y &=& 0 & -p_t - \Delta p &=& 0 \\ \partial_\nu y + \alpha\, y &=& \beta\, u & \partial_\nu p + \alpha\, p &=& 0 \\ y(0) &=& y_0 & p(T) &=& y(T) - y_\Omega \end{array} \qquad u = \mathbb{P}_{[u_a,u_b]}\Big\{-\frac{1}{\lambda}\beta\, p\Big\}.} \tag{3.54}$$

If $\lambda = 0$, then the projection formula has to be replaced by:

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

Special case: $u_a = -\infty$, $u_b = \infty$ (no control constraints). In this case, the projection formula yields $u = -\lambda^{-1}\beta\, p$. Hence, u can be eliminated from the state equation, and we obtain the following *forward-backward system* of two parabolic problems for the unknown functions y and p:

$$
\boxed{
\begin{array}{rcl@{\qquad}rcl}
y_t - \Delta y &=& 0 & -p_t - \Delta p &=& 0 \\
\partial_\nu y + \alpha\, y &=& -\beta^2\, \lambda^{-1} p & \partial_\nu p + \alpha\, p &=& 0 \\
y(0) &=& y_0 & p(T) &=& y(T) - y_\Omega.
\end{array}
} \tag{3.55}
$$

The solution of such systems is not an easy task. Similar equations arise as Hamiltonian systems in the optimal control of ordinary differential equations, where they are solved by means of (multiple) shooting techniques. In the case of partial differential equations, the large number of variables originating from the spatial discretization presents an additional challenge that makes the direct solution of the above optimality system difficult. One possible approach is to apply multigrid methods; see, e.g., Borzi [**Bor03**] and Borzi and Kunisch [**BK01**]. Direct solution as an elliptic system is also promising (see the recommendations concerning numerical methods starting on page 170).

3.6.3. Optimal nonstationary heat source. We recall problem (3.38)–(3.40) on page 155, which in shortened form reads

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

subject to $u \in U_{ad}$ and to the state system

$$
\boxed{
\begin{array}{rcl}
y_t - \Delta y &=& \beta\, u \\
\partial_\nu y &=& 0 \\
y(0) &=& 0.
\end{array}
}
$$

Invoking the operator $G_Q : L^2(Q) \to W(0,T)$ introduced in (3.36) on page 152, we can express the solution y to the state system in the form

$$y = G_Q(\beta\, u).$$

The cost functional involves the observation $y_{|\Sigma}$. The observation operator $E_\Sigma : y \mapsto y_{|\Sigma}$ is a continuous linear mapping from $W(0,T)$ into $L^2\big(0,T;L^2(\Gamma)\big) \cong L^2(\Sigma)$, which entails that the *control-to-observation operator* $S : u \mapsto y_{|\Sigma}$ defined in (3.41) on page 156 is continuous from $L^2(Q)$ into $L^2(\Sigma)$. Hence, the problem is equivalent to the problem $\min_{u\in U_{ad}} f(u)$, where f is the reduced functional introduced in (3.42) on page 156. As in (3.50), we obtain the following as the necessary optimality condition for $\bar u$:

$$0 \le \big(S\bar u - y_\Sigma\,,\, Su - S\bar u\big)_{L^2(\Sigma)} + \lambda\,\big(\bar u\,,\, u - \bar u\big)_{L^2(Q)} \quad \forall\, u \in U_{ad},$$

that is, upon substituting $\bar{y}_{|\Sigma} = S\bar{u}$ and $y_{|\Sigma} = Su$,

$$0 \le \iint_\Sigma (\bar{y} - y_\Sigma)(y - \bar{y})\, ds\, dt + \lambda \iint_Q \bar{u}\,(u - \bar{u})\, dx\, dt \quad \forall\, u \in U_{ad}. \tag{3.56}$$

It is evident how the adjoint state p must be defined, namely as the weak solution to the parabolic problem

$$\boxed{\begin{aligned} -p_t - \Delta p &= 0 \\ \partial_\nu p &= \bar{y} - y_\Sigma \\ p(T) &= 0. \end{aligned}}$$

By virtue of Lemma 3.17, it has a unique weak solution p.

Theorem 3.21. *A control $\bar{u} \in U_{ad}$ is optimal for the optimal nonstationary heat source problem* (3.38)–(3.40) *on page 155 if and only if it satisfies, together with the adjoint state p defined above, the variational inequality*

$$\iint_Q (\beta\, p + \lambda\, \bar{u})(u - \bar{u})\, dx\, dt \ge 0 \quad \forall\, u \in U_{ad}.$$

Proof: This assertion is again a direct consequence of Theorem 3.18, with the specifications $a_\Sigma = \bar{y} - y_\Sigma$, $a_\Omega = 0$, $a_Q = 0$, $b_Q = \beta$, $b_\Sigma = 0$, and $b_Q = 0$. The steps are similar to those in the proof of Theorem 3.20. □

As in Section 3.6.2, the variational inequality just proved can be transformed into an equivalent pointwise minimum principle for $\bar{u}$ or, if $\lambda > 0$, into a projection formula. In particular, if $\lambda > 0$,

$$\bar{u}(x,t) = \mathbb{P}_{[u_a(x,t),u_b(x,t)]} \left\{ -\frac{1}{\lambda}\, \beta(x,t)\, p(x,t) \right\} \quad \text{for a.e. } (x,t) \in Q.$$

3.6.4. Differential operators in divergence form *.

Statement of the problem and existence of optimal controls. The parabolic problems studied so far have been comparatively simple, since we have confined ourselves for methodological reasons to the Laplacian. However, the theory can easily be extended to more general equations. To this end, we recall the uniformly elliptic differential operator in divergence form introduced on page 37,

$$\mathcal{A}\, y(x) = -\sum_{i,j=1}^{N} D_i\big(a_{ij}(x)\, D_j\, y(x)\big),$$

under the assumptions made there. We consider the optimal control problem

(3.57)
$$\begin{aligned}\min\, J(y,v,u) \;&=\; \frac{\lambda_\Omega}{2}\,\|y(T)-y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda_Q}{2}\,\|y-y_Q\|^2_{L^2(Q)} \\ &\quad + \frac{\lambda_\Sigma}{2}\,\|y-y_\Sigma\|^2_{L^2(\Sigma)} + \frac{\lambda_v}{2}\,\|v\|^2_{L^2(Q)} + \frac{\lambda_u}{2}\,\|u\|^2_{L^2(\Sigma)},\end{aligned}$$

subject to the parabolic initial-boundary value problem

(3.58)
$$\boxed{\begin{aligned} y_t + \mathcal{A}\,y + c_0\,y &= \beta_Q\,v && \text{in } Q \\ \partial_{\nu_\mathcal{A}} y + \alpha\,y &= \beta_\Sigma\,u && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega \end{aligned}}$$

and the control constraints

(3.59)
$$\begin{aligned} v_a(x,t) \;\le\; v(x,t) \;\le\; v_b(x,t) \quad &\text{for a.e. } (x,t)\in Q \\ u_a(x,t) \;\le\; u(x,t) \;\le\; u_b(x,t) \quad &\text{for a.e. } (x,t)\in \Sigma. \end{aligned}$$

Here, in addition to the quantities introduced in Assumption 3.14 on page 153, we are given nonnegative constants λ_Ω, λ_Q, λ_Σ, λ_v, and λ_u, as well as coefficient functions $\beta_Q \in L^\infty(Q)$ and $\beta_\Sigma \in L^\infty(\Sigma)$.

For clearer exposition, the (x,t) dependence of all functions has been suppressed in the parabolic problem. Observe also that the operator $\mathcal{A}$ itself does not depend on t; for simplicity, we have dispensed with the time dependence of the coefficient functions a_{ij}. Note, however, that the existence result of Theorem 5.1 in Ladyzhenskaya et al. [**LSU68**] allows time-dependent coefficients $a_{ij}(x,t)$ under appropriate smoothness assumptions; see also Wloka [**Wlo87**]. The definition of weak solutions to problem (3.58) reads as follows:

Definition. *A function $y \in W_2^{1,0}(Q)$ is said to be a* weak solution *to* (3.58) *if the following variational equation holds for all $w \in W_2^{1,1}(Q)$ such that $w(\cdot,T)=0$:*

$$\begin{aligned}\iint_Q y(x,t)\,w_t(x,t)\,dx\,dt \;&=\; \iint_Q \sum_{i,j=1}^{N} a_{ij}(x)\,D_i y(x,t)\,D_j w(x,t)\,dx\,dt \\ &\quad + \iint_Q \big(c_0(x,t)\,y(x,t) - \beta_Q(x,t)\,v(x,t)\big)\,w(x,t)\,dx\,dt \\ &\quad + \iint_\Sigma \big(\alpha(x,t)\,y(x,t) - \beta_\Sigma(x,t)\,u(x,t)\big)\,w(x,t)\,ds(x)dt \\ &\quad - \int_\Omega y_0(x)\,w(x,0)\,dx.\end{aligned}$$

The previous definition of weak solutions is evidently a special case. We now introduce the family of bilinear forms $a[t;\cdot,\cdot] : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ for $t \in [0,T]$,

$$\begin{aligned} a[t;y,w] &= \int_\Omega \sum_{i,j=1}^{N} \big(a_{ij}(x)\, D_i y(x)\, D_j w(x) + c(x,t)\, y(x)\, w(x) \big)\, dx \\ &\quad + \int_\Gamma \alpha(x,t)\, y(x)\, w(x)\, ds(x). \end{aligned}$$

Then the above weak formulation can be rewritten in the following somewhat simpler form: for all $w \in W_2^{1,1}(Q)$ such that $w(T) = 0$, we have, with $H = L^2(\Omega)$,

(3.60)
$$\begin{aligned} \int_0^T \big\{ -\big(y(t)\,,\, w_t(t)\big)_H + a[t;y(t),w(t)] \big\}\, dt &= \int_0^T \big(\beta_Q(t)\, v(t)\,,\, w(t)\big)_H\, dt \\ &\quad + \int_0^T \big(\beta_\Sigma(t)\, u(t)\,,\, w(t)\big)_{L^2(\Sigma)}\, dt + \big(y_0\,,\, w(0)\big)_H\,. \end{aligned}$$

Again, we have suppressed the dependence on x.

By virtue of Theorem 7.9 on page 373, under the above assumptions the initial-boundary value problem (3.58) has for any triple $(v,u,y_0) \in L^2(Q) \times L^2(\Sigma) \times L^2(\Omega)$ a unique weak solution $y \in W_2^{1,0}(Q)$. Moreover, there is a constant $c_p > 0$, which does not depend on the choice of (v,u,y_0), such that

$$\|y\|_{W(0,T)} \le c_p \Big(\|v\|_{L^2(Q)} + \|u\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)} \Big). \tag{3.61}$$

As in Theorem 3.12, it can be shown that (possibly after a modification on a set of zero measure) y belongs to $W(0,T)$. Hence, the mapping $(v,u,y_0) \mapsto y$ defines a continuous linear operator from $L^2(Q) \times L^2(\Sigma) \times L^2(\Omega)$ into $W(0,T)$. In particular, the following mappings are continuous:

$$\begin{aligned} (v,u,y_0) &\mapsto y && \text{from } L^2(Q) \times L^2(\Sigma) \times L^2(\Omega) && \text{into } L^2(Q), \\ (v,u,y_0) &\mapsto y|_\Sigma && \text{from } L^2(Q) \times L^2(\Sigma) \times L^2(\Omega) && \text{into } L^2(\Sigma), \\ (v,u,y_0) &\mapsto y(T) && \text{from } L^2(Q) \times L^2(\Sigma) \times L^2(\Omega) && \text{into } L^2(\Omega). \end{aligned}$$

With this information in hand, we can argue as in the preceding sections to arrive at the following result.

Conclusion. *Under the above assumptions, the control problem* (3.57)–(3.59) *has optimal controls* $\bar v$ *and* $\bar u$. *These are uniquely determined if one of the following conditions is fulfilled: either* $\lambda_v > 0$ *and* $\lambda_u > 0$, *or* $\lambda_Q > 0$ *and* β_Q *and* β_Σ *are nonzero almost everywhere.*

Remark. As in Section 2.3.3, it is possible to subdivide the boundary as $\Gamma = \Gamma_0 \cup \Gamma_1$ and prescribe homogeneous boundary data for y on Γ_0. We leave it to the reader to work out the details.

Necessary optimality conditions. We state the optimality conditions for problem (3.57)–(3.59) without proof, since the lines of argument follow closely those of the problems discussed previously. Let us consider optimal controls $\bar{v} \in L^2(Q)$ and $\bar{u} \in L^2(\Sigma)$ with associated state $\bar{y}$. The corresponding adjoint state $\bar{p}$ is the unique weak solution to the adjoint system

$$\boxed{\begin{aligned} -p_t + \mathcal{A}\, p + c_0\, p &= \lambda_Q\,(\bar{y} - y_Q) && \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} p + \alpha\, p &= \lambda_\Sigma\,(\bar{y} - y_\Sigma) && \text{in } \Sigma \\ p(T) &= \lambda_\Omega\,\big(\bar{y}(T) - y_\Omega\big) && \text{in } \Omega. \end{aligned}}$$

The necessary and sufficient optimality condition is given by the variational inequalities

$$\begin{aligned} \iint_Q \big(\beta_Q(x,t)\, p(x,t) + \lambda_v\, \bar{v}(x,t)\big)\big(v(x,t) - \bar{v}(x,t)\big)\, dx\, dt &\geq 0 \quad \forall v \in V_{ad} \\ \iint_\Sigma \big(\beta_\Sigma(x,t)\, p(x,t) + \lambda_u\, \bar{u}(x,t)\big)\big(u(x,t) - \bar{u}(x,t)\big)\, ds\, dt &\geq 0 \quad \forall u \in U_{ad}, \end{aligned}$$

which again can be expressed in terms of pointwise relations or projection formulas. Here, U_{ad} is defined as before, and V_{ad} is the set of all $v \in L^2(Q)$ that respect the constraint $v_a(x,t) \leq v(x,t) \leq v_b(x,t)$ for almost every $(x,t) \in Q$.

3.7. Numerical methods

We first explain the projected gradient method and the formulation of a finite-dimensional reduced problem, since these techniques are easily implemented for tests. Even today, gradient techniques are still the method of choice for complex problems, for instance in three-dimensional spatial domains. At the end of this section we also provide the reader with a brief overview of recommended, more efficient methods.

3.7.1. Projected gradient methods. We sketch the projected gradient method for the problem of finding the optimal nonstationary boundary temperature. Parabolic problems with distributed control can be treated similarly; see also the elliptic case with distributed control. We have to solve the optimal control problem

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

subject to $u_a \le u \le u_b$ and

$$\boxed{\begin{aligned} y_t - \Delta y &= 0 \\ \partial_\nu y + \alpha\, y &= \beta\, u \\ y(0) &= y_0. \end{aligned}} \tag{3.62}$$

As before, let $S : L^2(\Sigma) \to L^2(\Omega),\ u \mapsto y(T)$, be the operator that, for the homogeneous initial datum $y_0 = 0$, assigns to each control u the final value $y(T)$ of the solution to the above initial-boundary value problem. We also denote by $\hat{y}$ the solution that corresponds to the inhomogeneous initial datum y_0 and the control $u = 0$. Evidently,

$$y(x,T) = (Su)(x) + \hat{y}(x,T),$$

and the optimal control problem becomes a quadratic Hilbert space optimization problem,

$$\min_{u\in U_{ad}} f(u) := \frac{1}{2}\,\|Su + \hat{y}(T) - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Sigma)}.$$

The derivative of f at an iterate u_n is given by

$$f'(u_n)\,v = \iint_\Sigma \big(\beta(x,t)\,p_n(x,t) + \lambda\,u_n(x,t)\big)\,v(x,t)\;ds\,dt,$$

where p_n is the solution to the following adjoint equation:

$$\begin{aligned} -p_t - \Delta p &= 0 \\ \partial_\nu p + \alpha p &= 0 \\ p(T) &= y_n(T) - y_\Omega. \end{aligned} \tag{3.63}$$

By the Riesz representation theorem, we obtain the usual representation of the reduced gradient,

$$f'(u_n) = \beta\,p_n + \lambda\,u_n.$$

The algorithm proceeds as follows: suppose that the iterates $u_1, \dots, u_n$ have already been determined.

S1 (*New state*) Solve the state system (3.62) with $u := u_n$ for $y =: y_n$.

S2 (*New descent direction*) Calculate the associated adjoint state p_n by solving (3.63). Take as descent direction the negative gradient

$$v_n = -f'(u_n) = -(\beta\,p_n + \lambda\,u_n).$$

S3 (*Step size control*) Determine the optimal step size s_n by solving

$$f(\mathbb{P}_{[u_a,u_b]}\{u_n + s_n\, v_n\}) = \min_{s>0} f\big(\mathbb{P}_{[u_a,u_b]}\{u_n + s\, v_n\}\big).$$

S4 Put $u_{n+1} := \mathbb{P}_{[u_a,u_b]}\{u_n + s_n\, v_n\}$, $n := n+1$, GO TO **S1**.

The method is completely analogous to that for the elliptic case. The projection step is necessary, since $u_n + s_n\, v_n$ may not be admissible. Although the method converges only slowly, it is easy to implement and thus very suitable for numerical tests. Note also that parabolic problems require much more computational effort than elliptic ones, since we have time as an additional variable. Therefore, gradient methods are still useful alternatives to methods with higher order of convergence. For a detailed analysis of the method, we refer the reader to [**GS80**] and [**HPUU09**].

3.7.2. Derivation of the reduced problem. If problem (3.62) has to be solved several times, say for different initial values y_0, final values y_Ω, or regularization parameters λ, then reducing the problem to a statement involving u only may be worthwhile. This is also the case when the control has the form (3.64) with only a few functions e_i.

For simplicity and clarity, we imagine (as in the elliptic case) that the state y can be determined exactly for a given control u and that all integrals that arise can be evaluated exactly.

As in the elliptic case, we assume that the control $u = u(x,t)$ is, in terms of fixed ansatz functions e_i, of the form

$$u(x,t) = \sum_{i=1}^{m} u_i\, e_i(x,t). \tag{3.64}$$

Example. Let $N = 2$ and $\Omega = (0,1)^2$, and imagine that the (one-dimensional) boundary Γ is unrolled onto the interval $[0,4]$ on the real axis. If x varies over the boundary Γ, then we can interpret u as a function of the arc length s and the time t. Then u is defined on the rectangle $[0,4] \times [0,T]$, which we split up into $n_s \cdot n_t$ nonoverlapping subrectangles.

We take the control u to be a step function that is constant on each of the subrectangles. In this case, each basis function e_i equals unity on exactly one subrectangle and is zero otherwise; evidently, we have $m = n_s \cdot n_t$ different basis functions. As mentioned before, the above ansatz can be given a priori without discretization; this is often the case in practical applications. ◇

The functions $y_i(x,T) := (Se_i)(x)$, $i = 1,\dots,m$, have to be computed in advance as the final values of the solutions $y = y_i$ to the initial-boundary

value problems

$$\begin{aligned} y_t &= \Delta y \\ \partial_\nu y + \alpha\, y &= \beta\, e_i \\ y(0) &= 0. \end{aligned}$$

In addition, we need to calculate $\hat{y}(x,T)$, the final value of the solution to the initial-boundary value problem

$$\begin{aligned} \hat{y}_t &= \Delta \hat{y} \\ \partial_\nu \hat{y} + \alpha\, \hat{y} &= 0 \\ \hat{y}(0) &= y_0. \end{aligned}$$

Substituting this representation into the cost functional, we obtain the reduced cost function f_m that depends only on the vector $\vec{u} = (u_1, \ldots, u_m)^\top$,

$$f_m(\vec{u}) = \frac{1}{2} \left\| \sum_{i=1}^m u_i\, y_i(T) - y_\Omega + \hat{y}(T) \right\|^2_{L^2(\Omega)} + \frac{\lambda}{2} \left\| \sum_{i=1}^m u_i\, e_i \right\|^2_{L^2(\Sigma)}.$$

With these specifications, the finite-dimensional approximation of the original optimal control problem reads

$$(P_m) \qquad \min_{\vec{u}_a \le \vec{u} \le \vec{u}_b} f_m(\vec{u}).$$

Here, the inequality $\vec{u}_a \le \vec{u} \le \vec{u}_b$ is to be understood in the componentwise sense. We also tacitly assume that the restriction $u_a \le u_i \le u_b$ by means of constants u_a and u_b is a meaningful substitute for the original constraint $u_a \le u(x,t) \le u_b$. This is certainly the case for the step functions e_i introduced in the above example. Alternatively, the restrictions $u_a \le u_i \le u_b$ can be postulated a priori in the case of more general ansatz functions e_i.

A straightforward calculation yields

$$\begin{aligned} f_m(\vec{u}) &= \frac{1}{2} \|\hat{y}(T) - y_\Omega\|^2_{L^2(\Omega)} + \sum_{i=1}^m u_i \left(\hat{y}(T) - y_\Omega\,,\, y_i(T)\right)_{L^2(\Omega)} \\ &\quad + \frac{1}{2} \sum_{i,j=1}^m u_i u_j \left(y_i(T)\,,\, y_j(T)\right)_{L^2(\Omega)} + \frac{\lambda}{2} \sum_{i,j=1}^m u_i u_j \left(e_i\,,\, e_j\right)_{L^2(\Sigma)}. \end{aligned}$$

Consequently, (P_m) is—up to the constant $\frac{1}{2}\|\hat{y}(T) - y_\Omega\|^2_{L^2(\Omega)}$—equivalent to the finite-dimensional *reduced quadratic optimization problem*

$$(3.65) \qquad \boxed{\begin{array}{l} \min \left\{ \frac{1}{2} \vec{u}^\top \big(C + \lambda D\big)\, \vec{u} + \vec{a}^\top\, \vec{u} \right\} \\ \\ \vec{u}_a \;\le\; \vec{u} \;\le\; \vec{u}_b. \end{array}}$$

In order to set up this problem, the following quantities must be computed beforehand:

$$\begin{aligned} \vec{a} &= (a_i), \quad & a_i &= \big(\hat{y}(T) - y_\Omega\,,\, y_i(T)\big)_{L^2(\Omega)} \\ C &= (c_{ij}), \quad & c_{ij} &= \big(y_i(T)\,,\, y_j(T)\big)_{L^2(\Omega)} \\ D &= (d_{ij}), \quad & d_{ij} &= \big(e_i\,,\, e_j\big)_{L^2(\Sigma)}. \end{aligned}$$

In the example of the step functions, we have $(e_i\,,\,e_j)_{L^2(\Sigma)} = \delta_{ij}\,\|e_i\|^2_{L^2(\Sigma)}$, and D is a diagonal matrix, $D = \operatorname{diag}\big(\|e_i\|^2_{L^2(\Sigma)}\big)$.

The reduced problem can be solved by using a standard code of quadratic optimization, e.g., quadprog in MATLAB. Alternatively, an active-set strategy could be implemented. The internet website of NEOS (NEOS Server for Optimization) contains a list of other available codes.

The approach described here is worthwhile only if m is not too large, since the determination of all of the involved quantities requires the numerical solution of $m+1$ parabolic initial-boundary value problems. In principle, the method works as long as the numerical solution of the heat equation can be handled, that is, possibly also for spatially three-dimensional domains. It can be recommended whenever u is a priori given as the linear combination (3.64) of a small number of ansatz functions with unknown (control) components u_i.

Recommended numerical methods. We now give an overview of some methods that have been applied successfully to numerous problems and can be recommended for the numerical solution of linear-quadratic parabolic optimal control problems.

The unconstrained case. If the choice of the control u is unrestricted, then the coupled system (3.55) on page 162 consisting of state and adjoint systems can be solved, provided it can be handled numerically. This does not present major difficulties in one-dimensional domains. Many available codes can also handle simple two-dimensional geometries. In the case of domains in higher dimensions, multigrid techniques can be tried; see [**Bor03**] or [**BK01**].

Primal-dual active set strategies. These techniques are among the most often used methods in recent years; see Ito and Kunisch [**IK00**], Bergounioux et al. [**BIK99**], Kunisch and Rösch [**KR02**], as well as the detailed exposition in Ito and Kunisch [**IK08**]. The paper [**KR02**] treats the case of a general continuous linear mapping $u \mapsto y = S\,u$ and therefore covers parabolic problems, while the other references deal only with elliptic problems.

As in the elliptic case, at each iteration step one updates active sets for the upper and lower box constraints; the control is fixed in the next step

by taking the corresponding upper or lower threshold value. To determine the remaining values, one again solves a forward-backward parabolic system without constraints.

Direct solution of the optimality system (3.54). In the presence of control constraints, it is also very promising to solve the optimality system (3.54) directly. To this end, we substitute $u = \mathbb{P}_{[u_a,u_b]}\{-\lambda^{-1}\beta\, p\}$ in the state equation, using the representation $\mathbb{P}_{[u_a,u_b]}(z) = \max\{u_a, \min\{u_b, z\}\}$, to arrive at the following nonlinear system for y and p:

$$\begin{aligned} y_t - \Delta y &= 0 \\ \partial_\nu y + \alpha\, y &= \beta\, \max\{u_a, \min\{u_b, \{-\lambda^{-1}\beta\, p\}\}\} \\ y(0) &= y_0, \\ -p_t - \Delta p &= 0 \\ \partial_\nu p + \alpha\, p &= 0 \\ p(T) &= y(T) - y_\Omega. \end{aligned}$$

The only points of non-differentiability of the maximum and minimum functions appearing in this system are the points u_a and u_b; these functions are, however, globally *Newton differentiable*; see Ito and Kunisch [**IK08**]. These facts explain the successes that have been reported in the literature for the direct numerical solution of the above nonsmooth system. Another possibility is to approximate the maximum and minimum to very high accuracy by smooth functions (which in some codes is done automatically); the (smooth) nonlinear system of two parabolic problems thus generated is then solved numerically. Neitzel et al. [**NPS09**] report good experiences with this technique using existing software packages.

Optimization after full discretization. This method, often referred to as *discretize then optimize*, consists (as in the elliptic case) of a full discretization of both the parabolic problem and the cost functional, leading to a (large) optimization problem. It is easily performed, in particular, for spatially one-dimensional parabolic problems; in this case, existing solvers for finite-dimensional quadratic optimization problems may be employed.

3.8. Derivation of Fourier expansions

For the sake of completeness, we derive in this section the Fourier series (3.14). To this end, we consider the weak solution $y \in W_2^{1,0}(Q) \cap W(0,T)$

to the parabolic problem

$$\begin{aligned} y_t(x,t) - y_{xx}(x,t) &= f(x,t) \quad \text{in } Q = (0,1)\times(0,T) \\ y_x(0,t) &= 0 \quad \text{in } (0,T) \\ y_x(1,t) + \alpha\, y(1,t) &= u(t) \quad \text{in } (0,T) \\ y(x,0) &= y_0(x) \quad \text{in } (0,1), \end{aligned} \tag{3.66}$$

for fixed $\alpha > 0$, $T > 0$, and given functions $f \in L^2(Q)$, $u \in L^2(0,T)$, and $y_0 \in L^2(0,1)$. The (normalized) eigenfunctions $\{v_n(x)\}_{n=1}^{\infty}$ of the differential operator $A = -\partial^2/\partial x^2$ with the homogeneous boundary conditions

$$\frac{\partial v_n}{\partial x}(0) = 0, \qquad \frac{\partial v_n}{\partial x}(1) + \alpha v_n(1) = 0 \tag{3.67}$$

form a complete orthonormal system in $L^2(0,1)$; see, e.g., Tychonov and Samarski [**TS64**]. We aim to express y in the form of a series expansion,

$$y(x,t) = \sum_{n=1}^{\infty} v_n(x) z_n(t), \tag{3.68}$$

where the functions v_n and z_n are yet to be determined. First, we construct the eigenfunctions v_n. Let $\{\lambda_n\}_{n=1}^{\infty}$ denote the eigenvalues of A, i.e., $Av_n = \lambda_n v_n$. Then v_n satisfies the homogeneous boundary conditions (3.67) and

$$v_n''(x) + \lambda_n v_n(x) = 0 \quad \forall\, x \in [0,1]. \tag{3.69}$$

The ansatz $v_n(x) = c_1 \cos(\sqrt{\lambda}\, x) + c_2 \sin(\sqrt{\lambda}\, x)$ yields, upon taking (3.67) into account, that $c_2 = 0$; consequently, $\alpha \cos(\sqrt{\lambda}) = \sqrt{\lambda} \sin(\sqrt{\lambda})$. Putting $\mu := \sqrt{\lambda}$, we find that μ has to satisfy the equation

$$\mu \tan \mu = \alpha. \tag{3.70}$$

This equation has a countably infinite number of positive solutions μ_n, which we arrange as an increasing sequence $\{\mu_n\}_{n=1}^{\infty}$. The associated functions $\cos(\mu_n x)$ satisfy the orthogonality relations

$$\int_0^1 \cos(\mu_n x) \cos(\mu_\ell x)\, dx = \begin{cases} N_n, & n = \ell \\ 0, & n \neq \ell, \end{cases}$$

with $N_n := \frac{1}{2} + \frac{\sin(2\mu_n)}{4\mu_n}$. Hence, the functions

$$v_n(x) = \frac{1}{\sqrt{N_n}} \cos(\mu_n x)$$

form an orthonormal system that, owing to the theory of Sturm–Liouville eigenvalue problems, is complete (see, e.g., [**TS64**]). The initial condition in

(3.66) and the ansatz (3.68) imply that, after formally interchanging limit and summation,

$$y_0(x) = \lim_{t \downarrow 0} y(x,t) = \sum_{n=1}^{\infty} v_n(x) z_n(0). \tag{3.71}$$

On the other hand, by virtue of the theory of Fourier series in Hilbert spaces,

$$y_0(x) = \sum_{n=1}^{\infty} y_{0,n} v_n(x),$$

with the Fourier coefficients

$$y_{0,n} = \int_0^1 v_n(x) y_0(x)\, dx. \tag{3.72}$$

Comparison of the coefficients shows that

$$z_n(0) = \int_0^1 v_n(x) y_0(x)\, dx.$$

Next, we use the weak formulation (3.25) (on page 140) of the parabolic problem (3.66) in the spatially one-dimensional case and use $v(x,t) = \varphi(t) v_n(x)$, where $\varphi \in H^1(0,T)$ and $\varphi(T) = 0$, as the test function. It follows that

$$\begin{aligned} &- \int_0^T \int_0^1 y(x,t) \varphi'(t) v_n(x)\, dx\, dt + \int_0^T \int_0^1 y_x(x,t) \varphi(t) v_n'(x)\, dx\, dt \\ &\qquad + \int_0^T y(1,t) \varphi(t) v_n(1)\, dt - \int_0^T y(0,t) \varphi(t) v_n(0)\, dt \\ &\quad = \int_0^T \int_0^1 f(x,t) \varphi(t) v_n(x)\, dx\, dt + \int_0^T u(t) \varphi(t) v_n(1)\, dt \\ &\qquad + \int_0^1 y_0(x) \varphi(0) v_n(x)\, dx. \end{aligned} \tag{3.73}$$

Substitution of the series expansion (3.68) in the first integral of (3.73) yields, upon invoking the orthogonality relations,

$$- \int_0^T \int_0^1 y(x,t) \varphi'(t) v_n(x)\, dx\, dt = - \int_0^T z_n(t) \varphi'(t)\, dt. \tag{3.74}$$

The second integral can be transformed as follows, using (3.67), (3.69), the definition of μ_n, and the pairwise orthogonality of the v_n:

(3.75)
$$\begin{aligned}\int_0^T \int_0^1 y_x(x,t)\varphi(t)v_n'(x)\,dx\,dt &= \int_0^T \big(y(1,t)\varphi(t)v_n'(1) - y(0,t)\varphi(t)v_n'(0)\big)dt \\ &\quad - \int_0^T \int_0^1 y(x,t)\varphi(t)v_n''(x)\,dx\,dt \\ &= \int_0^T -\alpha\, y(1,t)\varphi(t)v_n'(1)\,dt + \int_0^T \int_0^1 y(x,t)\varphi(t)\mu_n^2 v_n(x)\,dx\,dt \\ &= \int_0^T -\alpha\, y(1,t)\varphi(t)v_n'(1)\,dt + \int_0^T \mu_n^2 z_n(t)\varphi(t)\,dt.\end{aligned}$$

Putting (3.74) and (3.75) into (3.73) yields that for all $\varphi \in H^1(0,T)$ such that $\varphi(T) = 0$,

(3.76)
$$\begin{aligned}&-\int_0^T z_n(t)\varphi'(t)\,dt + \int_0^T \mu_n^2 z_n(t)\varphi(t)\,dt \\ &= \int_0^T \left(\int_0^1 f(x,t)v_n(x)\,dx + u(t)v_n(1)\right)\varphi(t)\,dt + \int_0^1 y_0(x)v_n(x)\,dx\,\varphi(0).\end{aligned}$$

This is none other than the weak formulation of the ordinary initial value problem

$$\begin{aligned} z_n'(t) + \mu_n^2 z_n(t) &= F_n(t) \quad \text{in } (0,T) \\ z_n(0) &= z_{0,n}, \end{aligned} \tag{3.77}$$

where

$$\begin{aligned} F_n(t) &= \int_0^1 f(x,t)v_n(x)\,dx + u(t)v_n(1) \\ z_n(0) &= \int_0^1 y_0(x)v_n(x)\,dx. \end{aligned} \tag{3.78}$$

The solution to (3.77) is given by the variation of constants formula

$$z_n(t) = e^{-\mu_n^2 t} z_n(0) + \int_0^t e^{-\mu_n^2 (t-s)} F_n(s)\,ds,$$

whence, owing to (3.78),

$$\begin{aligned} z_n(t) = &\int_0^1 y_0(x)v_n(x)e^{-\mu_n^2 t}dx + \int_0^t \int_0^1 e^{-\mu_n^2(t-s)} f(x,t)v_n(x)\,dx \\ &+ \int_0^t e^{-\mu_n^2(t-s)} u(t)v_n(1)\,ds. \end{aligned} \tag{3.79}$$

Finally, we replace the integration variable x in (3.79) by ξ and insert this expression for z_n into (3.68). Interchanging the summation and integration then yields

$$
\begin{aligned}
y(x,t) &= \int_0^1 \sum_{n=0}^{\infty} v_n(x)v_n(\xi)e^{-\mu_n^2 t}y_0(\xi)\,d\xi \\
&\quad + \int_0^t \int_0^1 \sum_{n=0}^{\infty} v_n(x)v_n(\xi)e^{-\mu_n^2(t-s)}f(\xi,s)\,d\xi\,ds \\
&\quad + \int_0^t \sum_{n=0}^{\infty} v_n(x)v_n(1)e^{-\mu_n^2(t-s)}u(s)\,ds.
\end{aligned} \tag{3.80}
$$

By the definition of the Green's function G, this is the asserted formula (3.14) for $\alpha > 0$. The $\alpha = 0$ case can be treated in a similar way. The reader will be asked to do this in Exercise 3.11.

Remarks. In the above derivation, we have repeatedly interchanged limit or integration with summation. This needs careful justification, since the infinite series under the integrals in (3.80) are not uniformly convergent. However, for fixed t, the function $s \mapsto \sum_{n=0}^{\infty} e^{-\mu_n^2(t-s)}$ is integrable over $[0,t]$. Indeed, since $\mu_n \sim (n-1)\pi^2$ as $n \to \infty$, we have

$$
\int_0^t \sum_{n=0}^{\infty} e^{-\mu_n^2(t-s)}ds = \sum_{n=0}^{\infty} \frac{1}{\mu_n^2}(1 - e^{-\mu_n^2 t}) < \sum_{n=0}^{\infty} \frac{1}{\mu_n^2} < \infty.
$$

Consequently, if f and u are continuous or at least bounded, then the partial sums of the series containing f and u are easily seen to be majorized by an integrable function. Moreover, they converge pointwise on $[0,t)$. Hence, we can infer from Lebesgue's dominated convergence theorem that integration and summation may be interchanged in this case. The treatment of the integral containing y_0 and relation (3.71) follow easily from the L^2 theory of Fourier series; indeed, the sequence of Fourier coefficients $\{z_n(0)\}$ of y_0 is square summable, and the continuous functions $z_n(t)$ can be estimated as above. We leave the details to the reader. Finally, if f and u are unbounded, then they are approximated by sequences of continuous functions in $L^2(Q)$ and $L^2(0,T)$, respectively.

If y_0, f, and u are sufficiently smooth, then (3.80) even defines a classical solution y to the parabolic initial-boundary value problem; see, e.g., Tychonov and Samarski [**TS64**].

3.9. Linear continuous functionals as right-hand sides *

Parabolic equations can be written more generally than in the previous sections as equations in $L^2(0,T;V^*)$. To this end, we once more consider the initial-boundary value problem

$$\boxed{\begin{aligned} y_t + \mathcal{A}\,y + c_0\,y &= f \quad \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y + \alpha\,y &= g \quad \text{on } \Sigma \\ y(0) &= y_0 \quad \text{in } \Omega, \end{aligned}} \tag{3.81}$$

with given functions $f \in L^2\big(0,T;L^2(\Omega)\big)$, $g \in L^2\big(0,T;L^2(\Gamma)\big)$ and coefficient functions $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma)$. The elliptic differential operator $\mathcal{A}$ is defined as in (2.19) on page 37, with bounded and measurable coefficients a_{ij} that obey the symmetry condition and the uniform ellipticity condition (2.20). The weak formulation for $y \in W(0,T)$ reads as follows: $y(0) = y_0$ and, for all $v \in L^2(0,T;V)$,

$$\int_0^T \big(y_t(t)\,,\,v(t)\big)_{V^*,V}\,dt + \int_0^T a\big[y(t),v(t)\big]\,dt \;=\; \int_0^T \big(F(t)\,,\,v(t)\big)_{V^*,V}\,dt.$$

Here, $a = a[y,v]$ is the bilinear form associated with $\mathcal{A}$, introduced on page 38. Moreover, $V = H^1(\Omega)$, and the vector-valued function $F : [0,T] \to V^*$ is defined by

$$F(t)\,v = \int_\Omega f(x,t)\,v(x)\;dx + \int_\Gamma g(x,t)\,v(x)\,ds(x).$$

It was proved in Section 2.13 that the bilinear form a generates a continuous linear operator $A : V \to V^*$ upon taking $a[y,v] = (A\,y\,,\,v)_{V^*,V}$. Therefore, the weak form of (3.81) can be rewritten as follows: $y(0) = y_0$ and, for all $v \in L^2(0,T;V)$,

$$\begin{aligned} &\int_0^T \big(y_t(t)\,,\,v(t)\big)_{V^*,V}\,dt + \int_0^T \big(A\,y(t)\,,\,v(t)\big)_{V^*,V}\,dt \\ &\quad = \int_0^T \big(F(t)\,,\,v(t)\big)_{V^*,V}\,dt. \end{aligned}$$

Since f and g are square integrable vector-valued functions, we have

$$\|F(\cdot)\|_{L^2(0,T;V^*)} \le \tilde{c}\,\big(\|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)}\big),$$

so that all integrals in the above relation exist. Now, the preceding equation is equivalent to

$$\int_0^T \big(y_t(t) + A\,y(t) - F(t)\,,\,v(t)\big)_{V^*,V}\,dt = 0 \qquad \forall v \in L^2(0,T;V),$$

where y_t is to be understood as a (regular) vector-valued distribution belonging to $L^2(0,T;V^*)$. But this implies that

$$y_t(t) + A\,y(t) - F(t) = 0 \quad \text{in } L^2(0,T;V^*),$$

so the initial-boundary value problem (3.81) finally becomes

$$\boxed{\begin{aligned} y'(t) + A\,y(t) &= F(t) \in V^* \quad \text{for a.e. } t \in (0,T) \\ y(0) &= y_0. \end{aligned}}$$

Here, any arbitrary functional $F \in L^2(0,T;V^*)$ is permitted on the right-hand side. Now observe that under our assumptions we have, for almost every $t \in (0,T)$,

$$\begin{aligned} \big|a[y,v]\big| &\le \alpha_0 \,\|y\|_V \|v\|_V \\ a[v,v] &\ge \beta\,\|v\|^2_{H^1(\Omega)} - \beta_0\,\|v\|^2_{L^2(\Omega)} \end{aligned} \tag{3.82}$$

for all y, $v \in V$, with constants $\beta > 0$ and $\beta_0 \in \mathbb{R}$.

Theorem 3.22. *Suppose that* (3.82) *is satisfied. Then the evolution problem*

$$\begin{aligned} y'(t) + A\,y(t) &= F(t) \in V^* \qquad \textit{for a.e. } t \in (0,T) \\ y(0) &= y_0 \end{aligned}$$

has for any $F \in L^2(0,T;V^*)$ *and any* $y_0 \in H = L^2(\Omega)$ *a unique solution* $y \in W(0,T)$. *Moreover, there exists a constant* $c_P > 0$ *such that*

$$\|y\|_{W(0,T)} \le c_P \left(\|F\|_{L^2(0,T;V^*)} + \|y_0\|_H\right).$$

A more general version of this theorem (which also applies to the case of time-dependent coefficients) and its proof can be found, e.g., in Gajewski et al. [**GGZ74**] and Wloka [**Wlo87**].

An application. By virtue of Theorem 3.22, the adjoint state of a parabolic problem can be interpreted as a Lagrange multiplier associated with the parabolic differential equation. To this end, the differential equation $y' + A\,y - F = 0$ is regarded as a constraint in the range space $L^2(0,T;V^*)$. Since the mapping $y \mapsto y' + Ay$ is surjective, the Karush–Kuhn–Tucker theorem, Theorem 6.3 on page 330, yields the existence of a Lagrange multiplier $z^* = p \in L^2(0,T;V^*)^* = L^2(0,T;V)$, where the latter equality follows from $(V^*)^* = V$. Expositions of this technique in the theory of optimal control can be found in, e.g., Lions [**Lio71**], Hinze et al. [**HPUU09**], and Neittaanmäki and Tiba [**NT94**].

3.10. Exercises

3.1 The function $y(x,t) = \dfrac{e^t}{\sqrt{x}}$ belongs to the space $C([0,T],L^1(0,1))$. Compute its norm. To which of the spaces $L^p(0,T,L^q(0,1))$, $1 \le p,q \le \infty$, does this function belong?

3.2 Prove the existence of an optimal control for the problem (3.1)–(3.3) on page 119 with inhomogeneous initial state y_0.

3.3 Extend the notion of a weak solution to the initial-boundary value problem (3.58) on page 164 to the problem with mixed boundary conditions:

$$\begin{aligned} y_t + \mathcal{A}\, y + c_0\, y &= \beta_Q\, v && \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y + \alpha\, y &= \beta_\Sigma\, u && \text{on } \Sigma_1 \\ y &= 0 && \text{on } \Sigma_0 \\ y(0) &= y_0 && \text{in } \Omega. \end{aligned}$$

Here, $\Sigma_i = \Gamma_i \times (0,T)$, $i = 1, 2$, where the boundary pieces Γ_i are defined as in Section 2.3.3.

3.4 Determine the first-order necessary optimality conditions for the problem (3.57)–(3.59) on page 164 by means of the formal Lagrange method.

3.5 Investigate the problem (3.1)–(3.3) on page 119 with the extended cost functional

$$\begin{aligned} \tilde{J}(y,u,v) \;:=\; & J(y,u,v) + \iint_Q a_Q(x,t)\, y(x,t)\, dx\, dt \\ & + \iint_\Sigma a_\Sigma(x,t)\, y(x,t)\, ds(x) dt \\ & + \iint_Q v_Q(x,t)\, v(x,t)\, dx\, dt + \iint_\Sigma u_\Sigma(x,t)\, u(x,t)\, ds(x) dt, \end{aligned}$$

where J is the cost functional defined in (3.1) and the functions a_Q, $v_Q \in L^2(Q)$ and a_Σ, $v_\Sigma \in L^2(\Sigma)$ are prescribed. Do optimal controls exist for this problem? Derive the first-order necessary optimality conditions.

3.6 Consider the initial value control problem

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

subject to

$$\begin{aligned} y_t - \Delta y &= 0 && \text{in } Q \\ \partial_\nu y + y &= 0 && \text{on } \Sigma \\ y(0) &= w && \text{in } \Omega \end{aligned}$$

and $w \in L^2(\Omega)$, $|w(x)| \le 1$ for almost every $x \in \Omega$. Suppose that Assumption 3.14 on page 153 holds. Show the existence of an optimal control and derive the necessary optimality conditions.

3.7 Write a MATLAB program for the numerical solution of the initial-boundary value problem

$$
(3.83)\qquad
\begin{aligned}
y_t(x,t) - y_{xx}(x,t) &= 0 && \text{in } (0,\ell)\times(0,T)\\
y_x(0,t) &= 0 && \text{in } (0,T)\\
y_x(\ell,t) + y(\ell,t) &= u(t) && \text{in } (0,T)\\
y(x,0) &= y_0(x) && \text{in } (0,\ell).
\end{aligned}
$$

Use the implicit Euler method with time step τ for the discretization of $(0,T)$ and the difference quotient (2.91) on page 97 with step size h for the discretization of $(0,\ell)$. Take the control u to be a step function corresponding to the partition of $(0,T)$, and use the values of the given function y_0 at the spatial grid points.

3.8 Use the program from the preceding exercise to establish the finite-dimensional reduced problem for the optimal control problem

$$\min J(y,u) := \frac{1}{2}\|y(T) - y_\Omega\|^2_{L^2(0,1)} + \frac{\lambda}{2}\|u\|^2_{L^2(0,T)},$$

subject to (3.83) and the box constraints $|u(t)| \le 1$. Employ the method described in Section 3.7.2. Use the MATLAB code quadprog to solve the problem for the values chosen by Schittkowski [**Sch79**]: $\ell = 1, T = 1.58$, $y_\Omega(x) = 0.5\,(1-x^2)$, $y_0(x) = 0$, $\lambda = 10^{-3}$. Take the time step to be $\tau = 1/100$ and fit the spatial step size h accordingly.

3.9 Solve the reduced problem from the preceding exercise also for the choices $\lambda = 10^{-k}$, $k = -1, 0 \ldots 5$, and $\lambda = 0$. Solve the same problem for the function $y_\Omega(x) = 0.5\,(1-x)$, and halve the time step τ several times. What do you observe? Interpret your findings in light of Theorem 3.7 on page 133.

3.10 Develop, by slightly modifying the code written in Exercise 3.7, a program for the solution of the adjoint problem corresponding to the above optimal control problem. Compute the adjoint states associated with the optimal final states $y(T)$ obtained in Exercises 3.8 and 3.9. Verify the necessary optimality conditions numerically using the projection formula.

3.11 Use the method of Section 3.8 to construct the Green's function (3.14) in the case where $\alpha = 0$.

Chapter 4

Optimal control of semilinear elliptic equations

4.1. Preliminary remarks

In the previous chapters, we confined ourselves to linear partial differential equations and thus excluded many important applications. From now on, we will also consider nonlinear equations. A typical example is given by the optimal control problem

$$\min J(y,u) := \frac{1}{2}\int_\Omega \left|y(x) - y_\Omega(x)\right|^2 dx + \frac{\lambda}{2}\int_\Omega |u(x)|^2\, dx, \tag{4.1}$$

subject to

$$\begin{aligned} -\Delta y + y + y^3 &= u && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma \end{aligned} \tag{4.2}$$

and

$$u_a \le u(x) \le u_b \quad \text{for a.e. } x \in \Omega. \tag{4.3}$$

The elliptic equation occurring in problem (4.2) is *semilinear*. Once more, we will have to discuss relevant questions such as well-posedness of (4.2), existence of optimal controls, first-order necessary optimality conditions, and numerical methods. We should expect, however, that the corresponding analysis will become more difficult.

One might argue that, for instance, the necessary optimality conditions can easily be derived using the Lagrange technique. Indeed, they can be formulated in terms of an adjoint system which, in the sense of Chapter 2, comprises the adjoint problem corresponding to problem (4.2) linearized at $\bar{y}$:

$$\begin{aligned} -\Delta p + p + 3\bar{y}^2\, p &= \bar{y} - y_\Omega && \text{in } \Omega \\ \partial_\nu p &= 0 && \text{on } \Gamma. \end{aligned} \tag{4.4}$$

Moreover, we have, just as in the linear-quadratic case, the projection formula

$$\bar{u}(x) = \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\lambda}p(x)\right\}.$$

However, it becomes immediately evident that already the use of the space $H^1(\Omega)$ may create serious problems: indeed, we need to show the differentiability of nonlinear mappings like $y(\cdot) \mapsto y(\cdot)^3$ for our analysis, and it is by no means obvious in which function spaces this should be done.

There is another unpleasant fact: the above optimal control problem is not a convex one, even though the cost functional is convex. This arises from the fact that the elliptic state equation is nonlinear. Consequently, the first-order necessary optimality conditions are no longer sufficient, and there is a need to separately consider sufficient *second-order* optimality conditions. Unfortunately, unexpected difficulties arise in their analysis that have to be overcome.

4.2. A semilinear elliptic model problem

4.2.1. Motivation of the upcoming approach. To motivate the next steps, we consider the semilinear elliptic boundary value problem (4.2) in a bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$.

Since the function $y \mapsto y^3$ is monotone increasing, it will follow from Theorem 4.4 that a unique solution $y \in H^1(\Omega)$ exists for every given $u \in L^2(\Omega)$. By Theorem 7.1 on page 355, the embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ is continuous for $N = 3$, so that $y^3 \in L^2(\Omega)$. Moreover, as will be shown on page 230, the *Nemytskii operator* $\Phi : y(\cdot) \mapsto y^3(\cdot)$ is a Fréchet differentiable mapping from $L^6(\Omega)$ into $L^2(\Omega)$. This property will be needed to derive necessary optimality conditions. Finally, it follows from Lemma 2.35 on page 107 that the operator $A := -\Delta + I$ maps $V = H^1(\Omega)$ continuously into V^*.

Summarizing, we may rewrite the above boundary value problem as an equation in V^*,

$$A\, y + B\, \Phi(y) = B\, u,$$

where B denotes the embedding operator from $L^2(\Omega)$ into V^*.

In this way, the problem (4.2) can be treated in the state space $H^1(\Omega)$. This method also works if y^3 is replaced by the stronger nonlinearity y^5; indeed, we will see in Section 4.3.3 that $\Phi(y) = y^5$ is a differentiable mapping from $L^6(\Omega)$ into $L^{\frac{6}{5}}(\Omega)$ and thus, as the reader can check, also from V into V^*.

This method, while being adequate for many problems, is limited to cases in which Φ maps V into V^*. This requires growth conditions as in Section 4.3, which for instance are satisfied for $\Phi(y) = y^3$ and $\Phi(y) = y^5$ but not for $\Phi(y) = \exp(y)$. In fact, if $y \in H^1(\Omega)$, we cannot even expect $\exp(y) \in L^1(\Omega)$. We usually also have to impose restrictions on the dimension N of Ω.

However, we will see that under natural conditions the solution y is continuous on $\bar{\Omega}$, provided $u \in L^r(\Omega)$ for a sufficiently large r. If this is the case, then growth conditions are superfluous, and so are restrictions on the dimension such as $N \leq 3$.

There are two more reasons to look for continuous solutions y. First, the above method becomes more complicated in the case of parabolic problems in $W(0,T)$, since the degree of integrability of $y = y(x,t)$ on $\Omega \times (0,T)$ is lower than in the elliptic case; second, the continuity of the state y will be needed anyway for the treatment of state constraints in Chapter 6. It is therefore worthwhile to prove continuity or, at least, boundedness of the state y.

We proceed as follows. First, we treat the elliptic boundary value problem in $H^1(\Omega)$ under strong boundedness conditions. Then we show that the solution is actually bounded or continuous and that the boundedness conditions can be weakened. To this end, we investigate a more general class of problems. Monotone increasing nonlinearities of the types $\Phi(y) = y^3$ or $\Phi(y) = \exp(y)$ are too special for many applications: if, for instance, $\Omega = \Omega_1 \cup \Omega_2$ stands for a body composed of two materials having different physical constants κ_1 and κ_2, then a nonlinearity of the form

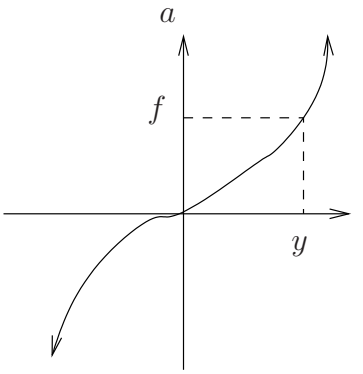
$$d(x,y) = \begin{cases} \kappa_1 y^3 & x \in \Omega_1 \\ \kappa_2 y^3 & x \in \Omega_2 \end{cases}$$

can be adequate. Evidently, d is no longer continuous in x, but is still bounded and measurable. Motivated by the above considerations, we treat as a model problem the elliptic boundary value problem

$$\boxed{\begin{array}{rcll} \mathcal{A}\, y + c_0(x)\, y + d(x,y) & = & f & \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y + \alpha(x)\, y + b(x,y) & = & g & \text{on } \Gamma. \end{array}} \tag{4.5}$$

Here, Ω, Γ, $c_0 \geq 0$, and $\alpha \geq 0$ are defined as in Section 2.1, and the nonlinear functions d and b are given. The elliptic differential operator $\mathcal{A}$ is assumed to take the form (2.19) on page 37, and the functions f and g will play the role of controls. This class of elliptic problems, while not being overly complicated, still exhibits the essential difficulties associated with nonlinear equations. Under suitable assumptions, we will be able to avoid the occurrence of the functions c_0 and/or α.

4.2.2. Solutions in $H^1(\Omega)$. We begin our analysis by investigating the existence and uniqueness of solutions to the semilinear elliptic boundary value problem (4.5) in the space $H^1(\Omega)$. To this end, we employ the theory of *monotone operators*.



f-point y of a monotone function a.

The basic idea is simple: if a continuous function $a : \mathbb{R} \to \mathbb{R}$ is strictly monotone increasing with $\lim_{x\to\pm\infty} a(x) = \pm\infty$, then the equation $a(y) = f$ has for any $f \in \mathbb{R}$ a uniquely determined solution $y \in \mathbb{R}$. This simple principle generalizes to equations $A\,y = f$ in Banach spaces.

In the following, V denotes a real separable Hilbert space, e.g., $V = H^1(\Omega)$ or $V = H_0^1(\Omega)$. Recall that a Banach space is said to be *separable* if it contains a countable dense subset.

Definition. *An operator $A : V \to V^*$ is said to be* monotone *if*

$$\big(Ay_1 - Ay_2\,,\, y_1 - y_2\big)_{V^*,V} \geq 0 \quad \forall\, y_1,\, y_2 \in V.$$

It is said to be strictly monotone *if equality can occur only if $y_1 = y_2$. We say that A is* coercive *if*

$$\frac{\big(Ay\,,\, y\big)_{V^*,V}}{\|y\|_V} \to \infty \quad \text{as } \|y\|_V \to \infty.$$

A is said to be hemicontinuous *if the real-valued function $\varphi : [0,1] \to \mathbb{R}$, $t \mapsto \big(A(y+tv)\,,\, w\big)_{V^*,V}$, is continuous on $[0,1]$ for all fixed $y, v, w \in V$. Finally, if there exists some $\beta_0 > 0$ such that*

$$\big(Ay_1 - Ay_2\,,\, y_1 - y_2\big)_{V^*,V} \geq \beta_0\, \|y_1 - y_2\|_V^2 \quad \forall\, y_1,\, y_2 \in V,$$

then A is said to be strongly monotone.

Theorem 4.1 (Main theorem on monotone operators)**.** *Let V be a separable Hilbert space, and let $A : V \to V^*$ be monotone, coercive, and hemicontinuous. Then the equation $Ay = f$ has for every $f \in V^*$ a solution $y \in V$. The set of all solutions is bounded, closed, and convex. If A is strictly monotone, then y is uniquely determined. If A is moreover strongly monotone, then the inverse $A^{-1} : V^* \to V$ is a Lipschitz continuous mapping.*

The above theorem is due to Browder and Minty. Its proof can be found in, e.g., Zeidler [**Zei90b**]. We apply it to problem (4.5) in the space $V = H^1(\Omega)$. To do this, we first have to define the notion of a weak solution to the nonlinear elliptic boundary value problem (4.5). The idea is simple: we bring the nonlinear terms $d(x, y)$ and $b(x, y)$ in (4.5) to the right-hand sides of the equations, thus obtaining a boundary value problem with the right-hand sides $\tilde{f} = f - d(\cdot, y)$ and $\tilde{g} = g - b(\cdot, y)$, respectively, and *linear* differential operators on the left-hand sides. For this purpose, we use the variational formulation for linear boundary value problems.

At this point, a problem arises if $b(x, y)$ or $d(x, y)$ is unbounded (e.g., for nonlinearities like y^k or e^y): elements of $y \in H^1(\Omega)$ need not be bounded. Without further assumptions, it is therefore unclear to which function spaces $d(x, y)$ and $b(x, y)$ should belong. Initially, we postulate that d and b be bounded on their respective domains; then $d(x, y)$ and $b(x, y)$ will be bounded, even if y is not.

Assumption 4.2.
(i) *$\Omega \subset \mathbb{R}^N$, $N \geq 2$, is a bounded Lipschitz domain with boundary Γ, and $\mathcal{A}$ is an elliptic differential operator of the form* (2.19) *(see page 37) with bounded and measurable coefficient functions a_{ij} that satisfy the symmetry condition and the condition* (2.20) *of uniform ellipticity.*

(ii) *The functions $c_0 : \Omega \to \mathbb{R}$ and $\alpha : \Gamma \to \mathbb{R}$ are bounded, measurable and almost everywhere nonnegative. Assume that at least one of these functions does not vanish almost everywhere, that is, $\|c_0\|_{L^\infty(\Omega)} + \|\alpha\|_{L^\infty(\Gamma)} > 0$.*

(iii) *The functions $d = d(x, y) : \Omega \times \mathbb{R} \to \mathbb{R}$ and $b = b(x, y) : \Gamma \times \mathbb{R} \to \mathbb{R}$ are bounded and measurable with respect to $x \in \Omega$ and $x \in \Gamma$, respectively, for every fixed $y \in \mathbb{R}$. Moreover, they are continuous and monotone increasing in y for almost every $x \in \Omega$ and $x \in \Gamma$, respectively.*

It follows from this assumption, in particular, that $d(x, 0)$ and $b(x, 0)$ are bounded and measurable in Ω and Γ, respectively. In view of the problem of unboundedness, we initially make a further assumption.

Assumption 4.3. *For almost every $x \in \Omega$ (respectively, $x \in \Gamma$) we have $d(x, 0) = 0$ (respectively, $b(x, 0) = 0$). Moreover, b and d are globally*

bounded, that is, there is a constant $M > 0$ such that for any $y \in \mathbb{R}$ we have

$$|b(x,y)| \le M \quad \text{for a.e. } x \in \Gamma, \qquad |d(x,y)| \le M \quad \text{for a.e. } x \in \Omega. \tag{4.6}$$

The differential equation is associated with the bilinear form

$$a[y,v] := \int_\Omega \sum_{i,j=1}^N a_{ij}(x) D_i y \, D_j v \, dx + \int_\Omega c_0 \, y \, v \, dx + \int_\Gamma \alpha \, y \, v \, ds. \tag{4.7}$$

Definition. *Suppose that Assumptions 4.2 and 4.3 hold. A function $y \in H^1(\Omega)$ is called a* weak solution *to* (4.5) *if we have, for every $v \in H^1(\Omega)$,*

$$a[y,v] + \int_\Omega d(x,y) \, v \, dx + \int_\Gamma b(x,y) \, v \, ds = \int_\Omega f \, v \, dx + \int_\Gamma g \, v \, ds. \tag{4.8}$$

Invoking the main theorem on monotone operators, we can prove the following well-posedness result.

Theorem 4.4. *Suppose that Assumptions 4.2 and 4.3 hold. Then the elliptic boundary value problem* (4.5) *has for any pair $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ of right-hand sides a unique weak solution $y \in H^1(\Omega)$. Moreover, there is some constant $c_M > 0$, which is independent of d, b, f, and g, such that*

$$\|y\|_{H^1(\Omega)} \le c_M \big(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \big). \tag{4.9}$$

Proof. We apply the main theorem on monotone operators in $V = H^1(\Omega)$.

(i) *Definition of a monotone operator $A : V \to V^*$*

It follows from Section 2.13 that the bilinear form (4.7) generates a continuous linear operator $A_1 : V \to V^*$ through the relation

$$(A_1 y \, , \, v)_{V^*,V} = a[y,v].$$

This is the linear part of the nonlinear operator A. The first nonlinear part of A is formally defined by the identity $(A_2 y)(x) := d(x, y(x))$. We have to make this formal "definition" precise: any $y \in V$ belongs to $L^2(\Omega)$. Owing to the strong assumptions imposed on d, the function $x \mapsto d(x, y(x))$ is measurable (by continuity of d in y) and bounded (by boundedness of d). We thus have $d(\cdot, y) \in L^\infty(\Omega)$ for all $y \in V$. Therefore, the linear functional F_d given by

$$F_d(v) = \int_\Omega d\big(x, y(x)\big) \, v(x) \, dx$$

is continuous on V and thus belongs to V^*. In this sense, using the canonical isomorphism from V into V^*, we may identify $d(\cdot, y(\cdot))$ with $F_d \in V^*$. In conclusion, we may define $A_2 : V \to V^*$ by putting $A_2\, y = F_d$. The third part A_3, which corresponds to the nonlinearity b, can be defined similarly. Indeed, the linear functional

$$F_b(v) = \int_\Gamma b\big(x, y(x)\big)\, v(x)\, ds(x)$$

also belongs to V^* and can be identified with $b(\cdot, y(\cdot))$. In this sense, $A_3\, y = F_b$. The sum of the three operators yields the operator A, i.e., $A = A_1 + A_2 + A_3$.

(ii) *Monotonicity of A*

We show that each of the operators A_i, $1 \le i \le 3$, is monotone, so that this property then also holds for A. First, A_1 is monotone, since $a[y, y] \ge 0$ for all $y \in V$. Next, we consider A_2. Owing to the monotonicity of d in y, we have $\big(d(x, y_1) - d(x, y_2)\big)(y_1 - y_2) \ge 0$ for all $y_1,\, y_2 \in \mathbb{R}$ and all x. Therefore, for all $y \in H^1(\Omega)$,

$$\begin{aligned}&(A_2(y_1) - A_2(y_2)\,,\, y_1 - y_2)_{V^*,V}\\ &= \int_\Omega (d(x, y_1(x)) - d(x, y_2(x)))(y_1(x) - y_2(x))\, dx \ge 0\,.\end{aligned}$$

Note that the boundedness condition for d guarantees that the function $x \mapsto d(x, y_1(x)) - d(x, y_2(x))$ is square integrable, so that the above integral exists. In conclusion, A_2 is monotone. The monotonicity of A_3 follows from analogous reasoning.

(iii) *Coercivity of A*

A_1 is coercive, since the assumptions made on c_0 and α imply, as in the proof of Theorem 2.6 on page 35, that

$$(A_1\, v\,,\, v)_{V^*,V} \;=\; a[v, v] \;\ge\; \beta_0\, \|v\|_V^2 \quad \forall\, v \in V.$$

The operators A_2 and A_3 contribute nonnegative terms that do not destroy the coercivity. Here, the assumption $d(x, 0) = b(x, 0) = 0$ is exploited. Indeed, we have, by the monotonicity of d,

$$\begin{aligned}(A_2\, v\,,\, v)_{V^*,V} &= \int_\Omega d(x, v(x))\, v(x)\, dx\\ &= \int_\Omega (d(x, v(x)) - d(x, 0))(v(x) - 0)\, dx \ge 0\,.\end{aligned}$$

A similar estimate holds for A_3, and this proves the claim that $A = A_1 + A_2 + A_3$ is coercive.

(iv) *Hemicontinuity of A*

The operator A_1 is linear and thus hemicontinuous. For A_2 we argue as follows: we put

$$\varphi(t) := \big(A_2(y + t\,v)\,,\,w\big)_{V^*,V} = \int_\Omega d\big(x, y(x) + t\,v(x)\big)\,w(x)\,dx.$$

Now let $\tau \in \mathbb{R}$ be fixed, and let $\{t_n\}_{n=1}^\infty \subset \mathbb{R}$ be some sequence such that $t_n \to \tau$ as $n \to \infty$. We need to show that $\varphi(t_n) \to \varphi(\tau)$ as $n \to \infty$.

Since by Assumption 4.2 d is continuous with respect to y for almost every $x \in \Omega$, it follows that

$$f_n(x) := d\big(x, y(x) + t_n\,v(x)\big)\,w(x) \;\to\; d\big(x, y(x) + \tau\,v(x)\big)\,w(x) =: f(x)$$

pointwise almost everywhere in Ω. Moreover, the sequence $\{f_n\}_{n=1}^\infty$ is also pointwise almost everywhere majorized by an integrable function; indeed, it follows from (4.6) that

$$\big|d\big(x, y(x) + t_n\,v(x)\big)\,w(x)\big| \le M\,\big|w(x)\big| \qquad \text{for a.e. } x \in \Omega,$$

where $w \in L^2(\Omega)$. Thus, we can infer from Lebesgue's dominated convergence theorem that

$$\int_\Omega d\big(x, y(x) + t_n\,v(x)\big)\,w(x)\,dx \to \int_\Omega d\big(x, y(x) + \tau\,v(x)\big)\,w(x)\,dx,$$

and hence $\varphi(t_n) \to \varphi(t)$ as $n \to \infty$. The operator A_3 can be treated analogously.

(v) *Well-posedness of the solution*

Existence and uniqueness of a weak solution $y \in H^1(\Omega)$ now follow directly from the main theorem on monotone operators. Since A is obviously strongly monotone, the asserted estimate also holds. However, it is not clear why the estimate does not depend on d or b. Therefore, we give a direct proof of it. To this end, in the variational equation (4.8) we take y itself as the test function to obtain

$$a[y,y] + (A_2\,y\,,\,y)_{V^*,V} + (A_3\,y\,,\,y)_{V^*,V} = \int_\Omega f(x)\,y(x)\,dx + \int_\Gamma g(x)\,y(x)\,ds(x).$$

Since the terms containing A_2 and A_3 are nonnegative, it follows that

$$a[y,y] \le \int_\Omega f(x)\,y(x)\,dx + \int_\Gamma g(x)\,y(x)\,ds(x).$$

This is the reason why the asserted estimate does not depend on d or b. Now, we estimate the bilinear form $a[y,y]$ from below by the H^1 norm and

the right-hand side from above by the Cauchy–Schwarz inequality. We then obtain

$$\begin{aligned}\beta_0 \|y\|^2_{H^1(\Omega)} &\leq \|f\|_{L^2(\Omega)} \|y\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|y\|_{L^2(\Gamma)} \\ &\leq c \big(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\big) \|y\|_{H^1(\Omega)}\,,\end{aligned}$$

whence the asserted estimate follows. This concludes the proof of the theorem. □

Remarks.
(i) The above theorem does not apply directly to problem (4.2), since $d(x, y) = y^3$ does not meet Assumption 4.3. But Assumption 4.3 was only instrumental in guaranteeing that $d(\cdot, y(\cdot)) \in L^2(\Omega)$, which, as mentioned above, is true for $d(x, y(x)) = y(x)^3$ if $y \in H^1(\Omega)$ and $N \leq 3$.

(ii) The above proof only made use of the fact that f and g generate continuous linear functionals on V and therefore can be identified with elements of V^*. But for this to be the case the square integrability is not necessary; see page 40. In particular, the result remains valid for data $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$ whenever $r > \frac{N}{2}$ and $s > N - 1$, respectively; owing to Theorem 4.7 below, y is then even continuous on $\bar{\Omega}$. For $N \in \{2, 3\}$ this includes also cases where r, $s < 2$.

(iii) Further techniques for the treatment of nonlinear elliptic equations can be found in the monographs by Barbu [**Bar93**], Lions [**Lio69**], Ladyzhenskaya and Ural'ceva [**LU73**], Neittaanmäki et al. [**NST06**], and Zeidler [**Zei90b, Zei95**].

4.2.3. Continuity of solutions. In this section, we follow ideas developed by E. Casas in [**Cas93**]. We begin our analysis by using a technique due to Stampacchia to show that the weak solution $y \in H^1(\Omega)$ is actually even essentially bounded, provided that the functions f and g have "better" properties than just square integrability. The main result of this section will be Theorem 4.8 on the continuity of y.

Theorem 4.5. *Suppose that Assumptions 4.2 and 4.3 hold, and let $r > N/2$ and $s > N - 1$. Then for any pair $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$, there exists a unique weak solution $y \in H^1(\Omega)$ to the boundary value problem* (4.5). *We have $y \in L^\infty(\Omega)$, and there is some constant $c_\infty > 0$, which does not depend on d, b, f, or g, such that*

$$\|y\|_{L^\infty(\Omega)} \leq c_\infty \big(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\Gamma)}\big). \tag{4.10}$$

The proof of this theorem will be given in Section 7.2.2, beginning on page 358. As the reader will be asked to verify in Exercise 4.1, we have the inequality

$$\|y\|_{L^\infty(\Gamma)} \leq \|y\|_{L^\infty(\Omega)} \quad \forall\, y \in H^1(\Omega) \cap L^\infty(\Omega).$$

Therefore, from (4.10) the same estimate for $\|y\|_{L^\infty(\Gamma)}$ follows.

It is noteworthy that the estimate (4.10) does not depend on the nonlinearities d and b. The reason behind this is their monotonicity. It is therefore quite natural to presume that the postulated boundedness of the nonlinearities is dispensable. This is indeed the case, and without this boundedness assumption it is still possible to show that the solution y is continuous on $\bar{\Omega}$. To this end, we need another preparatory result.

Lemma 4.6 (**[Cas93]**). *Suppose that $\Omega \subset \mathbb{R}^N$ is a bounded Lipschitz domain, and let $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$ with $r > N/2$ and $s > N-1$ be given. Then the weak solution y to the Neumann problem*

$$\begin{aligned} \mathcal{A}\, y + y &= f \\ \partial_{\nu_{\mathcal{A}}} y &= g \end{aligned}$$

is continuous on $\bar{\Omega}$. Moreover, there is some constant $c(r,s) > 0$, which does not depend on f or g, such that

$$\|y\|_{C(\bar{\Omega})} \le c(r,s)\,\big(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\Gamma)}\big).$$

Proof: For simplicity, we prove the assertion under the additional assumption that Ω has a $C^{1,1}$ boundary and that the coefficients a_{ij} belong to $C^{0,1}(\bar{\Omega})$. References for the general case will be given below, after this proof.

The existence of a unique weak solution $y \in H^1(\Omega)$ follows from Theorem 2.6 on page 35 (cf. the remarks on page 40 concerning data with $r < 2$ or $s < 2$). Owing to Theorem 4.5, applied with $c_0 = 1$ and $d = b = \alpha = 0$, y is essentially bounded and satisfies the estimate (4.10). It remains to show that y is continuous on $\bar{\Omega}$, since then $\|y\|_{L^\infty(\Omega)} = \|y\|_{C(\bar{\Omega})}$, which implies the validity of the asserted estimate.

To this end, observe that under the above regularity assumption for Γ, the spaces $C^\infty(\Omega)$ and $C^\infty(\Gamma)$ are dense in $L^r(\Omega)$ and $L^s(\Gamma)$, respectively. We may therefore choose functions $f_n \in C^\infty(\Omega)$ and $g_n \in C^\infty(\Gamma)$, $n \in \mathbb{N}$, such that

$$\|f_n - f\|_{L^r(\Omega)} \to 0 \quad \text{and} \quad \|g_n - g\|_{L^s(\Gamma)} \to 0 \qquad \text{as } n \to \infty.$$

Now denote by y_n the unique weak solution to the above Neumann problem with right-hand sides f_n and g_n, $n \in \mathbb{N}$. Owing to the assumed regularity of the boundary Γ and the coefficient functions a_{ij}, we may apply the regularity results from Grisvard **[Gri85]** collected in Section 2.14.3 to conclude that $y_n \in W^{2,r}(\Omega)$. Since $W^{2,r}(\Omega)$ is for $r > N/2$ continuously embedded in $C(\bar{\Omega})$ (see Theorem 7.1 on page 355), it follows that $y_n \in C(\bar{\Omega})$. Moreover, the difference $y - y_n$ solves the boundary value problem

$$\begin{aligned} \mathcal{A}(y - y_n) + y - y_n &= f - f_n \\ \partial_{\nu_{\mathcal{A}}}(y - y_n) &= g - g_n. \end{aligned}$$

Recalling the definition of the sequences $\{f_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ and invoking Theorem 4.5 once more (in particular, the estimate (4.10)), we conclude that $\|y_n - y\|_{L^\infty(\Omega)} \to 0$ as $n \to \infty$. In particular, $\{y_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L^\infty(\Omega)$ and, since all the terms y_n of the sequence are continuous on $\bar{\Omega}$, also in $C(\bar{\Omega})$. Hence, $\{y_n\}_{n=1}^{\infty}$ has a limit in $C(\bar{\Omega})$, which obviously must be y. This concludes the proof of the assertion. □

The proof for the case of L^∞ coefficients a_{ij} and Lipschitz domains is due to Casas [**Cas93**]. It was extended to more general situations by Alibert and Raymond [**AR97**]. Recent results by Griepentrog [**Gri02**] on the regularity of solutions to elliptic boundary value problems include the above lemma as a special case. This will be explained in Section 7.2.1, following page 356. The Hölder continuity of the solution for mixed (Dirichlet–Neumann) boundary conditions was recently proved by Haller-Dintelmann et al. [**HDMRS09**].

We now drop the postulate of Assumption 4.3 that the nonlinearities b and d be bounded. In this situation, we call $y \in H^1(\Omega) \cap L^\infty(\Omega)$ a *weak solution* to the boundary value problem (4.5) if it satisfies the variational equality (4.8).

Theorem 4.7. *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain, and let $r > N/2$ and $s > N-1$. Suppose also that Assumption 4.2 holds, and that $b(x,0) = 0$ and $d(x,0) = 0$ for almost every $x \in \Gamma$ and $x \in \Omega$, respectively. Then the semilinear boundary value problem (4.5) has for any pair $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$ a unique weak solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$. Moreover, y is continuous on $\bar{\Omega}$, and there is some constant $c_\infty > 0$, which does not depend on d, b, f, or g, such that*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \le c_\infty \big(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\Gamma)}\big). \tag{4.11}$$

Proof: We again follow the argument in Casas [**Cas93**]. First, we show that the requirement that d and b be bounded is dispensable. To this end, consider for arbitrary $k > 0$ the *cut-off* function

$$d_k(x,y) = \begin{cases} d(x,k) & \text{if} \quad y > k \\ d(x,y) & \text{if} \quad |y| \le k \\ d(x,-k) & \text{if} \quad y < -k. \end{cases}$$

In the same way, we define a cut-off function b_k for b. The functions b_k and d_k are uniformly bounded and satisfy Assumption 4.3. We can thus infer from Theorem 4.5 that the elliptic boundary value problem

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

possesses a unique weak solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$. Moreover, y satisfies the estimate (4.10), which does not depend on d_k or b_k and therefore not on k. Now choose $k > c_\infty \left(\|f\|_{L^r(\Omega)} + \|g\|_{L^s(\Gamma)}\right)$. Then, by virtue of (4.10), $|y(x)| \le k$ for almost every $x \in \Omega$ and almost every $x \in \Gamma$. Therefore, $d_k(x, y(x)) = d(x, y(x))$ and $b_k(x, y(x)) = b(x, y(x))$ almost everywhere. Consequently, y is a solution to problem (4.5).

Next, we show the continuity of y. To this end, we rewrite the nonlinear boundary value problem solved by y in the form

$$\begin{aligned} \mathcal{A}\, y + y &= f + y - c_0\, y - d_k(x, y) && \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y &= g - \alpha\, y - b_k(x, y) && \text{on } \Gamma. \end{aligned}$$

Since y is by (4.10) essentially bounded on both Ω and Γ, the right-hand sides of this problem belong to $L^r(\Omega)$ and $L^s(\Gamma)$, respectively. Hence, we can infer from Lemma 4.6 that y is continuous on $\bar{\Omega}$.

It remains to show that y is unique. Indeed, owing to the monotonicity of d and b with respect to y, any solution $\bar{y} \in H^1(\Omega) \cap L^\infty(\Omega)$ to problem (4.5) is at the same time the uniquely determined solution to the above cut-off system for any $k > \|\bar{y}\|_{L^\infty(\Omega)}$. But this obviously entails that there can be at most one solution in the space $H^1(\Omega) \cap L^\infty(\Omega)$, which concludes the proof of the assertion. □

As the final step, we now demonstrate that the postulate $d(x, 0) = b(x, 0) = 0$ is also dispensable; this was used in the proof of Theorem 4.4 to guarantee monotonicity of the operator A.

Theorem 4.8. *The assertion of Theorem 4.7 remains valid without the assumption $b(x, 0) = d(x, 0) = 0$, provided that the estimate* (4.11) *is replaced by*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \le c_\infty \left(\|f - d(\cdot, 0)\|_{L^r(\Omega)} + \|g - b(\cdot, 0)\|_{L^s(\Gamma)}\right). \tag{4.12}$$

Proof: We rewrite the boundary value problem in the form

$$\begin{aligned} \mathcal{A}\, y + c_0(x)\, y + d(x, y) - d(x, 0) &= f(x) - d(x, 0) && \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y + \alpha(x)\, y + b(x, y) - b(x, 0) &= g(x) - b(x, 0) && \text{on } \Gamma. \end{aligned}$$

The functions $y \mapsto d(x, y) - d(x, 0)$ and $y \mapsto b(x, y) - b(x, 0)$ vanish at zero. Moreover, Assumption 4.2 obviously implies that $d(\cdot, 0) \in L^r(\Omega)$ and $b(\cdot, 0) \in L^s(\Gamma)$. Hence, Theorem 4.7 applies with the right-hand sides $f - d(\cdot, 0)$ and $g - b(\cdot, 0)$, yielding the validity of (4.12). The assertion is thus proved. □

4.2.4. Weakening of the assumptions. So far, we have studied the semilinear elliptic model problem in the form (4.5) on page 183. The required coercivity of the elliptic operator has been guaranteed by the properties of the coefficient functions c_0 and α. In particular, the choice $d(x, y) = 0$ and $b(x, y) = 0$ has been possible. This raises the question of under what conditions the functions c_0 and α can be omitted, since the coercivity of the nonlinear operator follows from the properties of d and b alone.

As characteristic examples, let us consider the semilinear Neumann problems

$$\begin{aligned} -\Delta y + e^y &= f && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma \end{aligned} \tag{4.13}$$

and

$$\begin{aligned} -\Delta y + y^3 &= f && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma. \end{aligned} \tag{4.14}$$

We investigate whether unique solutions exist in $H^1(\Omega) \cap L^\infty(\Omega)$ that depend continuously on the right-hand side f.

It is easy to see that this is not the case for problem (4.13). In fact, the function $y(x) \equiv c$ solves (4.13) with the right-hand side $f(x) \equiv e^c$, and we have $f \to 0$ as $c \to -\infty$. However, for $f = 0$ there can be no solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$. Indeed, if there were such a solution, then we could use the test function $v \equiv 1$ in the variational formulation to find that $\int_\Omega e^{y(x)}\, dx = 0$, which is a contradiction. Observe that this problem does not arise if a homogeneous Dirichlet condition is given: the operator $\mathcal{A} = -\Delta$ is coercive in $H^1_0(\Omega)$.

In contrast to this, the boundary value problem (4.14) is well posed in $H^1(\Omega) \cap L^\infty(\Omega)$. This will be a consequence of the next theorem, which was communicated to me by E. Casas. It concerns the boundary value problem

$$\boxed{\begin{aligned} \mathcal{A}\, y + d(x, y) &= 0 && \text{in } \Omega \\ \partial_{\nu_\mathcal{A}} y + b(x, y) &= 0 && \text{on } \Gamma. \end{aligned}} \tag{4.15}$$

Here, $\mathcal{A}$ is defined as in (4.5), and the given right-hand sides f and g are incorporated into $d(x, y)$ and $b(x, y)$, respectively.

Assumption 4.9. *The domain Ω and the linear differential operator $\mathcal{A}$ satisfy the conditions stated in Assumption 4.2 on page 185. The functions $d = d(x, y) : \Omega \times \mathbb{R} \to \mathbb{R}$ and $b = b(x, y) : \Gamma \times \mathbb{R} \to \mathbb{R}$ are measurable with respect to x for every $y \in \mathbb{R}$, and are monotone increasing and continuous in y*

for almost all $x \in \Omega$ and $x \in \Gamma$, respectively. Moreover, for any $M > 0$ there are functions $\psi_M \in L^r(\Omega)$ with $r > N/2$ and $\phi_M \in L^s(\Gamma)$ with $s > N - 1$ such that

$$\begin{aligned} |d(x,y)| &\le \psi_M(x) \quad \text{for a.e. } x \in \Omega, \ \text{whenever } |y| \le M; \\ |b(x,y)| &\le \phi_M(x) \quad \text{for a.e. } x \in \Gamma, \ \text{whenever } |y| \le M. \end{aligned} \tag{4.16}$$

Finally, one of the following two conditions holds:

(i) *There exist a set $E_d \subset \Omega$ with positive measure and constants $M_d > 0$ and $\lambda_d > 0$ such that the following inequalities hold:*

$$\begin{aligned} d(x,y_1) < d(x,y_2) \quad \forall\, x \in E_d, \quad \forall\, y_1 < y_2; \\ (d(x,y) - d(x,0))\, y \ge \lambda_d\, |y|^2 \quad \forall\, x \in E_d, \quad \forall\, |y| > M_d. \end{aligned} \tag{4.17}$$

(ii) *There exist a set $E_b \subset \Gamma$ with positive measure and constants $M_b > 0$ and $\lambda_b > 0$ such that the following inequalities hold:*

$$\begin{aligned} b(x,y_1) < b(x,y_2) \quad \forall\, x \in E_b, \quad \forall\, y_1 < y_2; \\ (b(x,y) - b(x,0))\, y \ge \lambda_b\, |y|^2 \quad \forall\, x \in E_b, \quad \forall\, |y| > M_b. \end{aligned} \tag{4.18}$$

Theorem 4.10. *Suppose Assumption 4.9 holds. Then the boundary value problem* (4.15) *has a unique weak solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$. The weak solution is continuous on $\bar{\Omega}$.*

Proof: We follow an idea of E. Casas and consider for $n \in \mathbb{N}$ the boundary value problem

$$\begin{aligned} \mathcal{A}\, y + n^{-1} y + d(x,y) &= 0 \qquad \text{in } \Omega \\ \partial_{\nu_{\mathcal{A}}} y + b(x,y) &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

In analogy to Theorem 4.8, under these slightly modified assumptions there is also a unique weak solution $y_n \in H^1(\Omega) \cap L^\infty(\Omega)$ to this problem; in this connection, we put $c_0(x) = n^{-1}$. From Theorem 7.6 on page 363, we can infer that there is some constant $K > 0$ such that

$$\|y_n\|_{L^\infty(\Omega)} \le K \quad \forall\, n \in \mathbb{N}.$$

Since $\|y_n\|_{L^\infty(\Gamma)} \le \|y_n\|_{L^\infty(\Omega)}$ for all $n \in \mathbb{N}$, the sequence $\{\|y_n\|_{L^\infty(\Gamma)}\}$ is also bounded. But then the L^2 norms of the functions y_n and of their traces on the boundary are bounded, too.

Next, we put $v = y_n$ in the weak formulation of the above problem, for $n \in \mathbb{N}$. Using the monotonicity of d and b and invoking inequality (2.20) on

page 37, we obtain from (4.16) that

$$\gamma_0 \int_\Omega |\nabla y|^2 dx \le \int_\Omega \psi_K \, |y_n| \, dx + \int_\Gamma \phi_K \, |y_n| \, ds \le c \quad \text{for all } n \in \mathbb{N}.$$

Since $\|y_n\|_{L^\infty(\Omega)} \le K$ for all $n \in \mathbb{N}$, it follows that $\{\|y_n\|_{H^1(\Omega)}\}$ is also bounded. We may therefore select a subsequence $\{y_{n_k}\}$ such that with some $y \in H^1(\Omega) \cap L^\infty(\Omega)$ we have $y_{n_k} \to y$ weakly in $H^1(\Omega)$ and, by compact embedding, strongly in $L^2(\Omega)$.

By virtue of the boundedness of $\{y_{n_k}\}$ in $L^\infty(\Omega)$ and in $L^\infty(\Gamma)$, we may apply Lebesgue's dominated convergence theorem to conclude that $d(\cdot, y_{n_k}) \to d(\cdot, y)$ strongly in $L^2(\Omega)$ and $b(\cdot, y_{n_k}) \to b(\cdot, y)$ strongly in $L^2(\Gamma)$. Hence, taking the limit as $k \to \infty$ in the above sequence of boundary value problems, we find that y is a solution to (4.15).

The uniqueness of the solution follows from a standard argument: suppose that we are given two weak solutions $y_1, y_2 \in H^1(\Omega) \cap L^\infty(\Omega)$. Testing the difference between the corresponding equations by $v = y_1 - y_2$, we obtain that

$$\begin{aligned} (4.19) \qquad & \gamma_0 \int_\Omega |\nabla(y_1 - y_2)|^2 \, dx + \int_\Omega (d(x, y_1) - d(x, y_2)) \, (y_1 - y_2) \, dx \\ & + \int_\Gamma (b(x, y_1) - b(x, y_2)) \, (y_1 - y_2) \, ds \, \le \, 0. \end{aligned}$$

Owing to the monotonicity of d and b, all three summands are nonnegative and therefore must vanish. But then $|\nabla(y_1 - y_2)(x)| = 0$, and thus $y_1(x) - y_2(x) = c$ for almost every $x \in \Omega$, with some $c \in \mathbb{R}$ (cf. Zeidler **[Zei90a]**, Problem 21.31a). Since $y_1 - y_2 \in H^1(\Omega)$ is equivalent in the Lebesgue sense to the continuous function $y \equiv c$, it follows from the trace theorem that also $y_1(x) - y_2(x) = c$ for almost every $x \in \Gamma$.

If $c \neq 0$, then without loss of generality we may assume that $y_1(x) < y_2(x)$ for almost every $x \in \Omega$ and almost every $x \in \Gamma$. From (4.17) and (4.18) it then follows that at least one of the last two summands in (4.19) must be positive, and we have a contradiction. Consequently, $c = 0$, and thus $y_1 = y_2$.

Finally, the continuity of the solution is a consequence of Lemma 4.6.

□

If the functions $d(x, y)$ and $b(x, y)$ are not only increasing in y but also differentiable with respect to y for almost all x, then the following conditions are sufficient for (4.17) and (4.18) to hold:

There are measurable sets $E_d \subset \Omega$ and $E_b \subset \Gamma$ of positive measure, as well as constants $\lambda_d > 0$ and $\lambda_b > 0$, such that

$$d_y(x,y) \geq \lambda_d \ \ \forall x \in E_d, \forall y \in \mathbb{R}; \qquad b_y(x,y) \geq \lambda_b \ \ \forall x \in E_b, \forall y \in \mathbb{R}. \tag{4.20}$$

If one of these conditions holds, then Theorem 4.8 applies to the boundary value problem (4.15): for instance, we put $c_0(x) := \chi(E_d)\lambda_d$ and write

$$d(x,y) = c_0(x)y + (d(x,y) - c_0(x)y) = c_0(x)y + \tilde{d}(x,y).$$

Then $\tilde{d}$ is increasing with respect to y, and $\|c_0\|_{L^\infty(\Omega)} \neq 0$. In a similar way, b can be transformed using $\alpha(x) := \chi(E_b)\lambda_b$. With this, the boundary value problem (4.15) attains the form (4.5), and the assumptions of Theorem 4.8 are met with either of the two conditions in (4.20).

4.3. Nemytskii operators

4.3.1. Continuity of Nemytskii operators. Any nonlinearity $d(x,y)$ generates for a given function y a new function by putting $z(x) = d(x, y(x))$. Such operators are called *superposition operators* or *Nemytskii operators*. Quite unexpectedly, it is a nontrivial task to study the differentiability properties of such operators.

Examples. The following mappings $y(\cdot) \mapsto z(\cdot)$ define Nemytskii operators:

$$z(x) = \big(y(x)\big)^3, \quad z(x) = a(x)\,\big(y(x)\big)^3, \quad z(x) = \sin\big(y(x)\big),$$
$$z(x) = \big(y(x) - a(x)\big)^2.$$

The first mapping occurs in the superconductivity example, the third is the archetypical example for illustrating possible difficulties, and the fourth appears as an integrand in our quadratic cost functionals. The corresponding generating nonlinearities are evidently given by

$$d(y) = y^3, \ d(x,y) = a(x)\,y^3, \ d(y) = \sin(y), \ d(x,y) = \big(y - a(x)\big)^2. \tag{4.21}$$

◇

All nonlinearities of the above type that appear in this book may depend on two variables, namely the *domain variable* x and the *function variable* for which a control or state function is inserted. The function variable will usually be denoted by y, u, v, or w. In the case of controls distributed in the domain, the nonlinearity $d = d(x,y)$, $x \in \Omega$, (d for **d**istributed) is used; likewise, for boundary controls the function $b = b(x,y)$, $x \in \Gamma$, (b for **b**oundary) appears.

In parabolic problems we have, in addition to the spatial variable x, the time variable t, so that (x,t) represents the domain variable. To simplify

the next definition, we denote by E the set in which the domain variable varies. In the case of elliptic problems we have $E = \Omega$ or $E = \Gamma$, while in the parabolic case $E = \Omega \times (0,T)$ or $E = \Gamma \times (0,T)$. We generally assume that the set E is bounded and Lebesgue measurable.

The numerical analysis of nonlinear optimal control problems requires the determination of first- and second-order derivatives of Nemytskii operators. In this section, we begin our analysis with continuity and first-order derivatives. Second-order derivatives will be discussed later in connection with sufficient optimality conditions.

Definition. *Let $E \subset \mathbb{R}^m$, $m \in \mathbb{N}$, be a bounded and measurable set, and let $\varphi = \varphi(x,y) : E \times \mathbb{R} \to \mathbb{R}$ be a function. The mapping Φ given by*

$$\Phi\big(y\big) = \varphi\big(\cdot, y(\cdot)\big),$$

which assigns to a function $y : E \to \mathbb{R}$ the function $z : E \to \mathbb{R}$, $z(x) = \varphi\big(x, y(x)\big)$, is called a Nemytskii operator *or* superposition operator.

The analysis of Nemytskii operators in L^p spaces with $1 \le p < \infty$ necessitates more or less restrictive growth conditions on $\varphi(x,y)$ with respect to y. Since all control and state functions to be studied in the following will be uniformly bounded, we can work in L^∞ and thus with simpler conditions that are met, for example, by all elementary functions defined on the whole real line.

Definition.
(i) *A function $\varphi = \varphi(x,y) : E \times \mathbb{R} \to \mathbb{R}$ is said to satisfy the* Carathéodory condition *if it is measurable with respect to x for any fixed $y \in \mathbb{R}$ and continuous with respect to y for almost every fixed $x \in E$.*

(ii) *φ is said to satisfy the* boundedness condition *if there exists a constant $K > 0$ such that*

$$|\varphi(x,0)| \le K \quad \text{for a.e } x \in E. \tag{4.22}$$

(iii) *φ is said to be* locally Lipschitz continuous *with respect to y if for any constant $M > 0$ there is a constant $L(M) > 0$ such that for almost every $x \in E$ and all y, $z \in [-M, M]$ we have the estimate*

$$|\varphi(x,y) - \varphi(x,z)| \le L(M)\,|y - z|. \tag{4.23}$$

The following fact is readily seen: if φ satisfies the boundedness condition and is locally Lipschitz continuous, then for any $M > 0$ there is some $c_M > 0$ such that $|\varphi(x,y)| \le c_M$ for all $y \in [-M, M]$ and almost all $x \in E$.

Examples. All functions $\varphi \in C^1(\mathbb{R})$ satisfy the above conditions. The same holds for $\varphi(x,y) := a_1(x) + a_2(x)\, b(y)$ if $a_i \in L^\infty(E)$ and $b \in C^1(\mathbb{R})$. Also, the functions from (4.21) comply with the conditions provided that $a \in L^\infty(E)$. $\diamond$

Lemma 4.11. *Suppose that the function $\varphi = \varphi(x,y) : E \times \mathbb{R} \to \mathbb{R}$ is measurable with respect to $x \in E$ for every $y \in \mathbb{R}$, and suppose that φ satisfies the boundedness condition and is locally Lipschitz continuous with respect to y. Then the associated Nemytskii operator Φ is continuous in $L^\infty(E)$. Moreover, for all $r \in [1,\infty]$, we have*

$$\|\Phi(y) - \Phi(z)\|_{L^r(E)} \;\le\; L(M)\, \|y - z\|_{L^r(E)}$$

for all $y, z \in L^\infty(E)$ such that $\|y\|_{L^\infty(E)} \le M$ and $\|z\|_{L^\infty(E)} \le M$.

Proof: Let $y \in L^\infty(E)$ be given. Then there is some $M > 0$ such that $|y(x)| \le M$ for almost every $x \in E$. By virtue of the conditions (4.22) and (4.23), we have, for almost every $x \in E$,

$$|\varphi\big(x, y(x)\big)| \le |\varphi\big(x,0\big)| + |\varphi\big(x, y(x)\big) - \varphi\big(x,0\big)| \le K + L(M)M\,.$$

Hence, $\Phi\big(y(\cdot)\big) = \varphi\big(\cdot, y(\cdot)\big) \in L^\infty(E)$, and Φ maps $L^\infty(E)$ into itself.

Now let $y, z \in L^\infty(E)$ be given such that $\|y\|_{L^\infty(E)} \le M$ and $\|z\|_{L^\infty(E)} \le M$. Then it follows from the local Lipschitz continuity that, for any $1 \le r < \infty$,

$$\begin{aligned} \int_E |\varphi\big(x, y(x)\big) - \varphi\big(x, z(x)\big)|^r \, dx \;&\le\; L(M)^r \int_E |y(x) - z(x)|^r \, dx \\ &=\; L(M)^r \, \|y - z\|^r_{L^r(E)}. \end{aligned}$$

The asserted estimate is thus proved for $1 \le r < \infty$. For $r = \infty$ the argument is even simpler. Hence, Φ is also locally Lipschitz continuous in $L^\infty(E)$. $\square$

If φ is moreover *uniformly bounded* and Lipschitz continuous on the whole real line $\mathbb{R}$, then Φ is Lipschitz continuous in any of the spaces $L^r(E)$, that is, there is some $L > 0$ such that

$$\|\Phi(y) - \Phi(z)\|_{L^r(E)} \le L\, \|y - z\|_{L^r(E)} \quad \forall\, y,\, z \in L^r(E).$$

This follows immediately from the above proof.

Example. $\Phi(y) = \sin\big(y(\cdot)\big)$.

The associated Nemytskii operator Φ is generated by $\varphi(x,y) = \sin(y)$. Obviously,

$$|\sin(y)| \le 1 \;\text{ and }\; |\sin(y) - \sin(z)| \le |y - z| \quad \forall\, y,\, z \in \mathbb{R}.$$

The function $\varphi(x,y) = \sin(y)$ is thus globally bounded and Lipschitz continuous with constant 1. Hence, Φ is globally Lipschitz continuous, and

$$\big\| \sin\big(y(\cdot)\big) - \sin\big(z(\cdot)\big) \big\|_{L^r(E)} \;\le\; \|y - z\|_{L^r(E)} \quad \forall\, y,\, z \in L^r(E). \qquad \diamond$$

4.3.2. Differentiability of Nemytskii operators.

Assumptions on the nonlinearities. To obtain differentiability properties of Nemytskii operators, higher regularity conditions have to be imposed on the function φ. To this end, let us agree upon the following terminology:

Definition. *Let $E \subset R^m$, $m \in \mathbb{N}$, be a bounded set, and let $\varphi = \varphi(x,y) : E \times \mathbb{R} \to \mathbb{R}$ denote a function of the domain variable x and the function variable y. Suppose that φ is k-times differentiable with respect to y for almost every $x \in E$. We say that φ satisfies the* boundedness condition of order k *if there exists some $K > 0$ such that*

$$\big|D_y^l\, \varphi(x,0)\big| \le K \quad \forall\, 0 \le l \le k, \quad \text{for a.e. } x \in E. \tag{4.24}$$

We say that φ satisfies the local Lipschitz condition of order k *if for every $M > 0$ there is some (Lipschitz) constant $L(M) > 0$ such that*

$$\big|D_y^k\, \varphi(x,y_1) - D_y^k\, \varphi(x,y_2)\big| \le L(M)\, |y_1 - y_2| \tag{4.25}$$

for all $y_i \in \mathbb{R}$ such that $|y_i| \le M$, $i = 1, 2$.

If φ depends only on the second variable, i.e. $\varphi = \varphi(y)$, then the above two conditions are equivalent to the local Lipschitz continuity of $\varphi^{(k)}$,

$$\big|\varphi^{(k)}(y_1) - \varphi^{(k)}(y_2)\big| \le L(M)\, |y_1 - y_2| \quad \forall\, y_i \in \mathbb{R} \text{ such that } |y_i| \le M, \;\; i = 1, 2.$$

Remark. It is easily seen that the validity of both the boundedness condition and the local Lipschitz condition of order k implies local boundedness and local Lipschitz continuity of all derivatives up to order k: indeed, if $1 \le l \le k$, for all y with $|y| \le M$ it follows that

$$\big|D_y^l\, \varphi(x,y)\big| \le \big|D_y^l\, \varphi(x,y) - D_y^l\, \varphi(x,0)\big| + \big|D_y^l\, \varphi(x,0)\big| \le L(M)\, |y| + K \le K(M).$$

Therefore, $D_y^l\, \varphi$ is locally bounded, and the derivative of order $l-1$ is locally Lipschitz continuous; for instance, local Lipschitz continuity of order 2 implies local Lipschitz continuity of order 1, since the mean value theorem yields for $|y_i| \le M$, $i = 1, 2$, that

$$\big|D_y\, \varphi(x,y_1) - D_y\, \varphi(x,y_2)\big| = \big|D_y^2\, \varphi(x,y_\vartheta)(y_1 - y_2)\big| \le 2\, K(M)\, M,$$

with an intermediate point y_ϑ between y_1 and y_2. Applying this argument inductively, we can work from $l = k$ down to $l = 0$, that is, to φ itself.

First-order derivatives in $L^\infty(E)$. The *differentiability* of Φ seems to be perfectly clear: if φ is continuously differentiable with respect to y, then the associated Nemytskii operator Φ is also expected to be differentiable. As we

will discover below, this expectation is justified only in principle: everything depends on an appropriate choice of the function spaces.

To simplify the following exposition, we will use the following familiar abbreviations for partial derivatives: we write $\varphi_y := D_y\,\varphi = \partial\varphi/\partial y$ and $\varphi_{yy} := D_y^2\varphi = \partial^2\varphi/\partial y^2$.

Now, for every fixed x let φ be continuously differentiable with respect to y. If the *Fréchet* derivative of Φ at y exists, then it can be determined as the *Gâteaux* derivative

$$\begin{aligned}\big(\Phi'(y)h\big)(x) &= \lim_{t\to 0}\frac{1}{t}\Big[\varphi\big(x,y(x)+t\,h(x)\big)-\varphi\big(x,y(x)\big)\Big]\\ &= \frac{d}{dt}\varphi\big(x,y(x)+t\,h(x)\big)\big|_{t=0} = \varphi_y\big(x,y(x)\big)h(x),\end{aligned}\tag{4.26}$$

where the limit evidently exists for each fixed x. But this fact does not provide any information on whether the limit exists in the sense of a suitable L^r space, nor does it indicate to which space the function $\varphi_y\big(\cdot,y(\cdot)\big)$ and its product with h should belong. In other words, mere pointwise existence of the above limit does not yet suffice to guarantee Fréchet differentiability.

Example: Sine operator. The sine function is infinitely differentiable and all of its derivatives are uniformly bounded. In view of these nice smoothness and boundedness properties, we employ the sine function as a test case in the seemingly simplest space, namely the Hilbert space $L^2(E)$. The corresponding Nemytskii operator $\Phi\big(y(\cdot)\big) = \sin\big(y(\cdot)\big)$ is globally Lipschitz continuous in $L^2(E)$, as the example in the preceding section shows. We conjecture that Φ is also Fréchet differentiable. By (4.26), the derivative must be given by

$$\big(\Phi'(y)\,h\big)(x) = \cos\big(y(x)\big)\,h(x).\tag{4.27}$$

This seems to fit, since the cosine function is bounded by 1 and thus the right-hand side of (4.27) defines a continuous linear operator in $L^2(E)$ that assigns to each $h\in L^2(E)$ the product $\cos\big(y(\cdot)\big)\,h(\cdot)\in L^2(E)$.

Quite unexpectedly, however, our conjecture is wrong. In spite of all the nice smoothness properties of the sine function, the sine operator cannot be Fréchet differentiable in any of the spaces $L^p(E)$, $1\le p<\infty$. This disappointing fact follows from a well-known result which asserts that Φ is Fréchet differentiable in the space $L^p(E)$ for $1\le p<\infty$ if and only if φ is an affine function with respect to y, that is to say, $\varphi(x,y)=\varphi_0(x)+\varphi_1(x)\,y$ with some $\varphi_0\in L^p(E)$ and some $\varphi_1\in L^\infty(E)$; see Krasnoselskii et al. [**KZPS76**]. This fact, which will be demonstrated below for the case of the sine operator, is a big obstacle that renders the optimal control theory of nonlinear problems more difficult.

The non-differentiability of the sine operator is easily verified, as the interested reader can check in Exercise 4.4(i). This is particularly simple at the zero element of the space $L^p(0,1)$ with $1 \le p < \infty$. To see this, let $h \in L^p(0,1)$ be given. Taylor's theorem with integral remainder, applied to the sine function, yields

$$\begin{aligned} \sin\big(0+h(x)\big) &= \sin(0) + \cos(0)\, h(x) \\ &\quad + \int_0^1 \big[\cos\big(0 + s\, h(x)\big) - \cos(0)\big]\, h(x)\, ds \\ &= 0 + h(x) + r(x). \end{aligned}$$

Here, $h(x)$ is regarded as a real number for fixed x. For the increment h, we choose the step function

$$h(x) = \begin{cases} 1 & \text{in } [0,\varepsilon] \\ 0 & \text{in } (\varepsilon,1] \end{cases}$$

with $0 < \varepsilon < 1$ and pass to the limit as $\varepsilon \downarrow 0$. The remainder $r(x)$ can be immediately read off without the integral remainder: we have $r(x) = \sin(h(x)) - h(x)$. Therefore, for $x \in [0,\varepsilon]$,

$$r(x) = \sin(1) - 1 = c \neq 0,$$

so that

$$r(x) = \begin{cases} c & \text{in } [0,\varepsilon] \\ 0 & \text{in } (\varepsilon,1]. \end{cases}$$

If the sine operator were Fréchet differentiable at the zero function, then we would have

$$\frac{\|r\|_{L^p(0,1)}}{\|h\|_{L^p(0,1)}} \to 0 \quad \text{as} \quad \|h\|_{L^p(0,1)} \to 0\,.$$

However, we obtain that

$$\frac{\|r\|_{L^p(0,1)}}{\|h\|_{L^p(0,1)}} = \frac{\Big(\int_0^\varepsilon |r(x)|^p dx\Big)^{\frac{1}{p}}}{\Big(\int_0^\varepsilon |h(x)|^p dx\Big)^{\frac{1}{p}}} = \frac{c\,\varepsilon^{\frac{1}{p}}}{\varepsilon^{\frac{1}{p}}} = c \neq 0.$$

Fortunately, the situation is not completely hopeless. Indeed, it will follow from the next lemma that the sine operator is Fréchet differentiable at least in the space $L^\infty(E)$. Moreover, we conclude from the above calculation that it ought to be differentiable from $L^{p_1}(0,1)$ into $L^{p_2}(0,1)$ for $1 \le p_2 < p_1$; indeed, then the corresponding quotient $\varepsilon^{\frac{1}{p_2}-\frac{1}{p_1}}$ tends to zero as ε approaches zero; therefore, the above contradiction no longer exists. The reader will be asked to show the differentiability between this pair of spaces in Exercise 4.4(ii). ◇

Lemma 4.12. *Suppose that the function φ is measurable with respect to $x \in E$ for every $y \in \mathbb{R}$ and differentiable with respect to y for almost every $x \in E$. Moreover, let both the boundedness condition* (4.24) *and the local Lipschitz condition* (4.25) *of order $k = 1$ be satisfied. Then the Nemytskii operator Φ associated with φ is Fréchet differentiable in $L^\infty(E)$, and we have*

$$\big(\Phi'(y)\,h\big)(x) = \varphi_y\big(x, y(x)\big)\,h(x) \quad \text{for a.e. } x \in E \text{ and all } h \in L^\infty(E).$$

Proof: Let y, $h \in L^\infty(E)$ be arbitrary, and choose $M > 0$ such that $|y(x)| \le M$ and $|h(x)| \le M$ for almost every $x \in E$. Then

$$\varphi\big(x, y(x) + h(x)\big) - \varphi\big(x, y(x)\big) = \varphi_y\big(x, y(x)\big)h(x) + r\big(y, h\big)(x)$$

with the remainder

$$r\big(y, h\big)(x) = \int_0^1 \Big[\varphi_y\big(x, y(x) + s\,h(x)\big) - \varphi_y\big(x, y(x)\big)\Big]\,ds\,h(x).$$

By the Lipschitz continuity of φ_y, we can estimate, for almost all $x \in E$,

$$\begin{aligned}\big|r\big(y, h\big)(x)\big| \;&\le\; L(2M)\int_0^1 s\,|h(x)|\,ds\;|h(x)| \;\le\; \frac{L(2M)}{2}\;|h(x)|^2 \\ &\le\; \frac{L(2M)}{2}\;\|h\|^2_{L^\infty(E)}.\end{aligned}$$

Therefore, $\|r(y,h)\|_{L^\infty(E)} \le c\,\|h\|^2_{L^\infty(E)}$ and thus

$$\frac{\|r(y,h)\|_{L^\infty(E)}}{\|h\|_{L^\infty(E)}} \to 0 \quad \text{as } \|h\|_{L^\infty(E)} \to 0.$$

The desired convergence of the remainder is thus shown. Moreover, since $\varphi_y\big(\cdot, y(\cdot)\big)$ is by the boundedness condition for φ_y bounded, the multiplication operator $h(\cdot) \mapsto \varphi_y\big(\cdot, y(\cdot)\big)h(\cdot)$ is a continuous linear mapping from $L^\infty(E)$ into itself. In this connection, note that the measurability of φ_y follows from the measurability of φ, since the derivative with respect to y comes from a limit of measurable functions. With this, all properties of the Fréchet derivative are proved. □

Conclusion. *Every function $\varphi \in C^2(\mathbb{R})$ that depends only on y generates a Fréchet differentiable Nemytskii operator in $L^\infty(E)$; indeed, φ' is locally Lipschitz continuous.*

Since the implicit function theorem will be applied later, we now introduce the notion of *continuous* differentiability where the operator $F'(u)$ depends continuously on u.

Definition. *Let $F : \mathcal{U} \to V$ be a Fréchet differentiable mapping in an open neighborhood $\mathcal{U}$ of a point $\bar{u} \in U$. F is said to be* continuously Fréchet differentiable *at $\bar{u}$ if the mapping $u \mapsto F'(u)$ from $\mathcal{U}$ into $\mathcal{L}(U, V)$ is continuous at $\bar{u}$, that is, if*

$$\|u - \bar{u}\|_U \to 0 \;\Rightarrow\; \|F'(u) - F'(\bar{u})\|_{\mathcal{L}(U,V)} \to 0.$$

F is said to be continuously Fréchet differentiable in $\mathcal{U}$ *if it is continuously Fréchet differentiable at every $\bar{u} \in \mathcal{U}$.*

Lemma 4.13. *Suppose that the assumptions of Lemma 4.12 hold. Then the Nemytskii operator Φ is continuously Fréchet differentiable in $L^\infty(E)$.*

Proof: Let $\bar{y} \in L^\infty(E)$ be arbitrary but fixed, and let $\|y_n - \bar{y}\|_{L^\infty(E)} \to 0$, where $y_n \in L^\infty(E)$ for all $n \in \mathbb{N}$. We have to show that

$$\|\Phi'(y_n) - \Phi'(\bar{y})\|_{\mathcal{L}(L^\infty(E))} \to 0 \quad \text{as } n \to \infty.$$

Evidently, there is some $M > 0$ such that

$$\|\bar{y}\|_{L^\infty(E)} + \|y_n\|_{L^\infty(E)} \le M \quad \forall\, n \in \mathbb{N}.$$

Using the local Lipschitz continuity of φ_y, we obtain

$$\begin{aligned}
&\|\Phi'(y_n) - \Phi'(\bar{y})\|_{\mathcal{L}(L^\infty(E))} \\
&\quad = \sup_{\|v\|_{L^\infty(E)}=1} \big\| \big[\varphi_y\big(\cdot, y_n(\cdot)\big) - \varphi_y\big(\cdot, \bar{y}(\cdot)\big)\big] v(\cdot) \big\|_{L^\infty(E)} \\
&\quad \le \big\|\varphi_y\big(\cdot, y_n(\cdot)\big) - \varphi_y\big(\cdot, \bar{y}(\cdot)\big)\big\|_{L^\infty(E)} \le L(M)\, \|y_n - \bar{y}\|_{L^\infty(E)} \to 0.
\end{aligned}$$

Hence, Φ is (Lipschitz) continuously differentiable. □

Example. On $C[0,1]$ and for $k \in \mathbb{N}$, $k \ge 2$, we consider the mapping

$$\Phi\big(y(\cdot)\big) = y(\cdot)^k.$$

The directional derivative at $\bar{y} \in C[0,1]$ in the direction $y \in C[0,1]$ is obviously

$$\Phi'(\bar{y})\, y = k\, \bar{y}^{k-1}\, y.$$

We can identify $\Phi'(\bar{y})$ with the function $k\, \bar{y}^{k-1}$. Now let $\{y_n\}_{n=1}^\infty \subset C[0,1]$ be any sequence such that $y_n \to y$ in $C[0,1]$. Again, there is some $M > 0$ such that

$$\|\bar{y}\|_{C[0,1]} + \|y_n\|_{C[0,1]} \le M \quad \forall\, n \in \mathbb{N}.$$

Using the mean value theorem, we have with suitable $\theta_n(x) \in (0,1)$ the estimate

$$\begin{aligned}
\big\|\Phi'(y_n) - \Phi'(\bar y)\big\|_{\mathcal{L}(C[0,1])} &= \max_{\|y\|_{C[0,1]}=1} \big\|\big(\Phi'(y_n) - \Phi'(\bar y)\big)\, y\big\|_{C[0,1]} \\
&= \max_{\|y\|_{C[0,1]}=1} \big\|k\,(y_n^{k-1} - \bar y^{k-1})\, y\big\|_{C[0,1]} \;\le\; k\,\big\|y_n^{k-1} - \bar y^{k-1}\big\|_{C[0,1]} \\
&\le\; k(k-1) \sup_{x\in[0,1]} \big|\big(y_n + \theta_n(y_n - \bar y)\big)(x)\big|^{k-2}\, \big|\big(y_n - \bar y\big)(x)\big| \\
&\le\; k(k-1)\,\big(\|y_n\|_{C[0,1]} + \|\bar y\|_{C[0,1]}\big)^{k-2}\, \|y_n - \bar y\|_{C[0,1]} \\
&\le\; k(k-1)\, M^{k-2}\, \|y_n - \bar y\|_{C[0,1]} \;\to\; 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Consequently, the mapping $y \mapsto \Phi'(y)$ is continuous, so that Φ is continuously differentiable. ◇

4.3.3. Derivatives in other L^p spaces *. For completeness, we now collect some properties of Nemytskii operators in L^p spaces with $1 \le p < \infty$. These will, for instance, be used for the analysis of the elliptic equation $-\Delta y + y + y^3 = u$ in $H^1(\Omega)$. The proofs can be found in the monographs [**AZ90**] and [**KZPS76**] and in the papers [**App88**] and [**GKT92**]. For the analysis of Nemytskii operators in Sobolev or Hölder spaces, we refer the reader to [**AZ90**] and [**Goe92**].

Continuity. Let a bounded and measurable set $E \subset \mathbb{R}^n$ be given, and assume that $\varphi = \varphi(x,y)$ satisfies the Carathéodory condition. Then the Nemytskii operator $\Phi(y) := \varphi(\cdot, y(\cdot))$ maps $L^p(E)$ into $L^q(E)$ for $1 \le q \le p < \infty$ if and only if there are functions $\alpha \in L^q(E)$ and $\beta \in L^\infty(E)$ such that the *growth condition*

$$|\varphi(x,y)| \le \alpha(x) + \beta(x)\,|y|^{\frac{p}{q}} \tag{4.28}$$

is satisfied. Moreover, the operator Φ is for $q < \infty$ automatically continuous if it maps $L^p(E)$ into $L^q(E)$; see [**AZ90**].

Differentiability. In addition, let the partial derivative $\varphi_y(x,y)$ exist for almost every $x \in E$, and assume that the Nemytskii operator generated by $\varphi_y(x,y)$ maps $L^p(E)$ into $L^r(E)$. If $1 \le q < p < \infty$ satisfies the condition

$$r = \frac{p\,q}{p-q}, \tag{4.29}$$

then Φ is Fréchet differentiable from $L^p(E)$ into $L^q(E)$, and we have

$$\big(\Phi'(y)\,h\big)(x) = \varphi_y\big(x, y(x)\big)\, h(x).$$

For the proof of the differentiability, we have to show that (4.28) holds with q replaced by r and φ replaced by φ_y. This is plausible, since the product $\varphi_y\, h$ will have to belong to $L^q(E)$ for $h \in L^p(E)$. We thus determine the conjugate exponent s of p/q from the equation $1/s + q/p = 1$ and estimate, using Hölder's inequality (2.25) on page 43, that

$$\begin{aligned}\int_E |\varphi_y(x, y(x))|^q |h(x)|^q \, dx \;&\le\; \Big(\int_E |\varphi_y(x, y(x))|^{qs} \, dx \Big)^{\frac{1}{s}} \Big(\int_E |h(x)|^{q\frac{p}{q}} \, dx \Big)^{\frac{q}{p}} \\ &= \Big(\int_E |\varphi_y(x, y(x))|^{r} \, dx \Big)^{\frac{1}{s}} \Big(\int_E |h(x)|^{p} \, dx \Big)^{\frac{q}{p}}.\end{aligned}$$

By assumption, both integrals are finite.

Example. Let $\Omega \subset R^N$ be a bounded domain, $k \ge 1$ an integer, and Φ the Nemytskii operator generated by $\varphi(y) = y^k$. In connection with equation (4.2) on page 181, we want to know for which values of k the operator Φ is differentiable from $L^6(\Omega)$ into $L^{6/5}(\Omega)$. By (4.29), the derivative φ_y has to map the space $L^6(\Omega)$ into $L^r(\Omega)$ with

$$r = \frac{6\,\frac{6}{5}}{6 - \frac{6}{5}} = \frac{3}{2}.$$

We have $|\varphi_y| = k\,|y|^{k-1}$. In view of the growth condition (4.28), we postulate that $k - 1 \le p/r$ with $p = 6$ and $r = 3/2$; hence

$$k - 1 \le \frac{6}{r} = 4,$$

and thus $k \le 5$. Therefore, for $k \le 5$ $\;y(\cdot)^k$ is differentiable from $L^6(\Omega)$ into $L^{6/5}(\Omega)$. ◇

4.4. Existence of optimal controls

4.4.1. General assumptions for this chapter. Owing to the required assumptions on the nonlinearities, the theory involving nonlinear equations and cost functionals may become confusing. Depending on the problem—for instance, the existence of optimal controls, necessary first-order conditions, or sufficient second-order optimality conditions—the requirements differ and would have to be specified anew in each section. To avoid this, we list a set of assumptions to hold throughout the remainder of this chapter, which is in fact too strong for most results. We will discuss at the relevant places which parts of the assumptions are dispensable; see also the remark at the end of this section.

In the following, the real-valued functions $d(x, y)$, $b(x, y)$, $\varphi(x, y)$, and $\psi(x, u)$, which depend on a *domain variable* $x \in E$ and a real *function*

variable y or u, will repeatedly occur. Here, the specifications $E = \Omega$ and $E = \Gamma$ are possible. Moreover, thresholds u_a, u_b, v_a, $v_b : E \to \mathbb{R}$ for the controls will be prescribed.

Assumption 4.14.
(**i**) $\Omega \subset \mathbb{R}^N$ *is a bounded Lipschitz domain.*

(**ii**) *The functions* $d = d(x,y)$, $\varphi = \varphi(x,y) : \Omega \times \mathbb{R} \to \mathbb{R}$, $b = b(x,y) : \Gamma \times \mathbb{R} \to \mathbb{R}$, *and* $\psi = \psi(x,u) : E \times \mathbb{R} \to \mathbb{R}$, *where* $E = \Omega$ *or* $E = \Gamma$, *are measurable with respect to* x *for every* $y \in \mathbb{R}$ *(respectively,* $u \in \mathbb{R}$*) and twice differentiable with respect to* y *(respectively,* u*) for almost every* $x \in \Omega$ *(respectively,* $x \in \Gamma$*). Moreover, they satisfy the boundedness and local Lipschitz conditions* (4.24)–(4.25) *of order* $k = 2$*; for* φ *this means, for example, that there are constants* $K > 0$ *and* $L(M) > 0$ *such that for almost every* $x \in \Omega$ *we have*

$$|\varphi(x,0)| + |\varphi_y(x,0)| + |\varphi_{yy}(x,0)| \le K,$$
$$|\varphi_{yy}(x,y_1) - \varphi_{yy}(x,y_2)| \le L(M)\,|y_1 - y_2| \quad \forall\, y_1,\, y_2 \in [-M,M].$$

(**iii**) *Additionally,* $d_y(x,y) \ge 0$ *for almost every* $x \in \Omega$ *and all* $y \in \mathbb{R}$, *and* $b_y(x,y) \ge 0$ *for almost every* $x \in \Gamma$ *and all* $y \in \mathbb{R}$. *Moreover, there are sets* $E_d \subset \Omega$ *and* $E_b \subset \Gamma$ *of positive measure and constants* $\lambda_d > 0$ *and* $\lambda_b > 0$ *such that*

$$d_y(x,y) \ge \lambda_d \ \ \forall x \in E_d, \forall y \in \mathbb{R}; \qquad b_y(x,y) \ge \lambda_b \ \ \forall x \in E_b, \forall y \in \mathbb{R}.$$

(**iv**) *The bounds* u_a, u_b, v_a, $v_b : E \to \mathbb{R}$ *belong to* $L^\infty(E)$ *for* $E = \Omega$ *or* $E = \Gamma$ *and satisfy the conditions* $u_a(x) \le u_b(x)$ *and* $v_a(x) \le v_b(x)$ *for almost every* $x \in E$.

As mentioned before, the above set of assumptions is too restrictive. In fact, for the existence of optimal controls the conditions in (**ii**) concerning the derivatives of φ and ψ are dispensable; these conditions, including Lipschitz continuity, are needed only for the functions themselves (order $k = 0$). On the other hand, we have to postulate that ψ is convex with respect to u. For first-order necessary optimality conditions, (**ii**) needs to be postulated up to order $k = 1$ only, while Assumption 4.14 is needed in its entirety for second-order conditions and for SQP methods.

Example. The following functions satisfy the above assumption:

$$\varphi(x,y) = a(x)\,y + \beta(x)\,(y - y_\Omega(x))^2 \quad \text{with } a,\, \beta,\, y_\Omega \in L^\infty(\Omega),$$

$$d(x,y) = c_0(x)\,y + y^k \quad \text{with odd } k \in \mathbb{N} \text{ and } c_0(x) \geq 0 \quad \text{in } \Omega$$
$$\text{such that } \|c_0\|_{L^\infty(\Omega)} > 0,$$
$$d(x,y) = c_0(x)\,y + \exp(a(x)\,y) \quad \text{with } c_0 \in L^\infty(\Omega),\ c_0(x) \geq 0,$$
$$\|c_0\|_{L^\infty(\Omega)} > 0 \quad \text{and } a \in L^\infty(\Omega),\ a(x) \geq 0.$$

◇

Under Assumption 4.14, the existence result of Theorem 4.8 on page 192 applies to the subsequent elliptic boundary value problems. We rewrite d in the form

$$d(x,y) = c_0(x)y + (d(x,y) - c_0(x)y) = c_0(x)y + \tilde{d}(x,y), \tag{4.30}$$

with $c_0 = \chi(E_d)\,\lambda_d$. Then $\tilde{d}$ satisfies Assumption 4.2 on page 185. Analogously, we can rewrite b, defining $\alpha := \chi(E_b)\,\lambda_b$.

4.4.2. Distributed control. We exemplify this case by investigating the optimal control problem

$$\min J(y,u) := \int_\Omega \varphi\big(x, y(x)\big)\,dx + \int_\Omega \psi\big(x, u(x)\big)\,dx, \tag{4.31}$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + d(x,y) & = & u & \text{in } \Omega \\ \partial_\nu y & = & 0 & \text{on } \Gamma \end{array}} \tag{4.32}$$

and

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Omega. \tag{4.33}$$

We recall that problems in which the control occurs as a source term on the right-hand side of the partial differential equation are termed *distributed control problems*. Here, the set of admissible controls is given by

$$U_{ad} = \big\{u \in L^\infty(\Omega) \,:\, u_a(x) \leq u(x) \leq u_b(x) \ \text{ for a.e. } x \in \Omega\big\}.$$

Definition. *A control $\bar{u} \in U_{ad}$ is said to be* optimal *if it satisfies, together with the associated* optimal state $\bar{y} = y(\bar{u})$*, the inequality*

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

A control is said to be locally optimal *in the sense of $L^r(\Omega)$ if there exists some $\varepsilon > 0$ such that the above inequality holds for all $u \in U_{ad}$ such that $\|u - \bar{u}\|_{L^r(\Omega)} \leq \varepsilon$.*

Before stating the first result on the existence of optimal controls, we note two properties of the functionals

$$F(y) = \int_\Omega \varphi\big(x, y(x)\big)\, dx, \qquad Q(u) = \int_\Omega \psi\big(x, u(x)\big)\, dx. \tag{4.34}$$

Both functionals are composed of a Nemytskii operator and a continuous linear integral operator from $L^1(\Omega)$ into $\mathbb{R}$. By virtue of Lemma 4.11 on page 198, F is Lipschitz continuous on the set $\{y \in L^2(\Omega) : \|y\|_{L^\infty(\Omega)} \le M\}$, for any fixed but arbitrary $M > 0$. The same holds for Q on U_{ad}, since U_{ad} is bounded with respect to the L^∞ norm. Moreover, the reader will be asked in Exercise 4.5 to show that Q is convex on U_{ad} provided that ψ satisfies the convexity condition stated in (4.35) below.

Theorem 4.15. *Suppose that Assumption 4.14 holds, and assume that ψ is convex in u, that is,*

$$\psi\big(x, \lambda\, u + (1-\lambda)v\big) \le \lambda\, \psi(x,u) + (1-\lambda)\, \psi(x,v) \tag{4.35}$$

for almost every $x \in \Omega$, all u, $v \in \mathbb{R}$, and every $\lambda \in (0,1)$. Then the problem (4.31)–(4.33) *with distributed control has at least one optimal control $\bar u$ with associated optimal state $\bar y = y(\bar u) \in H^1(\Omega) \cap C(\bar\Omega)$.*

Proof: We first bring the elliptic equation in (4.32) into the form (4.5) on page 183, using the transformation from (4.30). Recalling the remarks following (4.32), we may apply Theorem 4.8 on page 192 with $\tilde d$ in place of d. Consequently, the state problem (4.32) has for every control $u \in U_{ad}$ a uniquely determined state $y = y(u) \in H^1(\Omega) \cap C(\bar\Omega)$.

Next, observe that U_{ad} is a bounded subset of $L^\infty(\Omega)$ and thus bounded in any space $L^r(\Omega)$ for $r > N/2$. Without loss of generality, we may assume $r \ge 2$. Hence, we may employ the estimate (4.11) on page 191 to conclude that there is some constant $M > 0$ such that

$$\|y(u)\|_{C(\bar\Omega)} \le M$$

for all states $y(u)$ that correspond to a control $u \in U_{ad}$.

By virtue of Assumption 4.14, the functional $J(y,u) = F(y) + Q(u)$ is bounded from below. Therefore, the infimum

$$j = \inf_{u \in U_{ad}} J\big(y(u), u\big)$$

exists. Let $\big\{(y_n, u_n)\big\}_{n=1}^\infty$ be a minimizing sequence, that is, let $u_n \in U_{ad}$ and $y_n = y(u_n)$, for $n \in \mathbb{N}$, be such that $J(y_n, u_n) \to j$ as $n \to \infty$.

We now interpret U_{ad} as a subset of $L^r(\Omega)$. The reader will be asked in Exercise 4.6 to verify that U_{ad} is nonempty, closed, bounded, and convex in $L^r(\Omega)$. Since $L^r(\Omega)$ is a reflexive Banach space, it follows from Theorem

2.11 that U_{ad} is weakly sequentially compact. Hence, there exists a sequence, without loss of generality $\{u_n\}_{n=1}^{\infty}$ itself, that converges weakly in $L^r(\Omega)$ to some $\bar{u} \in U_{ad}$, i.e.,

$$u_n \rightharpoonup \bar{u} \quad \text{as } n \to \infty.$$

Now we have found a candidate for the optimal control, but this has yet to be proved. To this end, we have to show that the state sequence $\{y_n\}_{n=1}^{\infty}$ converges in a suitable sense. This is not as straightforward as in the linear-quadratic case. To begin with, consider the sequence

$$z_n(\cdot) = d\big(\cdot, y_n(\cdot)\big), \ n \in \mathbb{N}.$$

Now recall that $\|y_n\|_{L^\infty(\Omega)} \leq M$ for all $n \in \mathbb{N}$. Then $\{z_n\}_{n=1}^{\infty}$ is also bounded in $L^\infty(\Omega)$ and, a fortiori, in $L^r(\Omega)$. Therefore, a subsequence, without loss of generality $\{z_n\}_{n=1}^{\infty}$ itself, converges weakly in $L^r(\Omega)$ to some $z \in L^r(\Omega)$.

Next, observe that y_n solves the boundary value problem

$$\begin{aligned} -\Delta y_n + y_n &= R_n \\ \partial_\nu y_n &= 0, \end{aligned}$$

where the right-hand side $R_n := -d(x, y_n) + u_n$ converges weakly in $L^r(\Omega)$ to $-z + \bar{u}$. By virtue of Theorem 2.6 on page 35, the mapping $R_n \mapsto y_n$ is linear and continuous from $L^2(\Omega)$ into $H^1(\Omega)$, and since $r \geq 2$, also from $L^r(\Omega)$ into $H^1(\Omega)$. Since every continuous linear operator is also weakly continuous, $\{y_n\}_{n=1}^{\infty}$ must converge weakly in $H^1(\Omega)$ to some $\bar{y} \in H^1(\Omega)$, i.e.

$$y_n \rightharpoonup \bar{y} \quad \text{as } n \to \infty.$$

Moreover, since $H^1(\Omega)$ is by Theorem 7.4 on page 356 compactly embedded in $L^2(\Omega)$, we also have the strong convergence

$$\|y_n - \bar{y}\|_{L^2(\Omega)} \to 0 \quad \text{as } n \to \infty.$$

The function $\bar{y}$ is the natural candidate for the desired optimal state.

After these somewhat lengthy preliminaries, the rest of the proof is straightforward. First, recall that $|y_n(x)| \leq M \ \forall x \in \bar{\Omega}$. Now, the set $\big\{v \in L^r(\Omega) : \|v\|_{L^\infty(\Omega)} \leq M\big\}$ is bounded, closed, and convex, and thus weakly sequentially closed. Therefore, $\bar{y}$ belongs to this set.

We now aim to show that $\bar{y}$ is the weak solution associated with $\bar{u}$. Once this is shown, we also know that $y \in C(\bar{\Omega})$. Now observe that, by Lemma 4.11 on page 198, it follows from the boundedness of $\{y_n\}_{n=1}^{\infty}$ in $L^\infty(\Omega)$ that

$$\|d(\cdot, y_n) - d(\cdot, \bar{y})\|_{L^2(\Omega)} \leq L(M)\, \|y_n - \bar{y}\|_{L^2(\Omega)}, \tag{4.36}$$

and thus $d(\cdot, y_n) \to d(\cdot, \bar{y})$ in $L^2(\Omega)$. Moreover, the sequence $\{u_n\}_{n=1}^{\infty}$ also converges weakly in $L^2(\Omega)$ to $\bar{u}$.

Now, we have, for any $n \in \mathbb{N}$,

$$\int_\Omega \nabla y_n \cdot \nabla v \, dx + \int_\Omega d(\cdot, y_n)\, v \, dx = \int_\Omega u_n \, v \, dx \quad \forall\, v \in H^1(\Omega).$$

Passing to the limit as $n \to \infty$, we see that $y_n \rightharpoonup \bar{y}$ in $H^1(\Omega)$ yields the convergence of the first integral; moreover, from $y_n \to \bar{y}$ in $L^2(\Omega)$ and $\|y_n\|_{L^\infty(\Omega)} \le M$ the convergence of the second integral follows, and from $u_n \rightharpoonup \bar{u}$ in $L^r(\Omega)$ that of the third integral. In summary, we have

$$\int_\Omega \nabla \bar{y} \cdot \nabla v \, dx + \int_\Omega d(\cdot, \bar{y})\, v \, dx = \int_\Omega \bar{u} \, v \, dx \quad \forall\, v \in H^1(\Omega).$$

In other words, $\bar{y}$ is the weak solution corresponding to the right-hand side $\bar{u}$, that is, $\bar{y} = y(\bar{u})$.

The proof is still not complete: it remains to show the optimality of $\bar{u}$. At this point, one might be tempted to believe that the continuity of the functional Q already suffices to conclude from the convergence $u_n \rightharpoonup \bar{u}$ that $Q(u_n) \to Q(\bar{u})$. Unfortunately, however, nonlinear continuous functionals need not be weakly continuous, so this line of argument is inconclusive.

Here, the convexity of the functional Q comes to the rescue. In fact, by Theorem 2.12 on page 47, Q is therefore weakly lower semicontinuous, that is,

$$u_n \rightharpoonup \bar{u} \qquad \Rightarrow \qquad \liminf_{n\to\infty} Q(u_n) \ge Q(\bar{u}).$$

In summary, we have

$$\begin{aligned} j &= \lim_{n\to\infty} J(y_n, u_n) = \lim_{n\to\infty} F(y_n) + \liminf_{n\to\infty} Q(u_n) \\ &= F(\bar{y}) + \liminf_{n\to\infty} Q(u_n) \ge F(\bar{y}) + Q(\bar{u}) = J(\bar{y}, \bar{u}). \end{aligned}$$

By definition of the infimum j, we therefore must have $J(\bar{y}, \bar{u}) = j$, which proves the optimality. □

Analogous reasoning yields the existence of an optimal control for zero Dirichlet boundary conditions.

Remarks. In the above proof, the following two conclusions would have been wrong:

(i) To conclude that $y_n \rightharpoonup y \;\Rightarrow\; d(\cdot, y_n) \rightharpoonup d(\cdot, y)$ without knowing the strong convergence $y_n \to y$: indeed, nonlinear mappings need not be weakly continuous.

(ii) To conclude that $y_n \to y$ in $L^2(\Omega) \;\Rightarrow\; d(\cdot, y_n) \to d(\cdot, y)$ without knowing that $\|y_n\|_{L^\infty(\Omega)} \le M \;\; \forall\, n \in \mathbb{N}$.

Observe that only the boundedness and Lipschitz condition of order $k = 0$ in Assumption 4.14 have to be postulated for φ and ψ for the theorem to be valid.

Since the state equation is nonlinear, the optimization problem is nonconvex with respect to u. Therefore, the uniqueness of the optimal control $\bar{u}$ cannot be shown without imposing additional assumptions. Theoretically, arbitrarily many global and local minima are possible. The following examples demonstrate how strange things can be even for the simplest nonlinear optimization problems in Banach spaces. We will come back to this issue later in the chapter; these examples are not really optimal control problems, since they do not have a differential equation as constraint.

Examples.

(i) Consider the problem

$$\min f(u) := -\int_0^1 \cos(u(x))dx, \qquad 0 \le u(x) \le 2\pi, \quad u \in L^\infty(0,1). \tag{4.37}$$

Evidently, -1 is the optimal value, attained for instance at $\bar{u} \equiv 0$. But there are uncountably many other global solutions, namely, all measurable functions u that attain only the values 0 and 2π. These global minima can be arbitrarily close to each other with respect to the L^2 norm, while their L^∞ distance is always $2\,\pi$.

(ii) The situation is similar (cf. Alt and Malanowski [**AM93**]) for the problem

$$\min \int_0^1 |u^2(x) - 1|^2\, dx, \qquad |u(x)| \le 1, \quad u \in L^\infty(0,1).$$

◇

Boundary control. Under Assumption 4.14, the existence of at least one solution to the boundary control problem (4.49)–(4.51) to be discussed on page 218 can be shown in a similar way.

4.5. The control-to-state operator

For all of the problems addressed in this book, a unique state y is assigned to the control. In the linear-quadratic case, we assigned to the state y the part $S\,u$ that actually occurs in the cost functional. Here, we no longer follow this approach, simply because the diversity of the spaces and dual spaces involved would necessitate a very technical exposition. We therefore only consider the mapping $u \mapsto y$, which is generally denoted by G. The range of G will always be a subset of $Y = C(\bar{\Omega}) \cap H^1(\Omega)$. As before, we will often write $y(u)$ in place of $G(u)$, presuming that any confusion with the value $y(x)$ of y at the space point x will be clarified by the context. We begin our analysis with the control acting as a source in the domain.

4.5.1. Distributed control. We consider the state problem (4.32)

$$\boxed{\begin{aligned} -\Delta y + d(x,y) &= u && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma. \end{aligned}}$$

By Theorem 4.8 on page 192, for any control $u \in U := L^r(\Omega)$ with $r > N/2$, there exists a unique state $y \in Y = H^1(\Omega) \cap C(\bar\Omega)$, provided that the corresponding assumptions are met (which we shall take to be the case). We denote the associated control-to-state operator by $G : U \to Y$, $G(u) = y$.

Theorem 4.16. *Suppose that Assumption 4.14 on page 206 holds for Ω and d. Then G is a Lipschitz continuous mapping from $L^r(\Omega)$, $r > N/2$, into $H^1(\Omega) \cap C(\bar\Omega)$, that is, there is a constant $L > 0$ such that*

$$\|y_1 - y_2\|_{H^1(\Omega)} + \|y_1 - y_2\|_{C(\bar\Omega)} \le L \, \|u_1 - u_2\|_{L^r(\Omega)}$$

whenever $u_i \in L^r(\Omega)$ and $y_i = G(u_i)$, $i = 1, 2$.

Proof: By virtue of Theorem 4.10 on page 194, $y_i \in C(\bar\Omega)$ for $i = 1, 2$. Subtracting the equations satisfied by y_1 and y_2, we see that

$$\begin{aligned} -\Delta(y_1 - y_2) + d(x, y_1) - d(x, y_2) &= u_1 - u_2 \\ \partial_\nu (y_1 - y_2) &= 0. \end{aligned} \tag{4.38}$$

Evidently, we have

$$d\big(x, y_1(x)\big) - d\big(x, y_2(x)\big) = -\int_0^1 \frac{d}{ds}\, d\big(x, y_1(x) + s\,(y_2(x) - y_1(x))\big)\, ds$$

$$= \int_0^1 d_y\big(x, y_1(x) + s\,\big(y_2(x) - y_1(x)\big)\big)\, ds\ \big(y_1(x) - y_2(x)\big).$$

Owing to the continuity of the functions d_y, y_1, and y_2, the integral in the second line defines a bounded and measurable function $c_0 = c_0(x)$, which is nonnegative since the mapping $y \mapsto d(x, y)$ is increasing. On E_d the integrand satisfies

$$d_y\big(x, y_1(x) + s\,\big(y_2(x) - y_1(x)\big)\big) \ge \lambda_d > 0,$$

so that $c_0(x) \ge \lambda_d$ for almost every $x \in E_d$. Of course, c_0 also depends on y_1 and y_2, but this is immaterial for the argumentation to follow.

Now let $y = y_1 - y_2$ and $u = u_1 - u_2$. Then, in view of (4.38),

$$\begin{aligned} -\Delta y + c_0(x)\, y &= u \\ \partial_\nu y &= 0. \end{aligned}$$

Because $c_0 \geq 0$, the function $c_0(x)\, y$ is increasing with respect to y. We may therefore apply Theorem 4.7 on page 191 with the above c_0 and the specifications $d(x, y) = 0$, $f = u$, and $g = b = \alpha = 0$. Invoking (4.11), we immediately deduce that

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq L\, \|u\|_{L^r(\Omega)}$$

for all u and the associated y. Since the above boundary value problem is linear, we may take $u = u_1 - u_2$ and $y = y_1 - y_2$ to arrive at the asserted estimate. □

Remark. In both the preceding and the next theorem, only the boundedness and Lipschitz conditions of order $k = 1$ from Assumption 4.14 are needed for the asserted results to be valid.

Next, we determine the Fréchet derivative of the control-to-state operator at a fixed point $\bar{u}$. In the later applications, $\bar{u}$ will be a locally optimal control. We have the following result.

Theorem 4.17. *Suppose that Assumption 4.14 on page 206 holds. Then for every $r > N/2$ the control-to-state operator G is Fréchet differentiable from $L^r(\Omega)$ into $H^1(\Omega) \cap C(\bar{\Omega})$. Its directional derivative at $\bar{u} \in L^r(\Omega)$ in the direction u is given by*

$$G'(\bar{u})\, u = y,$$

where y denotes the weak solution to the boundary value problem linearized at $\bar{y} = G(\bar{u})$:

$$\begin{aligned} -\Delta y + d_y(x, \bar{y})\, y &= u \quad \text{in } \Omega \\ \partial_\nu y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.39}$$

Proof: We have to show that

$$G(\bar{u} + u) - G(\bar{u}) = D\, u + r(\bar{u}, u)$$

with a continuous linear operator $D : L^r(\Omega) \to H^1(\Omega) \cap C(\bar{\Omega})$ and a mapping r that satisfies

$$\frac{\|r(\bar{u}, u)\|_{H^1(\Omega)\cap C(\bar{\Omega})}}{\|u\|_{L^r(\Omega)}} \to 0 \quad \text{as } \|u\|_{L^r(\Omega)} \to 0.$$

Here, we have put $\|r\|_{H^1(\Omega)\cap C(\bar{\Omega})} := \|r\|_{H^1(\Omega)} + \|r\|_{C(\bar{\Omega})}$. It then follows that $G'(\bar{u}) = D$.

The boundary value problems satisfied by $\bar{y} = y(\bar{u})$ and $\tilde{y} = y(\bar{u} + u)$ read, respectively,

$$\begin{aligned} -\Delta \bar{y} + d(x, \bar{y}) &= \bar{u} & \qquad -\Delta \tilde{y} + d(x, \tilde{y}) &= \bar{u} + u \\ \partial_\nu \bar{y} &= 0 & \qquad \partial_\nu \tilde{y} &= 0. \end{aligned}$$

Subtracting them yields

$$\begin{aligned} -\Delta(\tilde{y}-\bar{y}) + \overbrace{d(x,\tilde{y})-d(x,\bar{y})}^{=d_y(x,\bar{y})(\tilde{y}-\bar{y})+r_d} &= u \\ \partial_\nu(\tilde{y}-\bar{y}) &= 0. \end{aligned}$$

The Nemytskii operator $\Phi(y) = d\big(\cdot, y(\cdot)\big)$ is, by Lemma 4.12 on page 202, Fréchet differentiable from $C(\bar{\Omega})$ into $L^\infty(\Omega)$. Therefore,

$$\Phi(\tilde{y}) - \Phi(\bar{y}) = d\big(\cdot,\tilde{y}(\cdot)\big) - d\big(\cdot,\bar{y}(\cdot)\big) = d_y\big(\cdot,\bar{y}(\cdot)\big)\,\big(\tilde{y}(\cdot)-\bar{y}(\cdot)\big) + r_d, \tag{4.40}$$

with a remainder r_d such that

$$\frac{\|r_d\|_{L^\infty(\Omega)}}{\|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})}} \to 0 \quad \text{as } \|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})} \to 0.$$

The reader will be asked in Exercise 4.7 to show that this implies that

$$\tilde{y} - \bar{y} = y + y_\rho,$$

with the solution y to (4.39) and a remainder y_ρ that solves the boundary value problem

$$\begin{aligned} -\Delta y_\rho + d_y(\cdot,\bar{y})\, y_\rho &= -r_d \\ \partial_\nu y_\rho &= 0. \end{aligned} \tag{4.41}$$

In this connection, recall that $d_y(x,\bar{y}) \geq \lambda_d > 0$ in E_d, so that this problem is uniquely solvable. From the Lipschitz continuity shown in Theorem 4.16 it follows that

$$\|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})} + \|\tilde{y}-\bar{y}\|_{H^1(\Omega)} \leq L\,\|u\|_{L^r(\Omega)}.$$

Moreover,

$$\frac{\|r_d\|_{L^\infty(\Omega)}}{\|u\|_{L^r(\Omega)}} = \frac{\|r_d\|_{L^\infty(\Omega)}}{\|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})}}\,\frac{\|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})}}{\|u\|_{L^r(\Omega)}} \leq \frac{\|r_d\|_{L^\infty(\Omega)}}{\|\tilde{y}-\bar{y}\|_{C(\bar{\Omega})}}\,L,$$

and thus $\|r_d\|_{L^\infty(\Omega)} = o(\|u\|_{L^r(\Omega)})$. By (4.41), we also have

$$\|y_\rho\|_{C(\bar{\Omega})} + \|y_\rho\|_{H^1(\Omega)} = o(\|u\|_{L^r(\Omega)}).$$

Denoting the continuous linear mapping $u \mapsto y$ by D, we conclude that

$$G(\bar{u}+u) - G(\bar{u}) = \tilde{y} - \bar{y} = D\,u + y_\rho = D\,u + r(\bar{u},u),$$

where $r(\bar{u},u) = y_\rho$ has the required properties. This concludes the proof of the assertion. □

Conclusion. *A fortiori, G is Fréchet differentiable from $L^\infty(\Omega)$ into $H^1(\Omega)\cap C(\bar{\Omega})$.*

4.5.2. Boundary control. In this case, the situation is quite similar. In fact, under Assumption 4.14 on page 206, G is continuously Fréchet differentiable from $U = L^s(\Gamma)$ into $Y = H^1(\Omega) \cap C(\bar{\Omega})$ for all $s > N - 1$. The directional derivative at $\bar{u} \in L^s(\Gamma)$ in the direction u is given by

$$G'(\bar{u})\, u = y,$$

where y denotes the weak solution to the boundary value problem linearized at $\bar{y} = G(\bar{u})$, that is,

$$\begin{aligned} -\Delta y &= 0 \\ \partial_\nu y + b_y(x, \bar{y})\, y &= u. \end{aligned}$$

4.6. Necessary optimality conditions

4.6.1. Distributed control. In the following, let $\bar{u} \in L^\infty(\Omega)$ denote some (in the sense of the L^∞ norm) locally optimal control of the problem (4.31)–(4.33) on page 207. We derive the first-order necessary conditions that have to be obeyed by $\bar{u}$ and the associated state $\bar{y}$.

For y we can write $y(u) = G(u)$, with the control-to-state operator $G : L^\infty(\Omega) \to H^1(\Omega) \cap C(\bar{\Omega})$. The cost functional thus attains the form

$$J(y, u) = J\big(G(u), u\big) = F\big(G(u)\big) + Q(u) =: f(u),$$

where F and Q are defined as in (4.34). Under Assumption 4.14, f is a Fréchet differentiable functional in $L^\infty(\Omega)$; indeed, F, Q, and G are, by virtue of Lemma 4.12 on page 202 and Theorem 4.17, Fréchet differentiable.

Suppose now that $\bar{u}$ is locally optimal, U_{ad} is convex, and $u \in U_{ad}$ is arbitrarily chosen. Then for all sufficiently small $\lambda > 0$ the convex combination $v := \bar{u} + \lambda(u - \bar{u})$ belongs to both U_{ad} and the ε-neighborhood of $\bar{u}$ in which $f(\bar{u}) \le f(v)$ holds. Hence, for all sufficiently small $\lambda > 0$,

$$f\big(\bar{u} + \lambda(u - \bar{u})\big) \ge f(\bar{u}).$$

Division by λ and passage to the limit as $\lambda \downarrow 0$ lead to the variational inequality of Lemma 2.21 on page 63. We have thus shown the following result.

Lemma 4.18. *Suppose that Assumption 4.14 on page 206 holds, and let $\bar{u}$ be a locally optimal control for problem* (4.31)–(4.33). *Then we have the variational inequality*

$$f'(\bar{u})(u - \bar{u}) \ge 0 \quad \forall\, u \in U_{ad}. \tag{4.42}$$

The above result is valid for all nonlinear functionals of the type $f(u) = J\big(y(u), u\big)$ that we subsequently consider. Using the chain rule, we immediately see that the directional derivative at $\bar{u}$ in the direction h is given

by

$$\begin{aligned} f'(\bar u)\,h &= F'\big(G(\bar u)\big)\,G'(\bar u)\,h + Q'(\bar u)\,h \\ &= F'(\bar y)\,y + Q'(\bar u)\,h \\ &= \int_\Omega \varphi_y\big(x,\bar y(x)\big)\,y(x)\,dx + \int_\Omega \psi_u\big(x,\bar u(x)\big)\,h(x)\,dx. \end{aligned} \tag{4.43}$$

Here, $y = G'(\bar u)\,h$ is by virtue of Theorem 4.17 the weak solution to the linearized boundary value problem

$$\begin{aligned} -\Delta y + d_y(x,\bar y)\,y &= h \\ \partial_\nu y &= 0. \end{aligned} \tag{4.44}$$

Next, we define the adjoint state $p \in H^1(\Omega) \cap C(\bar\Omega)$ as the unique weak solution to the *adjoint equation*

$$\boxed{\begin{aligned} -\Delta p + d_y(x,\bar y)\,p &= \varphi_y\big(x,\bar y(x)\big) && \text{in } \Omega \\ \partial_\nu p &= 0 && \text{on } \Gamma. \end{aligned}} \tag{4.45}$$

Lemma 4.19. *Let y be the weak solution to problem* (4.44) *for given $h \in L^2(\Omega)$, and let p be the adjoint state defined as the weak solution to problem* (4.45)*. Then*

$$\int_\Omega \varphi_y\big(x,\bar y(x)\big)\,y(x)\,dx = \int_\Omega p(x)h(x)\,dx.$$

Proof: The assertion follows directly from Lemma 2.31 on page 74, with the specifications $a_\Omega(x) = \varphi_y\big(x,\bar y(x)\big)$, $c_0(x) = d_y\big(x,\bar y(x)\big)$ (observe that $d_y(\cdot,\bar y(\cdot)) \neq 0$), $\beta_\Omega = 1$, and $a_\Gamma = \alpha = \beta_\Gamma = 0$. □

As a simple conclusion, the following expression for the directional derivative of the reduced functional f at $\bar u$ in the direction $h \in L^\infty(\Omega)$ results:

$$f'(\bar u)\,h = \int_\Omega \big(p(x) + \psi_u(x,\bar u(x))\big)\,h(x)\,dx. \tag{4.46}$$

Moreover, we obtain the desired necessary optimality condition.

Theorem 4.20. *Suppose that Assumption 4.14 holds. Then every locally optimal control $\bar u$ for problem* (4.31)–(4.33) *satisfies, together with the associated adjoint state $p \in H^1(\Omega) \cap C(\bar\Omega)$ defined by* (4.45)*, the variational inequality*

$$\int_\Omega (p(x) + \psi_u(x,\bar u(x))(u(x) - \bar u(x))\,dx \geq 0 \quad \forall\, u \in U_{ad}. \tag{4.47}$$

As in Chapter 2, we can reformulate the variational inequality in terms of a minimum principle.

Conclusion. *Suppose that Assumption 4.14 holds, and suppose that $\bar{u}$ is a locally optimal control of problem* (4.31)–(4.33) *with associated adjoint state p. Then for almost every $x \in \Omega$ the minimum of the problem*

$$\min_{u_a(x) \le v \le u_b(x)} \big\{ \big(p(x) + \psi_u(x, \bar{u}(x)) \big)\, v \big\} \tag{4.48}$$

is attained at $v = \bar{u}(x)$.

Special case: The function $\psi(x,u) = \frac{\lambda}{2}\, u^2$, with $\lambda > 0$, obviously satisfies the assumptions. Clearly, $\psi_u(x,u) = \lambda\, u$; hence, the minimum for the problem

$$\min_{u_a(x) \le v \le u_b(x)} \big\{ \big(p(x) + \lambda\, \bar{u}(x) \big)\, v \big\}$$

is attained at $v = \bar{u}(x)$ for almost every $x \in \Omega$. Therefore, we have the projection formula as in (2.58) on page 70:

$$\boxed{\bar{u}(x) = \mathbb{P}_{[u_a(x), u_b(x)]} \left\{ -\frac{1}{\lambda}\, p(x) \right\} \quad \text{for a.e. } x \in \Omega.}$$

If u_a and u_b are continuous, then so is $\bar{u}$; in fact, we have $p \in H^1(\Omega) \cap C(\bar{\Omega})$, and the projection operator maps continuous functions to continuous ones. If, in addition, $u_a, u_b \in H^1(\Omega)$, then, by the same token, $\bar{u} \in H^1(\Omega) \cap C(\bar{\Omega})$.

Example. Consider the "superconductivity" problem

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

subject to $-2 \le u(x) \le 2$ and

$$\begin{aligned} -\Delta y + y + y^3 &= u \\ \partial_\nu y &= 0. \end{aligned}$$

This is a special case of problem (4.31)–(4.33) on page 207, with the specifications

$$\varphi(x,y) = \frac{1}{2}\, \big(y - y_\Omega(x) \big)^2, \quad \psi(x,u) = \frac{\lambda}{2}\, u^2, \quad d(x,y) = y + y^3.$$

If we assume $y_\Omega \in L^\infty(\Omega)$, then all requirements (measurability, boundedness, differentiability, monotonicity of d, convexity of ψ with respect to u)

are met. Theorem 4.15 on page 208 yields the existence of an optimal control $\bar{u}$. The adjoint equation for p reads

$$\begin{aligned} -\Delta p + p + 3\,\bar{y}^2\,p &= \bar{y} - y_\Omega \\ \partial_\nu p &= 0. \end{aligned}$$

Together with the solution $p \in H^1(\Omega) \cap C(\bar{\Omega})$ to the adjoint system, $\bar{u}$ must obey the variational inequality

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

In the case where $\lambda > 0$, the usual projection relation follows, and we have $\bar{u} \in H^1(\Omega) \cap C(\bar{\Omega})$ provided that $u_a,\, u_b \in H^1(\Omega) \cap C(\bar{\Omega})$. If $\lambda = 0$, then obviously $\bar{u}(x) = -2$ sign $p(x)$. ◇

Test example. The reader will be asked in Exercise 4.8 to verify that the necessary first-order optimality conditions are fulfilled for $\bar{u}(x) \equiv 2$ in the case where $\lambda = 1$ and $y_\Omega = 9$. ◇

Remark. In the above example, the assumption $y_\Omega \in L^r(\Omega)$, $r > N/2$, would be sufficient. Then it also follows from our regularity results that p and thus also the optimal control $\bar{u}$ are continuous if $\lambda > 0$. In addition, the boundedness and Lipschitz conditions from Assumption 4.14 are needed only up to order $k = 1$.

4.6.2. Boundary control. Consider the problem

$$\min J(y,u) := \int_\Omega \varphi\big(x, y(x)\big)\,dx + \int_\Gamma \psi\big(x, u(x)\big)\,ds(x), \tag{4.49}$$

subject to

$$\boxed{\begin{aligned} -\Delta y + y &= 0 \quad \text{in } \Omega \\ \partial_\nu y + b(x,y) &= u \quad \text{on } \Gamma \end{aligned}} \tag{4.50}$$

and

$$u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Gamma. \tag{4.51}$$

Here, we put

$$U_{ad} = \big\{u \in L^\infty(\Gamma) \,:\, u_a(x) \leq u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Gamma\big\}.$$

Under Assumption 4.14 on page 206, the control-to-state operator $G : u \mapsto y$ maps $L^\infty(\Gamma)$ into $H^1(\Omega) \cap C(\bar{\Omega})$. This would also be the case if $-\Delta + I$ were replaced by $-\Delta$. However, below we will discuss an example in which absence of the term I would lead to problems in the adjoint equation when $\bar{y} = 0$.

The necessary optimality conditions for locally optimal controls $\bar{u}$ can be derived in a similar way as in the case of distributed controls, except that G and G' have a different meaning. First, it results in the variational inequality

(4.52)
$$\int_\Omega \varphi_y\big(x,\bar{y}(x)\big)\,y(x)\,dx + \int_\Gamma \psi_u\big(x,\bar{u}(x)\big)\,\big(u(x)-\bar{u}(x)\big)\,ds(x) \geq 0 \quad \forall u \in U_{ad},$$

with the directional derivative $y = G'(\bar{u})(u-\bar{u})$ at $\bar{u}$ in the direction $u-\bar{u}$. In analogy to Theorem 4.17 on page 213, one obtains that y solves the boundary value problem linearized at $\bar{y}$:

(4.53)
$$\begin{aligned} -\Delta y + y &= 0 \\ \partial_\nu y + b_y(x,\bar{y})\,y &= u-\bar{u}. \end{aligned}$$

The adjoint state p is defined as the weak solution to the *adjoint equation*

(4.54)
$$\boxed{\begin{aligned} -\Delta p + p &= \varphi_y\big(x,\bar{y}(x)\big) && \text{in } \Omega \\ \partial_\nu p + b_y(x,\bar{y})\,p &= 0 && \text{on } \Gamma. \end{aligned}}$$

We have $p \in H^1(\Omega)\cap C(\bar{\Omega})$, since the right-hand side φ_y belongs to $L^\infty(\Omega)$. For this conclusion to hold, $\varphi_y \in L^r(\Omega)$, for some $r > N/2$, already suffices. Invoking Lemma 2.31 on page 74 with the specifications $a_\Omega(x) = \varphi_y\big(x,\bar{y}(x)\big)$, $\alpha(x) = b_y\big(x,\bar{y}(x)\big)$, $\beta_\Omega = 0$, and $\beta_\Gamma = 1$, we find that

$$\int_\Omega \varphi_y\big(x,\bar{y}(x)\big)\,y(x)\,dx = \int_\Gamma p(x)\,\big(u(x)-\bar{u}(x)\big)\,ds(x) \quad \forall\, u \in L^2(\Gamma).$$

Substituting this into the variational inequality (4.52) yields the following result.

Theorem 4.21. *Suppose that Assumption 4.14 on page 206 holds, and let $\bar{u}$ be a locally optimal control for the boundary control problem* (4.49)–(4.51)*, with associated adjoint state p defined as the solution to the boundary value problem* (4.54)*. Then $\bar{u}$ satisfies the variational inequality*

(4.55)
$$\int_\Gamma \big(p(x)+\psi_u(x,\bar{u}(x))\big)\big(u(x)-\bar{u}(x)\big)\,ds(x) \geq 0 \quad \forall\, u \in U_{ad}.$$

Again, the variational inequality can be rephrased in terms of a pointwise minimum principle. Since this is completely analogous to the case of distributed controls, we omit the details. As in the distributed control case, the directional derivative of $f(u) = J\big(G(u),u\big)$ at $\bar{u}$ in the direction u can

be expressed in the form

$$f'(\bar u)\,u = \int_\Gamma \big(p(x) + \psi_u(x,\bar u(x))\big)\,u(x)\,ds(x). \tag{4.56}$$

Example. Consider the boundary control problem

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

subject to

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

and $0 \le u(x) \le 1$.

This is a special case of the boundary control problem (4.49)–(4.51), with

$$\varphi(x,y) = \frac{1}{2}\,\big(y - y_\Omega(x)\big)^2, \quad \psi(x,u) = \frac{\lambda}{2}\,u^2, \quad b(x,y) = y^3|y|.$$

Actually, the boundary condition is of Stefan–Boltzmann type, because $y^3|y| = y^4$ for nonnegative y. However, unlike y^4, the function $y^3|y|$ is monotone increasing. Once more, we assume $y_\Omega \in L^\infty(\Omega)$. Obviously, the requested properties concerning measurability, differentiability, monotonicity of b, and convexity of ψ with respect to u all hold. Therefore, in analogy to Theorem 4.15 on page 208, at least one optimal control $\bar u$ exists.

A simple calculation shows that $b'(y) = 4\,y^2\,|y|$. Hence, the adjoint boundary value problem (4.54) becomes

$$\begin{aligned} -\Delta p + p &= \bar y - y_\Omega \\ \partial_\nu p + 4\,\bar y^2\,|\bar y|\,p &= 0, \end{aligned}$$

and it follows that, with its solution p, for all u such that $0 \le u(x) \le 1$ for almost every $x \in \Gamma$ the inequality

$$\int_\Gamma (\lambda\bar u + p)(u - \bar u)\,ds \ge 0$$

holds.

4.7. Application of the formal Lagrange method

The formal Lagrange method is a powerful tool for the derivation of optimality conditions also for the nonconvex problems treated in this chapter. One easily arrives at the correct result, which can then be verified rigorously. We

demonstrate this approach for the following rather general optimal control problem:

$$\text{(4.57)}\quad \min\, J(y,v,u) := \int_\Omega \varphi\big(x,y(x),v(x)\big)\,dx + \int_\Gamma \psi\big(x,y(x),u(x)\big)\,ds(x),$$

subject to

$$\text{(4.58)}\qquad \boxed{\begin{array}{rcll} -\Delta y + d(x,y,v) &=& 0 & \text{in } \Omega \\ \partial_\nu y + b(x,y,u) &=& 0 & \text{on } \Gamma \end{array}}$$

and

$$\text{(4.59)}\qquad \begin{array}{rcccll} v_a(x) &\le& v(x) &\le& v_b(x) & \text{for a.e. } x \in \Omega \\ u_a(x) &\le& u(x) &\le& u_b(x) & \text{for a.e. } x \in \Gamma. \end{array}$$

The cost functional, containing also boundary values of y, is more general than before. Since the method will be explained only formally, we do not state precise assumptions on the involved quantities. Clearly, the functions φ, ψ, d, and b have to be measurable with respect to x and differentiable with respect to y, u, and v. In addition, at least the monotonicity of d and b in y must be postulated.

Since the controls u and v appear nonlinearly in the state problem, a general existence result for optimal controls cannot be expected. We therefore assume that locally optimal controls $\bar u$ and $\bar v$ exist, for which the necessary optimality conditions have to be derived.

In analogy to the problems studied earlier, we denote the sets of admissible controls by $V_{ad} \subset L^\infty(\Omega)$ and $U_{ad} \subset L^\infty(\Gamma)$. The Lagrangian function is formally introduced as

$$\begin{aligned} \mathcal{L}(y,v,u,p) &= J(y,v,u) - \int_\Omega \big(-\Delta y + d(\cdot,y,v)\big)\,p\,dx \\ &\qquad - \int_\Gamma \big(\partial_\nu y + b(\cdot,y,u)\big)\,p\,ds. \end{aligned}$$

It is apparent that the Lagrangian $\mathcal{L}$ is not meaningful if the state space $Y = H^1(\Omega) \cap C(\bar\Omega)$ is used. Therefore, we use formal integration by parts and redefine $\mathcal{L}$ by

$$\text{(4.60)}\quad \begin{aligned} \mathcal{L}(y,v,u,p) &:= J(y,v,u) - \int_\Omega \big(\nabla y \cdot \nabla p + d(\cdot,y,v)\,p\big)\,dx \\ &\qquad - \int_\Gamma b(\cdot,y,u)\,p\,ds. \end{aligned}$$

This definition anticipates that the Lagrange multiplier for the boundary condition will eventually coincide with the boundary values of the multiplier

p for the differential equation. Therefore, we have in all integrals the same function p. However, we caution the reader to use different multipliers p_1 and p_2 for the differential equation and the boundary condition if there is any doubt. In particular, this should be done in the case of a Dirichlet-type boundary control.

We expect there to exist a function $p \in H^1(\Omega) \cap C(\bar{\Omega})$ satisfying the following conditions:

$$\begin{aligned} &D_y\mathcal{L}(\bar{y},\bar{v},\bar{u},p)\,y = 0 && \forall\, y \in H^1(\Omega) \\ &D_v\mathcal{L}(\bar{y},\bar{v},\bar{u},p)(v-\bar{v}) \geq 0 && \forall\, v \in V_{ad} \\ &D_u\mathcal{L}(\bar{y},\bar{v},\bar{u},p)(u-\bar{u}) \geq 0 && \forall\, u \in U_{ad}. \end{aligned}$$

We explain only briefly how these relations are exploited, because the technique is completely analogous to that used in Chapter 2. The first relation yields

$$\begin{aligned} D_y\mathcal{L}(\bar{y},\bar{v},\bar{u},p)\,y \;=\; &\int_\Omega \varphi_y(\cdot,\bar{y},\bar{v})\,y\,dx + \int_\Gamma \psi_y(\cdot,\bar{y},\bar{u})\,y\,ds \\ -\int_\Omega \nabla y\cdot\nabla p\,dx - &\int_\Omega d_y(\cdot,\bar{y},\bar{v})\,y\,p\,dx - \int_\Gamma b_y(\cdot,\bar{y},\bar{u})\,y\,p\,ds \;=\; 0 \end{aligned}$$

for every $y \in H^1(\Omega)$. This is nothing but the variational formulation for the weak solution p to the linear boundary value problem

$$\boxed{\begin{aligned} -\Delta p + d_y(\cdot,\bar{y},\bar{v})\,p &= \varphi_y(\cdot,\bar{y},\bar{v}) \\ \partial_\nu p + b_y(\cdot,\bar{y},\bar{u})\,p &= \psi_y(\cdot,\bar{y},\bar{u}), \end{aligned}} \tag{4.61}$$

which we interpret as the *adjoint equation*. Its solution p, the adjoint state, exists whenever b_y, d_y, φ_y, ψ_y are bounded and measurable and b_y, d_y are nonnegative and not both zero almost everywhere. From the second and third relations we deduce the variational inequalities

$$\begin{aligned} \int_\Omega \big(\varphi_v(\cdot,\bar{y},\bar{v}) - p\,d_v(\cdot,\bar{y},\bar{v})\big)\,(v-\bar{v})\,dx &\geq 0 \qquad \forall v \in V_{ad} \\ \int_\Gamma \big(\psi_u(\cdot,\bar{y},\bar{u}) - p\,b_u(\cdot,\bar{y},\bar{u})\big)\,(u-\bar{u})\,ds &\geq 0 \qquad \forall u \in U_{ad}. \end{aligned} \tag{4.62}$$

The optimality system of the optimal control problem (4.57)–(4.59) then consists of the state equation, the adjoint equation, the two variational inequalities, and the inclusions $u \in U_{ad}$, $v \in V_{ad}$.

Example. As an illustration, let us investigate the problem

$$\min\, J(y,u,v) := \int_\Omega \left[y^2 + y_\Omega\,y + \lambda_1\,v^2 + v_\Omega\,v\right]\,dx + \int_\Gamma \lambda_2\,u^8\,ds,$$

subject to

$$\begin{aligned} -\Delta y + y + e^y &= v && \text{in } \Omega \\ \partial_\nu y + y^4 &= u^4 && \text{on } \Gamma \end{aligned}$$

and

$$-1 \le v(x) \le 1, \qquad 0 \le u(x) \le 1,$$

with prescribed functions y_Ω, $v_\Omega \in L^\infty(\Omega)$. Since the nonlinearity y^4 of Stefan–Boltzmann type is not monotone increasing, it does not fit into the general theory. We therefore replace y^4 by the increasing function $|y|\,y^3$, which is identical to y^4 for $y \ge 0$. Then Theorem 4.7 on page 191 applies, yielding for each pair u and v of admissible controls the existence of a unique solution $y \in H^1(\Omega) \cap C(\bar\Omega)$ to the state problem.

In spite of the occurrence of the nonlinear function u^4 in the boundary condition, the existence of an optimal control can be proved. To this end, we replace u^4 by the new control $\tilde u$. Then we have $\tilde u$ in the boundary condition and $\lambda_2 \tilde u^2$ in the cost functional, while the constraint for u becomes $0 \le \tilde u \le 1$. The problem thus transformed has, in analogy to Theorem 4.15 on page 208, a pair of optimal controls $\bar v$ and $\tilde u$, whence we obtain for the original problem the optimal controls $\bar v$ and $\bar u = \tilde u^{1/4}$.

The optimal control problem under study is a special case of the class of problems (4.57)–(4.59), with the specifications

$$\begin{aligned} &\varphi(x,y,v) = y^2 + y_\Omega(x)\,y + v_\Omega(x)\,v + \lambda_1\,v^2, \quad \psi(x,y,u) = \lambda_2\,u^8, \\ &d(x,y,v) = y + e^y - v, \quad b(x,y,u) = |y|y^3 - u^4. \end{aligned}$$

The boundedness of y_Ω and v_Ω has been assumed solely for the purpose of fitting our problem into this general class of problems. The following optimality conditions for arbitrary locally optimal controls $\bar v$ and $\bar u$ evidently remain valid for merely square integrable functions y_Ω and v_Ω:

Adjoint problem:

$$\begin{aligned} -\Delta p + p + e^{\bar y}\,p &= 2\,\bar y + y_\Omega && \text{in } \Omega \\ \partial_\nu p + 4\,\bar y^2|\bar y|\,p &= 0 && \text{on } \Gamma. \end{aligned}$$

Variational inequalities:

$$\begin{aligned} \int_\Omega (2\lambda_1\,\bar v + v_\Omega + p)(v - \bar v)\,dx &\ge 0 \quad \forall\, v \in V_{ad} \\ \int_\Gamma (8\lambda_2\,\bar u^7 + 4\bar u^3\,p)(u - \bar u)\,ds &\ge 0 \quad \forall\, u \in U_{ad}. \end{aligned}$$

4.8. Pontryagin's maximum principle *

4.8.1. Hamiltonian functions. All of the first-order necessary optimality conditions derived above have, with respect to the controls, the form of a variational inequality, obtained by differentiating the Lagrangian with respect to the control variable.

The famous *Pontryagin maximum principle* avoids the differentiation with respect to the control. It was first proved for optimal control problems involving ordinary differential equations (see Pontryagin et al. [**PBGM62**]). An extension to the case of semilinear parabolic partial differential equations, using the integral equation method, is due to von Wolfersdorf [**vW76, vW77**]; he employed a general technique developed by Bittner [**Bit75**].

More general results for semilinear elliptic and parabolic problems have been established in recent years, beginning with Bonnans and Casas [**BC91**]. Later, the maximum principle was extended to include state constraints; we refer the reader to Casas [**Cas86**] and Bonnans and Casas [**BC95**] for elliptic equations, and to Casas [**Cas97**] and Raymond and Zidani [**RZ99**] for parabolic equations. We also refer in this context to the book [**LY95**] by Li and Yong.

We demonstrate the use of the maximum principle for the optimal control problem (4.57)–(4.59):

$$\min\, J(y,u,v) := \int_\Omega \varphi(x,y,v)\,dx + \int_\Gamma \psi(x,y,u)\,ds,$$

subject to

$$\begin{aligned} -\Delta y + d(x,y,v) &= 0 \qquad \text{in } \Omega \\ \partial_\nu y + b(x,y,u) &= 0 \qquad \text{on } \Gamma \end{aligned}$$

and

$$v_a(x) \le v(x) \le v_b(x) \ \text{ for a.e. } x \in \Omega, \quad u_a(x) \le u(x) \le u_b(x) \ \text{ for a.e. } x \in \Gamma.$$

We choose $Y = H^1(\Omega) \cap C(\bar\Omega)$ as the state space and start from the Lagrangian function (4.60). The integrands of its derivative-free terms are incorporated into two *Hamiltonian functions* that separately collect the terms involving Ω and Γ:

Definition. *The functions* $H^\Omega : \Omega \times \mathbb{R}^4 \to \mathbb{R}$ *and* $H^\Gamma : \Gamma \times \mathbb{R}^4 \to \mathbb{R}$ *defined by*

$$\begin{aligned} H^\Omega(x,y,v,p_0,p) &= p_0\,\varphi(x,y,v) - d(x,y,v)\,p \\ H^\Gamma(x,y,u,p_0,p) &= p_0\,\psi(x,y,u) - b(x,y,u)\,p \end{aligned} \tag{4.63}$$

are called Hamiltonian functions.

The Hamiltonians are functions of real variables; their arguments are not functions. The necessary optimality conditions for a locally optimal pair $(\bar u, \bar v)$ of controls, given by (4.61) and (4.62), can be rephrased more elegantly in terms of the Hamiltonian: the adjoint system is equivalent to

$$\begin{aligned} -\Delta p &= D_y H^\Omega(x, \bar y, \bar v, 1, p) && \text{in } \Omega \\ \partial_\nu p &= D_y H^\Gamma(x, \bar y, \bar u, 1, p) && \text{on } \Gamma, \end{aligned} \tag{4.64}$$

and the variational inequalities attain the pointwise form

$$\begin{aligned} D_v H^\Omega\big(x, \bar y(x), \bar v(x), 1, p(x)\big)\big(v - \bar v(x)\big) &\geq 0 \qquad \forall\, v \in [v_a(x), v_b(x)], \\ D_u H^\Gamma\big(x, \bar y(x), \bar u(x), 1, p(x)\big)\big(u - \bar u(x)\big) &\geq 0 \qquad \forall\, u \in [v_a(x), v_b(x)], \end{aligned}$$

for almost all $x \in \Omega$ and $x \in \Gamma$, respectively. From this, we immediately deduce *weak minimum principles*, for instance,

$$\begin{aligned} &\min_{v \in [v_a(x), v_b(x)]} \Big\{ D_v H^\Omega\big(x, \bar y(x), \bar v(x), 1, p(x)\big)\, v \Big\} \\ &\qquad = D_v H^\Omega\big(x, \bar y(x), \bar v(x), 1, p(x)\big)\, \bar v(x) \qquad \text{for a.e. } x \in \Omega. \end{aligned} \tag{4.65}$$

Hence, the minimum in (4.65) is almost everywhere attained at $\bar v(x)$.

4.8.2. The maximum principle. If the Hamiltonian H^Ω is convex with respect to v, then the weak minimum condition (4.65) is equivalent to the minimum condition

$$H^\Omega\big(x, \bar y(x), \bar v(x), 1, p(x)\big) = \min_{v \in [v_a(x), v_b(x)]} H^\Omega\big(x, \bar y(x), v, 1, p(x)\big) \tag{4.66}$$

for almost every $x \in \Omega$. In order to derive the corresponding *maximum* formulation, we multiply the Hamiltonian by -1 and introduce the negative adjoint state $q := -p$. In view of (4.64), q solves the adjoint equation

$$\begin{aligned} -\Delta q &= D_y H^\Omega(x, \bar y, \bar v, -1, q) && \text{in } \Omega \\ \partial_\nu q &= D_y H^\Gamma(x, \bar y, \bar u, -1, q) && \text{on } \Gamma. \end{aligned} \tag{4.67}$$

The minimum condition (4.66) then becomes the maximum condition

$$H^\Omega\big(x, \bar y(x), \bar v(x), -1, q(x)\big) = \max_{v \in [v_a(x), v_b(x)]} H^\Omega\big(x, \bar y(x), v, -1, q(x)\big) \tag{4.68}$$

for almost every $x \in \Omega$. An analogous result holds for the boundary control u.

Quite unexpectedly, the maximum condition (4.68) remains valid under natural assumptions without the convexity postulate used in the above argument. Then Pontryagin's maximum principle holds:

Definition. *The controls $\bar{u} \in U_{ad}$ and $\bar{v} \in V_{ad}$ obey* Pontryagin's maximum principle *if with $q_0 = -1$ and the adjoint state q defined by* (4.67) *the following* maximum conditions *hold for almost every $x \in \Omega$ and $x \in \Gamma$, respectively:*

(4.69)
$$\begin{aligned} H^{\Omega}\big(x, \bar{y}(x), \bar{v}(x), q_0, q(x)\big) &= \max_{v \in [v_a(x), v_b(x)]} H^{\Omega}\big(x, \bar{y}(x), v, q_0, q(x)\big), \\ H^{\Gamma}\big(x, \bar{y}(x), \bar{u}(x), q_0, q(x)\big) &= \max_{u \in [u_a(x), u_b(x)]} H^{\Gamma}\big(x, \bar{y}(x), u, q_0, q(x)\big). \end{aligned}$$

Global solutions to elliptic optimal control problems must obey Pontryagin's maximum principle if certain natural conditions are postulated. In comparison with the weak minimum conditions, the maximum principle has the advantage that no partial derivatives with respect to the controls are needed. If need be, with its help those among several solutions to (4.65) can be identified that are not minimizers. Also, functionals that cannot be differentiated with respect to the control can be handled. All the problems involving linear-quadratic controls in this chapter satisfy the requirements for the maximum principle to be valid. Therefore, in all of these cases the optimal controls must obey Pontryagin's maximum principle.

In the case of control problems involving partial differential equations, the maximum principle has so far been mostly of theoretical interest. Since numerical methods require, as a rule, derivatives with respect to the control, weak minimum conditions in the form of variational inequalities usually suffice.

4.9. Second-order derivatives

If F is a Fréchet differentiable mapping from an open set $\mathcal{U} \subset U$ into V, then $u \mapsto F'(u)$ defines an operator-valued mapping from $\mathcal{U}$ into $Z = \mathcal{L}(U, V)$. The question arises as to when this mapping is differentiable.

Definition. *Let $F : \mathcal{U} \subset U \to V$ be a Fréchet differentiable mapping. If the mapping $u \mapsto F'(u)$ is Fréchet differentiable at $u \in \mathcal{U}$, then F is said to be* twice Fréchet differentiable *at u. For the second derivative we write $F''(u) := (F')'(u)$.*

By definition, $F''(u)$ is a continuous linear operator from U into $Z = \mathcal{L}(U, V)$, that is, $F''(u) \in \mathcal{L}\big(U, \mathcal{L}(U, V)\big)$. This is a rather abstract mathematical object. Fortunately, we do not need to know the operator F'' itself, only how it is to be evaluated at given points. For any fixed direction $u_1 \in U$, the object $F''(u)u_1$ is already somewhat simpler: it is just a linear operator from $\mathcal{L}(U, V)$. Applying this linear operator to another direction $u_2 \in U$, we obtain an element $\big(F''(u)u_1\big)u_2$ of V. We will not have to deal with the

abstract quantity $F''(u)$, but with $F''(u)u_1$ or $\big(F''(u)u_1\big)u_2$ instead. We use the following notation:

$$F''(u)[u_1,u_2] := \big(F''(u)\,u_1\big)\,u_2, \qquad F''(u)\,v^2 := F''(u)[v,v].$$

With respect to u_1 and u_2, $F''(u)[u_1,u_2]$ is a symmetric and continuous bilinear form; see Cartan [**Car67**]. Taylor's theorem shows that for twice Fréchet differentiable mappings $F : U \to V$ we have the representation (cf. [**Car67**])

$$F(u+h) = F(u) + F'(u)\,h + \frac{1}{2}F''(u)\,h^2 + r_2^F(u,h),$$

where the second-order remainder r_2^F satisfies

$$\frac{\|r_2^F(u,h)\|_V}{\|h\|_U^2} \to 0 \quad \text{as } \|h\|_U \to 0.$$

We call the mapping F *twice continuously Fréchet differentiable* if the mapping $u \mapsto F''(u)$ is continuous, that is, if

$$\big\|F''(u) - F''(\bar u)\big\|_{\mathcal{L}(U,\mathcal{L}(U,V))} \to 0 \quad \text{whenever } \|u-\bar u\|_U \to 0.$$

Next, our interest turns to the question of how these norms can be calculated or at least estimated. For example, this can be done using the corresponding bilinear form. In fact, we have, by definition,

$$\begin{aligned}\big\|F''(u)\big\|_{\mathcal{L}(U,\mathcal{L}(U,V))} &= \sup_{\|u_1\|_U=1} \big\|F''(u)\,u_1\big\|_{\mathcal{L}(U,V)} \\ &= \sup_{\|u_1\|_U=1} \Big(\sup_{\|u_2\|_U=1} \big\|(F''(u)\,u_1)\,u_2\big\|_V \Big),\end{aligned}$$

and so, with the notation introduced above,

$$\big\|F''(u)\big\|_{\mathcal{L}(U,\mathcal{L}(U,V))} = \sup_{\|u_1\|_U=1,\,\|u_2\|_U=1} \big\|F''(u)[u_1,u_2]\big\|_V. \tag{4.70}$$

The equivalence between $F''(u)$ and the bilinear form $F''(u)[\cdot,\cdot]$ is discussed in, e.g., [**Car67**], [**KA64**], and [**Zei86**].

Calculation of $F''(u)$. In place of $F''(u)$, we work with the associated bilinear form. To this end, we first determine for fixed but arbitrary $u_1 \in U$ the directional derivative $F'(u)u_1$. Then we put $\tilde F(u) := F'(u)u_1$. Clearly, $\tilde F$ maps U into V. For its directional derivative $\tilde F'(u)u_2$ in the direction u_2,

we easily find that

$$\begin{aligned}\tilde{F}'(u)u_2 = \frac{d}{dt}\tilde{F}(u+t\,u_2)\big|_{t=0} &= \frac{d}{dt}(F'(u+t\,u_2)u_1)\big|_{t=0}\\ &= \big(F''(u+t\,u_2)u_1\big)u_2\big|_{t=0} = \big(F''(u)\,u_1\big)u_2\\ &= F''(u)[u_1,\,u_2].\end{aligned}$$

Example. We carry out this procedure (only formally at first) for the Nemytskii operator

$$\Phi(y) = \varphi\big(\cdot, y(\cdot)\big)$$

in the space $Y = L^\infty(E)$, tacitly assuming that the derivative Φ'' exists. This will be a consequence of Theorem 4.22 below, whose assumptions we suppose are satisfied. By Lemma 4.12 on page 202, we have

$$\big(\Phi'(y)y_1\big)(x) = \varphi_y\big(x, y(x)\big)\, y_1(x).$$

We now put

$$\tilde{\varphi}\big(x, y\big) := \varphi_y\big(x, y\big)\, y_1(x)$$

and define a new Nemytskii operator by

$$\tilde{\Phi}(y) = \tilde{\varphi}\big(\cdot, y(\cdot)\big).$$

Since $\tilde{\varphi}$ satisfies the assumptions of Lemma 4.12, $\tilde{\Phi}$ is differentiable. The directional derivative in the direction y_2 is given by

$$\big(\tilde{\Phi}'(y)y_2\big)(x) = \tilde{\varphi}_y\big(x, y(x)\big)\, y_2(x) = \varphi_{yy}\big(x, y(x)\big)\, y_1(x)\, y_2(x),$$

so, summarizing, we have

$$\big(\Phi''(y)[y_1, y_2]\big)(x) = \varphi_{yy}\big(x, y(x)\big)\, y_1(x)\, y_2(x). \qquad \diamond$$

As the subsequent theorem will show, we have indeed calculated a second-order Fréchet derivative. In view of the representation just derived, the abstract derivative $\Phi''(y)$ can be identified with a simple real-valued function, namely $\varphi_{yy}\big(x, y(x)\big)$. Moreover, the norms of the two quantities coincide, since

$$\begin{aligned}&\big\|\Phi''(y)\big\|_{\mathcal{L}(L^\infty(E),\mathcal{L}(L^\infty(E)))}\\ &\quad= \sup_{\|y_1\|_{L^\infty(E)}=\|y_2\|_{L^\infty(E)}=1} \big\|\Phi''(y)[y_1, y_2]\big\|_{L^\infty(E)}\\ &\quad= \sup_{\|y_1\|_{L^\infty(E)}=\|y_2\|_{L^\infty(E)}=1} \big\|\varphi_{yy}(\cdot, y)\, y_1\, y_2\big\|_{L^\infty(E)}\\ &\quad= \big\|\varphi_{yy}(\cdot, y)\big\|_{L^\infty(E)}.\end{aligned} \tag{4.71}$$

Theorem 4.22. *Suppose that the function $\varphi = \varphi(x,y) : E \times \mathbb{R} \to \mathbb{R}$ is measurable with respect to $x \in E$ for all $y \in \mathbb{R}$ and twice differentiable with respect to y for almost every $x \in E$. Let φ satisfy the boundedness and local Lipschitz conditions of order $k = 2$ from (4.24)–(4.25) on page 199. Then the Nemytskii operator Φ generated by φ is twice continuously Fréchet differentiable in $L^\infty(E)$, and the second derivative can be evaluated through*

$$\big(\Phi''(y)[y_1,y_2]\big)(x) = \varphi_{yy}\big(x,y(x)\big)\, y_1(x)\, y_2(x).$$

Proof: (i) We have to show that the representation of Φ'' asserted in the theorem in fact represents the first derivative of Φ', that is,

$$\frac{\|\Phi'(y+h) - \Phi'(y) - \Phi''(y)\,h\|_{\mathcal{L}(L^\infty(E))}}{\|h\|_{L^\infty(E)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Omega)} \to 0. \tag{4.72}$$

Moreover, it must be proved that $\Phi''(y)\,h$ belongs to $\mathcal{L}\big(L^\infty(E)\big)$ and that the linear mapping $h \mapsto \Phi''(y)\,h$ is bounded. Finally, we need to verify that the mapping $y \mapsto \Phi''(y)$ is in fact continuous. Right from the beginning we shall use the notation Φ'', even though this will be justified only later.

For the proof, let $y \in L^\infty(E)$ and $h \in L^\infty(E)$ be given. Evidently, the linear operator $A = \Phi''(y)\,h$ has to be defined as the multiplication operator $k \mapsto A\,k$,

$$\big(A\,k\big)(x) = \big(\varphi_{yy}\big(x,y(x)\big)h(x)\big)\; k(x).$$

The operator A corresponds to the given functions y and h, is generated by the bounded and measurable function $\varphi_{yy}\big(x,y(x)\big)\,h(x)$, and is obviously a continuous linear operator in $L^\infty(E)$. For a better overview, we list the correspondences between the abstract operators and the associated generating functions:

$$\begin{array}{lcll}
\Phi''(y) & \sim & \varphi_{yy}\big(x,y(x)\big) & \in \mathcal{L}\big(L^\infty(E), \mathcal{L}(L^\infty(E))\big) \\
\Phi''(y)\,h & \sim & \varphi_{yy}\big(x,y(x)\big)\,h(x) \quad \sim \quad A & \in \mathcal{L}\big(L^\infty(E)\big) \\
\Phi''(y)[h,k] & \sim & \varphi_{yy}\big(x,y(x)\big)\,h(x)\,k(x) \quad \sim \quad A\,k & \in L^\infty(E).
\end{array}$$

(ii) The mapping $\Phi''(y) : h \mapsto \Phi''(y)\,h =: A$ defines a continuous linear operator from $L^\infty(E)$ into $\mathcal{L}\big(L^\infty(E)\big)$: indeed, the linearity is evident, and the boundedness follows from (4.71), since $\varphi_{yy}(\cdot, y(\cdot))$ is bounded and measurable so that the last norm in (4.71) is finite.

(iii) Proof of (4.72): for the numerator we obviously have the estimate

$$\begin{aligned}&\big\|\Phi'(y+h)-\Phi'(y)-\Phi''(y)\,h\big\|_{\mathcal{L}(L^\infty(E))}\\&\quad=\sup_{\|k\|_{L^\infty(E)}=1}\big\|\big(\varphi_y(\cdot,y+h)-\varphi_y(\cdot,y)-\varphi_{yy}(\cdot,y)\,h\big)\,k\big\|_{L^\infty(E)}\\&\quad=\|\varphi_y(\cdot,y+h)-\varphi_y(\cdot,y)-\varphi_{yy}(\cdot,y)\,h\|_{L^\infty(E)}.\end{aligned}$$

The function $\tilde\varphi(x,y):=\varphi_y(x,y)$ generates a Fréchet differentiable Nemytskii operator whose derivative is generated by $\tilde\varphi_y(x,y)=\varphi_{yy}(x,y)$. Therefore,

$$\frac{\|\varphi_y(\cdot,y+h)-\varphi_y(\cdot,y)-\varphi_{yy}(\cdot,y)\,h\|_{L^\infty(E)}}{\|h\|_{L^\infty(E)}}\to 0\quad\text{as}\quad\|h\|_{L^\infty(E)}\to 0.$$

Dividing the above chain of inequalities by $\|h\|_{L^\infty(E)}$, we thus conclude that (4.72) is valid.

(iv) Continuity of the mapping $y\mapsto\Phi''(y)$: let $y_1,y_2\in L^\infty(\Omega)$ be given. As in the previous conclusions, we obtain

$$\begin{aligned}\big\|\Phi''(y_1)-\Phi''(y_2)\big\|_{\mathcal{L}(L^\infty(E),\mathcal{L}(L^\infty(E)))}&=\big\|\varphi_{yy}(\cdot,y_1)-\varphi_{yy}(\cdot,y_2)\big\|_{L^\infty(E)}\\&\le L(M)\,\|y_1-y_2\|_{L^\infty(E)},\end{aligned}$$

where $\max\{\|y_1\|_{L^\infty(E)},\|y_2\|_{L^\infty(E)}\}\le M$. This even proves local Lipschitz continuity. The theorem is thus proved. □

Second-order derivatives in other L^p spaces. The second derivative of $\Phi: L^p(E)\to L^q(E)$ exists for $1\le 2\,q<p<\infty$ if $y(\cdot)\mapsto\varphi_{yy}(\cdot,y(\cdot))$ maps $L^p(E)$ into $L^r(E)$, where

$$r=\frac{p\,q}{p-2\,q}\,; \tag{4.73}$$

see **[GKT92]**. In this case, we have

$$\big(\Phi''(y)[h_1,h_2]\big)(x)=\varphi_{yy}(x,y(x))\,h_1(x)\,h_2(x).$$

The relation (4.73) is evident, since for $h_1,h_2\in L^p(E)$ the product $h=h_1\,h_2$ belongs to $L^{\frac{p}{2}}(E)$; in light of formula (4.29) on page 204 for the first directional derivative in the direction h, the function φ_{yy} must map into $L^r(E)$, where

$$r=\frac{\frac{p}{2}\,q}{\frac{p}{2}-q}=\frac{p\,q}{p-2\,q}\,.$$

Example. We consider the sine operator $y(\cdot)\mapsto\sin(y(\cdot))$, which maps any space $L^p(E)$ into $L^q(E)$ for $1\le q\le p\le\infty$. The same holds for the

cosine operator $y(\cdot) \mapsto \cos(y(\cdot))$. Sufficient for the differentiability of the sine operator from $L^p(E)$ into $L^q(E)$ is the condition $1 \le q < p$; in fact, in this case $r = pq/(p-q) < \infty$, and it is evident that the cosine operator maps $L^p(E)$ into $L^r(E)$. Similar reasoning shows that for the existence of the second derivative one needs $q < p/2$. $\diamond$

4.10. Second-order optimality conditions

4.10.1. Basic ideas for sufficient optimality conditions. For the convex problems studied in Chapters 2 and 3, any control satisfying the first-order necessary optimality conditions is automatically globally optimal. Indeed, for convex problems the necessary optimality conditions are also sufficient. In the nonconvex case, derivatives of higher order have to be employed to guarantee local optimality.

To begin with, recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ has a local minimum at $\bar{u} \in \mathbb{R}^n$ if, in addition to the first-order necessary condition $f'(\bar{u}) = 0$, the Hessian matrix $f''(\bar{u})$ is positive definite; that is, there is some $\delta > 0$ such that

$$h^\top f''(\bar{u})\, h \ge \delta\, |h|^2 \qquad \forall\, h \in \mathbb{R}^n.$$

In infinite-dimensional spaces, the situation is quite similar, except that the theory is more challenging. We begin our analysis with a simple result that in many situations does not work in function spaces, since its assumptions cannot be satisfied in the spaces under investigation. In particular, the condition (4.74) postulated for *all* $h \in U$ is usually overly restrictive; it needs to be weakened.

Theorem 4.23. *Let U be a Banach space, let $C \subset U$ be convex, and suppose that the functional $f : U \to \mathbb{R}$ is twice continuously Fréchet differentiable in an open neighborhood of $\bar{u} \in C$. Let the control $\bar{u}$ satisfy the first-order necessary condition*

$$f'(\bar{u})(u - \bar{u}) \ge 0 \qquad \forall\, u \in C,$$

and assume there is some $\delta > 0$ such that

$$f''(\bar{u})[h,h] \ge \delta\, \|h\|_U^2 \quad \forall\, h \in U. \tag{4.74}$$

Then there are constants $\varepsilon > 0$ and $\sigma > 0$ such that we have the quadratic growth condition

$$f(u) \ge f(\bar{u}) + \sigma\, \|u - \bar{u}\|_U^2 \quad \text{for all } u \in C \text{ such that } \|u - \bar{u}\|_U \le \varepsilon.$$

In particular, f has a local minimum in C at $\bar{u}$.

Proof: The proof is the same as that in a finite-dimensional space, and we use the abbreviation $f''(\bar{u})\, h^2 := f''(\bar{u})[h,h]$. Consider the function $F : [0,1] \to \mathbb{R}$,

$F(s) := f(\bar u + s(u-\bar u))$. Then $f(u) = F(1)$ and $f(\bar u) = F(0)$, and the Taylor expansion

$$F(1) = F(0) + F'(0) + \frac{1}{2}F''(\theta), \quad \theta \in (0,1) \tag{4.75}$$

yields

$$\begin{aligned} f(u) &= f(\bar u) + f'(\bar u)(u-\bar u) + \frac{1}{2}\, f''\big(\bar u + \theta(u-\bar u)\big)(u-\bar u)^2 \\ &\geq f(\bar u) + \frac{1}{2}\, f''\big(\bar u + \theta(u-\bar u)\big)(u-\bar u)^2 \\ &= f(\bar u) + \frac{1}{2}\, f''(\bar u)(u-\bar u)^2 + \frac{1}{2}\, \big[f''\big(\bar u + \theta(u-\bar u)\big) - f''\big(\bar u\big)\big](u-\bar u)^2. \end{aligned}$$

Setting $h = u - \bar u$ in (4.74), we find that

$$f''(\bar u)(u-\bar u)^2 \geq \delta \, \|u-\bar u\|_U^2.$$

Moreover, the second derivative of f is continuous, and thus there is some $\varepsilon > 0$ such that the last summand in the above chain of inequalities can be bounded from above by $\delta \, \|u-\bar u\|_U^2/4$, provided that $\|u-\bar u\|_U \leq \varepsilon$.

In summary, we obtain for $\|u-\bar u\|_U \leq \varepsilon$ that

$$f(u) \geq f(\bar u) + \frac{\delta}{2}\, \|u-\bar u\|_U^2 - \frac{\delta}{4}\, \|u-\bar u\|_U^2 \geq f(\bar u) + \frac{\delta}{4}\, \|u-\bar u\|_U^2,$$

whence the assertion follows with the choice $\sigma := \delta/4$. □

This theorem will be applicable to optimal control problems involving semilinear partial differential equations if the control-to-state operator G is twice continuously differentiable as a mapping from L^2 into the state space and if the control appears only linearly or quadratically in the cost functional and only linearly in the differential equation. If G maps from L^2 into $C(\bar\Omega)$ or $C(\bar Q)$, then this simple situation can be expected to occur; see Sections 4.10.6 and 5.7.4.

Example. We consider the optimal control problem

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

subject to

$$\begin{aligned} -\Delta y + e^y &= u \qquad \text{in } \Omega \\ y_{|\Gamma} &= 0 \qquad \text{on } \Gamma \end{aligned}$$

and

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

We assume that $\dim \Omega = N \leq 3$ and that functions u_a, $u_b \in L^\infty(\Omega)$ and $y_\Omega \in L^2(\Omega)$ are prescribed. We have $2 > N/2 = 3/2$ and can use the fact that $\mathcal{A} = -\Delta$ is a coercive operator in $H_0^1(\Omega)$. Therefore, the assertion of Theorem 4.10 on page 194 remains valid for this elliptic problem with zero boundary condition of Dirichlet type (see, however, the counterexample on page 193 for Neumann boundary conditions). The mapping $G : u \mapsto y$ is therefore twice continuously Fréchet differentiable from $L^2(\Omega)$ into $C(\bar{\Omega})$. Hence, the reduced functional $f : L^2(\Omega) \to \mathbb{R}$,

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

is also twice continuously differentiable. As in (4.60) on page 221, we define the Lagrangian function $\mathcal{L} : (H_0^1(\Omega) \cap C(\bar{\Omega})) \times L^2(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ by

$$\mathcal{L}(y, u, p) = J(y, u) - \int_\Omega \big(\nabla y \cdot \nabla p + e^y - u\big)\, p \, dx.$$

Then $\mathcal{L}$ is also twice continuously differentiable. Now let $\bar{u}$ be a control that, together with the state $\bar{y} = G(\bar{u})$ and the adjoint state p, satisfies the first-order optimality conditions for the above problem.

In Theorem 4.25, we will prove the representation

$$f''(\bar{u})[h_1, h_2] = \mathcal{L}''(\bar{y}, \bar{u}, p)[(y_1, h_1), (y_2, h_2)], \tag{4.76}$$

where y_i, with $i = 1$ or 2, denotes the solution to the problem linearized at $\bar{y}$,

$$\begin{aligned} -\Delta y + e^{\bar{y}}\, y &= h \qquad \text{in } \Omega \\ y_{|\Gamma} &= 0 \qquad \text{on } \Gamma, \end{aligned} \tag{4.77}$$

with $h = h_i$. If, in addition, the triple $(\bar{y}, \bar{u}, p)$ satisfies with some $\delta > 0$ the *second-order sufficient condition*

$$\mathcal{L}''(\bar{y}, \bar{u}, p)(y, h)^2 \geq \delta \|h\|^2$$

for all $h \in L^2(\Omega)$ and associated solutions y to (4.77), then the preceding theorem and (4.76) yield the local optimality of $\bar{u}$ in the sense of $L^2(\Omega)$. $\diamond$

The condition (4.76) has been postulated for all $h \in L^2(\Omega)$, which is an overly restrictive requirement. In fact we know, for example, that $h(x) = u(x) - \bar{u}(x) \geq 0$ for almost every x such that $\bar{u}(x) = u_a(x)$ and, analogously, that $h(x) \leq 0$ for almost every x such that $\bar{u}(x) = u_b(x)$. Such sign conditions restrict the set of admissible directions h for which (4.76) must be postulated. The set of the directions h to be taken into account can be restricted even further by *strongly active restrictions*. The problem

$$\min_{u \in [-1,1]} -u^2,$$

which will be treated in Section 4.10.5, demonstrates this in a very simple way.

The choice of admissible directions h influences only the strength of the second-order sufficient condition and not its general applicability. More restrictive are regularity aspects of the corresponding partial differential equation, since they limit the applicability of the L^2 technique: in fact, while we could admit a three-dimensional domain in the above example, we would have to postulate $N \le 2$ in the case of a boundary control under otherwise unchanged assumptions. For comparable parabolic problems, only distributed controls with $N = 1$ can be handled, while boundary controls cannot be treated in L^2.

We will explain the reasons for this difficulty in the next section. Later, we will introduce a two-norm technique to overcome these problems.

4.10.2. The two-norm discrepancy. The following example demonstrates that a careless application of Theorem 4.23 may easily lead to erroneous conclusions in the infinite-dimensional case. Here, no differential equation is prescribed, but the control function u occurs non-quadratically in the cost functional.

(Counter)Example. We reconsider the problem (4.37) discussed earlier,

$$\min_{0\le u(x)\le 2\pi} f(u) := -\int_0^1 \cos\big(u(x)\big)\,dx.$$

Since, as a rule, the analysis is easiest in a Hilbert space setting, an obvious choice of control space is $U = L^2(0,1)$. Then we have $C = \{u \in U : 0 \le u(x) \le 2\pi \quad \text{for a.e. } x \in (0,1)\}$, and $\bar u \equiv 0$ is an obvious global minimizer. Let us check whether the optimality conditions are satisfied by $\bar u$. The first-order necessary condition,

$$f'(\bar u)(u-\bar u) = \int_0^1 \sin\big(\bar u(x)\big)\,\big(u(x)-\bar u(x)\big)\,dx = \int_0^1 \sin(0)\,u(x)\,dx = 0,$$

holds trivially. For the second-order sufficient condition, we find after a formal calculation that

$$f''(\bar u)\,u^2 = \int_0^1 \cos(0)\,u^2(x)\,dx = 1\cdot\int_0^1 u^2(x)\,dx = 1\cdot\|u\|^2_{L^2(0,1)}$$

for all $u \in L^2(0,1)$. Hence, the condition is satisfied with $\delta = 1$. Theorem 4.23 then implies the existence of some constant $\sigma > 0$ such that $f(u) \ge f(\bar u) + \sigma\,\|u-\bar u\|^2_{L^2(0,1)}$ for every $u \in C$ that is sufficiently close to $\bar u \equiv 0$ with respect to the L^2 norm.

This cannot be true. Indeed, for any arbitrarily small $\varepsilon > 0$ the function

$$u_\varepsilon(x) = \begin{cases} 2\pi & \text{if } 0 \le x \le \varepsilon \\ 0 & \text{if } \varepsilon < x \le 1 \end{cases}$$

also yields the global minimum value -1 and, therefore, is globally optimal. Moreover,

$$\begin{aligned} \|u_\varepsilon - \bar{u}\|_{L^2(0,1)} &= \Big(\int_0^1 |u_\varepsilon(x)|^2 \, dx \Big)^{1/2} = \Big(\int_0^\varepsilon (2\pi)^2 \, dx \Big)^{1/2} \\ &= 2\pi\sqrt{\varepsilon} \to 0 \quad \text{as } \varepsilon \downarrow 0. \end{aligned}$$

This contradicts the quadratic growth condition, which requires that, for sufficiently small $\varepsilon > 0$, u_ε must yield a larger value than $\bar{u}$.

What is wrong? The mistake is hidden in a nice trap into which we fell by tacitly assuming and taking for granted that the cosine functional f is twice continuously Fréchet differentiable in the space $L^2(0,1)$. Below we will demonstrate that this is not true at the point $u \equiv 0$, while the treatment of the general case will be the subject of Exercise 4.9. ◇

If the Banach space $L^\infty(0,1)$ is chosen for U in place of $L^2(0,1)$, then f is twice continuously Fréchet differentiable. Here, we have a typical example of the well-known *two-norm discrepancy*: in $L^2(0,1)$, f does not meet the requirement concerning differentiability, while $f''(\bar{u})$ is positive definite,

$$f''(\bar{u})\, u^2 \ge \delta \, \|u\|^2_{L^2(0,1)}.$$

On the other hand, the second derivative of f exists and is continuous in the space $L^\infty(0,1)$, but there cannot exist a $\delta > 0$ such that $f''(\bar{u})\, u^2 \ge \delta \, \|u\|^2_{L^\infty(0,1)}$ for every $u \in L^\infty(0,1)$.

Could this be a case in which no second-order sufficient condition can be satisfied? No. Fortunately, there is a way out that was discovered by Ioffe [**Iof79**]: one has to work with *two different* norms.

To this end, we examine the remainder of second order in the Taylor expansion of f in the space $L^\infty(0,1)$. Using the known series expansion of cosine at $u(x)$ and the integral form of the remainder for real-valued functions (see Heuser [**Heu08**], Eq. (168.6)), we obtain that

$$f(u+h) \;=\; -\int_0^1 \cos\big(u(x) + h(x)\big)\, dx$$

$$= \int_0^1 \Big[-\cos\big(u(x)\big) + \sin\big(u(x)\big)\, h(x) + \int_0^1 (1-s)\cos\big(u(x)+s\,h(x)\big)\, h^2(x)\, ds \Big]\, dx.$$

On the other hand, by the definition of Fréchet derivatives,

$$f(u+h) = \int_0^1 \Big[-\cos\big(u(x)\big) + \sin\big(u(x)\big)\, h(x) + \frac{1}{2}\cos\big(u(x)\big)\, h^2(x) \Big]\, dx + r_2^f(u,h),$$

where $r_2^f(u,h)$ denotes the second-order remainder of f at u in the direction h. Comparing the two representations for $f(u+h)$, we find that

(4.78)
$$r_2^f(u,h) = \int_0^1 \int_0^1 (1-s)\, \big[\cos\big(u(x)+s\,h(x)\big) - \cos\big(u(x)\big)\big]\, h^2(x)\, ds\, dx.$$

This representation is a special case of the general version of Taylor's theorem with integral remainder (cf. Cartan **[Car67]**, Thm. 5.6.1).

It is advantageous to use the integral form of the remainder instead of the simpler Lagrangian form: in this way, a discussion of the measurability with respect to x of the intermediate points $\theta = \theta(x)$, which arise for instance in the expansion

$$\cos(u(x)+h(x)) = \cos(u(x)) - \sin(u(x)+\theta(x)\,h(x))\, h(x),$$

can be avoided.

Example. We insert here the proof that f is not twice Fréchet differentiable with respect to $L^2(0,1)$ at $u \equiv 0$. To this end, we calculate $r_2^f(0,h)$ as on page 200 for

$$h(x) = \begin{cases} 1 & \text{in } [0,\varepsilon] \\ 0 & \text{in } (\varepsilon,1] \end{cases}$$

and take $\varepsilon \downarrow 0$. We get

$$r_2^f(0,h) = \int_0^\varepsilon \int_0^1 (1-s)\, \big(\cos(0+s) - \cos(0)\big) ds\, dx = \varepsilon \left(\frac{1}{2} - \cos(1)\right) = \varepsilon\, c$$

with $c \neq 0$, and thus

$$\frac{r_2^f(0,h)}{\|h\|^2_{L^2(0,1)}} = \varepsilon\, c\, \left\{ \int_0^\varepsilon 1^2\, dx \right\}^{-1} = c.$$

This expression does not tend to zero as $\varepsilon \downarrow 0$, while $\|h\|_{L^2(0,1)}$ does. Consequently, the cosine functional cannot be twice Fréchet differentiable at the point $u \equiv 0$. $\diamond$

In order to overcome the two-norm discrepancy, we estimate $r_2^f(u,h)$ as follows:

$$
\begin{aligned}
(4.79) \quad |r_2^f(u,h)| &\leq \int_0^1 \int_0^1 (1-s)\, s\, |h(x)|\, h^2(x)\, ds\, dx \\
&\leq \frac{1}{6}\, \|h\|_{L^\infty(0,1)} \int_0^1 h^2(x)\, dx \;\leq\; \frac{1}{6}\, \|h\|_{L^\infty(0,1)}\, \|h\|^2_{L^2(0,1)}.
\end{aligned}
$$

Conclusion. *The second-order remainder of the cosine functional satisfies*

$$
(4.80) \qquad \boxed{\frac{|r_2^f(u,h)|}{\|h\|^2_{L^2(0,1)}} \to 0 \qquad \textit{as} \qquad \|h\|_{L^\infty(0,1)} \to 0.}
$$

To verify this, we just have to divide (4.79) by $\|h\|^2_{L^2(0,1)}$. Notice that in the important estimate (4.80) two different norms occur, which is characteristic of the treatment of the two-norm discrepancy.

We are now in a position to conclude the cosine example. We find that at $\bar{u} \equiv 0$,

$$
\begin{aligned}
f(0+h) &= f(0) + f'(0)\, h + \frac{1}{2}\, f''(0)\, h^2 + r_2^f(0,h) \\
&= f(0) + 0 + \frac{1}{2}\, \|h\|^2_{L^2(0,1)} + r_2^f(0,h) \\
&= f(0) + \|h\|^2_{L^2(0,1)} \Big(\frac{1}{2} + \frac{r_2^f(0,h)}{\|h\|^2_{L^2(0,1)}} \Big) \\
&\geq f(0) + \|h\|^2_{L^2(0,1)} \Big(\frac{1}{2} - \frac{1}{6}\, \|h\|_{L^\infty(0,1)} \Big) \\
&\geq f(0) + \frac{1}{3}\, \|h\|^2_{L^2(0,1)},
\end{aligned}
$$

whenever $\|h\|_{L^\infty(0,1)} \leq \varepsilon = 1$. Thus, a quadratic growth condition with respect to the L^2 norm is satisfied in a sufficiently small L^∞ neighborhood of $\bar{u} \equiv 0$, whence we can conclude that $\bar{u}$ is a locally optimal solution in the sense of L^∞. Of course, this is nothing new, since we knew already that $\bar{u}$ is in fact even globally optimal.

In the following, the method just illustrated will be applied to optimal control problems, leading to results of the kind described above.

4.10.3. Distributed control. In this section, we investigate second-order conditions for the problem (4.31)–(4.33):

$$\min J(y,u) := \int_\Omega \varphi(x,y(x))\,dx + \int_\Omega \psi(x,u(x))\,dx,$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + d(x,y) & = & u & \text{in } \Omega \\ \partial_\nu y & = & 0 & \text{on } \Gamma \end{array}}$$

and

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

We assume that the control $\bar u \in U_{ad}$ satisfies, together with the associated state $\bar y = G(\bar u)$ and the adjoint state p, the first-order necessary optimality conditions (4.45)–(4.47) on pages 216 and 217. Then the pair $(\bar y, \bar u)$ need not be optimal, but is our candidate for which local optimality is to be shown. Since we work in the control space $L^\infty(\Omega)$, no restrictions on $\dim \Omega = N$ are needed.

It should be emphasized that the validity of the second-order sufficient optimality conditions can only be verified once $(\bar y, \bar u)$ is known. The situation is much the same as in the minimization of real-valued functions of several real variables: after the (usually numerical) determination of a solution to the optimality system, we have to check that the sufficient optimality condition is fulfilled, which then guarantees local optimality. Observe, however, that the numerical method will usually exhibit reasonable convergence behavior only if it starts in the vicinity of a critical point satisfying the sufficient conditions. One can only hope to find a starting point in such a neighborhood.

Remark. It is a difficult task to verify a second-order sufficient condition in a function space numerically. In fact, one has to use positive definiteness obtained for a finite-dimensional approximation to the infinite-dimensional case. There is, however, a numerical method due to Rösch and Wachsmuth [**RW08**] that can be applied to certain classes of problems.

Second-order derivatives.

The second derivative of the control-to-state mapping. As before, we denote by G the control-to-state mapping $u \mapsto y$ for the above elliptic boundary value problem. We consider G as a mapping between $L^\infty(\Omega)$ and $H^1(\Omega) \cap C(\bar\Omega)$. First, we show existence and continuity of the second-order Fréchet derivative of G.

Theorem 4.24. *Suppose that Assumption 4.14 on page 206 holds. Then the operator $G : L^\infty(\Omega) \to H^1(\Omega) \cap C(\bar\Omega)$ is twice continuously Fréchet differentiable. The second derivative $G''(u)$ is given by*

$$G''(u)[u_1, u_2] = z,$$

where z is the unique weak solution to the elliptic boundary value problem

$$\begin{aligned} -\Delta z + d_y(x,y)\, z &= -d_{yy}(x,y)\, y_1\, y_2 \\ \partial_\nu z &= 0, \end{aligned} \tag{4.81}$$

where $y = G(u)$ and $y_i = G'(u)\, u_i \in H^1(\Omega)$ for $i = 1, 2$.

Proof: (i) *Existence of the second derivative.*
By virtue of Theorem 4.17 on page 213, we know that G is Fréchet differentiable. To show the existence of the second derivative, we apply the implicit function theorem. To this end, we transform the elliptic boundary value problem for $y = G(u)$ into a suitable form. For this purpose, let $R : L^\infty(\Omega) \to H^1(\Omega) \cap C(\bar\Omega)$ denote the solution operator of the linear elliptic boundary value problem

$$\begin{aligned} -\Delta y + y &= v && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma. \end{aligned}$$

We regard R as an operator with range in $C(\bar\Omega)$. The equation $y = G(u)$ means that

$$\begin{aligned} -\Delta y + y &= u - d(x,y) + y && \text{in } \Omega \\ \partial_\nu y &= 0 && \text{on } \Gamma. \end{aligned} \tag{4.82}$$

In terms of R, this means that $y = R\big(u - d(\cdot, y) + y\big)$ or, more precisely,

$$y - R\big(u - \Phi(y)\big) = 0, \tag{4.83}$$

where $\Phi : C(\bar\Omega) \to L^\infty(\Omega)$ denotes the Nemytskii operator generated by $d(\cdot, y) - y$. Obviously, if $y \in C(\bar\Omega)$ solves (4.83), then y automatically lies in the range of R and hence belongs to $H^1(\Omega)$ and is a weak solution to (4.82). Therefore, (4.82) and (4.83) are equivalent. Next, we define the operator $F : C(\bar\Omega) \times L^\infty(\Omega) \to C(\bar\Omega)$,

$$F(y,u) = y - R\big(u - \Phi(y)\big).$$

Since Φ is, by Theorem 4.22 on page 229, twice continuously differentiable and R is linear and continuous, it follows from the chain rule that F is also twice continuously Fréchet differentiable.

Moreover, the derivative $D_y F(y,u)$ is surjective: in fact, the equality

$$D_y F(y,u)\, w = v$$

is equivalent to $w + R\,\Phi'(y)\,w = v$ and, upon putting $\tilde{w} := w - v$, to $\tilde{w} = -R\,\Phi'(y)(\tilde{w} + v)$. A straightforward calculation, using the definition of R, shows that the latter equation is equivalent to the boundary value problem

$$\begin{aligned} -\Delta\tilde{w} + d_y(x,y)\,\tilde{w} &= -d_y(x,y)\,v + v \\ \partial_\nu\,\tilde{w} &= 0, \end{aligned}$$

which for every $v \in C(\bar{\Omega})$ has a unique solution $\tilde{w} \in H^1(\Omega) \cap C(\bar{\Omega})$.

In summary, all assumptions of the implicit function theorem are satisfied and, therefore, the equation $F(y,u) = 0$ has a unique solution $y = G(u)$ for any u in a suitable open neighborhood of $\bar{u}$. This is nothing new, since we have shown already that a unique solution $y = G(u)$ exists even for all $u \in U_{ad}$. However, the implicit function theorem also yields that G inherits the smoothness properties of F (see, e.g., [**Car67**]); therefore, G is twice continuously Fréchet differentiable.

(ii) *Calculation of* $G''(u)$.
Taking $y = G(u)$ in the definition of F, we see that

$$F\big(G(u),u\big) = G(u) - R\,u + R\,\Phi\big(G(u)\big) = 0.$$

Differentiation in the direction u_1 yields, by the chain rule,

$$G'(u)\,u_1 - R\,u_1 + R\,\Phi'\big(G(u)\big)G'(u)\,u_1 = 0$$

or, upon defining the operator $K : L^\infty(\Omega) \to C(\bar{\Omega})$, $K(u) := G'(u)\,u_1$,

$$K(u) - R\,u_1 + R\,\Phi'\big(G(u)\big)\,K(u) = 0.$$

Next, we calculate the directional derivative in the direction u_2. Using the product and chain rules, we obtain that

$$K'(u)\,u_2 + R\,\Phi''\big(G(u)\big)\big[K(u), G'(u)\,u_2\big] + R\,\Phi'\big(G(u)\big)\,K'(u)\,u_2 = 0.$$

Since $K'(u)\,u_2 = G''(u)[u_1,u_2]$ (cf. the scheme for the evaluation of second-order derivatives explained on page 227), this is in turn equivalent to the equation

$$\begin{aligned} G''(u)[u_1,u_2] \;+\; R\,\Phi''\big(G(u)\big)\big[G'(u)\,u_1, G'(u)\,u_2\big] \\ + R\,\Phi'\big(G(u)\big)\,G''(u)[u_1,u_2] \;=\; 0. \end{aligned}$$

With $z := G''(u)[u_1,u_2]$, we thus have

$$z + R\,\big(\Phi'(y)\,z + \Phi''(y)[y_1,y_2]\big) = 0,$$

which by the definition of R means that z is the unique solution to the boundary value problem

$$\begin{aligned} -\Delta z + z &= -d_y(x,y)\, z - d_{yy}(x,y)\, y_1\, y_2 \\ \partial_\nu\, z &= 0. \end{aligned}$$

With this, the proof of the theorem is complete. □

The second derivative of the cost functional. Under Assumption 4.14 on page 206, the cost functional J is twice continuously Fréchet differentiable in $Y \times L^\infty(\Omega)$. By virtue of the chain rule, $f : u \mapsto J\big(G(u), u\big)$ is then also twice continuously Fréchet differentiable in $L^\infty(\Omega)$. The second derivative can be calculated as in the preceding proof. First, we obtain

$$f'(u)\, u_1 = D_y J\big(G(u), u\big) G'(u)\, u_1 + D_u J\big(G(u), u\big)\, u_1.$$

Next, we calculate the directional derivative of $\tilde f(u) := f'(u)\, u_1$ in the direction u_2. Invoking the product and chain rules, we find that

$$\begin{aligned} f''(u)[u_1, u_2] &= \tilde f'(u)\, u_2 \\ &= D_y^2 J\big(G(u), u\big)\big[G'(u)\, u_1, G'(u)\, u_2\big] \\ &\quad + D_u D_y J\big(G(u), u\big)\big[G'(u)\, u_1, u_2\big] + D_y J\big(G(u), u\big) G''(u)[u_1, u_2] \\ &\quad + D_y D_u J\big(G(u), u\big)\big[u_1, G'(u) u_2\big] + D_u^2 J\big(G(u), u\big)[u_1, u_2] \\ &= J''(y,u)\big[(y_1, u_1), (y_2, u_2)\big] + D_y J(y,u) G''(u)[u_1, u_2]. \end{aligned} \tag{4.84}$$

To simplify the terms, we again use the abbreviations $z := G''(u)[u_1, u_2]$ and $y_i := G'(u) u_i$, $i = 1, 2$. With this, we obtain the expression

$$D_y J(y,u)\, z = \int_\Omega \varphi_y\big(x, y(x)\big)\, z(x)\, dx,$$

which can be easily transformed using the adjoint state p. In fact, p is defined as the unique weak solution to the boundary value problem

$$\begin{aligned} -\Delta p + d_y(\cdot, y)\, p &= \varphi_y(\cdot, y) && \text{in } \Omega \\ \partial_\nu p &= 0 && \text{on } \Gamma. \end{aligned} \tag{4.85}$$

By virtue of Theorem 4.24, z solves the boundary value problem (4.81) whose right-hand side $\tilde u := -d_{yy}(\cdot, y)\, y_1\, y_2$ can be regarded as a "control". Specifying the involved quantities appropriately, in particular putting $a_\Omega = \varphi_y$, $\beta_\Omega = 1$, and $v = \tilde u$, we conclude from Lemma 2.31 on page 74 that

$$D_y J(y,u)\, z = \int_\Omega p\, \tilde u\, dx = -\int_\Omega p\, d_{yy}(x,y)\, y_1\, y_2\, dx. \tag{4.86}$$

Using this in (4.84) finally yields

$$f''(u)[u_1,u_2] = J''(y,u)\big[(y_1,u_1),(y_2,u_2)\big] - \int_\Omega p\, d_{yy}(x,y)\, y_1\, y_2\, dx. \tag{4.87}$$

This expression can be further simplified using the Lagrangian function.

Definition. *The Lagrangian function associated with the problem* (4.31)–(4.33) *on page 207 is defined by*

$$\mathcal{L}(y,u,p) = \int_\Omega \big(\,\varphi(x,y) + \psi(x,u) - (d(x,y)-u)\,p\,\big)\,dx - \int_\Omega \nabla y\cdot\nabla p\,dx.$$

We simplify the notation for the second derivative of $\mathcal{L}$ with respect to (y,u) by setting

$$\mathcal{L}''(y,u,p)[(y_1,u_1),(y_2,u_2)] := D^2_{(y,u)}\,\mathcal{L}(y,u,p)[(y_1,u_1),(y_2,u_2)].$$

Here, the increment (y_i,u_i) indicates that the derivative is to be understood with respect to (y,u). It follows from (4.87) that

(4.88)

$$\begin{aligned} f''(u)[u_1,u_2] &= \int_\Omega \big(\varphi_{yy}(x,y)\,y_1y_2 - p\,d_{yy}(x,y)\,y_1y_2 + \psi_{uu}(x,u)\,u_1u_2\big)\,dx \\ &= \mathcal{L}''(y,u,p)\big[(y_1,u_1),(y_2,u_2)\big]. \end{aligned}$$

Again, the Lagrangian function has proved to be a powerful tool for the calculation of derivatives with respect to u. Summarizing, we have shown the following result.

Theorem 4.25. *Suppose that Assumption 4.14 holds. Then the reduced functional* $f : L^\infty(\Omega) \to \mathbb{R}$,

$$f(u) = J(y,u) = J\big(G(u),u\big),$$

is twice continuously Fréchet differentiable. The second derivative of f *can be expressed in the form*

$$f''(u)[u_1,u_2] = \mathcal{L}''(y,u,p)\big[(y_1,u_1),(y_2,u_2)\big].$$

Here, y *is the state associated with* u, p *is the corresponding adjoint state, and* $y_i = G'(u)\,u_i$, $i = 1,2$, *denote the solutions to the linearized problems*

$$\begin{aligned} -\Delta y_i + d_y(x,y)\,y_i &= u_i \qquad \text{in } \Omega \\ \partial_\nu y_i &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

An auxiliary result. The treatment of the two-norm discrepancy requires estimates of a somewhat technical nature.

Lemma 4.26. *Suppose that Assumption 4.14 holds, and let the functional* $f : L^\infty(\Omega) \to \mathbb{R}$ *be given by*

$$f(u) = J(y,u) = J\big(G(u),u\big).$$

Then for each $M > 0$ *there exists a constant* $L(M) > 0$ *such that*

(4.89)
$$|f''(u+h)[u_1,u_2] - f''(u)[u_1,u_2]| \le L(M)\,\|h\|_{L^\infty(\Omega)}\,\|u_1\|_{L^2(\Omega)}\|u_2\|_{L^2(\Omega)}$$

for all u, h, u_1, $u_2 \in L^\infty(\Omega)$ *such that* $\max\big\{\|u\|_{L^\infty(\Omega)}, \|h\|_{L^\infty(\Omega)}\big\} \le M$.

Proof: (i) *Transformation using the Lagrangian.*

We put $y = G(u)$ and $y_h = G(u+h)$, and denote the corresponding adjoint states by p and p_h. Moreover, let $y_i = G'(u)\,u_i$ and $y_{i,h} = G'(u+h)\,u_i$ for $i = 1, 2$. The existence of the second derivative $f''(u)$ has been proved in Theorem 4.25. Invoking relation (4.88), we find that

$$\begin{aligned}
(4.90)\quad & f''(u+h)[u_1,u_2] - f''(u)[u_1,u_2] \\
&= \mathcal{L}''(y_h,u+h,p_h)\big[(y_{1,h},u_1),(y_{2,h},u_2)\big] \\
&\quad - \mathcal{L}''(y,u,p)\big[(y_1,u_1),(y_2,u_2)\big] \\
&= \int_\Omega \big(\varphi_{yy}(x,y_h)\,y_{1,h}\,y_{2,h} - \varphi_{yy}(x,y)\,y_1\,y_2\big)\,dx \\
&\quad - \int_\Omega p_h\,d_{yy}(x,y_h)\,y_{1,h}\,y_{2,h}\,dx + \int_\Omega p\,d_{yy}(x,y)\,y_1\,y_2\,dx \\
&\quad + \int_\Omega \big(\psi_{uu}(x,u+h) - \psi_{uu}(x,u)\big)\,u_1\,u_2\,dx.
\end{aligned}$$

(ii) *Estimation of* $y_{i,h} - y_i$ *and* $p_h - p$.

Owing to Theorem 4.16 on page 212, the operator G is Lipschitz continuous from $L^\infty(\Omega)$ into $C(\bar\Omega)$, that is,

$$(4.91)\qquad \|y_h - y\|_{C(\bar\Omega)} \le C_L\,\|h\|_{L^\infty(\Omega)}\,.$$

Hence, with a constant $c(M) > 0$ that depends on M,

$$(4.92)\qquad \|d_y(x,y_h) - d_y(x,y)\|_{L^\infty(\Omega)} \le c(M)\,\|h\|_{L^\infty(\Omega)}.$$

In the following, $c > 0$ will always denote a generic constant. The derivatives y_i and $y_{i,h}$ solve, respectively, the elliptic equations

$$\begin{aligned} -\Delta y_i + d_y(x,y)\, y_i &= u_i \\ -\Delta y_{i,h} + d_y(x,y_h)\, y_{i,h} &= u_i \end{aligned} \tag{4.93}$$

with Neumann boundary conditions. Since d is increasing with respect to y, we deduce from Theorem 4.7 on page 191 the existence of a constant $c > 0$, which is independent of h, such that

$$\|y_i\|_{H^1(\Omega)} \le c\,\|u_i\|_{L^2(\Omega)}, \qquad \|y_{i,h}\|_{H^1(\Omega)} \le c\,\|u_i\|_{L^2(\Omega)}. \tag{4.94}$$

The difference $y_{i,h} - y_i$ satisfies the equation

$$\begin{aligned} &-\Delta\,(y_{i,h} - y_i) + d_y(x,y)\,(y_{i,h} - y_i) \\ &\quad = -\big(d_y(x,y_h) - d_y(x,y)\big)\, y_{i,h}. \end{aligned} \tag{4.95}$$

The L^2 norm of the right-hand side can be estimated by means of (4.92) and (4.94):

$$\|(d_y(x,y_h) - d_y(x,y))\, y_{i,h}\|_{L^2(\Omega)} \le c\,\|h\|_{L^\infty(\Omega)}\|u_i\|_{L^2(\Omega)}. \tag{4.96}$$

Using the solution properties of the linear equation (4.95) and recalling that $d_y(x,y) \ge \lambda_d > 0$ on E_d, we find that

$$\|y_i - y_{i,h}\|_{H^1(\Omega)} \le c\,\|h\|_{L^\infty(\Omega)}\|u_i\|_{L^2(\Omega)}. \tag{4.97}$$

The difference of the adjoint states obeys the equation

$$\begin{aligned} &-\Delta\,(p_h - p) + d_y(x,y)\,(p_h - p) \\ &\quad = \varphi_y(\cdot,y_h) - \varphi_y(\cdot,y) + \big(d_y(x,y) - d_y(x,y_h)\big)\, p_h. \end{aligned}$$

Owing to (4.91), the $C(\bar\Omega)$ norms of y and y_h are uniformly bounded. This then also applies to the adjoint states p_h, because the boundedness of y_h is inherited by the right-hand side $\varphi_y(\cdot,y_h)$ of the adjoint equation (4.85). Hence, the right-hand side of the above equation can be estimated as follows:

$$\begin{aligned} &\|\varphi_y(\cdot,y_h) - \varphi_y(\cdot,y)\|_{L^\infty(\Omega)} + \|d_y(x,y) - d_y(x,y_h)\|_{L^\infty(\Omega)}\,\|p_h\|_{L^\infty(\Omega)} \\ &\qquad \le c\,\|y_h - y\|_{L^\infty(\Omega)} \le c\,\|h\|_{L^\infty(\Omega)}. \end{aligned}$$

Finally, from the difference of the adjoint equations for p and p_h we obtain that

$$\|p_h - p\|_{L^\infty(\Omega)} \le c\,\|h\|_{L^\infty(\Omega)}. \tag{4.98}$$

(iii) *Proof of the assertion.* The terms in (4.90) can now be estimated. For instance, we have

$$\begin{aligned}
&\|\varphi_{yy}(\cdot,y_h)\,y_{1,h}\,y_{2,h}-\varphi_{yy}(\cdot,y)\,y_1\,y_2\|_{L^1(\Omega)}\\
&\quad\le\|(\varphi_{yy}(\cdot,y_h)-\varphi_{yy}(\cdot,y))\,y_1y_2\|_{L^1(\Omega)}+\|\varphi_{yy}(\cdot,y_h)\,(y_{1,h}y_{2,h}-y_1y_2)\|_{L^1(\Omega)}\\
&\quad\le\|\varphi_{yy}(\cdot,y_h)-\varphi_{yy}(\cdot,y)\|_{L^\infty(\Omega)}\,\|y_1\|_{L^2(\Omega)}\,\|y_2\|_{L^2(\Omega)}\\
&\qquad+\|\varphi_{yy}(\cdot,y_h)\|_{L^\infty(\Omega)}\,\big(\|y_{1,h}(y_{2,h}-y_2)\|_{L^1(\Omega)}+\|(y_{1,h}-y_1)\,y_2\|_{L^1(\Omega)}\big)\\
&\quad\le c\,\|h\|_{L^\infty(\Omega)}\,\|u_1\|_{L^2(\Omega)}\,\|u_2\|_{L^2(\Omega)}+c\,\|y_{1,h}\|_{L^2(\Omega)}\,\|y_{2,h}-y_2\|_{L^2(\Omega)}\\
&\qquad+c\,\|y_2\|_{L^2(\Omega)}\,\|y_{1,h}-y_1\|_{L^2(\Omega)}\ \le\ c\,\|h\|_{L^\infty(\Omega)}\,\|u_1\|_{L^2(\Omega)}\,\|u_2\|_{L^2(\Omega)}.
\end{aligned}$$

The other integrals in (4.90) are estimated similarly, and (4.89) follows. This concludes the proof of the lemma. □

Second-order optimality conditions.

Second-order necessary conditions. For the derivation of necessary conditions, we follow **[CT99]**. The minimum principle (4.48) on page 217 yields for the solution to problem (4.31)–(4.33) on page 207 the representation

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

Hence, $\bar u$ is determined on the set of all x such that $|p(x)+\psi_u\big(x,\bar u(x)\big)|>0$, and higher-order conditions are only of interest on its complement. Now let

$$A_0(\bar u)=\big\{x\in\Omega:\ \big|p(x)+\psi_u\big(x,\bar u(x)\big)\big|>0\big\}. \tag{4.99}$$

For any $u\in U_{ad}$, $u(x)-\bar u(x)$ is nonpositive if $\bar u(x)=u_b(x)$ and nonnegative if $\bar u(x)=u_a(x)$. These facts motivate the following notation.

Definition (Critical cone). *The set $C_0(\bar u)$ is defined as the set of all $h\in L^\infty(\Omega)$ such that*

$$h(x)\begin{cases}=0 & \text{if}\quad x\in A_0(\bar u)\\ \ge 0 & \text{if}\quad x\notin A_0(\bar u)\ \text{and}\ \bar u(x)=u_a(x)\\ \le 0 & \text{if}\quad x\notin A_0(\bar u)\ \text{and}\ \bar u(x)=u_b(x).\end{cases}$$

Hence, h can be chosen freely on the inactive set $\{x\in\Omega:\ u_a(x)<\bar u(x)<u_b(x)\}$.

Theorem 4.27. *Suppose that Assumption 4.14 on page 206 holds, and let $\bar u$ be a locally optimal control for the problem* (4.31)–(4.33) *on page 207. Then*

$$f''(\bar u)\,h^2 \ge 0 \qquad \forall\, h \in C_0(\bar u). \tag{4.100}$$

Proof. Let $h \in C_0(\bar u)$ be arbitrary. Observe that even for very small $t > 0$ we cannot guarantee that $\bar u + t\,h$ belongs to U_{ad}. We therefore introduce the sets

$$I_n = \big\{x \in \Omega :\ u_a(x) + 1/n \le \bar u(x) \le u_b(x) - 1/n\big\}, \quad n \in \mathbb{N},$$

and consider the functions $h_n := \chi_n\, h$, where

$$\chi_n(x) = \begin{cases} 1 & \text{if} \quad x \in I_n \text{ or } \big[\bar u(x) \in \{u_a(x), u_b(x)\} \\ & \qquad\qquad \text{and } u_b(x) - u_a(x) \ge 1/n\big] \\ 0 & \text{if} \quad \bar u(x) \in \big(u_a(x), u_a(x) + 1/n\big) \cup \big(u_b(x) - 1/n, u_b(x)\big). \end{cases}$$

Hence, $\chi_n(x) = 0$ also if $\bar u(x) \in \big\{u_a(x), u_b(x)\big\}$ and $u_b(x) - u_a(x) < 1/n$, and we have $u = \bar u + t\,h_n \in U_{ad}$ for sufficiently small $t > 0$. Therefore,

$$\begin{aligned} 0 \ &\le\ J(y,u) - J(\bar y, \bar u) \ =\ f(u) - f(\bar u) \\ &=\ f'(\bar u)\,t\,h_n + \frac{1}{2}\,f''(\bar u)\,t^2\,h_n^2 + r_2^f(\bar u, t h_n), \end{aligned}$$

with the second-order remainder r_2^f of f. By virtue of the minimum condition (4.48) on page 217, h and also h_n vanish at almost all points where $p + \psi_u(\cdot, \bar u)$ differs from zero. Therefore, $f'(\bar u)\,h_n = \big(p + \psi_u(\cdot, \bar u)\,,\, h_n\big)_{L^2(\Omega)} = 0$, and we obtain after division by t^2 that

$$0 \le \frac{1}{2} f''(\bar u)\,h_n^2 + t^{-2}\,r_2^f(\bar u, t h_n). \tag{4.101}$$

Passage to the limit as $t \downarrow 0$ shows that $f''(\bar u)\,h_n^2 \ge 0$. Now observe that $h_n(x) \to h(x)$ pointwise almost everywhere as $n \to \infty$, and $h_n(x)^2 \le h(x)^2$ pointwise everywhere for all $n \in \mathbb{N}$. We may therefore employ Lebesgue's dominated convergence theorem to deduce that $h_n \to h$ in $L^2(\Omega)$. Hence, passing to the limit as $n \to \infty$, we can conclude the validity of the assertion from the continuity of the quadratic form $f''(\bar u)\,h^2$ in $L^2(\Omega)$. □

The second-order condition (4.100) has been formulated in terms of the abstract operator f'', i.e., not with an explicitly given function. From Theorem 4.25, we obtain the following, in the theory of optimization more popular, form.

Lemma 4.28. *The second-order necessary condition* (4.100) *is equivalent to*

$$\mathcal{L}''(\bar{y},\bar{u},p)(y,h)^2 \geq 0 \qquad \forall\, h \in C_0(\bar{u}),$$

where $y = y(h) \in H^1(\Omega)$ *is the solution to the associated linearized problem*

$$\begin{aligned} -\Delta y + d_y(x,\bar{y})\, y &= h \\ \partial_\nu y &= 0. \end{aligned}$$

Second-order sufficient conditions. For the formulation of the sufficient conditions, we introduce the following cone:

$$\begin{aligned} C(\bar{u}) = \big\{ u \in L^\infty(\Omega) :\ & u(x) \geq 0 \ \text{ if } \ \bar{u}(x) = u_a(x) \\ & \text{and} \quad u(x) \leq 0 \ \text{ if } \ \bar{u}(x) = u_b(x) \big\}. \end{aligned} \tag{4.102}$$

The following condition is an example of a *second-order sufficient condition.*

There exists some $\delta > 0$ *such that*

$$f''(\bar{u})\, u^2 \geq \delta \, \|u\|^2_{L^2(\Omega)} \quad \forall\, u \in C(\bar{u}). \tag{4.103}$$

By (4.88) on page 242, this is equivalent to the condition

(4.104)

$$\int_\Omega \Big\{ \big(\varphi_{yy}(x,\bar{y}) - p\, d_{yy}(x,\bar{y})\big)\, y^2 + \psi_{uu}(x,\bar{u})\, u^2 \Big\}\, dx \geq \delta \, \|u\|^2_{L^2(\Omega)}$$

for all $u \in C(\bar{u})$ *and* $y \in H^1(\Omega)$ *such that*

$$\begin{aligned} -\Delta y + d_y(x,\bar{y})\, y &= u \\ \partial_\nu y &= 0. \end{aligned}$$

Remark. The above sufficient condition is too restrictive. In fact, in comparison with $C_0(\bar{u})$, the cone $C(\bar{u})$ is too large. The gap between $C(\bar{u})$ and $C_0(\bar{u})$ can be narrowed by invoking strongly active constraints; see Section 4.10.5. However, the above form is frequently used, in particular, as a usual assumption in the convergence analysis of numerical methods.

If the pair $(\bar{y},\bar{u})$ satisfies both the first-order necessary conditions and the second-order sufficient condition (4.104), then $\bar{u}$ is a locally optimal control in the sense of $L^\infty(\Omega)$, as the following result shows.

Theorem 4.29. *Suppose that Assumption 4.14 on page 206 holds. Let the control* $\bar{u} \in U_{ad}$*, together with the associated state* $\bar{y} = G(\bar{u})$ *and the adjoint state* p*, satisfy the first-order necessary optimality conditions stated in Theorem 4.20 on page 216. If, in addition,* $(\bar{y},\bar{u})$ *satisfies the second-order*

sufficient condition (4.104)*, then there exist constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that we have the quadratic growth condition*

$$J(y,u) \geq J(\bar{y},\bar{u}) + \sigma \|u-\bar{u}\|^2_{L^2(\Omega)} \quad \forall\, u \in U_{ad} \ \text{ with } \ \|u-\bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon,$$

where $y = G(u)$*. In particular,* $\bar{u}$ *is a locally optimal control in the sense of* $L^\infty(\Omega)$*.*

Proof: The proof is almost identical to that for the cosine functional, so we will be brief. We have

$$J(y,u) = f(u) = f(\bar{u}) + f'(\bar{u})(u-\bar{u}) + \frac{1}{2} f''\big(\bar{u}+\theta(u-\bar{u})\big)\,(u-\bar{u})^2$$

with $\theta \in (0,1)$. In view of the first-order necessary condition, the first-order term is nonnegative. Indeed, it follows from the variational inequality in Theorem 4.20 (see also formula (4.43)) that

$$f'(\bar{u})(u-\bar{u}) = \int_\Omega \big(p + \psi_u(\cdot,\bar{u})\big)(u-\bar{u})\,dx \geq 0.$$

Next, note that $u - \bar{u} \in C(\bar{u})$. We estimate the second-order term from below. We obtain

$$\begin{aligned}
&f''\big(\bar{u}+\theta(u-\bar{u})\big)\,(u-\bar{u})^2 \\
&\quad = f''(\bar{u})\,(u-\bar{u})^2 + \big[f''\big(\bar{u}+\theta\,(u-\bar{u})\big) - f''(\bar{u})\big]\,(u-\bar{u})^2 \\
&\quad \geq \delta\,\|u-\bar{u}\|^2_{L^2(\Omega)} - L\,\|u-\bar{u}\|_{L^\infty(\Omega)}\,\|u-\bar{u}\|^2_{L^2(\Omega)} \\
&\quad \geq \frac{\delta}{2}\|u-\bar{u}\|^2_{L^2(\Omega)},
\end{aligned}$$

provided that $\|u-\bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$ for some sufficiently small $\varepsilon > 0$. Here, we have used (4.103) and the estimate (4.89) on page 243, as well as the fact that all $u \in U_{ad}$ are bounded in $L^\infty(\Omega)$ by a common constant $M > 0$. In summary, we have

$$J(y,u) \geq f(\bar{u}) + \frac{\delta}{4}\,\|u-\bar{u}\|^2_{L^2(\Omega)} = J(\bar{y},\bar{u}) + \sigma\,\|u-\bar{u}\|^2_{L^2(\Omega)}$$

with $\sigma = \delta/4$, provided that $\|u-\bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$ for a sufficiently small $\varepsilon > 0$. □

In analogy to Lemma 4.28, we rewrite (4.104) in the following more commonly used form:

Lemma 4.30. *The second-order sufficient condition* (4.104) *is equivalent to*

$$\mathcal{L}''(\bar{y},\bar{u},p)(y,u)^2 \geq \delta\,\|u\|^2_{L^2(\Omega)}$$

for every $u \in C(\bar{u})$ *and every* $y \in H^1(\Omega)$ *such that*

$$\begin{aligned} -\Delta y + d_y(x,\bar{y})\, y &= u \qquad \text{in } \Omega \\ \partial_\nu y &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

4.10.4. Boundary control. In order to elucidate the second-order optimality conditions, we exemplarily investigate the boundary control problem (4.49)–(4.51):

$$\min J(y,u) := \int_\Omega \varphi\big(x, y(x)\big)\, dx + \int_\Gamma \psi\big(x, u(x)\big)\, ds(x),$$

subject to

$$\boxed{\begin{aligned} -\Delta y &= 0 \qquad \text{in } \Omega \\ \partial_\nu y + b(x,y) &= u \qquad \text{on } \Gamma \end{aligned}}$$

and

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

To this end, we use the associated Lagrangian function

$$\mathcal{L}(y,u,p) = \int_\Omega \big(\, \varphi(x,y) - \nabla y \cdot \nabla p \big)\, dx + \int_\Gamma \big(\psi(x,u) - (b(x,y) - u)\, p\,\big)\, ds.$$

The second-order sufficient condition postulates the existence of some $\delta > 0$ such that

$$\text{(4.105)} \qquad \left\{ \begin{array}{l} \mathcal{L}''(\bar{y},\bar{u},p)(y,u)^2 \ge \delta\, \|u\|^2_{L^2(\Gamma)} \\ \text{for all } u \in C(\bar{u}) \text{ and all } y \in H^1(\Omega) \text{ such that} \\ \qquad\qquad -\Delta y = 0 \\ \qquad \partial_\nu y + b_y(x,\bar{y})\, y = u. \end{array} \right.$$

Here, the cone $C(\bar{u})$ is defined as in (4.102), except that Ω has to be replaced by Γ. As an explicit expression for $\mathcal{L}''$, we obtain

$$\begin{aligned} \mathcal{L}''(\bar{y},\bar{u},p)(y,u)^2 &= \int_\Omega \varphi_{yy}(x,\bar{y})\, y^2\, dx - \int_\Gamma b_{yy}(x,\bar{y})\, p\, y^2\, ds \\ &\quad + \int_\Gamma \psi_{uu}(x,\bar{u})\, u^2\, ds. \end{aligned}$$

Theorem 4.31. *Suppose that Assumption 4.14 holds, and suppose that the control* $\bar{u} \in U_{ad}$*, together with the associated state* $\bar{y} = G(\bar{u})$ *and the adjoint state* p*, satisfies the first-order necessary optimality condition stated in Theorem 4.21 on page 219. If, in addition, the pair* $(\bar{y},\bar{u})$ *satisfies the*

second-order sufficient condition (4.105)*, then there are constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that we have the quadratic growth condition*

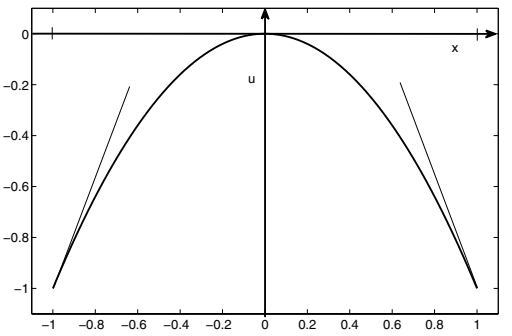
$$J(y,u) \geq J(\bar{y},\bar{u}) + \sigma \, \|u-\bar{u}\|^2_{L^2(\Gamma)}$$

for every $u \in U_{ad}$ *such that* $\|u-\bar{u}\|_{L^\infty(\Gamma)} \leq \varepsilon$*, where* $y = G(u)$*. In particular,* $\bar{u}$ *is locally optimal in the sense of* $L^\infty(\Gamma)$*.*

The proof is completely analogous to that for the case of distributed controls.

4.10.5. Inclusion of strongly active constraints *. In the previously established results, the gap between necessary and sufficient second-order conditions is too large. In fact, in the necessary condition the nonnegativity of $\mathcal{L}''$ is postulated on the critical cone $C_0(\bar{u})$, which is usually smaller than the cone $C(\bar{u})$ appearing in the second-order sufficient optimality condition. In $C_0(\bar{u})$ the controls vanish on the strongly active set, while in $C(\bar{u})$ they only have to obey sign restrictions. Therefore, the second-order sufficient condition is actually an overly restrictive requirement. By including strongly active constraints, this gap can be closed to a certain extent. To this end, one additionally considers *first-order sufficient conditions.*

Example. The nonconvex function $f : \mathbb{R} \to \mathbb{R}$, $f(u) = -u^2$, has two minimizers $u_1 = -1$ and $u_2 = 1$ in $U_{ad} = [-1,1]$. At both points, the first-order necessary condition $f'(u_i)(u-u_i) \geq 0$ for all $u \in [-1,1]$ holds, and we even have $|f'(u_i)| = 2 > 0$; see the figure.



First-order sufficient conditions.

Second-order conditions of the previously used type cannot be valid here, since f is concave. However, they are not needed, because the u_i satisfy the *first-order sufficient optimality conditions*, given by $|f'(u_i)| \neq 0$. This already implies local optimality: for instance, for u_1 we have, for any $h \in (0,2)$,

$$f(-1+h) = f(-1) + f'(-1)\,h + r(h) = f(-1) + 2h - h^2 > -1 = f(-1).$$

Hence, $u_1 = -1$ is locally (and here even globally) optimal. ◇

Analogous constructions can be applied in function spaces to weaken second-order sufficient optimality conditions.

A simplified example in function space. Suppose that the function φ satisfies the conditions stated in Assumption 4.14 on page 206. We consider the minimization problem

$$\min_{u\in U_{ad}} f(u) := \int_\Omega \varphi\big(x, u(x)\big)\, dx,$$

where $U_{ad} = \big\{u \in L^\infty(\Omega) \,:\, u_a(x) \le u(x) \le u_b(x) \ \text{ for a.e. } x \in \Omega\big\}$. Let the function $\bar u \in U_{ad}$ obey the first-order necessary condition

$$\int_\Omega \varphi_u\big(x, \bar u(x)\big)\big(u(x) - \bar u(x)\big)\, dx \ge 0 \qquad \forall\, u \in U_{ad}.$$

Then, almost everywhere in Ω, we must have

$$\varphi_u\big(x, \bar u(x)\big) \begin{cases} \ge 0 & \text{if} \quad \bar u(x) = u_a(x) \\ \le 0 & \text{if} \quad \bar u(x) = u_b(x). \end{cases}$$

Definition. *For arbitrary but fixed $\tau \ge 0$, let*

$$A_\tau(\bar u) = \big\{x \in \Omega \,:\quad |\varphi_u(x, \bar u(x))| > \tau\big\}.$$

Then $A_\tau(\bar u)$ is said to be the set of strongly active constraints *or, for short, the* strongly active set*.*

In the special case of $\tau = 0$, we obtain the set $A_0(\bar u)$ defined in (4.99). The above definition is due to Dontchev et al. [**DHPY95**]. In the example under discussion, we suppose the following sufficient second-order optimality condition to be valid: there exist constants $\delta > 0$ and $\tau > 0$ such that

$$f''(\bar u)\, h^2 \ge \delta\, \|h\|^2_{L^2(\Omega)}$$

for all $h \in L^\infty(\Omega)$ such that

$$h(x) \begin{cases} = 0 & \text{if } x \in A_\tau \\ \ge 0 & \text{if } x \notin A_\tau \text{ and } \bar u(x) = u_a(x) \\ \le 0 & \text{if } x \notin A_\tau \text{ and } \bar u(x) = u_b(x). \end{cases}$$

We therefore postulate that $\varphi_{uu}\big(x, \bar u(x)\big) \ge \delta$ on $\Omega\backslash A_\tau$ and $|\varphi_u\big(x, \bar u(x)\big)| > \tau$ on A_τ, and hence the positive definiteness of $f''(\bar u)$ is assumed only for a proper subset of the set $C(\bar u)$ defined in (4.102) on page 247. We claim that this postulate, in combination with the first-order necessary condition, is sufficient for the local optimality of $\bar u$. This can be seen as follows.

For $u \in U_{ad}$ sufficiently close to $\bar u$, say, for $\|u - \bar u\|_{L^\infty(\Omega)} \le \varepsilon$, we have the Taylor expansion

$$f(\bar u + h) - f(\bar u) = f'(\bar u)\, h + \frac{1}{2}\, f''(\bar u)\, h^2 + \frac{1}{2}\, \big(f''(\bar u + \theta\, h) - f''(\bar u)\big)\, h^2,$$

with $h = u - \bar{u}$ and a suitable $\theta = \theta(x) \in (0,1)$.

We split h into two parts, $h = h_1 + h_2$, in such a way that $h_2(x) = 0$ on A_τ and $h_1(x) = 0$ on $\Omega \backslash A_\tau$. The function h_1 exploits the first-order sufficient conditions, while for h_2 the positive definiteness of f'' applies. Obviously, $h_1(x) \geq 0$ whenever $\bar{u}(x) = u_a$, and $h_1(x) \leq 0$ whenever $\bar{u}(x) = u_b$. With a remainder $r(u,h)$ of second order, we obtain

$$\begin{aligned} f(\bar{u}+h) - f(\bar{u}) &= \int_\Omega \varphi_u\big(x, \bar{u}(x)\big)\, h(x)\, dx \\ &\quad + \frac{1}{2} \int_\Omega \varphi_{uu}\big(x, \bar{u}(x)\big)\, h^2(x)\, dx + r(u,h). \end{aligned}$$

Now note that $\varphi_u\big(x, \bar{u}(x)\big)\, h(x) \geq 0$ for almost every $x \in \Omega$. Hence, invoking the fact that $h_1(x) h_2(x) = 0$, we obtain, with the abbreviations $\varphi_u(x) := \varphi_u\big(x, \bar{u}(x)\big)$ and $\varphi_{uu}(x) := \varphi_{uu}\big(x, \bar{u}(x)\big)$, that

$$\begin{aligned} & f(\bar{u}+h) - f(\bar{u}) \\ &\geq \int_{A_\tau} \varphi_u(x)\, h_1(x)\, dx + \frac{1}{2} \int_\Omega \varphi_{uu}(x) \big(h_1(x) + h_2(x)\big)^2\, dx \,+\, r(u,h) \\ &= \int_{A_\tau} \Big(|\varphi_u(x)\, h_1(x)| + \frac{1}{2}\, \varphi_{uu}(x)\, h_1(x)^2 \Big)\, dx \\ &\quad + \frac{1}{2} \int_{\Omega \backslash A_\tau} \varphi_{uu}(x)\, h_2(x)^2\, dx \,+\, r(u,h) \\ &\geq \int_{A_\tau} \Big(\tau\, |h_1(x)| - \frac{1}{2}\, \|\varphi_{uu}\|_{L^\infty(\Omega)}\, |h_1(x)|^2 \Big)\, dx \\ &\quad + \frac{\delta}{2} \int_{\Omega \backslash A_\tau} h_2(x)^2\, dx \,+\, r(u,h). \end{aligned}$$

For sufficiently small $\varepsilon \in (0,1)$, we have, for almost every $x \in \Omega$,

$$\|\varphi_{uu}\|_{L^\infty(\Omega)}\, |h_1(x)| \leq \|\varphi_{uu}\|_{L^\infty(\Omega)}\, \varepsilon \leq \tau \quad \text{and} \quad |h_1(x)| \geq h_1(x)^2 .$$

Hence, for sufficiently small $\varepsilon > 0$ it follows that

$$\begin{aligned} f(\bar{u}+h) - f(\bar{u}) &\geq \frac{\tau}{2} \int_\Omega |h_1(x)|\, dx + \frac{\delta}{2} \int_\Omega h_2(x)^2\, dx \,+\, r(u,h) \\ &\geq \min\Big\{ \frac{\tau}{2}, \frac{\delta}{2} \Big\} \int_\Omega \big(h_1(x)^2 + h_2(x)^2 \big)\, dx \,+\, r(u,h) \\ &= \frac{1}{2}\, \min\{\tau, \delta\} \int_\Omega h^2(x)\, dx \,+\, r(u,h) \\ &\geq \|h\|^2_{L^2(\Omega)} \left(\frac{1}{2}\, \min\{\tau, \delta\} - \frac{|r(u,h)|}{\|h\|^2_{L^2(\Omega)}} \right). \end{aligned}$$

As in (4.80) on page 237, we have

$$\frac{|r(u,h)|}{\|h\|^2_{L^2(\Omega)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Omega)} \to 0,$$

whence we conclude that

$$f(\bar u + h) - f(\bar u) \geq \sigma \|h\|^2_{L^2(\Omega)}$$

with $\sigma = \frac{1}{4}\min\{\tau, \delta\}$, provided that $\varepsilon > 0$ is sufficiently small and $\|\bar u - u\|_{L^\infty(\Omega)} \leq \varepsilon$. In other words, $\bar u$ is locally optimal, which proves our claim.

Strongly active constraints in elliptic optimal control problems. Obviously, the above method can also be employed to weaken the sufficient optimality conditions in control problems involving partial differential equations. In this connection, we refer the reader to [**CTU96**] for elliptic problems with control constraints, to [**CM02a**] for the case where additional constraints in integral form are imposed, and to [**CDlRT08**] for pointwise state constraints.

We discuss the use of strongly active constraints for the distributed control problem (4.31)–(4.33):

$$\min J(y,u) := \int_\Omega \varphi\big(x, y(x)\big)\,dx + \int_\Omega \psi\big(x, u(x)\big)\,dx,$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + d(x,y) & = & u & \text{in } \Omega \\ \partial_\nu y & = & 0 & \text{on } \Gamma \end{array}}$$

and

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

We assume that a control $\bar u \in U_{ad}$ with associated state $\bar y$ is given that satisfies the first-order necessary optimality condition

$$\int_\Omega \big(p(x) + \psi_u(x, \bar u(x)\big)\big(u(x) - \bar u(x)\big)\,dx \geq 0 \quad \forall\, u \in U_{ad}. \tag{4.106}$$

We define for fixed $\tau \geq 0$ the *strongly active set*

$$A_\tau(\bar u) = \big\{x \in \Omega : \big|p(x) + \psi_u\big(x, \bar u(x)\big)\big| > \tau\big\}$$

and the *τ-critical cone* $C_\tau(\bar u) = \big\{u \in L^\infty(\Omega) : u \text{ satisfies (4.107)}\big\}$, where

$$u(x)\begin{cases} = 0 & \text{if} \quad x \in A_\tau(\bar u) \\ \geq 0 & \text{if} \quad x \notin A_\tau(\bar u) \ \text{ and } \ \bar u(x) = u_a \\ \leq 0 & \text{if} \quad x \notin A_\tau(\bar u) \ \text{ and } \ \bar u(x) = u_b. \end{cases} \tag{4.107}$$

The following condition is called the *second-order sufficient condition*:

(4.108)
$$\int_\Omega \left\{ \big(\varphi_{yy}(x,\bar y) - p\, d_{yy}(x,\bar y)\big)\, y^2 + \psi_{uu}(x,\bar u)\, u^2 \right\} dx \ge \delta\, \|u\|^2_{L^2(\Omega)}$$
for all $u \in C_\tau(\bar u)$ *and* $y \in H^1(\Omega)$ *such that*
$$\begin{aligned} -\Delta y + d_y(x,\bar y)\, y &= u \\ \partial_\nu y &= 0. \end{aligned}$$

Theorem 4.32. *Suppose that Assumption 4.14 holds, and let the control* $\bar u \in U_{ad}$*, together with the associated state* $\bar y$ *and the adjoint state* p*, satisfy the first-order necessary optimality condition stated in Theorem 4.20 on page 216. If, in addition, the pair* $(\bar y, \bar u)$ *obeys for some* $\tau > 0$ *the second-order sufficient condition* (4.108)*, then there exist constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that we have the quadratic growth condition*

$$J(y,u) \ge J(\bar y, \bar u) + \sigma \|u - \bar u\|^2_{L^2(\Omega)} \quad \forall\, u \in U_{ad} \quad \text{with} \ \ \|u - \bar u\|_{L^\infty(\Omega)} \le \varepsilon,$$

where $y = G(u)$*. In particular,* $\bar u$ *is locally optimal in the sense of* $L^\infty(\Omega)$*.*

A proof of this result can be found in [**CTU96**]. However, one can also argue as in the proof of Theorem 5.17 on page 292 for the parabolic case; therefore, we do not give the proof here.

The issue of the gap between necessary and sufficient second-order conditions is also the subject of the monograph by Bonnans and Shapiro [**BS00**], where various other results concerning the use of second-order derivatives in optimization theory can be found. We also refer the reader to Casas and Mateos [**CM02a**] and, in connection with state constraints, to [**CT02**] and [**CDlRT08**].

4.10.6. Cases without two-norm discrepancy. So far, we have used two different norms for the derivation of sufficient second-order optimality conditions, namely, the L^∞ norm for the differentiation and the L^2 norm for the positive definiteness of $f''(\bar u)$. This is not always necessary. Indeed, the two-norm discrepancy does not play any role if the following three conditions are met: the equation is only linear in the control, the control-to-state mapping G maps L^2 continuously into $C(\bar\Omega)$, and the cost functional is linear-quadratic with respect to u in a sense that is yet to be made precise. An example of this type has been studied on page 232. In the cases below, we will be able to work with the L^2 norm alone.

Distributed control. Let Ω be a bounded Lipschitz domain with $\dim \Omega = N \leq 3$, and suppose Assumption 4.14 holds. Then, by Theorem 4.16 on page 212, the control-to-state operator $G : u \mapsto y$ of the distributed control problem (4.31)–(4.33) on page 207 maps $L^2(\Omega)$ Lipschitz continuously into $C(\bar{\Omega})$. Hence, Theorem 4.24 on page 239 concerning G'' remains valid if $L^\infty(\Omega)$ is replaced by $L^2(\Omega)$; in fact, the argument carries over unchanged if we simply replace $L^\infty(\Omega)$ by $L^2(\Omega)$ in the proof (see also the discussion in the next section for $r := 2$). Consequently, under Assumption 4.14, G is twice continuously differentiable as a mapping from $L^2(\Omega)$ into $C(\bar{\Omega})$.

It remains to discuss the cost functional. To this end, we assume for simplicity that ψ is of the form

$$\psi(x, u) = \gamma_1(x)\, u + \gamma_2(x) u^2 \tag{4.109}$$

with functions γ_1, $\gamma_2 \in L^\infty(\Omega)$, where $\gamma_2 \geq 0$. Then the functional

$$\int_\Omega \big(\gamma_1(x)\, u(x) + \gamma_2(x) u(x)^2\big)\, dx$$

is obviously twice continuously Fréchet differentiable in $L^2(\Omega)$.

Remark. It can be shown that a sufficient second-order condition can only hold if $\gamma_2(x) \geq \delta > 0$ for almost every $x \in \Omega$; see, e.g., **[Trö00]**.

Based on these premises, we can replace the control space $L^\infty(\Omega)$ by $L^2(\Omega)$ in the (strong) second-order sufficient conditions.

Theorem 4.33. *Suppose that Assumption 4.14 on page 206 holds for the distributed control problem* (4.31)–(4.33)*, where* $\dim \Omega = N \leq 3$*. Let the control* $\bar{u} \in U_{ad} \subset L^2(\Omega)$*, together with the associated state* $\bar{y} = G(\bar{u})$ *and the adjoint state* p*, satisfy the first-order necessary optimality conditions stated in Theorem 4.20 on page 216. Moreover, let the function* $\psi = \psi(u)$ *be of the form* (4.109)*. If, in addition, the pair* $(\bar{y}, \bar{u})$ *satisfies the second-order sufficient condition* (4.103) *on page 247, then there exist constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that we have the quadratic growth condition*

$$J(y, u) \geq J(\bar{y}, \bar{u}) + \sigma \|u - \bar{u}\|^2_{L^2(\Omega)}$$

for every $u \in U_{ad}$ *with* $\|u - \bar{u}\|_{L^2(\Omega)} \leq \varepsilon$*, where* $y = G(u)$*. In particular,* $\bar{u}$ *is a locally optimal control in the sense of* $L^2(\Omega)$*.*

Boundary control. The situation is quite similar for the boundary control problem (4.49)–(4.51) on page 218, provided that Ω is a bounded two-dimensional Lipschitz domain. Then $G : u \mapsto y$ is twice continuously differentiable from $L^2(\Gamma)$ into $H^1(\Omega) \cap C(\bar{\Omega})$. Hence, with a corresponding

modification, Theorem 4.33 remains valid for $\Omega \subset \mathbb{R}^2$. In this case, the local optimality of $\bar{u}$ in the sense of $L^2(\Gamma)$ results.

4.10.7. Local optimality in $L^r(\Omega)$. The results on local optimality shown so far have a weakness when the two-norm discrepancy is present. Indeed, they only ensure the local optimality of $\bar{u}$ in the sense of $L^\infty(\Omega)$ or $L^\infty(\Gamma)$. Hence, if $\bar{u}$ has jump discontinuities, any function sufficiently close to $\bar{u}$ in the sense of the L^∞ norm has to exhibit the same jump behavior as $\bar{u}$; in particular, this must be the case in some L^∞ neighborhood of $\bar{u}$ in which $\bar{u}$ yields the (locally) smallest value of the cost functional. It would thus be of great benefit if one were able to show the local optimality of $\bar{u}$ with respect to the $L^r(\Omega)$ norm for some $1 \le r < \infty$; then such effects would not matter anymore.

We exemplarily address this problem for the case of distributed controls. We already know that G is for all $r > N/2$ continuously Fréchet differentiable as a mapping from $L^r(\Omega)$ into $C(\bar{\Omega}) \cap H^1(\Omega)$. Moreover, the linear solution operator $R : u \mapsto y$ of the boundary value problem

$$\begin{aligned} -\Delta y + y &= u \\ \partial_\nu y &= 0, \end{aligned}$$

which was introduced in the proof of Theorem 4.24 on page 239, also defines a continuous mapping from $L^r(\Omega)$ into $C(\bar{\Omega}) \cap H^1(\Omega)$. Consequently, the equation (4.83),

$$y - R\,(u - \Phi(y)) = F(y, u) = 0,$$

with $\Phi(y) = d(\cdot, y) - y$ is well posed in $(C(\bar{\Omega}) \cap H^1(\Omega)) \times L^r(\Omega)$. We have $F : (C(\bar{\Omega}) \cap H^1(\Omega)) \times L^r(\Omega) \to C(\bar{\Omega})$, and the implicit function theorem is applicable. Therefore, the solution operator G is twice continuously differentiable from $L^r(\Omega)$ into $C(\bar{\Omega}) \cap H^1(\Omega)$.

As in the last section, we now postulate that

$$\psi(x, u) = \gamma_1(x)\, u + \gamma_2(x)\, \lambda u^2.$$

A closer look at the steps leading up to Theorem 4.29 reveals that $\|h\|_{L^\infty(\Omega)}$ can always be replaced by $\|h\|_{L^r(\Omega)}$ without losing the validity of the respective estimates. Also, the proof of Lemma 4.26 on page 243 remains correct if the norm $\|u - \bar{u}\|_{L^\infty(\Omega)}$ there is replaced by $\|u - \bar{u}\|_{L^r(\Omega)}$. In summary, we have the following L^r version of Theorem 4.29.

Theorem 4.34. *Suppose that Assumption 4.14 holds, let $r > N/2$, and let the function ψ be of the form*

$$\psi(x, u) = \gamma_1(x)\, u + \gamma_2(x)\, \lambda\, u^2,$$

with $\gamma_i \in L^\infty(\Omega)$, $i = 1, 2$. Moreover, assume that the control $\bar{u} \in U_{ad}$, together with the associated state $\bar{y} = G(\bar{u})$ and the adjoint state p, satisfies the first-order necessary optimality condition stated in Theorem 4.20 on page 216. If, in addition, the (strong) second-order sufficient condition (4.104) *on page 247 is satisfied, then there exist constants $\varepsilon > 0$ and $\sigma > 0$ such that we have the quadratic growth condition*

$$J(y, u) \geq J(\bar{y}, \bar{u}) + \sigma \|u - \bar{u}\|^2_{L^2(\Omega)}$$

for all $u \in U_{ad}$ with $\|u - \bar{u}\|_{L^r(\Omega)} \leq \varepsilon$, where $y = G(u)$. In particular, $\bar{u}$ is a locally optimal control in the sense of $L^r(\Omega)$.

Remark. The above result does not directly generalize to problems in which either the control occurs nonlinearly in the differential equation or the cost functional does not have the requested quadratic form with respect to u. In such cases, additional assumptions have to be imposed; see **[CTU96]**. Also, the L^r optimality was derived without reference to strongly active constraints. Otherwise, the analysis becomes more delicate; in this regard, we refer to, e.g., **[TW06]** for the case involving the Navier–Stokes equations.

4.11. Numerical methods

4.11.1. Projected gradient methods. In principle, this method does not differ from that for linear-quadratic parabolic problems, which was described in the section beginning on page 166. However, the nonlinearity of the equation leads to an additional complication in step S1 that renders the method rather unattractive: if the approximation u_n is known, we have to solve the semilinear boundary value problem (4.50) for the associated state y_n. Usually, this has to be done by an iterative technique, for instance, by Newton's method. It therefore makes sense to apply a method of Newton type instead of the projected gradient method. One such technique is the SQP method, which will be discussed in detail in the next subsection.

Also, the choice of the step size is rather costly; even in the case without constraints, it can no longer be determined analytically. One has to be content with finding a step size s_n that yields a sufficiently large descent. This can be done either by bisection or by Armijo's rule; see Section 2.12.2.

4.11.2. Basic idea of the SQP method. To motivate the SQP method (**S**equential **Q**uadratic **P**rogramming method), we first study a problem in the space $\mathbb{R}^n$:

$$\min f(u), \qquad u \in C, \tag{4.110}$$

where $f \in C^2(\mathbb{R}^n)$ and $C \subset \mathbb{R}^n$ is nonempty, closed, and convex. Initially, we treat the case *without constraints*, that is, we take $C = \mathbb{R}^n$. Then the

first-order necessary optimality condition for a local solution $\bar{u}$ to (4.110) reads

$$f'(\bar{u}) = 0. \tag{4.111}$$

This equation can be solved via Newton's method, provided that the conditions for its convergence are met. If the iterate u_n is known, the next iterate $u = u_{n+1}$ is obtained as the solution to the system of linear equations

$$f'(u_n) + f''(u_n)(u - u_n) = 0. \tag{4.112}$$

To guarantee the unique solvability of this system, the matrix $f''(u_n)$ has to be nonsingular, that is, because we look for a local minimum, positive definite. In other words, the validity of the second-order sufficient optimality condition is a natural requirement for Newton's method to converge to a local minimizer.

We now take a different look at Newton's method. It is apparent that equation (4.112) is none other than the first-order necessary optimality condition for the linear-quadratic optimization problem

$$\min \left\{ f'(u_n)^\top (u - u_n) + \frac{1}{2} (u - u_n)^\top f''(u_n)(u - u_n) \right\}. \tag{4.113}$$

If the Hessian $f''(u_n)$ is positive definite, then the cost functional is strictly convex, so that this minimization problem has a unique solution u_{n+1}. In fact, it does not make any difference whether we solve the quadratic optimization problem (4.113) or the system of linear equations (4.112), since they are equivalent. Thus, Newton's method for the solution of the nonlinear system (4.111) can alternatively be performed as the solution of a sequence of quadratic optimization problems (4.113), that is, as an *SQP method*.

This observation is the key to the treatment of the problem *with constraints*. Instead of equation (4.111), we then have the variational inequality

$$f'(\bar{u})^\top (u - \bar{u}) \geq 0 \quad \forall u \in C. \tag{4.114}$$

A direct application of the classical Newton method is not possible. However, we can easily add the constraint $u \in C$ to problem (4.113); that is, in order to obtain the next iterate u_{n+1}, we solve the minimization problem

$$\boxed{\min_{u \in C} \left\{ f'(u_n)^\top (u - u_n) + \frac{1}{2} (u - u_n)^\top f''(u_n)(u - u_n) \right\}}. \tag{4.115}$$

If the Hessian $f''(u_n)$ is positive definite, then (4.115) is uniquely solvable. The solution of the sequence of quadratic optimization problems (4.115)

can be interpreted as a Newton method for a generalized equation. Just like the classical Newton method, this method converges locally quadratically. Sufficient conditions for its local convergence to $\bar{u}$ are, for example, the positive definiteness of $f''(\bar{u})$ and the regularity requirement $f \in C^{2,1}$. In this connection, we refer the interested reader to Alt [**Alt02**], Robinson [**Rob80**], and Spellucci [**Spe93**]. Presentations of the analysis of Newton's method for equations in function spaces can be found in Deuflhard [**Deu04**] and Kantorovich and Akilov [**KA64**]. Klatte and Kummer [**KK02**] discuss generalizations to merely Lipschitz continuous functions that are relevant to optimization problems.

Direct application to optimal control problems. The basic idea just described generalizes directly to the case where $\mathbb{R}^n$ is replaced by a Banach space U. For instance, let $U = L^\infty(\Gamma)$ and $C = U_{ad}$, and let the functional

$$f(u) = J\big(y(u), u\big) = \int_\Omega \varphi\big(x, y(x)\big)\, dx + \int_\Gamma \psi\big(x, u(x)\big)\, ds(x)$$

be given, where $y = y(u)$ denotes the weak solution to the elliptic boundary value problem (4.50),

$$\begin{aligned} -\Delta y &= 0 \qquad \text{in } \Omega \\ \partial_\nu y + b(x,y) &= u \qquad \text{on } \Gamma. \end{aligned}$$

The derivatives $f'(u_n)$ and $f''(u_n)$ are determined as in formula (4.56) on page 220 and Theorem 4.25 on page 242, respectively, using the Lagrangian function. Starting from $u_n \in U_{ad}$, we obtain $u = u_{n+1}$ as the solution to the quadratic optimal control problem

$$\min_{u \in U_{ad}} \Big\{ f'(u_n)(u - u_n) + \frac{1}{2} f''(u_n)(u - u_n)^2 \Big\}.$$

There is a small but essential difference between this "SQP" method and the familiar SQP method from the literature on nonlinear optimization (this is the reason why the method is sometimes also called *Newton's method*): once the new iterate u_{n+1} is calculated, the new state y_{n+1} is determined as the solution to a nonlinear elliptic boundary value problem, $y_{n+1} = y(u_{n+1}) = G(u_{n+1})$. The calculation of y_{n+1} could again be done using Newton's method. This additional effort is avoided by using, instead of the rule $y_{n+1} = G(u_{n+1})$, its linearization

$$y_{n+1} = y_n + G'(u_n)(u_{n+1} - u_n),$$

which amounts to solving just a linear equation. We will investigate this type of SQP method in the next subsection.

4.11.3. The SQP method for elliptic problems. Here, we discuss the method for the distributed control problem (4.31)–(4.33):

$$\min J(y,u) := \int_\Omega \varphi\big(x,y(x)\big)\,dx + \int_\Omega \psi\big(x,u(x)\big)\,dx,$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + d(x,y) &=& u & \text{in } \Omega \\ \partial_\nu y &=& 0 & \text{on } \Gamma \end{array}}$$

and

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

We postulate that Assumption 4.14 holds. We aim to determine a local reference solution $(\bar y, \bar u)$ that is supposed to satisfy the second-order sufficient optimality conditions (4.104). As before, p denotes the associated adjoint state. The triple $(\bar y, \bar u, p)$ solves the *optimality system*

$$\begin{array}{rcl@{\qquad}rcl} -\Delta y + d(x,y) &=& u & -\Delta p + d_y(x,y)\,p &=& \varphi_y(x,y) \\ \partial_\nu y &=& 0 & \partial_\nu p &=& 0 \end{array}$$

$$\int_\Omega \big(\psi_u(x,u) + p\big)(v-u)\;dx \ge 0 \qquad \forall\, v \in U_{ad}.$$

As in the motivation of the SQP method, we first consider the unconstrained case $U_{ad} = L^\infty(\Omega)$. Then we have the equation $\psi_u(\cdot,u) + p = 0$ instead of the variational inequality, and hence the optimality system

$$\begin{array}{rcl@{\qquad}rcl} -\Delta y + d(x,y) &=& u & -\Delta p + d_y(x,y)\,p &=& \varphi_y(x,y) \\ \partial_\nu y &=& 0 & \partial_\nu p &=& 0 \end{array}$$

$$\psi_u(x,u) + p = 0.$$

This nonlinear system for the unknowns (y,u,p) can be solved by means of Newton's method; see Deuflhard **[Deu04]**. To this end, suppose that the iterates (y_i,u_i,p_i), $1 \le i \le n$, have already been determined. Then the new iterate $(y_{n+1},u_{n+1},p_{n+1})$ is the solution to the optimality system linearized at (y_n,u_n,p_n).

To find the latter, recall that linearization of a mapping F just means to make the approximation $F(y) \approx F(y_n) + F'(y_n)(y-y_n)$. If we perform this for the first equation, we find the linearized equation

$$-\Delta y_n - \Delta(y-y_n) + d(x,y_n) + d_y(x,y_n)(y-y_n) - u_n - (u-u_n) = 0,$$

that is,

$$-\Delta y + d(x, y_n) + d_y(x, y_n)(y - y_n) - u = 0.$$

Hence, the linear terms are preserved. We thus obtain, as optimality system for the determination of $(y, u, p) = (y_{n+1}, u_{n+1}, p_{n+1})$, the equations

(4.116)

$$\begin{aligned} -\Delta y + d(x, y_n) + d_y(x, y_n)(y - y_n) &= u \\ \partial_\nu y &= 0 \\ -\Delta p + d_y(x, y_n)\, p + p_n\, d_{yy}(x, y_n)(y - y_n) &= \varphi_y(x, y_n) \\ &\quad + \varphi_{yy}(x, y_n)(y - y_n) \\ \partial_\nu p &= 0 \\ \psi_u(x, u_n) + \psi_{uu}(x, u_n)(u - u_n) + p &= 0. \end{aligned}$$

Obviously, this is just the optimality system for the problem

$$\begin{aligned} \min \Big\{ & \int_\Omega \big(\varphi_y(x, y_n)\,(y - y_n) + \psi_u(x, u_n)\,(u - u_n)\big)\, dx \\ & -\frac{1}{2} \int_\Omega p_n\, d_{yy}(x, y_n)(y - y_n)^2\, dx \\ & +\frac{1}{2} \int_\Omega \big(\varphi_{yy}(x, y_n)(y - y_n)^2 + \psi_{uu}(x, u_n)(u - u_n)^2\big)\, dx \Big\}, \end{aligned}$$

subject to $u \in L^2(\Omega)$ and

$$\begin{aligned} -\Delta y + d(x, y_n) + d_y(x, y_n)(y - y_n) &= u \\ \partial_\nu y &= 0. \end{aligned}$$

This problem is in turn equivalent to the linear-quadratic problem

$$\min \Big\{ J'(y_n, u_n)(y - y_n, u - u_n) + \frac{1}{2}\, \mathcal{L}''(y_n, u_n, p_n)(y - y_n, u - u_n)^2 \Big\},$$

subject to

$$\begin{aligned} -\Delta y + d(x, y_n) + d_y(x, y_n)(y - y_n) &= u \\ \partial_\nu y &= 0. \end{aligned}$$

Thus we may solve this problem instead of the system (4.116).

While (4.116) as a *system of equations* does not directly generalize to the case with constraints $u \in U_{ad}$, this is evidently true for the above problem:

one only has to add the constraints. With box constraints, we have to solve in the nth iteration step the following problem:

(QP_n)

$$\min \left\{ J'(y_n,u_n)(y-y_n,u-u_n) + \frac{1}{2}\mathcal{L}''(y_n,u_n,p_n)(y-y_n,u-u_n)^2 \right\}$$

subject to

$$\begin{aligned} -\Delta y + d(x,y_n) + d_y(x,y_n)(y-y_n) &= u \\ \partial_\nu y &= 0 \end{aligned}$$

and

$$u_a \le u \le u_b.$$

As a result, we obtain the new control u_{n+1}, the new state y_{n+1}, and then also the associated adjoint state p_{n+1}. The reader will be asked in Exercise 4.10 to determine the boundary value problem to be satisfied by p_{n+1}. With this, the algorithm of the SQP method is described. A number of questions arise: Does (QP_n) have a solution (y_{n+1},u_{n+1})? If so, is it unique? Does the method converge, and if so, what is the order of convergence?

Theorem 4.35. *Suppose that Assumption 4.14 on page 206 holds for the distributed control problem* (4.31)–(4.33). *Moreover, let the triple* $(\bar y,\bar u,p)$ *satisfy the necessary optimality conditions and the second-order sufficient optimality condition* (4.104) *on page 247. Then there is some convergence radius* $\varrho>0$ *such that the SQP method, starting from an initial guess* (y_0,u_0,p_0) *with*

$$\max\left\{ \|y_0-\bar y\|_{C(\bar\Omega)},\ \|u_0-\bar u\|_{L^\infty(\Omega)},\ \|p_0-p\|_{C(\bar\Omega)} \right\} < \varrho,$$

generates a uniquely determined sequence $\{(y_n,u_n,p_n)\}_{n=1}^\infty$ *of iterates. Moreover, there is a constant* $c_N>0$ *such that*

$$\begin{aligned} &\big\|(y_{n+1},u_{n+1},p_{n+1}) - (\bar y,\bar u,p)\big\|_{C(\bar\Omega)\times L^\infty(\Omega)\times C(\bar\Omega)} \\ &\quad\le c_N \big\|(y_n,u_n,p_n) - (\bar y,\bar u,p)\big\|^2_{C(\bar\Omega)\times L^\infty(\Omega)\times C(\bar\Omega)} \qquad \forall\, n\in\mathbb{N}. \end{aligned}$$

The SQP method is therefore locally quadratically convergent under the given assumptions. Here, we have assumed the stronger sufficient optimality conditions (4.104), which do not invoke strongly active inequality constraints. This is common practice in convergence proofs for SQP methods. In **[Ung97]**, a proof based on the Newton method for generalized equations can be found. Another, to a large extent analogous, proof for the parabolic case is given in **[Trö99]**.

In the case of *boundary controls*, the structure and theory of the SQP method is completely analogous to that for distributed controls. The same is true if homogeneous boundary conditions of Dirichlet type are considered instead of Neumann conditions.

Remark. It requires considerable additional effort to develop the basic idea of the SQP method described here into reliable software tools. For instance, techniques for globalization have to be incorporated, and the solution of the quadratic subproblems has to be linked to the outer iteration in an efficient way. For details, we refer the reader to the relevant literature. However, the basic scheme described above shows excellent performance for the test examples given in this book.

4.12. Exercises

4.1 Let Ω be a bounded Lipschitz domain. Prove the inequality

$$\|y\|_{L^\infty(\Gamma)} \le \|y\|_{L^\infty(\Omega)} \quad \forall\, y \in H^1(\Omega) \cap L^\infty(\Omega).$$

Hint: Use the fact that $H^1(\Omega) \cap C(\bar\Omega)$ is dense in $H^1(\Omega)$. Choose a sequence $\{y_n\} \subset H^1(\Omega) \cap C(\bar\Omega)$ such that $\|y_n - y\|_{H^1(\Omega)} \to 0$ as $n \to \infty$, and project this sequence onto the interval $[-c, c]$, where $c = \|y\|_{L^\infty(\Omega)}$. The projection operator is continuous in $H^1(\Omega) \cap C(\bar\Omega)$. Apply Theorem 2.1 on page 29.

4.2 Examine the uniqueness question for the semilinear elliptic boundary value problem (4.5) on page 183. Show that under Assumption 4.2 on page 185 there can be at most one solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$.

4.3 Let $E \subset \mathbb{R}^N$ be a bounded and measurable set. For which spaces $L^q(E)$ is the Nemytskii operator $y(\cdot) \mapsto \sin(y(\cdot))$ Fréchet differentiable from $L^2(E)$ into $L^q(E)$?

4.4 (i) Prove that the Nemytskii operator $y(\cdot) \mapsto \sin(y(\cdot))$ is not Fréchet differentiable in any of the spaces $L^p(0,T)$ with $1 \le p < \infty$. *Hint*: The requested property of the remainder is already violated for step functions.

(ii) Show that this operator is Fréchet differentiable from $L^{p_1}(0,T)$ into $L^{p_2}(0,T)$ whenever $1 \le p_2 < p_1 \le \infty$.

4.5 Suppose that Assumption 4.14 on page 206 holds. Verify without the use of Lemma 4.11 that the functionals F and Q defined in (4.34), namely

$$F(y) = \int_\Omega \varphi\big(x, y(x)\big)\, dx, \qquad Q(u) = \int_\Omega \psi\big(x, u(x)\big)\, dx,$$

are Lipschitz continuous in their domains of definition. Show also that Q is convex.

4.6 Let $E \subset \mathbb{R}^N$ be bounded and measurable, and let $u_a, u_b \in L^\infty(E)$ be given such that $u_a(x) \le u_b(x)$ for almost every $x \in E$. Show that the set

$$U_{ad} = \big\{u \in L^r(E) \,:\, u_a(x) \le u(x) \le u_b(x) \ \text{ for a.e } \ x \in E\big\}$$

is nonempty, closed, bounded, and convex in $L^r(E)$ whenever $1 \le r \le \infty$.

4.7 Prove the representation used in (4.40):

$$\Phi(\tilde{y}) - \Phi(\bar{y}) = d(\cdot, \tilde{y}(\cdot)) - d(\cdot, \bar{y}(\cdot)) = d_y(\cdot, \bar{y}(\cdot))\,(\tilde{y}(\cdot) - \bar{y}(\cdot)) + r_d,$$

with a remainder r_d that satisfies $\|r_d\|_{C(\bar{\Omega})}/\|\tilde{y} - \bar{y}\|_{C(\bar{\Omega})} \to 0$ as $\|\tilde{y} - \bar{y}\|_{C(\bar{\Omega})} \to 0$.

4.8 Show that for $\lambda \in (0, 1]$ and $y_\Omega \equiv 9$ $\bar{u} \equiv 2$ satisfies the necessary optimality conditions for the "superconductivity" problem defined on page 217. Are the sufficient second-order optimality conditions satisfied?

4.9 Show that the functional f defined by

$$f(u) = \int_0^1 \cos\big(u(x)\big)\,dx$$

is not twice Fréchet differentiable in $L^2(0, 1)$. Use the hint given in Exercise 4.4. In which of the spaces $L^p(0, 1)$ does a second-order Fréchet derivative exist?

4.10 Derive the adjoint equation solved by the adjoint state p_{n+1} for the linear-quadratic problem (QP_n) defined on page 262.

Chapter 5

Optimal control of semilinear parabolic equations

In this chapter we study, in analogy to the elliptic case, optimal control problems for semilinear parabolic problems. While the existence and regularity theory for solutions to parabolic problems differs in many aspects from that for elliptic problems, the optimization theory for parabolic problems is rather similar. Hence, although we give a rather comprehensive treatment of the state problems, we do not go into as much detail as in Chapter 4 regarding the optimization theory. We can afford this, since the corresponding proofs are more or less the same as in the elliptic case.

5.1. The semilinear parabolic model problem

The state problems to be studied in the subsequent sections are all special cases of the following general initial-boundary value problem:

$$\boxed{\begin{aligned} y_t + \mathcal{A}\,y + d(x,t,y) &= f && \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y + b(x,t,y) &= g && \text{on } \Sigma \\ y(\cdot,0) &= y_0 && \text{in } \Omega. \end{aligned}} \tag{5.1}$$

Here, $T > 0$ is a given final time, and we have set $Q := \Omega \times (0,T)$ and $\Sigma := \Gamma \times (0,T)$. As before, $\mathcal{A}$ is the uniformly elliptic differential operator

defined in (2.19) on page 37, and $\partial_{\nu_{\mathcal{A}}}$ denotes the corresponding outward conormal derivative. For clarity, we usually suppress as in (5.1) the variables x and t in the function y and in the given data. The functions d and b are defined as in the elliptic case.

In this section, we present relevant properties of the initial-boundary value problem (5.1), the main result being Theorem 5.5 on the existence and uniqueness of a continuous weak solution. These results, which are due to Casas **[Cas97]** and Raymond and Zidani **[RZ98]**, form the basis of the corresponding optimal control theory. We generally assume the following:

Assumption 5.1. $\Omega \subset \mathbb{R}^N$, $N \geq 1$, *is a bounded Lipschitz domain (for* $N = 1$ *a bounded open interval). The function* $d = d(x,t,y) : Q \times \mathbb{R} \to \mathbb{R}$ *is measurable with respect to* $(x,t) \in Q$ *for any fixed* $y \in \mathbb{R}$. *Similarly, let* $b = b(x,t,y) : \Sigma \times \mathbb{R} \to \mathbb{R}$ *satisfy the same condition with* Σ *in place of* Q. *Moreover,* d *and* b *are monotone increasing with respect to* y *for almost every* $(x,t) \in Q$ *and* $(x,t) \in \Sigma$, *respectively.*

As standard spaces for the treatment of linear parabolic initial-boundary value problems, we have so far used $W_2^{1,0}(Q)$ and $W(0,T)$. Evidently, if $y \in W(0,T)$ or $y \in W_2^{1,0}(Q)$, then the functions $d(x,t,y(x,t))$ or $b(x,t,y(x,t))$ may be unbounded and possibly not integrable, unless further assumptions are imposed. For the proof of the existence of a unique solution to (5.1), we initially make the following additional assumption:

Assumption 5.2. *The function* $d = d(x,t,y) : Q \times \mathbb{R} \to \mathbb{R}$ *is uniformly bounded and globally Lipschitz continuous with respect to* y *for almost every* $(x,t) \in Q$, *that is, there are constants* $K > 0$ *and* $L > 0$ *such that for almost all* $(x,t) \in Q$ *and all* y_1, $y_2 \in \mathbb{R}$ *we have*

$$|d(x,t,0)| \leq K \tag{5.2}$$

$$|d(x,t,y_1) - d(x,t,y_2)| \leq L\,|y_1 - y_2|. \tag{5.3}$$

The function $b = b(x,t,y) : \Sigma \times \mathbb{R} \to \mathbb{R}$ *is assumed to satisfy the same condition with* Σ *in place of* Q.

The weak formulation (3.25) on page 140, valid for linear problems, is extended to the nonlinear case as follows.

Definition. *Suppose Assumptions 5.1 and 5.2 hold. A function* $y \in W_2^{1,0}(Q)$ *is said to be a* weak solution *to* (5.1) *if*

$$-\iint_Q y\, v_t \,dx\,dt + \iint_Q \Big(\sum_{i,j=1}^{N} a_{ij}(x)\, D_i y\, D_j v + d(x,t,y)\, v \Big)\, dx\,dt \tag{5.4}$$

$$+\iint_\Sigma b(x,t,y)\,v\,ds\,dt=\iint_Q f\,v\,dx\,dt+\iint_\Sigma g\,v\,ds\,dt+\int_\Omega y_0\,v(\cdot,0)\,dx$$

for every $v\in W_2^{1,1}(Q)$ *such that* $v(x,T)=0$.

Relation (5.4) is called the *weak* or *variational formulation* of the initial-boundary value problem (5.1). The following result is the parabolic analogue of Theorem 4.4 on page 186.

Lemma 5.3. *Suppose that Assumptions 5.1 and 5.2 hold. Then for every given triple* $f\in L^2(Q)$, $g\in L^2(\Sigma)$, *and* $y_0\in L^2(\Omega)$ *the initial-boundary value problem* (5.1) *has a unique weak solution* $y\in W_2^{1,0}(Q)$.

The above lemma will be proved in Section 7.3.1, beginning on page 373. However, Assumption 5.2 is much too restrictive and excludes many important applications. For instance, nonlinearities such as $d(y)=y^n$, $n>1$, fail to satisfy it. For this reason, one works with the more general conditions of *local* boundedness and Lipschitz continuity.

Assumption 5.4. *The function* $d=d(x,t,y):Q\times\mathbb{R}\to\mathbb{R}$ *satisfies on* $E=Q$ *the boundedness condition* (5.2) *and is for every* $(x,t)\in E$ *locally Lipschitz continuous with respect to* y, *that is, for any* $M>0$ *there is some* $L(M)>0$ *such that*

(5.5)

$$|d(x,t,y_1)-d(x,t,y_2)|\le L(M)\,|y_1-y_2|\quad\forall\,y_i\in\mathbb{R}\ \text{with}\ |y_i|\le M,\ \ i=1,2.$$

The same is assumed to hold for $b=b(x,t,y):\Sigma\times\mathbb{R}\to\mathbb{R}$ *on* $E=\Sigma$.

Analogously to Theorem 4.5 on page 189, it can be shown without the strong Assumption 5.2 that, for given data f and g from suitable L^p spaces, there exists a unique solution in $W(0,T)\cap L^\infty(Q)$. For this result to be valid, the boundedness of d and b is not needed, and the weaker Assumption 5.4 suffices. Therefore, the following generalization of the notion of weak solution is meaningful.

Definition. *A function* $y\in W_2^{1,0}(Q)\cap L^\infty(Q)$ *is said to be a weak solution to* (5.1) *if the variational formulation* (5.4) *holds for all* $v\in W_2^{1,1}(Q)$ *satisfying the additional condition* $v(x,T)=0$.

The corresponding existence and uniqueness results in connection with optimal control problems have been proved by Casas [**Cas97**] and Raymond and Zidani [**RZ99**].

Theorem 5.5. *Suppose that Assumptions 5.1 and 5.4 hold. Then the semilinear parabolic initial-boundary value problem*

$$\begin{aligned} y_t + \mathcal{A}\, y + d(x,t,y) &= f && \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y + b(x,t,y) &= g && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega \end{aligned}$$

has a unique weak solution $y \in W(0,T) \cap C(\bar{Q})$ for any triple $f \in L^r(Q)$, $g \in L^s(\Sigma)$, and $y_0 \in C(\bar{\Omega})$ with $r > N/2+1$ and $s > N+1$. Moreover, there is a constant $c_\infty > 0$, which is independent of d, b, f, g, and y_0, such that

$$\begin{aligned} \|y\|_{W(0,T)} + \|y\|_{C(\bar{Q})} \;\le\; c_\infty \big(&\|f - d(\cdot,0)\|_{L^r(Q)} \\ &+ \|g - b(\cdot,0)\|_{L^s(\Sigma)} + \|y_0\|_{C(\bar{\Omega})} \big). \end{aligned} \tag{5.6}$$

The basic idea of the proof is the following. By virtue of Lemma 5.3, one obtains as in the elliptic case a unique solution to the problem with cutoff functions d_k and b_k. Using techniques from [**LSU68**], one then shows that $\|y_k\|_{L^\infty(Q)}$ is bounded from above by a constant that does not depend on k. The continuity of the solution is a consequence of Lemma 7.12 on page 378, which is the parabolic analogue of Theorem 4.8 on page 192. To be able to apply this lemma, we bring the bounded and measurable functions $d(x,t,y)$ and $b(x,t,y)$ to the right-hand sides of the differential equation and the boundary condition, respectively; in this way, we obtain data from $L^r(Q)$ and $L^s(\Sigma)$, respectively. Finally, the estimate for $\|y\|_{W(0,T)}$ in (5.6) follows from Lemma 7.10.

Remark. If $y_0 \in L^\infty(\Omega)$ only, then $y \in C(\bar{Q})$ can no longer be expected. In this case, one only obtains the regularity $y \in C((0,T] \times \bar{\Omega}) \cap L^\infty(Q)$; see Raymond and Zidani [**RZ99**]. In particular, this concerns the regularity of the adjoint states, because in their final condition a function may occur that is merely bounded and measurable. In this case, the norm $\|y\|_{C(\bar{Q})}$ in the estimate (5.6) must be replaced by $\|y\|_{L^\infty(Q)}$.

5.2. Basic assumptions for the chapter

For better readability, we now impose a set of assumptions for this chapter that is sufficiently strong for all the following theorems to hold. However, several of the results are valid under much weaker assumptions. It will become evident in the individual theorems which parts of the general assumption are dispensable. Moreover, for the sake of a shorter exposition we confine ourselves to the Laplacian $-\Delta$, although all of the following results remain valid for the general elliptic operator $\mathcal{A}$ studied so far.

Besides $d = d(x,t,y)$ and $b = b(x,t,y)$, the following quantities occur: in the cost functionals the functions $\phi = \phi(x,y)$, $\varphi = \varphi(x,t,y,v)$, and $\psi = \psi(x,t,y,u)$, and in the control constraints the threshold functions u_a, u_b, v_a, and v_b, which all depend on (x,t). The "function variables" will be y, whenever the state $y(x,t)$ is to be inserted, as well as v and u for the controls $v(x,t)$ and $u(x,t)$. Some of these may not come up, but this will not contradict the assumptions to follow.

Again, we use the notation $Q := \Omega \times (0,T)$ and $\Sigma := \Gamma \times (0,T)$, for some fixed final time $T > 0$.

Assumption 5.6.
(**i**) $\Omega \subset \mathbb{R}^N$ *is a bounded Lipschitz domain.*

(**ii**) *The functions*

$$\begin{aligned} &d = d(x,t,y) : Q \times \mathbb{R} \to \mathbb{R}, && \phi = \phi(x,y) : \Omega \times \mathbb{R} \to \mathbb{R}, \\ &\varphi = \varphi(x,t,y,v) : Q \times \mathbb{R}^2 \to \mathbb{R}, && b = b(x,t,y) : \Sigma \times \mathbb{R} \to \mathbb{R}, \\ &\psi = \psi(x,t,y,u) : \Sigma \times \mathbb{R}^2 \to \mathbb{R} \end{aligned}$$

are measurable with respect to (x,t) for all $y, v, u \in \mathbb{R}$ and, for almost every (x,t) in Q or Σ, twice differentiable with respect to y, v, and u. Moreover, they satisfy the boundedness and local Lipschitz conditions (4.24)–(4.25) *of order $k = 2$; this means that for φ, for example, there exist some $K > 0$ and a constant $L(M) > 0$ for any $M > 0$ such that we have, with the objects $\nabla\varphi$ and φ'' explained below,*

$$\begin{aligned} &|\varphi(x,t,0,0)| + |\nabla\varphi(x,t,0,0)| + |\varphi''(x,t,0,0)| \le K, \\ &|\varphi''(x,t,y_1,v_1) - \varphi''(x,t,y_2,v_2)| \le L(M)\,\big\{|y_1 - y_2| + |v_1 - v_2|\big\}, \end{aligned}$$

for almost every $(x,t) \in Q$ and any y_i, $v_i \in [-M,M]$, $i = 1,2$.

(**iii**) *We have $d_y(x,t,y) \ge 0$ for almost every $(x,t) \in Q$ and $b_y(x,t,y) \ge 0$ for almost every $(x,t) \in \Sigma$. Moreover, $y_0 \in C(\bar\Omega)$.*

(**iv**) *The bounds u_a, u_b and v_a, $v_b : E \to R$ belong to $L^\infty(E)$ for $E = \Sigma$ and $E = Q$, respectively, and $u_a(x,t) \le u_b(x,t)$ and $v_a(x,t) \le v_b(x,t)$ for almost every $(x,t) \in E$.*

Remark. The gradient $\nabla\varphi$ and the Hessian φ'' in (ii) are defined by

$$\nabla\varphi = \begin{bmatrix} \varphi_y \\ \varphi_v \end{bmatrix}, \qquad \varphi'' = \begin{bmatrix} \varphi_{yy} & \varphi_{yv} \\ \varphi_{vy} & \varphi_{vv} \end{bmatrix}.$$

For these quantities, $|\cdot|$ represents an arbitrary norm in $\mathbb{R}^2$ or $\mathbb{R}^{2\times 2}$, respectively. Assumption 5.6 is, for instance, satisfied with functions d, $b \in C^3(\mathbb{R})$ that only

depend on the function variable, such as $d(y) = y^k$ where $k \in \mathbb{N}$ is odd or $d(y) = \exp(y)$. A typical example of an admissible φ is given by

$$\varphi(x,t,y,v) = \alpha(x,t)\,(y - y_Q(x,t))^2 + \beta(x,t)\,(v - v_Q(x,t))^2,$$

with α, β, y_Q, $v_Q \in L^\infty(Q)$; see also the remarks following Assumption 4.14 on page 206 for the elliptic case.

5.3. Existence of optimal controls

We begin the study of parabolic optimal control problems by showing the existence of optimal controls. In order to be able to treat several cases at the same time, we consider a problem that contains both distributed and boundary controls as well as a cost functional that combines observations on the boundary, within the domain, and at the final time. More precisely, we consider the problem

(5.7)
$$\begin{aligned}\min J(y,v,u) \quad := \quad & \int_\Omega \phi\big(x, y(x,T)\big)\,dx + \iint_Q \varphi\big(x,t,y(x,t),v(x,t)\big)\,dx\,dt \\ & + \iint_\Sigma \psi\big(x,t,y(x,t),u(x,t)\big)\,ds\,dt,\end{aligned}$$

subject to

(5.8)
$$\boxed{\begin{aligned} y_t - \Delta y + d(x,t,y) &= v && \text{in } Q \\ \partial_\nu y + b(x,t,y) &= u && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega \end{aligned}}$$

and

(5.9)
$$\begin{aligned} v_a(x,t) \le v(x,t) \le v_b(x,t) \qquad & \text{for a.e. } (x,t) \in Q \\ u_a(x,t) \le u(x,t) \le u_b(x,t) \qquad & \text{for a.e. } (x,t) \in \Sigma. \end{aligned}$$

If one of the two controls is not to occur, one can enforce pure boundary control by putting $v_a = v_b = 0$, or pure distributed control by putting $u_a = u_b = 0$. As the sets of admissible controls, we define

$$\begin{aligned} V_{ad} &= \big\{v \in L^\infty(Q) \,:\, v_a(x,t) \le v(x,t) \le v_b(x,t) \quad \text{for a.e. } (x,t) \in Q\big\} \\ U_{ad} &= \big\{u \in L^\infty(\Sigma) \,:\, u_a(x,t) \le u(x,t) \le u_b(x,t) \quad \text{for a.e. } (x,t) \in \Sigma\big\}. \end{aligned}$$

In the following, we denote by $y = y(v,u)$ the state associated with the control $(v,u) \in V_{ad} \times U_{ad}$.

Definition. *A pair of controls* $(\bar v, \bar u) \in V_{ad} \times U_{ad}$ *is said to be* optimal, *and* $\bar y = y(\bar v, \bar u)$ *the associated* optimal state, *if*

$$J\big(y(\bar v, \bar u), \bar v, \bar u\big) \le J\big(y(v,u), v, u\big) \quad \forall\, (v,u) \in V_{ad} \times U_{ad}.$$

A pair $(\bar v, \bar u) \in V_{ad} \times U_{ad}$ *is said to be* locally optimal *in the sense of* $L^r(Q) \times L^s(\Sigma)$ *if there is some* $\varepsilon > 0$ *such that the above inequality holds for all* $(v,u) \in V_{ad} \times U_{ad}$ *such that* $\|v - \bar v\|_{L^r(Q)} + \|u - \bar u\|_{L^s(\Sigma)} \le \varepsilon$.

In the next theorem, convexity of φ and ψ with respect to the control variables will be needed. For φ, for example, this requirement means that

$$\varphi\big(x,t,y, \lambda\, v_1 + (1-\lambda) v_2\big) \le \lambda\, \varphi\big(x,t,y,v_1\big) + (1-\lambda)\, \varphi\big(x,t,y,v_2\big)$$

for almost every $(x,t) \in Q$, all $y, v_1, v_2 \in \mathbb{R}$, and every $\lambda \in (0,1)$. The convexity of ψ with respect to the control variable is understood similarly.

Theorem 5.7. *Suppose that Assumption 5.6 holds, and let* φ *and* ψ *be convex with respect to* v *and* u, *respectively. Then the optimal control problem* (5.7)–(5.9) *has at least one optimal pair* $(\bar v, \bar u)$ *with associated optimal state* $\bar y = y(\bar v, \bar u)$.

Proof: The proof is similar to that of Theorem 4.15 for elliptic problems. We can therefore afford to be brief. By virtue of Theorem 5.5, the state equation (5.8) has for any pair of admissible controls a unique associated state $y = y(v,u) \in W(0,T) \cap C(\bar Q)$. Now observe that $V_{ad} \times U_{ad}$ is a bounded subset of $L^\infty(Q) \times L^\infty(\Sigma)$ and, a fortiori, also of $L^r(Q) \times L^s(\Sigma)$ for $r > N/2 + 1$ and $s > N + 1$. In view of estimate (5.6), we can thus conclude that there is some $M > 0$ such that

$$\|y(v,u)\|_{C(\bar Q)} \le M \quad \forall\, (v,u) \in V_{ad} \times U_{ad}. \tag{5.10}$$

Owing to Assumption 5.6 and (5.10), and since U_{ad} and V_{ad} are bounded, the functional J is bounded from below and therefore has a finite infimum j. By the reflexivity of $L^r(Q) \times L^s(\Sigma)$, we may select a minimizing sequence $\big\{(v_n, u_n)\big\}_{n=1}^\infty$ that converges weakly in this space to some limit $(\bar v, \bar u)$:

$$v_n \rightharpoonup \bar v, \quad u_n \rightharpoonup \bar u \quad \text{as } n \to \infty.$$

Since $V_{ad} \times U_{ad}$ is closed and convex, we have $(\bar v, \bar u) \in V_{ad} \times U_{ad}$, so it is an admissible pair.

Next, the strong convergence of the state sequence in a suitable space has to be shown, which requires a little more effort than in the elliptic case. To this end, define the sequences of functions $z_n(x,t) = -d(x,t,y_n(x,t))$ and $w_n(x,t) = -b(x,t,y_n(x,t))$, $n \in \mathbb{N}$. By (5.10) and Assumption 5.6,

these sequences are almost everywhere uniformly bounded. Hence, there are subsequences, without loss of generality $\{z_n\}_{n=1}^\infty$ and $\{w_n\}_{n=1}^\infty$ themselves, that converge weakly in $L^r(Q)$ and $L^s(\Sigma)$, respectively, to limits z and w.

We now regard the semilinear parabolic initial-boundary value problem as a linear problem with right-hand sides $z_n + v_n$ and $w_n + u_n$:

$$\begin{aligned} y_{n,t} - \Delta y_n &= z_n + v_n \\ \partial_\nu y_n &= w_n + u_n \\ y_n(0) &= y_0. \end{aligned} \tag{5.11}$$

These right-hand sides converge weakly to $z + \bar{v}$ and $w + \bar{u}$, respectively. Since the solution mapping $(v, u) \mapsto y(v, u)$ of the linear parabolic problem is weakly continuous, we can infer that the state sequence converges weakly in $W(0, T)$ to some limit $\bar{y} \in W(0, T)$,

$$y_n \rightharpoonup \bar{y} \quad \text{as } n \to \infty.$$

At this point, we employ a regularity result from **[Gri07a, Gri07b]**, which asserts that for zero initial condition $y_0 := 0$ the mapping $(v, u) \mapsto y(v, u)$ maps $L^r(Q) \times L^s(\Sigma)$ continuously into the space of Hölder continuous functions $C^{0,\kappa}(\bar{Q})$ for some $\kappa \in (0, 1)$.

Now let $\hat{y} \in C(\bar{Q})$ denote the (fixed) contribution to y_n that solves the linear parabolic problem with initial value y_0, homogeneous right-hand side and homogeneous boundary condition in (5.11). Then the sequence $\{y_n - \hat{y}\}_{n=1}^\infty$ is weakly convergent in $C^{0,\kappa}(\bar{Q})$. Since $C^{0,\kappa}(\bar{Q})$ is by the Arzelà–Ascoli theorem compactly embedded in $C(\bar{Q})$, the sequence also converges strongly in $C(\bar{Q})$. Now, $\hat{y} \in C(\bar{Q})$, and thus

$$y_n \to \bar{y} \quad \text{as } n \to \infty$$

with some $\bar{y} \in C(\bar{Q})$. Hence, we even have uniform convergence. This simplifies the following arguments relative to the elliptic case (where, however, we could work with simpler methods).

Owing to the assumed local Lipschitz continuity of d and b, we can infer that

$$\begin{aligned} d(\cdot, \cdot, y_n) &\to d(\cdot, \cdot, \bar{y}) \quad \text{strongly in } L^\infty(Q) \text{ and in } L^2(Q), \\ b(\cdot, \cdot, y_n) &\to b(\cdot, \cdot, \bar{y}) \quad \text{strongly in } L^\infty(\Sigma) \text{ and in } L^2(\Sigma). \end{aligned}$$

As in the elliptic case, we now use the variational formulation of the parabolic initial-boundary value problem to conclude that $\bar{y}$ is the weak solution associated with the pair $(\bar{v}, \bar{u})$, that is, $\bar{y} = y(\bar{v}, \bar{u})$. To this end,

note that we have, for every test function $w \in W_2^{1,1}(Q)$ with $w(T) = 0$,

$$\begin{aligned} &- \iint_Q y_n \, w_t \, dx \, dt + \iint_Q \big(\nabla y_n \cdot \nabla w + d(x,t,y_n) \, w \big) \, dx \, dt \\ &\quad + \iint_\Sigma b(x,t,y_n) \, w \, ds \, dt \\ &= \iint_Q v_n \, w \, dx \, dt + \iint_\Sigma u_n \, w \, ds \, dt + \int_\Omega y_0 \, w(\cdot,0) \, dx. \end{aligned}$$

Passage to the limit as $n \to \infty$, using the convergences shown above, yields that

$$\begin{aligned} &- \iint_Q \bar{y} \, w_t \, dx \, dt + \iint_Q \big(\nabla \bar{y} \cdot \nabla w + d(x,t,\bar{y}) \, w \big) \, dx \, dt \\ &\quad + \iint_\Sigma b(x,t,\bar{y}) \, w \, ds \, dt \\ &= \iint_Q \bar{v} \, w \, dx \, dt + \iint_\Sigma \bar{u} \, w \, ds \, dt + \int_\Omega y_0 \, w(\cdot,0) \, dx, \end{aligned}$$

that is, $\bar{y}$ is indeed a weak solution.

It remains to show the optimality of $(\bar{v}, \bar{u})$. To do this, we have to invoke the lower semicontinuity of the cost functional. It is apparent that the arguments used in the proof of the elliptic case carry over unchanged. The assertion is thus completely proved. □

Remark. Obviously, only the boundedness and Lipschitz conditions of order $k = 0$ from Assumption 5.6 were needed in the above proof.

5.4. The control-to-state operator

In this section, we prove the continuity and differentiability of the control-to-state mapping. Again, we treat the cases of boundary and distributed controls simultaneously, by considering the problem (5.8):

$$\boxed{\begin{aligned} y_t - \Delta y + d(x,t,y) &= v && \text{in } Q \\ \partial_\nu y + b(x,t,y) &= u && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega. \end{aligned}}$$

Again, we denote the control-to-state mapping by $G : V \times U := L^r(Q) \times L^s(\Sigma) \to Y := W(0,T) \times C(\bar{Q})$, $(v,u) \mapsto y$. We generally assume that $r > N/2 + 1$ and $s > N + 1$.

By virtue of Lemma 4.12 on page 202, the Nemytskii operators $y(\cdot) \mapsto d\big(\cdot,\cdot,y(\cdot)\big)$ and $y(\cdot) \mapsto b\big(\cdot,\cdot,y(\cdot)\big)$ are continuously differentiable from $C(\bar{Q})$ into $L^\infty(Q)$ and $L^\infty(\Sigma)$, respectively. By Theorem 5.5, the operator G assigns to each pair of controls $(v,u) \in V \times U$ a unique state $y \in Y$. As preparation for the proof of differentiability, we first show the Lipschitz continuity of G.

Theorem 5.8. *Under Assumption 5.6, for $r > N/2+1$ and $s > N+1$ the mapping G is Lipschitz continuous from $L^r(Q)\times L^s(\Sigma)$ into $W(0,T)\cap C(\bar{Q})$; that is, there exists some $L > 0$ such that*

$$\|y_1 - y_2\|_{W(0,T)} + \|y_1 - y_2\|_{C(\bar{Q})} \le L\left(\|v_1 - v_2\|_{L^r(Q)} + \|u_1 - u_2\|_{L^s(\Sigma)}\right)$$

for all $(v_i,u_i) \in L^r(Q)\times L^s(\Sigma)$ and associated states $y_i = G(v_i,u_i)$, $i = 1,2$.

Proof: Owing to Theorem 5.5, $y_i \in C(\bar{Q})$ for $i = 1,2$. Subtracting the initial-boundary value problems corresponding to y_1 and y_2, we find that the differences $y = y_1 - y_2$, $u = u_1 - u_2$, and $v = v_1 - v_2$ satisfy the problem

$$\begin{aligned} y_t - \Delta y + d(x,t,y_1) - d(x,t,y_2) &= v \\ \partial_\nu y + b(x,t,y_1) - b(x,t,y_2) &= u \\ y(0) &= 0. \end{aligned} \tag{5.12}$$

By the fundamental theorem of calculus, we have for $y_1, y_2 \in \mathbb{R}$ the identity

$$d(x,t,y_1) - d(x,t,y_2) = \Big(\int_0^1 d_y\big(x,t,y_2 + s\,(y_1 - y_2)\big)\,ds\Big)\,(y_1 - y_2).$$

Now, we put $y_i = y_i(x,t)$, $i = 1,2$, in this identity. Since d_y is nonnegative, the above integral becomes a nonnegative function $\delta = \delta(x,t) \in L^\infty(Q)$. An analogous representation holds for b with an integral term $\beta = \beta(x,t) \ge 0$. We thus have the initial-boundary value problem

$$\begin{aligned} y_t - \Delta y + \delta(x,t)\,y &= v \\ \partial_\nu y + \beta(x,t)\,y &= u \\ y(0) &= 0. \end{aligned}$$

Note that δ and β depend on y_1 and y_2, but this is immaterial for our arguments. In fact, in view of the boundedness and nonnegativity of δ and β, the functions $\tilde{d}(x,t,y) := \delta(x,t)\,y$ and $\tilde{b}(x,t,y) := \beta(x,t)\,y$ are increasing in y and vanish at $y = 0$. By virtue of Theorem 5.5 on page 268, the solution y is unique, and its norm does not depend on δ or β. Since $\tilde{d}(x,t,0) = \tilde{b}(x,t,0) = 0$, it follows from the estimate (5.6) on page 268 that the asserted estimate

$$\|y\|_{W(0,T)} + \|y\|_{C(\bar{Q})} \le L\left(\|v\|_{L^r(Q)} + \|u\|_{L^s(\Sigma)}\right)$$

is valid. This concludes the proof of the assertion. □

For the proof of differentiability, we consider a fixed $(\bar v, \bar u)$, which in the applications will be a locally optimal pair of controls.

Theorem 5.9. *Suppose that Assumption 5.6 on page 269 holds. Then for $r > N/2 + 1$ and $s > N + 1$ the control-to-state operator G is Fréchet differentiable from $L^r(Q) \times L^s(\Sigma)$ into $W(0,T) \cap C(\bar Q)$. The directional derivative in the direction (v,u) is given by*

$$G'(\bar v, \bar u)\,(v,u) = y,$$

where, with the state $\bar y = G(\bar v, \bar u)$ corresponding to $(\bar v, \bar u)$, y is the weak solution to the initial-boundary value problem linearized at $\bar y$:

$$\boxed{\begin{array}{rcll} y_t - \Delta y + d_y(x,t,\bar y)\,y &=& v & \text{in } Q \\ \partial_\nu y + b_y(x,t,\bar y)\,y &=& u & \text{on } \Sigma \\ y(0) &=& 0 & \text{in } \Omega. \end{array}} \tag{5.13}$$

Proof: We subtract the parabolic problem solved by $\bar y = G(\bar v, \bar u)$ from the one solved by $\tilde y = G(\bar v + v, \bar u + u)$ to obtain the initial-boundary value problem

$$\begin{aligned} (\tilde y - \bar y)_t - \Delta(\tilde y - \bar y) + d(x,t,\tilde y) - d(x,t,\bar y) &= v \\ \partial_\nu(\tilde y - \bar y) + b(x,t,\tilde y) - b(x,t,\bar y) &= u \\ (\tilde y - \bar y)(0) &= 0. \end{aligned}$$

The Nemytskii operators $\Phi : y \mapsto d\big(\cdot,\cdot,y(\cdot)\big)$ and $\Psi : y \mapsto b\big(\cdot,\cdot,y(\cdot)\big)$ are, by Lemma 4.12 on page 202, Fréchet differentiable in $L^\infty(Q)$ and $L^\infty(\Sigma)$, respectively. Therefore,

$$\begin{aligned} \Phi(\tilde y) - \Phi(\bar y) &= d_y\big(\cdot,\cdot,\bar y(\cdot)\big)\,\big(\tilde y(\cdot) - \bar y(\cdot)\big) + r_d, \\ \Psi(\tilde y) - \Psi(\bar y) &= b_y\big(\cdot,\cdot,\bar y(\cdot)\big)\,\big(\tilde y(\cdot) - \bar y(\cdot)\big) + r_b, \end{aligned}$$

with remainders r_d, r_b satisfying

$$\begin{aligned} \|r_d\|_{L^\infty(Q)} / \|\tilde y - \bar y\|_{L^\infty(Q)} \to 0 \quad &\text{as} \quad \|\tilde y - \bar y\|_{L^\infty(Q)} \to 0, \\ \|r_b\|_{L^\infty(\Sigma)} / \|\tilde y - \bar y\|_{L^\infty(\Sigma)} \to 0 \quad &\text{as} \quad \|\tilde y - \bar y\|_{L^\infty(\Sigma)} \to 0. \end{aligned}$$

We now write $\tilde y - \bar y$ with a remainder y_ρ in the form

$$\tilde y - \bar y = y + y_\rho,$$

where y is defined as in (5.13). The remainder y_ρ then solves the initial-boundary value problem

$$\begin{aligned} y_{\rho,t} - \Delta y_\rho + d_y(\cdot,\cdot,\bar{y})\, y_\rho &= -r_d \\ \partial_\nu y_\rho + b_y(\cdot,\cdot,\bar{y})\, y_\rho &= -r_b \\ y_\rho(0) &= 0. \end{aligned}$$

At this point, the Lipschitz continuity just shown comes into play: in fact, we have

$$\|\tilde{y} - \bar{y}\|_{C(\bar{Q})} \le L\, \|(v,u)\|_{L^r(Q)\times L^s(\Sigma)} \to 0.$$

The proof can now be concluded in a similar way as the proof of Theorem 4.17 on page 213. □

Conclusion. *G is Fréchet differentiable as a mapping from $L^\infty(Q)\times L^\infty(\Sigma)$ into $W(0,T)\cap C(\bar{Q})$.*

Remark. The preceding proof could also have been carried out with the aid of the implicit function theorem, without reference to the Lipschitz continuity of G. This technique will be applied in the proof of Theorem 5.15. However, the above argumentation is less abstract and not much longer, and it even yields the form of the derivative G'. Moreover, it is interesting to know that G is uniformly Lipschitz continuous.

Nonlinear controls. In some applications, the physical background leads to controls that occur nonlinearly. This is, for instance, the case for the heat conduction problem with Stefan–Boltzmann boundary condition,

$$\boxed{\begin{aligned} y_t - \Delta y &= 0 \\ \partial_\nu y + \beta(x,t)\,|y|\,y^3 &= u^4 \\ y(0) &= y_0. \end{aligned}} \tag{5.14}$$

Here, the mapping G is composed of the Nemytskii operator $u(\cdot) \mapsto u(\cdot)^4$ and the solution operator that assigns the solution to the semilinear initial-boundary value problem to the function $\tilde{u} := u(\cdot)^4$. The mapping $u(\cdot) \mapsto u(\cdot)^4$ is by Lemma 4.13 on page 203 Fréchet differentiable in $L^\infty(\Sigma)$ and, a fortiori, from $L^\infty(\Sigma)$ into $L^r(\Sigma)$, for any $r \ge 1$. Hence, the composition $G : u \mapsto y$ is Fréchet differentiable from $L^\infty(\Sigma)$ into $W(0,T)\cap C(\bar{Q})$.

If one wants G to be differentiable with the domain of definition $L^{\tilde{r}}(\Sigma)$ in place of $L^\infty(\Sigma)$, then $\tilde{r} > r > N/2+1$ has to be chosen sufficiently large so as to guarantee that the mapping $u \mapsto u^4$ is differentiable from $L^{\tilde{r}}(\Sigma)$ into $L^r(\Sigma)$. We do not enter into details here, because then growth conditions like those in Section 4.3.3 would be necessary. Differentiability is more easily obtained with controls belonging to $L^\infty(\Sigma)$. In particular, this is true for the treatment of the two-norm discrepancy in connection with second-order sufficient optimality conditions.

Special cases. For the sake of brevity, we have so far investigated the properties of G simultaneously for distributed and boundary controls. We now give an example for each of these cases in which the required conditions are fulfilled.

Distributed control. Consider the initial-boundary value problem

$$\boxed{\begin{aligned} y_t - \Delta y + d_0(x,t) + d_1(x,t)\, y^3 &= v \\ \partial_\nu y + b_0(x,t) + b_1(x,t)\, y &= 0 \\ y(0) &= 0. \end{aligned}} \tag{5.15}$$

Here, we assume $d_0 \in L^r(Q)$, $b_0 \in L^s(\Sigma)$, and that almost-everywhere nonnegative functions $d_1 \in L^\infty(Q)$ and $b_1 \in L^\infty(\Sigma)$ are given. Then $d(x,t,y) := d_0(x,t) + d_1(x,t)\, y^3$ and $b(x,t,y) := b_0(x,t) + b_1(x,t)\, y$ satisfy the assumptions of the above theorem. Hence, the mapping $v \mapsto y$ is for $r > N/2+1$ Fréchet differentiable from $L^r(Q)$ into $W(0,T) \cap C(\bar{Q})$.

Boundary control. Under analogous assumptions, the mapping $G : u \mapsto y$ for the problem with radiation boundary condition,

$$\boxed{\begin{aligned} y_t - \Delta y + d_0(x,t) + d_1(x,t)\, y &= 0 \\ \partial_\nu y + b_0(x,t) + b_1(x,t)\, |y|\, y^3 &= u \\ y(0) &= 0, \end{aligned}} \tag{5.16}$$

is continuously Fréchet differentiable from $L^s(\Sigma)$ into $W(0,T) \cap C(\bar{Q})$ whenever $s > N+1$.

5.5. Necessary optimality conditions

Once more, we consider the problem (5.7)–(5.9). We aim to derive the first-order necessary conditions for a locally optimal pair $(\bar{v}, \bar{u})$. It is clear that if we keep $u = \bar{u}$ fixed, then $\bar{v}$ must obey the necessary conditions for the

distributed control problem with variable v. Similarly, if $\bar{v}$ is kept fixed, then $\bar{u}$ satisfies the necessary conditions for the corresponding problem with boundary control u. We can therefore investigate the cases of distributed and boundary control separately at first.

5.5.1. Distributed control. Consider the optimal control problem

(5.17)
$$\min J(y,v) := \int_\Omega \phi\big(x,y(x,T)\big)\,dx + \iint_Q \varphi\big(x,t,y(x,t),v(x,t)\big)\,dx\,dt + \iint_\Sigma \psi\big(x,t,y(x,t)\big)\,ds\,dt,$$

subject to

(5.18)
$$\boxed{\begin{array}{rcll} y_t - \Delta y + d(x,t,y) &=& v & \text{in } Q \\ \partial_\nu y + b(x,t,y) &=& 0 & \text{on } \Sigma \\ y(0) &=& y_0 & \text{in } \Omega \end{array}}$$

and

(5.19)
$$v_a(x,t) \le v(x,t) \le v_b(x,t) \quad \text{for a.e. } (x,t) \in Q.$$

We have $y = y(v) = G(v)$ with the control-to-state operator $G : L^\infty(Q) \to W(0,T) \cap C(\bar{Q})$. Substituting this into J, we obtain the reduced cost functional f,

$$J(y,v) = J\big(G(v),v\big) =: f(v).$$

Under Assumption 5.6 on page 269, f is Fréchet differentiable in $L^\infty(Q)$, since J (by Lemma 4.12 on page 202) and G (by Theorem 5.9) are both differentiable.

Obviously, V_{ad} is convex. Hence, if $\bar{v}$ is locally optimal and $v \in V_{ad}$ is arbitrary, then we have for every sufficiently small $\lambda > 0$ the inequality

$$f\big(\bar{v} + \lambda(v - \bar{v})\big) - f(\bar{v}) \ge 0.$$

Dividing by λ and passing to the limit as $\lambda \downarrow 0$, we arrive, as in the elliptic case, at the following result.

Lemma 5.10. *Suppose that Assumption 5.6 on page 269 holds. Then every locally optimal control $\bar{v}$ for the problem* (5.17)–(5.19) *satisfies the variational inequality*

(5.20)
$$f'(\bar{v})(v - \bar{v}) \ge 0 \quad \forall\, v \in V_{ad}.$$

The derivative f' can be calculated using the chain rule. We obtain

$$
\begin{aligned}
(5.21)\qquad f'(\bar v)(v-\bar v) \;&=\; J_y(\bar y,\bar v)\,G'(\bar v)(v-\bar v) + J_v(\bar y,\bar v)(v-\bar v)\\
&= \int_\Omega \phi_y\big(x,\bar y(x,T)\big)\,y(x,T)\,dx\\
&\quad + \iint_Q \varphi_y\big(x,t,\bar y(x,t),\bar v(x,t)\big)\,y(x,t)\,dx\,dt\\
&\quad + \iint_\Sigma \psi_y\big(x,t,\bar y(x,t)\big)\,y(x,t)\,ds\,dt\\
&\quad + \iint_Q \varphi_v\big(x,t,\bar y(x,t),\bar v(x,t)\big)\,\big(v(x,t)-\bar v(x,t)\big)\,dx\,dt,
\end{aligned}
$$

where $y = G'(\bar v)(v-\bar v)$ is by Theorem 5.9 the solution to the linearized problem

$$
\begin{aligned}
y_t - \Delta y + d_y(x,t,\bar y)\,y &= v-\bar v\\
\partial_\nu y + b_y(x,t,\bar y)\,y &= 0\\
y(0) &= 0.
\end{aligned}
\qquad (5.22)
$$

As before, y can be eliminated from (5.21) by means of an adjoint state $p = p(x,t)$. Guided by the experience gained in the treatment of the elliptic case, we define the adjoint state as the solution to the *adjoint problem*

$$
\boxed{
\begin{aligned}
-p_t - \Delta p + d_y(x,t,\bar y)\,p &= \varphi_y(x,t,\bar y,\bar v)\\
\partial_\nu p + b_y(x,t,\bar y)\,p &= \psi_y(x,t,\bar y)\\
p(x,T) &= \phi_y\big(x,\bar y(x,T)\big),
\end{aligned}
}
\qquad (5.23)
$$

which belongs to $W(0,T)\cap L^\infty(Q)\cap C([0,T),C(\bar\Omega))$. In fact, the existence of a unique solution $p\in W(0,T)$ follows from Lemma 3.17 on page 157; the higher regularity of p is a consequence of Lemma 7.12 on page 378 and the subsequent remark on the L^∞ case. For this, the substitution $\tau := T-t$ has to be made. Moreover, if $\phi_y(x,y)$ is continuous in $\bar\Omega\times\mathbb{R}$, then the function $x\mapsto \phi_y\big(x,\bar y(x,T)\big)$ is also continuous in $\bar\Omega$. In this case, we even have $p\in W(0,T)\cap C(\bar Q)$.

Lemma 5.11. *Let y be the weak solution to the linearized problem* (5.22)*, and let p be the weak solution to* (5.23)*. Then we have, for all $v \in L^2(Q)$,*

$$\begin{aligned}\int_\Omega \phi_y\big(x,\bar y(x,T)\big)\,y(x,T)\,dx + \iint_Q \varphi_y\big(x,t,\bar y(x,t),\bar v(x,t)\big)\,y(x,t)\,dx\,dt \\ + \iint_\Sigma \psi_y\big(x,t,\bar y(x,t)\big)\,y(x,t)\,ds(x)dt \\ = \iint_Q p(x,t)\,\big(v(x,t)-\bar v(x,t)\big)\,dx\,dt.\end{aligned}$$

Proof: The assertion follows from Theorem 3.18 on page 158 with the specifications $a_\Omega(x) = \phi_y\big(x,\bar y(x,T)\big)$, $a_Q(x,t) = \varphi_y\big(x,t,\ \bar y(x,t),\bar v(x,t)\big)$, and $a_\Sigma(x,t) = \psi_y\big(x,t,\bar y(x,t)\big)$. □

From formula (5.21), we can thus deduce the form of the derivative $f'(\bar v)$:

$$f'(\bar v)\,v = \iint_Q \big(p + \varphi_v(x,t,\bar y,\bar v)\big)\,v\,dx\,dt. \tag{5.24}$$

Moreover, we obtain the desired necessary optimality condition:

Theorem 5.12. *Suppose that Assumption 5.6 on page 269 holds, and let $\bar v$ be locally optimal for the problem* (5.17)–(5.19)*. If $p \in W(0,T)\cap L^\infty(Q)$ is the associated adjoint state solving problem* (5.23)*, then the variational inequality*

$$\iint_Q \big(p + \varphi_v(x,t,\bar y,\bar v)\big)(v-\bar v)\,dx\,dt \ge 0 \qquad \forall\, v \in V_{ad} \tag{5.25}$$

is satisfied.

As in the case of elliptic problems, the variational inequality can be reformulated in terms of a minimum principle.

Conclusion. *Suppose that the assumptions of Theorem 5.12 hold, let $\bar v$ be locally optimal for problem* (5.7)–(5.9)*, and let p be the associated adjoint state. Then the minimum of the minimization problem*

$$\min_{v_a(x,t)\le v\le v_b(x,t)} \big\{\big(p(x,t) + \varphi_v(x,t,\bar y(x,t),\bar v(x,t))\big)\,v\big\} \tag{5.26}$$

is for almost every $(x,t)\in Q$ attained at $v = \bar v(x,t)$.

Special case: Let $\varphi(x,t,y,v) := \varphi(x,t,y) + \dfrac{\lambda}{2}\,v^2$, with some $\lambda > 0$. Then $\varphi_v(x,t,y,v) = \lambda\,v$; hence, the minimum of the problem

$$\min_{v_a(x,t) \le\, v \,\le v_b(x,t)} \big\{ \big(p(x,t) + \lambda\,\bar v(x,t)\big)\,v \big\}$$

is for almost every $(x,t) \in Q$ attained at $v = \bar v(x,t)$. This implies, as in formula (2.58) on page 70, that in the $\lambda > 0$ case we have for almost every $(x,t) \in Q$ the projection formula

$$\boxed{\bar v(x,t) = \mathbb{P}_{[v_a(x,t),v_b(x,t)]} \left\{ -\frac{1}{\lambda}\,p(x,t) \right\}.}$$

If v_a and v_b are continuous, then this implies that $\bar v \in C(\bar Q)$ if p is continuous. Sufficient for this to be true is the continuity of $\phi_y(x,y)$; in this case, any locally optimal control $\bar v$ must be continuous.

Example. We now discuss the "superconductivity" problem, which was already treated in the stationary situation, in the time-dependent case and with a slightly different cost functional:

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

subject to

$$\boxed{\begin{aligned} y_t - \Delta y + y^3 &= v \\ \partial_\nu y + \beta(x,t)\,y &= 0 \\ y(\cdot,0) &= y_0 \end{aligned}}$$

and

$$-1 \le v(x,t) \le 1 \quad \text{for a.e. } (x,t) \in Q.$$

Evidently, the above problem is a special case of problem (5.7)–(5.9) on page 270, with the specifications

$$\begin{aligned} \phi(x,y) &= \frac{1}{2}\big(y - y_\Omega(x)\big)^2, \quad \varphi(x,t,y,v) = \frac{\lambda}{2}\,v^2, \\ \psi(x,t,y) &= \frac{1}{2}\big(y - y_\Sigma(x,t)\big)^2, \quad d(x,t,y) = y^3, \quad b(x,t,y) = \beta(x,t)\,y. \end{aligned}$$

We assume that $y_0 \in C(\bar\Omega)$, $\beta \in L^\infty(\Sigma)$ with $\beta \ge 0$, $y_\Omega \in C(\bar\Omega)$, and $y_\Sigma \in L^\infty(\Sigma)$. Then all the required conditions (measurability with respect to (x,t), boundedness, differentiability, monotonicity of d, convexity of φ in v) are met, and Theorem 5.7 yields the existence of at least one (globally)

optimal control $\bar v$. The adjoint system solved by the associated adjoint state $p \in W(0,T) \cap C(\bar Q)$ reads

$$\begin{aligned} -p_t - \Delta p + 3\,\bar y^2\, p &= 0 \\ \partial_\nu p + \beta\, p &= \bar y - y_\Sigma \\ p(\cdot, T) &= \bar y(\cdot, T) - y_\Omega. \end{aligned}$$

Evidently, we have the variational inequality

$$\iint_Q (\lambda \bar v + p)(v - \bar v)\, dx\, dt \geq 0 \quad \forall\, v \in V_{ad},$$

from which, for $\lambda > 0$, the usual projection relation and the property $\bar v \in C(\bar Q)$ follow. For $\lambda = 0$, we have $\bar v(x,t) = -\operatorname{sign} p(x,t)$. ◇

5.5.2. Boundary control. The necessary optimality conditions for the corresponding boundary control problem are derived by similar reasoning. We consider the problem

$$\begin{aligned} (5.27) \quad \min J(y,u) \;:=\; & \int_\Omega \phi\big(x, y(x,T)\big)\, dx + \iint_Q \varphi\big(x,t,y(x,t)\big)\, dx\, dt \\ & + \iint_\Sigma \psi\big(x,t,y(x,t),u(x,t)\big)\, ds\, dt, \end{aligned}$$

subject to

$$(5.28) \qquad \boxed{\begin{aligned} y_t - \Delta y + d(x,t,y) &= 0 && \text{in } Q \\ \partial_\nu y + b(x,t,y) &= u && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega \end{aligned}}$$

and

$$(5.29) \qquad u_a(x,t) \leq u(x,t) \leq u_b(x,t) \quad \text{for a.e. } (x,t) \in \Sigma.$$

Then the control-to-state operator $G = G(u) : u \mapsto y(u)$ maps $L^\infty(\Sigma)$ into $W(0,T) \cap C(\bar Q)$, since $y_0 \in C(\bar\Omega)$. We may now follow the lines of argumentation employed in the case of distributed controls. The directional derivative of the reduced functional $f(u) = J\big(G(u), u\big)$ at $\bar u$ in the direction u is given by

$$(5.30) \qquad f'(\bar u)\, u = \iint_\Sigma \big(p + \psi_u(x,t,\bar y,\bar u)\big)\, u\, ds\, dt,$$

where $p \in W(0,T) \cap L^\infty(Q)$ is the solution to the adjoint problem

$$\begin{aligned} -p_t - \Delta p + d_y(x,t,\bar{y})\, p &= \varphi_y(x,t,\bar{y}) \\ \partial_\nu p + b_y(x,t,\bar{y})\, p &= \psi_y(x,t,\bar{y},\bar{u}) \\ p(x,T) &= \phi_y(x,\bar{y}(x,T)). \end{aligned} \tag{5.31}$$

In analogy to Theorem 5.12, we obtain the following result.

Theorem 5.13. *Suppose that Assumption 5.6 on page 269 holds. Let $\bar{u}$ be a locally optimal control for the boundary control problem* (5.27)–(5.29)*, and let $p \in W(0,T) \cap L^\infty(Q)$ be the associated adjoint state solving problem* (5.31)*. Then the variational inequality*

$$\iint_\Sigma \big(p + \psi_u(x,t,\bar{y},\bar{u})\big)(u-\bar{u})\, ds\, dt \geq 0 \qquad \forall u \in U_{ad} \tag{5.32}$$

is satisfied. Moreover, the minimum of the minimization problem

$$\min_{u_a(x,t) \leq u \leq u_b(x,t)} \big\{ \big(p(x,t) + \psi_u(x,t,\bar{y}(x,t),\bar{u}(x,t)\big)\, u \big\} \tag{5.33}$$

is for almost all $(x,t) \in \Sigma$ attained at $u = \bar{u}(x,t)$.

Example. Consider the problem

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

subject to

$$\boxed{\begin{aligned} y_t - \Delta y &= 0 \\ \partial_\nu y + y^3|y| &= u \\ y(0) &= y_0 \end{aligned}}$$

and

$$0 \leq u(x,t) \leq 1.$$

This is a special case of the above problem with the specifications

$$\phi(x,y) = \frac{1}{2}\,(y - y_\Omega(x))^2, \quad \varphi(x,t,y) = \frac{1}{2}\,(y - y_Q(x,t))^2,$$
$$\psi(x,t,y,u) = \frac{\lambda}{2}\, u^2, \quad b(x,t,y) = y^3|y|.$$

The boundary condition is of Stefan–Boltzmann type, since $y^3|y| = y^4$ for $y \geq 0$; we choose this form because the mapping $y \mapsto y^3|y|$ is increasing while $y \mapsto y^4$ is not. Here we also assume y_Ω, $y_0 \in C(\bar{\Omega})$. Then Assumption

5.6 holds, and ψ is convex with respect to u. We can thus infer from Theorem 5.7 that there exists at least one optimal control $\bar{u}$.

Since $b_y(y) = 4\,y^2\,|y|$, we obtain for the adjoint problem

$$\begin{aligned} -p_t - \Delta p &= \bar{y} - y_Q \\ \partial_\nu p + 4\,\bar{y}^2\,|\bar{y}|\,p &= 0 \\ p(\cdot,T) &= \bar{y}(\cdot,T) - y_\Omega. \end{aligned}$$

Together with its solution $p \in W(0,T) \cap C(\bar{Q})$, $\bar{u}$ satisfies the variational inequality

$$\iint_\Sigma (\lambda\bar{u} + p)(u - \bar{u})\,ds\,dt \geq 0 \quad \forall u \in U_{ad}.$$

Moreover, we have for $\lambda > 0$ the projection relation

$$\bar{u}(x,t) = \mathbb{P}_{[0,1]}\left\{-\frac{1}{\lambda}\,p(x,t)\right\} \quad \text{for a.e. } (x,t) \in \Sigma.$$

Since p is continuous, we can conclude that $\bar{u}$ is also continuous in $\bar{\Sigma}$. $\diamond$

The general case. Combining the results established for the cases of distributed and boundary control, we are finally in a position to state the necessary optimality condition for the general case. Here, the associated adjoint state p is the solution to the following adjoint problem:

$$\boxed{\begin{aligned} -p_t - \Delta p + d_y(x,t,\bar{y})\,p &= \varphi_y(x,t,\bar{y},\bar{v}) && \text{in } Q \\ \partial_\nu p + b_y(x,t,\bar{y})\,p &= \psi_y(x,t,\bar{y},\bar{u}) && \text{on } \Sigma \\ p(x,T) &= \phi_y\big(x,\bar{y}(x,T)\big) && \text{in } \Omega. \end{aligned}} \tag{5.34}$$

Theorem 5.14. *Suppose that Assumption 5.6 on page 269 is satisfied, let* $(\bar{v},\bar{u})$ *be a locally optimal pair for the problem* (5.7)–(5.9) *on page 270, and let* $p \in W(0,T) \cap L^\infty(Q)$ *be the associated adjoint state solving problem* (5.34). *Then the variational inequalities* (5.25) and (5.32) *and the minimum conditions* (5.26) and (5.33) *are satisfied.*

Proof: The necessary condition for v follows from the fact that $\bar{v}$ has to solve the general problem (5.7)–(5.9) for fixed $u = \bar{u}$, which in turn is a special case of the distributed control problem (5.17)–(5.19) if we put $b(x,t,y) := b(x,t,y) - \bar{u}(x,t)$. Similarly, $\bar{u}$ must solve the general problem for fixed $\bar{v}$. In this way, we obtain the full optimality system. $\square$

Remark. It ought to be clear that for all the first-order necessary optimality conditions established in this chapter, only the boundedness and Lipschitz conditions of order $k = 1$ from Assumption 5.6 are needed.

5.6. Pontryagin's maximum principle *

We will discuss Pontryagin's maximum principle for the following optimal control problem in which the control functions appear nonlinearly:

(5.35)
$$\begin{aligned}\min J(y,v,u) \;:=\; & \int_\Omega \phi\big(x,y(x,T)\big)\,dx + \iint_Q \varphi\big(x,t,y(x,t),v(x,t)\big)\,dx\,dt \\ & + \iint_\Sigma \psi\big(x,t,y(x,t),u(x,t)\big)\,ds(x)dt,\end{aligned}$$

subject to

(5.36)
$$\boxed{\begin{array}{rcll} y_t - \Delta y + d(x,t,y,v) &=& 0 & \text{in } Q \\ \partial_\nu y + b(x,t,y,u) &=& 0 & \text{on } \Sigma \\ y(0) &=& y_0 & \text{in } \Omega \end{array}}$$

and

(5.37)
$$\begin{aligned} v_a(x,t) \le v(x,t) \le v_b(x,t) \quad &\text{for a.e. } (x,t) \in Q \\ u_a(x,t) \le u(x,t) \le u_b(x,t) \quad &\text{for a.e. } (x,t) \in \Sigma. \end{aligned}$$

In analogy to the elliptic case, one introduces Hamiltonian functions.

Definition. *The functions* $H^Q : Q \times \mathbb{R}^4 \to \mathbb{R}$, $H^\Sigma : \Sigma \times \mathbb{R}^4 \to \mathbb{R}$, *and* $H^\Omega : \Omega \times \mathbb{R}^3 \to \mathbb{R}$, *given by*

$$\begin{aligned} H^Q(x,t,y,v,q_0,q) &= q_0\,\varphi(x,t,y,v) - q\,d(x,t,y,v) \\ H^\Sigma(x,t,y,u,q_0,q) &= q_0\,\psi(x,t,y,u) - q\,b(x,t,y,u) \\ H^\Omega(x,y,q_0,q) &= q_0\,\phi(x,y), \end{aligned}$$

are called Hamiltonian functions.

We define the adjoint state q as the solution to the *adjoint system*

(5.38)
$$\begin{aligned} -q_t - \Delta q + d_y(x,t,\bar y,\bar v)\,q &= q_0\,\varphi_y(x,t,\bar y,\bar v) \\ \partial_\nu q + b_y(x,t,\bar y,\bar u)\,q &= q_0\,\psi_y(x,t,\bar y,\bar u) \\ q(x,T) &= q_0\,\phi_y(x,\bar y(\cdot,T)). \end{aligned}$$

We have $q = -p$ with the solution p to (5.34), provided we put $q_0 = -1$. Using the above Hamiltonians, and putting $q_0 := -1$, we may rewrite the

adjoint system in the form

$$\boxed{\begin{aligned} -q_t - \Delta q &= D_y H^Q(x,t,\bar y,\bar v,-1,q) && \text{in } Q \\ \partial_\nu q &= D_y H^\Sigma(x,t,\bar y,\bar u,-1,q) && \text{on } \Sigma \\ q(\cdot,T) &= D_y H^\Omega(x,\bar y(\cdot,T),-1,q) && \text{in } \Omega. \end{aligned}} \tag{5.39}$$

Definition. *Let q be the adjoint state defined in* (5.39) *and let $q_0 = -1$. The controls $\bar v$ and $\bar u$ are said to satisfy* Pontryagin's maximum principle *if the maximum conditions*

$$\begin{aligned} &\max_{v_a(x,t)\le v\le v_b(x,t)} \big\{H^Q\big(x,t,\bar y(x,t),v,q_0,q(x,t)\big)\big\} \\ &\quad = H^Q\big(x,t,\bar y(x,t),\bar v(x,t),q_0,q(x,t)\big), \\ &\max_{u_a(x,t)\le u\le u_b(x,t)} \big\{H^\Sigma\big(x,t,\bar y(x,t),u,q_0,q(x,t)\big)\big\} \\ &\quad = H^\Sigma\big(x,t,\bar y(x,t),\bar u(x,t),q_0,q(x,t)\big) \end{aligned}$$

are satisfied for almost every $(x,t)\in Q$ and $(x,t)\in\Sigma$, respectively.

In other words, the maxima of H^Q and H^Σ must for almost every (x,t) be attained at $\bar v(x,t)$ and $\bar u(x,t)$, respectively. Under natural assumptions, it can be expected that (globally) optimal controls satisfy the maximum principle; see, e.g., the papers **[Cas97]**, **[LY95]**, **[RZ99]**, **[vW76]**, and **[vW77]** listed in the overview of relevant literature on the maximum principle at the beginning of Section 4.8.1.

5.7. Second-order optimality conditions

5.7.1. Second-order derivatives. We again consider the optimal control problem (5.7)–(5.9) on page 270. We have the following result.

Theorem 5.15. *Suppose that Assumption 5.6 holds. Then the control-to-state mapping $G:(v,u)\mapsto y$ associated with the initial-boundary value problem* (5.8) *is twice continuously Fréchet differentiable from $L^\infty(Q)\times L^\infty(\Sigma)$ into $W(0,T)\times C(\bar Q)$.*

Proof. As in the elliptic case, we apply the implicit function theorem. We first derive an operator equation for $y = G(v,u)$. To this end, we reformulate

the equation for y in the form

$$\begin{aligned} y_t - \Delta y &= v - d(x,t,y) && \text{in } Q \\ \partial_\nu y &= u - b(x,t,y) && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega. \end{aligned}$$

The linear part on the left-hand side is decomposed into three continuous linear solution operators $G_Q : L^\infty(Q) \to Y := W(0,T) \cap C(\bar{Q})$, $G_\Sigma : L^\infty(\Sigma) \to Y$, and $G_0 : C(\bar{\Omega}) \to Y$, which correspond to the linear initial-boundary value problem

$$\begin{aligned} y_t - \Delta y &= v && \text{in } Q \\ \partial_\nu y &= u && \text{on } \Sigma \\ y(0) &= w && \text{in } \Omega \end{aligned}$$

in the following way: we have

$$\begin{aligned} y = G_Q\, v \quad &\text{for } u = 0,\ w = 0 \\ y = G_\Sigma\, u \quad &\text{for } v = 0,\ w = 0 \\ y = G_0\, w \quad &\text{for } u = 0,\ v = 0. \end{aligned}$$

In the following, we regard these operators as mappings with range in $C(\bar{Q})$. It is apparent that the solution y to the nonlinear equation can be expressed in the form

$$y = G_Q\big(v - d(\cdot, y)\big) + G_\Sigma\,\big(u - b(\cdot, y)\big) + G_0\, y_0 \tag{5.40}$$

or, equivalently,

$$0 = y - G_Q\big(v - d(\cdot, y)\big) - G_\Sigma\,\big(u - b(\cdot, y)\big) - G_0\, y_0 =: F(y, v, u).$$

In this way, we avoid the discussion of differential operators and the use of the space $W(0,T)$. Evidently, F is a twice continuously Fréchet differentiable mapping from $C(\bar{Q}) \times L^\infty(Q) \times L^\infty(\Sigma)$ into $C(\bar{Q})$; indeed, G_Q, G_Σ, and G_0 are continuous linear mappings, and the Nemytskii operators $y \mapsto d(\cdot, y)$ and $y \mapsto b(\cdot, y)$ are twice continuously Fréchet differentiable from $C(\bar{Q})$ into $L^\infty(Q)$ and $L^\infty(\Sigma)$, respectively.

According to Exercise 5.1, the partial Fréchet derivative $F_y(y, v, u)$ is invertible in $C(\bar{Q})$. Hence, by the implicit function theorem, the equation $F(y, v, u) = 0$ has, in some open neighborhood of any arbitrarily chosen point $(\bar{y}, \bar{v}, \bar{u})$, a unique solution $y = y(v, u)$; moreover, the mapping $(v, u) \mapsto y$ is twice continuously differentiable. Since $y = G(v, u)$ is a solution, we conclude that G is twice continuously differentiable. □

Theorem 5.16. *Suppose that Assumption 5.6 on page 269 holds. Then the second derivative of G at (v,u) is given by the expression*

$$G''(v,u)\big[(v_1,u_1),(v_2,u_2)\big] = z,$$

with z being the uniquely determined weak solution to the parabolic initial-boundary value problem

$$\begin{aligned} z_t - \Delta z + d_y(x,t,y)\,z &= -d_{yy}(x,t,y)\left\{y_1y_2 + y_1w_2 + w_1y_2 + w_1w_2\right\} \\ \partial_\nu z + b_y(x,t,y)\,z &= -b_{yy}(x,t,y)\left\{y_1y_2 + y_1w_2 + w_1y_2 + w_1w_2\right\} \\ z(0) &= 0 \end{aligned}$$

with $y = G(v,u)$, where the functions y_i, $w_i \in W(0,T)$, $i = 1,2$, are the solutions to the following linearized initial-boundary value problems:

$$\begin{aligned} \partial y_i/\partial t - \Delta y_i + d_y(x,t,y)\,y_i &= 0 \\ \partial_\nu y_i + b_y(x,t,y)\,y_i &= u_i \\ y_i(0) &= 0, \end{aligned}$$

$$\begin{aligned} \partial w_i/\partial t - \Delta w_i + d_y(x,t,y)\,w_i &= v_i \\ \partial_\nu w_i + b_y(x,t,y)\,w_i &= 0 \\ w_i(0) &= 0. \end{aligned}$$

Proof: The mapping $G''(v,u)$ is given in terms of second-order partial derivatives in the form

$$\begin{aligned} G''(v,u)\big[(u_1,v_1),(u_2,v_2)\big] &= G_{uu}[u_1,u_2] + G_{uv}[u_1,v_2] + G_{vu}[v_1,u_2] \\ &\quad + G_{vv}[v_1,v_2], \end{aligned}$$

where the operators G_{uu}, G_{uv}, G_{vu}, and G_{vv} can be determined by suitable combinations of the directions u_i and v_i. For example, by choosing $v_1 = v_2 = 0$, we calculate G_{uu}. As in the preceding proof, we start from the representation (5.40),

$$y = G(v,u) = G_Q\left(v - d(\cdot,\cdot,G(v,u))\right) + G_\Sigma\left(u - b(\cdot,\cdot,G(v,u))\right) + G_0\,y_0.$$

Differentiating the left- and right-hand sides with respect to u in the direction u_1, we obtain that

$$\begin{aligned} G_u(v,u)\,u_1 &= -G_Q\,d_y\big(\cdot,\cdot,G(v,u)\big)\,G_u(v,u)\,u_1 + G_\Sigma\,u_1 \\ &\quad - G_\Sigma\,b_y\big(\cdot,\cdot,G(v,u)\big)\,G_u(v,u)\,u_1. \end{aligned}$$

A further differentiation in the direction u_2 yields the equation

$$\begin{aligned} G_{uu}(v,u)[u_1,u_2] &= -G_Q\big\{d_{yy}\big(\cdot,\cdot,G(v,u)\big)\,\big(G_u(v,u)\,u_1\big)\,\big(G_u(v,u)\,u_2\big) \\ &\qquad + d_y\big(\cdot,\cdot,G(v,u)\big)\,G_{uu}(v,u)[u_1,u_2]\big\} \\ &\quad -G_\Sigma\big\{b_{yy}\big(\cdot,\cdot,G(v,u)\big)\,\big(G_u(v,u)u_1\big)\,\big(G_u(v,u)u_2\big) \\ &\qquad + b_y\big(\cdot,\cdot,G(v,u)\big)\,G_{uu}(v,u)[u_1,u_2]\big\}. \end{aligned}$$

Putting $y = G(v,u)$, $y_i = G_u(v,u)\,u_i$, and $z_{uu} := G_{uu}(v,u)[u_1,u_2]$ in this equation, we find that

$$\begin{aligned} z_{uu} &= -G_Q\big\{d_{yy}(\cdot,\cdot,y)\,y_1y_2 \,+\, d_y(\cdot,\cdot,y)\,z_{uu}\big\} \\ &\quad -G_\Sigma\big\{b_{yy}(\cdot,\cdot,y)\,y_1y_2 + b_y(\cdot,\cdot,y)\,z_{uu}\big\}, \end{aligned}$$

where, by virtue of Theorem 5.9, y_1 and y_2 solve the initial-boundary value problems defined in the assertion. By the definitions of G_Q, G_Σ, and G_0, the function z_{uu} solves the initial-boundary value problem

$$\begin{aligned} z_t - \Delta z + d_y(x,t,y)\,z &= -d_{yy}(x,t,y)\,y_1y_2 \\ \partial_\nu\, z + b_y(x,t,y)\,z &= -b_{yy}(x,t,y)\,y_1y_2 \\ z(0) &= 0. \end{aligned}$$

We have thus obtained the first contribution z_{uu} to the representation of z asserted in the theorem. The remaining three contributions are constructed by the same procedure: $G_{uv}[u_1,v_2]$ with $u_2 = 0$ and $v_1 = 0$, $G_{vu}[v_1,u_2]$ using $u_1 = 0$ and $v_2 = 0$, and $G_{vv}[v_1,v_2]$ with $u_1 = u_2 = 0$. Superposition of the four contributions finally yields the function z from the statement of the theorem and the asserted form of G''. □

With the above theorem, the ground is prepared to state sufficient optimality conditions and to prove their sufficiency for local optimality. We could do this for distributed and boundary controls simultaneously, but this would necessitate a rather complicated exposition. Therefore, we choose to treat the two cases separately, giving a proof only for distributed controls. The case of boundary controls can be handled analogously.

5.7.2. Distributed controls. Once again, we consider the optimal control problem (5.17)–(5.19) on page 278. Let the control $\bar v \in V_{ad}$ satisfy the associated first-order necessary optimality conditions, and let p denote the corresponding adjoint state defined by (5.23). Then we have the variational

inequality (5.25),

$$\iint_Q \big(p + \varphi_v(x,t,\bar y,\bar v)\big)(v-\bar v)\,dx\,dt \ge 0 \quad \forall\, v \in V_{ad}.$$

Second-order necessary optimality conditions can be obtained in the same way as in the elliptic case; we do not discuss this further here. The most convenient way to state second-order sufficient conditions is to make use of the Lagrangian function, which for the problem (5.17)–(5.19) on page 278 takes the form

$$\begin{aligned}\mathcal{L}(y,v,p) &= J(y,v) - \iint_Q \big(\big(y_t + d(x,t,y) - v\big)\,p \;+\; \nabla y\cdot\nabla p\big)\,dx\,dt \\ &\quad - \iint_\Sigma b(x,t,y)\,p\,ds\,dt.\end{aligned}$$

The explicit expression for the second derivative of $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L}''(\bar y,\bar v,p)(y,v)^2 &= J''(\bar y,\bar v)(y,v)^2 - \iint_Q p\,d_{yy}(x,t,\bar y)\,y^2\,dx\,dt \\ &\quad - \iint_\Sigma p\,b_{yy}(x,t,\bar y)\,y^2\,ds\,dt,\end{aligned}$$

where

$$\begin{aligned}J''(\bar y,\bar v)(y,v)^2 &= \int_\Omega \phi_{yy}(x,\bar y(x,T))\,y(x,T)^2\,dx + \iint_\Sigma \psi_{yy}(x,t,\bar y)\,y^2\,ds\,dt \\ &\quad + \iint_Q \begin{bmatrix} y \\ v \end{bmatrix}^\top \begin{bmatrix} \varphi_{yy}(x,t,\bar y,\bar v) & \varphi_{yv}(x,t,\bar y,\bar v) \\ \varphi_{vy}(x,t,\bar y,\bar v) & \varphi_{vv}(x,t,\bar y,\bar v) \end{bmatrix} \begin{bmatrix} y \\ v \end{bmatrix} dx\,dt.\end{aligned}$$

For pedagogic reasons, we began our investigations in the elliptic case with sufficient optimality conditions that do not invoke strongly active constraints and thus are usually too restrictive; weaker conditions were studied only later. Here, we do not take this detour and incorporate strongly active constraints right from the beginning. Similarly to the elliptic case, we make the following definition:

Definition. *For given $\tau \ge 0$ the set*

$$A_\tau(\bar v) = \big\{(x,t) \in Q :\, \big|p(x,t) + \varphi_v\big(x,t,\bar y(x,t),\bar v(x,t)\big)\big| > \tau\big\}$$

is called the set of strongly active constraints *for $\bar v$.*

As was explained on page 250 for an optimization problem in $\mathbb{R}$, it does not make sense to postulate the positive definiteness of the second derivative

of the Lagrangian in $A_\tau(\bar v)$. The set $A_\tau(\bar v)$ is fundamental for the introduction of the τ-critical cone, which is defined just as in the elliptic case. It is the set of those controls for which the positive definiteness of $\mathcal{L}''$ holds upon their insertion into $\mathcal{L}''$ together with the solution y to the linearized initial-boundary value problem

$$\begin{aligned} y_t - \Delta y + d_y(x,t,\bar y)\, y &= v && \text{in } Q \\ \partial_\nu y + b_y(x,t,\bar y)\, y &= 0 && \text{on } \Sigma \\ y(0) &= 0 && \text{in } \Omega. \end{aligned} \tag{5.41}$$

Definition. *The* τ-critical cone $C_\tau(\bar v)$ *is the set of all* $v \in L^\infty(Q)$ *satisfying*

$$v(x,t) \begin{cases} = 0 & \text{if } (x,t) \in A_\tau(\bar v) \\ \geq 0 & \text{if } \bar v(x,t) = v_a \text{ and } (x,t) \notin A_\tau(\bar v) \\ \leq 0 & \text{if } \bar v(x,t) = v_b \text{ and } (x,t) \notin A_\tau(\bar v). \end{cases} \tag{5.42}$$

Depending on the sign of $p + \varphi_v$, on $A_\tau(\bar v)$ we have either $\bar v = v_a$ or $\bar v = v_b$, whence the above sign conditions are derived. One may put $v = 0$ at points where the gradient of the cost functional, that is, of the function $p + \varphi_v(x,t,\bar y,\bar v)$, is at least τ in absolute value. We must assume $\tau > 0$, a restriction that is not needed in the finite-dimensional case: indeed, in the finite-dimensional case every component of a vector lying in the critical cone for which the corresponding component of the gradient is nonzero must vanish. The counterexample constructed by J. Dunn [**Dun98**] shows that this does not hold in function spaces. Therefore, the second-order sufficient optimality condition is formulated as follows.

There exist constants $\delta > 0$ and $\tau > 0$ such that

$$\begin{aligned} &\mathcal{L}''(\bar y,\bar v,p)(y,v)^2 \geq \delta \,\|v\|^2_{L^2(Q)} \\ &\text{for all } v \in C_\tau(\bar v) \text{ and } y \in W(0,T) \text{ solving } (5.41). \end{aligned} \tag{5.43}$$

Theorem 5.17. *Suppose that Assumption 5.6 holds, and assume that the pair* $(\bar y,\bar v)$ *obeys the first-order necessary optimality conditions of Theorem 5.12 and all the constraints of the problem* (5.17)–(5.19)*. Moreover, suppose there exist constants* $\delta > 0$ *and* $\tau > 0$ *such that the positive definiteness condition* (5.43) *is fulfilled. Then there are constants* $\varepsilon > 0$ *and* $\sigma > 0$ *such that every* $v \in V_{ad}$ *with* $\|v - \bar v\|_{L^\infty(Q)} \leq \varepsilon$*, together with the associated solution* $y(v)$ *of problem* (5.18) *on page 278, satisfies the quadratic growth condition*

$$J(y,v) \geq J(\bar y,\bar v) + \sigma \,\|v - \bar v\|^2_{L^2(Q)}.$$

In particular, $\bar v$ is locally optimal in the sense of $L^\infty(Q)$.

Proof: (i) *Preliminaries.*
As before, $G : L^\infty(Q) \to W(0,T) \cap C(\bar Q)$, $v \mapsto y$, denotes the control-to-state operator. We know already that G is twice continuously Fréchet differentiable. Now let $f(v) := J\big(y(v), v\big) = J\big(G(v), v\big)$. In analogy to Theorem 4.25 on page 242, we again have

$$f''(\bar v)[v_1, v_2] = \mathcal{L}''(\bar y, \bar v, p)\big[(y_1, v_1), (y_2, v_2)\big], \tag{5.44}$$

where p is the adjoint state associated with $(\bar y, \bar v)$ and $y_i := G'(\bar v)v_i$, $i = 1, 2$, denote the solutions to the linearized problem with right-hand sides v_i. With this notation, $f''(\bar v)$ can be estimated in terms of the L^2 norms of the increments:

(5.45)
$$\begin{aligned} |f''(\bar v)[v_1, v_2]| &\le \big|\mathcal{L}''(\bar y, \bar v, p)\big[(y_1, v_1), (y_2, v_2)\big]\big| \\ &\le c\big\{\|y_1\|_{W(0,T)}\,\|y_2\|_{W(0,T)} + \|y_1\|_{W(0,T)}\,\|v_2\|_{L^2(Q)} \\ &\quad + \|y_2\|_{W(0,T)}\,\|v_1\|_{L^2(Q)} + \|v_1\|_{L^2(Q)}\,\|v_2\|_{L^2(Q)}\big\} \\ &\le c\,\|v_1\|_{L^2(Q)}\,\|v_2\|_{L^2(Q)}. \end{aligned}$$

Here, we have used the continuity of the operator $G'(\bar v)$ in the representation $y_i = G'(\bar v)\,v_i$ as a mapping from $L^2(Q)$ into $W(0,T)$, and $c > 0$ denotes a generic constant. The estimate just shown will be used repeatedly in step (iii) below.

Note also that $f'(\bar v)$ can be expressed, with $g := p + \varphi_v(\cdot, \cdot, \bar y, \bar v)$, in the form

$$f'(\bar v)h = \iint_Q g(x,t)\,h(x,t)\,dx\,dt.$$

(ii) *Taylor expansion.*
Let $v(\cdot) \in V_{ad}$ with $\|v - \bar v\|_{L^\infty(Q)} \le \varepsilon$ be arbitrary. For almost every $(x,t) \in Q$, we have the pointwise variational inequality

$$g(x,t)\big(v - \bar v(x,t)\big) \ge 0 \qquad \forall v \in \big[v_a(x,t), v_b(x,t)\big].$$

Hence, with $h(x,t) = v(x,t) - \bar{v}(x,t)$,

$$\begin{aligned} f(v) - f(\bar{v}) &= f'(\bar{v})\, h + \frac{1}{2} f''(\bar{v})\, h^2 + r_2^f \\ &\geq \iint_{A_\tau(\bar{v})} g(x,t)\, h(x,t)\, dx\, dt + \frac{1}{2} f''(\bar{v})\, h^2 + r_2^f \\ &\geq \tau \iint_{A_\tau(\bar{v})} |h(x,t)|\, dx\, dt + \frac{1}{2} f''(\bar{v})\, h^2 + r_2^f. \end{aligned}$$

Here, $r_2^f = r_2^f(\bar{v}, h)$ denotes the second-order remainder in the Taylor expansion of f. We now make the decomposition $h := h_0 + h_1$, where

$$h_0(x,t) := \begin{cases} h(x,t) & \text{if} \quad (x,t) \notin A_\tau \\ 0 & \text{if} \quad (x,t) \in A_\tau. \end{cases}$$

By construction, we have $h_0 \in C_\tau(\bar{v})$, since h_0 satisfies the sign conditions of the critical cone. With these functions, it follows that

$$f(v) - f(\bar{v}) \geq \tau \iint_{A_\tau(\bar{v})} |h(x,t)|\, dx\, dt + \frac{1}{2} f''(\bar{v})\, (h_0 + h_1)^2 + r_2^f. \tag{5.46}$$

(iii) *Estimation of* $f''(\bar{v})\,(h_0 + h_1)^2$.
Invoking (5.43), $h_0 \in C_\tau(\bar{v})$, and the representation (5.44), we readily see that

$$\frac{1}{2} f''(\bar{v})\, h_0^2 \;\geq\; \frac{\delta}{2}\, \|h_0\|^2_{L^2(Q)}.$$

Using Young's inequality, we conclude from (5.45) that, with a generic constant $c > 0$,

$$\begin{aligned} \left| f''(\bar{v})[h_0, h_1] \right| &\leq c\, \|h_0\|_{L^2(Q)}\, \|h_1\|_{L^2(Q)} \;\leq\; \frac{\delta}{4}\, \|h_0\|^2_{L^2(Q)} + c\, \|h_1\|^2_{L^2(Q)} \\ &\leq \frac{\delta}{4}\, \|h_0\|^2_{L^2(Q)} + c\, \|h_1\|_{L^1(Q)}\, \|h_1\|_{L^\infty(Q)} \\ &\leq \frac{\delta}{4}\, \|h_0\|^2_{L^2(Q)} + c_1\, \varepsilon\, \|h_1\|_{L^1(Q)}, \end{aligned}$$

because $\|h\|_{L^\infty(Q)} \leq \varepsilon$. By the same token,

$$\left| \frac{1}{2} f''(\bar{v})\, h_1^2 \right| \leq c\, \|h_1\|^2_{L^2(Q)} \leq c_2\, \varepsilon\, \|h_1\|_{L^1(Q)}.$$

Summarizing, we obtain after combining the above inequalities that

$$\begin{aligned}\frac{1}{2}f''(\bar v)\,(h_0+h_1)^2 &\geq \frac{\delta}{2}\,\|h_0\|^2_{L^2(Q)} - \left[\frac{\delta}{4}\,\|h_0\|^2_{L^2(Q)} + (c_1+c_2)\,\varepsilon\,\|h_1\|_{L^1(Q)}\right]\\ &\geq \frac{\delta}{4}\,\|h_0\|^2_{L^2(Q)} - (c_1+c_2)\,\varepsilon\,\|h_1\|_{L^1(Q)}.\end{aligned}$$

Now we choose $\varepsilon > 0$ so small that $\varepsilon\,(c_1+c_2) \leq \tau/2$. Since $h_1 = 0$ on $\Omega \setminus A_\tau$, we can infer that

$$\|h_1\|_{L^1(Q)} = \iint_{A_\tau(\bar v)} |h_1|\,dx\,dt.$$

Substituting the above estimates into (5.46) then yields

$$\begin{aligned}f(v) - f(\bar v) &\geq \tau \iint_{A_\tau(\bar v)} |h|\,dx\,dt - \frac{\tau}{2}\iint_{A_\tau(\bar v)} |h|\,dx\,dt + \frac{\delta}{4}\,\|h_0\|^2_{L^2(Q)} + r_2^f\\ &\geq \frac{\tau}{2}\iint_{A_\tau(\bar v)} |h|\,dx\,dt + \frac{\delta}{4}\,\|h_0\|^2_{L^2(Q)} + r_2^f.\end{aligned}$$

Now we choose $\varepsilon \leq 1$, which can be done without loss of generality. Then, by the definition of h, we have $|h(x,t)| \geq h(x,t)^2$. Since

$$\|h_0\|^2_{L^2(Q)} = \iint_{Q\setminus A_\tau(\bar v)} h^2\,dx\,dt,$$

we finally obtain that

$$\begin{aligned}f(v) - f(\bar v) &\geq \frac{\tau}{2}\iint_{A_\tau(\bar v)} h^2\,dx\,dt + \frac{\delta}{4}\,\|h\|^2_{L^2(Q\setminus A_\tau)} + r_2^f\\ &\geq \min\left\{\frac{\tau}{2},\frac{\delta}{4}\right\}\,\|h\|^2_{L^2(Q)} + r_2^f.\end{aligned}$$

In Exercise 5.2, the reader will be asked to show that the remainder $r_2^f(\bar v, h)$ satisfies, as in (4.79) on page 237,

$$\frac{r_2^f(\bar v,h)}{\|h\|^2_{L^2(Q)}} \to 0 \qquad \text{as} \quad \|h\|_{L^\infty(Q)} \to 0.$$

Hence, for sufficiently small $\varepsilon > 0$ we have the estimate

$$f(v) - f(\bar v) \geq \frac{1}{2}\,\min\left\{\frac{\tau}{2},\frac{\delta}{4}\right\}\,\|h\|^2_{L^2(Q)} = \sigma\,\|h\|^2_{L^2(Q)},$$

which is the asserted quadratic growth. □

Remarks.

(i) The assertion of the above theorem is valid, in particular, if the definiteness condition (5.43) is postulated for the larger set of all (y, v) satisfying both the linearized system (5.41) and the sign conditions in (5.42), but not the condition $v = 0$ on $A_\tau(\bar{v})$. These stronger second-order sufficient conditions, which are often assumed in the convergence proofs of numerical methods, have been used by us before in the study of elliptic problems.

(ii) Another, more elegant, approach to deriving the second-order sufficient optimality conditions which avoids the need for $\delta > 0$ and $\tau > 0$ can be found in Casas et al. [**CDlRT08**]. However, that second-order condition is equivalent to the one used here.

5.7.3. Boundary control. In view of the analogy, we present the sufficient conditions for boundary controls without proof. We consider the optimal control problem (5.27)–(5.29) on page 282. The Lagrangian function is defined, in analogy to the case of distributed controls, by

$$\begin{aligned}\mathcal{L}(y,v,p) &= J(y,v) - \iint_Q \big((y_t + d(x,t,y))\,p + \nabla y \cdot \nabla p\big)\,dx\,dt \\ &\quad - \iint_\Sigma (b(x,t,y) - u)\,p\,ds\,dt.\end{aligned}$$

Its second derivative is given by

$$\begin{aligned}\mathcal{L}''(\bar{y},\bar{u},p)(y,v)^2 &= J''(\bar{y},\bar{u})(y,u)^2 - \iint_Q p\,d_{yy}(x,t,\bar{y})\,y^2\,dx\,dt \\ &\quad - \iint_\Sigma p\,b_{yy}(x,t,\bar{y})\,y^2\,ds\,dt,\end{aligned}$$

where

$$J''(\bar{y},\bar{u})(y,u)^2 = \int_\Omega \phi_{yy}(x,\bar{y}(x,T))\,y(x,T)^2\,dx + \iint_Q \varphi_{yy}(x,t,\bar{y})\,y^2\,dx\,dt$$
$$+ \iint_\Sigma \begin{bmatrix} y \\ u \end{bmatrix}^\top \begin{bmatrix} \psi_{yy}(x,t,\bar{y},\bar{u}) & \psi_{yu}(x,t,\bar{y},\bar{u}) \\ \psi_{uy}(x,t,\bar{y},\bar{u}) & \psi_{uu}(x,t,\bar{y},\bar{u}) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} ds\,dt.$$

The control $\bar{u} \in U_{ad}$ is assumed to comply with the first-order necessary optimality conditions. With the associated adjoint state p, defined by the adjoint problem (5.31) on page 283, the variational inequality (5.32) is satisfied:

$$\iint_\Sigma \big(p + \psi_u(x,t,\bar{y},\bar{u})\big)(u - \bar{u})\,ds\,dt \geq 0 \quad \forall\, u \in U_{ad}.$$

For given $\tau \geq 0$, the set

$$A_\tau(\bar u) = \big\{(x,t) \in \Sigma : \big|p(x,t) + \psi_u\big(x,t,\bar y(x,t),\bar u(x,t)\big)\big| > \tau\big\}$$

is called the *set of strongly active constraints* for $\bar u$. The linearized problem reads

$$\begin{aligned} y_t - \Delta y + d_y(x,t,\bar y)\, y &= 0 \qquad \text{in } Q \\ \partial_\nu y + b_y(x,t,\bar y)\, y &= u \qquad \text{on } \Sigma \\ y(0) &= 0 \qquad \text{in } \Omega, \end{aligned} \tag{5.47}$$

and the τ-critical cone $C_\tau(\bar u)$ consists of all $u \in L^\infty(\Sigma)$ such that

$$u(x,t) \begin{cases} = & 0 \ \text{ if } \ (x,t) \in A_\tau(\bar u) \\ \geq & 0 \ \text{ if } \ \bar u(x,t) = u_a \ \text{ and } \ (x,t) \notin A_\tau(\bar u) \\ \leq & 0 \ \text{ if } \ \bar u(x,t) = u_b \ \text{ and } \ (x,t) \notin A_\tau(\bar u). \end{cases} \tag{5.48}$$

Again, we postulate the following as the *second-order sufficient condition*:

$$\begin{aligned} &\mathcal{L}''(\bar y,\bar u,p)(y,u)^2 \geq \delta\, \|u\|^2_{L^2(\Sigma)} \\ &\text{for all } u \in C_\tau(\bar u) \text{ and } y \in W(0,T) \text{ satisfying } (5.47). \end{aligned} \tag{5.49}$$

Theorem 5.18. *Suppose that Assumption 5.6 on page 269 holds, and assume that the pair $(\bar y,\bar u)$ complies with both the first-order necessary optimality conditions stated in Theorem 5.13 on page 283 and the constraints of the optimal boundary control problem. Moreover, let there be constants $\delta > 0$ and $\tau > 0$ such that the definiteness condition* (5.49) *is fulfilled. Then there exist constants $\varepsilon > 0$ and $\sigma > 0$ such that for every $u \in U_{ad}$ with $\|u - \bar u\|_{L^\infty(\Sigma)} \leq \varepsilon$ and associated state y, the quadratic growth condition*

$$J(y,u) \geq J(\bar y,\bar u) + \sigma\, \|u - \bar u\|^2_{L^2(\Sigma)}$$

holds. In particular, $\bar u$ is locally optimal.

5.7.4. A case without two-norm discrepancy. For parabolic problems also, the two-norm discrepancy does not occur if for L^2 controls the regularity of the state and the differentiability of the nonlinearities fit with each other. For instance, this is the case for nonlinearities that are quadratic in a certain sense so that the second-order remainder vanishes. Such an example will be given in Section 5.10.2 by considering control of the Navier–Stokes equations. Also for the phase field model with cubic nonlinearities to be studied in Section 5.10.1, a two-norm discrepancy does not occur.

In the case of problem (5.7)–(5.9) on page 270 with the very general nonlinearities $d(x,t,y)$ and $b(x,t,y)$, the two-norm discrepancy can be excluded for the following constellation: distributed control with a suitable

cost functional and a one-dimensional spatial domain $\Omega = (0, \ell)$. We take as an example the problem

$$\min \left\{ \int_0^\ell \phi\big(x, y(x,T)\big)\, dx + \int_0^T \big(\psi_1(t, y(0,t)) + \psi_2(t, y(\ell,t))\big)\, dt \right. \\ \left. + \int_0^\ell \int_0^T \varphi(x,t,y,v)\, dx\, dt \right\},$$

subject to

$$\begin{array}{rcll} y_t(x,t) - y_{xx}(x,t) + d(x,t,y(x,t)) & = & v(x,t) & \text{in } (0,\ell)\times(0,T) \\ -y_x(0,t) + b_1(t, y(0,t)) & = & 0 & \text{in } (0,T) \\ y_x(\ell,t) + b_2(t, y(\ell,t)) & = & 0 & \text{in } (0,T) \\ y(x,0) & = & y_0(x) & \text{in } (0,\ell) \end{array}$$

and

$$v_a(x,t) \le v(x,t) \le v_b(x,t).$$

So that the two-norm discrepancy does not come into play, the cost functional must be linear-quadratic with respect to v. We therefore postulate the following form for φ:

$$\varphi(x,t,y,v) = \varphi_1(x,t,y) + \varphi_2(x,t,y)\, v + \lambda(x,t)\, v^2. \tag{5.50}$$

For $d = 0$ and $b_1 = b_2 = 0$, that is, for the linear parabolic problem, right-hand sides $v \in L^r(Q)$ of the differential equation are mapped into $W(0,T) \cap C(\bar{Q})$, provided that y_0 is continuous and $r > N/2 + 1$. The latter holds for $N = 1$ and $r = 2$. Hence, with a slight modification of the proof of Theorem 5.15 on page 286, it can be shown that the control-to-state operator G is twice continuously Fréchet differentiable as a mapping from $L^2(Q)$ into $W(0,T) \cap C(\bar{Q})$. The linear operators used there are defined as $G_Q : L^2(Q) \to Y$, $G_\Sigma : L^\infty(\Sigma) \to Y$, and $G_\Omega : C(\bar{\Omega}) \to Y$.

Consequently, we may employ $L^2(Q)$ controls for distributed control problems in one-dimensional domains. Unfortunately, this is not possible for boundary control problems.

By modifying Assumption 5.6 on page 269 appropriately to suit the one-dimensional case considered here, the reader will be able to determine in Exercise 5.3 the regularity conditions that the functions ϕ, φ_i, and ψ_i must obey in order for the cost functional to be twice continuously Fréchet differentiable in $C(\bar{Q}) \times L^2(Q)$. In particular, λ needs to be bounded and

measurable, and we have to postulate $\lambda(x,t) \geq \delta > 0$ for a sufficient second-order optimality condition to hold. The assertion of Theorem 5.17 then holds with $L^2(Q)$ in place of $L^\infty(Q)$.

5.8. Test examples

In this section, we are going to discuss two test examples for nonlinear parabolic control problems in which the constructed solution satisfies sufficient optimality conditions. For this purpose, we use the same approach as for elliptic problems: the optimal quantities $\bar{u}$ and $\bar{y}$ and the associated adjoint state p are chosen a priori; then certain linear parts of the cost functional and constant terms in the initial-boundary value problem are fitted in such a way that both the optimality system and a second-order sufficient optimality condition are satisfied.

The first test example is constructed in such a way that the sufficient condition holds in the entire control-state space. In the second, more complicated example, the sufficient conditions will only be satisfied where the control constraints are not strongly active. In addition, a state constraint in the form of an integral inequality will be prescribed.

5.8.1. A test example with control constraints. Let $T > 0$, $\ell > 0$, and $Q = (0,\ell) \times (0,T)$. We consider the spatially one-dimensional optimal control problem

$$\begin{aligned} \min J(y,u) := \frac{1}{2}\int_0^\ell \big|y(x,T) - y_\Omega(x)\big|^2\,dx - \int_0^T a_y(t)\,y(\ell,t)\,dt \\ + \int_0^T \big(a_u(t)\,u(t) + \frac{\lambda}{2}(u(t))^2\big)\,dt, \end{aligned} \tag{5.51}$$

subject to

(5.52)

$$\boxed{\begin{array}{rcll} y_t(x,t) - y_{xx}(x,t) & = & 0 & \text{in } (0,\ell)\times(0,T) \\ -y_x(0,t) & = & 0 & \text{in } (0,T) \\ y_x(\ell,t) + y(\ell,t) & = & b(t) + u(t) - \varphi\big(y(\ell,t)\big) & \text{in } (0,T) \\ y(x,0) & = & a(x) & \text{in } (0,\ell) \end{array}}$$

and

$$0 \leq u(t) \leq 1 \qquad \text{for a.e. } t \in (0,T). \tag{5.53}$$

Here, the following quantities are prescribed:

$$\ell = \frac{\pi}{4}, \quad T = 1, \quad \lambda = \frac{\sqrt{2}}{2}\left(e^{2/3} - e^{1/3}\right), \quad \varphi(y) = y\,|y|^3,$$
$$y_\Omega(x) = \left(e + e^{-1}\right)\cos(x), \quad a_y(t) = e^{-2t}, \quad a_u(t) = \frac{\sqrt{2}}{2}\,e^{1/3},$$
$$a(x) = \cos(x), \quad b(t) = \frac{1}{4}\,e^{-4t} - \min\left\{1, \max\left\{0, \frac{e^t - e^{1/3}}{e^{2/3} - e^{1/3}}\right\}\right\}.$$

Obviously, this is an optimal control problem for the one-dimensional heat equation with boundary condition of Stefan–Boltzmann type. The reader will have the opportunity in Exercise 5.4 to verify that the following triple of functions satisfies the first-order necessary optimality conditions:

$$\begin{aligned} \bar{u}(t) &= \min\left\{1, \max\left\{0, \frac{e^t - e^{1/3}}{e^{2/3} - e^{1/3}}\right\}\right\}, \\ \bar{y}(x,t) &= e^{-t}\cos(x), \quad p(x,t) = -e^t\cos(x). \end{aligned}$$

Here, the adjoint state p solves the adjoint system

$$\begin{aligned} -p_t(x,t) - p_{xx}(x,t) &= 0 \\ p_x(0,t) &= 0 \\ p_x(\ell,t) + \left[1 + \varphi'\left(\bar{y}(\ell,t)\right)\right] p(\ell,t) &= -a_y(t) \\ p(x,T) &= \bar{y}(x,T) - y_\Omega(x). \end{aligned} \tag{5.54}$$

In particular, the projection relation for $\bar{u}$ has to be verified. For further details, we refer the interested reader to the paper **[ART02]**, from which this example was taken.

The basic idea for the construction of $\bar{u}$ is the following: the graph of an interesting optimal control should reach both the upper and the lower bounds, connecting them smoothly. In view of the projection relation, $\bar{u}$ has to be a multiple of the adjoint state p between the bounds. Moreover, exponential functions of time are very good candidates for solving the heat conduction equation explicitly, especially for p. These considerations, and an appropriate adjustment of constants, lead to the above choice of the function $\bar{u}$, which is depicted on page 312.

To prove local optimality, we show that a second-order sufficient optimality condition is fulfilled. The (formal) Lagrangian function reads

$$\begin{aligned}\mathcal{L} \;=\; & J(y,u) - \int_0^T \int_0^\ell \big(y_t - y_{xx}\big)\, p\, dx\, dt - \int_0^\ell \big(y(x,0) - a(x)\big)\, p(x,0)\, dx \\ & + \int_0^T y_x(0,t)\, p(0,t)\, dt - \int_0^T \big(y_x(\ell,t) + y(\ell,t) - b(t) - u(t)\big)\, p(\ell,t)\, dt \\ & - \int_0^T \varphi\big(y(\ell,t)\big)\, p(\ell,t)\, dt,\end{aligned}$$

which is a special case of the Lagrangian for more general boundary control problems that was introduced on page 295. Since $\Gamma = \{0,1\}$, we have

$$\begin{aligned}\int_\Sigma p\, \partial_\nu y\, ds\, dt &= \int_0^T \big(p(0,t)\partial_\nu y(0,t) + p(\ell,t)\partial_\nu y(\ell,t)\big)\, dt \\ &= \int_0^T \big(-p(0,t)y_x(0,t) + p(\ell,t)y_x(\ell,t)\big)\, dt.\end{aligned}$$

Moreover, since $\varphi''(y) = 12\, y^2$,

$$\begin{aligned}\mathcal{L}''(\bar y,\bar u,p)(y,u)^2 &= \|y(T)\|^2_{L^2(0,\ell)} + \lambda\, \|u\|^2_{L^2(0,T)} \\ &\qquad -12 \int_0^T p(\ell,t)\, \bar y(\ell,t)^2\, y(\ell,t)^2\, dt\end{aligned}$$

and thus, since $p(x,t) \le 0$,

$$\mathcal{L}''(\bar y,\bar u,p)(y,u)^2 \ge \lambda\, \|u\|^2_{L^2(0,T)} \tag{5.55}$$

for all square integrable functions y and u. Theorem 5.18 yields the local optimality of $\bar u$ in the sense of $L^\infty(0,T)$. Observe, however, that a very strong form of the second-order sufficient conditions holds; we therefore should not be too surprised at the following result.

Theorem 5.19. *The pair $(\bar y, \bar u)$ defined above is (globally) optimal for the optimal control problem* (5.51)–(5.53).

Proof: Let (y,u) be another admissible pair. Invoking the variational inequality satisfied by the local minimizer $\bar u$ of the reduced functional f, we find that, with some function $\theta = \theta(x) \in (0,1)$,

$$f(u) \ge f(\bar u) + \frac{1}{2} f''(\bar u + \theta(u - \bar u))(u - \bar u)^2.$$

As in formula (5.44) on page 292, we have

$$f''(\bar u+\theta(u-\bar u))(u-\bar u)^2 = \mathcal{L}''(y_\theta,u_\theta,p_\theta)(y,u-\bar u)^2$$
$$\geq -\int_0^T p_\theta(\ell,t)\,\varphi''\big(y_\theta(\ell,t)\big)\,y(\ell,t)^2 dt.$$

Here, $u_\theta=\bar u+\theta(u-\bar u)$, y_θ is the associated state, p_θ is the adjoint state, and y denotes the solution to the linearized problem at y_θ corresponding to the control $u-\bar u$. Since φ'' is nonnegative, the assertion $f(u)\geq f(\bar u)$ will follow once we are able to show that all possible adjoint states p_θ are nonpositive.

We sketch the proof of this claim. To this end, we first show the following claim: for any $u_\theta\in U_{ad}$, the associated state y_θ satisfies the inequality

$$y_\theta(x,T)-y_\Omega(x)<0 \quad \forall x\in\bar\Omega.$$

Indeed, we immediately see that $y_\Omega\geq (e+e^{-1})\cos(\pi/4)>3\sqrt{2}/2>2$, $b(t)+u_\theta(t)\leq 1.25$, and $y_0(x)=\cos(x)\leq 1$. Moreover, since all the given data of the problem are nonnegative, it follows from standard comparison principles for parabolic problems, as in **[RZ99]**, that the state y_θ is nonnegative. Hence, $y_\theta^3\,|y_\theta|=y_\theta^4\geq 0$. But then

$$(y_{\theta,x}+y_\theta)(\ell,t)=b(t)+u(t)-y_\theta^4(\ell,t)\leq 1.25.$$

By the maximum principle, the maximum of the linear parabolic initial-boundary value problem with boundary condition $y_{\theta,x}+y_\theta=1.25$ can be attained only at the boundary (i.e., at $x\in\{0,\ell\}$) or at $t=0$. Consequently, we must have $y_\theta\leq\max\{1,1.25\}=1.25$, and thus $y_\theta(x,T)-y_\Omega\leq 1.25-2<0$, which proves the claim.

Hence, if we replace $\bar y$ by y_θ in the adjoint problem (5.54), then all four right-hand sides are nonpositive. Moreover, we have $1+\varphi'(y_\theta)\geq 0$ in the boundary condition. It then follows from the aforementioned comparison principles for parabolic problems that p_θ is in fact nonpositive. This concludes the proof of the assertion. □

5.8.2. A problem with a state constraint of integral form *. In **[MT02]**, the following optimal control problem was investigated numerically with respect to second-order sufficient conditions:

(5.56)

$$\min J(y,u) := \frac12\iint_Q \alpha(x,t)\big|y(x,t)-y_Q(x,t)\big|^2\,dx\,dt+\frac{\lambda}{2}\int_0^T|u(t)|^2\,dt$$
$$+\int_0^T\big(a_y(t)\,y(\ell,t)+a_u(t)\,u(t)\big)\,dt,$$

subject to

$$
\boxed{\begin{array}{rcll}
y_t(x,t) - y_{xx}(x,t) &=& e_Q & \text{in } Q \\
y_x(0,t) &=& 0 & \text{in } (0,T) \\
y_x(\ell,t) + y(\ell,t)^2 &=& e_\Sigma(t) + u(t) & \text{in } (0,T) \\
y(x,0) &=& 0 & \text{in } (0,\ell)
\end{array}} \tag{5.57}
$$

and the control and state constraints

$$
0 \le u(t) \le 1 \qquad \text{for a.e. } t \in (0,T), \tag{5.58}
$$

$$
\iint_Q y(x,t)\,dx\,dt \le 0. \tag{5.59}
$$

Here, $T > 0$, $\ell > 0$, and $\lambda > 0$ are given, and we have put $Q := (0,\ell)\times(0,T)$. The functions α, y_Q, $e_Q \in L^\infty(Q)$ and a_y, a_u, $e_\Sigma \in L^\infty(0,T)$ are prescribed. As before,

$$
U_{ad} = \big\{ u \in L^\infty(0,T) \,:\, 0 \le u(t) \le 1 \quad \text{for a.e. } t \in (0,T) \big\}
$$

denotes the set of admissible controls. Now observe that the nonlinearity $y \mapsto y^2$ is not monotone increasing. We could circumvent this difficulty by considering the function $y \mapsto y\,|y|$ instead, in order to guarantee well-posedness of the state y; but this would be immaterial, since we will construct a nonnegative solution anyway.

First-order necessary conditions. Suppose that $\bar u$ is a locally optimal control for the above problem, and let $\bar y$ be the associated state. In addition to the control constraints, a state constraint is also to be respected. Therefore, the necessary optimality conditions derived so far do not apply directly. In the next chapter, we will establish the so-called Karush–Kuhn–Tucker conditions, which are valid under a certain regularity assumption. The test solution is constructed in such a way that these assumptions hold. Hence, it follows from Theorem 6.3 on page 330 that there exist an adjoint state $p \in W(0,T) \cap C(\bar Q)$ and a nonnegative Lagrange multiplier $\mu \in \mathbb{R}$ satisfying the adjoint problem

$$
\begin{array}{rcll}
-p_t - p_{xx} &=& \alpha\,(\bar y - y_Q) + \mu & \text{in } Q \\
p_x(0,t) &=& 0 & \text{in } (0,T) \\
p_x(\ell,t) + 2\,\bar y(\ell,t)\,p(\ell,t) &=& a_y(t) & \text{in } (0,T) \\
p(x,T) &=& 0 & \text{in } (0,\ell)
\end{array} \tag{5.60}
$$

and the variational inequality

$$\int_0^T \big(\lambda\,\bar u(t) + p(\ell,t) + a_u(t)\big)\,\big(u(t)-\bar u(t)\big)\,dt \ge 0 \quad \forall u \in U_{ad}. \tag{5.61}$$

In addition, the *complementary slackness condition*

$$\mu \iint_Q \bar y(x,t)\,dx\,dt = 0 \tag{5.62}$$

must hold; see Casas and Mateos [**CM02a**]. The result also follows from Theorem 6.3 on page 330. Note that the variational inequality (5.61) is equivalent to the projection relation

$$\bar u(t) = \mathbb{P}_{[0,1]}\left\{-\frac{1}{\lambda}\big(p(\ell,t)+a_u(t)\big)\right\} \qquad \text{for a.e. } t \in (0,T), \tag{5.63}$$

where $\mathbb{P}_{[0,1]} : \mathbb{R} \to [0,1]$ denotes pointwise projection onto the interval $[0,1]$.

These optimality conditions can also be obtained using the formal Lagrange method. In this case the Lagrangian function reads

$$\begin{aligned}\mathcal{L}(y,u,p,\mu) \;=\; & J(y,u) - \iint_Q (y_t - y_{xx} - e_Q)\,p\,dx\,dt + \iint_Q \mu\,y(x,t)\,dx\,dt \\ & - \int_0^T \big(y_x(\ell,t) + y(\ell,t)^2 - u(t) - e_\Sigma(t)\big)\,p(\ell,t)\,dt,\end{aligned}$$

where we assume that the homogeneous initial and boundary conditions for y are incorporated in the choice of state space. Then the conditions (5.60)–(5.61) follow from the requirements that $D_y\mathcal{L}(\bar y,\bar u,p,\mu)\,y = 0$ for all admissible increments y and $D_u\mathcal{L}(\bar y,\bar u,p,\mu)\,(u-\bar u) \ge 0$ for all $u \in U_{ad}$.

For the construction of the test example, we fix the data as follows:

$T = 1, \quad \ell = \pi, \quad \lambda = 0.004, \quad a_u(t) = \lambda + 1 - (1+2\lambda)\,t,$

$$\alpha(x,t) \;=\; \begin{cases} \alpha_0 \in \mathbb{R}, & t \in \big[0,\frac14\big] \\ 1, & t \in \big(\frac14,1\big], \end{cases}$$

$$y_Q(x,t) \;=\; \begin{cases} \dfrac{1}{\alpha(x,t)}\,(1-(2-t)\,\cos(x)), & t \in \big[0,\frac12\big] \\ \dfrac{1}{\alpha(x,t)}\,\big(1-\big(2-t-\alpha(x,t)\,(t-0.5)^2\big)\big)\,\cos(x), & t \in \big(\frac12,1\big], \end{cases}$$

$$a_y(t) \;=\; \begin{cases} 0, & t \in \big[0,\frac12\big] \\ 2(t-0.5)^2\,(1-t), & t \in \big(\frac12,1\big], \end{cases}$$

$$e_Q(t) = \begin{cases} 0, & t \in \left[0, \frac{1}{2}\right] \\ (t^2 + t - 0.75)\cos(x), & t \in \left(\frac{1}{2}, 1\right], \end{cases}$$

$$e_\Sigma(t) = \begin{cases} 0, & t \in \left[0, \frac{1}{2}\right] \\ (t - 0.5)^4 - (2t - 1), & t \in \left(\frac{1}{2}, 1\right]. \end{cases}$$

Theorem 5.20. *With the above data, the quantities*

$$\bar{u}(t) = \max\{0, 2t - 1\}, \quad \bar{y}(x,t) = \begin{cases} 0, & t \in \left[0, \frac{1}{2}\right] \\ (t - 1/2)^2 \cos(x), & t \in \left(\frac{1}{2}, 1\right], \end{cases}$$

$$p(x,t) = (1 - t)\cos(x), \qquad \mu = 1$$

satisfy the first-order necessary optimality conditions for the problem (5.56)–(5.59).

Proof: Evidently, the state and adjoint problems are satisfied. Since the integral of the cosine function over $[0, \pi]$ vanishes, the constraint (5.59) is active, which entails that the complementarity condition (5.62) is valid. Moreover, $\bar{u}$ is an admissible control. It remains to verify the variational inequality (5.61). Invoking (5.63), we find that

$$-\frac{1}{\lambda}\left(p(\pi,t) + a_u(t)\right) = 2t - 1 \begin{cases} < 0, & t \in \left[0, \frac{1}{2}\right) \\ > 0, & t \in \left(\frac{1}{2}, 1\right]. \end{cases}$$

Therefore,

$$\mathbb{P}_{[0,1]}\left\{-\frac{1}{\lambda}\left(p(\pi,t) + a_u(t)\right)\right\} = \max\left\{0, 2t - 1\right\} = \bar{u}(t).$$

The assertion now follows from the equivalence between the variational inequality and the projection relation. □

Second-order sufficient optimality conditions. In order to incorporate strongly active constraints, we define, for given $\tau > 0$,

$$A_\tau(\bar{u}) = \left\{t \in (0,T) \,:\, |\lambda\,\bar{u}(t) + p(\ell,t) + a_u(t)| > \tau\right\}.$$

Owing to the variational inequalities, either $\bar{u}(t) = 0$ or $\bar{u}(t) = 1$ must hold on $A_\tau(\bar{u})$, depending on the sign of $\bar{u}(t) + p(\ell,t) + a_u(t)$. The second derivative $\mathcal{L}''$ is given by

(5.64)

$$\mathcal{L}''(\bar{y},\bar{u},p,\mu)(y,u)^2 = \iint_Q \alpha\, y^2\, dx\, dt + \int_0^T \left(-2\,p(\ell,t)\,y(\ell,t)^2 + \lambda\,u(t)^2\right) dt.$$

We also assume that the state constraint (5.59) is active. Otherwise, this constraint is meaningless, and the sufficient conditions coincide with those for pure control constraints. The second-order sufficient optimality condition reads as follows: there exist constants $\delta > 0$ and $\tau > 0$ such that

$$\mathcal{L}''(\bar{y}, \bar{u}, p, \mu)(y, u)^2 \geq \delta \int_0^T u^2 \, dt \tag{5.65}$$

for all $y \in W(0,T)$ and $u \in L^2(0,T)$ satisfying the following conditions:

$$\begin{aligned} y_t - y_{xx} &= 0 \\ y_x(0,t) &= 0 \\ y_x(\ell,t) + 2\,\bar{y}(\ell,t)\,y(\ell,t) &= u(t) \\ y(x,0) &= 0, \end{aligned} \tag{5.66}$$

$$u(t) \begin{cases} = 0 & \text{if } t \in A_\tau(\bar{u}) \\ \geq 0 & \text{if } \bar{u}(t) = 0 \text{ and } t \notin A_\tau(\bar{u}) \\ \leq 0 & \text{if } \bar{u}(t) = 1 \text{ and } t \notin A_\tau(\bar{u}), \end{cases} \tag{5.67}$$

$$\iint_Q y(x,t)\,dx\,dt = 0.$$

Sufficient conditions in the test example. Let us verify that our test solution $(\bar{y}, \bar{u})$ satisfies the above conditions. In fact, by construction $\bar{u}$ attains the lower bound 0 on $\left[0, \frac{1}{2}\right)$; the constraint is strongly active there, since for $b < \frac{1}{2}$ and $t \in [0,b]$ we have

$$\lambda\,\bar{u}(t) + p(\pi,t) + a_u(t) = p(\pi,t) + a_u(t) = -\lambda\,(2t-1) > -\lambda\,(2b-1).$$

Therefore, $[0,b] \subset A_\tau(\bar{u})$ for $\tau = |\lambda(2b-1)|$.

For the verification of the second-order sufficient conditions it thus suffices to show the definiteness condition (5.65) for all (y,u) such that the linearized problem (5.66) is satisfied and such that $u = 0$ on $[0,b]$ for some arbitrary but fixed $b \in \left(0, \frac{1}{2}\right)$. We are still free to choose α. We put

$$\alpha(x,t) = \begin{cases} \alpha_0, & 0 \leq t \leq b \\ 1, & b < t \leq 1. \end{cases} \tag{5.68}$$

Theorem 5.21. *Let α be chosen as in* (5.68) *with some $b \in \left[0, \frac{1}{2}\right)$. Then the second-order sufficient condition is satisfied for any $\alpha_0 \in \mathbb{R}$ by $(\bar{y}, \bar{u}, p, \mu)$.*

Proof: We choose (y,u) in such a way that u vanishes on $[0,b]$ and y solves the initial-boundary value problem (5.66). Then $y(x,t)=0$ on $[0,b]$. Therefore

$$\begin{aligned}(5.69)\quad &\mathcal{L}''(\bar y,\bar u,p,\mu)(y,u)^2\\ &=\int_0^\pi\int_b^1 |y|^2\,dx\,dt+\lambda\int_0^1|u|^2\,dt-2\int_0^1 p(\pi,t)\,|y(\pi,t)|^2\,dt\\ &\ge\lambda\int_0^1|u|^2\,dt-2\int_0^1\big(-(1-t)\big)\,|y(\pi,t)|^2\,dt\ge\lambda\int_0^1|u|^2\,dt,\end{aligned}$$

whence the definiteness condition (5.65) follows. □

Observe that α_0 has not been assumed positive. Obviously, for $\alpha_0\ge 0$, $\mathcal{L}''$ would be uniformly positive definite on the entire space $W(0,1)\times L^2(0,1)$; hence, in this case the definiteness condition would be met in a very strong sense. We now choose a partially negative α so that $\mathcal{L}''$ becomes indefinite.

Theorem 5.22. *For any sufficiently large negative α_0 there exists a pair (y,u) with the following properties: $u\ge 0$, y solves the linearized problem* (5.66)*, and*

$$(5.70)\qquad \mathcal{L}''(\bar y,\bar u,p,\mu)(y,u)^2<0.$$

Proof: We fix $b\in\big(0,\frac12\big)$ and set

$$u(t)=\begin{cases}1 & \text{in } [0,b]\\ 0 & \text{in } (b,1].\end{cases}$$

Then y cannot vanish identically, so $\int_0^\pi\int_0^b|y|^2\,dx\,dt>0$. Therefore,

$$(5.71)\qquad \alpha_0\int_0^\pi\int_0^b|y|^2\,dx\,dt\to-\infty$$

as $\alpha_0\to-\infty$. Consequently, the expression (5.69) becomes negative for sufficiently large negative α_0, since the additional integral term (5.71) occurs. □

For numerical purposes, one needs a rough estimate of how small α_0 has to be chosen so as to guarantee that $\mathcal{L}''(\bar y,\bar u,p,\mu)(y,u)^2<0$. We must have

$$\alpha_0\int_0^\pi\int_0^b y^2\,dx\,dt+\int_0^\pi\int_b^1 y^2\,dx\,dt+\int_0^1 2\,(1-t)\,y^2\,(\pi,t)\,dt+\lambda\int_0^1 u^2\,dt<0,$$

and therefore we postulate that

(5.72)
$$|\alpha_0| > \frac{\int_0^\pi \int_b^1 y^2\,dx\,dt + \int_0^1 2(1-t)y^2(\pi,t)\,dt + \lambda \int_0^1 u^2 dt}{\int_0^\pi \int_0^b y^2\,dx\,dt} =: \frac{I_1 + I_2 + I_3}{I_0}.$$

Here, $b \in \left[0, \frac{1}{2}\right)$ can be chosen arbitrarily, say $b = \frac{1}{4}$. We evaluate the integrals I_j with this choice and with

$$u(t) = \begin{cases} 1 & \text{in } \left[0, \frac{1}{4}\right] \\ 0 & \text{in } \left(\frac{1}{4}, 1\right]. \end{cases}$$

The corresponding state y solves (5.66), the homogeneous heat conduction equation with zero initial condition, zero boundary condition at $x = 0$, and

$$y_x(\pi,t) = \begin{cases} 1 & \text{in } \left[0, \frac{1}{4}\right] \\ -2\,\bar{y}(\pi,t) & \text{in } \left(\frac{1}{4}, 1\right]. \end{cases}$$

A numerical evaluation of the integrals I_j, $0 \le j \le 3$, yields $I_0 = 0.0103271$, $I_1 = 0.0401844$, $I_2 = 0.0708107$, $I_3 = 0.001$, and thus

$$\frac{I_1 + I_2 + I_3}{I_0} = 10.845.$$

In this example, the second-order sufficient conditions are fulfilled, even though the bilinear form $\mathcal{L}''(\bar{y}, \bar{u}, p, \mu)$ is not positive definite in the entire space; it is negative definite in some sets where the control constraints are strongly active. This has been enforced by an analytical construction.

The question arises as to whether the second-order conditions can be verified numerically, because an explicit solution of the problem is usually impossible. Such a numerical solution of the problem has to be found by means of a discretization, for instance by using finite differences in time and finite elements in space.

For the discretized problem, the so-called *reduced Hessian matrix* can be determined. This is the Hessian of the cost functional taken with respect to the discrete control u, after y has been eliminated using the discrete linearized problem. It also accounts for the active constraints; see Kelley **[Kel99]**. If the eigenvalues of the Hessian are sufficiently positive, then one can trust that the sufficient conditions are fulfilled—but this is no proof. The above method has been tested numerically in **[MT02]**.

5.9. Numerical methods

The numerical methods presented in this section are concerned with the optimal control problem (5.7)–(5.9):

$$\begin{aligned}\min J(y,v,u) \;:=\; & \int_\Omega \phi\big(x,y(x,T)\big)\,dx + \iint_Q \varphi\big(x,t,y(x,t),v(x,t)\big)\,dx\,dt \\ & + \iint_\Sigma \psi\big(x,t,y(x,t),u(x,t)\big)\,ds\,dt,\end{aligned}$$

subject to

$$\boxed{\begin{aligned} y_t - \Delta y + d(x,t,y) &= v && \text{in } Q \\ \partial_\nu y + b(x,t,y) &= u && \text{on } \Sigma \\ y(0) &= y_0 && \text{in } \Omega \end{aligned}}$$

and

$$\begin{aligned} v_a(x,t) \le v(x,t) \le v_b(x,t) \quad & \text{for a.e. } (x,t) \in Q \\ u_a(x,t) \le u(x,t) \le u_b(x,t) \quad & \text{for a.e. } (x,t) \in \Sigma. \end{aligned}$$

We present the methods only formally, assuming that all of the functions and data involved are so smooth that the expressions that arise are meaningful.

5.9.1. Gradient methods. We begin with the projected gradient method introduced in Chapter 2. To this end, we consider the reduced cost functional $f(v,u) = J(y(v,u),v,u)$. The Fréchet derivative of f at (v_n,u_n), with associated adjoint state p_n, is given by

$$\begin{aligned} f'(v_n,u_n)\,(v,u) \;=\; & \iint_Q \big(\varphi_v(x,t,y_n,v_n) + p_n\big)\,v\,dx\,dt \\ & + \iint_\Sigma \big(\psi_u(x,t,y_n,u_n) + p_n\big)\,u\,ds\,dt. \end{aligned}$$

Here, the adjoint state p_n is the solution to the initial-boundary value problem

$$\begin{aligned} -p_t - \Delta p + d_y(x,t,y_n)\,p &= \varphi_y(x,t,y_n,v_n) \\ \partial_\nu p + b_y(x,t,y_n)\,p &= \psi_y(x,t,y_n,u_n) \\ p(x,T) &= \phi_y\big(x,y_n(x,T)\big). \end{aligned} \tag{5.73}$$

Suppose now that the controls (v_j,u_j), $1 \le j \le n$, have already been calculated. Then the next iterate (v_{n+1},u_{n+1}) is determined as follows:

S1 Find the state $y_n = y(v_n, u_n)$ by solving the initial-boundary value problem

$$\begin{aligned} y_t - \Delta y + d(x,t,y) &= v_n \\ \partial_\nu y + b(x,t,y) &= u_n \\ y(0) &= y_0. \end{aligned}$$

S2 Determine the adjoint state p_n by solving the adjoint problem (5.73).

S3 (*Descent directions*) Put

$$h_n := -(\varphi_v(\cdot, y_n, v_n) + p_n), \quad r_n := -(\psi_u(\cdot, y_n|_\Sigma, u_n) + p_n|_\Sigma).$$

S4 (*Step size control*) Determine s_n from

$$\min_{s>0} \; f\big(\mathbb{P}_V(v_n + s\,h_n), \mathbb{P}_U(u_n + s\,r_n)\big).$$

Here, $\mathbb{P}_V$ and $\mathbb{P}_U$ denote the pointwise projections onto the admissible sets V_{ad} and U_{ad}, respectively, that is, $\mathbb{P}_V = \mathbb{P}_{[v_a,v_b]}$ and $\mathbb{P}_U = \mathbb{P}_{[u_a,u_b]}$.

S5 (*New controls*) Set

$$(v_{n+1}, u_{n+1}) := \big(\mathbb{P}_V(u_n + s_n\,h_n), \mathbb{P}_U(v_n + s_n\,r_n)\big), \quad n := n+1; \quad \text{GO TO } \mathbf{S1}.$$

Remark. In contrast to the corresponding method for semilinear elliptic problems, the nonlinear problem in step **S1** is comparatively easy to solve numerically. Here, semi-explicit methods can be employed: to this end, take an equidistant subdivision $t_i = i\,\tau$, with $i = 0, \ldots, k$ and $\tau = T/k$, of the time interval $[0, T]$, and denote the approximation of $y_n(x, t_i)$ to be found by $w_i(x)$. Then $w_{i+1}(x)$ is determined as the solution to the *linear* elliptic boundary value problem

$$\begin{aligned} w_{i+1} - \tau\,\Delta w_{i+1} &= w_i - \tau\,d(x, t_{i+1}, w_i) + \tau\,v_n(\cdot, t_{i+1}) && \text{in } \Omega \\ \partial_\nu w_{i+1} &= -b(x, t_{i+1}, w_i) + u_n(\cdot, t_{i+1}) && \text{on } \Gamma, \end{aligned}$$

for $i = 0, \ldots, k-1$. In other words, k linear elliptic boundary value problems have to be solved numerically.

5.9.2. The SQP method. The SQP method follows the same lines as in the elliptic case. Therefore, the individual steps are not explained in further detail. We again consider the optimal control problem (5.7)–(5.9) studied above.

Let the iterates v_n, u_n, y_n, and p_n be determined. In the next iteration step, we have to solve the quadratic optimal control problem

$$\begin{aligned}(5.74)\quad \min \tilde{J}(y,v,u) \;:=\; & J'(y_n,v_n,u_n)(y-y_n,v-v_n,u-u_n)\\ & +\frac{1}{2}\,\mathcal{L}''(y_n,v_n,u_n,p_n)(y-y_n,v-v_n,u-u_n)^2,\end{aligned}$$

subject to

$$\begin{aligned} y_t-\Delta y+d(x,t,y_n)+d_y(x,t,y_n)(y-y_n) &= v && \text{in } Q\\ \partial_\nu y+b(x,t,y_n)+b_y(x,t,y_n)(y-y_n) &= u && \text{on } \Sigma\\ y(0) &= y_0 && \text{in } \Omega \end{aligned}\tag{5.75}$$

and

$$\begin{aligned} v_a(x,t)\le v(x,t)\le v_b(x,t) \quad &\text{for a.e. } (x,t)\in Q\\ u_a(x,t)\le u(x,t)\le u_b(x,t) \quad &\text{for a.e. } (x,t)\in \Sigma. \end{aligned}\tag{5.76}$$

For brevity, the Lagrangian is written down only formally:

$$\mathcal{L}=J-\iint_Q\big(y_t-\Delta y-d(x,t,y)-v\big)p\,dx\,dt-\iint_\Sigma\big(\partial_\nu y+b(x,t,y)-u\big)p\,ds\,dt.$$

Note that after calculation of the second derivative, the terms y_t, $-\Delta y$ and $\partial_\nu y$ that are only formally defined no longer appear anyway. A tedious but straightforward calculation yields for the quadratic cost functional $\tilde{J}$ the expression

$$\begin{aligned}\tilde{J}(y,v,u)=&\int_\Omega\Big\{\phi_y\big(x,y_n(\cdot,T)\big)\big(y(\cdot,T)-y_n(\cdot,T)\big)\\ &\qquad+\frac{1}{2}\,\phi_{yy}\big(x,y_n(\cdot,T)\big)\big|y(\cdot,T)-y_n(\cdot,T)\big|^2\Big\}\,dx\\ &+\iint_Q\left\{\begin{bmatrix}\varphi_y\\ \varphi_v\end{bmatrix}\cdot\begin{bmatrix}y-y_n\\ v-v_n\end{bmatrix}+\frac{1}{2}\begin{bmatrix}y-y_n\\ v-v_n\end{bmatrix}^\top\begin{bmatrix}\varphi_{yy}&\varphi_{yv}\\ \varphi_{vy}&\varphi_{vv}\end{bmatrix}\begin{bmatrix}y-y_n\\ v-v_n\end{bmatrix}\right\}dx\,dt\\ &+\iint_\Sigma\left\{\begin{bmatrix}\psi_y\\ \psi_u\end{bmatrix}\cdot\begin{bmatrix}y-y_n\\ u-u_n\end{bmatrix}+\frac{1}{2}\begin{bmatrix}y-y_n\\ u-u_n\end{bmatrix}^\top\begin{bmatrix}\psi_{yy}&\psi_{yu}\\ \psi_{uy}&\psi_{uu}\end{bmatrix}\begin{bmatrix}y-y_n\\ u-u_n\end{bmatrix}\right\}ds\,dt\\ &-\frac{1}{2}\iint_Q p_n\,d_{yy}(x,t,y_n)(y-y_n)^2\,dx\,dt-\frac{1}{2}\iint_\Sigma p_n\,b_{yy}(x,t,y_n)(y-y_n)^2\,ds\,dt,\end{aligned}$$

where the first- and second-order partial derivatives of φ and ψ also depend on the variables (x,t,y_n,v_n) and (x,t,y_n,u_n), respectively.

The solution of the quadratic problem (5.74)–(5.76) yields the new iterates $(y_{n+1},v_{n+1},u_{n+1})$. These functions are substituted into the adjoint

problem, which is solved for $p = p_{n+1}$. Here, the derivative of $\tilde{J}$ with respect to y occurs on the right-hand side, in accordance with the respective domains of integration; the adjoint system therefore reads

$$\begin{aligned} -p_t - \Delta p + d_y(x,t,y_n)\,p &= \varphi_y(x,t,y_n,v_n) + \varphi_{yy}(x,t,y_n,v_n)(y-y_n) \\ &\quad -p_n\, d_{yy}(x,t,y_n)(y-y_n) \\ \partial_\nu p + b_y(x,t,y_n)\,p &= \psi_y(x,t,y_n,u_n) + \psi_{yy}(x,t,y_n,u_n)(y-y_n) \\ &\quad -p_n\, b_{yy}(x,t,y_n)(y-y_n) \\ p(T) &= \phi_y\big(x, y_n(\cdot,T)\big) \\ &\quad +\phi_{yy}\big(x, y_n(\cdot,T)\big)\big(y(\cdot,T) - y_n(\cdot,T)\big). \end{aligned}$$

In the practical implementation of the numerical method, this adjoint state is often obtained as a byproduct, so that it is not necessary to solve the above problem. The SQP method described above converges locally quadratically to a local minimizer in the sense of the L^∞ norms of y, v, u, and p, provided the second-order sufficient optimality conditions are fulfilled; see [**Trö99**] and the references given there.

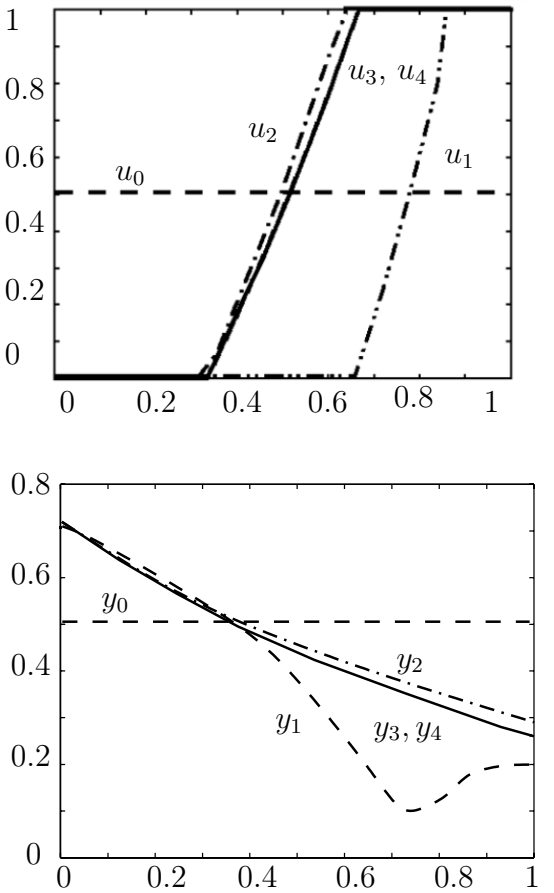
Example. The test example (5.51)–(5.53) on page 298 was solved with the initial guesses

$$u_0(t) \equiv y_0(x,t) \equiv p_0(x,t) \equiv 0.5.$$

The numerical solution of the parabolic problem was performed by an implicit finite difference scheme using a second-order approximation of the boundary condition. The step sizes for time and space were both chosen to be 1/200, and the controls were chosen as piecewise constant functions on the time grid.

The sequences of controls $u_n(t)$ and boundary values $y_n(\ell,t)$ are depicted in the figures that follow. Already after four iteration steps, the accuracy was higher than could be expected from numerical approximation of the partial differential equation. Graphically, a further refinement did not yield any improvement.

References. For the convergence analysis of the projected gradient method, we refer the reader to [**GS80**] and [**Kel99**]. In [**GS80**], other commonly used methods for the solution of discretized optimal control problems are discussed in detail. Parabolic problems with box constraints for the control can be solved very efficiently by the *projected Newton method*; see, e.g., [**HPUU09**], [**KS94**], [**KS95**], and [**Hei96**].



Controls $u_n(t)$ and boundary values $y_n(\ell, t)$.

Many papers, especially in the parabolic case, report on the use of the SQP method; see [**GT97a**], [**KS92**], and [**LS00**]. Schur complement techniques for the effective application of SQP methods to discretized parabolic problems are discussed in [**SW98**] and [**MS00**]. Besides SQP methods, trust-region and interior-point methods, as well as hybrid techniques derived from them, are also gaining importance in the solution of nonlinear parabolic problems. Among numerous references, we mention [**UUH99**], [**UU00**], and [**WS04**] as representatives. Further references are given, e.g., in the review article [**HV01**].

In the case of nonlinear parabolic problems in multi-dimensional domains, the dimensionality of the discretized problems is very high. It is possible to use model reduction techniques to establish problems of much smaller dimension that exhibit important properties of the original problem and yet are relatively accurate. Among these techniques, there is the so-called *POD method* (POD for **p**roper **o**rthogonal

decomposition), which plays an important role especially for flow problems. Applications of this technique can be found in, e.g., [**AH01**], [**HV03**], [**KV01**], [**KV02**], and [**Vol01**]. Another model reduction technique, which is commonly used for linear problems, is the *balanced truncation* method; see, e.g., [**Ant05**] and the list of references given there.

5.10. Further parabolic problems *

In this section, we demonstrate that the methods developed so far carry over to more general semilinear parabolic problems. We only sketch and explain the principal features of the approach, without entering into theoretical details that are beyond the scope of this textbook and which can be looked up in original papers. Both of the following examples will make it clear that eventually each class of nonlinear partial differential equations has to be tackled by techniques that are specifically tailored to it.

5.10.1. A phase field model. Here, we investigate an optimal control problem for the phase field model from Section 1.3.2. We establish the optimality system using the formal Lagrange method. The aim of the optimization is to approximate a desired temperature evolution $u_Q(x,t)$ and a desired solidification process $\varphi_Q(x,t)$:

$$\min\ J(u,\varphi,f) := \frac{\alpha}{2}\iint_Q |u-u_Q|^2\,dx\,dt + \frac{\beta}{2}\iint_Q |\varphi-\varphi_Q|^2\,dx\,dt + \frac{\lambda}{2}\iint_Q |f|^2\,dx\,dt, \tag{5.77}$$

subject to

$$\boxed{\begin{aligned} u_t + \frac{\ell}{2}\varphi_t &= \kappa\,\Delta u + f && \text{in } Q\\ \tau\,\varphi_t &= \xi^2\,\Delta\varphi + g(\varphi) + 2\,u && \text{in } Q\\ \partial_\nu u &= 0, \quad \partial_\nu\varphi = 0 && \text{on } \Sigma\\ u(\cdot,0) &= u_0, \quad \varphi(\cdot,0) = \varphi_0 && \text{in } \Omega \end{aligned}} \tag{5.78}$$

and

$$f_a(x,t) \le f(x,t) \le f_b(x,t) \quad \text{for a.e. } (x,t)\in Q. \tag{5.79}$$

Here, the following data are prescribed: functions u_Q, $\varphi_Q \in L^2(Q)$, constants $\alpha \ge 0$, $\beta \ge 0$, $\lambda \ge 0$, positive constants ℓ, κ, τ, ξ, and a polynomial $g = g(z) = a\,z + b\,z^2 - c\,z^3$ with bounded coefficient functions $a(x,t)$, $b(x,t)$, and $c(x,t) \ge \bar{c} > 0$. Moreover, $\Omega \subset \mathbb{R}^N$, $N \in \{2,3\}$, is a bounded domain of

class C^2, and the initial data u_0 and φ_0 are assumed to belong to $W^{2,\infty}(\Omega)$ and to satisfy the compatibility conditions $\partial_\nu \varphi_0 = \partial_\nu u_0 = 0$ on $\Gamma = \partial\Omega$.

The distributed heat source $f \in L^\infty(Q)$ is the control variable, and the states are given by the temperature u and the phase function φ. The states are considered in the normed spaces

$$W_p^{2,1}(Q) = \Big\{ u \in L^p(Q) \,:\, u, \frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i \partial x_j}, \frac{\partial u}{\partial t} \in L^p(Q) \text{ for } i,\, j = 1, \ldots, N \Big\},$$

endowed with the corresponding norm.

It has been shown in [**HJ92**] that the parabolic initial-boundary value system (5.78) admits for any given $f \in L^q(Q)$ with $q \geq 2$ a unique state pair $(\varphi, u) \in W_q^{2,1}(Q) \times W_p^{2,1}(Q)$, where

$$p = \begin{cases} \frac{5q}{5-2q} & \text{if } 2 \leq q < \frac{5}{2} \text{ and } N = 3 \\ \text{arbitrary in } \mathbb{R}_+ & \text{if } \frac{5}{2} \leq q \text{ and } N = 3 \text{ or } q \geq 2 \text{ and } N = 2. \end{cases}$$

According to [**HT99**], the mapping $G : f \mapsto (u, \varphi)$ is twice continuously Fréchet differentiable from $F := L^2(Q)$ into $Y := W_2^{1,2}(Q) \times W_2^{1,2}(Q)$. This follows from the fact that the Nemytskii operator $\varphi(\cdot,\cdot) \mapsto g(\varphi(\cdot,\cdot))$ is twice continuously differentiable from $W_2^{1,2}(Q)$ into $L^2(Q)$. The latter property is a consequence of the embedding $W_2^{1,2}(Q) \hookrightarrow L^6(Q)$ for $N \leq 3$ and the special form of g as a third-degree polynomial in φ.

For the derivation of the necessary optimality conditions, we introduce the Lagrangian function, following the lines of Section 3.1. We set

$$\begin{aligned} \mathcal{L}(u, \varphi, f, p, \psi) \;:=\; & J(u, \varphi, f) - \iint_Q \left[\left(u_t + \frac{\ell}{2}\, \varphi_t - f \right) p + \kappa\, \nabla u \cdot \nabla p \right] dx\, dt \\ & - \iint_Q \left((\tau\, \varphi_t - g(\varphi) - 2\, u)\, \psi + \xi^2\, \nabla \varphi \cdot \nabla \psi \right) dx\, dt, \end{aligned}$$

which already accounts for the homogeneous Neumann boundary conditions. The set of admissible controls is given by

$$F_{ad} = \big\{ f \in L^2(Q) :\, f_a(x,t) \leq f(x,t) \leq f_b(x,t) \quad \text{for a.e. } (x,t) \in Q \big\}.$$

The Lagrange method yields the following first-order necessary optimality conditions:

$$\begin{aligned} D_u \mathcal{L}(\bar u, \bar\varphi, \bar f, p, \psi)\, u &= 0 \qquad \forall u :\, u(0) = 0 \\ D_\varphi \mathcal{L}(\bar u, \bar\varphi, \bar f, p, \psi)\, \varphi &= 0 \qquad \forall \varphi :\, \varphi(0) = 0 \\ D_f \mathcal{L}(\bar u, \bar\varphi, \bar f, p, \psi)(f - \bar f) &\geq 0 \qquad \forall f \in F_{ad}. \end{aligned}$$

The zero initial conditions for the directions u and φ are prescribed so that $\bar{u} + s\,u$ and $\bar{\varphi} + s\,\varphi$ satisfy the initial conditions for $s \in \mathbb{R}$. The first of the above conditions yields, after integration by parts with respect to t,

$$\iint_Q \big(\alpha\,(\bar{u} - u_Q)\,u - p\,u_t - \kappa\,\nabla u \cdot \nabla p + 2\,\psi\,u\big)\,dx\,dt = 0$$

for all u with $u(0) = 0$. This is just the weak formulation of the initial-boundary value problem

$$\begin{aligned} -p_t &= \kappa\,\Delta p + 2\,\psi + \alpha\,(\bar{u} - u_Q) && \text{in } Q \\ \partial_\nu p &= 0 && \text{on } \Sigma \\ p(T) &= 0 && \text{in } \Omega. \end{aligned}$$

By the same token, we obtain from the second condition that

$$\begin{aligned} -\frac{\ell}{2}\,p_t - \tau\,\psi_t &= \xi^2\,\Delta\psi + g'(\bar{\varphi})\,\psi + \beta\,(\bar{\varphi} - \varphi_Q) && \text{in } Q \\ \partial_\nu \psi &= 0 && \text{on } \Sigma \\ \psi(T) &= 0 && \text{in } \Omega. \end{aligned}$$

The third condition yields the variational inequality

$$\iint_Q (\lambda \bar{f} + p)(f - \bar{f})\,dx\,dt \geq 0 \qquad \forall f \in F_{ad},$$

from which pointwise minimum principles or projection relations may again be derived. Further details can be looked up in original papers.

References. The existence of optimal controls and first-order necessary optimality conditions were derived in [**HJ92**], and the gradient method for the numerical solution of the problem was described in [**CH91**]. Second-order sufficient optimality conditions and the convergence of the SQP method were treated in [**HT99**]. Since the nonlinearity g is a polynomial of degree three, the two-norm discrepancy does not arise, and one can work in the space $Y := W_2^{1,2}(Q) \times W_2^{1,2}(Q) \times L^2(Q)$. The numerical solution of a concrete optimally controlled solidification problem was presented in [**Hei97**] and in [**HS94**]. The model reduction of such a problem is the subject of [**Vol01**]. The optimal control of more general (thermodynamically consistent) phase field models was studied in, e.g., [**SZ92**] and [**LS07**].

5.10.2. Nonstationary Navier–Stokes equations.

Problem formulation and definitions. The control of flows is an important and very active field of research in both mathematics and engineering. The great interest in such problems is motivated by various applications.

For instance, the behavior of the flow past an airfoil can be improved by controlled blowing or suction of air at the airfoil surface. Flow control can also prevent loss of contact of the water flow with the rudder of a ship or reduce the noise produced by jet engines. Optimal shape design of cardiac valves would enable flow patterns in artificial hearts to be improved.

In this section, we exemplarily discuss a simplified problem in which an optimal velocity field is to be generated in a spatial domain $\Omega \subset \mathbb{R}^2$ via controllable forces acting in Ω. Problems of this type arise, for example, in the flow control of magnetohydrodynamic processes. In mathematical terms, the problem reads

(5.80)
$$\min J(u,f) := \frac{1}{2} \iint_Q \big|u(x,t) - u_Q(x,t)\big|^2 dx\,dt + \frac{\lambda}{2} \iint_Q \big|f(x,t)\big|^2\,dx\,dt,$$

subject to

(5.81)
$$\begin{aligned} u_t - \frac{1}{Re}\,\Delta u + (u\cdot\nabla)\,u + \nabla p &= f && \text{in } Q \\ \operatorname{div} u &= 0 && \text{in } Q \\ u &= 0 && \text{on } \Sigma \\ u(\cdot,0) &= u_0 && \text{in } \Omega, \end{aligned}$$

and constraints on the control f,

$$f_{a,i}(x,t) \le f_i(x,t) \le f_{b,i}(x,t) \qquad \text{for a.e. } (x,t) \in Q,\ \ i = 1,2. \tag{5.82}$$

We restrict our analysis to two-dimensional spatial domains Ω, since only for this case is there a satisfactory result concerning well-posedness of the problem (5.81). We are given two-dimensional vector functions u_Q, f_a, $f_b \in L^2(0,T;L^2(\Omega)^2)$ with $f_a(x,t) \le f_b(x,t)$ almost everywhere, a divergence-free vector function $u_0 \in L^2(\Omega)^2$, and constants $T > 0$, $\lambda \ge 0$, and $Re > 0$. The force density $f \in L^2(0,T;L^2(\Omega)^2)$ plays the role of the control, while the associated state u represents a velocity field. In this connection, the unknown p denotes the pressure (not the adjoint state).

In practical applications, the *Reynolds number* Re can be very large, which makes the numerical solution of (5.81) a difficult task. In fluid mechanics, the number $1/Re$ is viscosity of the fluid, usually denoted by ν; since, however, ν always denotes the outward unit normal to Γ in this book, we use the letter κ instead, that is, we put $\kappa := 1/Re$.

In the following, we denote by F_{ad} the admissible set, which consists of all controls $f \in L^2(0,T;L^2(\Omega)^2)$ that obey the above constraints. Denoting as before the scalar product in $\mathbb{R}^2$ by the dot $\cdot$, we introduce the spaces

$V := \{v \in H^1(\Omega)^2 : \operatorname{div} v = 0\}$, endowed with the scalar product

$$(u\,,\,v) = \sum_{i=1}^{2} \nabla u_i \cdot \nabla v_i,$$

and $H = \{v \in L^2(\Omega)^2 : \operatorname{div} v = 0\}$. Note that for functions $v \in H$ the requirement $\operatorname{div} v = 0$ is to be understood in the sense of $(v\,,\,\nabla z) = 0$ for all $z \in H^1(\Omega)^2$. In addition, we set

$$\mathcal{W}(0,T) := \{v = (v_1, v_2) \in W(0,T)^2 : \operatorname{div} v_i(t) = 0,\ \ i = 1,2, \\ \text{for a.e. } t \in (0,T)\}\,.$$

The abstract formulation of the state problem is introduced by means of the *trilinear form* $b : V \times V \times V \to \mathbb{R}$,

$$b(u,v,w) = \big((u\cdot\nabla)\,v\,,\,w\big)_{L^2(\Omega)} = \int_\Omega \sum_{i,j=1}^{2} u_i\, w_j\, D_i v_j\, dx.$$

We have $b(u,v,w) = -b(u,w,v)$ and

$$|b(u,v,w)| \le C\,\|u\|_{L^4(\Omega)^2}\,\|v\|_{H^1(\Omega)^2}\,\|w\|_{L^4(\Omega)^2};$$

see Temam **[Tem79]**. From the first relation it follows that $b(u,v,v) = 0$.

In an analogous way as in Section 2.13.1, the operator $A\,u := -\Delta u$ generates a continuous linear mapping from $L^2(0,T;V)$ into $L^2(0,T;V^*)$. Moreover, the operator B defined by

$$\int_0^T \big(B(u)(t)\,,\,w(t)\big)_{V^*,V}\,dt := \int_0^T b\big(u(t),u(t),w(t)\big)\,dt$$

maps $L^2(0,T;V)$ into $L^1(0,T;V^*)$.

Definition. *A vector-valued function $u \in L^2(0,T;V)$ with $u_t \in L^1(0,T;V^*)$ is called a* weak solution *to the initial-boundary value problem* (5.81) *if $u(0) = u_0$ and, in the sense of $L^1(0,T;V^*)$,*

$$u_t + \kappa\,A\,u + B(u) = f.$$

Theorem 5.23 ([**Tem79**])**.** *For any pair $f \in L^2(0,T;V^*)$ and $u_0 \in H$, there exists a unique weak solution $u \in \mathcal{W}(0,T)$ to the problem* (5.81).

Observe that the above theorem yields more regularity than the postulated $u_t \in L^1(0,T;V^*)$. Moreover, it can be shown that the control-to-state mapping $G : L^2(0,T;V^*) \to \mathcal{W}(0,T)$, $f \mapsto u$, is twice Fréchet differentiable.

Optimality conditions. The derivation of first- and second-order optimality conditions can in principle be performed using the methods presented earlier in this chapter. To this end, the solvability of the associated linearized problem and the regularity of its solution have to be investigated. The corresponding analysis may be found in the papers that are referred to at the end of this section. Instead of providing a rigorous analysis, we once more employ the formal Lagrange method to determine what kind of first-order necessary conditions can be expected. These conditions turn out to coincide with those established rigorously in the relevant literature. We introduce the following Lagrangian function:

$$\begin{aligned}\mathcal{L}(u,p,f,w,q) \;&=\; J(u,f) + \iint_Q q \operatorname{div} u \, dx\, dt \\ &- \iint_Q \big(u_t - \kappa\,\Delta u + (u\cdot\nabla)\,u + \nabla p - f\big)\cdot w \, dx\, dt.\end{aligned}$$

While the homogeneous boundary condition for u is encoded in implicit form for the sake of brevity, this is not done with the condition $\operatorname{div} u = 0$, although it would be possible. By virtue of the formal Lagrange method, we expect that at a locally optimal $(\bar u, \bar p, \bar f, w, q)$ the following relations ought to be satisfied: $D_u\mathcal{L} = 0$ for all u with $u(0) = 0$, $D_p\mathcal{L} = 0$, and $D_f\mathcal{L}\,(f-\bar f) \ge 0$ for all $f \in F_{ad}$.

From the second equality it follows that

$$\begin{aligned}0 = D_p\mathcal{L}(\bar u,\bar p,\bar f,w,q)\,p &= -\iint_Q w\cdot\nabla p \, dx\, dt \\ &= -\iint_\Sigma p\, w\cdot\nu \, ds\, dt + \iint_Q p \operatorname{div} w \, dx\, dt\end{aligned}$$

for all sufficiently smooth p. Letting p vary over $C_0^\infty(Q)$, we immediately see that

$$\operatorname{div} w = 0 \quad \text{in } Q,$$

whence we conclude that also $w\cdot\nu = 0$. This relation will be satisfied anyway, since we will see below that $w|_\Sigma = 0$.

Next, we conclude from the condition $D_u\mathcal{L} = 0$ that

(5.83)
$$0 = D_u\mathcal{L}(\bar{u}, \bar{p}, \bar{f}, w, q)\, u = \iint_Q (\bar{u} - u_Q) \cdot u \, dx\, dt - \iint_Q w \cdot u_t \, dx\, dt$$
$$+ \iint_Q q \operatorname{div} u \, dx\, dt + \iint_Q \kappa\, w \cdot \Delta u - \iint_Q \big((\bar{u} \cdot \nabla)\, u + (u \cdot \nabla)\, \bar{u}\big) \cdot w \, dx\, dt$$

for all sufficiently smooth u with $u(0) = 0$ and $u|_\Sigma = 0$. Since $u|_\Sigma = 0$, we infer from the third Green's formula that

$$\iint_Q w \cdot \Delta u \, dx\, dt = \iint_Q \sum_{i=1}^{2} w_i \, \Delta u_i \, dx\, dt$$
$$= \iint_\Sigma \sum_{i=1}^{2} w_i \, \partial_\nu u_i \, ds\, dt + \iint_Q u \cdot \Delta w \, dx\, dt.$$

Moreover,

$$\iint_Q \big((\bar{u} \cdot \nabla)\, u + (u \cdot \nabla)\, \bar{u}\big) \cdot w \, dx\, dt$$
$$= b(\bar{u}, u, w) + b(u, \bar{u}, w) = -b(\bar{u}, w, u) + b(u, \bar{u}, w)$$
$$= \iint_Q \sum_{i,j=1}^{2} \big(- \bar{u}_i \, (D_i w_j)\, u_j + u_i \, (D_i \bar{u}_j)\, w_j \big)\, dx\, dt$$
$$= \iint_Q \Big(- \big((\bar{u} \cdot \nabla)\, w\big) \cdot u + \big((\nabla \bar{u})^\top\, w\big) \cdot u \Big)\, dx\, dt,$$

with the matrix $\nabla \bar{u} := (\nabla \bar{u}_1 \ \nabla \bar{u}_2)^\top$. In addition, because $u(0) = 0$,

$$\iint_Q w \cdot u_t \, dx\, dt = \int_\Omega w(x,T)\, u(x,T)\, dx - \iint_Q u \cdot w_t \, dx\, dt.$$

Invoking all of the above rearrangements of terms, we infer from (5.83) that

$$0 = \iint_Q \Big\{ \bar{u} - u_Q + w_t + \kappa\, \Delta\, w - (\nabla \bar{u})^\top\, w + (\bar{u} \cdot \nabla)\, w - \nabla q \Big\} \cdot u \, dx\, dt$$
$$+ \iint_\Sigma \sum_{i=1}^{2} w_i \, \partial_\nu u_i \, ds\, dt - \int_\Omega w(x,T)\, u(x,T)\, dx,$$

for all relevant u. As before, we deduce from the fact that u, $u(T)$, and $\partial_\nu u_i$ can be freely chosen in Q, Ω, and Σ, respectively, that w is a weak solution to the *adjoint problem*

$$\boxed{\begin{aligned} -w_t - \kappa\,\Delta\, w + (\nabla \bar u)^\top\, w - (\bar u \cdot \nabla)\, w + \nabla q &= \bar u - u_Q && \text{in } Q \\ \operatorname{div} w &= 0 && \text{in } Q \\ w &= 0 && \text{on } \Sigma \\ w(T) &= 0 && \text{in } \Omega. \end{aligned}}$$

Finally, from the derivative of $\mathcal{L}$ with respect to f we infer the *variational inequality*

$$\iint_Q \big(\lambda\, \bar f + w\big) \cdot (f - \bar f)\, dx\, dt \geq 0 \quad \forall f \in F_{ad}.$$

The solution (w, q) to the adjoint problem is to be understood as a weak solution. Here, w enjoys less regularity than expected. In fact, we can guarantee merely that $w \in W^{4/3}(0,T;V)$, where

$$W^{4/3}(0,T;V) = \big\{w \in L^2(0,T,V) : w_t \in L^{4/3}(0,T,V^*)\big\}.$$

The regularity $w \in C([0,T],H)$ can only be obtained under additional assumptions.

References. Since the theory and numerical treatment of optimal flow control problems is currently a very active research area, there are numerous papers devoted to this subject. One of the first groundbreaking papers on the theory of first-order necessary optimality conditions for this class of problems was [**AT90**]. More recent works are, for instance, [**Cas95**], [**GM99**], [**GM00**], [**Hin99**], [**HK01**], and [**Rou02**]. We also refer to the overview given in [**Gun95**]. The technically more difficult case of boundary control was treated in [**HK04**]. Second-order sufficient optimality conditions were studied in [**Hin99**] and [**HK01**] in connection with numerical methods, and [**RT03**] contains the proof that optimal controls of stationary problems depend Lipschitz continuously on perturbations. A weaker version of the second-order conditions, using strongly active control constraints, was presented in [**TW06**]. Model reduction by POD methods is the subject of the papers [**AH01**] and [**KV02**]. An overview of numerical methods for the optimal control of flows was given in [**Gun03**]. Second-order techniques such as the SQP method were addressed in, e.g., [**Hin99**] and [**HK04**]. New grid adaption techniques were proposed in [**BKR00**] and in the review paper [**BR01**]. Various other contributions

to the numerical analysis of optimal control problems are presented or cited in [**HPUU09**].

5.11. Exercises

5.1 Show that the partial Fréchet derivative $F_y(y, v, u)$ defined in the proof of Theorem 5.15 is continuously invertible in $C(\bar{Q})$.

5.2 Show that the second-order remainder r_2^f of the function $f(u) := J(y(u), u)$ used in Theorem 5.17 on page 291 satisfies

$$\frac{r_2^f(\bar{v}, h)}{\|h\|^2_{L^2(Q)}} \to 0 \qquad \text{as } \|h\|_{L^\infty(Q)} \to 0.$$

5.3 Prove that the functional

$$J = \int_0^\ell \phi\big(x, y(x,T)\big)\,dx + \int_0^T \big\{\psi_1\big(t, y(0,t)\big) + \psi_2\big(t, y(\ell,t)\big)\big\}\,dt + \int_0^\ell \int_0^T \varphi(x,t,y,v)\,dx\,dt$$

is twice continuously Fréchet differentiable in $C([0,\ell]\times[0,T]) \times L^2(0,T)$, provided that φ has the form

$$\varphi(x,t,y,v) = \varphi_1(x,t,y) + \varphi_2(x,t,y)\,v + \lambda(x,t)\,v^2,$$

the functions ψ_i and φ_i are sufficiently smooth, and λ is bounded and measurable.

5.4 Verify that the functions $\bar{u}$, $\bar{y}$, and p defined in the test example (5.51)–(5.53) on page 298 jointly satisfy the state problem (5.52), the adjoint problem (5.54), and the corresponding projection relation for $\bar{u}$.

5.5 Examine whether the Nemytskii operator $\Phi : y \mapsto y^3$ from $H^1(\Omega)$ into $L^2(\Omega)$ has first and second Fréchet derivatives. Use the results stated in Section 4.3.3.

Chapter 6

Optimization problems in Banach spaces

6.1. The Karush–Kuhn–Tucker conditions

6.1.1. Convex problems.

The Lagrange multiplier rule. The formal Lagrange method, which was employed repeatedly in the previous chapters, has a rigorous mathematical foundation. In this section, we introduce the basics of this theory needed for understanding problems with state constraints. The corresponding proofs and further results can be found in texts dealing with optimization in general spaces. The theory of convex problems is described in Balakrishnan [**Bal65**], Barbu and Precupanu [**BP78**], and Ekeland and Temam [**ET74**]; nonconvex differentiable problems are treated in, e.g., Ioffe and Tihomirov [**IT79**], Jahn [**Jah94**], Luenberger [**Lue69**], and Tröltzsch [**Trö84b**].

There are numerous books dealing with the theory and numerical treatment of nonlinear differentiable finite-dimensional optimization problems. In this connection, we refer the interested reader to Alt [**Alt02**], Gill et al. [**GMW81**], Grossmann and Terno [**GT97b**], Kelley [**Kel99**], Luenberger [**Lue84**], Nocedal and Wright [**NW99**], Polak [**Pol97**], and Wright [**Wri93**], to name just a few.

In the following, we generally assume that U and Z are real Banach spaces, $G : U \to Z$ is in general a nonlinear mapping, and $C \subset U$ is a nonempty and convex set.

Definition. *A convex set $K \subset Z$ is said to be a* convex cone *if $\lambda z \in K$ whenever $z \in K$ and $\lambda > 0$.*

Any convex cone induces a partial ordering $\geq_K$ in the space Z:

Definition. *Let $K \subset Z$ be a convex cone. We write $z \geq_K 0$ if and only if $z \in K$. Analogously, we write $z \leq_K 0$ if and only if $-z \in K$.*

The elements in K are said to be *nonnegative*. Note, however, that nonnegativity in the sense of this definition does not imply the usual nonnegativity in the set of real numbers, as the following example shows.

Example. Let $Z = \mathbb{R}^3$, and let $K = \{z \in \mathbb{R}^3 : z_1 = 0,\, z_2 \leq 0,\, z_3 \geq 0\}$. Then K is evidently a convex cone, but $z \geq_K 0$ implies nonnegativity only for z_3. ◇

The next definition enables us to introduce a notion of "nonnegativity" also in dual spaces. This notion will be needed for defining Lagrange multipliers, because they are elements of dual spaces.

Definition. *Let $K \subset Z$ be a convex cone. Then the set*

$$K^+ = \big\{z^* \in Z^* : \langle z^*, z\rangle_{Z^*,Z} \geq 0 \quad \forall z \in K\big\}$$

is called the dual cone *of K.*

Examples.
(i) Let $Z = L^2(\Omega)$ with a bounded domain $\Omega \subset \mathbb{R}^N$, and let

$$K = \big\{z \in L^2(\Omega) : z(x) \geq 0 \text{ for a.e. } x \in \Omega\big\}.$$

Here, we have $Z = Z^*$ by the Riesz representation theorem and $K^+ = K$ according to Exercise 6.1.

(ii) Let Z be a Banach space and let $K = \{0\}$. Then $z \geq_K 0$ if and only if $z = 0$, and thus $K^+ = Z^*$; in fact, for any $z^* \in Z^*$ we have $\langle z^*, 0\rangle_{Z^*,Z} = 0 \geq 0$.

(iii) If $K = Z$, then all elements of Z are nonnegative. Hence, $K^+ = \{0\}$ with the zero functional $0 \in Z^*$. ◇

Below, we consider the following optimization problem in a Banach space:

(6.1)
$$\boxed{\begin{aligned} &\min f(u), \\ &G(u) \leq_K 0, \quad u \in C. \end{aligned}}$$

The constraints in (6.1) are viewed differently: as a "complicated" inequality $G(u) \leq_K 0$, which is to be eliminated by means of a Lagrange multiplier, and a "simple" constraint $u \in C$, which is accounted for explicitly. This motivates the following definition.

Definition. *The function* $L : U \times Z^* \to \mathbb{R}$,

$$L(u, z^*) = f(u) + \langle z^* , G(u) \rangle_{Z^*,Z}, \tag{6.2}$$

is called the Lagrangian function. *Any* $(\bar{u}, z^*) \in U \times K^+$ *satisfying the chain of inequalities*

$$L(\bar{u}, v^*) \leq L(\bar{u}, z^*) \leq L(u, z^*) \quad \forall\, u \in C, \ \ \forall\, v^* \in K^+ \tag{6.3}$$

is called a saddle point *of* L. *If this is the case,* z^* *is said to be a* Lagrange multiplier *associated with* $\bar{u}$.

In the previous chapters, when dealing with the optimal control of partial differential equations we denoted the Lagrangian by $\mathcal{L}$. To facilitate the distinction, we use the letter L here. The existence of saddle points is most easily shown for *convex* optimization problems.

Definition. *Let* U *be a Banach space, and let the convex cone* $K \subset Z$ *induce the partial ordering* $\geq_K$ *in the Banach space* Z. *An operator* $G : U \to Z$ *is said to be* convex *(with respect to* $\leq_K$*) if*

$$G(\lambda\, u + (1 - \lambda)\, v) \leq_K \lambda\, G(u) + (1 - \lambda)\, G(v) \quad \forall\, u,\, v \in U, \quad \forall\, \lambda \in (0, 1).$$

Evidently, every linear operator is convex. In the following, we write the strict inequality $z <_K 0$ if and only if $-z$ is an *interior point* of K, that is, $z <_K 0 \Leftrightarrow -z \in \operatorname{int} K$.

Theorem 6.1. *Suppose that a convex functional* $f : U \to \mathbb{R}$, *a convex operator* $G : U \to Z$, *and a solution* $\bar{u}$ *to the problem* (6.1) *are given. Moreover, let there exist some* $\tilde{u} \in C$ *such that* $G(\tilde{u}) <_K 0$, *that is,*

$$-G(\tilde{u}) \in \operatorname{int} K. \tag{6.4}$$

Then there is some $z^* \in K^+$ *such that* $(\bar{u}, z^*)$ *is a saddle point of the Lagrangian* L. *In addition, we have the complementary slackness condition*

$$\langle z^* , G(\bar{u}) \rangle_{Z^*,Z} = 0. \tag{6.5}$$

The proof of the above theorem can be found in, e.g., Luenberger [**Lue69**]. In the literature, the condition (6.4) is usually referred to as the *Slater condition*. It can only be satisfied if the cone K has nonempty interior. This excludes, for instance, the case $K = \{0\}$, which corresponds to the equality constraint $G(u) = 0$. In this case, the above theorem fails to apply, but other existence results concerning Lagrange multipliers are available. The lack of interior points is a much more serious problem in the following situation.

Example. Consider the natural nonnegative cone in $Z = L^2(0,1)$,

$$K = \big\{ z(\cdot) \in L^2(0,1) \,:\, z(x) \geq 0 \ \text{ for a.e. } \ x \in (0,1) \big\}.$$

Quite unexpectedly, we have $\operatorname{int} K = \emptyset$. How can this be possible? One is tempted to believe that, for instance, $z(x) \equiv 1$ is an interior point of K. Unfortunately, this is not true. In fact, the sequence $\{v_n\}_{n=1}^{\infty} \subset L^2(\Omega)$ with

$$v_n(x) = \begin{cases} \ \ 1 & \text{in} \quad \big[0\,,\, 1 - \frac{1}{n}\big) \\ -1 & \text{in} \quad \big[1 - \frac{1}{n}\,,\, 1\big], \end{cases}$$

while obviously converging to z with respect to the L^2 norm, is not contained in K. Consequently, $z \notin \operatorname{int} K$. This undesired behavior is simply a consequence of the fact that the L^2 norm, and likewise any other L^p norm with $1 \leq p < \infty$, measures an integral and not the maximal absolute value of a function. This fact constitutes a major obstacle in the treatment of optimization problems in function spaces. ◇

Theorem 6.2. *Suppose that the mappings f and G in Theorem 6.1 are Gâteaux differentiable at $\bar{u}$. Then we have the variational inequality*

$$D_u L(\bar{u}, z^*)\,(u - \bar{u}) \geq 0 \quad \forall\, u \in C.$$

Here and in the following, D_u again denotes the partial Gâteaux or Fréchet derivative with respect to u. The assertion is an immediate consequence of the saddle point condition (6.3), which implies that $\bar{u}$ solves the problem without the constraint $G(u) \leq_K 0$, namely

$$L(\bar{u}, z^*) = \min_{u \in C} L(u, z^*).$$

The associated variational inequality reads, in explicit form,

$$f'(\bar{u})(u - \bar{u}) + \langle z^*\,,\, G'(\bar{u})(u - \bar{u}) \rangle_{Z^*,Z} \geq 0 \quad \forall\, u \in C$$

or, equivalently,

$$\langle f'(\bar{u}) + G'(\bar{u})^* z^*\,,\, u - \bar{u} \rangle_{U^*,U} \geq 0 \quad \forall\, u \in C.$$

In the unconstrained case where $C = U$, we get the equation

$$f'(\bar{u}) + G'(\bar{u})^* z^* = 0 \in U^*.$$

Examples. We illustrate the application and limitations of the above theorems by means of simple examples that do not involve partial differential equations.

One-sided box constraints in $L^2(0,1)$. Let $u_d \in L^2(0,1)$ be given. We consider the minimization problem

$$\min f(u) := \frac{1}{2} \int_0^1 \big|u(x) - u_d(x)\big|^2 \, dx, \tag{6.6}$$

subject to

$$u(x) \le 0 \qquad \text{for a.e. } x \in (0,1).$$

The above problem is a special case of problem (6.1), with the specifications $U = Z = L^2(0,1)$ and $G = I$ (the identity mapping). The associated convex cone K is the set of almost-everywhere nonnegative elements of $L^2(0,1)$, and we have $C = U$.

The problem has a unique minimizer $\bar{u}$, namely,

$$\bar{u}(x) = \min\{u_d(x), 0\}.$$

We investigate whether there exists an associated Lagrange multiplier. The corresponding Lagrangian function reads

$$L(u,\mu) = f(u) + \big(\mu, G(u)\big)_{L^2(0,1)} = \int_0^1 \Big(\frac{1}{2}\big(u(x) - u_d(x)\big)^2 + \mu(x)\, u(x)\Big)\, dx.$$

Here, by the Riesz representation theorem, the functional $z^* \in Z^*$ has been identified with some function $\mu \in L^2(0,1)$. We search for a Lagrange multiplier $\mu \in L^2(0,1)$. Since $\operatorname{int} K = \emptyset$, Theorem 6.1 does not apply. Instead, the Lagrange multiplier is constructed using a pointwise approach. To this end, recall that, owing to Lemma 2.21 on page 63, we have the variational inequality

$$\int_0^1 \big(\bar{u}(x) - u_d(x)\big)\big(u(x) - \bar{u}(x)\big)\, dx \ge 0 \qquad \forall\ u(\cdot) \le 0.$$

This can only be true if the implications

$$\begin{aligned} \bar{u}(x) < 0 \quad &\Rightarrow \quad \bar{u}(x) - u_d(x) = 0 \\ \bar{u}(x) = 0 \quad &\Rightarrow \quad \bar{u}(x) - u_d(x) \le 0 \end{aligned}$$

are valid almost everywhere. But then $\bar{u}(x) - u_d(x)$ must be nonpositive almost everywhere. Since we have used arguments of this kind repeatedly in Chapter 2, we do not explain this in detail here. We now define

$$\mu(x) := -f'(\bar{u})(x) = -\big(\bar{u}(x) - u_d(x)\big).$$

Then, owing to the above implications, we have $\mu \geq 0$ as well as

$$\mu(x)\,\bar{u}(x) = 0 \quad \text{for a.e. } x \in (0,1),$$

which is the pointwise form of the complementary slackness condition (6.5). Finally, it follows from the definition of μ that $\mu = -f'(\bar{u})$, that is,

$$f'(\bar{u}) + \mu = 0,$$

which, in turn, is equivalent to the equation $D_u L(\bar{u}, \mu) = 0$. Hence, the function μ defined above is a Lagrange multiplier. ◇

Two-sided box constraints in $L^2(0,1)$. We now consider the above minimization problem with the same functional f as in (6.6), but this time with constraints from both above and below:

$$\min f(u), \tag{6.7}$$

subject to

$$-1 \leq u(x) \leq 1 \quad \text{for a.e. } x \in (0,1).$$

Again, we put $C = U = L^2(0,1)$, and we cast the constraints in the form

$$u(x) - 1 \leq 0, \qquad -u(x) - 1 \leq 0.$$

We then have to choose $Z = L^2(0,1) \times L^2(0,1)$ and $K = L^2(0,1)_+ \times L^2(0,1)_+$, where $L^2(0,1)_+$ denotes the set of almost-everywhere nonnegative elements of $L^2(0,1)$. The convex operator $G : L^2(0,1) \to L^2(0,1) \times L^2(0,1)$ is defined by

$$G(u) = \begin{pmatrix} u(\cdot) - 1 \\ -u(\cdot) - 1 \end{pmatrix}.$$

Although the function $\tilde{u}(x) \equiv 0$ obeys both inequalities strictly, we again have $\operatorname{int} K = \emptyset$ and thus cannot employ Theorem 6.1. However, the construction used in Section 1.4.7 works. The Lagrangian is now given by

$$\begin{aligned} L(u,\mu) &= L(u,\mu_a,\mu_b) \\ &= \frac{1}{2}\|u - u_d\|^2_{L^2(0,1)} + (-u - 1\,,\,\mu_a)_{L^2(0,1)} + (u - 1\,,\,\mu_b)_{L^2(0,1)}. \end{aligned}$$

We make the pointwise definitions

$$\begin{aligned} \mu_a(x) &= \big(f'(x)\big)_+ = \big(\bar{u}(x) - u_d(x)\big)_+ \\ \mu_b(x) &= \big(f'(x)\big)_- = \big(\bar{u}(x) - u_d(x)\big)_-, \end{aligned} \tag{6.8}$$

where, as usual, $z_+ = (z + |z|)/2$ and $z_- = (|z| - z)/2$. Obviously, μ_a and μ_b are nonnegative. The reader will be asked in Exercise 6.2 to check that the arguments from Section 1.4.7 carry over almost unchanged to give

$$D_u L(\bar{u}, \mu) = f'(\bar{u}) + \mu_b - \mu_a = 0$$

and the slackness conditions

$$(-\bar{u} - 1\,,\, \mu_a)_{L^2(0,1)} = (\bar{u} - 1\,,\, \mu_b)_{L^2(0,1)} = 0.$$

Consequently, μ_a and μ_b are Lagrange multipliers for $\bar{u}$.

If we assume $u_d \in L^\infty(0,1)$, then both multipliers belong to $L^\infty(0,1)$. This nice byproduct of the pointwise construction follows from the fact that $\bar{u} - u_d \in L^\infty(0,1)$. ◇

Remark. The problem with two-sided constraints could also be considered in the space $L^\infty(0,1)$, since in this case every admissible control u is automatically bounded and measurable. Moreover, the cone K of nonnegative functions in $L^\infty(0,1)$ has interior points, and $\tilde{u}(x) \equiv 0$ satisfies the Slater condition. Theorem 6.1 then yields the existence of Lagrange multipliers $\mu_a, \mu_b \in L^\infty(0,1)^*$. However, we do not gain much benefit from this result, since $L^\infty(0,1)^*$ is a space of continuous linear functionals that need not even be measures.

6.1.2. Differentiable problems.

Lagrange multiplier rules and constraint qualifications. We now investigate the problem (6.1) without assuming f and G to be convex. We consider

$$\min f(u), \qquad G(u) \leq_K 0, \quad u \in C,$$

where C is still convex. Instead of convexity, we postulate the Fréchet differentiability of f and G. We use the same Lagrangian function $L = L(u, z^*)$ as in Section 6.1.1, but, in view of the nonconvexity, we can no longer expect a saddle point property to be valid. Therefore, Lagrange multipliers are defined in a slightly different way.

Definition. *Let $\bar{u} \in U$ be admissible. We call $\bar{u}$ a* local solution *of the minimization problem* (6.1) *if there is some $\varepsilon > 0$ such that*

$$f(\bar{u}) \le f(u) \qquad \forall\, u \in C \ \text{ with } \ G(u) \le_K 0 \ \text{ and } \ \|u - \bar{u}\|_U \le \varepsilon.$$

Definition. *Let $\bar{u}$ be a local solution to the problem* (6.1). *Then any $z^* \in K^+$ satisfying the conditions*

$$D_u L(\bar{u}, z^*)(u - \bar{u}) \ \ge \ 0 \qquad \forall\, u \in C \tag{6.9}$$

$$\langle z^* \,,\, G(\bar{u}) \rangle_{Z^*,Z} \ = \ 0 \tag{6.10}$$

is called a Lagrange multiplier *associated with $\bar{u}$.*

In order that the existence of such a Lagrange multiplier be guaranteed, a so-called *constraint qualification* must be postulated. Since such a condition involves the locally optimal control itself, it usually cannot be verified without knowledge of this function. There are various constraint qualifications. A rather general one, which suffices for our purposes, is the *Zowe–Kurcyusz condition* (see Zowe and Kurcyusz **[ZK79]**).

Definition. *Suppose that $\bar{u} \in C$ with $G(\bar{u}) \le_K 0$ is given. We call the sets*

$$C(\bar{u}) := \big\{ \alpha\,(u - \bar{u}) \,:\, \alpha \ge 0,\, u \in C \big\}, \qquad K(\bar{z}) := \big\{ \beta\,(z - \bar{z}) \,:\, \beta \ge 0,\, z \in K \big\}$$

the conical hulls *to C and K at $\bar{u}$ and $\bar{z}$, respectively. The condition*

$$\boxed{G'(\bar{u})\, C(\bar{u}) + K\big(-G(\bar{u})\big) = Z} \tag{6.11}$$

is called the Zowe–Kurcyusz constraint qualification.

The above relation is obviously equivalent to saying that for any $z \in Z$ the equation

$$\alpha\, G'(\bar{u})(u - \bar{u}) + \beta\,\big(v + G(\bar{u})\big) = z \tag{6.12}$$

is solvable with suitable $u \in C$, $v \ge_K 0$, $\alpha \ge 0$, and $\beta \ge 0$. Recall that $v \ge_K 0$ if and only if $v \in K$.

Theorem 6.3. *Let $\bar{u}$ be a local solution to problem* (6.1), *and let f and G be continuously Fréchet differentiable in an open neighborhood of $\bar{u}$. If*

the constraint qualification (6.11) *holds, then there exists a Lagrange multiplier* $z^* \in Z^*$ *associated with* $\bar{u}$*. Moreover, the set of Lagrange multipliers associated with* $\bar{u}$ *is bounded.*

The proof of this multiplier rule is due to Zowe and Kurcyusz [**ZK79**]. From (6.9) it follows that

$$\langle f'(\bar{u}) + G'(\bar{u})^* \, z^* \, , \, u - \bar{u} \rangle_{U^*,U} \geq 0 \quad \forall \, u \in C. \tag{6.13}$$

Remark. Sometimes it is difficult or even meaningless to establish $G'(\bar{u})^*$ in explicit form. Then (6.13) is replaced by the equivalent inequality

$$f'(\bar{u}) \, (u - \bar{u}) + \langle z^* \, , \, G'(\bar{u}) \, (u - \bar{u}) \rangle_{Z^*,Z} \geq 0 \qquad \forall \, u \in C.$$

Example. Of particular interest is the minimization problem with both equality and set constraints:

$$\min \, f(u), \quad G(u) = 0, \quad u \in C, \tag{6.14}$$

where f, G, and C are defined as before. In this special case, the constraint qualification (6.11) reads

$$G'(\bar{u}) \, C(\bar{u}) = Z. \tag{6.15}$$

If it is satisfied, then a Lagrange multiplier $z^* \in Z^*$ exists such that the variational inequality (6.13) is valid. The complementary slackness condition (6.10) is meaningless for equality constraints.

In Section 6.1.3, we will apply this result to the special case of $G(u) = A \, y - B \, v = 0$, where $A : Y \to Y^*$ is a continuously invertible operator representing an elliptic differential operator, y denotes the state, and $v \in V_{ad} \subset V$ is the control.

In this case, we have $Z := Y^*$, $U := Y \times V$, and $C := Y \times V_{ad}$. The constraint qualification is always satisfied, since the equation

$$G'(\bar{u}) \, (u - \bar{u}) = A \, (y - \bar{y}) + B \, (v - \bar{v}) = z$$

is solvable for any $z \in Z = Y^*$ with $v = \bar{v}$ and $y = A^{-1} \, z + \bar{y}$. The element $u - \bar{u} = (y - \bar{y}, \bar{v} - \bar{v})$ belongs to the cone $C(\bar{u})$. ◇

Discussion of the Zowe–Kurcyusz constraint qualification. In the following, we illustrate the application of condition (6.11) for various types of constraints, first in the general situation, and then for pointwise constraints in function spaces.

Pure equality constraints $G(u) = 0$. With $C = U$ and $K = \{0\}$, (6.11) becomes

$$G'(\bar{u})\, U = Z. \tag{6.16}$$

In other words, the operator $G'(\bar{u})$ must be surjective. This surjectivity requirement comes from the classical Lagrange multiplier rule for equality constraints. The relation (6.13) attains the form

$$f'(\bar{u}) + G'(\bar{u})^*\, z^* = 0. \tag{6.17}$$

Inequality constraints. Let the constraints be given as in (6.1). If the minimizer $\bar{u}$ satisfies $G(\bar{u}) <_K 0$, that is, if $-G(\bar{u}) \in \operatorname{int} K$, then the constraint qualification (6.11) is fulfilled; the reader will be asked to verify this in Exercise 6.3. Since the constraint is not active, this case is not interesting.

The following *linearized Slater condition* is sufficient for the Zowe–Kurcyusz constraint qualification (6.11) to hold:

$$\boxed{\exists\, \tilde{u} \in C \,:\, G(\bar{u}) + G'(\bar{u})(\tilde{u} - \bar{u}) <_K 0.} \tag{6.18}$$

This is easily seen: the Zowe–Kurcyusz condition postulates for any $z \in Z$ the existence of constants $\alpha \geq 0$, $\beta \geq 0$ and elements $k \in K$, $u \in C$ such that the equation

$$\alpha\big(G'(\bar{u})(u - \bar{u}) + \beta(k + G(\bar{u})\big) = z$$

is valid. To show this, put $\alpha = \beta$, $u = \tilde{u}$, and $\bar{z} = G(\bar{u}) + G'(\bar{u})(\tilde{u} - \bar{u})$. Then the above equation reduces to $\alpha(\bar{z} + k) = z$ and, since K is a cone, to

$$\alpha\bar{z} + q = z,$$

with $q \in K$. Now we choose α so large that $z - \alpha\bar{z} \geq_K 0$. This is possible, because by (6.18) $\bar{z}$ lies in the interior of $-K$. With this, we satisfy the above condition with the choice $q = z - \alpha\bar{z} \geq_K 0$.

If both K and C have interior points, then (6.18) is equivalent to the following condition (cf. Penot [**Pen82**]):

$$\exists\, h \in \operatorname{int} C(\bar{u}) \,:\, G(\bar{u}) + G'(\bar{u})\, h <_K 0. \tag{6.19}$$

Equality and inequality constraints. Suppose that the constraints have the form

$$G_1(u) = 0, \quad G_2(u) \leq_K 0, \quad u \in C.$$

Then the following condition is sufficient for (6.11) to hold (cf. [**HPUU09**], Lemma 1.14): $G_1'(\bar{u})$ is surjective, and

$$\exists\, h \in C(\bar{u}) \,:\, G_1'(\bar{u})\, h = 0, \quad G_2(\bar{u}) + G_2'(\bar{u})\, h <_K 0. \tag{6.20}$$

As the following examples will show, the applicability of the Zowe–Kurcyusz constraint qualification to inequality constraints in function spaces is essentially restricted to cones of nonnegative functions with nonempty interior.

One-sided box constraints for u. We begin our analysis with a problem involving one-sided constraints:

$$\min\, f(u) := \int_\Omega \psi\big(x, u(x)\big)\, dx, \qquad u(x) \leq u_b(x) \quad \text{for a.e. } x \in \Omega. \tag{6.21}$$

Here, $u_b \in L^\infty(\Omega)$ is given, and the function ψ is sufficiently smooth and satisfies a suitable growth condition in order to guarantee that the given integral functional f be continuously differentiable in $U = L^2(\Omega)$. The above minimization problem is of the form

$$\min f(u), \qquad G(u) \leq_K 0,$$

with $G(u)(x) := u(x) - u_b(x)$. As an affine continuous operator, G is differentiable from U into $Z = U$. The cone K is given by the set of almost-everywhere nonnegative elements of $L^2(\Omega)$.

In this case, the Zowe–Kurcyusz constraint qualification (6.11) is satisfied: in view of $C = L^2(\Omega)$, we have $C(\bar{u}) = L^2(\Omega)$. And since $G'(\bar{u})$ is the identity mapping, for any $z \in L^2(\Omega)$ there is some $u \in L^2(\Omega) = C(\bar{u})$ such that $G'(\bar{u})\, u = z$: one simply chooses $u = z$.

Consequently, Theorem 6.3 may be applied in $L^2(\Omega)$, where, in view of the Riesz representation theorem, every $z^* \in L^2(\Omega)^*$ can be identified with some $\mu \in L^2(\Omega)$. Hence, for any local solution $\bar{u}$ there exists some almost-everywhere nonnegative multiplier $\mu \in L^2(\Omega)$ such that $f'(\bar{u}) + G'(\bar{u})^*\, \mu = 0$. We may identify $f'(\bar{u})$ with the function $\psi_u(\cdot, \bar{u}(\cdot)) \in L^2(\Omega)$, and $G'(\bar{u})^*$ is the identity operator in $L^2(\Omega)$. We therefore find that

$$\psi_u\big(x, \bar{u}(x)\big) + \mu(x) = 0, \quad \mu(x) \geq 0, \qquad \text{for a.e. } x \in \Omega.$$

In this example, the Zowe–Kurcyusz condition was applicable even though the cone of nonnegative functions in $L^2(\Omega)$ had empty interior. This is in a certain sense an exceptional case. Alternatively, we could have constructed the multiplier directly as in (6.8) by setting $\mu(x) = \big(\psi_u(x, \bar{u}(x))\big)_-$.

Two-sided box constraints for u. We consider the same problem as above, but this time with the two-sided control constraints

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

with bounded and measurable functions $u_a \le u_b$. We fit these constraints into the abstract framework of (6.1) by choosing the operator G to be of the form

$$G(u) = \begin{pmatrix} u_a - u \\ u - u_b \end{pmatrix}.$$

Evidently, G is a continuously differentiable mapping from $L^2(\Omega)$ into $L^2(\Omega) \times L^2(\Omega)$. However, the Zowe–Kurcyusz constraint qualification cannot be directly satisfied in the form (6.11), as can be shown with a little effort. Again, we have the problem that the cone of nonnegative functions in $L^2(\Omega)$ has empty interior. It would also not be helpful to work in $L^\infty(\Omega)$ instead, since then we would at best obtain measures as multipliers for the control constraints. As in (6.8), a possible way out is to define Lagrange multipliers by

$$\mu_a(x) := \psi_u\big(x, \bar{u}(x)\big)_+, \qquad \mu_b(x) := \psi_u\big(x, \bar{u}(x)\big)_-,$$

with which the Karush–Kuhn–Tucker conditions are fulfilled.

Second-order optimality conditions. The scope of the Karush–Kuhn–Tucker theory in Banach spaces also encompasses second-order necessary and sufficient optimality conditions. For illustration, we only discuss the problem (6.14):

$$\min f(u), \quad G(u) = 0, \quad u \in C,$$

additionally assuming that f and G are twice continuously Fréchet differentiable. Suppose that $\bar{u}$ satisfies, together with $z^* \in Z^*$, the first-order necessary condition

$$f'(\bar{u})(u - \bar{u}) + \langle z^*\,,\, G'(\bar{u})(u - \bar{u})\rangle_{Z^*,Z} \ge 0 \qquad \forall\, u \in C. \tag{6.22}$$

Moreover, let there exist some $\delta > 0$ such that

$$L''(\bar{u}, z^*)[u,u] := f''(\bar{u})[u,u] + \langle z^*\,,\, G''(\bar{u})[u,u]\,\rangle_{Z^*,Z} \ge \delta\, \|u\|_U^2 \tag{6.23}$$

for all $u \in C(\bar{u})$ such that

$$G'(\bar{u})\, u = 0. \tag{6.24}$$

Lemma 6.4. *If $\bar{u}$ is admissible for problem* (6.14) *and the conditions* (6.22)–(6.24) *are fulfilled, then $\bar{u}$ is locally optimal for* (6.14).

These second-order sufficient optimality conditions follow from general results due to Maurer and Zowe [**MZ79, Mau81**]. The lemma applies only to problems in which the two-norm discrepancy does not play a role. Since we have not provided any further information concerning the structure of the set C, we are not in a position to define and make use of strongly active constraints. The conditions above are thus too restrictive. In the case of inequality constraints of the form $G(u) \leq_K 0$, first-order sufficient optimality conditions can also be employed; see [**MZ79**]. For partial differential equations with state constraints, we refer to [**CDlRT08**]. In the case of pointwise constraints in function spaces, usually strongly active sets in the sense of Dontchev et al. [**DHPY95**] are used for this purpose.

6.1.3. A semilinear elliptic problem. Let $\Omega \subset \mathbb{R}^3$, $N \leq 3$, be a bounded Lipschitz domain. For given $v \in L^2(\Omega)$, we consider the elliptic boundary value problem

$$\begin{aligned} -\Delta y + y + y^3 &= v \quad \text{in } \Omega \\ \partial_\nu y &= 0 \quad \text{on } \Gamma. \end{aligned}$$

As shown on pages 181 to 183, this problem is easy to handle in the state space $Y = H^1(\Omega)$. Introducing the mapping $A : Y \to Y^*$ generated by the elliptic operator $-\Delta + I$, the Nemytskii operator $\Phi : Y \to V = L^2(\Omega)$, $y(\cdot) \mapsto y(\cdot)^3$, and the embedding operator $B : L^2(\Omega) \to Y^*$, we can transform the above boundary value problem into the equation $A\, y + B\, \Phi(y) = B\, v$ in Y^*.

In the following, we are going to demonstrate how Theorem 6.3 and Lemma 6.4 can be applied to a corresponding optimal control problem. To this end, we study the minimization of

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

subject to the above elliptic state problem and to the control constraint $-1 \leq v(x) \leq 1$ for almost every $x \in \Omega$. With the embedding operator $E_Y : H^1(\Omega) \to L^2(\Omega)$ and the admissible set

$$V_{ad} = \{v \in L^2(\Omega) : \, -1 \leq v(x) \leq 1 \quad \text{for a.e. } x \in \Omega\},$$

we obtain the problem

$$\min J(y,v) := \frac{1}{2}\,\|E_Y\,y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|v\|^2_{L^2(\Omega)}, \tag{6.25}$$

subject to

$$A\,y + B\,(\Phi(y) - v) = 0, \qquad v \in V_{ad}. \tag{6.26}$$

Obviously, this is a special case of the problem (6.14),

$$\min J(u), \qquad G(u) = 0, \quad u \in C,$$

with the specifications $U := Y \times V$, $u := (y,v)$, $G : Y \to Y^*$, $G(u) := A\,y + B\,(\Phi(y) - v)$, and $C := Y \times V_{ad}$.

First-order necessary conditions. According to the results established in Section 4.3.3, the Nemytskii operator Φ is continuously differentiable from $H^1(\Omega)$ into $L^2(\Omega)$. This implies that the operator G is continuously Fréchet differentiable from Y into Y^*.

Moreover, we claim that for any $\bar u = (\bar y, \bar v) \in U$, $G'(\bar u)$ is a surjective mapping. To prove this claim, we first observe that $G'(\bar u)(y,v) = A\,y + B\,(\Phi'(\bar y)\,y - v)$. Now consider, for an arbitrary right-hand side $z \in V^*$, the equation

$$A\,y + B\,(\Phi'(\bar y)\,y - v) = z. \tag{6.27}$$

On the left-hand side, we have the differential operator $\tilde A$,

$$\tilde A\,y = -\Delta\,y + y + 3\,\bar y^2\,y.$$

By virtue of Theorem 4.7 on page 191, $\bar y$ is continuous and thus bounded. Hence, the coefficient function $c_0(x) = 1 + 3\,\bar y(x)^2$ is both positive and bounded. It therefore follows that the bilinear form $a[\cdot,\cdot]$ generated by $\tilde A$ meets the conditions of the Lax–Milgram lemma. From this, we readily deduce that the equation (6.27) admits a solution for any $z \in Z = Y^*$: we simply put $v = 0$ and determine the unique solution $y \in V$ to the equation $A\,y + B\,\Phi'(\bar y)\,y = z$. This proves the claim, and $G'(\bar u)$ is thus surjective.

In addition, in Exercise 6.5 the interested reader will have the opportunity to check that the constraint qualification (6.15) on page 331 is also fulfilled. Consequently, Theorem 6.3 applies, yielding the existence of a Lagrange multiplier $z^* \in Z^* = (Y^*)^* = Y$. Putting $p := -z^*$, we obtain for the associated Lagrangian function

$$L(u,p) = L(y,v,p) = J(y,v) - \big(A\,y + B\,(\Phi'(\bar y)\,y - v)\,,\,p\big)_{Y^*,Y}\,.$$

The first minus sign is needed so that the adjoint problem will take on the form familiar to us from the previous chapters. If $\bar{u} = (\bar{y}, \bar{v})$ is locally optimal, that is, if $\bar{v}$ is a locally optimal control, then, by virtue of Theorem 6.3, the variational inequality

$$D_{(y,v)}L(\bar{y}, \bar{v}, p)(y - \bar{y}, v - \bar{v}) \geq 0 \quad \forall\, (y, v) \in C$$

is valid. In terms of y, this evidently means that $D_y L(\bar{y}, \bar{v}, p)\, y = 0$ for all $y \in Y$, that is,

$$(\bar{y} - y_\Omega\,,\, y) - (A^* p\,,\, y) - (B^*\, p\,,\, \Phi'(\bar{y})\, y) = 0 \ \ \forall\, y \in H^1(\Omega).$$

Now, $B^* = E_Y$ and $A^* = A$. Therefore,

$$\bar{y} - y_\Omega - A\, p - \Phi'(\bar{y})^* B^*\, p = 0,$$

and the Lagrange multiplier p turns out to be the unique solution to the adjoint boundary value problem

$$\begin{aligned} -\Delta p + p + 3\,\bar{y}^2 p &= \bar{y} - y_\Omega \qquad \text{in } \Omega \\ \partial_\nu\, p &= 0 \qquad\qquad \text{on } \Gamma. \end{aligned}$$

Evaluating the above variational inequality for $\bar{v}$, we find that

$$D_v L(\bar{y}, \bar{v}, p)\,(v - \bar{v}) \geq 0 \quad \forall\, v \in V_{ad},$$

and we finally arrive at the variational inequality

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

The above optimality conditions also resulted from the analysis of the optimal control problem (4.31)–(4.33) on page 207, which was performed in the state space $H^1(\Omega) \cap C(\bar{\Omega})$. The method presented here has the advantage that the conditions can be derived directly from the optimization theory in Banach spaces. It only works in this simple form if the given nonlinearity is differentiable in $H^1(\Omega)$. Since $H^1(\Omega) \not\subset C(\bar{\Omega})$ for $N \geq 2$, it does not apply to problems with pointwise state constraints.

Remark. In applying the general result Theorem 6.3 in function spaces, usually a compromise has to be made between two conflicting restraints: in order that the constraint qualification be valid and, at the same time, the nonlinearities be differentiable, the range space Z should not be too large; on the other hand, Z also should not be too small, since otherwise the dual space Z^* becomes too large in the sense that it contains functions of low regularity that can no longer be interpreted as (weakly differentiable) solutions to adjoint problems.

Second-order sufficient conditions. Since the functional J and the Nemytskii operator Φ are twice continuously Fréchet differentiable in $H^1(\Omega) \times L^2(\Omega)$ and $H^1(\Omega)$, respectively, so is the Lagrangian. Thus, in view of Lemma 6.4, the following condition is sufficient for local optimality: the pair $(\bar{u}, \bar{v})$ satisfies both the first-order necessary conditions and the definiteness condition

$$L''(\bar{y}, \bar{v}, p)\big[(y, v), (y, v)\big] \geq \delta \left(\|y\|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Omega)} \right)$$

for all pairs (y, v) satisfying the boundary value problem

$$\begin{aligned} -\Delta y + y + 3\,\bar{y}^2\, y &= v \qquad \text{in } \Omega \\ \partial_\nu\, y &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

Then $\bar{v}$ is locally optimal in the sense of the norm of $L^2(\Omega)$. The above definiteness condition is already valid if we merely have, with a modified δ,

$$L''(\bar{y}, \bar{v}, p)\big[(y, v), (y, v)\big] \geq \delta \, \|v\|^2_{L^2(\Omega)}.$$

The explicit expression for the second derivative L'' is

$$L''(\bar{y}, \bar{v}, p)\big[(y, v), (y, v)\big] = \|y\|^2_{L^2(\Omega)} + \lambda \, \|v\|^2_{L^2(\Omega)} - 6 \int_\Omega p\, \bar{y}\, y^2 \, dx.$$

6.2. Control problems with state constraints

State constraints naturally arise in many applications. A typical example is that of heating problems in which the temperature is forbidden to exceed or fall short of certain prescribed threshold values. Such problems raise interesting, and in parts still unsolved, mathematical questions. Here, we shall only briefly address some basic ideas in order to enable the reader to consult the relevant literature for a more in-depth study. For a comprehensive treatment of the elliptic case, we refer the reader to Neittaanmäki et al. [**NST06**]. For simplicity, we confine ourselves to elliptic problems; the theory for parabolic problems is quite similar.

The necessary optimality conditions to be proved below may also be derived from Pontryagin's maximum principle for state-constrained elliptic problems; for this purpose, the maximum condition is transformed into a variational inequality. In the case of boundary controls, the corresponding maximum principle was proved by Alibert and Raymond [**AR97**] and by Casas [**Cas93**]. The same applies to state-constrained parabolic problems, which were treated in Casas [**Cas97**] and in Raymond and Zidani [**RZ99**]. However, the proof of Pontryagin's maximum principle is very technical, while the optimality conditions to be presented here can be obtained much

more simply by means of the Lagrange method in Banach spaces. This technique was employed also in the works of Casas [**Cas86**] and Tröltzsch [**Trö84b**]. In the following, we apply it to derive first-order necessary conditions. We do not pursue second-order sufficient conditions, referring the reader to the papers [**CTU00**] and [**RT00**]. Thus far, second-order sufficient conditions for problems with pointwise state constraints in the whole domain could only be shown for low-dimensional domains; see Casas et al. [**CDlRT08**].

6.2.1. Convex problems.

An elliptic problem with pointwise state constraints. Let $\Omega \subset \mathbb{R}^N$ denote a bounded Lipschitz domain. We consider the optimal control problem

$$\min J(y,u) := \frac{1}{2}\,\|y - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)}, \tag{6.28}$$

subject to

$$\boxed{\begin{aligned} -\Delta y + y &= u && \text{in } \Omega \\ \partial_\nu\, y &= 0 && \text{on } \Gamma \end{aligned}} \tag{6.29}$$

and the constraints

$$\begin{aligned} &u_a(x) \le u(x) \le u_b(x) && \text{for a.e. } x \in \Omega, \\ &y(x) \le 0 && \forall\, x \in \bar\Omega. \end{aligned} \tag{6.30}$$

We are given $y_\Omega \in L^2(\Omega)$, $\lambda \ge 0$, and $u_a, u_b \in L^\infty(\Omega)$ such that $u_a \le u_b$ almost everywhere. In addition to the usual box constraints for the control, a *pointwise state constraint* for y is imposed. By virtue of Lemma 4.6 on page 190, for every $u \in L^r(\Omega)$ with $r > N/2$, the elliptic boundary value problem (6.29) has a unique weak solution $y \in H^1(\Omega) \cap C(\bar\Omega)$. We therefore choose for the controls u the space $U = L^r(\Omega)$ with some arbitrary $r > N/2$.

The control-to-state mapping $G:\ u \mapsto y$ is considered as a map between two different pairs of spaces: in the cost functional, y appears as an L^2 function; there, we define $y = S\,u = E_Y\,G\,u$, with the embedding operator $E_Y : H^1(\Omega) \hookrightarrow L^2(\Omega)$. We then have $S : L^r(\Omega) \to L^2(\Omega)$.

In the state constraint (6.30), the continuity of y is exploited, because the cone K of nonnegative functions in $C(\bar\Omega)$ has interior points. Here, we regard $u \mapsto y$ as a mapping from $L^r(\Omega)$ into $C(\bar\Omega)$. To avoid the introduction of further notation, we denote this mapping also by G, although we had introduced $H^1(\Omega) \cap C(\bar\Omega)$ as the range space of G. From this, no confusion should arise.

With the specifications $U = L^r(\Omega)$, $Z = C(\bar{\Omega})$, and $K = \big\{y \in C(\bar{\Omega}) : y(x) \geq 0 \quad \forall x \in \bar{\Omega}\big\}$, the optimal control problem (6.28)–(6.30) then turns out to be a special case of the optimization problem (6.1), namely,

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

subject to

$$G(u) \leq_K 0, \qquad u \in C,$$

where $C := U_{ad} = \big\{u \in L^r(\Omega) : u_a(x) \leq u(x) \leq u_b(x) \text{ for a.e. } x \in \Omega\big\}$.

Through the representation

$$\langle z^*\,,\, y\rangle_{Z^*,Z} = \int_{\bar{\Omega}} y(x)\,d\mu(x),$$

the dual space Z^* can be identified with the space $M(\bar{\Omega})$ of all regular Borel measures μ defined on $\bar{\Omega}$; see Alt [**Alt99**]. Note that in view of the compactness of $\bar{\Omega}$, the notion of *Radon measure*, which is also commonly used, is equivalent to the notion of regular Borel measure.

We thus identify z^* with $\mu \in M(\bar{\Omega})$. The nonnegativity $z^* \geq_{K^+} 0$ then means that

$$\int_{\bar{\Omega}} y(x)\,d\mu(x) \geq 0 \quad \forall\, y(\cdot) \geq 0,$$

which is equivalent to nonnegativity of the measure μ. The Lagrangian function is given by

$$\begin{aligned} L(u,\mu) &= f(u) + \int_{\bar{\Omega}} (G\,u)(x)\,d\mu(x) \\ &= \frac{1}{2}\,\|S\,u - y_\Omega\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)} + \int_{\bar{\Omega}} y(x)\,d\mu(x). \end{aligned}$$

To guarantee the existence of a Lagrange multiplier, we now suppose that the Slater condition is valid: we assume that there is some control $\tilde{u} \in U_{ad}$ such that the associated state $\tilde{y} = G\,\tilde{u}$ satisfies the strict inequality

$$\tilde{y}(x) < 0 \quad \forall\, x \in \bar{\Omega}. \tag{6.31}$$

Since $\tilde{y}$ is continuous on $\bar{\Omega}$, there is some $\delta > 0$ such that $\tilde{y}(x) \leq \delta < 0$ for all $x \in \bar{\Omega}$. In other words, $-\tilde{y} \in \operatorname{int} K$, and the Slater condition formulated in Theorem 6.1 is fulfilled. Hence, for any solution $\bar{u}$ to the given problem there exists an associated Lagrange multiplier $\mu \in M(\bar{\Omega})$ such that $(\bar{u}, \mu)$ is a saddle point of the Lagrangian. By virtue of Theorem 6.2, it follows that

$$D_u L(\bar{u}, \mu)(u - \bar{u}) \geq 0 \quad \forall\, u \in U_{ad},$$

and hence

$$f'(\bar{u})(u-\bar{u}) + \int_{\bar{\Omega}} G'(\bar{u})(u-\bar{u})\, d\mu \geq 0. \tag{6.32}$$

The first term on the left-hand side can be reformulated as in Chapter 4 by means of an adjoint state p_1:

$$f'(\bar{u})(u-\bar{u}) = \int_{\Omega} \big(p_1 + \lambda\,\bar{u}\big)\,(u-\bar{u})\, dx.$$

Here, $p_1 \in H^1(\Omega)$ is the weak solution to the elliptic boundary value problem

$$\boxed{\begin{aligned} -\Delta p_1 + p_1 &= \bar{y} - y_\Omega && \text{in } \Omega \\ \partial_\nu\, p_1 &= 0 && \text{on } \Gamma. \end{aligned}} \tag{6.33}$$

We aim to deal with the second term in a similar way. Since $G'(\bar{u})$ maps $L^r(\Omega)$ continuously into $C(\bar{\Omega})$, the dual operator $G'(\bar{u})^*$ maps $M(\bar{\Omega})$ continuously into $L^r(\Omega)^* = L^{r'}(\Omega)$, where $r' = r/(r-1)$. Hence, there exists some function $p_2 \in L^{r'}(\Omega)$ such that

$$\begin{aligned} \int_{\bar{\Omega}} G'(\bar{u})(u-\bar{u})\, d\mu &= \big\langle \mu\,,\, G'(\bar{u})(u-\bar{u})\big\rangle_{M(\bar{\Omega}),C(\bar{\Omega})} \\ = \big\langle G'(\bar{u})^*\,\mu\,,\, u-\bar{u}\big\rangle_{L^{r'}(\Omega),L^r(\Omega)} &= \int_{\bar{\Omega}} p_2\,(u-\bar{u})\, dx. \end{aligned} \tag{6.34}$$

We now decompose μ into $\mu_\Omega + \mu_\Gamma$, where, as indicated by the subscripts, μ_Ω and μ_Γ have their supports in Ω and Γ, respectively. Owing to the linearity of G, we have $G'(\bar{u})(u-\bar{u}) = G\,(u-\bar{u}) = y - \bar{y}$, where y and $\bar{y}$ denote the states associated with u and $\bar{u}$, respectively. Hence, p_2 satisfies the relation

$$\int_{\bar{\Omega}} (y-\bar{y})\, d\mu = \int_{\Omega} (y-\bar{y})\, d\mu_\Omega + \int_{\Gamma} (y-\bar{y})\, d\mu_\Gamma = \int_{\bar{\Omega}} p_2\,(u-\bar{u})\, dx.$$

A comparison with the variational inequality (2.67) on page 74 shows how the function p_2 has to be defined. Formally, we must have

$$\begin{aligned} -\Delta p_2 + p_2 &= \mu_\Omega && \text{in } \Omega \\ \partial_\nu\, p_2 &= \mu_\Gamma && \text{on } \Gamma. \end{aligned} \tag{6.35}$$

More precisely, p should be a function that satisfies the variational equation

$$\int_{\Omega} (\nabla p \cdot \nabla v + p\, v)\, dx + \int_{\Gamma} p\, v\, ds = \int_{\Omega} v\, d\mu_\Omega + \int_{\Gamma} v\, d\mu_\Gamma$$

for all $v \in H^1(\Omega)$ which are so smooth that all the integrals appearing in the equation are meaningful. In this way, we have arrived at an elliptic boundary value problem with a measure on the right-hand side. Note that the measure μ generates a functional that does not belong to $H^1(\Omega)^*$, in general. Since the elements $v \in H^1(\Omega)$ are not necessarily continuous, not every functional from $C(\bar{\Omega})^*$ can be applied to v. Moreover, the integrability properties of the gradients ∇p and ∇y have to match each other. In conclusion, the definition of a "solution" to (6.35) requires a somewhat different formulation. For the unique solvability of (6.35) in a suitable sense, we refer the reader to Theorem 7.7 on page 366. With $p = p_1 + p_2$, one then arrives at the following result.

Theorem 6.5. *Suppose that the control $\bar{u}$, together with the associated state $\bar{y}$, is optimal for the problem* (6.28)–(6.30)*, and suppose that the Slater condition* (6.31) *is fulfilled. Then there exist a regular Borel measure $\mu \in M(\bar{\Omega})$ of the form $\mu = \mu_\Omega + \mu_\Gamma$, where $\mu_\Omega := \mu|_\Omega$ and $\mu_\Gamma := \mu|_\Gamma$, and an associated adjoint state $p \in W^{1,s}(\Omega)$, with $s \in \big[1, \frac{N}{N-1}\big)$ arbitrary, such that the adjoint problem* (6.36)*, the variational inequality* (6.37)*, and the complementarity condition* (6.38) *are satisfied. We thus have:*

$$\begin{aligned} -\Delta p + p &= \bar{y} - y_\Omega + \mu_\Omega && \text{in } \Omega \\ \partial_\nu p &= \mu_\Gamma && \text{on } \Gamma, \end{aligned} \tag{6.36}$$

$$\int_\Omega (\lambda\, \bar{u} + p)\,(u - \bar{u})\, dx \geq 0 \quad \forall\, u \in U_{ad}, \tag{6.37}$$

$$\mu \geq 0, \qquad \int_{\bar{\Omega}} \bar{y}(x)\, d\mu(x) = 0. \tag{6.38}$$

Proof: The variational inequality with $p = p_1 + p_2$ follows from the relations (6.32)–(6.34). The representation of p as the solution to the adjoint problem is a consequence of the fact that p_2 solves, by Theorem 7.7, the elliptic boundary value problem (6.35). Finally, the complementarity condition results from the relation (6.5) of the Lagrange multiplier rule. □

In the above theorem, only the state constraints have been eliminated by using a Lagrange multiplier, while the box constraints for the control have been accounted for by the variational inequality. By means of the pointwise construction (6.8) on page 329, these constraints can also be included through multipliers μ_a, $\mu_b \in L^s(\Omega)$. To this end, we set

$$\mu_a(x) := (\lambda\, \bar{u}(x) + p(x))_+, \qquad \mu_b(x) := (\lambda\, \bar{u}(x) + p(x))_-.$$

Then, by definition,

$$\mu_a \geq 0, \quad \mu_b \geq 0, \quad \lambda\,\bar{u} + p + \mu_b - \mu_a = 0,$$

and the usual pointwise analysis of the variational inequality (6.37) yields that for almost all $x \in \Omega$ the complementarity conditions

$$(u_a(x) - \bar{u}(x))\,\mu_a(x) = 0, \qquad (\bar{u}(x) - u_b(x))\,\mu_b(x) = 0$$

are valid. In summary, we have derived the following optimality system for the quantities u, y, μ_a, μ_b, and μ:

$$\begin{array}{rcl@{\qquad}rcl}
-\Delta y + y &=& u & -\Delta p + p &=& y - y_\Omega + \mu_\Omega \\
\partial_\nu y &=& 0, & \partial_\nu p &=& \mu_\Gamma, \\
\multicolumn{6}{c}{\lambda\,u + p + \mu_b - \mu_a = 0,} \\
\multicolumn{6}{c}{\mu \geq 0, \quad \displaystyle\int_{\bar\Omega} y(x)\,d\mu(x) = 0,} \\
\multicolumn{6}{l}{\mu_a(x) \geq 0, \quad \big(u_a(x) - u(x)\big)\,\mu_a(x) = 0 \quad \text{for a.e. } x \in \Omega,} \\
\multicolumn{6}{l}{\mu_b(x) \geq 0, \quad \big(u(x) - u_b(x)\big)\,\mu_b(x) = 0 \quad \text{for a.e. } x \in \Omega.}
\end{array}$$

Remark. Also in this case, the formal Lagrange method leads to the correct result: in fact, defining the Lagrangian function as

$$\begin{aligned}
\mathcal{L}(u, y, p, \mu_a, \mu_b, \mu) \;&:=\; J(y, u) + \int_{\bar\Omega} y\,d\mu \\
&- \int_\Omega \big\{ \nabla y \cdot \nabla p - (y - u)\,p - (u_a - u)\,\mu_a - (u - u_b)\,\mu_b \big\}\,dx,
\end{aligned}$$

we find that the equality $D_y\mathcal{L} = 0$ leads to the adjoint system while $D_u\mathcal{L} = 0$, in combination with the usual complementarity conditions, yields the other relations.

Minimization of the state at a point. In the next convex problem, no state constraints will be prescribed. We will see, however, that simple linear optimal control problems with box constraints for the control may already lead to adjoint differential equations involving measures if the cost functional does not have the usual integral form. Hence this problem, which is a boundary control problem for a change, fits into this chapter.

We again consider a bounded Lipschitz domain $\Omega \subset \mathbb{R}^N$, $N \geq 2$, and assume that functions α, u_a, $u_b \in L^\infty(\Gamma)$ are given such that $\alpha \geq 0$ and $u_a \leq u_b$ almost everywhere in Γ. We change the problem of achieving an optimal stationary temperature distribution a little bit by specifying that

the temperature y be minimized at a prescribed point $x_0 \in \Omega$. We thus consider the following problem:

$$\min J(y) := y(x_0), \tag{6.39}$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + y & = & 0 & \text{in } \Omega \\ \partial_\nu y + \alpha\, y & = & u & \text{on } \Gamma \end{array}} \tag{6.40}$$

and

$$u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Gamma. \tag{6.41}$$

At first glance, the problem (6.39)–(6.41), being only linear, seems to be simpler than the linear-quadratic optimal control problems investigated in Chapter 2. It turns out, however, that from a theoretical viewpoint it is just as difficult.

We consider y in the state space $Y = H^1(\Omega) \cap C(\bar\Omega)$, so that the value $y(x_0)$ will be well defined. The existence of at least one optimal control $\bar u$ is easily proved: the existence result of Theorem 4.15 remains valid for the linear optimal control problem under study, although the cost functional is not of integral type as in (4.31) on page 207. However, this property of the cost functional was not used in the proof of the theorem; in fact, we only needed its continuity, which is evident for (6.39). To guarantee that $y \in C(\bar\Omega)$, we consider the set U_{ad} of admissible controls in some space $L^p(\Gamma)$ for $p > N - 1$.

Difficulties first arise in the derivation of necessary optimality conditions, since the cost functional is not of integral type. Application of the formal Lagrange method to determine the adjoint problem fails, as the reader may verify. However, by means of the Dirac measure $\mu = \delta_{x_0}$, the cost functional can be written in the integral form

$$y(x_0) = f(y) = \int_\Omega y(x)\, d\mu(x)$$

Clearly, f is continuous on Y.

Proceeding as in the previous section, we now make use of the linear solution operator $G : u \mapsto y$, which maps $L^p(\Gamma)$ continuously into Y. We obtain

$$f'(\bar u)(u - \bar u) = \int_\Omega (y - \bar y)\, d\mu = \int_\Omega G\,(u - \bar u)\, d\mu \ge 0 \quad \forall\, u \in U_{ad}.$$

This inequality is of the same form as (6.32), except that μ represents the derivative of J, not a multiplier.

The next steps are completely analogous to those taken in the preceding problem: we define $p \in W^{1,s}(\Omega)$ as the solution to the adjoint system

$$\begin{aligned} -\Delta p + p &= \delta_{x_0} \\ \partial_\nu p + \alpha p &= 0. \end{aligned} \tag{6.42}$$

From Theorem 7.7 on page 366, we infer that

$$\int_\Omega (y - \bar{y})\, d\mu = \int_\Gamma p\,(u - \bar{u})\, ds.$$

Consequently, $\bar{u}$ must obey the variational inequality

$$\int_\Gamma p\,(u - \bar{u})\, ds \geq 0 \quad \forall\, u \in U_{ad}. \tag{6.43}$$

The relations (6.42) and (6.43) constitute the necessary optimality conditions.

Best approximation in the maximum norm. We now discuss a convex problem with nonsmooth cost functional which can be transformed into a convex differentiable problem with state constraints. Here, the distance to the desired target function y_Ω is no longer measured in terms of the L^2 norm, but rather with respect to the maximum norm. We thus consider the optimal control problem

$$\min\, J(y,u) := \|y - y_\Omega\|_{C(\bar{\Omega})} + \frac{\lambda}{2}\, \|u\|^2_{L^2(\Gamma)}, \tag{6.44}$$

subject to

$$\boxed{\begin{aligned} -\Delta y + y &= 0 \quad && \text{in } \Omega \\ \partial_\nu y + \alpha y &= u \quad && \text{on } \Gamma \end{aligned}}$$

and

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

Here, $y_\Omega \in C(\bar{\Omega})$ is prescribed, and all the other data are chosen as in (6.39)–(6.41). The constant λ is nonnegative and hence may vanish. Evidently, the cost functional in (6.44) is not differentiable. In addition, it is not even well defined in the space $H^1(\Omega)$, because elements of $H^1(\Omega)$ do not have to be continuous. In order that $y \in C(\bar{\Omega})$ automatically, we assume that $u_a, u_b \in L^\infty(\Gamma)$. Then u is bounded and measurable, and Theorem 4.7 on page 191 ensures the desired continuity of y.

By means of a simple and often used trick, we transform the given problem into a linear-quadratic and thus differentiable optimization problem. To this end, we put

$$\eta = \max_{x \in \bar{\Omega}} \big| y(x) - y_\Omega(x) \big|.$$

Evidently, we have

$$-\eta \le y(x) - y_\Omega(x) \le \eta \quad \forall\, x \in \bar{\Omega}.$$

Therefore, our problem can be rewritten in the equivalent form

$$\min \left\{ \eta + \frac{\lambda}{2} \|u\|^2_{L^2(\Gamma)} \right\}, \tag{6.45}$$

subject to

$$\eta \in \mathbb{R},$$

$$\begin{aligned} -\Delta y + y &= 0 \quad \text{in } \Omega \\ \partial_\nu y + \alpha\, y &= u \quad \text{on } \Gamma, \end{aligned} \tag{6.46}$$

the control constraints

$$u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Gamma, \tag{6.47}$$

and the *state constraints*

$$\begin{aligned} y(x) &\le y_\Omega(x) + \eta \\ -y(x) &\le -y_\Omega(x) + \eta \end{aligned} \qquad \text{for all } x \in \bar{\Omega}. \tag{6.48}$$

We have thus removed the nondifferentiability at the expense of adding pointwise state constraints. But we have already demonstrated for the problem (6.28)–(6.30) how such state constraints can be handled. The mathematically rigorous derivation of the necessary optimality conditions will be a task given to the reader in Exercise 6.3. It should be noted that the Slater condition (6.4) can always be satisfied with sufficiently large $\eta > 0$.

Formal derivation of the optimality conditions. For simplicity, we employ the formal Lagrange method, which leads to the correct result. At first, we eliminate only the state constraints by means of Lagrange multipliers, but not the elliptic boundary value problem or the box constraints for the control. In this way, the state-constrained problem is transformed into a problem that only has control constraints and thus can be formally treated

using the theory of Chapter 2. To this end, we introduce as the Lagrangian function

$$\begin{aligned} L(y,u,\eta,\mu_1,\mu_2) &= \eta + \frac{\lambda}{2}\|u\|^2_{L^2(\Gamma)} + \int_{\bar\Omega}(y - y_\Omega - \eta)\,d\mu_1 \\ &\quad + \int_{\bar\Omega}(-y + y_\Omega - \eta)\,d\mu_2, \end{aligned}$$

with regular Borel measures μ_1 and μ_2. The reason for this choice is given by the following interpretation of our problem: we minimize the cost functional (6.45), subject to the state constraints (6.48) and the convex constraint $(y,u,\eta) \in C$, where

$$C := \{(y,u,\eta) \in H^1(\Omega) \times L^2(\Gamma) \times \mathbb{R} : \ y \text{ and } u \text{ satisfy (6.46) and (6.47)}\}.$$

The associated Lagrange multiplier rule follows from Theorem 6.1 and Theorem 6.2. Since the postulated constraint qualification is satisfied with sufficiently large η, there exist Lagrange multipliers $\mu_1,\ \mu_2 \in M(\bar\Omega)$ satisfying the saddle point condition (6.3). By means of these multipliers, we can eliminate the state constraints: in fact, by (6.3) the triple $(\bar y, \bar u, \bar\eta)$ solves the following optimal control problem without state constraints:

$$\min \tilde J(y,u,\eta) := L(y,u,\eta,\mu_1,\mu_2),$$

subject to

$$\begin{aligned} -\Delta y + y &= 0 \qquad \text{in } \Omega \\ \partial_\nu y + \alpha y &= u \qquad \text{on } \Gamma, \end{aligned}$$

$$\eta \in \mathbb{R}, \qquad u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Gamma.$$

The controls are the variable $\eta \in \mathbb{R}$ and the function $u \in U_{ad} = \{u \in L^2(\Gamma) : u_a(x) \le u(x) \le u_b(x) \text{ for a.e. } x \in \Gamma\}$.

The new problem has only box constraints for the control u. The corresponding theory was treated in Chapter 2. If the necessary optimality conditions for problems with cost functionals of integral type are formally applied, then one obtains the derivative of the cost functional with respect to y on the right-hand side of the adjoint problem for p; more precisely, the part defined in Ω appears in the elliptic equation, while the part defined on Γ occurs in the boundary condition. We therefore have to determine the

derivative $D_y \tilde{J}$ for fixed multipliers μ_1 and μ_2. It follows that

$$\begin{aligned} 0 \;=\; D_y \tilde{J}(\bar{y},\bar{u},\bar{\eta})\, y \;&=\; \int_{\bar{\Omega}} y(x)\,\big(d\mu_1(x) - d\mu_2(x)\big) \\ &=\; \int_{\Omega} y\,(d\mu_1 - d\mu_2)|_{\Omega} + \int_{\Gamma} y\,(d\mu_1 - d\mu_2)|_{\Gamma}. \end{aligned}$$

Therefore, the adjoint problem is given by

$$\begin{aligned} -\Delta p + p &= (\mu_1 - \mu_2)|_{\Omega} \qquad \text{in } \Omega \\ \partial_\nu\, p + \alpha\, p &= (\mu_1 - \mu_2)|_{\Gamma} \qquad \text{on } \Gamma. \end{aligned} \tag{6.49}$$

By virtue of Theorem 7.7 on page 366, it has a unique solution $\bar{p} \in W^{1,s}(\Omega)$, with $s \in \big[1, \frac{N}{N-1}\big)$.

To derive the other necessary conditions, we make use of the Lagrangian function $\mathcal{L}$ associated with the minimization problem for $\tilde{J}$, which is given by

$$\mathcal{L}(y,u,\eta) = \tilde{J}(\bar{y},\bar{u},\bar{\eta}) - \int_{\Omega} \nabla y \cdot \nabla p\, dx - \int_{\Omega} y\, p\, dx - \int_{\Gamma} (\alpha\, y - u)\, p\, ds.$$

We must have $D_u \mathcal{L}(\bar{y},\bar{u},\bar{\eta})(u-\bar{u}) \geq 0$ for all $u \in U_{ad}$, that is,

$$\int_{\Gamma} (\lambda \bar{u} + p)(u - \bar{u})\, dx \geq 0 \quad \forall u \in U_{ad}. \tag{6.50}$$

Moreover, we have

$$D_\eta\, \mathcal{L} = D_\eta\, L = 1 - \int_{\bar{\Omega}} d\mu_1 - \int_{\bar{\Omega}} d\mu_2,$$

and thus the condition $D_\eta \mathcal{L}(\bar{y},\bar{u},\bar{\eta}) = 0$ implies that

$$\int_{\bar{\Omega}} (d\mu_1 + d\mu_2) = 1. \tag{6.51}$$

In summary, $(\bar{y},\bar{u},\bar{\eta})$ has to satisfy the optimality system consisting of the state problem (6.46), the adjoint problem (6.49), the constraints (6.47) and (6.48), the complementary slackness conditions

$$\int_{\bar{\Omega}} (\bar{y} - y_\Omega - \bar{\eta})\, d\mu_1 = \int_{\bar{\Omega}} (-\bar{y} + y_\Omega - \bar{\eta})\, d\mu_2 = 0,$$

the variational inequality (6.50), and equation (6.51).

The above derivation was only formal, since we treated the integral functional $\tilde{J}$ as if it were a continuous linear functional on $L^2(\Omega)$ with respect

to y. However, if we take the same approach as in the problem of minimizing the value of the state at a given point, then the minimization of $\tilde{J}$ without state constraints leads to the same result in a mathematically more rigorous way.

6.2.2. A nonconvex problem. We now give another illustrative example of the use of Theorem 6.3. We consider the optimal control problem for a semilinear elliptic equation:

$$\min\ J(y,u) := \int_\Omega \varphi\big(x,y(x)\big)\,dx + \int_\Omega \psi\big(x,u(x)\big)\,dx, \tag{6.52}$$

subject to

$$\boxed{\begin{array}{rcll} -\Delta y + d(x,y) &=& u & \text{in } \Omega \\ \partial_\nu y &=& 0 & \text{on } \Gamma \end{array}} \tag{6.53}$$

and to the box and state constraints

$$u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Omega, \tag{6.54}$$

$$y(x) \le 0 \qquad \forall x \in \bar{\Omega}. \tag{6.55}$$

We suppose that Assumption 4.14 on page 206 holds. By the nonlinearity of the mapping $G : u \mapsto y$, the above problem is a nonconvex one. This would also be the case if a convex quadratic cost functional were chosen in place of the functional in (6.52). By Theorem 4.17 on page 213, the mapping $G : u \mapsto y$ is continuously differentiable from $L^r(\Omega)$ into $H^1(\Omega) \cap C(\bar{\Omega})$, for $r > N/2$. We therefore fix some $r > N/2$ and put $f(u) = J(G(u), u)$. Then the above problem becomes an optimization problem in a Banach space, namely,

$$\min f(u), \qquad u \in C, \quad G(u) \le_K 0, \tag{6.56}$$

where

$$\begin{aligned} C &= \{u \in L^r(\Omega) :\ u_a(x) \le u(x) \le u_b(x) \quad \text{for a.e. } x \in \Omega\} \\ K &= \{y \in C(\bar{\Omega}) : y(x) \ge 0 \quad \forall\, x \in \bar{\Omega}\}. \end{aligned}$$

This problem is a special case of the problem (6.1), with the specifications $U = L^r(\Omega)$ and $Z = C(\bar{\Omega})$. K has nonempty interior in Z. Therefore, it makes sense to postulate the validity of the linearized Slater condition (6.18) on page 332: we require that there are some $\varepsilon > 0$ and some $\tilde{u} \in C$ such that, with $y = G'(\bar{u})(\tilde{u} - \bar{u})$ and $\bar{y} = G(\bar{u})$,

$$\bar{y}(x) + y(x) \le -\varepsilon \quad \forall\, x \in \bar{\Omega}. \tag{6.57}$$

The function y thus defined is the solution to the linearized problem

$$\begin{aligned} -\Delta y + d_y(x,\bar{y})\, y &= \tilde{u} - \bar{u} \qquad \text{in } \Omega \\ \partial_\nu y &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

Under the condition (6.57), the Zowe–Kurcyusz constraint qualification (6.11) on page 330 is fulfilled. The multiplier z^* associated with the state constraint $y \le 0$ can be identified with some $\mu \in M(\bar{\Omega})$, and the Lagrangian L corresponding to problem (6.56) is given by

$$L(u,\mu) = f(u) + \int_{\bar{\Omega}} G(u)\, d\mu = f(u) + \int_\Omega G(u)\, d\mu_\Omega + \int_\Gamma G(u)\, d\mu_\Gamma,$$

where μ_Ω and μ_Γ are the restrictions of $\mu \in M(\bar{\Omega})$ to Ω and Γ, respectively. From Theorem 6.3, we deduce the following necessary optimality conditions.

Lemma 6.6. *Suppose that $\bar{u}$ is a local solution to problem* (6.56) *that satisfies the constraint qualification* (6.57)*. Then there is a nonnegative regular Borel measure $\mu \in M(\bar{\Omega})$ such that*

$$\begin{aligned} D_u L(\bar{u},\mu)(u-\bar{u}) &\ge 0 \quad \forall\, u \in C \\ \int_{\bar{\Omega}} \bar{y}(x)\, d\mu(x) &= 0. \end{aligned} \tag{6.58}$$

Substituting the derivative $D_u L$ of the above Lagrangian into (6.58) yields

$$f'(\bar{u})(u-\bar{u}) + \int_\Omega G'(\bar{u})(u-\bar{u})\, d\mu_\Omega + \int_\Gamma G'(\bar{u})(u-\bar{u})\, d\mu_\Gamma \ge 0 \tag{6.59}$$

for all $u \in C$. We thus obtain the following result.

Lemma 6.7. *Under the above assumptions, $\bar{u}$ solves the linear optimal control problem*

$$\begin{aligned} \min j(y,u) := &\int_\Omega \varphi_y\big(x,\bar{y}(x)\big)\, y(x)\, dx + \int_\Omega \psi_u\big(x,\bar{u}(x)\big)\, u(x)\, dx \\ &+ \int_\Omega y(x)\, d\mu_\Omega(x) \; + \int_\Gamma y(x)\, d\mu_\Gamma(x), \end{aligned} \tag{6.60}$$

subject to

$$\begin{aligned} -\Delta y + d_y(x,\bar{y})\, y &= u, \qquad u_a(x) \le u(x) \le u_b(x), \\ \partial_\nu y &= 0. \end{aligned}$$

Proof: The assertion follows from the expression for $G'(\bar{u})$, upon rearranging (6.59) for $\bar{u}$. □

By the above lemma, $\bar{u}$ solves a convex problem in which only the box constraints for the control occur. From this point onward, the optimality conditions can be derived as in the preceding section. To this end, we define the adjoint state p as the solution to

$$\boxed{\begin{aligned} -\Delta p + d_y(x,\bar{y})\, p &= \varphi_y(x,\bar{y}) + \mu_\Omega && \text{in } \Omega \\ \partial_\nu p &= \mu_\Gamma && \text{on } \Gamma. \end{aligned}} \tag{6.61}$$

Owing to Theorem 7.7, the above elliptic problem has a unique solution p that belongs to $W^{1,s}(\Omega)$ for every $s < N/(N-1)$. With this p, the variational inequality

$$\int_\Omega \big(p(x) + \psi_u\big(x,\bar{u}(x)\big)\big)\big(u(x) - \bar{u}(x)\big)\, dx \geq 0 \qquad \forall\, u \in C \tag{6.62}$$

is satisfied. The first-order necessary optimality conditions are thus derived in the form of a variational inequality for $\bar{u}$. Alternatively, one can transform the variational inequality (6.62) by means of Lagrange multipliers into a set of equations and formulate all optimality conditions in terms of a Karush–Kuhn–Tucker system.

To this end, we define Lagrange multipliers μ_a and μ_b associated with the box constraints for u by

$$\mu_a(x) = \big(p(x) + \psi_u\big(x,\bar{u}(x)\big)\big)_+, \qquad \mu_b(x) = \big(p(x) + \psi_u\big(x,\bar{u}(x)\big)\big)_-.$$

By virtue of the embedding result of Theorem 7.1 on page 355, these pointwise multipliers define functions belonging to $L^q(\Omega)$ for any $q < N/(N-2)$. As has been explained repeatedly before, one can now transform the variational inequality into an equation plus complementary slackness conditions. The reader will be asked in Exercise 6.4 to produce the corresponding argument. One then arrives at the following result.

Theorem 6.8. *Suppose that $\bar{u}$ is a locally optimal control for the problem* (6.52)–(6.55) *with associated state $\bar{y}$, and suppose that the constraint qualification* (6.57) *is fulfilled. Then there exist μ_a and μ_b belonging to $L^q(\Omega)$ for all $q < N/(N-2)$, some $\mu \in M(\bar{\Omega})$, and an adjoint state p belonging to $W^{1,s}(\Omega)$ for all $s < N/(N-1)$ such that $u = \bar{u}$, $y = \bar{y}$, p, μ_a, μ_b, and μ satisfy the following optimality system:*

$$
\begin{array}{rcl@{\qquad}rcl}
-\Delta y + d(x,y) & = & u & -\Delta p + d_y(x,y)\,p & = & \varphi_y(x,y) + \mu_\Omega \\
\partial_\nu y & = & 0 & \partial_\nu p & = & \mu_\Gamma
\end{array}
$$

$$
p + \psi_u(x,u) + \mu_b - \mu_a = 0
$$

$$
\mu \geq 0, \quad \int_{\bar\Omega} y(x)\, d\mu(x) = 0
$$

$$
\begin{array}{ll}
\mu_a(x) \geq 0, & \big(u_a(x) - u(x)\big)\,\mu_a(x) = 0 \quad \textit{for a.e. } x \in \Omega \\
\mu_b(x) \geq 0, & \big(u(x) - u_b(x)\big)\,\mu_b(x) = 0 \quad \textit{for a.e. } x \in \Omega.
\end{array}
$$

References. State constraints have attracted much interest because of their importance in various applications. We only mention a small selection of the numerous relevant papers.

In the case of elliptic problems, first-order necessary optimality conditions were treated, for instance, in [**AR97**], [**BC91**], [**BC95**], [**Cas86**], and [**Cas93**], while the papers [**CM02a**], [**CT99**], [**CTU00**], [**MT06**], and [**CDlRT08**] deal with second-order conditions. For a comprehensive treatment of the elliptic case, we refer the reader to Neittaanmäki et al. [**NST06**]. In the parabolic case, we refer to [**Cas97**], [**Mac81**], [**Mac82**], [**Mac83a**], [**RZ98**], [**RZ99**], and [**Trö84b**] for first-order optimality conditions, and to [**GT93**], [**RT00**], and [**CDlRT08**] for second-order conditions. The structure of Lagrange multipliers for state constraints was studied in [**BK02a**].

Error estimates for finite element approximations and state constraints were investigated in [**CM02b**] and [**TT96**]. A more detailed exposition of this subject, as well as references to the relevant literature, can be found in [**HPUU09**]. Numerical techniques for the solution of state-constrained elliptic problems are given in, e.g., [**BK02a**], [**BK02b**], [**GR01**], [**MRT06**], [**MT06**], [**MM01**], and [**MM00**]. For parabolic problems, we refer to [**AM84**], [**AM89**], [**LS00**], and [**Trö84a**]. A comprehensive treatment of numerical methods for the optimal control of partial differential equations is contained in [**IK08**] and [**HPUU09**].

Further references for optimal control problems. A comprehensive treatment of the optimal control theory for linear-quadratic elliptic, parabolic, and hyperbolic problems can be found in the standard textbook [**Lio71**]. Various practical applications were presented in [**But69**] and [**But75**]. Nonlinear problems were treated in, e.g., [**Bar93**], [**HPUU09**], [**Lio69**], [**NT94**], [**NST06**], and [**Tib90**]. Numerical methods for the optimal control of flow problems were discussed in [**Gun03**]. In [**NST06**], further results concerning the theory of elliptic problems with state constraints were given. The use of strongly continuous semigroups, in place of weak solutions, for parabolic problems can be found in [**BDPDM92**],

[**BDPDM93**], [**Fat99**], [**LT00a**], and [**LT00b**]. Riccati techniques for the solution of control problems, as well as results concerning controllability and observability, are covered comprehensively in [**LT00a**] and [**LT00b**]. An introduction to the modern theory of controllability and stabilizability is given in [**Cor07**]. In this context, coupled systems of partial differential equations were treated in [**Las02**]. The use of Riccati operators was also addressed in [**BDPDM92**], [**BDPDM93**], and [**Lio71**].

6.3. Exercises

6.1 Determine the dual cone K^+ to the cone K of almost-everywhere nonnegative functions in $L^p(\Omega)$ for $1 \leq p < \infty$.

6.2 Prove that the functions μ_1 and μ_2 defined in relation (6.8) on page 329 are Lagrange multipliers associated with the optimal control $\bar{u}$ of problem (6.7).

6.3 Derive the first-order necessary optimality conditions for the optimal control problem with maximum norm functional defined on page 345.

6.4 Derive the optimality system stated in Theorem 6.8 on page 351 by means of the variational inequality (6.62) on page 351.

6.5 Show that for the elliptic problem (6.25)–(6.26) on page 336, every solution $(\bar{y}, \bar{u})$ satisfies the constraint qualification (6.15) on page 331.

Chapter 7

Supplementary results on partial differential equations

7.1. Embedding results

The usefulness of Sobolev spaces is to a large extent determined by embedding results and trace theorems that will be provided in this chapter. We follow the standard text by Adams [**Ada78**].

Theorem 7.1. *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain. Moreover, let $1 < p < \infty$, and let m be a nonnegative integer. Then the following embeddings exist and are continuous:*

- *for $m\,p < N$:* $W^{m,p}(\Omega) \hookrightarrow L^q(\Omega)$ *if* $1 \le q \le \dfrac{N\,p}{N - m\,p}$
- *for $m\,p = N$:* $W^{m,p}(\Omega) \hookrightarrow L^q(\Omega)$ *if* $1 \le q < \infty$
- *for $m\,p > N$:* $W^{m,p}(\Omega) \hookrightarrow C(\bar{\Omega})$.

In particular, if $\Omega \subset \mathbb{R}^2$, then $H^1(\Omega) = W^{1,2}(\Omega) \hookrightarrow L^q(\Omega)$ for all $1 \le q < \infty$, and if $\Omega \subset \mathbb{R}^3$, then $H^1(\Omega) \hookrightarrow L^6(\Omega)$. The smoothness properties of boundary values are described by the following result.

Theorem 7.2. *Let $m \in \mathbb{N}$ with $m > 0$, and let the boundary Γ be of class $C^{m-1,1}$. Then for $m\,p < N$ the trace operator τ is continuous from $W^{m,p}(\Omega)$ into $L^r(\Gamma)$, provided that $1 \le r \le \frac{(N-1)\,p}{N-m\,p}$. If $m\,p = N$, then τ is continuous for all $1 \le r < \infty$.*

The above theorems follow from Theorem 5.4 and Theorem 5.22 in [**Ada78**]. We also refer to [**Eva98**] and [**Wlo82**]. Collections of results on Sobolev spaces can be found in [**Fur99**] and [**GGZ74**]. For an extension of the above theorems to noninteger m, see [**Ada78**], Theorems 7.57 and 7.53 and Remark 7.56, as well as the comprehensive treatment in [**Tri95**]. By means of fractional-order Sobolev spaces, one can obtain a more precise characterization of the trace mapping. From Theorem 7.53 in [**Ada78**], we have the following result for integers $m \geq 1$:

Theorem 7.3. *Suppose that Ω is a domain of class C^m, and let $1 < p < \infty$. Then the trace operator τ is continuous from $W^{m,p}(\Omega)$ onto $W^{m-1/p,p}(\Gamma)$.*

In particular, the continuity of the mapping $\tau : H^1(\Omega) \to H^{1/2}(\Gamma)$ follows; τ is even surjective. The following result was used in the existence proof for optimal controls.

Theorem 7.4 (Rellich)**.** *Suppose that Ω is a bounded Lipschitz domain, and let $1 \leq p < \infty$ and $m \in \mathbb{N}$, with $m > 0$. Then every bounded set in $W^{m,p}(\Omega)$ is relatively compact in $W^{m-1,p}(\Omega)$.*

The above property is called a *compact embedding*. In particular, bounded subsets of $H^1(\Omega)$ are relatively compact in $L^2(\Omega)$. We remark that the above results remain valid under somewhat weaker requirements on the boundary Γ (regular boundaries, boundaries satisfying a cone condition); see, e.g., [**Ada78**], [**GGZ74**], or [**Gri85**].

7.2. Elliptic equations

In this section, we will first discuss the proof of Lemma 4.6 in the case where the coefficient functions of the differential operator are nonsmooth and the domain is Lipschitz. Then we will prove the results concerning essential boundedness of the solution to the semilinear elliptic boundary value problem (4.5) in Section 4.2. We will also present an existence result for elliptic problems with measures as data.

7.2.1. Elliptic regularity and continuity of solutions. Owing to Lemma 4.6, the weak solution y to the linear elliptic boundary value problem

$$\begin{aligned} \mathcal{A}\, y + y &= f \\ \partial_{\nu_{\mathcal{A}}} y &= g \end{aligned}$$

with given data $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$, where $r > N/2$ and $s > N-1$, is continuous in $\bar{\Omega}$ and thus belongs to $H^1(\Omega) \cap C(\bar{\Omega})$.

This result was proved on page 190 under the simplifying assumption that the coefficient functions a_{ij} and the boundary Γ of Ω are all sufficiently smooth. However, it remains valid for the elliptic differential operator $\mathcal{A}$ introduced on page 37 if the coefficient functions a_{ij} belong to $L^\infty(\Omega)$ and Ω is a bounded Lipschitz domain. This can be verified using the following line of argument that was communicated to me by J. Griepentrog.

Let $c_0 \in L^\infty(\Omega)$ be nonnegative almost everywhere, and let $\|c_0\|_{L^\infty(\Omega)} > 0$. In the following, we put $V := H^1(\Omega)$ and $G := \bar{\Omega}$. We then consider the elliptic operator $L \in \mathcal{L}(V, V^*)$ defined by

$$\langle Ly\,,\, v\rangle_{V^*,V} = \int_\Omega \Big(\sum_{i,j=1}^N a_{ij}(x) D_i y(x) D_j v(x) + c_0(x) y(x) v(x) \Big)\, dx.$$

Then the linear elliptic Neumann boundary value problem $Lu = F$ has for every functional $F \in V^*$ a unique solution $y \in V$.

In **[Gri02]**, Theorem 4.12, it was proved that for any $\omega \in [0, N]$ two families of *Sobolev–Campanato spaces* $W_0^{1,2,\omega}(G) \subset V$ and $Y^{-1,2,\omega}(G) \subset V^*$ with the following property can be found: there exists a constant $\bar{\omega} \in (N-2, N)$ such that the restriction of L to $W_0^{1,2,\omega}(G)$ is a linear isomorphism between $W_0^{1,2,\omega}(G)$ and $Y^{-1,2,\omega}(G)$, for any $\omega \in [0, \bar{\omega})$.

Here (i.e., in the case of homogeneous Neumann data), $W_0^{1,2,\omega}(G)$ coincides with the Sobolev–Campanato space

$$W^{1,2,\omega}(\Omega) = \big\{ u \in V : |\nabla u| \in \mathcal{L}^{2,\omega}(\Omega) \big\}$$

of all functions in V that have weak derivatives in the Campanato space $\mathcal{L}^{2,\omega}(\Omega)$. In this context, the fact that for $\omega \in (N-2, N)$ the space $W_0^{1,2,\omega}(G)$ is continuously embedded in the Hölder space $C^{0,\kappa}(G)$ with $\kappa = (N-\omega)/2$ is of particular importance. We only have to guarantee that under the given assumptions f and g belong to $Y^{-1,2,\omega}(G)$. We need $\omega \in (N-2, N)$ for this property to hold.

By Theorem 3.9 in **[Gri02]**, for any $\omega \in [0, N)$ all those functionals $F \in V^*$ that belong to $Y^{-1,2,\omega}(G)$ can be represented in the form

$$\langle F\,,\, \varphi\rangle_{V^*,V} = \int_\Omega \sum_{i=1}^m f_i(x) D_i\varphi(x)\, dx + \int_\Omega f(x)\varphi(x)\, dx + \int_\Gamma g(x)\varphi(x)\, ds(x),$$

where

$$f_1, \dots, f_n \in \mathcal{L}^{2,\omega}(\Omega), \quad f \in \mathcal{L}^{2N/(N+2),\omega N/(N+2)}(\Omega),$$
$$g \in \mathcal{L}^{2(N-1)/N,\omega(N-1)/N}(\Gamma).$$

The mapping $(f_1, \dots, f_N, f, g) \mapsto F$ defines a continuous linear operator.

To apply these results, we need the connection between Campanato spaces and the usual Lebesgue spaces; see Remark 3.10 in **[Gri02]**:

(i) If $q \geq 2$ and $\omega_q = N(1-2/q)$, then $L^q(\Omega)$ is continuously embedded in $\mathcal{L}^{2,\omega_q}(\Omega)$. In particular, $\omega_q > N-2$ if $q > N$.

(ii) If $r \geq 2N/(N+2)$ and $\omega_r = 2+N(1-2/r)$, then $L^r(\Omega)$ is continuously embedded in $\mathcal{L}^{2N/(N+2),\omega_r N/(N+2)}(\Omega)$. In particular, $\omega_r > N-2$ if $r > N/2$.

(iii) If $s \geq 2(N-1)/N$ and $\omega_s = 1+(N-1)(1-2/s)$, then $L^s(\Gamma)$ is continuously embedded in $\mathcal{L}^{2(N-1)/N,\omega_s(N-1)/N}(\Gamma)$. In particular, $\omega_s > N-2$ if $s > N-1$.

The latter two conditions, that is, $r > N/2$ and $s > N-1$, are just the assumptions of Lemma 4.6.

7.2.2. Stampacchia's method. In this section, we prove the validity of Theorem 4.5 on page 189, that is, the boundedness of the solution to the elliptic problem (4.5) on page 183:

$$\begin{aligned} \mathcal{A}\,y + c_0(x)\,y + d(x,y) &= f \qquad \text{in} \quad \Omega \\ \partial_{\nu_{\mathcal{A}}} y + \alpha(x)\,y + b(x,y) &= g \qquad \text{on} \quad \Gamma. \end{aligned} \tag{7.1}$$

To this end, we employ a method due to Stampacchia. It makes use of the following auxiliary result (cf. Kinderlehrer and Stampacchia **[KS80]**, Lemma B.1).

Lemma 7.5. *Let $k_0 \in \mathbb{R}$, and suppose that φ is a nonnegative and nonincreasing function defined in $[k_0, \infty)$ and having the following property: for every $h > k \geq k_0$,*

$$\varphi(h) \leq \frac{C}{(h-k)^a}\ \varphi(k)^b$$

with constants $C > 0$, $a > 0$, and $b > 1$. Then $\varphi(k_0+\delta) = 0$, where

$$\delta^a = C\,\varphi(k_0)^{b-1}\ 2^{\frac{ab}{b-1}}. \tag{7.2}$$

Proof of Theorem 4.5: We modify a proof given in **[Sta65]** and **[KS80]** for homogeneous Dirichlet boundary conditions to deal with the present case.

(i) <u>*Preliminaries*</u>

Existence and uniqueness of a solution $y \in H^1(\Omega)$ follow from Theorem 4.4 on page 186. It remains to show the boundedness and the validity of the estimate (4.10) on page 189. To this end, we will test the solution y to (4.5)

in the variational formulation with the part of y that is larger than $k > 0$ in absolute value, and then show that this part vanishes for sufficiently large k.

In the statement of the theorem, integrability properties of f and g were postulated. Here, we denote the orders of integrability by $\tilde{r}$ and $\tilde{s}$, respectively. We thus have $f \in L^{\tilde{r}}(\Omega)$ and $g \in L^{\tilde{s}}(\Gamma)$, where $\tilde{r} > N/2$ and $\tilde{s} > N-1$.

We first assume $N \geq 3$ and explain at the end of the proof which modifications have to made for the case of $N = 2$. We fix some $\lambda \in \left(1, \frac{N-1}{N-2}\right)$ sufficiently close to unity such that

$$\tilde{r} > r := \frac{N}{N - \lambda(N-2)}, \qquad \tilde{s} > s := \frac{N-1}{N-1-\lambda(N-2)}.$$

Since $N \geq 3$, and owing to the choice of λ, we obviously have $r > 1$ and $s > 1$. If we succeed in proving the result for r and s, then it will be valid for all $\tilde{r} > r$ and $\tilde{s} > s$. The conjugate exponents r' and s' for r and s are given by

$$\frac{1}{r'} = 1 - \frac{1}{r} = \lambda \frac{N-2}{N}, \qquad \frac{1}{s'} = 1 - \frac{1}{s} = \lambda \frac{N-2}{N-1}. \tag{7.3}$$

Below, we will use the embedding estimates

(7.4)

$$\|v\|_{L^p(\Omega)} \leq c \|v\|_{H^1(\Omega)} \quad \text{for} \quad \frac{1}{p} = \frac{1}{2} - \frac{1}{N} = \frac{N-2}{2N} = \frac{1}{2\lambda r'},$$

$$\|v\|_{L^q(\Gamma)} \leq c \|v\|_{H^1(\Omega)} \quad \text{for} \quad \frac{1}{q} = \frac{1}{2} - \frac{1}{2(N-1)} = \frac{N-2}{2(N-1)} = \frac{1}{2\lambda s'}.$$

Since $2r' \leq p$ and $2s' \leq q$, this implies that

$$\|v\|_{L^{2r'}(\Omega)} \leq c \|v\|_{H^1(\Omega)}, \quad \|v\|_{L^{2s'}(\Gamma)} \leq c \|v\|_{H^1(\Omega)}.$$

Next, we define for each $k > 0$ a function $v_k \in H^1(\Omega)$, such that

$$v_k(x) = \begin{cases} y(x) - k & \text{if} \quad y(x) \geq k \\ 0 & \text{if} \quad |y(x)| < k \\ y(x) + k & \text{if} \quad y(x) \leq -k. \end{cases}$$

We aim to show that v_k vanishes almost everywhere for sufficiently large k, which then implies the boundedness of y. For the sake of brevity, we suppress the subscript k, writing v_k simply as v. We introduce the sets

$$\Omega(k) = \big\{x \in \Omega : |y(x)| \geq k\big\}, \quad \Gamma(k) = \big\{x \in \Gamma : |(\tau\, y)(x)| \geq k\big\},$$

where $\tau\, y$ denotes the trace of y on Γ.

(ii) *Consequences of the monotonicity assumptions*

We now derive the inequality (7.6) below. First, we claim that

$$\int_\Omega d(x,y)\, v\, dx \geq 0, \quad \int_\Gamma b(x,y)\, v\, ds \geq 0. \tag{7.5}$$

To see this, let $\Omega_+(k) := \{x : y(x) > k\}$. Using the monotonicity of d with respect to y and the fact that $d(x,0) = 0$, we find that

$$\int_{\Omega_+(k)} d(x,y)\, v\, dx = \int_{\Omega_+(k)} d(x,y)\ (y-k)\, dx$$

$$= \int_{\Omega_+(k)} d(x, y-k+k)\, (y-k)\, dx \geq \int_{\Omega_+(k)} d(x,y-k)\, (y-k)\, dx \geq 0.$$

By the same token,

$$\int_{\Omega_-(k)} d(x,y)\, v\, dx \geq 0,$$

where $\Omega_-(k) := \{x : y(x) < -k\}$. This proves the claim for the first integral in (7.5). The claim for the integral over Γ follows by analogous reasoning.

From the variational formulation for y, we infer that, with the bilinear form $a[y,v]$ defined in (4.7) on page 186,

$$a[y,v] + \int_\Omega d(x,y)\, v\, dx + \int_\Gamma b(x,y)\, v\, ds = \int_\Omega f\, v\, dx + \int_\Gamma g\, v\, ds,$$

whence, in view of (7.5),

$$a[y,v] \leq \int_\Omega f\, v\, dx + \int_\Gamma g\, v\, ds. \tag{7.6}$$

At this point, we can already see that the nonlinearities d and b will not play a role in the estimation.

(iii) *Estimation of* $\|v\|_{H^1(\Omega)}$

We claim that

$$a[v,v] \leq a[y,v]. \tag{7.7}$$

Indeed, we obviously have $D_i y = D_i v$ in $\Omega(k)$ and $v = 0$ in $\Omega \setminus \Omega(k)$, whence it follows that

$$\int_\Omega \sum_{i,j=1}^N a_{ij}(x) D_i y \, D_j v \, dx = \int_\Omega \sum_{i,j=1}^N a_{ij}(x) D_i v \, D_j v \, dx.$$

Moreover, since $y - k > 0$ in $\Omega_+(k)$, $y + k < 0$ in $\Omega_-(k)$, and $v = 0$ in $\Omega \setminus \Omega(k)$,

$$\begin{aligned}
\int_\Omega c_0 \, y \, v \, dx \; &= \int_{\Omega_+(k)} c_0 \, y \, (y-k) \, dx + \int_{\Omega_-(k)} c_0 \, y \, (y+k) \, dx \\
&= \int_{\Omega_+(k)} c_0 \left[(y-k)^2 + (y-k)k \right] dx + \int_{\Omega_-(k)} c_0 \left[(y+k)^2 - (y+k)k \right] dx \\
&\geq \int_\Omega c_0 \, v^2 \, dx.
\end{aligned}$$

The integral $\int_\Gamma \alpha \, y \, v \, ds$ can be treated similarly. From (7.6) and (7.7) and the coercivity properties of the elliptic boundary value problem, we conclude that, with some $\beta > 0$,

$$\beta \, \|v\|^2_{H^1(\Omega)} \leq \int_\Omega f \, v \, dx + \int_\Gamma g \, v \, ds. \tag{7.8}$$

(iv) *Estimation of both sides of* (7.8)

Now recall the first embedding inequality in (7.4) and Young's inequality $a \, b \leq \varepsilon \, a^2 + \frac{1}{4\varepsilon} \, b^2$ for all $a, b \in \mathbb{R}$ and $\varepsilon > 0$. With some generic constant $c > 0$, and using Hölder's inequality, we can estimate the first term of the right-hand side as follows:

$$\begin{aligned}
\Big| \int_\Omega f \, v \, dx \Big| \; &\leq \; \|f\|_{L^r(\Omega)} \, \|v\|_{L^{r'}(\Omega)} \\
&\leq \|f\|_{L^r(\Omega)} \left[\Big(\int_{\Omega(k)} |v|^{2r'} \, dx \Big)^{\frac{1}{2}} \Big(\int_{\Omega(k)} 1 \, dx \Big)^{\frac{1}{2}} \right]^{\frac{1}{r'}} \\
&\leq \|f\|_{L^r(\Omega)} \, \|v\|_{L^{2r'}(\Omega)} \, |\Omega(k)|^{\frac{1}{2r'}} \; \leq \; c \, \|f\|_{L^r(\Omega)} \, \|v\|_{H^1(\Omega)} \, |\Omega(k)|^{\frac{1}{2r'}} \\
&\leq c \, \|f\|^2_{L^r(\Omega)} \, |\Omega(k)|^{\frac{1}{r'}} + \varepsilon \, \|v\|^2_{H^1(\Omega)} \; = \; c \, \|f\|^2_{L^r(\Omega)} \, |\Omega(k)|^{\lambda \frac{2}{p}} + \varepsilon \, \|v\|^2_{H^1(\Omega)}.
\end{aligned}$$

The number $\varepsilon > 0$ is yet to be determined. By similar reasoning, for the boundary integral we find that

$$\begin{aligned}\Big|\int_\Gamma g\,v\,ds\Big| &\le \|g\|_{L^s(\Gamma)}\,\|v\|_{L^{s'}(\Gamma)} \le \|g\|_{L^s(\Gamma)}\,\|v\|_{L^{2s'}(\Gamma)}\,|\Gamma(k)|^{\frac{1}{2s'}}\\ &\le c\,\|g\|^2_{L^s(\Gamma)}\,|\Gamma(k)|^{\lambda\frac{2}{q}} + \varepsilon\,\|v\|^2_{H^1(\Omega)}.\end{aligned}$$

Now, by choosing $\varepsilon := \beta/4$, we may absorb into the left-hand side of (7.8) the terms $\varepsilon\,\|v\|^2_{H^1(\Omega)}$ occurring in the last two inequalities.

Next,we split the expression $\|v\|^2_{H^1(\Omega)}$ on the left-hand side of (7.8) into two equal parts. Invoking (7.8) and the two estimates given in (7.4), we find that

$$\Big(\int_{\Omega(k)} |v|^p\,dx\Big)^{\frac{2}{p}} + \Big(\int_{\Gamma(k)} |v|^q\,ds\Big)^{\frac{2}{q}} \le c\,\|v\|^2_{H^1(\Omega)},$$

whence, by the definition of v,

$$\Big(\int_{\Omega(k)} (|y|-k)^p\,dx\Big)^{\frac{2}{p}} + \Big(\int_{\Gamma(k)} (|y|-k)^q\,dx\Big)^{\frac{2}{q}} \le c\,\|v\|^2_{H^1(\Omega)}. \tag{7.9}$$

(v) *Application of Lemma 7.5*

Suppose that $h > k$. Then $\Omega(h) \subset \Omega(k)$ and $\Gamma(h) \subset \Gamma(k)$, and thus $|\Omega(h)| \le |\Omega(k)|$ and $|\Gamma(h)| \le |\Gamma(k)|$. Therefore,

$$\begin{aligned}\Big(\int_{\Omega(k)} (|y|-k)^p\,dx\Big)^{\frac{2}{p}} &\ge \Big(\int_{\Omega(h)} (|y|-k)^p\,dx\Big)^{\frac{2}{p}} \ge \Big(\int_{\Omega(h)} (h-k)^p\,dx\Big)^{\frac{2}{p}}\\ &= (h-k)^2\,|\Omega(h)|^{2/p}.\end{aligned}$$

The boundary integral is estimated similarly. Finally, we infer from (7.9) and (7.8) that

$$\begin{aligned}&(h-k)^2\left(|\Omega(h)|^{\frac{2}{p}} + |\Gamma(h)|^{\frac{2}{q}}\right)\\ &\le c\left(\|f\|^2_{L^r(\Omega)} + \|g\|^2_{L^s(\Gamma)}\right)\left(|\Omega(k)|^{\lambda\frac{2}{p}} + |\Gamma(k)|^{\lambda\frac{2}{q}}\right)\\ &\le c\left(\|f\|^2_{L^r(\Omega)} + \|g\|^2_{L^s(\Gamma)}\right)\left(|\Omega(k)|^{\frac{2}{p}} + |\Gamma(k)|^{\frac{2}{q}}\right)^{\lambda}.\end{aligned}$$

Here, we have used the fact that for all $a \ge 0$, $b \ge 0$, and $\lambda \ge 1$ the inequality $a^\lambda + b^\lambda \le (a+b)^\lambda$ is valid. Putting $\varphi(h) := |\Omega(h)|^{\frac{2}{p}} + |\Gamma(h)|^{\frac{2}{q}}$, we therefore

obtain the inequality

$$(h-k)^2\,\varphi(h) \le c\left(\|f\|^2_{L^r(\Omega)} + \|g\|^2_{L^s(\Gamma)}\right)\varphi(k)^\lambda,$$

for all $h > k \ge 0$. We now apply Lemma 7.5 with the specifications

$$a = 2,\ b = \lambda > 1,\ k_0 = 0,\ C = c\,(\|f\|^2_{L^r(\Omega)} + \|g\|^2_{L^s(\Gamma)}).$$

We obtain $\delta^2 = \tilde{c}\,(\|f\|^2_{L^r(\Omega)} + \|g\|^2_{L^s(\Gamma)})$, whence the assertion follows: in fact, $\varphi(\delta) = 0$ means that $|y(x)| \le \delta$ for almost every $x \in \Omega$ and $|(\tau\, y)(x)| \le d$ for almost every $x \in \Gamma$.

(vi) *Modification for the $N = 2$ case*

Let $r > N/2 = 1$ and $s > N - 1 = 1$ be the orders of integrability of f and g assumed in the theorem. In the case of $N = 2$, the embedding inequalities (7.4) are valid for all $p < \infty$ and $q < \infty$. We therefore define p and q with $\lambda > 1$ by

$$\frac{1}{p} = \frac{1}{2\lambda r'}, \qquad \frac{1}{q} = \frac{1}{2\lambda s'}.$$

With this specification, all conclusions subsequent to (7.4) in the $N \ge 3$ case carry over to $N = 2$, yielding the validity of the assertion. This concludes the proof. □

Remark. In the above proof, the boundedness of y on Γ has been shown directly, too. Since $\|y\|_{L^\infty(\Gamma)} \le \|y\|_{L^\infty(\Omega)}$, it would already follow from the boundedness of y; see Exercise 4.1.

The above method does not apply directly to the equation $-\Delta y + y^k = f$, with $k \in \mathbb{N}$ odd, subject to a Neumann boundary condition. We now discuss an extension of Stampacchia's method due to E. Casas. This technique applies to boundary value problems of the form considered in (4.15) on page 193:

$$\begin{aligned} \mathcal{A}y + d(x,y) &= 0 \quad \text{in} \quad \Omega \\ \partial_{\nu_\mathcal{A}} y + b(x,y) &= 0 \quad \text{on} \quad \Gamma. \end{aligned}$$

Theorem 7.6. *Suppose that Assumption 4.9 on page 193 is satisfied. For each $n \in \mathbb{N}$ let y_n denote the unique weak solution to the elliptic boundary value problem*

$$\begin{aligned} \mathcal{A}y + n^{-1}y + d(x,y) &= 0 \quad \textit{in} \quad \Omega \\ \partial_{\nu_\mathcal{A}} y + b(x,y) &= 0 \quad \textit{on} \quad \Gamma. \end{aligned}$$

Then there is some $K > 0$ *such that* $\|y_n\|_{L^\infty(\Omega)} \le K$ *for every* $n \in \mathbb{N}$.

Proof: We shall only explain how the preceding proof has to be modified. Since we have zero right-hand sides f and g, we cannot additionally assume that $d(x,0) = 0$ and $b(x,0) = 0$. In order to have the same situation as in the preceding proof, we cut off d and b at $k \in \mathbb{N}$ and at $-k$, respectively, to obtain the functions d_k and b_k defined on page 192. Putting

$$f(x) := -d_k(x,0), \quad g(x) := -b_k(x,0), \quad \tilde{d}(x,y) := d_k(x,y) - d_k(x,0),$$
$$\tilde{b}(x,y) := b_k(x,y) - b_k(x,0),$$

we then have, as in the preceding proof, $\tilde{d}(x,0) = 0$ and $\tilde{b}(x,0) = 0$, as well as $f \in L^r(\Omega)$ and $g \in L^s(\Gamma)$. Note that all of these functions do not depend on $n \in \mathbb{N}$.

We now assume that in Assumption 4.9 the inequality (i) is satisfied by d. If (ii) is valid instead, we can argue analogously using b.

Now let $y := y_n$, where $n \in \mathbb{N}$ is arbitrary but fixed. The following easily verified relations are essential for our method to work: as in the preceding proof, the monotonicity of d yields that on $\Omega \setminus E_d$, we have

$$\tilde{d}(x,y(x))v(x) = \begin{cases} (\tilde{d}(x,y(x)) - \tilde{d}(x,0))(y(x)-k) \ge 0, & x \in \Omega_+(k) \\ 0, & x \in \Omega \setminus \Omega(k) \\ (\tilde{d}(x,y(x)) - \tilde{d}(x,0))(y(x)+k) \ge 0, & x \in \Omega_-(k). \end{cases}$$

Hence, $\tilde{d}(x,y(x))v(x) \ge 0$ for all $x \in \Omega \setminus E_d$. Moreover, for all $x \in E_d$,

$$\tilde{d}(x,y(x))v(x) \begin{cases} \ge \lambda_d\,|y(x)|v(x) = \lambda_d\,|v(x)|^2, & |y(x)| \ge k \\ = \tilde{d}(x,y(x)) \cdot 0 = \lambda_d\,|v(x)|^2, & |y(x)| < k. \end{cases} \tag{7.10}$$

We may therefore modify the arguments employed between (7.5) and (7.6) as follows:

$$\begin{aligned} a[y,v] &\ge \gamma_0 \int_\Omega |\nabla v|^2\,dx + n^{-1}\int_\Omega |v|^2\,dx + \int_\Omega \tilde{d}(x,y)\,v\,dx + \int_\Gamma \tilde{b}(x,y)\,v\,ds \\ &\ge \gamma_0 \int_\Omega |\nabla v|^2\,dx + \int_\Omega \tilde{d}(x,y)\,v\,dx \\ &\ge \gamma_0 \int_\Omega |\nabla v|^2\,dx + \int_{E_d} \lambda_d\,|v|^2\,dx. \end{aligned}$$

By the generalized Poincaré inequality (2.15) on page 35, the latter expression is the square of a norm that is equivalent to the standard norm of $H^1(\Omega)$.

In this way, we eventually arrive at an inequality resembling (7.8) in the preceding proof, whence the subsequent arguments carry over unchanged. We thus obtain a bound $\delta > 0$ for $\|y\|_{L^\infty(\Omega)} = \|y_n\|_{L^\infty(\Omega)}$ that does not depend on n. The assertion thus follows with $K := \delta$. □

7.2.3. Elliptic problems with measures. Since in the case of pointwise state constraints measures occur on the right-hand sides of the associated adjoint problems, a corresponding extension of the theory of elliptic and parabolic problems is needed. The foundations of such a theory are due to Casas [**Cas86, Cas93**] and Alibert and Raymond [**AR97**] for elliptic problems, and to Casas [**Cas97**] and Raymond and Zidani [**RZ99**] for the parabolic case. In the following, we cite such a result for the boundary value problem

$$\boxed{\begin{array}{rcll} \mathcal{A}\,p + c_0\,p & = & \mu_\Omega & \text{in } \Omega \\ \partial_{\nu_\mathcal{A}} p + \alpha\,p & = & \mu_\Gamma & \text{on } \Gamma. \end{array}} \tag{7.11}$$

Here, $\Omega \subset \mathbb{R}^N$ is a bounded Lipschitz domain. The Borel measures μ_Ω and μ_Γ, which are supported in Ω and on Γ, respectively, are the corresponding restrictions of a regular Borel measure $\mu \in M(\bar\Omega)$; that is, we have $\mu = \mu_\Omega + \mu_\Gamma$. The differential operator $\mathcal{A}$ is the elliptic operator introduced in (2.19) on page 37; we assume that the coefficient functions $a_{ij} \in L^\infty(\Omega)$ obey the ellipticity condition (2.20) and (only for simplicity) the symmetry condition $a_{ij}(x) = a_{ji}(x)$. Moreover, we assume that we are given functions $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^\infty(\Gamma)$ such that $c_0 \geq 0$, $\alpha \geq 0$, and $\|\alpha\|_{L^\infty(\Gamma)} + \|c_0\|_{L^\infty(\Omega)} > 0$.

With the above elliptic problem, we associate the bilinear form

$$\begin{aligned} a[p,v] \;&=\; \int_\Omega \Big(\sum_{i,j=1}^N a_{ij}(x)\, D_i p(x)\, D_j v(x) + c_0(x)\, p(x)\, v(x) \Big)\, dx \\ &\quad + \int_\Gamma \alpha(x)\, p(x)\, v(x)\, ds(x). \end{aligned}$$

Moreover, we define for $r > N/2$ and $s > N-1$ the linear space

$$V^{r,s} = \big\{ v \in H^1(\Omega) :\, \mathcal{A}\,v \in L^r(\Omega),\, \partial_{\nu_\mathcal{A}}\, v \in L^s(\Gamma) \big\},$$

where $\mathcal{A}\,v$ is to be understood in the sense of distributions and $\partial_{\nu_\mathcal{A}}\, v$ is defined as in [**Cas93**].

Definition. *A function $p \in W^{1,\sigma}(\Omega)$, where $\sigma \geq 1$, is called a* weak solution *to the problem* (7.11) *if it satisfies the variational equality*

$$a[p,v] = \int_\Omega v(x)\, d\mu_\Omega(x) + \int_\Gamma v(x)\, d\mu_\Gamma(x) \quad \forall\, v \in C^1(\bar{\Omega}). \tag{7.12}$$

The following theorem is a special case of a result proved in **[Cas93]**.

Theorem 7.7. *Under the above assumptions, the boundary value problem* (7.11) *has a unique weak solution p such that $p \in W^{1,\sigma}(\Omega)$ for all $1 \leq \sigma < N/(N-1)$, with the property that for all $v \in V^{r,s}$ the integration by parts formula*

$$\int_\Omega p\,(\mathcal{A}\, v + c_0\, v)\, dx\, dt + \int_\Sigma p\,(\partial_{\nu_\mathcal{A}} v + \alpha\, v)\, ds\, dt = \int_\Omega v\, d\mu_\Omega + \int_\Gamma v\, d\mu_\Gamma$$

is valid. Moreover, there exists a constant $c_\sigma > 0$, which does not depend on μ, such that $\|p\|_{W^{1,\sigma}(\Omega)} \leq c_\sigma \|\mu\|_{M(\bar{\Omega})}$.

Remark. Weak solutions to (7.11) do not have to be unique, as the detailed discussion in **[AR97]** shows. Uniqueness is only ensured if the integration by parts formula (Green's formula) is valid. A definition of $\partial_{\nu_\mathcal{A}}\, p$ is given in **[Cas93]**; see also **[AR97]**. For a definition of the norm $\|\mu\|_{M(\bar{\Omega})}$, we refer the reader to **[Alt99]**.

7.3. Parabolic problems

7.3.1. Solutions in $W(0,T)$.

The linear problem. Following the monograph by Ladyzhenskaya et al. **[LSU68]**, we introduce the following spaces:

Definition. *We denote by $V_2(Q)$ the space $W_2^{1,0}(Q) \cap L^\infty\big(0,T;L^2(\Omega)\big)$, endowed with the norm*

$$\|y\|_{V_2(Q)} = \operatorname*{ess\,sup}_{t\in[0,T]} \|y(t)\|_{L^2(\Omega)} + \left(\iint_Q |\nabla_x y(x,t)|^2\, dx\, dt \right)^{1/2},$$

and by $V_2^{1,0}(Q)$ the space $W_2^{1,0}(Q) \cap C\big([0,T],L^2(\Omega)\big)$, endowed with the norm

$$\|y\|_{V_2^{1,0}(Q)} = \max_{t\in[0,T]} \|y(t)\|_{L^2(\Omega)} + \left(\iint_Q |\nabla_x y(x,t)|^2\, dx\, dt \right)^{1/2}.$$

Observe that in the case of intersections like $W_2^{1,0}(Q) \cap L^\infty\big(0,T;L^2(\Omega)\big)$, the elements of $W_2^{1,0}(Q)$ have to be identified with vector-valued functions belonging to $L^2\big(0,T;H^1(\Omega)\big)$; otherwise, one would be trying to compare real-valued functions with vector-valued ones.

For the sake of better readability, we once more write down the initial-boundary value problem (3.23) under investigation:

$$
\boxed{\begin{array}{rcll} y_t + \mathcal{A}\,y + c_0\,y &=& f & \text{in } Q = \Omega \times (0,T) \\ \partial_{\nu_{\mathcal{A}}} y + \alpha\,y &=& g & \text{on } \Sigma = \Gamma \times (0,T) \\ y(\cdot,0) &=& y_0 & \text{in } \Omega. \end{array}} \tag{7.13}
$$

Here, the uniformly elliptic differential operator $\mathcal{A}$ is defined as in (2.19) on page 37.

Theorem 7.8. *Suppose that $\Omega \subset \mathbb{R}^N$ is a bounded Lipschitz domain, and let functions $c_0 \in L^\infty(Q)$, $\alpha \in L^\infty(\Sigma)$ with $\alpha \geq 0$ almost everywhere on Σ, $y_0 \in L^2(\Omega)$, $f \in L^2(Q)$, and $g \in L^2(\Sigma)$ be given. Moreover, let the differential operator $\mathcal{A}$ have coefficients $a_{ij} \in L^\infty(\Omega)$ such that $a_{ij} = a_{ji}$ and such that the uniform ellipticity condition* (2.20) *on page 37 is fulfilled. Then the initial-boundary value problem* (7.13) *has a unique solution in $V_2(Q)$.*

Proof: The assertion follows from Theorem 5.1 in Chapter III of [**LSU68**]. Its proof indicates the changes that have to be made in comparison with the proof of Theorem 4.1 in Chapter III for the case of homogeneous Dirichlet boundary conditions. We follow the proof in [**LSU68**] and sketch the modifications for boundary conditions of the third kind that are needed to understand the proof of Lemma 5.3 concerning our semilinear problem.

Without loss of generality, we may assume that $c_0(x,t) \geq 0$ for almost every $(x,t) \in Q$. Indeed, if this property fails to hold, we simply substitute $y(x,t) = e^{\lambda\,t}\,\tilde{y}(x,t)$. Then in the resulting differential equation for $\tilde{y}$ the term $(\lambda + c_0)\,\tilde{y}$ occurs in place of $c_0\,y$, and this is nonnegative for sufficiently large $\lambda > 0$.

(i) *Galerkin approximation*

We set $V = H^1(\Omega)$ and $H = L^2(\Omega)$. Since V is a separable Hilbert space, we may choose a countable dense set of linearly independent elements $\{v_i\}_{i=1}^\infty$ in V. After possibly performing an orthogonalization process with respect to the scalar product of H, we may assume that $\{v_i\}_{i=1}^\infty$ forms an orthonormal system in H which is also complete in H.

For arbitrary but fixed $n \in \mathbb{N}$, we now determine approximations $y_n = y_n(x,t)$ through the ansatz

$$y_n(x,t) = \sum_{i=1}^{n} u_i^n(t)\, v_i(x), \tag{7.14}$$

with unknown functions $u_i^n : [0,T] \to \mathbb{R}$, $i = 1, \dots, n$. In the following, we write $(\cdot\,,\,\cdot)$ and $\|\cdot\|$ for the scalar product and the norm, respectively, in H, and $(\cdot\,,\,\cdot)_\Gamma$ for the scalar product in $L^2(\Gamma)$. We also define

$$\begin{aligned} a[t;y,v] \;:=\; & \int_\Omega \Big\{ \sum_{i,j=1}^{N} a_{ij}(x)\, D_i\, y(x)\, D_j\, v(x) + c_0(x,t)\, y(x)\, v(x) \Big\}\, dx \\ & + \int_\Gamma \alpha(x,t)\, y(x)\, v(x)\, ds(x). \end{aligned}$$

We interpret $y_n = y_n(\cdot,t)$ as a function with values in $H^1(\Omega)$. Multiplying the parabolic equation by v_j and integrating over Ω by parts, we find that

$$\Big(\frac{d}{dt}\, y_n(t)\,,\, v_j \Big) + a\big[t; y_n(t), v_j\big] = \big(f(t)\,,\, v_j\big) + \big(g(t)\,,\, v_j\big)_\Gamma \tag{7.15}$$

for almost every $t \in (0,T)$. The initial condition for y is equivalent to $\big(y(\cdot,0)\,,\, v\big) = \big(y_0\,,\, v\big)$ for all $v \in V$. Therefore, we postulate that

$$\big(y_n(\cdot,0)\,,\, v_j\big) = \big(y_0\,,\, v_j\big), \qquad 1 \le j \le n. \tag{7.16}$$

Substituting the ansatz (7.14) in (7.15) and using the orthonormality, we see that

$$\begin{aligned} \frac{d}{dt}\, u_j^n(t) + \sum_{i=1}^{n} u_i^n(t)\, a[t; v_i, v_j] \;&=\; b_j(t) \quad \text{for a.e. } t \in (0,T), \\ u_j^n(0) \;&=\; (y_0\,,\, v_j), \end{aligned} \tag{7.17}$$

for $j = 1, \dots, n$, with the given functions $b_j(t) = (f(t)\,,\, v_j) + (g(t)\,,\, v_j)_\Gamma$. Owing to Carathéodory's theorem, this initial value problem for a system of n linear ordinary differential equations on $[0,T]$ for the unknown vector function $u^n = (u_1^n, \dots, u_n^n)^\top$ has a unique absolutely continuous solution $u^n \in (H^1(0,T))^n$. Multiplying (7.15) by $u_j(t)$ and adding the resulting equations from $j = 1$ to $j = n$, we obtain, for almost every $t \in (0,T)$,

$$\Big(\frac{d}{dt}\, y_n(t)\,,\, y_n(t) \Big) + a\big[t; y_n(t), y_n(t)\big] = \big(f(t)\,,\, y_n(t)\big) + \big(g(t)\,,\, y_n(t)\big)_\Gamma. \tag{7.18}$$

Evidently, since $u^n \in (H^1(0,T))^n$, the function $y_n : [0,T] \to L^2(\Omega)$ is almost everywhere differentiable with respect to t in $(0,T)$.

(ii) *Estimates for* $\{y_n\}$

For any arbitrary but fixed $\tau \in (0,T]$, we have the identity

$$\int_0^\tau \left(\frac{d}{dt}\, y_n(t)\,,\, y_n(t) \right) dt = \frac{1}{2} \int_0^\tau \frac{d}{dt}\, \|y_n(t)\|^2\, dt = \frac{1}{2}\, \|y_n(\tau)\|^2 - \frac{1}{2}\, \|y_n(0)\|^2.$$

Integration of (7.18) over $[0,\tau]$ therefore yields that

$$\begin{aligned} (7.19) \qquad & \frac{1}{2}\, \|y_n(\tau)\|^2 + \int_0^\tau a[t; y_n(t), y_n(t)]\, dt \\ & = \frac{1}{2}\, \|y_n(0)\|^2 + \int_0^\tau \Big\{ \big(f(t)\,,\, y_n(t)\big) + \big(g(t)\,,\, y_n(t)\big)_\Gamma \Big\}\, dt. \end{aligned}$$

By virtue of Bessel's inequality, we have

$$(7.20) \qquad \|y_n(0)\|^2 = \sum_{j=1}^n |u_j^n(0)|^2 = \sum_{j=1}^n |(y_0\,,\, v_j)|^2 \le \|y_0\|^2.$$

Moreover, because $c_0(x,t) \ge 0$ and $\alpha(x,t) \ge 0$,

$$(7.21) \qquad a[t; v\,,\, v] \ge \gamma_0 \,\big\|\, |\nabla v|\, \big\|^2 \qquad \forall\, v \in V,$$

where γ_0 is as in (2.20) on page 37. Using some standard estimates (which may be omitted here) and invoking Gronwall's lemma, we can infer from (7.19) and (7.20) that

$$(7.22) \qquad \max_{t\in[0,T]} \|y_n(t)\| \le c\, \big(\|y_0\| + \|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)} \big).$$

Recalling (7.21), and inserting this estimate for y_n in $C([0,T],H)$ into (7.19) with $\tau = T$, we see that there is a constant $K > 0$ such that

$$(7.23) \qquad \|y_n\|_{C([0,T],H)} + \|y_n\|_{W_2^{1,0}(Q)} \le K \quad \forall\, n \in \mathbb{N}.$$

In particular, $\|y_n(t)\|_H^2 \le K^2$, and in view of the orthonormality,

$$(7.24) \qquad \sum_{i=1}^n |u_i^n(t)|^2 \le K^2 \quad \forall\, t \in [0,T],\ \forall\, n \in \mathbb{N}.$$

(iii) *Convergence properties of the sequences* $\{u_j^n\}$ *and* $\{y_n\}$

Owing to (7.24), we have $|u_j^n(t)| \le K$ for all t, j, and n. Moreover, it follows from (7.15) by integration over time that for any fixed $j \in \mathbb{N}$ the sequence $\{u_j^n\}_{n=1}^\infty$ forms an equicontinuous set in $C[0,T]$. Hence, the Arzelà–Ascoli theorem may be applied to any of these sequences of functions. We now combine this theorem with a suitable diagonal selection procedure to establish the existence of a subsequence $\{n_k\}_{k=1}^\infty$ of indices such that

$$\lim_{k\to\infty} u_j^{n_k} = u_j \quad \text{strongly in } C[0,T] \qquad \forall\, j \in \mathbb{N}.$$

To this end, we proceed as follows: first, we select a subsequence $\{u_1^{n_\ell}\}$, $\ell = 1, 2, \ldots$, that converges uniformly on $[0,T]$. We now choose for each j the element $u_j^{n_1}$ as the first term. Next, we consider the sequence $\{u_2^{n_\ell}\}$, $\ell = 2, 3, \ldots$, from which we again select a uniformly convergent subsequence $\{u_2^{n_{\ell_m}}\}$. Obviously, the sequence $\{u_1^{n_{\ell_m}}\}$, being a subsequence of $\{u_1^{n_\ell}\}$, also converges uniformly. Now we choose for each j the element $u_j^{n_{\ell_1}}$ as the second term; that is, we put $n_2 := n_{\ell_1}$.

Continuing this selection process inductively for all $j \in \mathbb{N}$, we obtain a subsequence $\{n_k\}_{k=1}^\infty$ of indices such that all the sequences $\{u_j^{n_k}\}$ converge uniformly on $[0,T]$. Observe that for any of the sequences $\{u_j^{n_k}\}$, at most the first $j-1$ terms do not belong to the selected convergent subsequences.

With the limit functions u_j thus constructed, we define the function

$$y(x,t) := \sum_{i=1}^\infty u_i(t)\, v_i(x), \qquad (x,t) \in Q.$$

It can then be shown that the sequence $\{y_{n_k}(\cdot,t)\}$ converges weakly in $L^2(\Omega)$ to $y(\cdot,t)$, uniformly with respect to $t \in [0,T]$. Owing to the weak lower sequential semicontinuity of the norm, we infer from the estimate (7.23) that $\|y(t)\| \le K$ for almost all t, which means that $y \in L^\infty(0,T;L^2(\Omega))$. Moreover, $y_{n_k}(0)$ even converges strongly in $L^2(\Omega)$ to y_0; indeed, we have, as $k \to \infty$,

$$\|y_{n_k}(0) - y_0\| = \left\| \sum_{i=1}^{n_k} u_i^{n_k}(0)\, v_i - \sum_{i=1}^\infty (y_0\,,\, v_i)\, v_i \right\| = \left\| \sum_{i=n_k+1}^\infty u_i(0)\, v_i \right\| \to 0,$$

since $\sum\limits_{i=1}^\infty |u_i(0)|^2 < \infty$ by (7.24). All of the derivations above can be found in full detail in [**LSU68**].

(iv) *y is a weak solution*

By virtue of the estimate (7.23), we may assume without loss of generality that $\{y_{n_k}\}_{k=1}^{\infty}$ converges weakly in $W^{1,0}(Q)$ to y. Next observe that we can take as test function in (7.15) any function v_m of the form

$$v_m(x,t) = \sum_{j=1}^{m} \alpha_j(t)\, v_j(x), \quad m \le n,$$

where $\alpha_j \in C^1[0,T]$ satisfies $\alpha_j(T) = 0$ for $1 \le j \le m$. It then follows from (7.18) that

$$\left(\frac{d}{dt}\, y_{n_k}(t)\,,\, v_m(t)\right) + a\big[t; y_{n_k}(t), v_m(t)\big] = \big(f(t)\,,\, v_m(t)\big) + \big(g(t)\,,\, v_m(t)\big)_\Gamma,$$

whence, upon integrating over $[0,T]$ by parts,

$$\begin{aligned} -\int_0^T \left(y_{n_k}(t)\,,\, \frac{d}{dt}\, v_m(t)\right) dt + \int_0^T a\big[t; y_{n_k}(t), v_m(t)\big]\, dt \\ = \iint_Q f\, v_m\, dx\, dt + \iint_\Sigma g\, v_m\, ds\, dt + \big(y_{n_k}(\cdot,0)\,,\, v_m(\cdot,0)\big). \end{aligned}$$

Now recall that $y_{n_k} \to y$ weakly in $W_2^{1,0}(Q)$ and $y_{n_k}(0) \to y_0$ strongly in $L^2(\Omega)$. Passage to the limit as $k \to \infty$ in the above equation therefore yields

$$\begin{aligned} -\int_0^T \left(y(t)\,,\, \frac{d}{dt}\, v_m(t)\right) dt + \int_0^T a[t; y(t), v_m(t)]\, dt \\ = \iint_Q f\, v_m\, dx\, dt + \iint_\Sigma g\, v_m\, ds\, dt + \int_\Omega y_0\, v_m(\cdot,0)\, dx. \end{aligned}$$

Finally, we use the fact that, by Lemma 4.12 in Chapter II of [**LSU68**], the set of all functions v_m of the above type is dense in the class of all functions from $W_2^{1,1}(Q)$ having zero final value. Therefore, y satisfies the variational formulation and is thus a weak solution. This concludes the proof of the assertion. □

Remark. The assumption that α be nonnegative is dispensable; see, e.g., Raymond and Zidani [**RZ98**].

We now turn to the uniqueness question. The proof of uniqueness of the solution is somewhat technical. It is based on the energy equality

(7.25)
$$\frac{1}{2}\,\|y(\tau)\|^2_{L^2(\Omega)} + \int_0^\tau a\big[t; y(t), y(t)\big]\,dt = \frac{1}{2}\,\|y(0)\|^2_{L^2(\Omega)} + \int_0^\tau \Big[\big(f(t)\,,\,y(t)\big)_{L^2(\Omega)} + \big(g(t)\,,\,y(t)\big)_{L^2(\Gamma)}\Big]\,dt.$$

Clearly, the above equality follows from the variational formulation for the solution when y itself is inserted as the test function. However, this is not allowed, since y does not necessarily have the properties $y \in W_2^{1,1}(Q)$ and $y(T) = 0$ that are required for test functions. In [**LSU68**], it was shown by means of averaged functions that the above energy balance equation is indeed valid, provided we have more regularity, namely, $y \in V_2^{1,0}(Q)$. Hence, if the assumptions of Theorem 7.8 hold and y is a solution belonging to $V_2^{1,0}(Q)$, then there exists a constant $c_P > 0$, which does not depend on f, g, or y_0, such that

(7.26)
$$\max_{t\in[0,T]} \|y(t)\|_{L^2(\Omega)} + \|y\|_{W_2^{1,0}(Q)} \le c_P\,\big(\|f\|_{L^2(Q)} + \|g\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)}\big).$$

Since the energy balance equation (7.25) has the same structure as equation (7.19), the inequality (7.26) can be derived from it in the same way as the estimates (7.22) and (7.23) for y_n were derived from (7.19).

By means of similar estimates, it can then be shown that in $W_2^{1,0}(Q)$ there is also at most one solution; see [**LSU68**], Chapter III, Theorem 3.3. The reason for this is the fact that the difference of two solutions solves the initial-boundary value problem with homogeneous and thus smooth data. In this way, an estimate resembling (7.25) can be employed, which finally leads to the conclusion that the difference of two solutions must vanish; see [**LSU68**], Chapter III, Theorem 3.2.

By virtue of Theorem 7.8, we know that there is at least one solution in $V_2(Q)$. In [**LSU68**], Chapter III, Theorem 4.2, it is shown that any weak solution in $V_2(Q)$ even belongs to $V_2^{1,0}(Q)$. In view of the uniqueness in $W_2^{1,0}(Q)$, we may thus conclude that the unique solution y belongs to $V_2^{1,0}(Q)$. This information makes it possible to derive the estimate (7.26) for the solution.

Remark. The results cited from [**LSU68**], which were proved there for homogeneous Dirichlet boundary conditions, carry over to the case of boundary conditions of the third kind.

In summary, we have the following main result.

Theorem 7.9. *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain, and suppose that functions $c_0 \in L^\infty(Q)$, $\alpha \in L^\infty(\Sigma)$, $y_0 \in L^2(\Omega)$, $f \in L^2(Q)$, and $g \in L^2(\Sigma)$ are given. Moreover, let the differential operator $\mathcal{A}$ satisfy the conditions stated in Theorem 7.8. Then the initial-boundary value problem* (7.13) *has a unique solution in $W_2^{1,0}(Q)$ that belongs to $V_2^{1,0}(Q)$. In addition, the solution satisfies the estimate* (7.26) *with a constant $c_P > 0$ that does not depend on f, g, or y_0.*

Remark. The proof becomes simpler if one works in the space $W(0,T)$ right from the beginning; see Lions [**Lio71**] or Wloka [**Wlo82**]. We have followed the arguments of Ladyzhenskaya et al. [**LSU68**] in order to be able to introduce the space $W_2^{1,0}(Q)$ first, in analogy to the treatment of weak solutions in the elliptic case.

The semilinear equation. We now study the semilinear parabolic initial-boundary value problem

$$\boxed{\begin{aligned} y_t + \mathcal{A}\,y + d(x,t,y) &= f && \text{in } Q \\ \partial_{\nu_\mathcal{A}} y + b(x,t,y) &= g && \text{on } \Sigma \\ y(\cdot,0) &= y_0 && \text{in } \Omega. \end{aligned}} \tag{7.27}$$

Here, we follow in parts the proof for the linear case, and also employ ideas from Gajewski et al. [**GGZ74**] and Wloka [**Wlo82**]. Initially, we impose the strong conditions of uniform boundedness and uniform Lipschitz continuity of d and b with respect to y. We prove Lemma 5.3, which was stated on page 267.

Lemma 5.3 *Suppose that Assumptions 5.1 and 5.2 from page 266 hold, and that the assumptions on $\mathcal{A}$ stated in Theorem 7.8 are satisfied. Then the initial-boundary value problem* (7.27) *has for any triple of data $f \in L^2(Q)$, $g \in L^2(\Sigma)$, and $y_0 \in L^2(\Omega)$ a unique weak solution $y \in W(0,T)$.*

Proof: (i) *Galerkin approximation*

We use the same notation as in the proof of Theorem 7.8; in particular, $\{v_n\}_{n=1}^\infty$ has the same meaning and properties. Again, we make the ansatz

$y_n(x,t) = \sum_{i=1}^{n} u_i^n(t)\, v_i(x)$. This time, the bilinear form a has the form

$$a[y,v] := \int_\Omega \sum_{i,j=1}^{N} a_{ij}(x)\, D_i\, y\, D_j\, v\, dx.$$

As in the proof of Theorem 7.8, we obtain from the parabolic problem the following initial value problem for a nonlinear system of ordinary differential equations:

$$\begin{aligned} \frac{d}{dt}\, u_j^n(t) + \sum_{i=1}^{n} u_i^n(t)\, a[v_i, v_j] + \Phi_j\big(t, u^n(t)\big) &= b_j(t), \\ u_j^n(0) &= (y_0\,,\, v_j), \end{aligned} \tag{7.28}$$

for $j = 1, \dots, n$. Here, we have set $b_j(t) := \big(f(t)\,,\, v_j\big) + \big(g(t)\,,\, v_j\big)_\Gamma$ and

$$\Phi_j(t,u) := \left(d\big(\cdot, t, \sum_{i=1}^{n} u_i\, v_i\big)\,,\, v_j \right) + \left(b\big(\cdot, t, \sum_{i=1}^{n} u_i\, v_i\big)\,,\, v_j \right)_\Gamma .$$

By assumption, the mapping $\Phi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ is uniformly Lipschitz continuous and uniformly bounded. Hence, owing to Carathéodory's theorem, for every $n \in \mathbb{N}$ the above initial value problem has a unique absolutely continuous solution $u^n(\cdot) \in (H^1(0,T))^n$ in $[0,T]$. As in the linear case, we obtain, for almost every $t \in (0,T)$,

$$\begin{aligned} &\left(\frac{d}{dt}\, y_n(t)\,,\, y_n(t) \right) + a\big[y_n(t), y_n(t)\big] + \big(d\big(\cdot, t, y_n(t)\big)\,,\, y_n(t)\big) \\ &\quad + \big(b\big(\cdot, t, y_n(t)\big)\,,\, y_n(t)\big)_\Gamma \;=\; \big(f(t)\,,\, y_n(t)\big) + \big(g(t)\,,\, y_n(t)\big)_\Gamma . \end{aligned} \tag{7.29}$$

(ii) *Estimates for* $\{y_n\}$

Without loss of generality, we may assume that $d(\cdot,\cdot,0) = b(\cdot,\cdot,0) = 0$ (otherwise, we subtract these terms from both sides of (7.27)). As in the derivation of (7.19), it follows from the monotonicity of d and b that

$$\big(d(\cdot,t,y_n)\,,\, y_n\big) = \big(d(\cdot,t,y_n) - d(\cdot,t,0)\,,\, y_n - 0\big) \geq 0.$$

An analogous inequality holds for b. We therefore have

$$\begin{aligned} &\frac{1}{2}\, \|y_n(\tau)\|^2 + \int_0^\tau a\big[t; y_n(t), y_n(t)\big]\, dt \\ &\quad \leq \frac{1}{2}\, \|y_n(0)\|^2 + \int_0^\tau \Big\{ \big(f(t)\,,\, y_n(t)\big) + \big(g(t)\,,\, y_n(t)\big)_\Gamma \Big\}\, dt, \end{aligned} \tag{7.30}$$

as well as the estimate (7.20) for $y_n(0)$. We thus obtain an analogous inequality in place of the equality (7.19). Since we estimated from above anyway in the further steps of the proof of Theorem 7.8, we may argue as in that proof to arrive at the estimate (7.23):

$$\|y_n\|_{C([0,T],H)} + \|y_n\|_{W_2^{1,0}(Q)} \le K \quad \forall\, n \in \mathbb{N}.$$

(iii) *Weak convergence of* $\{y_n\}$

From the above estimate, we conclude that some subsequence, without loss of generality $\{y_n\}_{n=1}^{\infty}$ itself, converges weakly in $W_2^{1,0}(Q)$ to some $y \in W_2^{1,0}(Q)$. Since d and b are nonlinear, we cannot conclude from this that $d(\cdot, y_n)$ (respectively, $b(\cdot, y_n)$) converges weakly to $d(\cdot, y)$ (respectively, $b(\cdot, y)$). However, we do know that these sequences are bounded in $L^2(Q)$ and $L^2(\Sigma)$, respectively. We may therefore assume without loss of generality that

$$d(\cdot, y_n) \rightharpoonup D \text{ in } L^2(Q), \qquad b(\cdot, y_n) \rightharpoonup B \text{ in } L^2(\Sigma), \tag{7.31}$$

with suitable $D \in L^2(Q)$ and $B \in L^2(\Sigma)$. Passing to the limit as $n \to \infty$, we find that y is the weak solution to a linear auxiliary problem: indeed, as in the linear case, it follows that for all functions $v_m(\cdot, t) = \sum_{i=1}^{m} \alpha_i(t)\, v_i(\cdot)$ where $\alpha_i \in C^1[0,T]$ and $\alpha_i(T) = 0$ for $1 \le i \le m$, we have

$$\begin{aligned} -\int_0^T \Big(y(t)\,,\, \frac{d}{dt}\, v_m(t)\Big)\, dt + \int_0^T \Big\{ a\big[y(t), v_m(t)\big] + \big(D(t)\,,\, v_m(t)\big) \\ + \big(B(t)\,,\, v_m(t)\big)_\Gamma \Big\}\, dt \\ = \iint_Q f\, v_m\, dx\, dt + \iint_\Sigma g\, v_m\, ds\, dt + \int_\Omega y_0\, v_m(\cdot, 0)\, dx\,. \end{aligned}$$

Since the set of all functions v_m of the above form is dense in $W^{1,1}(Q)$ (see **[LSU68]**, Chapter II, Lemma 4.12), we can infer that

$$\begin{aligned} -\int_0^T \Big(y(t)\,,\, \frac{d}{dt}\, v(t)\Big)\, dt + \int_0^T \Big\{ a\big[y(t), v(t)\big] + \big(D(t)\,,\, v(t)\big) \\ + \big(B(t)\,,\, v(t)\big)_\Gamma \Big\}\, dt \\ = \iint_Q f\, v\, dx\, dt + \iint_\Sigma g\, v\, ds\, dt + \int_\Omega y_0\, v(\cdot, 0)\, dx \end{aligned}$$

for all $v \in W^{1,1}(Q)$ with $v(T) = 0$. But this is none other than the variational equation for a weak solution y having initial value $y(0) = y_0$. Hence, if we were to succeed in showing that $D(x,t) = d(x,t,y(x,t))$ and

$B(x,t) = b(x,t,y(x,t))$ almost everywhere, then y would be the desired solution and the proof would be complete.

To this end, we put $Y := L^2(0,T;V)$. Then $Y^* = L^2(0,T;V^*)$, and we have

$$y' + w = F \tag{7.32}$$

in Y^*, where $F \in Y^*$ is defined by

$$\langle F\,,\,v\rangle_{Y^*,Y} = \iint_Q f\,v\,dx\,dt + \iint_\Sigma g\,v\,ds\,dt$$

and $w \in Y^*$ by

$$\langle w\,,\,v\rangle_{Y^*,Y} = \int_0^T \Big\{ a\big[y(t),v(t)\big] + \big(D(t)\,,\,v(t)\big) + \big(B(t)\,,\,v(t)\big)_\Gamma \Big\}\,dt.$$

(iv) *y is a weak solution to the nonlinear problem*

We follow the lines of the proof of Theorem 1.1 in Chapter VI of **[GGZ74]**. By construction, $y_n \in W(0,T)$ for all $n \in \mathbb{N}$. As in part (i) of the proof of Theorem 4.4, we define a monotone operator $A : W(0,T) \to L^2(0,T;V^*)$ by $A = A_1 + A_2 + A_3$, with $A_i : W(0,T) \to L^2(0,T;V^*)$, $y \mapsto z_i$, for $i = 1,2,3$, where

$$z_1(t) = a[y(t)\,,\,\cdot\,], \quad z_2(t) = d\big(\cdot,t,y(t)\big), \quad z_3(t) = b\big(\cdot,t,y(t)\big).$$

From (7.29), we infer that

$$\begin{aligned}\int_0^T \big(y_n'(t)\,,\,y_n(t)\big)_{V^*,V}\,dt + \int_0^T \big(A(y_n)(t)\,,\,y_n(t)\big)_{V^*,V}\,dt \\ = \iint_Q f\,y_n\,dx\,dt + \iint_\Sigma g\,y_n\,ds\,dt\end{aligned}$$

for every $n \in \mathbb{N}$. By virtue of formula (3.30) on page 148, and using the definition of F, we have

$$\int_0^T \big(A(y_n)(t)\,,\,y_n(t)\big)_{V^*,V}\,dt = \langle F\,,\,y_n\rangle_{Y^*,Y} + \frac{1}{2}\,\|y_n(0)\|_H^2 - \frac{1}{2}\,\|y_n(T)\|_H^2.$$

Evidently, $y_n(0) \to y_0$ strongly in $H = L^2(\Omega)$. Moreover, it follows from $y_n \rightharpoonup y$ in $W(0,T)$ and the continuity of the linear operator $y \mapsto y(T)$ in $W(0,T)$ that also $y_n(T) \rightharpoonup y(T)$ in H. Therefore,

$$\liminf_{n\to\infty} \|y_n(T)\|_H \geq \|y(T)\|_H,$$

whence, upon using (3.30) once more,

$$\begin{aligned}\overline{\lim_{n\to\infty}}\langle A(y_n)\,,\,y_n\rangle_{Y^*,Y} &\le \langle F\,,\,y\rangle_{Y^*,Y} + \frac{1}{2}\,\|y(0)\|_H^2 - \frac{1}{2}\,\|y(T)\|_H^2\\ &= \langle F\,,\,y\rangle_{Y^*,Y} - \langle y'\,,\,y\rangle_{Y^*,Y}\\ &= \langle w\,,\,y\rangle_{Y^*,Y}\end{aligned}$$

by (7.32). In summary, we have shown the relations

$$y_n \rightharpoonup y, \quad A(y_n) \rightharpoonup w, \quad \overline{\lim_{n\to\infty}}\langle A(y_n)\,,\,y_n\rangle \le \langle w\,,\,y\rangle.$$

Hence, we can infer from Lemma 7.11 below that $A(y) = w$, whence, finally, $D = d(\cdot, y)$ and $B = b(\cdot, y)$ follow. This concludes the proof of the assertion. We remark that an alternative way of concluding this proof can be found in Lions [**Lio69**]. □

The assumed uniform boundedness of the nonlinearities is too restrictive. Therefore, the above theorem has restricted applicability; it is more of an auxiliary result. In Theorem 5.5, the above restriction was removed. In the proof of Theorem 5.5, we made use of the following estimate instead:

Lemma 7.10. *The solution $y \in W(0,T)$ that exists by Lemma 5.3 satisfies the estimate*

$$\|y\|_{W(0,T)} \le c_P\left(\|f - d(\cdot,0)\|_{L^2(Q)} + \|g - b(\cdot,0)\|_{L^2(\Sigma)} + \|y_0\|_{L^2(\Omega)}\right), \tag{7.33}$$

with a constant $c_P > 0$ that depends neither on f nor on g.

Proof: Since $y \in W(0,T)$, we may test the nonlinear equation with y itself to obtain the nonlinear analogue of the energy balance equation (7.25):

$$\begin{aligned}\frac{1}{2}\,\|y(\tau)\|_{L^2(\Omega)}^2 + \int_0^\tau \Big\{ a\big[y(t),y(t)\big] &+ \big(d(\cdot,t,y(t))\,,\,y(t)\big)\\ &+ \big(b(\cdot,t,y(t))\,,\,y(t)\big)_\Gamma\Big\}\,dt\\ = \frac{1}{2}\,\|y(0)\|_{L^2(\Omega)}^2 + \int_0^\tau \Big\{ \big(f(t)\,,\,y(t)\big)_{L^2(\Omega)} &+ \big(g(t)\,,\,y(t)\big)_{L^2(\Gamma)}\Big\}\,dt.\end{aligned}$$

Now add to both sides of the above equation the expression

$$-\int_0^\tau \Big\{\big(d(\cdot,t,0)\,,\,y(t)\big) + \big(b(\cdot,t,0)\,,\,y(t)\big)_\Gamma\Big\}\,dt.$$

Applying Young's inequality to the terms on the right-hand side and invoking the monotonicity of b and d, we then easily conclude the validity of (7.33). This concludes the proof of the assertion. □

Finally, we provide the reader with the auxiliary result that was applied in the final step of the proof of Lemma 5.3. Its proof can be found in [**GGZ74**], Chapter III, Lemma 1.3, or in [**Zei95**].

Lemma 7.11. *Let Y be a reflexive Banach space, and suppose that $A : Y \to Y*$ is a monotone and demicontinuous operator. Moreover, let*

$$y_n \rightharpoonup y \text{ in } Y, \quad A(y_n) \rightharpoonup w \text{ in } Y^* \quad \text{as } n \to \infty$$

$$\text{and} \quad \overline{\lim_{n\to\infty}} \langle A(y_n)\,,\, y_n\rangle_{Y^*,Y} \le \langle w\,,\, y\rangle_{Y^*,Y}.$$

Then $A(y) = w$.

7.3.2. Continuity of solutions. In the following, $\mathcal{A}$ is the uniformly elliptic differential operator with coefficients in $L^\infty(\Omega)$ that was introduced in (2.19) on page 37. The continuity result below is a consequence of Theorem 6.8 in Griepentrog [**Gri07a**] on maximal parabolic regularity. Since a proper understanding of this result requires some knowledge of the Sobolev–Morrey spaces defined in [**Gri07b**] and their embedding properties, we will briefly comment on Griepentrog's results at the end of this section.

Lemma 7.12. *Let $\Omega \subset \mathbb{R}^N$ be a bounded Lipschitz domain, and suppose that there are given functions $f \in L^r(Q)$, $g \in L^s(\Sigma)$, and $y_0 \in C(\bar\Omega)$, where $r > N/2+1$ and $s > N+1$. Then the weak solution y to the linear parabolic initial-boundary value problem with Neumann boundary condition*

$$\begin{aligned} y_t + \mathcal{A}\, y &= f \qquad \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y &= g \qquad \text{on } \Sigma \\ y(0) &= y_0 \qquad \text{in } \Omega \end{aligned}$$

belongs to $W(0,T) \cap C(\bar Q)$. Moreover, there exists some constant $c(r,s) > 0$, which does not depend on f, g, or y_0, such that

$$\|y\|_{W(0,T)} + \|y\|_{C(\bar Q)} \le c(r,s)\,\big(\|f\|_{L^r(Q)} + \|g\|_{L^s(\Sigma)} + \|y_0\|_{C(\bar\Omega)}\big).$$

Proof: In realizing the idea for this proof, the author acknowledges J. Griepentrog's help.

(i) In the case where $y_{=}0$, the assertion follows from Theorem 6.8 in [**Gri07a**]; this will be explained later. Owing to the superposition principle, it therefore

remains to show the continuity of the solution y to the initial-boundary value problem

$$\begin{aligned} y_t + \mathcal{A}\, y &= 0 \qquad \text{in } Q \\ \partial_{\nu_{\mathcal{A}}} y &= 0 \qquad \text{on } \Sigma \\ y(0) &= y_0 \qquad \text{in } \Omega. \end{aligned}$$

The estimate then follows from the continuity of the respective solution mappings.

To this end, we adapt the notation from **[Gri07a, Gri07b]** and put $S := (0,T)$ and $V := H^1(\Omega)$. Moreover, we define the continuous linear operator $A : L^2(S;V) \to L^2(S;V^*)$,

$$\int_S \langle (Ay)(t)\,,\, v(t)\rangle_{V^*,V}\, dt = \int_S \int_\Omega \sum_{i,j=1}^N a_{ij}(x) D_i y(x,t) D_j v(x,t)\, dx\, dt. \tag{7.34}$$

With this, the linear initial-boundary value problem attains the form

$$y_t + A\, y = 0, \quad y(0) = y_0, \tag{7.35}$$

with homogeneous Neumann boundary conditions. Problem (7.35) has for any initial datum $y_0 \in L^2(\Omega)$ a unique solution $y \in W(0,T)$. In the following, we will, as in **[Gri07a, Gri07b]**, denote $W(0,T)$ by $W(S;V)$. We aim to show that the solution y belongs to $C(\bar{S}; C(\bar{\Omega})) = C(\bar{Q})$ if $y_0 \in C(\bar{\Omega})$.

To this end, let y_* and y^* denote the minimum and maximum, respectively, of y_0 in $\bar{\Omega}$. It is then a consequence of the maximum principle for parabolic equations and the structure of the operator A that the solution y of (7.35) satisfies

$$y_* \le y(x,t) \le y^* \qquad \text{for a.e. } (x,t) \in \Omega \times S. \tag{7.36}$$

To verify this, one can test the variational formulation of (7.35) with the functions $v = (y - y^*)_+ \in L^2(S;V)$ and $v = (y_* - y)_+ \in L^2(S;V)$. Since this argument, while not trivial, is rather standard in the theory of parabolic equations, we do not go into the details here.

(ii) Next, we approximate y_0 by a sequence of smooth initial data $y_{0,k}$ and show that the associated solutions y_k are continuous.

To this end, we use the fact that $y_0 \in C(\bar{\Omega})$. It follows from the classical Stone–Weierstrass theorem that there exists a sequence $y_{0,k} \in C^\infty(\bar{\Omega})$ of initial values that converges uniformly on $\bar{\Omega}$ to y_0 as $k \to \infty$. By virtue of Theorem 2.4 in **[Gri07b]**, the solutions $y_k \in W(S;V)$ of problem (7.35), which correspond to the regularized initial values $y_{0,k}$, converge as $k \to \infty$

in $W(S;V)$ to the solution y to problem (7.35) that is associated with the initial datum y_0. We claim that $y_k \in C(\bar\Omega \times \bar S)$ for all $k \in \mathbb{N}$.

For this purpose, we define $v_k(t) := y_{0,k}$ for all $t \in \bar S$ and $k \in \mathbb{N}$. The functions v_k are constant in time and smooth in space. Let $w_k \in W(S;V)$ denote the associated solutions to the problem

$$(w_k)_t + A\,w_k = -A\,v_k, \quad w_k(0) = 0. \tag{7.37}$$

Owing to Theorem 5.6 in **[Gri07b]**, every $A\,v_k$, $k \in \mathbb{N}$, belongs to the Sobolev–Morrey space $L_2^{N+2}(S;V^*)$. By virtue of the maximal parabolic regularity for problems of type (7.37) (see **[Gri07a]**, Theorem 6.8), there is some exponent $\omega \in (N, N+2)$ such that w_k belongs to the Sobolev–Morrey space $W^\omega(S;V)$ for every $k \in \mathbb{N}$. By Theorems 3.4 and 6.8 in **[Gri07b]**, the latter space is continuously embedded in $C(\bar S; C(\bar\Omega))$. Hence w_k, and thus also $y_k = w_k + v_k$, belongs to the space $C(\bar S; C(\bar\Omega))$ for every $k \in \mathbb{N}$, which proves our claim.

(iii) It is now easy to deduce the continuity of y. Indeed, for any $k, \ell \in \mathbb{N}$ the difference $y_k - y_\ell \in W(S;V)$ is the solution to problem (7.35) associated with the initial datum $y_{0,k} - y_{0,\ell}$. From step (ii), it follows that $y_k - y_\ell \in C(\bar S; C(\bar\Omega))$. Hence, applying (7.36) to $y_k - y_\ell$, we have the estimate

$$\min_{x\in\bar\Omega} \big(y_{0,k}(x) - y_{0,\ell}(x)\big) \le y_k(x,t) - y_\ell(x,t) \le \max_{x\in\bar\Omega} \big(y_{0,k}(x) - y_{0,\ell}(x)\big) \tag{7.38}$$

for all k, $\ell \in \mathbb{N}$ and all $(x,t) \in \bar\Omega \times \bar S$. Now, $\{y_{0,k}\}_{k=1}^\infty$ converges uniformly to y_0, and therefore is a Cauchy sequence in $C(\bar\Omega)$. But then it follows immediately from (7.38) that $\{y_k\}_{k=1}^\infty$ is a Cauchy sequence and thus convergent in $C(\bar S; C(\bar\Omega))$. Since $y_k \to y$ in $W(S;V)$ and y is the solution of (7.35) associated with y_0, we finally obtain that $y \in C(\bar S; C(\bar\Omega))$, which concludes the proof of the lemma. □

The above result can be found in **[Cas97]**. In **[RZ99]**, a detailed proof using strongly continuous semigroups was given under somewhat more restrictive smoothness assumptions on the coefficients of the differential operator and on the boundary Γ.

Meanwhile, the papers **[Gri07b, Gri07a]** on maximal parabolic regularity include the above result as a special case. In this connection, Ω can even be a bounded Lipschitz domain in the sense of Grisvard **[Gri85]**; this fact may become important, since two three-dimensional rectangular boxes laid on top of each other in the form of a cross define a Lipschitz domain in the sense of Grisvard but not in the sense of Nečas, and hence they do not

constitute a regular domain in the sense of our definition given in Section 2.2.2.

Remarks.

(i) If the coefficients of $\mathcal{A}$ and the boundary Γ are so smooth that there exists a Green's function $G(x, \xi, t)$ for the above linear initial-boundary value problem, then Lemma 7.12 can be proved rather easily. In fact, in this case y can be represented in the form

$$y(x,t) = \int_0^t \int_\Omega G(x,\xi,t-\tau)\, f(\xi,\tau)\, d\xi\, d\tau + \int_0^t \int_\Gamma G(x,\xi,t-\tau)\, g(\xi,\tau)\, ds(\xi) d\tau + \int_\Omega G(x,\xi,t)\, y_0(\xi)\, d\xi.$$

We have already met the spatially one-dimensional analogue of this formula, (3.13) on page 126. For $t > 0$ and x, $\xi \in \bar{\Omega}$ with $x \neq \xi$, the function G obeys an estimate of the form

$$|G(x,\xi,t)| \leq c_1\, t^{-N/2} \exp\left(-c_2\, \frac{|x-\xi|^2}{\tau}\right)$$

with positive constants c_1 and c_2. With this, the lemma can be proved directly by estimation; see **[Trö84b]**, Lemma 5.6.6.

(ii) The mapping $y_0 \mapsto y$ is continuous from $C(\bar{\Omega})$ into $C(\bar{Q})$ if, for instance, the elliptic differential operator generates a continuous semigroup in $C(\bar{\Omega})$. This property was shown in **[War06]** for the Laplacian and homogeneous variational boundary conditions, and in **[Nit09]** for the above operator $\mathcal{A}$. Inhomogeneous Dirichlet boundary conditions were treated in **[ABHN01]**.

Maximal regularity of parabolic problems. In **[Gri07b, Gri07a]**, parabolic equations of the type $u' + Au + Bu = F$ were investigated for zero initial conditions. For our purposes, the special case

$$u' + Au = F, \quad u(0) = 0 \tag{7.39}$$

suffices, since for $y_0 = 0$ the parabolic initial-boundary value problem from Lemma 7.12 can be rewritten in the form (7.39). To this end, the operator $A : L^2(0,T;V) \to L^2(0,T;V^*)$ introduced in (7.34) on page 379 is used. We consider (7.39) in the time interval $S = (0,T)$ and with $V = H^1(\Omega)$.

In this context, the *Morrey spaces* $L_2^\omega(S; L^2(\Omega))$, $L_2^\omega(S; L^2(\Gamma))$, and $L_2^\omega(S; V^*)$ defined in **[Gri07b]** are needed. We do not give a precise definition of these spaces here, since that would be beyond the scope of this book. Indeed, we are mainly concerned with their embedding properties and with regularity results for the solutions to (7.39). The space

$$L_2^\omega(S;V) = \{u \in L^2(S;V) \,:\, u \in L_2^\omega(S;L^2(\Omega)),\, |\nabla u| \in L_2^\omega(S;L^2(\Omega))\}$$

also plays a role. Finally, in the (for us) important case of the duality mapping $E : V \to V^*$, the *Sobolev–Morrey space*

$$W^\omega(S;V) = \{u \in L_2^\omega(S;V) \,:\, u' \in L_2^\omega(S;V^*)\}$$

is needed; see Definition 6.1 in **[Gri07b]** with $X = Y = V$. The interested reader will find a comprehensive discussion of all these Sobolev–Morrey spaces and their properties in **[Gri07b]**.

Remark. In **[Gri07a]**, when $G = \Omega \cup \Gamma \subset \mathbb{R}^N$ the notation $H_0^1(G)$ is used for the subspace of $H^1(\Omega)$ consisting of those functions that vanish on the Dirichlet boundary, that is, on the complement of the boundary portion $\Gamma \subset \partial\Omega$ on which Neumann data are prescribed. Since in our case $\Gamma = \partial\Omega$, the Dirichlet boundary is empty, and thus $H_0^1(G) = H^1(\Omega)$ and $H^{-1}(G) = H^1(\Omega)^* = V^*$.

Understanding of the results from **[Gri07a]** on maximal parabolic regularity and their application to the proof of Lemma 7.12 will be facilitated by the following facts.

(i) By Theorem 6.8 in **[Gri07a]**, there is some $\bar\omega \in (N, N+2]$ such that for any $\omega \in [0, \bar\omega)$ the restriction of the parabolic differential operator

$$P : u \mapsto u' + Au$$

to the space $\{u \in W^\omega(S;V) : u(0) = 0\}$ is a linear isomorphism between $\{u \in W^\omega(S;V) : u(0) = 0\}$ and $L_2^\omega(S;V^*)$.

(ii) The space $W^\omega(S;V)$ has the important property that for $\omega > N$ it is continuously embedded in some space $C^{0,\kappa}(\bar Q)$ of Hölder continuous functions; see Remark 6.1 in **[Gri07a]**. Then the assertion of Lemma 7.12 for the case of $y_0 = 0$ follows from the fact that under our assumptions on r and s the mapping $(f,g) \mapsto y$ is continuous from $L^r(Q) \times L^s(\Sigma)$ into $W^\omega(S;V)$ for some $\omega \in (N, \bar\omega)$. This can be deduced from the results in **[Gri07b]** in the following way:

(iii) Owing to Theorem 5.6 in **[Gri07b]**, for $\omega \in [0, N+2]$ the space $L_2^\omega(S;V^*)$ contains all those functionals $F \in L^2(S;V^*)$ that can be represented in the form

$$\int_S \langle F(t)\,,\, \varphi(t)\rangle_{V^*,V}\, dt \;=\; \int_S \int_\Omega \sum_{i=1}^m f_i(x,t)\, D_i\varphi(x,t)\, dx\, dt$$

$$+ \int_S \int_\Omega f(x,t)\, \varphi(x,t)\, dx\, dt + \int_S \int_\Gamma g(x,t)\, \varphi(x,t)\, ds(x) dt,$$

where

$$f_1, \dots, f_N \in L_2^{\omega}(S; L^2(\Omega)), \quad f \in L_2^{\omega-2}(S; L^2(\Omega)), \quad g \in L_2^{\omega-1}(S; L^2(\Gamma)),$$

and the mapping $(f_1, \dots, f_N, f, g) \mapsto F$ defines a continuous linear operator. The connection between these Morrey spaces and the usual Lebesgue spaces, in which our assumptions for the data f and g are formulated, is revealed by Remarks 3.4 and 3.7 in **[Gri07b]**. We obtain:

(a) If $q \geq 2$ and $\omega_q = (N+2)(1-2/q)$, then $L^q(Q)$ is continuously embedded in $L_2^{\omega_q}(S; L^2(\Omega))$, and we have $\omega_q > N$ if $q > N+2$.

(b) If $r \geq 2$ and $\omega_r = 2 + (N+2)(1-2/r)$, then $L^r(Q)$ is continuously embedded in $L_2^{\omega_r - 2}(S; L^2(\Omega))$, and we have $\omega_r > N$ if $r > N/2 + 1$.

(c) If $s \geq 2$ and $\omega_s = 1 + (N+1)(1-2/s)$, then $L^s(\Sigma)$ is continuously embedded in $L_2^{\omega_s - 1}(S; L^2(\Gamma))$, and we have $\omega_s > N$ if $s > N+1$.

Obviously, the last two conditions, namely $r > N/2+1$ and $s > N+1$, are exactly those imposed on f and g in Lemma 7.12.

Bibliography

[ABHN01] C. Arendt, C. Batty, M. Hieber, and F. Neubrander, *Vector-valued Laplace Transforms and Cauchy Problems*, Birkhäuser, Basel, 2001.

[ACT02] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl. **23** (2002), 201–229.

[Ada78] R. A. Adams, *Sobolev Spaces*, Academic Press, Boston, 1978.

[AEFR00] N. Arada, H. El Fekih, and J.-P. Raymond, *Asymptotic analysis of some control problems*, Asymptot. Anal. **24** (2000), 343–366.

[AH01] K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, Shape Optimization and Optimal Design, Lect. Notes Pure Appl. Math., vol. 216, Marcel Dekker, 2001, pp. 317–332.

[Alt99] H. W. Alt, *Lineare Funktionalanalysis*, Springer, Berlin, 1999.

[Alt02] W. Alt, *Nichtlineare Optimierung*, Vieweg, Wiesbaden, 2002.

[AM84] W. Alt and U. Mackenroth, *On the numerical solution of state constrained coercive parabolic optimal control problems*, Optimal Control of Partial Differential Equations (K.-H. Hoffmann and W. Krabs, eds.), Int. Ser. Numer. Math., vol. 68, Birkhäuser, 1984, pp. 44–62.

[AM89] ———, *Convergence of finite element approximations to state constrained convex parabolic boundary control problems*, SIAM J. Control Optim. **27** (1989), 718–736.

[AM93] W. Alt and K. Malanowski, *The Lagrange–Newton method for nonlinear optimal control problems*, Comput. Optim. Appl. **2** (1993), 77–100.

[Ant05] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, SIAM, Philadelphia, 2005.

[App88] J. Appell, *The superposition operator in function spaces – A survey*, Expo. Math. **6** (1988), 209–270.

[AR97] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim. **18** (1997), no. 3–4, 235–250.

[ART02] N. Arada, J.-P. Raymond, and F. Tröltzsch, *On an augmented Lagrangian SQP method for a class of optimal control problems in Banach spaces*, Comput. Optim. Appl. **22** (2002), 369–398.

[AT81] N. U. Ahmed and K. L. Teo, *Optimal Control of Distributed Parameter Systems*, North Holland, New York, 1981.

[AT90] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theor. Comput. Fluid Dyn. **1** (1990), 303–325.

[AZ90] J. Appell and P. P. Zabrejko, *Nonlinear Superposition Operators*, Cambridge University Press, Cambridge, 1990.

[Bal65] A. V. Balakrishnan, *Optimal control problems in Banach spaces*, SIAM J. Control **3** (1965), 152–180.

[Bar93] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[BBEFR03] F. Ben Belgacem, H. El Fekih, and J.-P. Raymond, *A penalized Robin approach for solving a parabolic equation with nonsmooth Dirichlet boundary condition*, Asymptot. Anal. **34** (2003), 121–136.

[BC91] F. Bonnans and E. Casas, *Une principe de Pontryagine pour le contrôle des systèmes semilinéaires elliptiques*, J. Differential Equations **90** (1991), 288–303.

[BC95] ———, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim. **33** (1995), 274–298.

[BDPDM92] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems, Vol. I*, Birkhäuser, Basel, 1992.

[BDPDM93] ———, *Representation and Control of Infinite Dimensional Systems, Vol. II*, Birkhäuser, Basel, 1993.

[Ber82] D. M. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim. **20** (1982), 221–246.

[Bet01] J. T. Betts, *Practical Methods for Optimal Control Using Nonlinear Programming*, SIAM, Philadelphia, 2001.

[BIK99] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim. **37** (1999), 1176–1194.

[Bit75] L. Bittner, *On optimal control of processes governed by abstract functional, integral and hyperbolic differential equations*, Math. Methods Oper. Res. **19** (1975), 107–134.

[BK01] A. Borzi and K. Kunisch, *A multigrid method for optimal control of time-dependent reaction diffusion processes*, Fast Solution of Discretized Optimization Problems (K. H. Hoffmann, R. Hoppe, and V. Schulz, eds.), Int. Ser. Numer. Math., vol. 138, Birkhäuser, 2001, pp. 513–524.

[BK02a] M. Bergounioux and K. Kunisch, *On the structure of the Lagrange multiplier for state-constrained optimal control problems*, Systems Control Lett. **48** (2002), 16–176.

[BK02b] ———, *Primal-dual active set strategy for state-constrained optimal control problems*, Comput. Optim. Appl. **22** (2002), 193–224.

[BKR00] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of partial differential equations: Basic concepts*, SIAM J. Control Optim. **39** (2000), 113–132.

[Bor03] A. Borzi, *Multigrid methods for parabolic distributed optimal control problems*, J. Comput. Appl. Math. **157** (2003), 365–382.

[BP78] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Bucharest, and Sijthoff & Noordhoff, Leyden, 1978.

[BR01] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer. **10** (2001), 1–102.

[Bra07] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Elasticity Theory,* 4th Edition, Cambridge University Press, Cambridge, 2007.

[BS94] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[BS96] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Springer, New York, 1996.

[BS00] F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[But69] A. G. Butkovskii, *Distributed Control Systems*, American Elsevier, New York, 1969.

[But75] ———, *Methods for the Control of Distributed Parameter Systems* (in Russian), Isd. Nauka, Moscow, 1975.

[Car67] H. Cartan, *Calcul Différentiel. Formes Différentielles*, Hermann, Paris, 1967.

[Cas86] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim. **4** (1986), 1309–1322.

[Cas92] ———, *Introduccion a las Ecuaciones en Derivadas Parciales*, Universidad de Cantabria, Santander, 1992.

[Cas93] ———, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim. **31** (1993), 993–1006.

[Cas95] ———, *Optimality conditions for some control problems of turbulent flow*, Flow Control (New York) (M. D. Gunzburger, ed.), Springer, 1995, pp. 127–147.

[Cas97] ———, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim. **35** (1997), 1297–1327.

[CDlRT08] E. Casas, J. C. De los Reyes, and F. Tröltzsch, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, SIAM J. Optim. **12** (2008), no. 2, 616–643.

[CH91] Z. Chen and K. H. Hoffmann, *Numerical solutions of the optimal control problem governed by a phase field model*, Estimation and Control of Distributed Parameter Systems (W. Desch, F. Kappel, and K. Kunisch, eds.), Int. Ser. Numer. Math., vol. 100, Birkhäuser, 1991, pp. 79–97.

[Cia78] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[CM02a] E. Casas and M. Mateos, *Second order sufficient optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim. **40** (2002), 1431–1454.

[CM02b] ———, *Uniform convergence of the FEM. Applications to state constrained control problems*, J. Comput. Appl. Math. **21** (2002), 67–100.

[CMT05] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl. **31** (2005), 193–220.

[Con90] J. B. Conway, *A Course in Functional Analysis*, Wiley & Sons, Warsaw/Dordrecht, 1990.

[Cor07] J. M. Coron, *Control and Nonlinearity*, American Mathematical Society, Providence, 2007.

[CR06] E. Casas and J.-P. Raymond, *Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations*, SIAM J. Control Optim. **45** (2006), 1586–1611.

[CT99] E. Casas and F. Tröltzsch, *Second-order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim. **39** (1999), 211–227.

[CT02] ———, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim. **13** (2002), 406–431.

[CTU96] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, Z. Anal. Anwendungen (ZAA) **15** (1996), 687–707.

[CTU00] ———, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim. **38** (2000), no. 5, 1369–1391.

[Deu04] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Springer Series in Computational Mathematics, vol. 35, Springer, Berlin, 2004.

[DHPY95] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Appl. Math. Optim. **31** (1995), 297–326.

[DT09] V. Dhamo and F. Tröltzsch, *Some aspects of reachability for parabolic boundary control problems with control constraints*, accepted for publication in: Comput. Optim. Appl. (2009).

[Dun98] J. C. Dunn, *On second order sufficient optimality conditions for structured nonlinear programs in infinite-dimensional function spaces*, Mathematical Programming with Data Perturbations (A. Fiacco, ed.), Marcel Dekker, 1998, pp. 83–107.

[ET74] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthier-Villars, 1974.

[ET86] K. Eppler and F. Tröltzsch, *On switching points of optimal controls for coercive parabolic boundary control problems*, Optimization **17** (1986), 93–101.

[Eva98] L. C. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, 1998.

[Fat99] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, 1999.

[Fri64] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, 1964.

[Fur99] A. V. Fursikov, *Optimal Control of Distributed Systems. Theory and Applications*, American Mathematical Society, Providence, 1999.

[Gal94] P. G. Galdi, *An Introduction to the Navier–Stokes Equations*, Springer, New York, 1994.

[GGZ74] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[GKT92] H. Goldberg, W. Kampowsky, and F. Tröltzsch, *On Nemytskij operators in L_p-spaces of abstract functions*, Math. Nachr. **155** (1992), 127–140.

[GM99] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier–Stokes flows with bounded distributed controls*, SIAM J. Control Optim. **37** (1999), 1913–1945.

[GM00] ———, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control*, SIAM J. Numer. Anal. **37** (2000), 1481–1512.

[GMW81] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.

[Goe92] M. Goebel, *Continuity and Fréchet-differentiability of Nemytskij operators in Hölder spaces*, Monatsh. Math. **113** (1992), 107–119.

[GR01] T. Grund and A. Rösch, *Optimal control of a linear elliptic equation with a supremum-norm functional*, Optim. Methods Softw. **15** (2001), 299–329.

[GR05] C. Grossmann and H.-G. Roos, *Numerische Behandlung partieller Differentialgleichungen*, Teubner, Wiesbaden, 2005.

[Gri85] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[Gri02] J. A. Griepentrog, *Linear elliptic boundary value problems with non-smooth data: Campanato spaces of functionals*, Math. Nachr. **243** (2002), 19–42.

[Gri07a] ———, *Maximal regularity for nonsmooth parabolic problems in Sobolev–Morrey spaces*, Adv. Differential Equations **12** (2007), 1031–1078.

[Gri07b] ———, *Sobolev–Morrey spaces associated with evolution equations*, Adv. Differential Equations **12** (2007), 781–840.

[GS77] K. Glashoff and E. W. Sachs, *On theoretical and numerical aspects of the bang-bang principle*, Numer. Math. **29** (1977), 93–113.

[GS80] W. A. Gruver and E. W. Sachs, *Algorithmic Methods in Optimal Control*, Pitman, London, 1980.

[GT93] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim. **31** (1993), 1007–1025.

[GT97a] ———, *On a SQP-multigrid technique for nonlinear parabolic boundary control problems*, Optimal Control: Theory, Algorithms and Applications (Dordrecht) (W. W. Hager and P. M. Pardalos, eds.), Kluwer Academic Publishers, 1997, pp. 154–177.

[GT97b] C. Grossmann and J. Terno, *Numerik der Optimierung*, Teubner, Stuttgart, 1997.

[Gun95] M. D. Gunzburger (ed.), *Flow Control*, Springer, New York, 1995.

[Gun03] ———, *Perspectives in Flow Control and Optimization*, SIAM, Philadelphia, 2003.

[GW76] K. Glashoff and N. Weck, *Boundary control of parabolic differential equations in arbitrary dimensions: Supremum-norm problems*, SIAM J. Control Optim. **14** (1976), 662–681.

[HDMRS09] R. Haller-Dintelmann, C. Meyer, J. Rehberg, and A. Schiela, *Hölder continuity and optimal control for nonsmooth elliptic problems*, Appl. Math. Optim. **60** (2009), 397–428.

[Hei96] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM J. Optim. **6** (1996), 373–417.

[Hei97] ———, *The numerical solution of a control problem governed by a phase field model*, Optim. Methods Softw. **7** (1997), 211–263.

[Heu08] H. Heuser, *Lehrbuch der Analysis, Teil 2*, Vieweg+Teubner, Wiesbaden, 2008.

[Hin99] M. Hinze, *Optimal and instantaneous control of the instationary Navier–Stokes equations*, Habilitation thesis, Technische Universität Berlin, 1999.

[Hin05] ———, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl. **30** (2005), 45–63.

[HJ92] K.-H. Hoffmann and L. Jiang, *Optimal control problem of a phase field model for solidification*, Numer. Funct. Anal. Optim. **13** (1992), 11–27.

[HK01] M. Hinze and K. Kunisch, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim. **40** (2001), 925–946.

[HK04] ———, *Second order methods for boundary control of the instationary Navier–Stokes system*, Z. Angew. Math. Mech. (ZAMM) **84** (2004), 171–187.

[HP57] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, 1957.

[HPUU09] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, vol. 23, Springer, Berlin, 2009.

[HS94] M. Heinkenschloss and E. W. Sachs, *Numerical solution of a constrained control problem for a phase field model*, Control and Estimation of Distributed Parameter Systems (W. Desch, F. Kappel, and K. Kunisch, eds.), Int. Ser. Numer. Math., vol. 118, Birkhäuser, 1994, pp. 171–188.

[HT99] M. Heinkenschloss and F. Tröltzsch, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation*, Control Cybernet. **28** (1999), no. 2, 178–211.

[HV01] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust-region SQP algorithms*, SIAM J. Optim. **12** (2001), 283–302.

[HV03] D. Hömberg and S. Volkwein, *Control of laser surface hardening by a reduced-order approach using proper orthogonal decomposition*, Math. Comput. Modelling **38** (2003), 1003–1028.

[IK96] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim. **34** (1996), 874–891.

[IK00] ———, *Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces*, Nonlinear Anal., Theory Methods Appl. **41** (2000), 591–616.

[IK08] ———, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.

[Iof79] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum. 3: Second order conditions and augmented duality*, SIAM J. Control Optim. **17** (1979), 266–288.

[IT79] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[Jah94] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 1994.

[JK81] D. S. Jerison and C. E. Kenig, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc., New Ser. **4** (1981), 203–207.

[KA64] L. V. Kantorovich and G. B. Akilov, *Functional Analysis in Normed Spaces*, Pergamon Press, Oxford, 1964.

[Kar77] A. Karafiat, *The problem of the number of switches in parabolic equations with control*, Ann. Pol. Math. **XXXIV** (1977), 289–316.

[Kel99] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.

[KK02] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications*, Kluwer Academic Publishers, Dordrecht, 2002.

[Kno77] G. Knowles, *Über das Bang-Bang-Prinzip bei Kontrollproblemen aus der Wärmeleitung*, Arch. Math. **29** (1977), 300–309.

[KR02] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim. **13** (2002), 321–334.

[Kra95] W. Krabs, *Optimal Control of Undamped Linear Vibrations*, Heldermann, Lemgo, 1995.

[Kre78] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, New York, 1978.

[KS80] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[KS92] F.-S. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl. **1** (1992), 113–135.

[KS94] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput. **15** (1994), 645–667.

[KS95] ———, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim. **33** (1995), 1731–1757.

[KV01] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math. **90** (2001), 117–148.

[KV02] ———, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM J. Numer. Anal. **40** (2002), 492–515.

[KV07] K. Kunisch and B. Vexler, *Constrained Dirichlet boundary control in L^2 for a class of evolution equations*, SIAM J. Control Optim. **46** (2007), 1726–1753.

[KZPS76] M. A. Krasnoselskii, P. P. Zabreiko, E. I. Pustylnik, and P. E. Sobolevskii, *Integral Operators in Spaces of Summable Functions*, Noordhoff, Leyden, 1976.

[Las02] I. Lasiecka, *Mathematical Control Theory of Coupled PDEs*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 75, SIAM, Philadelphia, 2002.

[Lio69] J. L. Lions, *Quelques Méthodes des Résolution des Problèmes aux Limites non Linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[Lio71] ———, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[LLS94] J. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser, Boston, 1994.

[LM72] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, vol. 1–3, Springer, Berlin, 1972.

[LS74] L. A. Lusternik and V. J. Sobolev, *Elements of Functional Analysis*, Wiley & Sons, New York, 1974.

[LS00] F. Leibfritz and E. W. Sachs, *Inexact SQP interior point methods and large scale optimal control problems*, SIAM J. Control Optim. **38** (2000), 272–293.

[LS07] C. Lefter and J. Sprekels, *Control of a phase field system modeling non-isothermal phase transitions*, Adv. Math. Sci. Appl. **17** (2007), 181–194.

[LSU68] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, 1968.

[LT00a] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories. I: Abstract Parabolic Systems*, Cambridge University Press, Cambridge, 2000.

[LT00b] ———, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories. II: Abstract Hyperbolic-like Systems over a Finite Time Horizon*, Cambridge University Press, Cambridge, 2000.

[LU73] O. A. Ladyzhenskaya and N. N. Ural'ceva, *Linear and Quasilinear Equations of Elliptic Type* (in Russian), Izd. Nauka, Moscow, 1973.

[Lue69] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[Lue84] ———, *Linear and Nonlinear Programming*, Addison Wesley, Reading, Massachusetts, 1984.

[LY95] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.

[Mac81] U. Mackenroth, *Time-optimal parabolic boundary control problems with state constraints*, Numer. Funct. Anal. Optim. **3** (1981), 285–300.

[Mac82] ———, *Convex parabolic boundary control problems with state constraints*, J. Math. Anal. Appl. **87** (1982), 256–277.

[Mac83a] ———, *On a parabolic distributed optimal control problem with restrictions on the gradient*, Appl. Math. Optim. **10** (1983), 69–95.

[Mac83b] ———, *Some remarks on the numerical solution of bang-bang type optimal control problems*, Numer. Funct. Anal. Optim. **5** (1983), 457–484.

[Mal81] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal control problems*, Appl. Math. Optim. **8** (1981), 69–95.

[Mau81] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Program. Study **14** (1981), 163–177.

[MM00] H. Maurer and H. D. Mittelmann, *Optimization techniques for solving elliptic control problems with control and state constraints. I: Boundary control*, J. Comput. Appl. Math. **16** (2000), 29–55.

[MM01] ———, *Optimization techniques for solving elliptic control problems with control and state constraints. II: Distributed control*, J. Comput. Appl. Math. **18** (2001), 141–160.

[MR04] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim. **43** (2004), 970–985.

[MRT06] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl. **33** (2006), 209–228.

[MS82] J. Macki and A. Strauss, *Introduction to Optimal Control Theory*, Springer, Berlin, 1982.

[MS00] B. Maar and V. Schulz, *Interior point multigrid methods for topology optimization*, Structural Optimization **19** (2000), 214–224.

[MT02] H. D. Mittelmann and F. Tröltzsch, *Sufficient optimality in a parabolic control problem*, Trends in Industrial and Applied Mathematics (Dordrecht) (A. H. Siddiqi and M. Kocvara, eds.), Kluwer Academic Publishers, 2002, pp. 305–316.

[MT06] C. Meyer and F. Tröltzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, Recent Advances in Optimization. Proceedings of the 12th French-German-Spanish Conference on Optimization held in Avignon, September 20–24, 2004 (A. Seeger, ed.), Lect. Notes Econ. Math. Syst., vol. 563, Springer, 2006, pp. 187–204.

[MZ79] H. Maurer and J. Zowe, *First- and second-order conditions in infinite-dimensional programming problems*, Math. Program. **16** (1979), 98–110.

[Nec67] J. Nečas, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Academia, Prague, 1967.

[Nit09] R. Nittka, *Regularity of solutions of linear second order elliptic and parabolic boundary value problems on Lipschitz domains*, arXiv:0906.5285v1, June 2009.

[NPS09] I. Neitzel, U. Prüfert, and T. Slawig, *Strategies for time-dependent PDE control with inequality constraints using an integrated modeling and simulation environment*, Numer. Algorithms **50** (2009), 241–269.

[NST06] P. Neittaanmäki, J. Sprekels, and D. Tiba, *Optimization of Elliptic Systems: Theory and Applications*, Springer, Berlin, 2006.

[NT94] P. Neittaanmäki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms, and Applications*, Marcel Dekker, New York, 1994.

[NW99] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[Paz83] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[PBGM62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[Pen82] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Program. Study **19** (1982), 167–199.

[Pol97] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Applied Math. Sciences, vol. 124, Springer, New York, 1997.

[Rob80] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res. **5** (1980), 43–62.

[Rös04] A. Rösch, *Error estimates for parabolic optimal control problems with control constraints*, Z. Anal. Anwendungen (ZAA) **23** (2004), 353–376.

[Rou02] T. Roubíček, *Optimization of steady-state flow of incompressible fluids*, Analysis and Optimization of Differential Systems (Boston) (V. Barbu, I. Lasiecka, D. Tiba, and C. Varsan, eds.), Kluwer Academic Publishers, 2002, pp. 357–368.

[RT00] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dyn. Syst. **6** (2000), 431–450.

[RT03] T. Roubíček and F. Tröltzsch, *Lipschitz stability of optimal controls for the steady state Navier–Stokes equations*, Control Cybernet. **32** (2003), 683–705.

[RW08] A. Rösch and D. Wachsmuth, *Numerical verification of optimality conditions*, SIAM J. Control Optim. **47** (2008), no. 5, 2557–2581.

[RZ98] J.-P. Raymond and H. Zidani, *Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim. **36** (1998), 1853–1879.

[RZ99] ———, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim. **39** (1999), 143–177.

[Sac78] E. W. Sachs, *A parabolic control problem with a boundary condition of the Stefan–Boltzmann type*, Z. Angew. Math. Mech. (ZAMM) **58** (1978), 443–449.

[Sch79] K. Schittkowski, *Numerical solution of a time-optimal parabolic boundary-value control problem*, J. Optimization Theory Appl. **27** (1979), 271–290.

[Sch80] E. J. P. G. Schmidt, *The bang-bang principle for the time-optimal problem in boundary control of the heat equation*, SIAM J. Control Optim. **18** (1980), 101–107.

[Sch89] ———, *Boundary control for the heat equation with non-linear boundary condition*, J. Differential Equations **78** (1989), 89–121.

[Spe93] P. Spellucci, *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser, Basel, 1993.

[Sta65] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier, Grenoble **15** (1965), 189–258.

[SW98] V. Schulz and G. Wittum, *Multigrid optimization methods for stationary parameter identification problems in groundwater flow*, Multigrid Methods V (W. Hackbusch and G. Wittum, eds.), Lect. Notes Comput. Sci. Eng., vol. 3, Springer, 1998, pp. 276–288.

[SZ92] J. Sprekels and S. Zheng, *Optimal control problems for a thermodynamically consistent model of phase-field type for phase transitions*, Adv. Math. Sci. Appl. **1** (1992), 113–125.

[Tan79] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

[Tem79] R. Temam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.

[Tib90] D. Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lect. Notes Math., vol. 1459, Springer, Berlin, 1990.

[Tri95] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, J. A. Barth, Heidelberg-Leipzig, 1995.

[Trö84a] F. Tröltzsch, *The generalized bang-bang principle and the numerical solution of a parabolic boundary-control problem with constraints on the control and the state*, Z. Angew. Math. Mech. (ZAMM) **64** (1984), 551–557.

[Trö84b] ———, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner Texte zur Mathematik, vol. 62, Teubner, Leipzig, 1984.

[Trö99] ———, *On the Lagrange-Newton-SQP method for the optimal control of semilinear parabolic equations*, SIAM J. Control Optim. **38** (1999), 294–312.

[Trö00] ———, *Lipschitz stability of solutions of linear-quadratic parabolic control problems with respect to perturbations*, Dyn. Contin. Discrete Impulsive Syst. **7** (2000), 289–306.

[TS64] A. N. Tychonov and A. A. Samarski, *Partial Differential Equations of Mathematical Physics, Vol. I*, Holden-Day, San Francisco, 1964.

[TT96] D. Tiba and F. Tröltzsch, *Error estimates for the discretization of state constrained convex control problems*, Numer. Funct. Anal. Optim. **17** (1996), 1005–1028.

[TW06] F. Tröltzsch and D. Wachsmuth, *Second-order sufficient optimality conditions for the optimal control of Navier–Stokes equations*, ESAIM: Control Optim. Calc. Var. **12** (2006), 93–119.

[Ung97] A. Unger, *Hinreichende Optimalitätsbedingungen 2. Ordnung und Konvergenz des SQP-Verfahrens für semilineare elliptische Randsteuerprobleme*, Ph.D. thesis, Technische Universität Chemnitz, 1997.

[UU00] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim. **38** (2000), 1938–1984.

[UUH99] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim. **37** (1999), 731–764.

[Vex07] B. Vexler, *Finite element approximation of elliptic Dirichlet optimal control problems*, Numer. Funct. Anal. Optim. **28** (2007), 957–973.

[Vol01] S. Volkwein, *Optimal control of a phase-field model using proper orthogonal decomposition*, Z. Angew. Math. Mech. (ZAMM) **81** (2001), 83–97.

[vW76] L. v. Wolfersdorf, *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type I*, Z. Angew. Math. Mech. (ZAMM) **56** (1976), 531–538.

[vW77] ———, *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type II*, Z. Angew. Math. Mech. (ZAMM) **57** (1977), 11–17.

[War06] M. Warma, *The Robin and Wentzell–Robin Laplacians on Lipschitz domains*, Semigroup Forum **73** (2006), no. 1, 10–30.

[Wer97] D. Werner, *Funktionalanalysis*, Springer, Berlin, 1997.

[Wlo82] J. Wloka, *Partielle Differentialgleichungen*, Teubner, Leipzig, 1982.

[Wlo87] ———, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

[Wou79] A. Wouk, *A Course of Applied Functional Analysis*, Wiley, New York, 1979.

[Wri93] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1993.

[WS04] M. Weiser and A. Schiela, *Function space interior point methods for PDE constrained optimization*, Proc. Appl. Math. Mech. (PAMM) **4** (2004), 43–46.

[Yos80] K. Yosida, *Functional Analysis*, Springer, New York, 1980.

[Zei86] E. Zeidler, *Nonlinear Functional Analysis and its Applications I: Fixed-point Theorems*, Springer, New York, 1986.

[Zei90a] ———, *Nonlinear Functional Analysis and its Applications II/A: Linear Monotone Operators*, Springer, New York, 1990.

[Zei90b] ———, *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*, Springer, New York, 1990.

[Zei95] ———, *Applied Functional Analysis and its Applications. Main Principles and their Applications*, Springer, New York, 1995.

[ZK79] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim. **5** (1979), 49–62.

Index

Optimal control theory is concerned with finding control functions that minimize cost functions for systems described by differential equations. The methods have found widespread applications in aeronautics, mechanical engineering, the life sciences, and many other disciplines.

Photograph by Elke Weiß

This book focuses on optimal control problems where the state equation is an elliptic or parabolic partial differential equation. Included are topics such as the existence of optimal solutions, necessary optimality conditions and adjoint equations, second-order sufficient conditions, and main principles of selected numerical techniques. It also contains a survey on the Karush-Kuhn-Tucker theory of nonlinear programming in Banach spaces.

The exposition begins with control problems with linear equations, quadratic cost functions and control constraints. To make the book self-contained, basic facts on weak solutions of elliptic and parabolic equations are introduced. Principles of functional analysis are introduced and explained as they are needed. Many simple examples illustrate the theory and its hidden difficulties. This start to the book makes it fairly self-contained and suitable for advanced undergraduates or beginning graduate students.

Advanced control problems for nonlinear partial differential equations are also discussed. As prerequisites, results on boundedness and continuity of solutions to semilinear elliptic and parabolic equations are addressed. These topics are not yet readily available in books on PDEs, making the exposition also interesting for researchers.

Alongside the main theme of the analysis of problems of optimal control, Tröltzsch also discusses numerical techniques. The exposition is confined to brief introductions into the basic ideas in order to give the reader an impression of how the theory can be realized numerically. After reading this book, the reader will be familiar with the main principles of the numerical analysis of PDE-constrained optimization.

ISBN 978-0-8218-4904-0

9 780821 849040

GSM/112

For additional information and updates on this book, visit

www.ams.org/bookpages/gsm-112