Juan Carlos De los Reyes

Numerical PDE-Constrained Optimization

Springer

Juan Carlos De los Reyes
Centro de Modelización Matemática (MODEMAT)
and Department of Mathematics
Escuela Politécnica Nacional Quito
Quito
Ecuador

ISSN 2190-8354 ISSN 2191-575X (electronic)
SpringerBriefs in Optimization
ISBN 978-3-319-13394-2 ISBN 978-3-319-13395-9 (eBook)
DOI 10.1007/978-3-319-13395-9

Library of Congress Control Number: 2014956766

Mathematics Subject Classification (2010): 49K20, 49K21, 49J20, 49J21, 49M05, 49M15, 49M37, 65K10, 65K15, 35J25, 35J60, 35J86.

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To María Soledad and Esteban

Preface

In recent years, the field of partial differential equation (PDE)-constrained optimization has received a significant impulse with large research projects being funded by different national and international agencies. A key ingredient for this success is related to the wide applicability that the developed results have (e.g., in crystal growth, fluid flow, or heat phenomena). In return, application problems gave rise to further deep theoretical and numerical developments. In particular, the numerical treatment of such problems has motivated the design of efficient computational methods in order to obtain optimal solutions in a manageable amount of time.

Although some books on optimal control of PDEs have been edited in the past years, they are mainly concentrated on theoretical aspects or on research-oriented results. At the moment, there is a lack of student accessible texts describing the derivation of optimality conditions and the application of numerical optimization techniques for the solution of PDE-constrained optimization problems. This text is devoted to fill that gap.

By presenting numerical optimization methods, their application to PDE-constrained problems, the resulting algorithms and the corresponding MATLAB codes, we aim to contribute to make the field of numerical PDE-constrained optimization accessible to advanced undergraduate students, graduate students, and practitioners.

Moreover, recent results in the emerging field of nonsmooth numerical PDE-constrained optimization are also presented. Up to the author's knowledge, such results are not part of any monograph yet. We provide an overview on the derivation of optimality conditions and on some solution algorithms for problems involving bound constraints, state constraints, sparsity enhancing cost functionals, and variational inequality constraints.

After an introduction and some preliminaries on the theory and approximation of partial differential equations, the theory of PDE-constrained optimization is presented.

Existence of optimal solutions and optimality conditions are addressed. We use a general framework that allows to treat both linear and nonlinear problems. First order optimality conditions are presented by means of both a reduced approach and a Lagrange multiplier methodology. The derivation is also illustrated with several examples, including linear and nonlinear ones. Also sufficient second-order conditions are developed and the application to semilinear problems is explained.

The next part of the book is devoted to numerical optimization methods. Classical methods (descent, Newton, quasi-Newton, sequential quadratic programming (SQP)) are presented in a general Hilbert-space framework and their application to the special structure of PDE-constrained optimization problems explained. Convergence results are presented explicitly for the PDE-constrained optimization structure. The algorithms are carefully described and MATLAB codes, for representative problems, are included.

The box-constrained case is addressed thereafter. This chapter focuses on bound constraints on the design (or control) variables. First- and second-order optimality conditions are derived for this special class of problems and solution techniques are studied. Projection methods are explained on basis of the general optimization algorithms developed in Chap. 4. In addition, the nonsmooth framework of primal-dual and semismooth Newton methods is introduced and developed. Convergence proofs, algorithms, and MATLAB codes are included.

In the last chapter, some representative nonsmooth PDE-constrained optimization problems are addressed. Problems with cost functionals involving the L^1-norm, with state constraints, or with variational inequality constraints are considered. Numerical strategies for the solution of such problems are presented together with the corresponding MATLAB codes.

This book is based on lectures given at the Humboldt-University of Berlin, at the University of Hamburg, and at the first *Escuela de Control y Optimización (ECOPT)*, a summer school organized together by the Research Center on Mathematical Modeling (MODEMAT) at EPN Quito and the Research Group on Analysis and Mathematical Modeling Valparaíso (AM2V) at USM Chile.

Contents

Chapter 1
Introduction

1.1 Introductory Examples

1.1.1 Optimal Heating

Let Ω be a bounded three-dimensional domain with boundary Γ, which represents a body that has to be heated. We may act along the boundary by setting a temperature $u = u(x)$ and, in that manner, change the temperature distribution inside the body. The goal of the problem consists in getting as close as possible to a given desired temperature distribution $z_d(x)$ in Ω.

Mathematically, the problem may be written as follows:

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

subject to:

$$\left.\begin{array}{ll} -\Delta y = 0 & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} = \rho(u-y) & \text{in } \Gamma, \end{array}\right\} \text{ State equation}$$

$$u_a \le u(x) \le u_b, \qquad \text{Control constraints}$$

where $u_a, u_b \in \mathbb{R}$ such that $u_a \le u_b$. The control constraints are imposed if there is a technological limitation on the maximum or minimum value of the temperature to be controlled. The scalar $\alpha > 0$ can be interpreted as a control cost, which, as a by-product, leads to more regular solutions of the optimization problem. The function $\rho(x)$

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_1

represents the heat transfer along the boundary. Quadratic objective functionals like $J(y,u)$ are known as *tracking type* costs.

The problem consists in finding an optimal control $u(x)$ and its associated state $y(x)$ such that $J(y,u)$ is minimized. This type of problems arise in several industrial control processes (see, e.g., [30, 45]) and in the design of energy efficient buildings [31].

1.1.2 Optimal Flow Control

Steady laminar incompressible fluid flow in a three-dimensional bounded domain Ω is modeled by the stationary Navier–Stokes equations:

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

where $y = y(x)$ stands for the velocity vector field at the position x, $p = p(x)$ for the pressure and $f = f(x)$ for a body force. The nonlinear term corresponds to the convection of the flow and is given by

$$(y\cdot\nabla)y = \sum_{i=1}^{3} y_i \begin{pmatrix} D_i y_1 \\ D_i y_2 \\ D_i y_3 \end{pmatrix}.$$

The scalar coefficient $Re > 0$ stands for the Reynolds number, a dimensionless quantity related to the apparent viscosity of the fluid. Existence of a solution to the stationary Navier–Stokes equations can, in fact, be argued only if the Reynolds number is sufficiently small so that the viscous term dominates the convective one.

The fluid flow may be controlled either by acting on the boundary (injection or suction) or by using a body force (e.g., gravitational, electromagnetic). If the aim is, for instance, to minimize the vorticity of the fluid by acting on the boundary of the domain, an optimization problem may be formulated in the following way:

$$\min J(y,u) = \frac{1}{2}\int_{\Omega} |\operatorname{curl} y(x)|^2\, dx + \frac{\alpha}{2}\|u\|_U^2$$

subject to:

$$-\frac{1}{Re}\Delta y + (y\cdot\nabla)y + \nabla p = f \qquad \text{in } \Omega,$$

$$\begin{aligned} \operatorname{div} y &= 0 && \text{in } \Omega, \\ y &= u && \text{on } \Gamma, \end{aligned}$$

where U stands for the boundary control space. In order to preserve the incompressibility of the fluid, the additional condition $\int_\Gamma u \cdot \mathbf{n}\, ds = 0$, with $\mathbf{n}$ the outward normal vector to Ω, has to be imposed on the control.

Examples of flow control problems include the design of airfoils [21, 26], the active control of separation [10], drag reduction [24], among others. Problems with control or state constraints have also been studied in the last years [13, 14, 57]. For more details on PDE-constrained optimization in the context of fluid flow, we refer to the monograph [25] and the references therein.

1.1.3 A Least Squares Parameter Estimation Problem in Meteorology

Data assimilation techniques play a crucial role in numerical weather prediction (NWP), making it possible to incorporate measurement information in the mathematical models that describe the behavior of the atmosphere. As a consequence, the quality of the predictions significantly increases and a larger prediction time-window may be obtained.

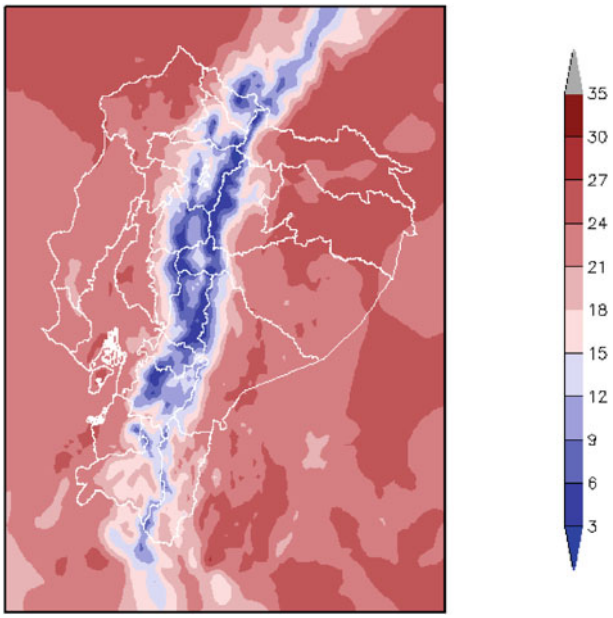

Fig. 1.1 Prediction of the surface temperature distribution in Ecuador

One of the widely used data assimilation methodologies is the so-called 4DVar. This variational approach treats the assimilation problem as a PDE-constrained optimization one, which can be stated in the following way:

$$\begin{aligned}\min J(y,u) = &\frac{1}{2}\sum_{i=1}^{n}[H(y(t_i)) - z_d(t_i)]^T R_i^{-1}[H(y(t_i)) - z_d(t_i)] \\ &+ \frac{1}{2}[u - y^b(t_0)]^T B^{-1}[u - y^b(t_0)]\end{aligned}$$

subject to:

$$\begin{aligned}&y(t) = M(y(t_0)), && \text{(system of PDEs)} \\ &y(t_0) = u, && \text{(initial condition)}\end{aligned}$$

where z_d are the observations obtained at different time steps t_i, y^b is the background vector, H is the observation operator, and R_i and B are the so-called observation and background error covariances, respectively.

The idea of 4DVar consists in solving the PDE-constrained optimization problem in order to obtain an initial condition for the atmospheric dynamics, which is subsequently used for the numerical simulations on a larger time-window. In this context, the use of efficient methods for the solution of the PDE-constrained optimization problem becomes crucial to obtain an operational procedure (see, e.g., [36]).

1.2 A Class of Finite-Dimensional Optimization Problems

We consider next a special class of finite-dimensional optimization problems, where the variable to be optimized has the following separable structure:

$$x = (y,u) \in \mathbb{R}^n,$$

where $u \in \mathbb{R}^l$ is a decision or control variable and $y \in \mathbb{R}^m$ is a state variable determined by solving the (possibly nonlinear) equation

$$e(y,u) = 0, \tag{1.1}$$

with $e : \mathbb{R}^n \to \mathbb{R}^m$.

We consider the optimization problem given by:

$$\begin{cases} \min\limits_{(y,u)\in\mathbb{R}^m\times U_{\text{ad}}} J(y,u) \\ \text{subject to:} \\ \quad e(y,u)=0, \end{cases} \tag{1.2}$$

with J and e twice continuously differentiable and $U_{\text{ad}} \subset \mathbb{R}^l$ a nonempty convex set. Existence of a solution to (1.2) can be obtained under suitable assumptions on J and e.

Let $\bar{x} = (\bar{y}, \bar{u})$ be a local optimal solution to (1.2). We further assume that

$$e_y(\bar{y}, \bar{u}) \text{ is a bijection.} \tag{1.3}$$

From the implicit function theorem (see, e.g., [12, p. 548]) we get the existence of a unique (at least locally) $y(u)$ such that

$$e(y(u), u) = 0 \tag{1.4}$$

in a neighborhood of $\bar{u}$, with the solution mapping $y(u)$ also being twice continuously differentiable.

If for each $u \in U_{\text{ad}}$ there is a unique solution $y(u)$ to (1.4), we may write the optimization problem in reduced form as:

$$\min_{u\in U_{\text{ad}}} f(u) = J(y(u), u). \tag{1.5}$$

Theorem 1.1. *Let $\bar{u}$ be a local optimal solution for (1.5). Then it satisfies the following variational inequality:*

$$f'(\bar{u})(v-\bar{u}) \geq 0, \qquad \textit{for all } v \in U_{\text{ad}}. \tag{1.6}$$

Proof. Let $v \in U_{\text{ad}}$. From the convexity of U_{ad} it follows that

$$tv + (1-t)\bar{u} = \bar{u} + t(v-\bar{u}) \in U_{\text{ad}}, \text{ for all } t \in [0,1].$$

Since $\bar{u}$ is locally optimal,

$$f(\bar{u} + t(v-\bar{u})) - f(\bar{u}) \geq 0, \text{ for } t \text{ sufficiently small.}$$

Dividing by t and passing to the limit we obtain:

$$f'(\bar{u})(v-\bar{u}) = \lim_{t\to 0} \frac{f(\bar{u} + t(v-\bar{u})) - f(\bar{u})}{t} \geq 0. \qquad \square$$

In the special case $U_{\text{ad}} = \mathbb{R}^l$, if $\overline{u} \in \mathbb{R}^l$ is an optimal solution to (1.5), we then obtain the necessary condition:

$$\nabla f(\overline{u})^\top h = \nabla_y J(y(\overline{u}),\overline{u})^\top [y'(\overline{u})h] + \nabla_u J(y(\overline{u}),\overline{u}))^\top h = 0, \tag{1.7}$$

for any $h \in \mathbb{R}^l$.

Definition 1.1. An element $p \in \mathbb{R}^m$ is called the adjoint state related to $\overline{u}$ if it solves the following adjoint equation:

$$e_y(y(\overline{u}),\overline{u})^\top p = \nabla_y J(y(\overline{u}),\overline{u}). \tag{1.8}$$

Theorem 1.2. *Let $(\overline{y},\overline{u})$ be a local optimal solution to (1.2), with $U_{ad} = \mathbb{R}^l$, and assume that (1.3) holds. Then there exists an adjoint state $p \in \mathbb{R}^m$ such that the following optimality system holds:*

$$e(\overline{y},\overline{u}) = 0, \tag{1.9a}$$

$$e_y(\overline{y},\overline{u})^\top p = \nabla_y J(\overline{y},\overline{u}), \tag{1.9b}$$

$$e_u(\overline{y},\overline{u})^\top p = \nabla_u J(\overline{y},\overline{u}). \tag{1.9c}$$

Proof. From the invertibility of $e_y(\overline{y},\overline{u})$ we obtain the existence of an adjoint state $p \in \mathbb{R}^m$ that solves (1.9b).

Since by the implicit function theorem the mapping $y(u)$ is twice continuously differentiable in a neighborhood of $\overline{u}$, we may take derivatives in

$$e(y(\bar{u}),\bar{u}) = 0$$

and obtain that

$$e_y(y(\bar{u}),\bar{u})[y'(\bar{u})h] + e_u(y(\bar{u}),\bar{u})h = 0, \tag{1.10}$$

for any $h \in \mathbb{R}^l$.

To obtain (1.9c) we compute the derivative of the reduced cost function in the following way:

$$\begin{aligned}\nabla f(\overline{u})^\top h &= \nabla_y J(y(\overline{u}),\overline{u})^\top [y'(\overline{u})h] + \nabla_u J(y(\overline{u}),\overline{u})^\top h \\ &= (e_y(\overline{y},\overline{u})^\top p)^\top [y'(\overline{u})h] + \nabla_u J(\overline{y},\overline{u})^\top h \\ &= p^\top (e_y(\overline{y},\overline{u})[y'(\overline{u})h]) + \nabla_u J(\overline{y},\overline{u})^\top h\end{aligned}$$

Thanks to equation (1.10), we then obtain that

$$\begin{aligned}\nabla f(\overline{u})^\top h &= -p^\top (e_u(\overline{y},\overline{u})h) + \nabla_u J(\overline{y},\overline{u})^\top h \\ &= -(e_u(\overline{y},\overline{u})^\top p)^\top h + \nabla_u J(\overline{y},\overline{u})^\top h\end{aligned}$$

which due to (1.7) yields

$$e_u(\bar{y},\bar{u})^\top p = \nabla_u J(\bar{y},\bar{u}). \qquad \square$$

Remark 1.1. From assumption (1.3) and proceeding as in the proof of Theorem 1.2 we get a formula for the gradient of the reduced cost function given by:

$$\nabla f(u) = -e_u(y,u)^\top p + \nabla_u J(y,u), \tag{1.11}$$

where y and p are solutions to

$$e(y,u) = 0 \ \text{ and } \ e_y(y,u)^\top p = \nabla_y J(y,u),$$

respectively.

One frequent choice for U_{ad} is given by so-called box constraints

$$U_{\text{ad}} = \{u \in \mathbb{R}^l : u_a \le u \le u_b\}, \tag{1.12}$$

where $u_a,\ u_b \in \mathbb{R}^l$ satisfy $u_a \le u_b$ componentwise. By rewriting (1.6), using (1.11), we obtain that

$$\left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u}),\bar{u}\right)_{\mathbb{R}^l} \le \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u}),u\right)_{\mathbb{R}^l}, \ \forall u \in U_{\text{ad}},$$

which implies that $\bar{u}$ is solution of

$$\min_{u\in U_{\text{ad}}} \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u}),u\right)_{\mathbb{R}^l} = \min_{u\in U_{\text{ad}}} \sum_{i=1}^{l} \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i u_i.$$

Thanks to the special structure of U_{ad} and the independence of the u_i's, it then follows that

$$\left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i \cdot \bar{u}_i = \min_{u_{a,i}\le u_i \le u_{b,i}} \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i \cdot u_i$$

for $i = 1,\dots,l$. Consequently,

$$\bar{u}_i = \begin{cases} u_{b,i} & \text{if } \ \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i < 0, \\ u_{a,i} & \text{if } \ \left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i > 0. \end{cases} \tag{1.13}$$

For the components where $\left(-e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)_i = 0$, no additional information is obtained.

Let us now define the multipliers:

$$\begin{aligned}\lambda_a &:= \max\left(\mathbf{0}, -e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right),\\ \lambda_b &:= \left|\min\left(\mathbf{0}, -e_u(\bar{y},\bar{u})^\top p + \nabla_u J(\bar{y},\bar{u})\right)\right|,\end{aligned} \tag{1.14}$$

where max, min, and $|\cdot|$ are considered componentwise. Then, from (1.13) it follows that

$$\begin{aligned}&\lambda_a \geq 0, \qquad u_a - \bar{u} \leq 0, \qquad (\lambda_a, u_a - \bar{u})_{\mathbb{R}^l} = 0,\\ &\lambda_b \geq 0, \qquad \bar{u} - u_b \leq 0, \qquad (\lambda_b, \bar{u} - u_b)_{\mathbb{R}^l} = 0,\end{aligned}$$

which is called a *complementarity system*. From (1.14) we then obtain that

$$\lambda_a - \lambda_b = \nabla_u J(\bar{y},\bar{u}) - e_u(\bar{y},\bar{u})^\top p,$$

which, together with the adjoint equation, implies the following theorem.

Theorem 1.3. *Let $(\bar{y},\bar{u})$ be an optimal solution for (1.2), with U_{ad} given by (1.12), and such that (1.3) holds. Then there exist multipliers $p \in \mathbb{R}^m$ and λ_a, $\lambda_b \in \mathbb{R}^l$ such that:*

$$\begin{cases} e(\bar{y},\bar{u}) = 0,\\ e_y(\bar{y},\bar{u})^T p = \nabla_y J(\bar{y},\bar{u}),\\ \nabla_u J(\bar{y},\bar{u}) - e_u(\bar{y},\bar{u})^\top p - \lambda_a + \lambda_b = 0,\\ \lambda_a \geq 0, \quad \lambda_b \geq 0,\\ \lambda_a^\top (u_a - \bar{u}) = \lambda_b^\top (\bar{u} - u_b) = 0,\\ u_a \leq \bar{u} \leq u_b. \end{cases}$$

The last necessary conditions are known as Karush–Kuhn–Tucker (KKT) conditions and constitute one of the cornerstones of nonlinear optimization theory.

Chapter 2
Basic Theory of Partial Differential Equations and Their Discretization

In this chapter we present some basic elements of the analysis of partial differential equations, and of their numerical discretization by finite differences. Our aim is to introduce some notions that enable the reader to follow the material developed in the subsequent chapters. Both the analysis and the numerical solution of partial differential equations (PDEs) are research areas by themselves, with a large amount of related literature. We refer, for instance, to the books [9, 19] for the analysis of PDEs and to, e.g., [23, 52] for their numerical approximation.

2.1 Notation and Lebesgue Spaces

Let X be a Banach space and let $\|\cdot\|_X$ be the associated norm. The topological dual of X is denoted by X' and the duality pair is written as $\langle\cdot,\cdot\rangle_{X',X}$. If X is, in addition, a Hilbert space, we denote by $(\cdot,\cdot)_X$ its inner product.

The set of bounded linear operators from X to Y is denoted by $\mathscr{L}(X,Y)$ or by $\mathscr{L}(X)$ if $X = Y$. The norm of a bounded linear operator $T : X \to Y$ is given by

$$\|T\|_{\mathscr{L}(X,Y)} := \sup_{v\in X, \|v\|_X=1} \|Tv\|_Y .$$

For $T \in \mathscr{L}(X,Y)$ we can also define an operator $T^* \in \mathscr{L}(Y',X')$, called the adjoint operator of T, such that

$$\langle w, Tv\rangle_{Y',Y} = \langle T^* w, v\rangle_{X',X}, \text{ for all } v \in X,\ w \in Y'$$

and $\|T\|_{\mathscr{L}(X,Y)} = \|T^*\|_{\mathscr{L}(Y',X')}$.

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_2

Definition 2.1. Let Ω be an open subset of $\mathbb{R}^N$ and $1 \leq p < \infty$. The set of p-integrable functions is defined by

$$L^p(\Omega) = \{u : \Omega \to \mathbb{R}; u \text{ is measurable and } \int_\Omega |u|^p \, dx < \infty\},$$

and the following norm is used: $\|u\|_{L^p} = (\int_\Omega |u(x)|^p \, dx)^{\frac{1}{p}}$.

Moreover, we also define the space

$$L^\infty(\Omega) = \{u : \Omega \to \mathbb{R}; u \text{ is measurable and } |u(x)| \leq C \text{ a.e. in } \Omega \text{ for some } C > 0\}$$

and endow it with the norm $\|u\|_{L^\infty} = \inf\{C : |u(x)| \leq C \text{ a.e. in } \Omega\}$.

Theorem 2.1 (Hölder). *Let $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then $uv \in L^1(\Omega)$ and*

$$\int_\Omega |uv| \, dx \leq \|u\|_{L^p} \|v\|_{L^q}.$$

The spaces $L^p(\Omega)$ are Banach spaces for $1 \leq p \leq \infty$ and reflexive for $1 < p < \infty$. For $L^2(\Omega)$, a scalar product can be defined by

$$(u, v)_{L^2} = \int_\Omega uv \, dx$$

and a Hilbert space structure is also obtained.

2.2 Weak Derivatives and Sobolev Spaces

Next, we study a weak differentiability notion which is crucial for the definition of Sobolev function spaces and for the variational study of PDEs.

Let $\Omega \subset \mathbb{R}^N, N = 2, 3$, be a bounded Lipschitz domain and consider functions $y, v \in C^1(\overline{\Omega})$. Utilizing Green's formula, we obtain the equivalence

$$\int_\Omega v(x) D_i y(x) \, dx = \int_\Gamma v(x) y(x) n_i(x) \, ds - \int_\Omega y(x) D_i v(x) \, dx,$$

where $n_i(x)$ denotes the i-th component of the exterior normal vector to Ω at the point $x \in \Gamma$ and ds stands for the Lebesgue surface measure at the boundary Γ. If, in addition, $v = 0$ on Γ, then

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

More generally, if higher order derivatives are involved, we obtain the following formula:

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

where $\alpha = (\alpha_1,\ldots,\alpha_N)$ is a multi-index and D^α denotes the differentiation operator with respect to the multi-index, i.e., $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x^{\alpha_1}\cdots\partial x^{\alpha_N}}$, with $|\alpha| = \sum_{i=1}^N \alpha_i$. The last equation is the starting point for the definition of a weaker notion of differentiable function, which takes advantage of the presence of the integral and the accompanying regular function $v(x)$.

Definition 2.2. Let $L^1_{\mathrm{loc}}(\Omega)$ denote the set of locally integrable functions on Ω, i.e., integrable on any compact subset of Ω. Let $y \in L^1_{\mathrm{loc}}(\Omega)$ and α be a given multi-index. If there exists a function $w \in L^1_{\mathrm{loc}}(\Omega)$ such that

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

for all $v \in C_0^\infty(\Omega)$, then w is called the derivative of y in the weak sense (or weak derivative), associated with α, and is denoted by $w = D^\alpha y$.

Example 2.1. $y(x) = |x|$ in $\Omega = (-1,1)$. The weak derivative of $y(x)$ is given by

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

Indeed, for $v \in C_0^\infty(-1,1)$,

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

Note that the value of y' at the point $x = 0$ is not important since the set $\{x = 0\}$ has zero measure.

Definition 2.3. Let $1 \le p < \infty$ and $k \in \mathbb{N}$. The space of functions $y \in L^p(\Omega)$ whose weak derivatives $D^\alpha y$, for $\alpha : |\alpha| \le k$, exist and belong to $L^p(\Omega)$ is denoted by $W^{k,p}(\Omega)$ and

is called Sobolev space. This space is endowed with the norm

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

If $p = \infty$, the space $W^{k,\infty}(\Omega)$ is defined in a similar way, but endowed with the norm

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

The spaces $W^{k,p}(\Omega)$ constitute Banach spaces, reflexive for $1 < p < +\infty$. In the special case $p = 2$ the Sobolev spaces are denoted by $H^k(\Omega) := W^{k,2}(\Omega)$.

A frequently used space is

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

endowed with the norm

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

and the scalar product

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

The space $H^1(\Omega)$ constitutes a Hilbert space with the provided scalar product.

Definition 2.4. The closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. The resulting space is endowed with the $W^{k,p}$-norm and constitutes a closed subspace of $W^{k,p}(\Omega)$.

Next, we summarize some important Sobolev spaces embedding results (see [12, Sect. 6.6] for further details).

Theorem 2.2. *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set with Lipschitz continuous boundary. Then the following continuous embeddings hold:*

1. *If $p < N$, $W^{1,p}(\Omega) \hookrightarrow L^{p^*}(\Omega)$, for $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$,*
2. *If $p = N$, $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$, for $1 \le q < +\infty$,*
3. *If $p > N$, $W^{1,p}(\Omega) \hookrightarrow C^{0,1-N/p}(\overline{\Omega})$.*

Theorem 2.3 (Rellich–Kondrachov). *Let $\Omega \subset \mathbb{R}^N$ be an open bounded set with Lipschitz continuous boundary. Then the following compact embeddings hold:*

1. *If* $p < N$, $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$, *for all* $1 \le q < p^*$ *with* $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$,
2. *If* $p = N$, $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$, *for all* $1 \le q < +\infty$,
3. *If* $p > N$, $W^{1,p}(\Omega) \hookrightarrow C(\overline{\Omega})$.

An important issue in PDEs is the value that the solution function takes at the boundary. If the function is continuous on Ω, then its boundary value can be determined by continuous extension. However, if the function is defined in an almost everywhere sense, then its boundary value has no specific sense, since the boundary has zero measure. The following result clarifies in which sense such a boundary value may hold (see [9, p. 315] for further details).

Theorem 2.4. *Let* Ω *be a bounded Lipschitz domain. There exists a bounded linear operator* $\tau\colon W^{1,p}(\Omega) \longrightarrow L^p(\Gamma)$ *such that*

$$(\tau y)(x) = y(x) \qquad \text{a.e. on } \Gamma,$$

for each $y \in C(\overline{\Omega})$.

Definition 2.5. The function τy is called the trace of y on Γ and τ is called the trace operator.

If Ω is a bounded Lipschitz domain, then it holds that

$$W_0^{1,p}(\Omega) = \left\{ y \in W^{1,p}(\Omega) : \tau y = 0 \text{ a.e. on } \Gamma \right\}.$$

In particular, $H_0^1(\Omega) = \left\{ y \in H^1(\Omega) : \tau y = 0 \text{ a.e. on } \Gamma \right\}$, which, thanks to the Poincaré inequality, can be endowed with the norm

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

2.3 Elliptic Problems

2.3.1 Poisson Equation

Consider the following classical PDE:

$$\begin{cases} -\Delta y = f & \text{in } \Omega, \\ \quad y = 0 & \text{on } \Gamma. \end{cases} \tag{2.1}$$

Existence of a unique solution $y \in C^2(\bar{\Omega})$ can be obtained by classical methods (see [19, Chap. 2]), under the assumption that the right hand side belongs to the space of continuous functions. In practice, however, it usually happens that the function on the right hand side has less regularity. To cope with that situation, an alternative (and weaker) notion of solution may be introduced.

Assuming enough regularity of y and multiplying (2.1) with a test function $v \in C_0^\infty(\Omega)$, we obtain the integral relation

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

which, using integration by parts, yields

$$\int_\Omega \nabla y \cdot \nabla v\, dx - \int_\Gamma v\, \partial_{\mathbf{n}} y\, ds = \int_\Omega f v\, dx,$$

where $\partial_{\mathbf{n}} y = \nabla y \cdot \mathbf{n} = \dfrac{\partial y}{\partial \mathbf{n}}$. Since $v = 0$ on Γ, it follows that

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

Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ and both terms in the previous equation are continuos with respect to the $H_0^1(\Omega)$-norm, then the equation holds for all $v \in H_0^1(\Omega)$.

Definition 2.6. A function $y \in H_0^1(\Omega)$ is called a weak solution for problem (2.1) if it satisfies the following variational formulation:

$$\int_\Omega \nabla y \cdot \nabla v\, dx = \int_\Omega f\, v\, dx, \qquad \forall v \in H_0^1(\Omega). \tag{2.2}$$

Existence of a unique solution to (2.2) can be proved by using the well-known Lax–Milgram theorem, which is stated next.

Theorem 2.5 (Lax–Milgram). *Let V be a Hilbert space and let $a(\cdot,\cdot)$ be a bilinear form such that, for all $y, v \in V$,*

$$|a(y,v)| \le C\|y\|_V \|v\|_V, \tag{2.3}$$
$$a(y,y) \ge \kappa \|y\|_V^2, \tag{2.4}$$

for some positive constants C and κ. Then, for every $\ell \in V'$, there exists a unique solution $y \in V$ to the variational equation

$$a(y,v) = \langle \ell, v \rangle_{V',V}, \text{ for all } v \in V. \tag{2.5}$$

Moreover, there exists a constant $\tilde{c}$, independent of ℓ, such that

$$\|y\|_V \leq \tilde{c}\|\ell\|_{V'}. \tag{2.6}$$

2.3.2 A General Linear Elliptic Problem

We consider the following general linear elliptic problem:

$$\begin{aligned} Ay + c_0 y &= f && \text{in } \Omega, \\ \partial_{n_A} y + \alpha y &= g && \text{on } \Gamma_1, \\ y &= 0 && \text{on } \Gamma_0, \end{aligned} \tag{2.7}$$

where A is an elliptic operator in divergence form:

$$Ay(x) = -\sum_{i,j=1}^{N} D_j(a_{ij}(x)D_i y(x)). \tag{2.8}$$

The coefficients $a_{ij} \in L^\infty(\Omega)$ satisfy the symmetry condition $a_{ij}(x) = a_{ji}(x)$ and the following ellipticity condition: $\exists \kappa > 0$ such that

$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \geq \kappa|\xi|^2, \qquad \forall \xi \in \mathbb{R}^n, \text{ for a.a. } x \in \Omega. \tag{2.9}$$

The operator ∂_{n_A} stands for the conormal derivative, i.e.,

$$\partial_{n_A} y(x) = \nabla y(x)^T n_A(x),$$

with $(n_A)_i(x) = \sum_{j=1}^{N} a_{ij}(x)n_j(x)$. Additionally $\Gamma = \Gamma_0 \uplus \Gamma_1$ and $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_1)$, $f \in L^2(\Omega)$, $g \in L^2(\Gamma_1)$.

By introducing the Hilbert space

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

and the bilinear form

$$a(y,v) := \int_\Omega \sum_{i,j=1}^{N} a_{ij}\, D_i y\, D_j v \, dx + \int_\Omega c_0 yv \, dx + \int_{\Gamma_1} \alpha yv \, ds, \tag{2.10}$$

the variational formulation of problem (2.7) is given in the following form: Find $y \in V$ such that

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

Theorem 2.6. *Let Ω be a bounded Lipschitz domain and $c_0 \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_1)$ given functions such that $c_0(x) \geq 0$ a.e. in Ω and $\alpha(x) \geq 0$ a.e. on Γ_1, respectively. If one of the following conditions holds:*

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

then there exist a unique weak solution $y \in V$ to problem (2.7). Additionally, there exists a constant $c_A > 0$ such that

$$\|y\|_{H^1} \leq c_A \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_1)} \right).$$

Proof. The proof makes use of the Lax–Milgram theorem and Friedrichs' inequality, and is left as an exercise for the reader. □

2.3.3 Nonlinear Equations of Monotone Type

An important class of nonlinear PDEs involve differential operators of monotone type. Such is the case, for instance, of equations that arise as necessary conditions in the minimization of energy functionals.

Let V be a separable, reflexive Banach space and consider the variational equation

$$\langle A(y), v \rangle_{V',V} = \langle \ell, v \rangle_{V',V}, \qquad \text{for all } v \in V, \tag{2.11}$$

where $\ell \in V'$ and the operator $A : V \to V'$ satisfies the following properties.

Assumption 2.1.

i) A is monotone*, i.e., for all $u, v \in V$,*

$$\langle A(u) - A(v), u - v \rangle_{V',V} \geq 0. \tag{2.12}$$

ii) A is hemicontinuous*, i.e., the function*

$$t \to \langle A(u + tv), w \rangle_{V',V}$$

is continuous on the interval $[0,1]$*, for all* $u,v,w \in V$.

iii) A is coercive*, i.e.,*

$$\lim_{\|u\|_V \to \infty} \frac{\langle A(u),u\rangle_{V',V}}{\|u\|_V} = +\infty. \tag{2.13}$$

Theorem 2.7 (Minty–Browder). *Let* $\ell \in V'$ *and* $A : V \to V'$ *be an operator satisfying Assumption 2.1. Then there exists a solution to the variational equation* (2.11)*. If A is strictly monotone, then the solution is unique.*

Proof. Since V is separable, there exists a basis $\{v_i\}_{i=1}^{\infty}$ of linearly independent vectors, dense in V. Introducing

$$V_n = \text{span}\{v_1,\ldots,v_n\},$$

we consider a solution $y_n \in V_n$ of the equation

$$\langle A(y_n),v_j\rangle_{V',V} = \langle \ell, v_j\rangle_{V',V}, \text{ for } j = 1\ldots,n. \tag{2.14}$$

By using the expression $y_n = \sum_{i=1}^n c_i v_i$, problem (2.14) can be formulated as a system of nonlinear equations in $\mathbb{R}^n$.

Thanks to the properties of A, we may use Brouwer's fixed point theorem (see, e.g., [12, p. 723]) and get existence of a solution to (2.14), with the additional bound:

$$\|y_n\|_V \le C,$$

with $C > 0$ a constant independent of n.

From Assumption 2.1 it follows that A is locally bounded [12, p. 740], which implies that there exist constants $r > 0$ and $\rho > 0$ such that

$$\|v\|_V \le r \qquad \Rightarrow \qquad \|A(v)\|_V \le \rho.$$

Consequently, it follows that

$$\begin{aligned}\langle A(y_n),v\rangle_{V',V} &\le \langle A(y_n),y_n\rangle_{V',V} - \langle A(v),y_n\rangle_{V',V} + \langle A(v),v\rangle_{V',V}\\ &= \langle \ell,y_n\rangle_{V',V} - \langle A(v),y_n\rangle_{V',V} + \langle A(v),v\rangle_{V',V}\\ &\le \|\ell\|_{V'}C + \rho C + \rho r,\end{aligned}$$

for all $n \ge 1$ and all $\|v\|_V \le r$, and, therefore, the sequence $\{A(y_n)\}$ is bounded in V'.

Thanks to the reflexivity of the spaces and the boundedness of the sequences, there exists a subsequence $\{y_m\}_{m\in\mathbb{N}}$ and limit points $y \in V$ and $g \in V'$ such that

$$y_m \rightharpoonup y \text{ weakly in } V \qquad \text{and} \qquad A(y_m) \rightharpoonup g \text{ weakly in } V'.$$

For any $k \geq 1$, we know that

$$\langle A(y_m), v_k \rangle_{V',V} = \langle \ell, v_k \rangle_{V',V}, \text{ for all } m \geq k.$$

Consequently,

$$\langle g, v_k \rangle_{V',V} = \lim_{m \to \infty} \langle A(y_m), v_k \rangle_{V',V} = \langle \ell, v_k \rangle_{V',V}$$

and, since the latter holds for all $k \geq 1$, we get that

$$\langle g, v \rangle_{V',V} = \langle \ell, v \rangle_{V',V}, \text{ for all } v \in V.$$

Therefore, $g = \ell$ in V'. Additionally,

$$\langle A(y_m), y_m \rangle_{V',V} = \langle \ell, y_m \rangle_{V',V} \to \langle \ell, y \rangle_{V',V}, \qquad \text{as } m \to \infty.$$

From the monotonicity of A,

$$\langle A(y_m) - A(v), y_m - v \rangle_{V',V} \geq 0, \forall v \in V.$$

By passing to the limit, we then get that

$$\langle l - A(v), y - v \rangle_{V',V} \geq 0, \forall v \in V.$$

Taking $v = y - tw$, with $t > 0$ and $w \in V$, it then follows that

$$\langle l - A(y - tw), w \rangle_{V',V} \geq 0, \forall w \in V.$$

Thanks to the hemicontinuity of A and taking the limit as $t \to 0$, we finally get that

$$A(y) = \ell \text{ in } V'. \qquad \square$$

Example 2.2 (A semilinear equation). Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain, $u \in L^2(\Omega)$ and consider the following nonlinear boundary value problem:

$$-\Delta y + y^3 = u \qquad \text{in } \Omega, \tag{2.15a}$$

$$y = 0 \qquad \text{on } \Gamma. \tag{2.15b}$$

Weak formulation of the PDE. Multiplying the state equation by a test function $v \in C_0^\infty(\Omega)$ and integrating yields

$$\int_\Omega -\Delta y \, v \, dx + \int_\Omega y^3 v \, dx = \int_\Omega uv \, dx.$$

Using integration by parts,

$$\int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx = \int_\Omega uv \, dx.$$

Since C_0^∞ is dense in $H_0^1(\Omega)$ and all terms are continuous with respect to v in the $H_0^1(\Omega)$ norm, we obtain the following variational formulation: Find $y \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx = \int_\Omega uv \, dx, \quad \forall v \in H_0^1(\Omega).$$

Indeed, thanks to the embedding $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$, for all $1 \leq p < +\infty$, it follows that $y^3 \in L^2(\Omega)$ and the second integral is well-defined.

Let us now define the operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$\langle A(y), v \rangle_{H^{-1}, H_0^1} := \int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx, \text{ for all } v \in H_0^1(\Omega).$$

Monotonicity. Let $v, w \in H_0^1(\Omega)$,

$$\begin{aligned} \langle A(v) - A(w), v - w \rangle_{H^{-1}, H_0^1} &= \int_\Omega |\nabla(v-w)|^2 \, dx + \int_\Omega v^3(v-w) - w^3(v-w) \, dx \\ &= \|v - w\|_{H_0^1}^2 + \int_\Omega (v-w)^2(v^2 + vw + w^2) \, dx. \end{aligned}$$

Since $v^2 + vw + w^2 \geq 0$ a.e. in Ω, it follows that

$$\langle A(v) - A(w), v - w \rangle_{H^{-1}, H_0^1} \geq \|v - w\|_{H_0^1}^2,$$

which implies the strict monotonicity of A.

Coercivity. Let $v \in H_0^1(\Omega)$,

$$\langle A(v), v \rangle_{H^{-1}, H_0^1} = \int_\Omega |\nabla v|^2 \, dx + \int_\Omega v^4 \, dx \geq \|v\|_{H_0^1}^2,$$

which implies that

$$\frac{\langle A(v), v \rangle_{H^{-1}, H_0^1}}{\|v\|_{H_0^1}} \to +\infty \text{ as } \|v\|_{H_0^1} \to +\infty.$$

Hemicontinuity.

$$\begin{aligned}\langle A(u+tv),w\rangle_{H^{-1},H_0^1} &= \int_\Omega \nabla(u+tv)\,\nabla w\,dx+\int_\Omega (u+tv)^3 w\,dx\\ &= \int_\Omega \nabla u\,\nabla w+u^3 w\,dx+t\int_\Omega \nabla v\,\nabla w+3u^2 vw\,dx\\ &\qquad +t^2\int_\Omega 3uv^2 w\,dx+t^3\int_\Omega v^3 w\,dx,\end{aligned}$$

which is continuous with respect to t.

Hence, all conditions of the Minty–Browder theorem are satisfied and there exists a unique solution $y\in H_0^1(\Omega)$ to the semilinear equation (2.15).

2.4 Discretization by Finite Differences

The basic idea of a finite difference discretization scheme consists in replacing the differential operators involved in the PDE with corresponding difference quotients involving the solution at different spatial points. This procedure leads to a system of equations in $\mathbb{R}^n$, that can be solved with different numerical techniques.

For simplicity, let us start with the one-dimensional case and consider the interval domain $\Omega=(0,1)$. The Poisson problem then becomes a boundary value ordinary differential equation (ODE) given by

$$-y''=f \text{ in } \Omega, \tag{2.16a}$$

$$y(0)=y(1)=0. \tag{2.16b}$$

Using a uniform spatial mesh, the discretization points are $x_j=jh,\ j=0,\dots n$, where $n\geq 2$ is an integer and $h=1/n$ is the mesh size step. The first derivative of the solution y at the inner discretization points $x_j=hj,\ j=1,\dots,n-1$ can then be approximated either by

forward differences:	$\dfrac{y_{j+1}-y_j}{h}$,
backward differences:	$\dfrac{y_j-y_{j-1}}{h}$,
or centered differences:	$\dfrac{y_{j+1}-y_{j-1}}{2h}$,

where $y_j := y(x_j),\ j = 0,\ldots,n$. For the second derivative, by applying subsequently forward and backward differences, the quotient

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2}$$

is obtained. The approximate solution to the boundary value problem (2.16), at the discretization points, then satisfies the following system of equations:

$$-\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = f(x_j), \quad j = 1,\ldots,n-1, \tag{2.17}$$

with $y_0 = y_n = 0$.

By defining the vectors $\mathbf{y} = (y_1,\ldots,y_{n-1})^T$ and $\mathbf{f} = (f_1,\ldots,f_{n-1})^T$, with $f_i := f(x_i)$, Eq. (2.17) can be written in the following matrix form:

$$A_h \mathbf{y} = \mathbf{f}, \tag{2.18}$$

where $A_h \in \mathscr{M}_{(n-1)}$ stands for the finite difference discretization matrix given by

$$A_h = h^{-2}\,\text{tridiag}_{n-1}(-1,2,-1) = h^{-2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}. \tag{2.19}$$

The matrix A_h is symmetric and positive definite. Indeed,

$$w^T A_h w = h^{-2}\left[w_1^2 + w_{n-1}^2 + \sum_{i=2}^{n-1}(w_i - w_{i-1})^2\right].$$

This implies, in particular, that (2.18) has a unique solution.

Consider now the two-dimensional bounded domain $\Omega = (0,1)^2 \subset \mathbb{R}^2$. Our aim is to find a solution to the Poisson problem:

$$\begin{cases} -\Delta y = f \text{ in } \Omega, \\ y = g \quad\ \text{on } \Gamma, \end{cases} \tag{2.20}$$

by using finite differences. Choosing the mesh size steps $h = \frac{1}{n}$ and $k = \frac{1}{m}$, with $n,\ m \in \mathbb{N}$, for the horizontal and vertical components, respectively, we get the mesh:

$$\overline{\Omega}_{hk} = \left\{(x_1^j, x_2^j) \colon x_1^i = ih,\ x_2^j = jh,\ i = 0,\ldots,n,\ j = 0,\ldots,m\right\}.$$

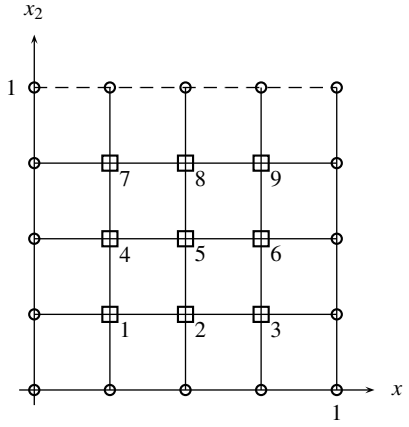

Fig. 2.1 Example of a grid with $m = n = 4$

Similarly to the one-dimensional case, but considering both spatial components, Eq. (2.20) can then be approximated in the following way:

$$\frac{1}{h^2}(-y_{i+1,j} + 2y_{i,j} - y_{i-1,j}) + \frac{1}{k^2}(-y_{i,j+1} + 2y_{i,j} - y_{i,j-1}) = f_{ij}, \tag{2.21}$$

for $i = 1, \ldots, n-1$; $j = 1, \ldots, m-1$. Utilizing a horizontal–vertical lexicographic order and taking $n = m$, the following block matrix is obtained:

$$A_h = h^{-2} \begin{pmatrix} B & -I & & \\ -I & B & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{pmatrix},$$

where $I \in \mathscr{M}_{(n-1)}$ stands for the identity matrix and $B \in \mathscr{M}_{(n-1)}$ is given by

$$B = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}.$$

The resulting discretization matrix A_h is also symmetric and positive definite in the multidimensional case. Moreover, it can be verified that such finite difference scheme is

consistent of order 2 with respect to h, with an approximation error of order 2 as well (see, e.g., [23, Chap. 2]).

Program: Finite Difference Matrix for the Laplace Operator

```
function [lap]=matrices(n,h)
d(n:n:n^2)=1;
d=d';
e=sparse(n^2,1);
e(1:n:(n-1)*n+1)=1;
b=ones(n^2,1);
a=[b,b-d,-4*b,b-e,b];
lap=-1/(h^2)*spdiags(a,[-n,-1,0,1,n],n^2,n^2);
```

Chapter 3
Theory of PDE-Constrained Optimization

3.1 Problem Statement and Existence of Solutions

Consider the following general optimization problem:

$$\begin{cases} \min\ J(y,u), \\ \text{subject to:} \\ \qquad e(y,u) = 0, \end{cases} \tag{3.1}$$

where $J\colon Y \times U \longrightarrow \mathbb{R}$, $e\colon Y \times U \longrightarrow W$, and Y, U, W are reflexive Banach spaces. We assume that there exists a unique solution $y(u)$ to $e(y,u) = 0$ and refer to the operator

$$\begin{aligned} G\colon U &\longrightarrow Y \\ u &\longmapsto y(u) = G(u), \end{aligned}$$

which assigns to each $u \in U$ the solution $y(u)$ to

$$e(y(u),u) = 0 \tag{3.2}$$

as *solution* or *control-to-state* operator.

Using this operator, we can write the optimization problem in reduced form as

$$\min_{u\in U} f(u) := J(y(u),u). \tag{3.3}$$

Hereafter we assume that $f\colon U \longrightarrow \mathbb{R}$ is bounded from below.

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_3

Definition 3.1. An element $\bar{u} \in U$ is called a global solution to (3.3) if $f(\bar{u}) \leq f(u)$, $\forall u \in U$. Further, $\bar{u}$ is called a local solution if there exists a neighborhood $V(\bar{u})$ of $\bar{u}$ in U such that

$$f(\bar{u}) \leq f(u), \qquad \forall u \in V(\bar{u}).$$

Definition 3.2. A functional $h\colon U \longrightarrow \mathbb{R}$ is called weakly lower semicontinuous (w.l.s.c) if for every weakly convergent sequence $u_n \rightharpoonup u$ in U it follows that

$$h(u) \leq \liminf_{n\to\infty} h(u_n).$$

Remark 3.1. If h is quasiconvex and continuous, then it is w.l.s.c (see [35, p. 15]). In addition, every convex functional is also quasiconvex.

Theorem 3.1. *If $f\colon U \longrightarrow \mathbb{R}$ is w.l.s.c and radially unbounded, i.e.,*

$$\lim_{\|u\|_U\to\infty} f(u) = +\infty, \tag{3.4}$$

then f has a global minimum.

Proof. Let $\{u_n\}_{n\in\mathbb{N}}$ be a minimizing sequence, i.e., $\{u_n\} \subset U$ and

$$\lim_{n\to\infty} f(u_n) = \inf_{u\in U} f(u).$$

Thanks to (3.4) it follows that the sequence $\{u_n\}$ is bounded. Since U is reflexive, there exists a subsequence $\{u_{n_k}\}_{k\in\mathbb{N}}$ of $\{u_n\}$ which converges weakly to a limit $\bar{u}$ as $k \to \infty$. Due to the weakly lower semi continuity of f it follows that

$$f(\bar{u}) \leq \liminf_{k\to\infty} f(u_{n_k}) = \inf_{u\in U} f(u).$$

Consequently, $\bar{u}$ is a global minimum. □

Example 3.1 (A linear–quadratic problem). Consider the following optimal heating problem:

$$\min J(y,u) = \frac{1}{2}\|y - z_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \tag{3.5a}$$

subject to:

$$-\Delta y = \beta u \qquad \text{in } \Omega, \tag{3.5b}$$

$$y = 0 \qquad \text{on } \Gamma, \tag{3.5c}$$

where $\Omega \subset \mathbb{R}^N$, $N = 2,3$, is a bounded Lipschitz domain, $\alpha > 0$, $z_d \in L^2(\Omega)$ and $\beta \in L^\infty(\Omega)$.
As control space we consider $U = L^2(\Omega)$ and, thanks to Theorem 2.4, there exists, for each $u \in U$, a unique weak solution for (3.5b)–(3.5c). The reduced functional $f\colon U \longrightarrow \mathbb{R}$ satisfies

$$f(u) = J(y(u),u) \geq \frac{\alpha}{2}\|u\|_{L^2}^2$$

and, consequently, is bounded from below and fulfills (3.4). Moreover f is convex and continuous, and, therefore, w.l.s.c. Consequently, there exists an optimal solution for (3.5).

3.2 First Order Necessary Conditions

3.2.1 Differentiability in Banach Spaces

Let U, V be two real Banach spaces and $F\colon U \longrightarrow V$ a mapping from U to V.

Definition 3.3. If, for given elements $u,h \in U$, the limit

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

exists, then $\delta F(u)(h)$ is called the *directional derivative* of F at u in direction h. If this limit exists for all $h \in U$, then F is called directionally differentiable at u.

Definition 3.4. If for some $u \in U$ and all $h \in U$ the limit

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

exists and $\delta F(u)$ is a continuous linear mapping from U to V, then $\delta F(u)$ is denoted by $F'(u)$ and is called the Gâteaux derivative of F at u, and F is called *Gâteaux differentiable* at u.

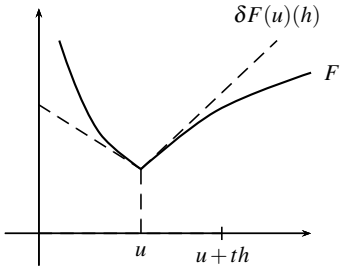

Fig. 3.1 Illustration of a directionally differentiable function

Example 3.2.

a) Let $U = C[0,1]$ and $f\colon U \longrightarrow \mathbb{R}$ be given through

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

Let also $h = h(x)$ be a function in $C[0,1]$. The directional derivative of f at u in direction h is then given by

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

Therefore, $\delta f(u)(h) = -\sin(u(1))h(1)$ and since $\delta f(u)$ is linear and continuous with respect to h, f is Gâteaux differentiable with its derivative given by

$$f'(u)h = -\sin(u(1))h(1).$$

b) Let H be a Hilbert space with scalar product $(\cdot,\cdot)_H$ and norm $\|\cdot\|_H$. Let $f\colon H \longrightarrow \mathbb{R}$ be defined by

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

The directional derivative of f at u in direction h is given by

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

Therefore $\delta f(u)(h) = 2(u,h)_H$, which is linear and continuous with respect to h. Consequently, f is Gâteaux differentiable.

Definition 3.5. If F is Gâteaux differentiable at $u \in U$ and satisfies in addition that

$$\lim_{\|h\|_U\to 0}\frac{\|F(u+h)-F(u)-F'(u)h\|_V}{\|h\|_U} = 0,$$

then $F'(u)$ is called the Fréchet derivative of F at u and F is called *Fréchet differentiable*.

Properties.

1. If $F\colon U \longrightarrow V$ is Gâteaux differentiable and $F'\colon U \longrightarrow \mathscr{L}(U,V)$ is also Gâteaux differentiable, then F is called twice Gâteaux differentiable and we write

   $$F''(u) \in \mathscr{L}(U,\mathscr{L}(U,V))$$

   for the second derivative of F at u.
2. If F is Fréchet differentiable at $u \in U$, then it is also Gâteaux differentiable at u.
3. If F is Fréchet differentiable at $u \in U$, then it is continuous at u.
4. *Chain rule*: Let $F\colon U \longrightarrow V$ and $G\colon V \longrightarrow Z$ be Fréchet differentiable at u and $F(u)$, respectively. Then

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

   is also Fréchet differentiable and its derivative is given by:

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

Let $C \subset U$ be a nonempty subset of a real normed space U and $f\colon C \subset U \longrightarrow \mathbb{R}$ a given functional, bounded from below. Consider the following problem:

$$\min_{u\in C} f(u). \tag{3.6}$$

Definition 3.6. For $u \in C$ the direction $v-u \in U$ is called admissible if there exists a sequence $\{t_n\}_{n\in\mathbb{N}}$, with $0 < t_n \to 0$ as $n \to \infty$, such that $u+t_n(v-u) \in C$ for every $n \in \mathbb{N}$.

Theorem 3.2. *Suppose that $\bar{u} \in C$ is a local minimum of (3.6) and that $v - \bar{u}$ is an admissible direction. If f is directionally differentiable at $\bar{u}$, in direction $v - \bar{u}$, then*

$$\delta f(\bar{u})(v - \bar{u}) \geq 0.$$

Proof. Since $\bar{u} \in C$ is a local minimum and $v - \bar{u}$ is feasible, for n sufficiently large we get that $\bar{u} + t_n(v - \bar{u}) \in V(\bar{u}) \cap C$ and

$$f(\bar{u}) \leq f(\bar{u} + t_n(v - \bar{u})),$$

which implies that

$$\frac{1}{t_n}(f(\bar{u} + t_n(v - \bar{u})) - f(\bar{u})) \geq 0.$$

By taking the limit as $n \to \infty$ on both sides,

$$\delta f(\bar{u})(v - \bar{u}) \geq 0. \qquad \square$$

Corollary 3.1. *Let $C = U$ and $\bar{u}$ be a local optimal solution for (3.6). If f is Gâteaux differentiable at $\bar{u}$, then*

$$f'(\bar{u})h = 0, \qquad \text{for all } h \in U.$$

Proof. Let $h \in U$ be arbitrary but fix. From Theorem 2.8 it follows, for $v = h + \bar{u}$, that

$$f'(\bar{u})h \geq 0$$

and, for $v = -h + \bar{u}$, that

$$f'(\bar{u})(-h) \geq 0,$$

which implies the result. $\square$

3.2.2 Optimality Condition

Let us now turn to PDE-constrained optimization problems and recall problem (3.1):

$$\begin{cases} \min J(y, u), \\ \text{subject to:} \\ \qquad e(y, u) = 0. \end{cases}$$

We assume that $J\colon Y \times U \longrightarrow \mathbb{R}$ and $e\colon Y \times U \longrightarrow W$ are continuously Fréchet differentiable. Further, we assume that the partial derivative of e with respect to y at

$(\bar{y},\bar{u})$ satisfies the following condition:

$$e_y(\bar{y},\bar{u}) \in \mathscr{L}(Y,W) \text{ is a bijection.} \tag{3.7}$$

From (3.7), existence of a (locally) unique solution $y(u)$ to the state equation $e(y,u)=0$, in a neighborhood of $(\bar{y},\bar{u})$, follows from the implicit function theorem (see, e.g., [12, p. 548]) and, moreover, the solution operator is also continuously Fréchet differentiable.

By taking the derivative, with respect to u, on both sides of the state equation

$$e(y(\bar{u}),\bar{u}) = 0,$$

we obtain

$$e_y(y(\bar{u}),\bar{u})\,y'(\bar{u})h + e_u(y(\bar{u}),\bar{u})\,h = 0, \tag{3.8}$$

where $y'(u)h$ denotes the derivative of the solution operator at u in direction h.

If $\bar{u}\in U$ is a local optimal solution to (3.3), we obtain from Corollary 3.1 the following necessary condition

$$f'(\bar{u})h = \underbrace{J_y(y(\bar{u}),\bar{u})\,y'(\bar{u})h}_{=\langle J_y(y(\bar{u}),\bar{u}),y'(\bar{u})h\rangle_{Y',Y}} + J_u(y(\bar{u}),\bar{u}))\,h = 0, \tag{3.9}$$

for all $h\in U$.

In order to make (3.9) more explicit we introduce the following definition.

Definition 3.7. An element $p\in W'$ is called the *adjoint state* related to $\bar{u}$ if it solves the following *adjoint equation*:

$$e_y(y(\bar{u}),\bar{u})^*\,p = J_y(y(\bar{u}),\bar{u}), \tag{3.10}$$

where $e_y(y(\bar{u}),\bar{u})^*$ denotes the adjoint operator of $e_y(y(\bar{u}),\bar{u})$.

Theorem 3.3. *Let $\bar{u}$ be a local optimal solution to (3.3) and $y(\bar{u})$ its associated state. If (3.7) holds, then there exists an adjoint state $p\in W'$ such that the following system of equations is satisfied:*

$$e(y(\bar{u}),\bar{u}) = 0, \tag{3.11a}$$
$$e_y(y(\bar{u}),\bar{u})^*\,p = J_y(y(\bar{u}),\bar{u}), \tag{3.11b}$$
$$e_u(y(\bar{u}),\bar{u})^*\,p = J_u(y(\bar{u}),\bar{u}). \tag{3.11c}$$

System (3.11) is called the optimality system for $\bar{u}$.

Proof. From (3.7) the surjectivity of $e_y(y(\bar{u}),\bar{u})^*$ follows (see, e.g., [9, p. 47]) and, therefore, there exists an adjoint state $p \in W'$ which solves (3.11b).

To obtain the result, we compute the derivative of the reduced cost functional as follows:

$$f'(\bar{u})h = \langle J_y(y(\bar{u}),\bar{u}), y'(\bar{u})h\rangle_{Y',Y} + J_u(y(\bar{u}),\bar{u})h.$$

Using the adjoint equation we get that

$$\begin{aligned} f'(\bar{u})h &= \langle e_y(y(\bar{u}),\bar{u})^* p, y'(\bar{u})h\rangle_{Y',Y} + J_u(y(\bar{u}),\bar{u})h \\ &= \langle p, e_y(y(\bar{u}),\bar{u})y'(\bar{u})h\rangle_{W',W} + J_u(y(\bar{u}),\bar{u})h. \end{aligned}$$

Finally, thanks to (3.8) and using the transpose of $e_u(\bar{y},\bar{u})$ we obtain

$$\begin{aligned} f'(\bar{u})h &= \langle p, -e_u(y(\bar{u}),\bar{u})h\rangle_{W',W} + J_u(y(\bar{u}),\bar{u})h \\ &= -\langle e_u(y(\bar{u}),\bar{u})^* p, h\rangle_{U',U} + J_u(y(\bar{u}),\bar{u})h. \end{aligned} \tag{3.12}$$

Consequently, from (3.9) it follows that

$$e_u(y(\bar{u}),\bar{u})^* p = J_u(y(\bar{u}),\bar{u}) \text{ in } U'. \qquad \square$$

Example 3.3. Consider again the heating problem given by

$$\begin{cases} \min J(y,u) = \frac{1}{2}\|y - z_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2, \\ \text{subject to:} \\ \qquad -\Delta y = \beta u \qquad \text{in } \Omega, \\ \qquad\quad y = 0 \qquad\quad \text{on } \Gamma. \end{cases}$$

The variational formulation of the state equation is given by: Find $y \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla v\, dx = \int_\Omega \beta u v\, dx, \qquad \forall v \in H_0^1(\Omega).$$

Consequently, $e\colon H_0^1(\Omega) \times L^2(\Omega) \longrightarrow H^{-1}(\Omega)$ is defined by

$$\langle e(y,u), v\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla y \cdot \nabla v\, dx - \int_\Omega \beta u v\, dx$$

and its partial derivative with respect to y is given by

$$\langle e_y(y,u)w, v\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla w \cdot \nabla v\, dx.$$

For a given function $\varphi \in H^{-1}(\Omega)$, equation

$$\langle e_y(y,u)w, v\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla w \cdot \nabla v \, dx = \langle \varphi, v\rangle_{H^{-1},H_0^1}, \forall v \in H_0^1(\Omega),$$

has a unique solution $w \in H_0^1(\Omega)$ and $\|w\|_{H_0^1} \leq C\|\varphi\|_{H^{-1}}$ for some constant $C > 0$ (*Lax–Milgram Theorem*). Consequently, (3.7) is satisfied.
In order to apply Theorem 3.3, we compute the remaining derivatives:

$$\begin{aligned} e_u(y,u)h &= -\beta h, \\ J_y(y,u) &= y - z_d, \\ J_u(y,u) &= \alpha u. \end{aligned}$$

The optimality system is then given through the following equations:

$$\begin{aligned} \int_\Omega \nabla y \cdot \nabla v \, dx &= \int_\Omega \beta u v \, dx, \quad \forall v \in H_0^1(\Omega), \\ \int_\Omega \nabla p \cdot \nabla v \, dx &= \int_\Omega (y - z_d) v \, dx, \quad \forall v \in H_0^1(\Omega), \\ -\beta p &= \alpha u, \qquad \text{a.e. in } \Omega, \end{aligned}$$

where we used that

$$\left\langle e_y(y(\bar{u}))w, v\right\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla w \cdot \nabla v \, dx = \int_\Omega \nabla v \cdot \nabla w \, dx = \left\langle w, e_y(\bar{y},\bar{u})^* v\right\rangle_{H_0^1,H^{-1}}$$

and, similarly,

$$(e_u(\bar{y},\bar{u})h, \phi)_{L^2} = \int_\Omega -\beta h \phi \, dx = \int_\Omega -h\beta\phi \, dx = (h, e_u(\bar{y},\bar{u})^*\phi)_{L^2}.$$

3.3 Lagrangian Approach

It is also possible to derive the optimality system (3.11) by using the Lagrangian approach. With such a procedure, a direct hint on what the adjoint equation looks like is obtained.

Consider again problem (3.3) with $J\colon Y \times U \longrightarrow \mathbb{R}$ and $e\colon Y \times U \longrightarrow W$. The Lagrangian functional associated to (3.3) is given by

$$\begin{aligned} \mathscr{L}\colon Y \times U \times W' &\longrightarrow \mathbb{R} \\ (y,u,p) &\longmapsto \mathscr{L}(y,u,p) = J(y,u) - \langle p, e(y,u)\rangle_{W',W}. \end{aligned}$$

By differentiating $\mathscr{L}(y,u,p)$ with respect to y, in direction w, we obtain that

$$\begin{aligned} \mathscr{L}_u(y,u,p)w &= J_y(y,u)w - \langle p, e_y(y,u)w\rangle_{W',W} \\ &= J_y(y,u)w - \langle e_y(y,u)^* p, w\rangle_{Y',Y}. \end{aligned}$$

Consequently, Eq. (3.11b) can also be expressed as

$$\mathscr{L}_y(\bar{y},\bar{u},p) = 0.$$

In a similar manner, by taking the derivative of $\mathscr{L}(y,u,p)$ with respect to u, in direction h, we obtain

$$\begin{aligned} \mathscr{L}_u(y,u,p)h &= J_u(y,u)h - \langle p, e_u(y,u)h\rangle_{W',W} \\ &= J_u(y,u)h - \langle e_u(y,u)^* p, h\rangle_{U',U} \end{aligned}$$

and, therefore, Eq. (3.11c) can be written as

$$\mathscr{L}_u(\bar{y},\bar{u},p) = 0.$$

Summarizing, the optimality system (3.11) can be written in the following way:

$$e(\bar{y},\bar{u}) = 0, \tag{3.13a}$$

$$\mathscr{L}_y(\bar{y},\bar{u},p) = 0, \tag{3.13b}$$

$$\mathscr{L}_u(\bar{y},\bar{u},p) = 0. \tag{3.13c}$$

Example 3.4 (The heating problem revisited).

$$\begin{cases} \min J(y,u) = \dfrac{1}{2}\|y - z_d\|_{L^2}^2 + \dfrac{\alpha}{2}\|u\|_{L^2}^2, \\ \text{subject to:} \\ \qquad -\Delta y = \beta u \qquad \text{in } \Omega, \\ \qquad\quad y = 0 \qquad\quad \text{on } \Gamma. \end{cases}$$

In this case $Y = H_0^1(\Omega)$, $U = L^2(\Omega)$, $W = H^{-1}(\Omega)$ and

$$\langle e(y,u),v\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla y \cdot \nabla v \, dx - \int_\Omega \beta uv \, dv = 0, \qquad \forall v \in H_0^1(\Omega).$$

The Lagrangian is then defined by

$$\begin{aligned}\mathscr{L}(y,u,p) &= J(y,u) - \langle p, e(y,u)\rangle_{W',W} \\ &= \frac{1}{2}\|y - z_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 - \int_\Omega \nabla y \cdot \nabla p \, dx + \int_\Omega \beta up \, dx.\end{aligned}$$

Next, we obtain the derivatives of $\mathscr{L}$ with respect to y and u and set them equal to zero. First, for the partial derivative with respect to y, we obtain

$$\begin{aligned}\mathscr{L}_y(y,u,p)w &= (y - z_d, w)_{L^2} - \int_\Omega \nabla w \cdot \nabla p \, dx \\ &= \int_\Omega (y - z_d) w dx - \int_\Omega \nabla p \cdot \nabla w \, dx = 0,\end{aligned}$$

which implies that

$$\int_\Omega \nabla p \cdot \nabla w \, dx = \int_\Omega (y - z_d) w \, dx, \qquad \forall w \in H_0^1(\Omega).$$

For the partial derivative with respect to u,

$$\begin{aligned}\mathscr{L}_u(y,u,p)h &= \alpha(u,h)_{L^2} + \int_\Omega \beta hp \, dx \\ &= \int_\Omega \alpha uh \, dx + \int_\Omega \beta ph \, dx = 0,\end{aligned}$$

which implies that,

$$\int_\Omega \alpha uh \, dx = -\int_\Omega \beta ph \, dx, \qquad \forall h \in L^2(\Omega)$$

and, therefore,

$$\alpha u = -\beta p \qquad \text{a.e. in } \Omega.$$

Altogether, we obtain the following optimality system:

$$\int_\Omega \nabla y \cdot \nabla v \, dx = \int_\Omega \beta uv \, dx, \qquad \forall v \in H_0^1(\Omega),$$

$$\int_\Omega \nabla p \cdot \nabla w \, dx = \int_\Omega (y - z_d) w \, dx, \qquad \forall w \in H_0^1(\Omega),$$
$$\alpha u = -\beta p \qquad \text{a.e. in } \Omega.$$

The Lagrangian approach is very helpful for complex nonlinear problems, where the structure of the adjoint equation is not easy to predict a priori. In the next example a prototypical semilinear problem is studied in depth. Although the Lagrangian approach is easily applicable, it should be carefully justified.

Example 3.5 (Optimal control of a semilinear equation).

$$\begin{cases} \min\, J(y,u) = \dfrac{1}{2}\,\|y - z_d\|_{L^2}^2 + \dfrac{\alpha}{2}\,\|u\|_{L^2}^2\,, \\ \text{subject to:} \\ \qquad -\Delta y + y^3 = u \qquad \text{in } \Omega, \\ \qquad\qquad y = 0 \qquad \text{on } \Gamma. \end{cases}$$

Weak formulation of the PDE. As was studied in Example 2.2., the variational formulation of the state equation is given in the following way: Find $y \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx = \int_\Omega uv \, dx, \quad \forall v \in H_0^1(\Omega).$$

Consequently, $e\colon H_0^1(\Omega) \times L^2(\Omega) \longrightarrow H^{-1}(\Omega)$ is defined by

$$\langle e(y,u), v \rangle_{H^{-1}, H_0^1} = \int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx - \int_\Omega uv \, dx,$$

for all $v \in H_0^1(\Omega)$. By Minty–Browder's theorem, there exists a unique solution to the PDE.

Differentiability. Since $y \in H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ with continuous injection, we consider the operator

$$\begin{aligned} N\colon L^6(\Omega) &\longrightarrow L^2(\Omega) \\ y &\longmapsto y^3. \end{aligned}$$

The Fréchet derivative of N is given by

$$N'(y)w = 3y^2 w.$$

Indeed,

$$\begin{aligned}\left\|(y+w)^3-y^3-3y^2w\right\|_{L^2} &= \left\|3yw^2+w^3\right\|_{L^2}\\ &\le 3\|y\|_{L^6}\|w\|_{L^6}^2+\|w\|_{L^6}^3 = O\left(\|w\|_{L^6}^2\right)\\ &= o\left(\|w\|_{L^6}\right).\end{aligned}$$

Moreover,

$$\begin{aligned}\left\|\left(N'(y+w)-N'(y)\right)v\right\|_{L^2} &= 3\left\|(y+w)^2-y^2\right\|_{L^2}\\ &= 3\|(2y+w)wv\|_{L^2}\\ &= 3\|2y+w\|_{L^6}\|w\|_{L^6}\|v\|_{L^6}\end{aligned}$$

$$\Rightarrow \left\|N'(y+w)-N'(y)\right\|_{\mathscr{L}(L^6(\Omega),L^2(\Omega))} \to 0 \text{ as } \|w\|_{L^6}\to 0,$$

which implies the continuity of the derivative.

Derivatives. The partial derivatives of $e(y,u)$ are given by

$$\begin{aligned}\langle e_y(y,u)w,v\rangle_{H^{-1},H_0^1} &= \int_\Omega \nabla w\cdot\nabla v\,dx+3\int_\Omega y^2wv\,dx,\\ \langle e_u(y,u)h,v\rangle_{H^{-1},H_0^1} &= -\int_\Omega hv\,dx.\end{aligned}$$

Lagrangian. The Lagrangian is defined by:

$$\begin{aligned}\mathscr{L}(y,u,p) = \frac{1}{2}\|y-z_d\|_{L^2}^2+\frac{\alpha}{2}\|u\|_{L^2}^2\\ -\int_\Omega \nabla y\cdot\nabla p\,dx+\int_\Omega y^3p\,dx+\int_\Omega up\,dx.\end{aligned}$$

Taking the partial derivative of the Lagrangian with respect to the state, we obtain that:

$$\begin{aligned}\mathscr{L}_y(y,u,p)w &= (y-z_d,w)-\int_\Omega \nabla w\cdot\nabla p\,dx-3\int_\Omega y^2wp\,dx\\ &= \int_\Omega (y-z_d)w\,dx-\int_\Omega \nabla p\cdot\nabla w\,dx-3\int_\Omega y^2pw\,dx=0,\end{aligned}$$

which implies that

$$\int_\Omega \nabla p\cdot\nabla w\,dx+3\int_\Omega y^2pw\,dx=\int_\Omega (y-z_d)w\,dx,\qquad \forall w\in H_0^1(\Omega).$$

On the other hand, taking the partial derivative with respect to u we get that:

$$\mathscr{L}_u(y,u,p)h = \alpha(u,h)_{L^2} + \int_\Omega hp\,dx = 0, \qquad \forall h \in L^2(\Omega)$$

$$\Rightarrow \quad \alpha u + p = 0 \quad \text{a.e. in } \Omega.$$

Optimality system.

$$\int_\Omega \nabla y \cdot \nabla v\,dx + \int_\Omega y^3 v\,dx = \int_\Omega uv\,dx, \qquad \forall v \in H_0^1(\Omega),$$

$$\int_\Omega \nabla p \cdot \nabla w\,dx + 3\int_\Omega y^2 pw\,dx = \int_\Omega (y - z_d)w\,dx, \qquad \forall w \in H_0^1(\Omega),$$

$$\alpha u + p = 0 \qquad \text{a.e. in } \Omega.$$

or, in strong form,

$$\begin{aligned} -\Delta y + y^3 &= u && \text{in } \Omega,\\ y &= 0 && \text{on } \Gamma,\\ -\Delta p + 3y^2 p &= y - z_d && \text{in } \Omega,\\ p &= 0 && \text{on } \Gamma,\\ \alpha u + p &= 0 && \text{a.e. in } \Omega. \end{aligned}$$

3.4 Second Order Sufficient Optimality Conditions

Within the theoretical framework developed so far, we can assure that if $(\bar{y},\bar{u})$ is a local optimal solution to (3.3) such that (3.7) holds, then $(\bar{y},\bar{u})$ is also a stationary point, i.e., it satisfies the optimality system (3.11). The following question arises: Is a stationary point also an optimal solution for (3.3)? Second order optimality conditions target precisely that question.

Consider again the general optimization problem (3.6)

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

where $C \subset U$ is a subset of a Banach space and f is a given functional.

Theorem 3.4. *Let U be a Banach space and $C \subset U$ a convex set. Let $f\colon U \longrightarrow \mathbb{R}$ be twice continuously Fréchet differentiable in a neighborhood of $\bar{u} \in U$. If $\bar{u}$ satisfies the first order necessary condition*

$$f'(\bar{u})(u - \bar{u}) \geq 0, \quad \forall u \in C, \tag{3.14}$$

and there exists some $\delta > 0$ *such that*

$$f''(u)[h]^2 \geq \delta \|h\|_U^2, \quad \forall h \in U, \tag{3.15}$$

then there exist constants $\varepsilon > 0$ *and* $\sigma > 0$ *such that*

$$f(u) \geq f(\bar{u}) + \sigma \|u - \bar{u}\|_U^2,$$

for all $u \in C : \|u - \bar{u}\|_U \leq \varepsilon$. *Therefore,* $\bar{u}$ *is a local minimum of* f *on* C.

Proof. Since f is twice Fréchet differentiable, a Taylor expansion can be used. Consequently, for some $\theta \in [0,1]$,

$$\begin{aligned} f(u) &= f(\bar{u}) + f'(\bar{u})(u-\bar{u}) + \frac{1}{2} f''(\bar{u} + \theta(u-\bar{u}))[u-\bar{u}]^2 \\ &\geq f(\bar{u}) + \frac{1}{2} f''(\bar{u} + \theta(u-\bar{u}))[u-\bar{u}]^2 \qquad \text{by (3.14)} \\ &= f(\bar{u}) + \frac{1}{2} f''(\bar{u})[u-\bar{u}]^2 + \frac{1}{2} \left[f''(\bar{u} + \theta(u-\bar{u})) - f''(\bar{u}) \right] [u-\bar{u}]^2. \end{aligned}$$

Since f is twice continuously Fréchet differentiable, there exists some $\varepsilon > 0$ such that

$$\|u - \bar{u}\| \leq \varepsilon \Rightarrow \left| \left[f''(\bar{u} + \theta(u-\bar{u})) - f''(\bar{u}) \right] [u-\bar{u}]^2 \right| \leq \frac{\delta}{2} \|u - \bar{u}\|_U^2.$$

Consequently,

$$\begin{aligned} f(u) &\geq f(\bar{u}) + \frac{1}{2} f''(\bar{u})[u-\bar{u}]^2 - \frac{\delta}{4} \|u - \bar{u}\|_U^2 \\ &\geq f(\bar{u}) + \frac{\delta}{4} \|u - \bar{u}\|_U^2, \qquad \text{by (3.15)}. \end{aligned}$$

The result follows by choosing $\sigma = \frac{\delta}{4}$. □

Condition (3.15) can be specified under additional properties of the optimization problem. In the case of PDE-constrained optimization, the positivity condition (3.15) is needed to hold for solutions of the linearized equation (3.8).

Theorem 3.5. *Let* $J : Y \times U \longrightarrow \mathbb{R}$ *and* $e : Y \times U \longrightarrow W$ *be twice continuously Fréchet differentiable such that* (3.7) *holds. Let* $(\bar{y}, \bar{u}, p)$ *be a solution to the optimality system* (3.11). *If there exists some constant* $\delta > 0$ *such that*

$$\mathscr{L}''_{(y,u)}[(w,h)]^2 \geq \delta \|h\|_U^2,$$

or, equivalently,

$$(w,h)\begin{pmatrix} J_{yy}(\bar{y},\bar{u}) & J_{yu}(\bar{y},\bar{u}) \\ J_{uy}(\bar{y},\bar{u}) & J_{uu}(\bar{y},\bar{u}) \end{pmatrix}\begin{pmatrix} w \\ h \end{pmatrix} - \left\langle p,(w,h)\begin{pmatrix} e_{yy}(\bar{y},\bar{u}) & e_{yu}(\bar{y},\bar{u}) \\ e_{uy}(\bar{y},\bar{u}) & e_{uu}(\bar{y},\bar{u}) \end{pmatrix}\begin{pmatrix} w \\ h \end{pmatrix}\right\rangle_{W',W} \geq \delta \|h\|_U^2, \quad (3.16)$$

for all $(w,h) \in Y \times U$ *that satisfy the equation*

$$e_y(\bar{y},\bar{u})w + e_u(\bar{y},\bar{u})h = 0,$$

then there exist constants $\varepsilon > 0$ *and* $\sigma > 0$ *such that*

$$J(y,u) \geq J(\bar{y},\bar{u}) + \sigma \|u-\bar{u}\|_U^2$$

for all $u \in U : \|u-\bar{u}\|_U \leq \varepsilon$.

Proof. First recall that

$$f'(\bar{u})h = \langle J_y(\bar{y},\bar{u}), y'(\bar{u})h\rangle_{Y',Y} + J_u(\bar{y},\bar{u})h.$$

The second derivative is then given by

$$f''(\bar{u})[h]^2 = J_{yy}(\bar{y},\bar{u})[y'(\bar{u})h]^2 + \langle J_y(\bar{y},\bar{u}), y''(\bar{u})[h]^2\rangle_{Y',Y} + \langle J_{yu}(\bar{y},\bar{u})[y'(\bar{u})h], h\rangle_{U',U} + \langle J_{uy}(\bar{y},\bar{u})[h], y'(\bar{u})h\rangle_{Y',Y} + J_{uu}(\bar{y},\bar{u})[h]^2. \quad (3.17)$$

On the other hand, by differentiating on both sides of the linearized equation (3.8) with respect to u, in direction h, we obtain that

$$e_{yy}(\bar{y},\bar{u})[y'(\bar{u})h]^2 + e_y(\bar{y},\bar{u})y''(\bar{u})[h]^2 + e_{yu}(\bar{y},\bar{u})[y'(\bar{u})h,h] + e_{yy}(\bar{y},\bar{u})[h,y'(\bar{u})h] + e_{uy}(\bar{y},\bar{u})[h]^2 = 0.$$

Therefore, using (3.11b) and the previous equation,

$$\begin{aligned} \langle J_y(\bar{y},\bar{u}), y''(\bar{u})[h]^2\rangle_{Y',Y} &= \langle p, e_y(\bar{y},\bar{u})y''(\bar{u})[h]^2\rangle_{W',W} \\ &= -\langle p, e_{yy}(\bar{y},\bar{u})[y'(\bar{u})h]^2 + e_{yu}(\bar{y},\bar{u})[y'(\bar{u})h,h] \\ &\quad + e_{uy}(\bar{y},\bar{u})[h,y'(\bar{u})h] + e_{uu}(\bar{y},\bar{u})[h]^2\rangle_{W',W.} \end{aligned}$$

Using the notation $w := y'(\bar{u})h$,

$$(w\ \ h)\begin{pmatrix} J_{yy}(\bar{y},\bar{u}) & J_{yu}(\bar{y},\bar{u}) \\ J_{uy}(\bar{u},\bar{y}) & J_{uu}(\bar{y},\bar{u}) \end{pmatrix}\begin{pmatrix} w \\ h \end{pmatrix} = J_{yy}(\bar{y},\bar{u})[w]^2 + J_{yu}(\bar{y},\bar{u})[w,h] + J_{uy}(\bar{y},\bar{u})[h,w] + J_{uu}(\bar{y},\bar{u})[h]^2,$$

and similarly for the second derivatives of e. Consequently, we obtain that $f''(\bar{u})[h]^2 \geq \delta \|h\|_U^2$ is equivalent to (3.16) and the result then follows from Theorem 3.4. □

Example 3.6. Consider again the semilinear optimal control problem:

$$\begin{cases} \min\ J(y,u) = \dfrac{1}{2}\|y - z_d\|_{L^2}^2 + \dfrac{\alpha}{2}\|u\|_{L^2}^2, \\ \text{subject to:} \\ \qquad \displaystyle\int_\Omega \nabla y \cdot \nabla v\, dx + \int_\Omega y^3 v\, dx = \int_\Omega uv\, dx, \qquad \forall v \in H_0^1(\Omega). \end{cases}$$

Recall that the first derivatives of e are given by

$$\langle e_y(y,u)w, v\rangle_{H^{-1},H_0^1} = \int_\Omega \nabla w \cdot \nabla v\, dx + 3\int_\Omega y^2 wv,$$
$$(e_u(y,u)h, v)_{L^2} = -\int_\Omega hv\, dx,$$

and the second derivatives are given by

$$\left\langle e_{yy}(y,u)[w]^2, v\right\rangle_{H^{-1},H_0^1} = 6\int_\Omega yw^2 v\, dx,$$

$$e_{yu}(y,u) = 0, \quad e_{uy}(y,u) = 0, \quad e_{uu}(y,u) = 0.$$

For the quadratic cost functional we get:

$$J_{yy}(\bar{y},\bar{u})[w]^2 = \|w\|_{L^2}^2, \qquad J_{yu}(y,u) = 0,$$
$$J_{uy}(y,u) = 0, \qquad J_{uu}(y,u)[h]^2 = \alpha\|h\|_{L^2}^2.$$

Condition (3.16) is, therefore, equivalent to

$$\|w\|_{L^2} + \alpha\|h\|_{L^2} - 6\int_\Omega yw^2 p\, dx \geq \delta\|h\|_{L^2}^2.$$

The sufficient optimality condition then holds if

$$\int_\Omega (1 - 6yp)w^2\, dx \geq 0.$$

Chapter 4
Numerical Optimization Methods

Consider the general optimization problem:

$$\min_{u \in U} J(y(u), u), \tag{4.1}$$

where U is a Hilbert space, $y(\cdot)$ is a partial differential equation (PDE) solution mapping and f is sufficiently smooth. The main strategy of optimization methods consists in, starting from an initial iterate u_0 and using first and second order information of the cost functional, moving along directions that lead to a decrease in the objective value. In this respect, the main questions are: How to choose the directions? How far to move along them?

In this chapter we present and analyze some infinite dimensional optimization methods for the solution of problem (4.1). Once the infinite dimensional algorithm is posed, the discretization of the partial differential equations is carried out. Such an approach is known as *optimize-then-discretize* in contrast to the *discretize-then-optimize* one, where the equations and the cost functional are first discretized and the problem is then solved using large-scale optimization tools.

The advantages of the *optimize-then-discretize* approach rely on a better understanding of the function space structure of the numerical algorithms. This can be of importance in problems where numerical difficulties may arise with direct discretization, as is the case of problems with low regularity multipliers.

Compactness of the involved operators also plays an important role in the convergence behavior of certain methods. The infinite dimensional Broyden–Fletcher–Goldfarb–Shanno (BFGS) method, for instance, converges locally with a superlinear rate if, in addition to the standard hypotheses, the difference between the initial Hessian approximation and the Hessian at the optimal solution is a compact operator.

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_4

Finally, a key concept in the framework of the *optimize-then-discretize* approach is *mesh independence*. Shortly, mesh independence implies that the convergence behavior (convergence rate and number of iterations) of an infinite dimensional method reflects the behavior of properly discretized problems, when the mesh size step is sufficiently small. Mesh independence turns out to be crucial for a robust behavior of the underlying method regardless of the size of the used grid. The infinite dimensional convergence theory for second order Newton type methods, for instance, enables to prove such a mesh independence property.

4.1 Descent Methods

Consider the optimization problem (3.1) in its reduced form:

$$\min_{u\in U} f(u) := J(y(u),u), \tag{4.2}$$

where U is a Hilbert space and $f : U \to \mathbb{R}$ is continuously Fréchet differentiable.

The main idea of descent methods consists in finding, at a given iterate u_k, a (descent) direction d_k such that

$$f(u_k + \alpha d_k) < f(u_k), \quad \text{for some } \alpha > 0, \tag{4.3}$$

based on first order information. Indeed, considering a linear model of the cost functional

$$f(u_k + \alpha d_k) \approx f(u_k) + \alpha(\nabla f(u_k), d_k)_U,$$

where $\nabla f(u_k)$ denotes the Riesz representative of the derivative of f, a descent direction may be chosen such that

$$d_k = \operatorname*{arg\,min}_{\|d\|_U=1} (\nabla f(u_k), d)_U. \tag{4.4}$$

Theorem 4.1. *Let $f : U \to \mathbb{R}$ be continuously Fréchet differentiable and $u_k \in U$ such that $\nabla f(u_k) \neq 0$. Then problem (4.4) has a unique solution given by*

$$\bar{d} = -\frac{\nabla f(u_k)}{\|\nabla f(u_k)\|_U}. \tag{4.5}$$

Proof. From the Cauchy–Schwarz inequality, we get for $d \in U$ with $\|d\|_U = 1$:

$$(\nabla f(u_k), d)_U \geq -\|\nabla f(u_k)\|_U \|d\|_U \geq -\|\nabla f(u_k)\|_U.$$

The equality is attained only if $d = -\frac{\nabla f(u_k)}{\|\nabla f(u_k)\|_U}$. □

Since f decreases most rapidly in the direction of $-\nabla f(u_k)$, the natural choice:

$$d_k = -\nabla f(u_k)$$

gives rise to the *steepest descent* method (also called *gradient* method). More generally, the (gradient related) descent direction d_k must satisfy the following condition:

$$-(\nabla f(u_k), d_k)_U \geq \eta \, \|\nabla f(u_k)\|_U \, \|d_k\|_U \tag{4.6}$$

for a fixed $\eta \in (0,1)$. Condition (4.6) is also referred to as *angle condition*.

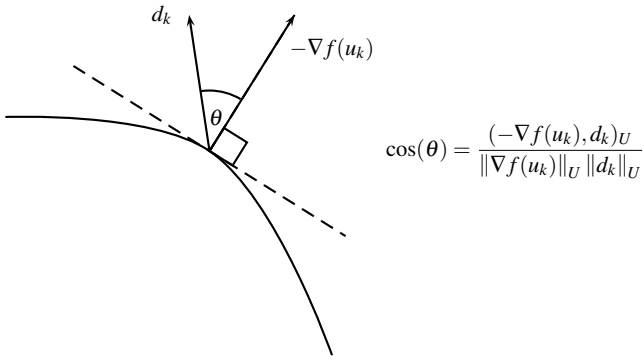

Fig. 4.1 Illustration of gradient related directions

Once the descent direction is determined, it is important to know how far to move in such direction, i.e., which *line search* parameter $\alpha_k > 0$ should be used. The ideal choice would be:

$$\alpha_k = \arg\min_{\alpha > 0} \{f(u_k + \alpha d_k)\}. \tag{4.7}$$

If such a minimization problem is easily solvable, one may choose α_k as the smallest positive root of the equation:

$$\frac{d}{d\alpha} f(u_k + \alpha d_k) = 0.$$

The latter constitutes a necessary optimality condition for (4.7).

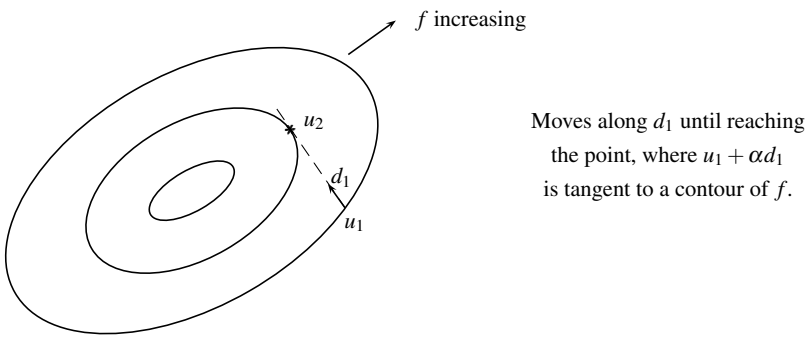

Fig. 4.2 Descent step with optimal line search parameter

Solving problem (4.7) may be too difficult in practice. Instead, the *line search* step sizes are chosen according to different strategies. In general, the following properties are required:

$$f(u_k + \alpha_k d_k) < f(u_k), \qquad \text{for all } k = 1, 2, \ldots, \tag{4.8}$$

$$f(u_k + \alpha_k d_k) - f(u_k) \xrightarrow{k\to\infty} 0 \Rightarrow \underbrace{\frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U}}_{= \frac{d}{d\alpha} f\left(u_k + \alpha \frac{d_k}{\|d_k\|}\right)} \xrightarrow{k\to\infty} 0. \tag{4.9}$$

A globally convergent descent algorithm can then be defined through the following steps:

Algorithm 1

1: Choose $u_0 \in U$ and set $k = 0$.
2: **repeat**
3: Choose a descent direction d_k such that (4.6) holds.
4: Determine α_k such that (4.8)– (4.9) hold.
5: Set $u_{k+1} = u_k + \alpha_k d_k$ and $k = k + 1$.
6: **until** stopping criteria.

Theorem 4.2. *Let f be continuously Fréchet differentiable and bounded from below. Let $\{u_k\}$, $\{d_k\}$, and $\{\alpha_k\}$ be sequences generated by Algorithm 1 with (4.6), (4.8), and (4.9) holding. Then*

$$\lim_{k\to\infty} \nabla f(u_k) = 0,$$

and every accumulation point of $\{u_k\}$ is a stationary point of f.

Proof. Since the sequence $\{f(u_k)\}$ is decreasing (thanks to (4.8)) and bounded from below, it converges to some value $\bar{f}$. Consequently,

$$\left|f(u_k+\alpha_k d_k) - f(u_k)\right| \le \underbrace{\left|f(u_k+\alpha_k d_k) - \bar{f}\right|}_{\xrightarrow{k\to\infty} 0} + \underbrace{\left|\bar{f} - f(u_k)\right|}_{\xrightarrow{k\to\infty} 0} \xrightarrow{k\to\infty} 0,$$

which, by condition (4.9), implies that

$$\lim_{k\to\infty} \frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U} = 0.$$

Thanks to (4.6)

$$0 \le \eta \|\nabla f(u_k)\|_U \le \frac{(-\nabla f(u_k), d_k)_U}{\|d_k\|_U} \xrightarrow{k\to\infty} 0,$$

which implies that

$$\lim_{k\to\infty} \nabla f(u_k) = 0.$$

If $\bar{u}$ is an accumulation point of $\{u_k\}$, then the continuity of ∇f implies that

$$\nabla f(\bar{u}) = \lim_{k\to\infty} \nabla f(u_k) = 0. \qquad \square$$

A quite popular line search strategy is the *Armijo rule* with backtracking, which consists in the following: Given a descent direction d_k of f at u_k, choose the largest $\alpha_k \in \left\{1, \frac{1}{2}, \frac{1}{4}, \ldots\right\}$ such that

$$f(u_k+\alpha_k d_k) - f(u_k) \le \gamma\alpha_k (\nabla f(u_k), d_k)_U,$$

where $\gamma \in (0,1)$ is a constant (typically $\gamma = 10^{-4}$). Using the notation $\phi(\alpha) = f(u_k + \alpha d_k)$, Armijo's rule may also be written as

$$\phi(\alpha) \le \phi(0) + \gamma\phi'(0). \tag{4.10}$$

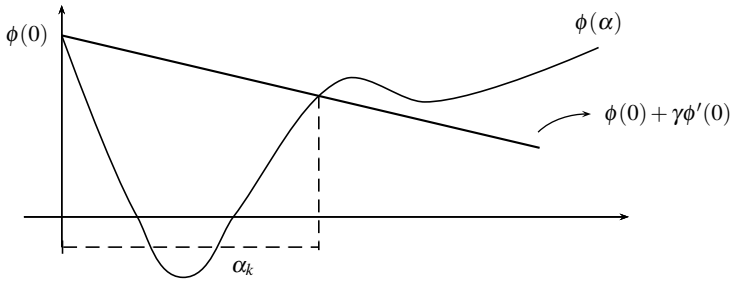

Fig. 4.3 Illustration of Armijo's rule

Lemma 4.1. *Let ∇f be uniformly continuous on the level set*

$$N_0^\rho = \{u+d : f(u) \le f(u_0), \|d\|_U \le \rho\},$$

for some $\rho > 0$. Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $u_k \in U$ satisfying $f(u_k) \le f(u_0)$ and all d_k satisfying

$$\frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U} \le -\varepsilon,$$

there holds

$$f(u_k + \alpha d_k) - f(u_k) \le \gamma\alpha(\nabla f(u_k), d_k)_U, \quad \forall \alpha \in \left[0, \frac{\delta}{\|d_k\|_U}\right].$$

Proof. We refer to [30, p. 102]. □

Proposition 4.1. *Let ∇f be uniformly continuous on N_0^ρ, for some $\rho > 0$. If the iterates generated by Algorithm 1, with $\{\alpha_k\}$ satisfying the Armijo rule, are such that*

$$\|d_k\|_U \ge \frac{-(\nabla f(u_k), d_k)_U}{\|d_k\|_U},$$

then $\{\alpha_k\}$ satisfies (4.9).

Proof. Assume there exists an infinite set K and $\varepsilon > 0$ such that

$$f(u_k + \alpha_k d_k) - f(u_k) \xrightarrow{k\to\infty} 0,$$

and

$$\frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U} \leq -\varepsilon, \qquad \forall k \in K.$$

Then

$$\|d_k\|_U \geq -\frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U} \geq \varepsilon > 0, \qquad \forall k \in K.$$

From Lemma 4.1 there exists some $\delta > 0$ such that

$$\begin{aligned} f(u_k + \alpha_k d_k) - f(u_k) \leq & \gamma\delta \frac{(\nabla f(u_k), d_k)_U}{\|d_k\|_U} \\ \leq & -\gamma\delta\varepsilon, \qquad \forall k \in K. \end{aligned}$$

Consequently,

$$f(u_k + \alpha_k d_k) - f(u_k) \not\longrightarrow 0,$$

which contradicts the hypothesis. □

An alternative line-search strategy is given by the Wolfe conditions:

$$f(u_k + \alpha_k d_k) - f(u_k) \leq \gamma\alpha_k (\nabla f(u_k), d_k)_U \,, \tag{4.11}$$

$$(\nabla f(u_k + \alpha_k d_k), d_k)_U \geq \beta (\nabla f(u_k), d_k)_U \,, \tag{4.12}$$

with $0 < \gamma < \beta < 1$. Condition (4.11) is similar to the sufficient decrease condition in Armijo's rule, while condition (4.12) is known as the curvature condition and guarantees that the slope of the function

$$\phi(\alpha) = f(u_k + \alpha d_k)$$

is less negative at the chosen α_k than at 0

$$\left.\frac{d\phi}{d\alpha}\right|_{\alpha_k} \geq \beta \left.\frac{d\phi}{d\alpha}\right|_0 .$$

Condition (4.12) is important in order to avoid very small values of α and ensures that the next chosen α_k is closer to a stationary point of problem (4.7).

A stronger version of the Wolfe conditions is obtained by replacing (4.12) with

$$|(\nabla f(u_k + \alpha_k d_k), d_k)_U| \leq \beta |(\nabla f(u_k), d_k)_U| \,, \tag{4.13}$$

in order to guarantee that the chosen step α_k actually lies in a neighborhood of a local minimizer of $\phi(\alpha)$.

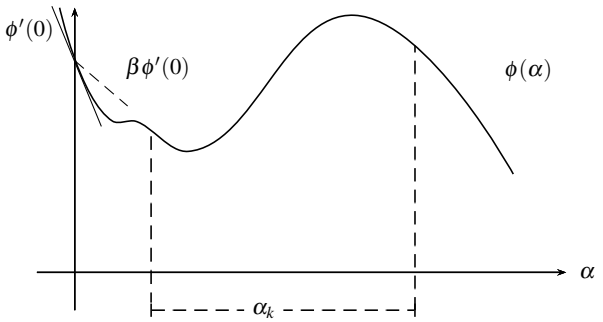

Fig. 4.4 Illustration of Wolfe's condition (4.12)

In the case of the PDE-constrained optimization problem (3.1), since the derivative of the cost functional is characterized by (see Theorem 3.3)

$$(\nabla f(u),h)_U = -\langle e_u(y(\bar{u}),\bar{u})^* p, h\rangle_{U',U} + J_u(y(\bar{u}),\bar{u})h, \tag{4.14}$$

where p solves the adjoint equation (3.11b) and y the state equation (3.11a), the complete steepest descent algorithm is given through the following steps:

Algorithm 2

1: Choose $u_0 \in U$ and solve

$$e(y,u_0) = 0, \qquad e_y(y_0,u_0)^* p = J_y(y_0,u_0),$$

to obtain (y_0,p_0). Set $k=0$.
2: **repeat**
3: Choose the descent direction $d_k = -\nabla f(u_k)$ according to (4.14).
4: Determine α_k such that (4.8)–(4.9) hold.
5: Set $u_{k+1} = u_k + \alpha_k d_k$ and solve sequentially

$$e(y,u_{k+1}) = 0, \qquad e_y(y_{k+1},u_{k+1})^* p = J_y(y_{k+1},u_{k+1}),$$

to obtain (y_{k+1},p_{k+1}). Set $k = k+1$.
6: **until** stopping criteria.

The verification of both Armijo's and Wolfe's rule requires the repetitive evaluation of the cost functional. Since in the case of PDE-constrained optimization such evaluation

involves the solution of a PDE, the studied line search strategies may become very costly in practice.

Under the stronger requirement that ∇f is Lipschitz continuous on N_0^ρ, for some $\rho > 0$, with Lipschitz constant $M > 0$, an alternative line search condition for the steepest descent method is given by

$$f(u_k + \alpha_k d_k) \leq f(u_k) - \frac{\eta^2}{2M} \|\nabla f(u_k)\|^2 .$$

Typically, a constant parameter $\alpha_k = \alpha \in (0,1)$, for all $k \geq 0$, is considered.

Example 4.1. For the linear quadratic problem:

$$\begin{cases} \min J(y,u) = \frac{1}{2}\|y - z_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 = f(u), \\ \text{subject to:} \\ \qquad -\Delta y = u \qquad \text{in} \quad \Omega, \\ \qquad y = 0 \qquad\quad \text{on} \quad \Gamma. \end{cases}$$

we obtain the following characterization of the gradient:

$$\nabla f(u) = \alpha u + p,$$

where p solves the adjoint equation

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

The general framework of descent methods allows several choices of directions d_k. A special class is given by directions of the following type:

$$d_k = -H_k^{-1} \nabla f(u_k),$$

where $\{H_k\}_{k\in\mathbb{N}} \subset \mathscr{L}(U)$ satisfy

$$m\|v\|_U^2 \leq (H_k v, v)_U \leq M\|v\|_U^2 , \qquad \text{for all } v \in U \text{ and all } k = 1,2,\ldots, \tag{4.15}$$

for some constants $0 < m < M$ independent of k. Condition (4.6) is then fulfilled for such a family. Indeed,

$$\begin{aligned}\frac{(-\nabla f(u_k),d_k)_U}{\|\nabla f(u_k)\|_U\,\|d_k\|_U} &= \frac{\left(-\nabla f(u_k),-H_k^{-1}\nabla f(u_k)\right)_U}{\|\nabla f(u_k)\|_U\,\left\|H_k^{-1}\nabla f(u_k)\right\|_U}\\ &= \frac{\left(H_kH_k^{-1}\nabla f(u_k),H_k^{-1}\nabla f(u_k)\right)_U}{\left\|H_kH_k^{-1}\nabla f(u_k)\right\|_U\,\left\|H_k^{-1}\nabla f(u_k)\right\|_U}\\ &\geq \frac{m\left\|H_k^{-1}\nabla f(u_k)\right\|_U^2}{M\left\|H_k^{-1}\nabla f(u_k)\right\|_U^2}\\ &= \frac{m}{M}=\eta.\end{aligned}$$

Remark 4.1. Although Theorem 4.2 provides a global convergence result, the local convergence of first order descent methods is typically slow (see [37, p. 45] for further details).

Program: Steepest Descent Method for Optimization of the Poisson Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Regularization parameter: ');

[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1);
res=1; iter=0; tol=1e-3;

while res >= tol
    iter=iter+1
    y=lap\u;         %%%%% State equation %%%%%
    p=lap\(y-z);     %%%%% Adjoint solver %%%%%
```

```
    beta=armijolq(u,y,p,z,alpha,lap);   %% Armijo line search %%
    uprev=u;
    u=u-beta*(p+alpha*u);   %%%%% Gradient step %%%%%
    res=l2norm(u-uprev)
end
```

Program: Armijo Line Search for the Linear-Quadratic Problem

```
function arm=armijolq(u,y,p,z,alpha,lap)
    countarm=0;
    gradcost=l2norm(p+alpha*u)^2;
    cost1=1/2*l2norm(y-z)^2+alpha/2*l2norm(u)^2;
    beta=1; armijo=1e5;

         while armijo > -1e-4*beta*gradcost
         beta=1/2^(countarm);
         uinc=u-beta*(p+alpha*u);
         yinc=lap\uinc;
         cost2=1/2*l2norm(yinc-z)^2+alpha/2*l2norm(uinc)^2;
         armijo=cost2-cost1;
         countarm=countarm+1
         end
     arm=beta;
```

4.2 Newton's Method

Another type of directions is obtained if, at each iteration, a quadratic model of the cost functional $f(u_k+d)$ is minimized with respect to the direction d. Using Taylor's expansion, such a quadratic model is given by

$$q(d) = f(u_k) + (\nabla f(u_k), d)_U + \frac{1}{2}(\nabla^2 f(u_k)d, d)_U, \tag{4.16}$$

where $\nabla^2 f(u_k)d$ stands for the Riesz representative of $f''(u_k)d$.

The minimizer d_k to (4.16) then satisfies the first order optimality condition

$$\nabla f(u_k) + \nabla^2 f(u_k)d_k = 0$$

or, equivalently,

$$\nabla^2 f(u_k) d_k = -\nabla f(u_k). \tag{4.17}$$

In this manner, assuming invertibility of the second derivative, the direction

$$d_k = -\left(\nabla^2 f(u_k)\right)^{-1} \nabla f(u_k) \tag{4.18}$$

is obtained, which leads to the iteration

$$u_{k+1} = u_k - \left(\nabla^2 f(u_k)\right)^{-1} \nabla f(u_k).$$

Equation (4.17) corresponds to the Newton iteration for the solution of the optimality condition

$$\nabla f(u) = 0.$$

If the second derivative of f at the iterates $\{u_k\}$ satisfies condition (4.15), convergence of the Newton iterates

$$u_{k+1} = u_k - \alpha_k \left(\nabla^2 f(u_k)\right)^{-1} \nabla f(u_k) \tag{4.19}$$

is obtained, according to Theorem 4.2.

Condition (4.15), however, is very difficult to be verified at each Newton iterate. Moreover, although $\nabla^2 f$ is "positive" at the solution, it does not have to be so at each iterate.

By observing that

$$\left(\nabla^2 f(u_k) v, v\right)_U = \left(\nabla^2 f(\bar{u}) v, v\right)_U + \left(\left[\nabla^2 f(u_k) - \nabla^2 f(\bar{u})\right] v, v\right)_U,$$

an alternative condition is obtained by assuming "positivity" of $\nabla^2 f(\bar{u})$ at a solution $\bar{u}$ and Lipschitz continuity of $\nabla^2 f$ in a neighborhood of $\bar{u}$.

Theorem 4.3. *Let $\bar{u} \in U$ be a local optimal solution to problem (4.2) and let f be twice continuously differentiable. Let $\nabla^2 f$ be Lipschitz continuous in a neighborhood $V(\bar{u})$ of $\bar{u}$ and*

$$(\nabla^2 f(\bar{u}) d, d)_U \geq \kappa \|h\|_{U,}^2 \qquad \forall h \in U, \tag{4.20}$$

for some constant $\kappa > 0$. Then there exists a constant $\delta > 0$ such that, if $\|u_0 - \bar{u}\|_U < \delta$, then:

a) the Newton iterates

$$u_{k+1} = u_k - \left(\nabla^2 f(u_k)\right)^{-1} \nabla f(u_k) \tag{4.21}$$

converge to $\bar{u}$,

b) there exists a constant $\bar{C} > 0$ such that

$$\|u_{k+1} - \bar{u}\|_U \leq \bar{C} \|u_k - \bar{u}\|_U^2. \tag{4.22}$$

Proof. Assuming that $\nabla f(\bar{u}) = 0$ and (for the moment) that $\nabla^2 f(u_k)$ is invertible for all k, we obtain, from the iterates given by (4.21), the estimate

$$\begin{aligned}
\|u_{k+1} - \bar{u}\|_U &= \left\| u_k - \left(\nabla^2 f(u_k)\right)^{-1} \nabla f(u_k) - \bar{u} \right\|_U \\
&= \left\| \left(\nabla^2 f(u_k)\right)^{-1} \left[\nabla^2 f(u_k)(u_k - \bar{u}) - \nabla f(u_k) + \nabla f(\bar{u}) \right] \right\| \\
&\leq \left\| \left(\nabla^2 f(u_k)\right)^{-1} \right\|_{\mathscr{L}(U)} \left\| \nabla f(\bar{u}) - \nabla f(u_k) + \nabla^2 f(u_k)(u_k - \bar{u}) \right\|_U .
\end{aligned}$$

From the mean value theorem we get that

$$\nabla f(\bar{u}) - \nabla f(u_k) = \int_0^1 \nabla^2 f(u_k + t(\bar{u} - u_k)) dt \ (\bar{u} - u_k),$$

which implies that

$$\begin{aligned}
&\left\| \nabla f(\bar{u}) - \nabla f(u_k) + \nabla^2 f(u_k)(u_k - \bar{u}) \right\|_U \\
&\qquad = \left\| \int_0^1 \left[\nabla^2 f(u_k + t(\bar{u} - u_k)) - \nabla^2 f(u_k) \right] dt \ (\bar{u} - u_k) \right\|_U \\
&\qquad \leq \int_0^1 \left\| \nabla^2 f(u_k + t(\bar{u} - u_k)) - \nabla^2 f(u_k) \right\|_{\mathscr{L}(U)} dt \, \|\bar{u} - u_k\|_U \\
&\qquad \leq \int_0^1 Lt \, \|\bar{u} - u_k\|_U \, dt \, \|\bar{u} - u_k\|_U \\
&\qquad \leq \frac{L}{2} \|\bar{u} - u_k\|_U^2 ,
\end{aligned}$$

where $L > 0$ is the Lipschitz constant for $\nabla^2 f$ on $V(\bar{u})$.

On the other hand, thanks to (4.20) there exists $\left(\nabla^2 f(\bar{u})\right)^{-1}$. By choosing the constant $\delta = \frac{1}{2L \left\| \left(\nabla^2 f(\bar{u})\right)^{-1} \right\|_{\mathscr{L}(U)}}$, we get that

$$\left\| \nabla^2 f(\bar{u}) - \nabla^2 f(u_0) \right\|_{\mathscr{L}(u)} \leq L \|\bar{u} - u_0\|_U < \frac{1}{2 \left\| \left(\nabla^2 f(\bar{u})\right)^{-1} \right\|_{\mathscr{L}(U)}},$$

which, by the theorem on inverse operators, implies that $\nabla^2 f(u_0)$ is invertible (see, e.g., [3, p. 51]). Moreover,

$$\begin{aligned}\left\|\left(\nabla^2 f(u_0)\right)^{-1}\right\|_{\mathscr{L}(U)} &\leq \frac{\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}}{1-\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}\left\|\nabla^2 f(\bar{u})-\nabla^2 f(u_0)\right\|_{\mathscr{L}(U)}} \\ &\leq 2\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}.\end{aligned}$$

Consequently,

$$\begin{aligned}\|u_1-\bar{u}\|_U &\leq \left\|\left(\nabla^2 f(u_0)\right)^{-1}\right\|_{\mathscr{L}(U)}\frac{L}{2}\|u_0-\bar{u}\|_U^2 \\ &\leq \left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)} \quad L \quad \frac{1}{2L\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}}\|u_0-\bar{u}\|_U \\ &\leq \frac{1}{2}\|u_0-\bar{u}\|_U.\end{aligned}$$

By induction we obtain the invertibility of $\nabla^2 f(u_k)$ and also that

$$\|u_k-\bar{u}\|_U < \delta \Rightarrow \|u_{k+1}-\bar{u}\|_U \leq \frac{1}{2}\|u_k-\bar{u}\|_U.$$

Therefore, the sequence $\{u_k\}$ converges toward $\bar{u}$ as $k \to \infty$. Additionally,

$$\left\|\left(\nabla^2 f(u_k)\right)^{-1}\right\|_{\mathscr{L}(U)} \leq 2\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}, \qquad \text{for all } k,$$

and, consequently,

$$\|u_{k+1}-\bar{u}\|_U \leq L\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}\|\bar{u}-u_k\|_{U.}^2$$

Taking $\bar{C} := L\left\|\left(\nabla^2 f(\bar{u})\right)^{-1}\right\|_{\mathscr{L}(U)}$, the result follows. □

The last result implies that the Newton method converges locally with quadratic rate.

Considering the special structure of the PDE-constrained optimization problems, the convergence result for Newton's method can be formulated as follows.

Theorem 4.4. *Let $(\bar{y},\bar{u})$ be a local optimal solution of the problem:*

$$\begin{cases}\min J(y,u) \\ \textit{subject to:} \\ \qquad e(y,u)=0,\end{cases}$$

where $J\colon Y\times U\longrightarrow\mathbb{R}$ and $e\colon Y\times U\longrightarrow W$ are twice continuously Fréchet differentiable with Lipschitz continuous second derivatives. Further, assume that:

$$e_y(y,u) \text{ is a bijection for all } (y,u) \text{ in a neighborhood of } (\bar{y},\bar{u}).$$

If there exists a constant $\kappa>0$ such that

$$\mathscr{L}''_{(y,u)}(\bar{y},\bar{u},p)[(w,h)]^2\geq\kappa\|h\|_U^2 \tag{4.23}$$

for all $(w,h)\in Y\times U$ satisfying

$$e_y(\bar{y},\bar{u})w+e_u(\bar{y},\bar{u})h=0,$$

then the Newton iterates converge locally quadratically.

Before proving Theorem 4.4, let us take a closer look at the structure of the corresponding Newton system. Similar to the proof of Theorem 3.5 we obtain that

$$f''(u)[h_1,h_2]=\mathscr{L}''_{(y,u)}(y,u,p)[(w_1,h_1),(w_2,h_2)],$$

for all $(w_i,h_i)\in Y\times U,\ i=1,2$, satisfying the linearized equation

$$e'(y,u)(w_i,h_i)=e_y(y,u)w_i+e_u(y,u)h_i=0, \tag{4.24}$$

where, in addition,

$$e(y,u)=0 \tag{4.25}$$

and

$$e_y(y,u)^*p=J_y(y,u). \tag{4.26}$$

Since $e'(y,u):Y\times U\to W$ is a continuous linear operator, it follows from the annihilator's lemma that

$$\ker(e'(y,u))^\perp=\operatorname{range}(e'(y,u)^*)$$

and, therefore, if for some $(\varphi_1,\varphi_2)\in Y'\times U'$

$$\mathscr{L}''_{(y,u)}(y,u,p)[(w_1,h_1),(w_2,h_2)]=\langle(\varphi_1,\varphi_2),(w_2,h_2)\rangle_{Y'\times U'}$$

for all $(w_2,h_2)\in\ker(e'(y,u))$, then there exists an element $\xi\in\ker(e'(y,u)^\perp)$ such that

$$\mathscr{L}''_{(y,u)}(y,u,p)[(w_1,h_1)]+(\xi_1,\xi_2)=(\varphi_1,\varphi_2)\qquad\text{in } Y'\times U'.$$

Moreover, there exists a unique $\pi\in W'$ such that

$$e'(y,u)^*\pi=\xi.$$

Therefore, in a neighborhood of $\bar{u}$,

$$f''(u)h_1 = \phi \in U'$$

is equivalent to

$$\langle \mathscr{L}''_{(y,u)}(y,u,p)[(w_1,h_1)] + e'(y,u)^*\pi, (v,h)\rangle_{Y'\times U'} = \langle \phi, h\rangle_{U'}, \tag{4.27}$$

for all $(v,h) \in Y \times U$. By taking, in particular, the cases $h = 0$ and $v = 0$ we arrive at the following system of equations:

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y,u,p) & e'(y,u)^* \\ e'(y,u) & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} w_1 \\ h_1 \end{pmatrix} \\ \pi \end{pmatrix} = \begin{pmatrix} 0 \\ \phi \\ 0 \end{pmatrix}. \tag{4.28}$$

Newton's method $\nabla^2 f(u)\delta_u = -\nabla f(u)$ for the PDE-constrained optimization problem is then given through the following steps:

Algorithm 3

1: Choose $u_0 \in V(\bar{u})$ and solve

$$e(y,u_0) = 0, \qquad e_y(y_0,u_0)^* p = J_y(y_0,u_0),$$

to obtain (y_0, p_0). Set $k = 0$.

2: **repeat**

3: Newton system: solve for $(\delta_y, \delta_u, \delta_\pi) \in Y \times U \times W'$

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y_k,u_k,p_k) & e'(y_k,u_k)^* \\ e'(y_k,u_k) & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_\pi \end{pmatrix} = \begin{pmatrix} 0 \\ e_u(y,u)^* p - J_u(y_k,u_k) \\ 0 \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$ and solve

$$\begin{aligned} &e(y,u_{k+1}) = 0, && \text{to obtain } y_{k+1} \in Y, \\ &e_y(y_{k+1},u_{k+1})^* p = J_y(y_{k+1},u_{k+1}), && \text{to obtain } p_{k+1} \in W'. \end{aligned}$$

5: Set $k = k+1$.

6: **until** Stopping criteria.

Proof. (Theorem 4.4). We have already argued that if

$$\nabla^2 f(u)\delta_u = -\nabla f(u)$$

has a unique solution $\delta_u \in U$, then there exists a unique $\pi \in W'$ such that (4.28) holds. Since condition (4.23) is equivalent to (4.20), the Newton iterates for the reduced problem are well defined and there exists a unique solution for system (4.28). □

Program: Newton Method for the Optimization of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Regularization parameter: ');
[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

lap=matrices(n,h);       %%%%% Laplacian %%%%%
u=sparse(n^2,1);         %%%%% Initial control %%%%%
y=semilinear(lap,u);     %%%%% Initial state %%%%%%

Y=spdiags(y,0,n^2,n^2);
p=(lap+3*Y.^2)\(y-z);    %%%%% Initial adjoint %%%%%
res=1; iter=0;
while res >= 1e-3
    iter=iter+1
    Y=spdiags(y,0,n^2,n^2); P=spdiags(p,0,n^2,n^2);

    A=[speye(n^2)-6*Y.*P sparse(n^2,n^2) lap+3*Y.^2
        sparse(n^2,n^2) alpha*speye(n^2) -speye(n^2)
        lap+3*Y.^2 -speye(n^2) sparse(n^2,n^2)];

    F=[sparse(n^2,1);-p-alpha*u;sparse(n^2,1)];

    delta=A\F;
    uprev=u;
    u=u+delta(n^2+1:2*n^2); %%%%% Control update %%%%%
    y=semilinear(lap,u);    %%%%% State equation %%%%%
    Y=spdiags(y,0,n^2,n^2);
    p=(lap+3*Y.^2)\(y-z);   %%%%% Adjoint equation %%%%%
    res=l2norm(u-uprev)
end
```

4.3 Quasi-Newton Methods

For the numerical solution of the reduced problem (4.1), we have so far considered descent directions of the type

$$d_k = -H_k^{-1} \nabla f(u_k).$$

For the choice $H_k^{-1} = I$ (steepest descent) a globally convergent behaviour is obtained, but possibly with a slow convergence rate. On the other hand, for $H_k^{-1} = \left(\nabla^2 f(u_k)\right)^{-1}$ (Newton) a fast local convergence is obtained, but with additional computational effort and without global convergence guarantee.

An alternative consists in considering operators H_k which approximate the second derivative and lead to a fast convergence rate, but at the same time preserve the positivity and lead to a globally convergent method.

As for finite dimensional problems, an alternative consists in approximating the second derivative by using an operator $B_{k+1} \in \mathscr{L}(U)$ that fulfills the secant equation:

$$B_{k+1} \underbrace{(u_{k+1} - u_k)}_{s_k} = \underbrace{\nabla f(u_{k+1}) - \nabla f(u_k)}_{z_k}.$$

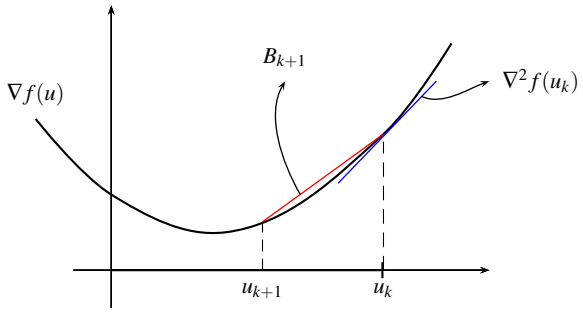

Fig. 4.5 Illustration of the secant equation

Since the operator B_{k+1} is not uniquely determined from the secant equation, additional criteria are needed for the construction of the operators. If $U = \mathbb{R}^n$, one possibility

consists in choosing rank-2 updates of the form:

$$B_{k+1} = B_k + \gamma_k w_k w_k^T + \beta_k v_k v_k^T,$$

as close as possible to the matrix B_k. Specifically, if B_{k+1} is chosen as solution of:

$$\begin{cases} \min\limits_{B} \ ||W(B^{-1} - B_k^{-1})W||_F \\ \text{subject to:} \\ \qquad B = B^T, \\ \qquad Bs_k = z_k, \end{cases} \tag{4.29}$$

where $\|\cdot\|_F$ stands for the Frobenius norm and W is a positive definite matrix such that $W^2 s_k = z_k$, then the solution to problem (4.29) is given by the BFGS update:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{z_k z_k^T}{z_k^T s_k}.$$

A generalization of the previous formula to Hilbert spaces yields:

$$B_{k+1} = B_k - \frac{B_k s_k \otimes B_k s_k}{(B_k s_k, s_k)_U} + \frac{z_k \otimes z_k}{(z_k, s_k)_U}, \tag{4.30}$$

where for $w, z \in U$, the operator $w \otimes z$ is defined by

$$(w \otimes z)(v) := (z, v)_U \, w.$$

The complete BFGS algorithm, without line-search, is then given through the following steps:

Algorithm 4

1: Choose $u_0 \in U$, $B_0 \in \mathscr{L}(U)$ symmetric, set $k = 0$.
2: **repeat**
3: Solve $B_k d_k = -\nabla f(u_k)$.
4: Update $u_{k+1} = u_k + d_k$.
5: Compute $\nabla f(u_{k+1})$.
6: Set $s_k = u_{k+1} - u_k, \quad z_k = \nabla f(u_{k+1}) - \nabla f(u_k)$.
7: Update $B_{k+1} = B_k - \dfrac{B_k s_k \otimes B_k s_k}{(B_k s_k, s_k)_U} + \dfrac{z_k \otimes z_k}{(z_k, s_k)_U}$.
8: Set $k = k + 1$.
9: **until** Stopping criteria.

Similar to the descent algorithm, the BFGS requires only one solution of the state equation and one solution of the adjoint equation.

Theorem 4.5. *Let f be twice Fréchet differentiable and $\nabla^2 f$ be Lipschitz continuous in a neighborhood of $\bar{u}$, with bounded inverse. Let B_0 be an initial positive operator. If $B_0 - \nabla^2 f(\bar{u})$ is compact, then the BFGS iterates converge q-superlinearly to $\bar{u}$ provided*

$$\|u_0 - \bar{u}\|_U \qquad \text{and} \qquad \left\|B_0 - \nabla^2 f(\bar{u})\right\|_{\mathscr{L}(U)}$$

are sufficiently small.

Under the choice of an appropriate line search parameter that guarantees the satisfaction of the curvature condition $(z_k, s_k)_U > 0$, global convergence of the BFGS is obtained. Typically, the Wolfe conditions are used for this purpose. As with Newton's method, the bounded inverse condition in Theorem 4.5 can be replaced by a convexity condition on the second derivative.

Remark 4.2. The compactness condition on the operator $B_0 - \nabla^2 f(\bar{u})$ is necessary for the superlinear convergence rate of the method [22, 39]. Indeed, an example in the space of square summable sequences l_2 is presented in [55], where only linear convergence of the BFGS is obtained. In the example, the quadratic function $f(x) = \frac{1}{2}x^T x$ and the initial sequence $x_0 = (2^{-0}, 2^{-1}, 2^{-2}, \dots)$ are considered. Additionally, the following starting infinite tridiagonal matrix is used:

$$B_0 = \begin{pmatrix} 5 & -2 & & \\ -2 & 5 & -2 & \\ & -2 & 5 & \ddots \\ & & \ddots & \ddots \end{pmatrix}.$$

Remark 4.3. With the same hypotheses as in Theorem 4.5, a mesh independence result for the BFGS can be obtained. Specifically, let U_h be a finite-dimensional Hilbert space that approximates U, and $u^h \in U_h$ be the solution to the discretized optimization problem:

$$\min_{u^h \in U_h} f_h(u^h).$$

Moreover, let the iteration indexes be defined by:

$$i(\varepsilon) := \min\{k \in \mathbb{N} : \|\nabla f(u_k)\|_U \leq \varepsilon\},$$
$$i_h(\varepsilon) := \min\{k \in \mathbb{N} : \|\nabla f_h(u_k^h)\|_{U_h} \leq \varepsilon\}.$$

If the discretization of the problem satisfies suitable conditions (see [38]) and the assumptions of Theorem 4.5 hold, then for each $\varepsilon > 0$ there exists h_ε such that

$$i(\varepsilon) - 1 \leq i_h(\varepsilon) \leq i(\varepsilon), \text{ for } h \leq h_\varepsilon. \tag{4.31}$$

This means that the number of required iterations of both the infinite dimensional algorithm and the discretized one differs at most by one, for h sufficiently small.

Example 4.2.

$$\begin{cases} \min J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega |y - z_d|^2 + \dfrac{\alpha}{2}\int_\Omega u^2\,dx, \\ \text{subject to:} \\ \qquad -\Delta y + y^3 = u, \\ \qquad\qquad y\big|_\Gamma = 0. \end{cases}$$

Adjoint equation:

$$\begin{aligned} -\Delta p + 3y^2 p &= y - z_d, && \text{in} \quad \Omega, \\ p &= 0, && \text{on} \quad \Gamma. \end{aligned}$$

Gradient of f:

$$\nabla f(u) = \alpha u + p.$$

Using a finite differences discretization and the simple integration formula

$$(u,v)_{L^2} \approx h^2 \mathbf{u}^T \mathbf{v},$$

we get that

$$(w \otimes z)v = (z,v)_{L^2} w \approx h^2 \mathbf{z}^T \mathbf{v}\, \mathbf{w} = h^2 \mathbf{w}\mathbf{z}^T\, \mathbf{v},$$

where the bold notation stands for the discretized vector in $\mathbb{R}^m$ of a function $u \in L^2(\Omega)$. The BFGS update is then given by:

$$\begin{aligned} B_{k+1} &= B_k - \frac{h^2 B_k \mathbf{s}_k (B_k \mathbf{s}_k)^T}{h^2 \mathbf{s}_k^T B_k \mathbf{s}_k} + \frac{h^2 \mathbf{z}_k \mathbf{z}_k^T}{h^2 \mathbf{z}_k^T \mathbf{s}_k} \\ &= B_k - \frac{B_k \mathbf{s}_k \mathbf{s}_k^T B_k}{\mathbf{s}_k^T B_k \mathbf{s}_k} + \frac{\mathbf{z}_k \mathbf{z}_k^T}{\mathbf{z}_k^T \mathbf{s}_k}, \end{aligned}$$

which coincides with the matrix update in $\mathbb{R}^m$.

If a finite element discretization is used, then

$$(u,v)_{L^2} \approx \mathbf{u}^T M \mathbf{v},$$

where M stands for the mass matrix corresponding to the discretization of $L_h^2(\Omega)$. Therefore,

$$(w \otimes z)v \approx (M^T \mathbf{z})^T \mathbf{v}\,\mathbf{w} = \mathbf{w}(M^T \mathbf{z})^T \mathbf{v}$$

and the BFGS update is the following:

$$B_{k+1} = B_k - \frac{B_k \mathbf{s}_k (M^T B_k \mathbf{s}_k)^T}{\mathbf{s}_k^T M B_k \mathbf{s}_k} + \frac{\mathbf{z}_k \mathbf{z}_k^T M}{\mathbf{s}_k^T M \mathbf{z}_k}.$$

Program: BFGS Method for the Optimization of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Tikhonov regularization parameter: ');
[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
%desiredstate=inline('x.*y','x','y');
desiredstate=inline('10*sin(5*x).*cos(4*y)','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1);
res=1; iter=0;
B=speye(n^2);   %%%%% Initial BFGS-matrix %%%%%

while res >= 1e-3
    iter=iter+1
    y=semilinear(lap,u);    %%%%% State equation %%%%%

    Y=spdiags(y,0,n^2,n^2);
    p=(lap+3*Y.^2)\(y-z);   %%%%% Adjoint solver %%%%%

    %%%%% BFGS matrix %%%%%
```

```
    if iter >= 2
    s=u-uprev;
    t=B*s;
    r=(p+alpha*u)-(pprev+alpha*uprev);
    B=B-1/(s'*t)*kron(t',t)+1/(s'*r)*kron(r',r);
    end

    delta=-B\(p+alpha*u);    %%%%% BFGS Direction %%%%%
    uprev=u; pprev=p;
    u=u+delta;               %%%%% BFGS step (no line search)%%%%%
    res=l2norm(u-uprev)
end
```

4.4 Sequential Quadratic Programming (SQP)

The Newton method studied previously provides a locally fast convergent approach for the solution of PDE-constrained optimization problems. The computational cost of Algorithm 3, however, appears to be high, since apart from the system in Step 3, many solutions of the state and adjoint equations have to be computed.

By considering the problem from a different viewpoint, namely the numerical solution of the optimality system, the sequential quadratic programming approach provides a locally quadratic convergent method for finding stationary points of PDE-constrained optimization problems.

The starting point of the SQP method is indeed the optimality system given by (3.13) or, equivalently,

$$\begin{cases} \mathscr{L}'_{(y,u)}(\bar{y},\bar{u},p) = 0, \\ -e(\bar{y},\bar{u}) = 0. \end{cases} \tag{4.32}$$

By applying a Newton method for solving the previous system of equations, we obtain the following linearized system:

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y_k,u_k,p_k) & -e'(y_k,u_k)^* \\ -e'(y_k,u_k) & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_p \end{pmatrix} = \begin{pmatrix} e_y(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e_u(y_k,u_k)^* p_k - J_u(y_k,u_k) \\ e(y_k,u_k) \end{pmatrix} \tag{4.33}$$

$$y_{k+1} = y_k + \delta_y, \qquad u_{k+1} = u_k + \delta_u, \qquad p_{k+1} = p_k + \delta_p.$$

Note that system (4.33) corresponds to the necessary and sufficient optimality condition of the following linear–quadratic problem:

$$\begin{cases} \min\limits_{(\delta_y,\delta_u)} \dfrac{1}{2}\mathscr{L}''_{(y,u)}(y_k,u_k,p_k)[(\delta_y,\delta_u)]^2 + \mathscr{L}'_{(y,u)}(y_k,u_k,p_k)(\delta_y,\delta_u), \\ \text{subject to:} \\ \qquad e_y(y_k,u_k)\delta_y + e_u(y_k,u_k)\delta_u + e(y_k,u_k) = 0. \end{cases} \tag{4.34}$$

Indeed, if $(y_k,u_k) \in V(\bar{y},\bar{u})$, where $(\bar{y},\bar{u})$ is an optimal solution to the PDE-constrained optimization problem such that $e'(\bar{y},\bar{u})$ is surjective and

$$\mathscr{L}''_{(y,u)}(\bar{y},\bar{u},p)[(w,h)]^2 \geq \kappa \|h\|_U^2,$$

for some $\kappa > 0$, and the second derivatives of J and e are Lipschitz continuous, then by proceeding as in the proof of Theorem 3.3, there exists $\delta_p \in W'$ such that system (4.33) holds.

System (4.33) is consequently well-posed and the SQP algorithm is given through the following steps:

Algorithm 5

1: Choose $(y_0,u_0,p_0) \in Y \times U \times W'$.
2: **repeat**
3: System: solve for $(\delta_y,\delta_u,\delta_\phi) \in Y \times U \times W'$

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y_k,u_k,p_k) & -e'(y_k,u_k)^* \\ -e'(y_k,u_k) & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_p \end{pmatrix} = \begin{pmatrix} e_y(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e_u(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e(y_k,u_k) \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$, $y_{k+1} = y_k + \delta_y$, $p_{k+1} = p_k + \delta p$ and $k = k+1$.
5: **until** Stopping criteria.

Remark 4.4. Since the SQP corresponds to the Newton method applied to the optimality system, it is also known as *Lagrange–Newton approach*. Local quadratic convergence of this approach can be proved similarly as for Newton's method. Moreover, a mesh independence principle can also be proved in this case if the discretization satisfies some technical assumptions [1]. The result also holds if inequality constraints are included [2].

Example 4.3. Consider the following semilinear problem:

$$\begin{cases} \min J(y,u) = \frac{1}{2}\int_\Omega |y-z_d|^2 dx + \dfrac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{subject to:} \\ \qquad \displaystyle\int_\Omega \nabla y \cdot \nabla v \, dx + \int_\Omega y^3 v \, dx = \int_\Omega uv \, dx, \quad \forall v \in H_0^1(\Omega). \end{cases}$$

By treating y an u independently, the following first and second derivatives of the Lagrangian can be computed:

$$\begin{aligned} \mathscr{L}'_{(y,u)}(y,u,p)(\delta_y,\delta_u) =& (y-z_d,\delta_y) - \int_\Omega \nabla p \cdot \nabla \delta_y \, dx - 3\int_\Omega y^2 \cdot \delta_y \cdot p \, dx \\ &+ \alpha(u,\delta_u) + \int_\Omega p\delta_u \, dx, \\ \mathscr{L}''_{(y,u)}(y,u,p)[(\delta_y,\delta_u),(w,h)] =& (\delta_y,w) - 6\int_\Omega y\cdot\delta_y\cdot w\cdot p \, dx + \alpha(\delta_u,h). \end{aligned}$$

Additionally, the action of the linearized equation operator and its adjoint are given through the following expressions:

$$\begin{aligned} \langle e'(y,u)(\delta_y,\delta_u),v\rangle_{W,W'} &= \int_\Omega \nabla\delta_y\cdot\nabla v \, dx + 3\int_\Omega y^2\delta_y v\,dx - \int_\Omega \delta_u v \, dx, \\ \langle e'(y,u)^*\delta_p,(w,h)\rangle_{Y'\times U'} &= \int_\Omega \nabla\delta_p\cdot\nabla w \, dx + 3\int_\Omega y^2\delta_p\cdot w\,dx - \int_\Omega \delta_p h \, dx. \end{aligned}$$

According to Algorithm 5, we shall solve the system

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y_k,u_k,p_k) & -e'(y_k,u_k)^* \\ -e'(y_k,u_k) & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_p \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} e_y(y_k,u_k)^*p_k - J_y(y_k,u_k) \\ e_u(y_k,u_k)^*p_k - J_u(y_k,u_k) \end{pmatrix} \\ e(y_k,u_k) \end{pmatrix},$$

which, by taking the expressions explicitly, is equivalent to

$$\begin{aligned} &\int_\Omega \delta_y w\,dx - 6\int_\Omega y_k p_k \delta_y w\,dx - \int_\Omega \nabla\delta_p\cdot\nabla w\,dx - 3\int_\Omega y_k^2\delta_p w\,dx \\ &\qquad = \textstyle\int_\Omega \nabla p_k\cdot\nabla w\,dx + 3\int_\Omega y_k^2 p_k w\,dx - \int_\Omega (y_k - z_d)w\,dx, \end{aligned}$$

$$\alpha\delta_u + \delta p = -p_k - \alpha u_k,$$

$$-\int_\Omega \nabla\delta_y\cdot\nabla v\,dx - \int_\Omega 3y_k^2\delta_y v\,dx + \int_\Omega \delta_u v\,dx = \int_\Omega \nabla y_k\cdot\nabla v\,dx + \int_\Omega y_k^3 v\,dx - \int_\Omega u_k v\,dx.$$

Finally, we obtain the iteration system:

$$\begin{pmatrix} I-6YP & 0 & -A-3Y^2 \\ 0 & \alpha I & I \\ -A-3Y^2 & I & 0 \end{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \\ \delta_p \end{pmatrix} = \begin{pmatrix} Ap_k+3Y^2p_k-y_k+z \\ -p_k-\alpha u_k \\ Ay_k+y_k^3-u_k \end{pmatrix}.$$

Program: SQP Method for the Optimization of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Regularization parameter: ');
[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); y=sparse(n^2,1); p=sparse(n^2,1);
res=1; iter=0;

while res >= 1e-3
    iter=iter+1

    %%%%% SQP step %%%%%
    Y=spdiags(y,0,n^2,n^2); P=spdiags(p,0,n^2,n^2);

    A=[speye(n^2)-6*Y.*P sparse(n^2,n^2) -lap-3*Y.^2
       sparse(n^2,n^2) alpha*speye(n^2) speye(n^2)
       -lap-3*Y.^2 speye(n^2) sparse(n^2,n^2)];

    F=[lap*p+3*Y.^2*p-y+z;-p-alpha*u;lap*y+y.^3-u];

    delta=A\F;
    uprev=u; yprev=y; pprev=p;
    y=y+delta(1:n^2);
    u=u+delta(n^2+1:2*n^2);
    p=p+delta(2*n^2+1:3*n^2);
    res=l2norm(u-uprev)+l2norm(y-yprev)+l2norm(p-pprev)
end
```

Chapter 5
Box-Constrained Problems

5.1 Problem Statement and Existence of Solutions

We consider the following type of optimization problems:

$$\begin{cases} \min J(y,u), \\ \text{subject to:} \\ \qquad e(y,u) = 0, \\ \qquad u \in U_{ad}, \end{cases} \tag{5.1}$$

where $J\colon Y \times U \longrightarrow \mathbb{R}$, $e\colon Y \times U \longrightarrow W$, with Y, U and W reflexive Banach spaces, and $U_{ad} \subset U$ is a closed, bounded, and convex set.

Further we assume that the state equation satisfies the following properties.

Assumption 5.1.

i) For each $u \in U_{ad}$, there exists a unique solution $y(u) \in Y$ to the state equation $e(y,u) = 0$.
ii) The set of solutions $\{y(u)\}$ is bounded in Y for $u \in U_{ad}$.
iii) If $u_n \rightharpoonup \hat{u}$ weakly in U, then the corresponding states $y(u_n) \rightharpoonup y(\hat{u})$ weakly in Y.

Theorem 5.1. *Let $J\colon Y \times U \longrightarrow \mathbb{R}$ be bounded from below and weakly lower semicontinuous (w.l.s.c.). Then there exists a global optimal solution for problem (5.1).*

Proof. Since J is bounded from below, there exists a sequence $\{(y_n,u_n)\} \subset \mathscr{T}_{ad} := \{(y,u) : u \in U_{ad}, e(y,u) = 0\}$ such that

$$\lim_{n\to\infty} J(y_n,u_n) = \inf_{(y,u)\in\mathscr{T}_{ad}} J(y,u).$$

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_5

Since U_{ad} is bounded and by Assumption 5.1 $\{y(u)\}$ is also bounded in Y, we may extract a subsequence $\{(y_{n_k}, u_{n_k})\}_{k\in\mathbb{N}}$ such that

$$(y_{n_k}, u_{n_k}) \rightharpoonup (\hat{y}, \hat{u}) \qquad \text{weakly in } Y \times U.$$

Since U_{ad} is convex and closed, it is weakly closed and therefore $\hat{u} \in U_{ad}$. Moreover, by Assumption 5.1 $\hat{y} = y(\hat{u})$ is a solution of the equation

$$e(\hat{y}, \hat{u}) = 0,$$

which implies that $(\hat{y}, \hat{u}) \in \mathscr{T}_{ad}$. Since the functional J is w.l.s.c. it follows that

$$J(\hat{y}, \hat{u}) \leq \liminf_{k\to\infty} J(y_{n_k}, u_{n_k}) = \inf_{(y,u)\in\mathscr{T}_{ad}} J(y, u).$$

Consequently, $(\hat{y}, \hat{u}) = (\bar{y}, \bar{u})$ corresponds to an optimal solution to (5.1). □

5.2 Optimality Conditions

Defining the solution operator

$$\begin{aligned} G\colon U &\longrightarrow Y \\ u &\longmapsto y(u) = G(u), \end{aligned}$$

we can reformulate the optimization problem in reduced form as:

$$\min_{u\in U_{ad}} f(u) = J(y(u), u). \tag{5.2}$$

Hereafter we assume that $J\colon Y \times U \longrightarrow \mathbb{R}$ and $e\colon Y \times U \longrightarrow W$ are continuously Fréchet differentiable. From Theorem 3.2, if $\bar{u} \in U_{ad}$ is a local optimal solution for (5.2), then it satisfies the variational inequality

$$f'(\bar{u})(v - \bar{u}) \geq 0,$$

for all admissible directions $v - \bar{u}$. Since U_{ad} is convex, the condition holds for all $v \in U_{ad}$.

Assuming that $e_y(\bar{y}, \bar{u})$ is a bijection and proceeding as in the proof of Theorem 3.3, we obtain the existence of an adjoint state $p \in W'$ such that the following optimality

system holds:

$$e(\bar{y},\bar{u}) = 0, \tag{5.3a}$$
$$e_y(\bar{y},\bar{u})^* p = J_y(\bar{y},\bar{u}), \tag{5.3b}$$
$$\langle J_u(\bar{y},\bar{u}) - e_u(\bar{y},\bar{u})^* p, v-\bar{u}\rangle_{U',U} \geq 0, \qquad \text{for all } v \in U_{ad}. \tag{5.3c}$$

Hereafter we will focus on the special case $U = L^2(\Omega)$, and

$$U_{ad} = \{u \in L^2(\Omega) : u_a \leq u(x) \leq u_b \text{ a.e in } \Omega\}, \tag{5.4}$$

with $u_a, u_b \in \mathbb{R}$ such that $u_a \leq u_b$.

By identifying U with its dual, inequality (5.3c) can be written as

$$(J_u(\bar{y},\bar{u}) - e_u(\bar{y},\bar{u})^* p, v-\bar{u})_U \geq 0, \qquad \forall v \in U_{ad}. \tag{5.5}$$

The box structure of the feasible set (5.4) allows to derive more detailed optimality conditions, which are the basis of the solution algorithms presented in this chapter.

Proposition 5.1. *Let $U = L^2(\Omega)$ and U_{ad} be defined by (5.4). Then inequality (5.5) is satisfied if and only if, for almost every $x \in \Omega$,*

$$(J_u(\bar{y},\bar{u})(x) - e_u(\bar{y},\bar{u})^* p(x))\,(v-\bar{u}(x)) \geq 0, \qquad \forall v \in \mathbb{R} : u_a \leq v \leq u_b. \tag{5.6}$$

Proof. Let $z(x) := J_u(\bar{y},\bar{u})(x) - e_u(\bar{y},\bar{u})^* p(x)$. Since $z \in L^2(\Omega)$, then almost every $x_0 \in \Omega$ is a Lebesgue point, i.e.,

$$\lim_{\rho \to 0^+} \frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} z(x)dx = z(x_0).$$

For ρ sufficiently small, the ball $B_\rho(x_0) \subset \Omega$ and the integral exists. Similarly, it follows that almost every $x_0 \in \Omega$ is a Lebesgue point of $\bar{u}$.

Let $x_0 \in \Omega$ be a common Lebesgue point of z and $\bar{u}$ and let $v \in [u_a, u_b]$. We define the function

$$u(x) = \begin{cases} v & \text{if } x \in B_\rho(x_0) \\ \bar{u}(x) & \text{elsewhere.} \end{cases}$$

It follows that $u \in U_{ad}$ and from inequality (5.5)

$$0 \leq \frac{1}{|B_\rho(x_0)|} \int_\Omega z(x)(u(x)-\bar{u}(x))dx = \frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} z(x)(v-\bar{u}(x))dx.$$

By taking the limit as $\rho \to 0$, it follows that

$$0 \leq z(x_0)(v-\bar{u}(x_0))$$

for almost all $x_0 \in \Omega$. □

Lemma 5.1. *Let U be a Hilbert space, $C \subset U$ be a nonempty, closed, and convex set. Then, for all $\varphi \in U$ and all $c > 0$, the following conditions are equivalent:*

$$u \in C \quad \text{and} \quad (\varphi, v-u)_U \geq 0, \quad \forall v \in C, \tag{5.7}$$

$$u = P(u - c\varphi), \tag{5.8}$$

where $P\colon U \longrightarrow U$ denotes the projection onto C.

Proof. (5.7)⇒ (5.8): Let $u_c := u - c\varphi$. Then we have that

$$(u_c - u, v-u)_U = -c(\varphi, v-u) \leq 0, \qquad \forall v \in C.$$

Consequently, $u = P(u_c)$.

(5.8)⇒ (5.7): Since $u = P(u_c) \in C$, it follows from the projection's characterization in Hilbert spaces that

$$(\varphi, v-u) = -\frac{1}{c}(u_c - u, v-u) \geq 0, \qquad \forall v \in C.$$ □

By introducing the multiplier

$$\lambda := J_u(\bar{y}, \bar{u}) - e_u(\bar{y}, \bar{u})^* p,$$

inequality (5.5) can be written as

$$\bar{u} = P_{U_{ad}}(\bar{u} - c\lambda).$$

Moreover, under the conditions of Proposition 5.1, inequality (5.6) can be written as

$$\bar{u}(x) = P_{[u_a, u_b]}(\bar{u}(x) - c\lambda(x)) \qquad \text{a.e in } \Omega,$$

where $P_{[u_a,u_b]}\colon \mathbb{R} \longrightarrow \mathbb{R}$ denotes the projection onto the interval $[u_a, u_b]$. By decomposing λ into its positive and negative parts, i.e.,

$$\lambda = \lambda_a - \lambda_b,$$

with $\lambda_a(x) = \max(0, \lambda(x))$ and $\lambda_b(x) = -\min(0, \lambda(x))$, inequality (5.6) can be written as:

$$J_u(\bar{y}, \bar{u}) - e_u(\bar{y}, \bar{u})^* p = \lambda_a - \lambda_b \qquad \text{a .e. in } \Omega, \tag{5.9a}$$

$$\lambda_a(x)(v-\bar{u}(x)) \geq 0 \qquad \text{a .e. in } \Omega, \forall v \in [u_a, u_b], \tag{5.9b}$$

$$\lambda_b(x)(v-\bar{u}(x)) \leq 0 \qquad \text{a .e. in } \Omega, \forall v \in [u_a, u_b], \tag{5.9c}$$

$$\lambda_a(x) \geq 0, \lambda_b(x) \geq 0 \qquad \text{a .e. in } \Omega, \tag{5.9d}$$

$$u_a \leq \bar{u} \leq u_b \qquad \text{a .e. in } \Omega. \tag{5.9e}$$

By taking $v = u_a$ in (5.9b) and $v = u_b$ in (5.9c), we obtain from (5.9d) and (5.9e) that

$$\lambda_a(u_a - \bar{u}) = \lambda_b(u_b - \bar{u}) = 0 \qquad \text{a .e. in } \Omega.$$

Consequently, inequality (5.6) is equivalent to the following *complementarity problem:*

$$u_a \leq \bar{u}(x) \leq u_b \qquad \text{a .e. in } \Omega, \tag{5.10a}$$

$$\lambda_a(x) \geq 0, \lambda_b(x) \geq 0 \qquad \text{a .e. in } \Omega, \tag{5.10b}$$

$$\lambda_a(x)(u_a - \bar{u}(x)) = \lambda_b(x)(u_b - \bar{u}(x)) = 0 \qquad \text{a .e. in } \Omega. \tag{5.10c}$$

The alternative formulations of the variational inequality (5.6) as a projection formula or as a complementarity system give rise to the development of different numerical strategies for the solution of problem (5.1). This will be the subject of the next sections of this chapter.

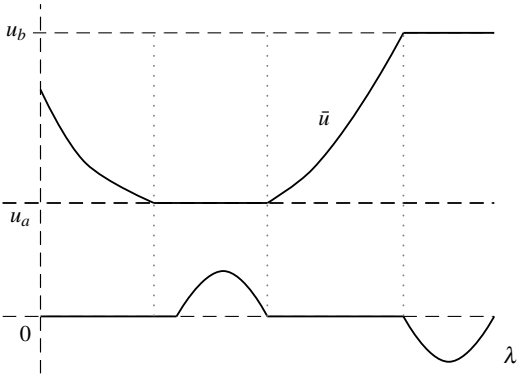

Fig. 5.1 Example of complementarity: the multiplier λ may take values different from 0 only on the sectors where the control is active

Second order sufficient optimality conditions for problem (5.2) follow from Theorem 3.4, if the condition

$$f''(\bar{u})[h]^2 \geq \delta \|h\|_U^2, \quad \forall h \in U, \tag{5.11}$$

holds for some $\delta > 0$. Moreover, a similar result than Theorem 3.3 for $(\bar{y}, \bar{u}, p)$ satisfying the optimality system (5.3) may be obtained.

Condition (5.11) may be however too strong, since it involves all possible directions $h \in U$. In the case $U = L^2(\Omega)$ and $U_{ad} = \{u \in L^2(\Omega) : u_a \leq u(x) \leq u_b \text{ a.e. in } \Omega\}$, the sufficient condition can be weakened by considering it on the cone of critical directions defined by:

$$C(\bar{u}) = \left\{ v \in L^2(\Omega) : \begin{array}{l} v(x) \geq 0 \ \text{ if } \bar{u}(x) = u_a \\ v(x) \leq 0 \ \text{ if } \bar{u}(x) = u_b \end{array} \right\}.$$

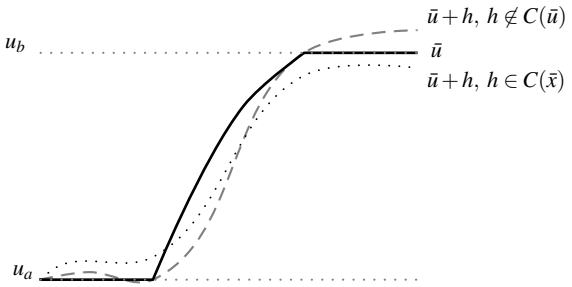

Fig. 5.2 Example of critical cone directions

The second order sufficient condition (SSC) is then given by

$$\mathscr{L}''_{(y,u)}(\bar{y}, \bar{u})[(w,h)]^2 \geq \delta \|h\|^2_{L^2}$$

for all $(w,h) \in Y \times C(\bar{u})$ that satisfy the equation

$$e_y(\bar{y}, \bar{u})w + e_u(\bar{y}, \bar{u})h = 0.$$

Example 5.1. Consider the following quadratic problem in $\mathbb{R}^2$:

$$\min_{x \in [0,1]^2} f(x) = x_1^2 + 4x_1x_2 + x_2^2.$$

The unique stationary point is given by the solution of the system

$$\nabla f(x)^T (y - \bar{x}) = \begin{pmatrix} 2x_1 + 4x_2 \\ 4x_1 + 2x_2 \end{pmatrix}^T (y - \bar{x}) \geq 0, \qquad \forall y \in [0,1]^2$$

which is given by $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. The Hessian is given by $H = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$, whose eigenvalues are $\lambda_1 = -1$, $\lambda_2 = 3$. H is clearly not positive definite. However, if we define the critical cone

$$C(\bar{x}) = \left\{ v \in \mathbb{R}^2 : v_i \begin{cases} \geq 0 & \text{if } \bar{x}_i = 0 \\ \leq 0 & \text{if } \bar{x}_i = 1 \end{cases} \right\},$$

it can be easily verified that

$$\begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \geq \|v\|^2_{\mathbb{R}^2},$$

for all $v \in C(\bar{x})$. Therefore, $\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ satisfies the (SSC).

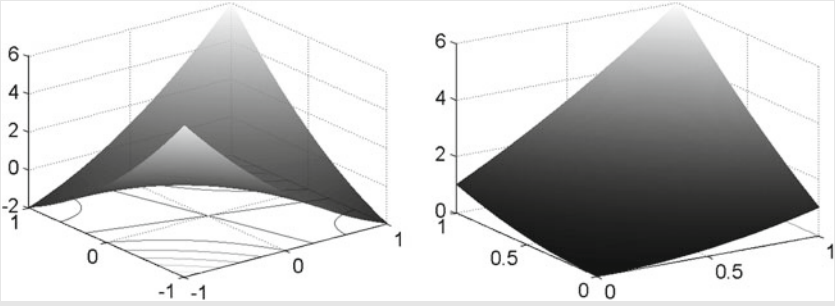

Fig. 5.3 Quadratic form on $[-1,1]^2$ (*left*) and restricted to the critical cone (*right*)

5.3 Projection Methods

The main idea of projection methods consists in making use of a descent direction computed for the problem without bound constraints and then project the new iterate onto the feasible set defined by the inequality constraints.

To be more specific, consider the reduced problem

$$\min_{u \in U_{ad}} f(u). \tag{5.12}$$

If we have some iterate u_k and compute some direction d_k, then the next iterate given by the related projection method, will be:

$$u_{k+1} = P_{U_{ad}}(u_k + \alpha_k d_k), \tag{5.13}$$

where $P_{U_{ad}}$ denotes the projection onto U_{ad} and $\alpha_k \in (0,1)$ is a line search parameter.

From equation (5.13) it can be observed that in order to make the update, a projection has to be computed. This may not be easy in general. Moreover, a modified line search strategy has to be implemented for the choice of α_k.

The general projected descent algorithm is given through the following steps:

Algorithm 6 Projected descent method

1: Choose $u_0 \in U_{ad}$ and set $k = 0$.
2: **repeat**
3: Compute a direction d_k for the problem on U.
4: Choose α_k by a projected line search rule such that

$$f(P_{U_{ad}}(u_k + \alpha_k d_k)) < f(u_k).$$

5: Set

$$u_{k+1} = P_{U_{ad}}(u_k + \alpha_k d_k) \qquad \text{and} \qquad k = k+1.$$

6: **until** stopping criteria.

Proposition 5.2. *Let U be a Hilbert space and let $f\colon U \longrightarrow \mathbb{R}$ be continuously Fréchet differentiable on a neighborhood of the closed convex set U_{ad}. Let $u_k \in U_{ad}$, $d_k = -\nabla f(u_k)$ and assume that ∇f is θ-order Hölder-continuous with modulus $L > 0$, for some $\theta \in (0,1]$, on the set*

$$\left\{(1-t)u_k + tP_{U_{ad}}(u_k^\alpha) : 0 \le t \le 1\right\},$$

where $u_k^\alpha := u_k - \alpha\nabla f(u_k)$. Then there holds

$$f(P_{U_{ad}}(u_k^\alpha)) - f(u_k) \le -\frac{1}{\alpha}\left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U^2 + L\left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U^{1+\theta}.$$

Proof. From the mean value theorem if follows that

$$\begin{aligned} f(P_{U_{ad}}(u_k^\alpha)) - f(u_k) &= (\nabla f(v_k^\alpha), P_{U_{ad}}(u_k^\alpha) - u_k)_U \\ &= \underbrace{(\nabla f(u_k), P_{U_{ad}}(u_k^\alpha) - u_k)_U}_{\text{a)}} + \underbrace{(\nabla f(v_k^\alpha) - \nabla f(u_k), P_{U_{ad}}(u_k^\alpha) - u_k)}_{\text{b)}} \end{aligned}$$

for some $v_k^\alpha = (1-t)u_k + tP_{U_{ad}}(u_k^\alpha)$.

For the term a), we obtain:

$$\begin{aligned}(\nabla f(u_k), P_{U_{ad}}(u_k^\alpha) - u_k)_U &= \frac{1}{\alpha}(u_k - u_k^\alpha, P_{U_{ad}}(u_k^\alpha) - u_k)_U \\ &= \frac{1}{\alpha}(\overbrace{P_{U_{ad}}(u_k)}^{=u_k} - P_{U_{ad}}(u_k^\alpha), P_{U_{ad}}(u_k^\alpha) - P_{U_{ad}}(u_k))_U \\ &\quad \underbrace{-\frac{1}{\alpha}(u_k^\alpha - P_{U_{ad}}(u_k^\alpha), P_{U_{ad}}(u_k^\alpha) - P_{U_{ad}}(u_k))_U}_{\leq 0} \\ &\leq -\frac{1}{\alpha}\left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U^2.\end{aligned}$$

For the term b),

$$\begin{aligned}(\nabla f(v_k^\alpha) - \nabla f(u_k), P_{U_{ad}}(u_k^\alpha) - u_k) &\leq \|\nabla f(v_k^\alpha) - \nabla f(u_k)\|_U \left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U \\ &\leq L\|v_k^\alpha - u_k\|_U^\theta \left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U \\ &\leq L\left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U^{1+\theta},\end{aligned}$$

since $\left\|v_k^\alpha - u_k\right\|_U \leq \left\|P_{U_{ad}}(u_k^\alpha) - u_k\right\|_U$. □

It then follows from Proposition 5.2 that the projected gradient direction is a descent direction for problem (5.12). A modified Armijo rule is given by: Choose the largest $\alpha_k \in \left\{1, \frac{1}{2}, \frac{1}{4}, \ldots\right\}$ for which

$$f(P_{U_{ad}}(u_k - \alpha_k \nabla f(u_k))) - f(u_k) \leq -\frac{\gamma}{\alpha_k}\left\|P_{U_{ad}}(u_k - \alpha_k \nabla f(u_k)) - u_k\right\|_U^2, \tag{5.14}$$

where $\gamma \in (0,1)$ is a given constant.

Thanks to Lemma 5.1, we can choose as stopping criteria

$$\left\|u_k - P_{U_{ad}}(u_k - \nabla f(u_k))\right\|_U < \varepsilon$$

for some $0 < \varepsilon \ll 1$. In the case of PDE-constrained optimization problems we can express the stopping criteria with help of the multiplier λ as

$$\left\|u_k - P_{U_{ad}}(u_k - c\lambda_k)\right\|_U < \varepsilon, \quad \text{for some } c > 0.$$

Theorem 5.2. *Let U be a Hilbert space, $f\colon U \longrightarrow \mathbb{R}$ be continuously Fréchet differentiable and $U_{ad} \subset U$ be nonempty, closed, and convex. Assume that $f(u_k)$ is bounded from below for the iterates generated by the gradient projected method with line search (5.14).*

If for some $\theta > 0$ and $\rho > 0$, ∇f is θ-order Hölder continuous on

$$N_0^\rho = \{u + d : f(u) \leq f(u_0), \|d\|_U \leq \rho\},$$

then

$$\lim_{k\to\infty} \|u_k - P_{U_{ad}}(u_k - \nabla f(u_k))\|_U = 0.$$

Proof. For the proof we refer to [30, p. 108]. □

For the particular case $U = L^2(\Omega)$ and $U_{ad} = \{u \in L^2(\Omega) : u_a \leq u(x) \leq u_b \text{ a. e. in } \Omega\}$, the projection formula is given, thanks to Proposition 5.1, by

$$P_{U_{ad}}(u)(x) = P_{[u_a,u_b]}(u(x)) = \max(u_a, \min(u(x), u_b)).$$

Remark 5.1. A mesh independence result for the projected gradient method is obtained in [40] for ordinary differential equation (ODE) optimal control problems. In this case, the iteration indexes are defined by:

$$i(\varepsilon) := \min\{k \in \mathbb{N} : \|u_k - u_{k-1}\|_U \leq \varepsilon\},$$
$$i_h(\varepsilon) := \min\{k \in \mathbb{N} : \|u_k^h - u_{k-1}^h\|_{U_h} \leq \varepsilon\}.$$

Under suitable assumptions on the discretization, the result establishes that for all $\varepsilon, \rho > 0$ there is $h_{\varepsilon,\rho}$ such that if $h < h_{\varepsilon,\rho}$, then

$$i(\varepsilon + \rho) \leq i_h(\varepsilon) \leq i(\varepsilon).$$

Remark 5.2. The application of projection methods considering other type of directions $d_k = -H_k^{-1}\nabla f(u_k)$ is by no means standard. For Newton directions

$$d_k = -(\nabla^2 f(u_k))^{-1}\nabla f(u_k),$$

for instance, the application of Algorithm 6 may not lead to descent in the objective function. To solve this problem in $\mathbb{R}^m$, the reduced Hessian

$$(\nabla_R^2 f(u))_{ij} = \begin{cases} \delta_{ij} & \text{if } i \in A(u) \text{ or } j \in A(u) \\ (\nabla^2 f(u))_{ij} & \text{otherwise} \end{cases}$$

where $A(u)$ denotes the set of active indexes, may be used instead of the full second order matrix (see, e.g., [37]). In [41] an infinite-dimensional variant is proposed for solving ODE control problems.

Program: Projected Gradient Method for the Optimization of the Poisson Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Tikhonov regularization parameter: ');
ua=input('Lower bound: '); ub=input('Upper bound: ');
Armijo=input('Line search (1=yes): ');
[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); res=1; iter=0;

while res >= 1e-3
    iter=iter+1
    y=lap\u;         %%%%% State equation %%%%%
    p=lap\(y-z);     %%%%% Adjoint solver %%%%%

        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        % Armijo line search
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        countarm=0;
        beta=1;
        gradcost=l2norm(max(ua,min(u-beta*(p+alpha*u),ub))-u);
        cost1=1/2*l2norm(y-z)^2+alpha/2*l2norm(u)^2;

         if Armijo==1
             armijo=1e5;
             while armijo > -1e-4/beta*gradcost^2
             beta=1/2^(countarm);

             uinc=max(ua,min(u-beta*(p+alpha*u),ub));
             yinc=lap\uinc;

             cost2=1/2*l2norm(yinc-z)^2+alpha/2*l2norm(uinc)^2;
             armijo=cost2-cost1;
```

```
            gradcost=l2norm(max(ua,min(u-beta*(p+alpha*u),ub))-u);
            countarm=countarm+1
            end
        end
    uprev=u;
    %%%% Projected gradient step %%%%
    u=max(ua,min(u-beta*(p+alpha*u),ub));
    res=l2norm(u-uprev)
end
```

5.4 Primal Dual Active Set Algorithm (PDAS)

Consider the following optimization problem:

$$\begin{cases} \min\ J(y,u), \\ \text{subject to:} \\ \qquad e(y,u)=0, \\ \qquad u \le u_b \text{ a.e. in } \Omega, \end{cases} \tag{5.15}$$

where $U = L^2(\Omega)$. Since in this case $u_a = -\infty$, the optimality condition (5.9) can be written in the following way:

$$\bar{u}(x) = P_{(-\infty,u_b]}(\bar{u}(x) - c\lambda(x)) \qquad \text{a.e. in } \Omega, \quad \forall c > 0, \tag{5.16}$$

which, since $\lambda(x) = -\lambda_b(x)$, implies that

$$\bar{u}(x) - P_{(-\infty,u_b]}(\bar{u}(x) + c\lambda_b(x)) = 0 \qquad \text{a.e. in } \Omega, \quad \forall c > 0.$$

Thanks to (5.16),

$$\begin{aligned} \lambda_b(x) &= \frac{1}{c}\left[\bar{u}(x) + c\lambda_b(x) - \min(u_b, \bar{u}(x) + c\lambda_b(x))\right] \\ &= \begin{cases} \frac{1}{c}(\bar{u}(x) + c\lambda_b(x) - u_b) & \text{if } \ u_b \le \bar{u}(x) + c\lambda_b(x) \\ 0 & \text{if } \ \text{not} \end{cases} \\ &= \frac{1}{c}\max(0, \bar{u}(x) + c\lambda_b(x) - u_b). \end{aligned}$$

Altogether, we obtain the following optimality system:

$$e(\bar{y},\bar{u}) = 0, \tag{5.17a}$$
$$e_y(\bar{y},\bar{u})^* p = J_y(\bar{y},\bar{u}), \tag{5.17b}$$
$$e_u(\bar{y},\bar{u})^* p = J_u(\bar{y},\bar{u}) + \lambda_b, \tag{5.17c}$$
$$\lambda_b(x) = c\max(0,\bar{u}(x) + \frac{1}{c}\lambda_b(x) - u_b), \qquad \text{for all } c > 0. \tag{5.17d}$$

Let us define the active and inactive sets at the solution $(\bar{y},\bar{u})$ by $A = \{x : \bar{u}(x) = u_b\}$ and $I = \Omega \setminus A$, respectively. The main idea of the PDAS strategy consists in considering equation (5.17d) in a iterative scheme:

$$\lambda_b^{k+1}(x) = c\max(0, u_k(x) + \frac{1}{c}\lambda_b^k(x) - u_b). \tag{5.18}$$

Considering, in addition, the complementarity relations given by (5.10), Eq. (5.18) leads to the following prediction of active and inactive sets at the next iteration:

$$A_{k+1} = \left\{ x | u_k(x) + \frac{\lambda_b^k(x)}{c} > u_b \right\}, \qquad I_{k+1} = \left\{ x | u_k(x) + \frac{\lambda_b^k(x)}{c} \leq u_b \right\}.$$

The complete algorithm is then given through the following steps:

Algorithm 7 Primal-dual active set method (PDAS)

1: Choose u_0, y_0 , λ_b^0 and $c > 0$. Set $k = 0$.
2: **repeat**
3: Determine the following subsets of Ω:

$$A_{k+1} = \left\{ x | u_k(x) + \frac{\lambda_b^k(x)}{c} > b \right\}, \qquad I_{k+1} = \Omega \setminus A_{k+1}.$$

4: Solve the following system:

$$(OS)_{k+1} = \begin{cases} e(y_{k+1},u_{k+1}) = 0, \\ e_y(y_{k+1},u_{k+1})^* p_{k+1} = J_y(y_{k+1},u_{k+1}), \\ u_{k+1} = u_b, \qquad \text{on } A_{k+1}, \\ \lambda_b^{k+1} = 0, \qquad \text{on } I_{k+1}, \\ e_u(y_{k+1},u_{k+1})^* p_{k+1} = J_u(y_{k+1},u_{k+1}) + \lambda_b^{k+1}. \end{cases}$$

and set $k = k + 1$.
5: **until** stopping criteria.

For the subsequent analysis of the PDAS algorithm we consider the following particular structure of the cost functional and the state equation:

$$J(y,u) = g(y) + \frac{\alpha}{2}\|u\|_{L^2}^2, \qquad e(y,u) = E(y) - u,$$

where $E\colon Y \longrightarrow L^2(\Omega)$ is continuously differentiable, with $E'(\bar{y})$ continuously invertible, and $g : Y \to \mathbb{R}$ is continuously differentiable. The system to be solved in each primal–dual iteration is then given by:

$$E(y_{k+1}) = u_{k+1}, \tag{5.19a}$$
$$E'(y_{k+1})^* p_{k+1} = g_y(y_{k+1}), \tag{5.19b}$$
$$\alpha u_{k+1} + \lambda_b^{k+1} + p_{k+1} = 0 \quad \text{a.e. in } \Omega, \tag{5.19c}$$
$$u_{k+1} = u_b \quad \text{on } A_{k+1}, \tag{5.19d}$$
$$\lambda_{k+1} = 0 \quad \text{on } I_{k+1}. \tag{5.19e}$$

Considering two consecutive iterates of the algorithm and choosing $c = \alpha$, it follows that

$$E(y_{k+1}) - E(y_k) = u_{k+1} - u_k = -\frac{1}{\alpha}\chi_{I_{k+1}}(p_{k+1} - p_k) + R^k,$$

where χ_C denotes the indicator function of a set C, and

$$R^k := \begin{cases} 0 & \text{on } A_k \cap A_{k+1}, \\ u_b + \frac{1}{\alpha}p_k = u_b - u_k < 0 & \text{on } I_k \cap A_{k+1}, \\ \frac{1}{\alpha}\lambda_b^k = -\frac{1}{\alpha}p_k - u_b \le 0 & \text{on } A_k \cap I_{k+1}, \\ 0 & \text{on } I_k \cap I_{k+1}. \end{cases}$$

For the global convergence of the PDAS method we define the following merit functional:

$$M(u,\lambda_b) = \alpha^2 \int_\Omega \max(0, u - u_b)^2 dx + \int_{A^+(u)} \min(0, \lambda_b)^2 dx,$$

where $A^+(u) = \{x : u \ge u_b\}$.

Theorem 5.3. *If there exists a constant $\rho \in [0,\alpha)$ such that*

$$\|p_{k+1} - p_k\|_{L^2} < \rho \left\|R^k\right\|_{L^2} \qquad \text{for every } k = 1,2,\ldots \tag{5.20}$$

then

$$M(u_{k+1}, \lambda_b^{k+1}) \le \alpha^{-2}\rho^2 M(u_k, \lambda_b^k)$$

for every $k = 1, 2, \ldots$

Proof. From (5.19) it follows that

$$\begin{aligned} -\lambda_b^{k+1} = \lambda_{k+1} = p_{k+1} + \alpha u_{k+1} \qquad & \text{on } A_{k+1}, \\ u_{k+1} + \frac{1}{\alpha} p_{k+1} = 0 \qquad & \text{on } I_{k+1}, \end{aligned}$$

which using the algorithm PDAS implies the following estimates:

On $A_{k+1} = \left\{ x | u_k + \frac{\lambda_b^k}{\alpha} > u_b \right\}$:

$$\begin{aligned} \lambda_{k+1} &= p_{k+1} - p_k + p_k + \alpha u_b \\ &= p_{k+1} - p_k + \begin{cases} \lambda_k & \text{on } A_{k+1} \cap A_k, \\ \alpha(u_b - u_k) & \text{on } A_{k+1} \cap I_k. \end{cases} \end{aligned}$$

Since $\lambda_k(x) < \alpha(u_k(x) - u_b)$ on A_{k+1} and $u_k(x) = u_b$ on A_k, it follows that $\lambda_k(x) < 0$ on $A_{k+1} \cap A_k$. In addition, since $\alpha(u_b - u_k)(x) < \lambda_b^k(x)$ on A_{k+1} and $\lambda_b^k(x) = 0$ on I_k, we get that $\alpha(u_b - u_k)(x) < 0$ on $A_{k+1} \cap I_k$. Consequently,

$$\begin{aligned} \lambda_{k+1}(x) = -\lambda_b^{k+1}(x) < p_{k+1}(x) - p_k(x) \qquad & \text{on } A_{k+1} \\ \Rightarrow \left| \min(0, \lambda_b^{k+1}(x)) \right| \leq |p_{k+1}(x) - p_k(x)| \qquad & \text{on } A_{k+1}. \end{aligned}$$

On $I_{k+1} = \left\{ x | u_k + \frac{\lambda_b^k}{\alpha} \leq u_b \right\}$:

$$\begin{aligned} u_{k+1}(x) - u_b &= \frac{1}{\alpha}(-p_{k+1} + p_k - p_k)(x) - u_b \\ &= \frac{1}{\alpha}\left(p_k(x) - p_{k+1}(x)\right) + \begin{cases} \frac{1}{\alpha}\lambda_b^k & \text{on } I_{k+1} \cap A_k, \\ u_k - u_b & \text{on } I_{k+1} \cap I_k. \end{cases} \end{aligned}$$

Since $\frac{1}{\alpha}\lambda_b^k(x) \leq u_b - u_k(x)$ on I_{k+1} and $u_b = u_k(x)$ on A_k, it follows that

$$\frac{1}{\alpha}\lambda_b^k(x) \leq 0 \qquad \text{on } A_k \cap I_{k+1}.$$

Also since $u_k(x) - u_b \leq -\frac{\lambda_b^k(x)}{\alpha}$ on I_{k+1} and $\lambda_b^k(x) = 0$ on I_k, we get that

$$u_k(x) - u_b \leq 0 \qquad \text{on } I_k \cap I_{k+1}.$$

Consequently,

$$u_{k+1}(x) - u_b \le \frac{1}{\alpha}(p_k(x) - p_{k+1}(x)) \qquad \text{on } I_{k+1}$$

$$\Rightarrow \left|\max(0, u_{k+1}(x) - u_b)\right| \le \frac{1}{\alpha}|p_{k+1}(x) - p_k(x)| \quad \text{on } I_{k+1}.$$

Since from (5.19) it also follows that

$$\lambda_b^{k+1}(x) = 0 \quad \text{on } I_{k+1} \qquad \text{and} \qquad u_{k+1}(x) = u_b \quad \text{on } A_{k+1},$$

we obtain that

$$\begin{aligned} M(u_{k+1}, \lambda_b^{k+1}) &\le \alpha^2 \int_{I_{k+1}} \max(0, u_{k+1} - u_b)^2 dx + \int_{A_{k+1}} \min(0, \lambda_b^{k+1})^2 dx \\ &\le \int_\Omega |p_{k+1} - p_k|^2 dx, \end{aligned}$$

which, thanks to (5.20), yields

$$M(u_{k+1}, \lambda_b^{k+1}) < \rho^2 \left\|R^k\right\|_U^2. \tag{5.21}$$

Additionally, from the structure of R^k, it follows that

$$\begin{aligned} &|R^k(x)| \le \max(0, u_k(x) - u_b) && \text{on } A_{k+1} \cap A_k, \\ &|R^k(x)| \le 0 && \text{on } A_{k+1} \cap I_k, \\ &|R^k(x)| \le \max(0, -\frac{1}{\alpha}\lambda_b^k) = \frac{1}{\alpha}|\min(0, \lambda_b^k)| && \text{on } I_{k+1} \cap A_k, \\ &|R^k(x)| \le 0 && \text{on } I_{k+1} \cap I_k. \end{aligned}$$

Integrating over Ω we get that

$$\begin{aligned} \left\|R^k\right\|_{L^2}^2 &\le \int_\Omega \max(0, u_k - u_b)^2 + \frac{1}{\alpha^2} \int_{A_k \cap I_{k+1}} |\min(0, \lambda_b^k)|^2 \\ &\le \alpha^{-2} M(u_k, \lambda_b^k). \end{aligned} \tag{5.22}$$

Consequently, from (5.21) and (5.22) the result follows. □

Corollary 5.1. *Under the hypothesis of Theorem 5.3, there exists* $(\bar{y}, \bar{u}, p, \lambda_b) \in Y \times U \times Y \times U$ *such that*

$$\lim_{k \to \infty} (y_k, u_k, p_k, \lambda_b^k) = (\bar{y}, \bar{u}, p, \lambda_b)$$

and $(\bar{y}, \bar{u}, p, \lambda_b)$ *satisfies the optimality system:*

$$E(\bar{y}) = \bar{u}, \tag{5.23a}$$
$$E'(\bar{y})^* p = g_y(\bar{y}), \tag{5.23b}$$
$$\alpha\bar{u} + p + \lambda_b = 0, \tag{5.23c}$$
$$\lambda_b = \max(0, \lambda_b + \alpha(\bar{u} - u_b)). \tag{5.23d}$$

Proof. From equations (5.20) and (5.22) we obtain that

$$\|p_{k+1} - p_k\|_{L^2}^2 < \left(\frac{\rho}{\alpha}\right)^2 M(u_k, \lambda_b^k),$$

which, thanks to (5.21), implies that

$$\|p_{k+1} - p_k\|_{L^2} < \left(\frac{\rho}{\alpha}\right) \rho \left\|R^{k-1}\right\|_{L^2},$$

and by induction

$$\|p_{k+1} - p_k\|_{L^2} < \left(\frac{\rho}{\alpha}\right)^k \rho \left\|R^0\right\|_{L^2}, \qquad \text{for } k = 1, 2, \ldots$$

Consequently, there exists some $p \in U$ such that

$$\lim_{k\to\infty} p_k = p \qquad \text{in } U.$$

Since, for $k \geq 1$,

$$A_{k+1} = \{x : -p_k(x) > \alpha u_b\}, \qquad I_{k+1} = \{x : -p_k(x) \leq \alpha u_b\},$$

it follows that

$$\begin{aligned} \lambda_b^{k+1} &= \begin{cases} -\alpha u_b - p_{k+1} & \text{on } A_{k+1}, \\ 0 & \text{on } I_{k+1}, \end{cases} \\ &= \max(0, -p_k - \alpha u_b) + \chi_{A_{k+1}} \cdot (p_k - p_{k+1}). \end{aligned}$$

Since $p_{k+1} - p_k \xrightarrow{k\to\infty} 0$ and $p_k \xrightarrow{k\to\infty} p$, the continuity of the max function from $L^2(\Omega) \to L^2(\Omega)$ implies that

$$\lim_{k\to\infty} \lambda_b^{k+1} = \lambda_b = \max(0, -p - \alpha u_b).$$

Finally, from the optimality condition

$$\alpha u_{k+1} + \lambda_b^{k+1} + p_{k+1} = 0,$$

there exists $\bar{u}$ such that $\lim_{k\to\infty} u_k = \bar{u}$. Thanks to the properties of the operator $E: Y \longrightarrow L^2(\Omega)$, we may pass to the limit in the PDAS optimality system and obtain (5.23). □

Remark 5.3.

1. If the operator $E: Y \longrightarrow L^2(\Omega)$ is linear, then the condition

$$\left\|E^{-1}\right\|^2_{\mathscr{L}(L^2(\Omega),Y)} < \alpha$$

   is sufficient for (5.20) to hold.
2. A frequently used stopping criteria for the PDAS algorithm is given by:

$$A_{k+1} = A_{k.}$$

   In the linear case, this choice is theoretically justified (see [6]).

5.5 Semismooth Newton Methods (SSN)

An alternative approach for the solution of system (5.17) consists in considering it as an operator equation

$$F(x) = 0, \tag{5.24}$$

where

$$\begin{aligned} F: X &\longrightarrow Z \\ x &\longmapsto F(x), \end{aligned}$$

with X and Z Banach spaces. If F would be Fréchet differentiable, a classical Newton type method could be used for solving (5.24). In the case of system (5.17), however, the max function is not Fréchet differentiable and a standard Newton scheme cannot be applied. The following question then arises: Is it possible to define a weaker differentiability notion for such a function such that a Newton type iterative scheme can be stated?

Definition 5.1. Let D be an open subset of a Banach space X. The mapping $F: D \subset X \to Z$ is called Newton differentiable on the open subset $V \subset D$ if there exists a generalized

derivative $G : V \to \mathscr{L}(X,Z)$ such that

$$\lim_{h\to 0} \frac{1}{\|h\|_X} \|F(x+h) - F(x) - G(x+h)h\|_Z = 0, \tag{5.25}$$

for every $x \in V$.

Example 5.2. Consider the absolute value function

$$\begin{aligned} f = |\cdot| \colon \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto |x|. \end{aligned}$$

The function is not differentiable at 0. However, by using the generalized derivative

$$g(x) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0, \end{cases}$$

we obtain for the case $x = 0$:

i) if $h > 0$: $\quad \big||x+h| - |x| - |h|\big| = 0,$
ii) if $h < 0$: $\quad \big||x+h| - |x| + |h|\big| = |-x-h-x+h| = 0.$

Consequently,

$$\lim_{h\to 0} \frac{1}{|h|} |f(x+h) - f(x) - g(x+h)h| = 0$$

and $|\cdot|$ is Newton differentiable.

Theorem 5.4. *Let $\bar{x}$ be a solution to (5.24), with F Newton differentiable in an open neighborhood V containing $\bar{x}$. If*

$$\big\|G(x)^{-1}\big\|_{\mathscr{L}(Z,X)} \leq C, \tag{5.26}$$

for some constant $C > 0$ and all $x \in V$, then the semismooth Newton (SSN) iteration

$$x_{k+1} = x_k - G(x_k)^{-1} F(x_k) \tag{5.27}$$

converges superlinearly to $\bar{x}$, provided that $\|x_0 - \bar{x}\|_X$ is sufficiently small.

Proof. Considering that $F(\bar{x}) = 0$ and the iterates given by (5.27) it follows that

$$\begin{aligned} \|x_{k+1} - \bar{x}\|_X &= \big\|x_k - G(x_k)^{-1} F(x_k) - \bar{x}\big\|_X \\ &= \big\|G(x_k)^{-1}(F(\bar{x}) - F(x_k) - G(x_k)(\bar{x} - x_k))\big\|_X \\ &\leq C \|F(x_k) - F(\bar{x}) - G(x_k)(x_k - \bar{x})\|_Z . \end{aligned} \tag{5.28}$$

Thanks to the Newton differentiability it then follows, for $\rho = \frac{1}{2C}$, that there exists a ball $B_\delta(\bar{x})$ such that if $x_k \in B_\delta(\bar{x})$, then

$$\|x_{k+1} - \bar{x}\|_X \leq C\rho \|x_k - \bar{x}\|_X = \frac{1}{2} \|x_k - \bar{x}\|_X .$$

Consequently, if $\|x_0 - \bar{x}\|_X < \delta$ then $x_k \in B_\delta(\bar{x})$, $\forall k \geq 1$, and

$$\lim_{k\to\infty} \|x_k - \bar{x}\|_X = 0.$$

Moreover, from (5.28) and the Newton differentiability, we get that

$$\lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|_X}{\|x_k - \bar{x}\|_X} \leq \lim_{k\to\infty} C \frac{\|F(x_k) - F(\bar{x}) - G(x_k)(x_k - \bar{x})\|_Z}{\|x_k - \bar{x}\|_X} = 0,$$

which implies the superlinear convergence rate. □

Proposition 5.3. *The mapping $max(0,\cdot)\colon L^q(\Omega) \longrightarrow L^p(\Omega)$ with $1 \leq p < q \leq \infty$ is Newton differentiable on $L^q(\Omega)$ with generalized derivative*

$$G_{max} : L^q(\Omega) \to \mathscr{L}(L^q(\Omega), L^p(\Omega))$$

given by:

$$G_{max}(y)(x) = \begin{cases} 1 & \text{if } y(x) > 0, \\ 0 & \text{if } y(x) \leq 0. \end{cases}$$

Proof. For the detailed proof we refer to [34, p. 237]. □

Considering system (5.23) with $E \in \mathscr{L}(Y,U)$, the system to be solved is given by

$$F(y,u,p,\lambda_b) = \begin{pmatrix} Ey - u \\ E^*p - g_y(y) \\ \alpha u + p + \lambda_b \\ \lambda_b - \max(0, \lambda_b + \alpha(u - u_b)) \end{pmatrix} = 0$$

and its generalized derivative by

$$G(y,u,p,\lambda_b)[(\delta_y,\delta_u,\delta_p,\delta_\lambda)] = \begin{pmatrix} E\delta_y - \delta_u \\ E^*\delta_p - g_{yy}(y)\delta_y \\ \alpha\delta_u + \delta_p + \delta_\lambda \\ \delta_\lambda - \chi_A(\delta_\lambda + \alpha\delta_u) \end{pmatrix},$$

where χ_A stands for the indicator function of the active set

$$A = \{x : \lambda_b + \alpha(u - u_b) \geq 0\}.$$

The Newton step is then given through the solution of the following system:

$$\begin{pmatrix} E & -I & 0 & 0 \\ -g_{yy}(y) & 0 & E^* & 0 \\ 0 & \alpha I & I & I \\ 0 & -\alpha\chi_A & 0 & \chi_I \end{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \\ \delta_p \\ \delta_\lambda \end{pmatrix} = - \begin{pmatrix} Ey-u \\ E^*p - g_y(y) \\ \alpha u + p + \lambda_b \\ \lambda_b - \max(0, \lambda_b + \alpha(u-u_b)) \end{pmatrix},$$

where χ_I denotes the indicator function of the inactive set $I = \Omega \backslash A$.

Theorem 5.5. *Consider system (5.23) with $E \in \mathscr{L}(Y, L^2(\Omega))$ and $g(y) = \frac{1}{2}\int_\Omega |y - z_d|^2\, dx$, where $z_d \in L^2(\Omega)$. Then the semismooth Newton method applied to (5.23) converges locally superlinearly.*

Proof. For the proof we refer the reader to [34, p. 242]. □

Remark 5.4. An important feature of the SSN applied to (5.23) is that, in the linear case, the iterates coincide with the ones of the PDAS algorithm.

Program: SSN Method for the Optimization of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Regularization parameter: ');
[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

ub=10*ones(n^2,1);  %%%%% Upper bound %%%%%
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); y=sparse(n^2,1);
p=sparse(n^2,1); lam=sparse(n^2,1);
res=1; iter=0;

while res >= 1e-3
    iter=iter+1
```

```
    %%%%% Semismooth Newton step %%%%%
    Y=spdiags(y,0,n^2,n^2); P=spdiags(p,0,n^2,n^2);
    Act=spdiags(spones(max(0,lam+alpha*(u-ub))),0,n^2,n^2);

    A=[lap+3*Y.^2 -speye(n^2) sparse(n^2,n^2) sparse(n^2,n^2)
       -speye(n^2)+6*Y.*P sparse(n^2,n^2) lap+3*Y.^2 sparse(n^2,n^2)
       sparse(n^2,n^2) alpha*speye(n^2) speye(n^2) speye(n^2)
       sparse(n^2,n^2) -alpha*Act sparse(n^2,n^2) speye(n^2)-Act];

    F=[ -lap*y-y.^3+u
        -lap*p-3*Y.^2*p+y-z
        -p-alpha*u-lam
        -lam+max(0,lam+alpha*(u-ub))];

    delta=A\F;
    uprev=u;    yprev=y;    pprev=p;
    y=y+delta(1:n^2);
    u=u+delta(n^2+1:2*n^2);
    p=p+delta(2*n^2+1:3*n^2);
    lam=lam+delta(3*n^2+1:4*n^2);
    res=l2norm(u-uprev)+l2norm(y-yprev)+l2norm(p-pprev)
end
```

Chapter 6
Nonsmooth PDE-Constrained Optimization

6.1 Sparse L^1-Optimization

Finite-dimensional optimization problems with cost functions involving the l^1-norm of the design variable are known for enhancing sparsity of the optimal solution. This has important consequences in problems where a large amount of data is present, like speech recognition, image restoration, or data classification (see, e.g., [53]).

In the context of PDE-constrained optimization, sparsity means that the infinite dimensional design variable is localized in its domain of action, i.e., it takes zero value in a large part of its domain. The function-space counterpart of the l^1-norm is the Lebesgue L^1-norm, and the corresponding problem is formulated in the following way:

$$\begin{cases} \min\ J(y,u) + \beta \|u\|_{L^1}, \\ \text{subject to:} \\ \qquad e(y,u) = 0, \end{cases} \tag{6.1}$$

where $\beta > 0$. The additional difficulties, compared to the problems treated in previous sections, arise from the non-differentiability of the L^1-norm.

By rewriting (6.1) in reduced form, we obtain the equivalent problem:

$$\min_{u \in U}\ J(y(u),u) + \beta \|u\|_{L^1}. \tag{6.2}$$

Although the reduced cost (6.2) is nonsmooth, it consists in the sum of a regular part and a convex non-differentiable term. Thanks to this structure, optimality conditions can still be established according to the following result.

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9_6

Theorem 6.1. *Let U be a Banach space, $j_1 : U \to \mathbb{R}$ Gâteaux differentiable and $j_2 : U \to \mathbb{R} \cup \{+\infty\}$ convex and continuous. If $\bar{u} \in U$ is a local optimal solution to*

$$\min_{u \in U} \; j_1(u) + j_2(u), \tag{6.3}$$

then it satisfies the following optimality condition:

$$j_1'(\bar{u})(v - \bar{u}) + j_2(v) - j_2(\bar{u}) \geq 0, \text{ for all } v \in U. \tag{6.4}$$

Proof. From the optimality of $\bar{u}$ we know that

$$j_1(\bar{u}) + j_2(\bar{u}) \leq j_1(w) + j_2(w), \text{ for all } w \in B_\delta(\bar{u}).$$

Taking, for $v \in U$ and $t > 0$ sufficiently small, $w = \bar{u} + t(v - \bar{u}) \in B_\delta(\bar{u})$, it follows that

$$\begin{aligned} 0 &\leq j_1(\bar{u} + t(v - \bar{u})) - j_1(\bar{u}) + j_2(\bar{u} + t(v - \bar{u})) - j_2(\bar{u}) \\ &\leq j_1(\bar{u} + t(v - \bar{u})) - j_1(\bar{u}) + t\; j_2(v) + (1 - t)\; j_2(\bar{u}) - j_2(\bar{u}). \end{aligned}$$

Dividing by t and taking the limit on both sides we get

$$0 \leq \frac{j_1(\bar{u} + t(v - \bar{u})) - j_1(\bar{u})}{t} + j_2(v) - j_2(\bar{u}),$$

which implies that

$$0 \leq j_1'(\bar{u})(v - \bar{u}) + j_2(v) - j_2(\bar{u}). \qquad \square$$

For the sake of readability, let us hereafter focus on the following tracking type cost term

$$J(y, u) = \frac{1}{2} \int_\Omega |y - z_d|^2 \, dx + \frac{\alpha}{2} \|u\|_U^2,$$

and the control space $U = L^2(\Omega)$. From Theorem 6.1 an optimality condition for (6.2) is given by the following variational inequality:

$$(y(\bar{u}) - z_d, y'(\bar{u})(v - \bar{u})) + \alpha(\bar{u}, v - \bar{u}) + \beta \|v\|_{L^1} - \beta \|\bar{u}\|_{L^1} \geq 0, \forall v \in U. \tag{6.5}$$

Moreover, it can be proved (see, e.g., [20, pp. 70–71]) that the optimality condition (6.5) is equivalent to the existence of a dual multiplier $\lambda \in U$ such that:

$$(y(\bar{u}) - z_d, y'(\bar{u})v) + (\alpha\bar{u} + \lambda, v) = 0, \qquad \text{for all } v \in U \tag{6.6a}$$

$$\lambda = \beta, \qquad \text{in } \{x \in \Omega : \bar{u} > 0\} \tag{6.6b}$$

$$|\lambda| \leq \beta, \qquad \text{in } \{x \in \Omega : \bar{u} = 0\} \tag{6.6c}$$

$$\lambda = -\beta, \qquad \text{in } \{x \in \Omega : \bar{u} < 0\}. \tag{6.6d}$$

Assuming bijectivity of $e_y(\bar{y}, \bar{u})$ and proceeding as in Theorem 3.3, it can be justified that there exists an adjoint state $p \in W'$ such that the following optimality system holds:

$$e(\bar{y}, \bar{u}) = 0, \tag{6.7a}$$
$$e_y(\bar{y}, \bar{u})^* p = \bar{y} - z_d, \tag{6.7b}$$
$$p + \alpha\bar{u} + \lambda = 0, \tag{6.7c}$$
$$\lambda = \beta, \qquad \text{in } \{x \in \Omega : \bar{u} > 0\}, \tag{6.7d}$$
$$|\lambda| \leq \beta, \qquad \text{in } \{x \in \Omega : \bar{u} = 0\}, \tag{6.7e}$$
$$\lambda = -\beta, \qquad \text{in } \{x \in \Omega : \bar{u} < 0\}. \tag{6.7f}$$

By using the *max* and *min* functions, the last three equations of the optimality system (6.7) can be reformulated in the following short way:

$$\bar{u} - \max(0, \bar{u} + c(\lambda - \beta)) - \min(0, \bar{u} + c(\lambda + \beta)) = 0, \tag{6.8}$$

for all $c > 0$. This equivalence can be verified by inspection and its proof can be traced back to [32].

Considering the first three equations of (6.7) and Eq. (6.8), a semismooth Newton method for the solution of the optimality system can be stated. Taking into account the generalized derivative of the max and min functions:

$$G_{max}(v)(x) = \begin{cases} 1 & \text{if } v(x) > 0, \\ 0 & \text{if } v(x) \leq 0, \end{cases} \qquad G_{min}(v)(x) = \begin{cases} 0 & \text{if } v(x) \geq 0, \\ 1 & \text{if } v(x) < 0, \end{cases} \tag{6.9}$$

a Newton update for Eq. (6.8) is given by

$$\begin{aligned} \delta_u - (\chi_{\{u+c(\lambda-\beta)>0\}} + \chi_{\{u+c(\lambda+\beta)<0\}})(\delta_u + c\delta_\lambda) \\ = -u + \max(0, u + c(\lambda - \beta)) + \min(0, u + c(\lambda + \beta)), \end{aligned} \tag{6.10}$$

where $\chi_{\{w>0\}}$ stands for the indicator function of the set $\{x : w(x) > 0\}$. With the choice $c = \alpha^{-1}$, the complete algorithm is given next.

Algorithm 8 Sparse optimization of a semilinear equation

1: Choose $(y_0, u_0, p_0, \lambda_0) \in Y \times U \times W' \times U$ and set $k = 0$.
2: **repeat**
3: Set $\chi_A := \chi_{\{|\alpha u_k + \lambda_k| > \beta\}}$ and solve for $(\delta_y, \delta_u, \delta_p, \delta_\lambda) \in Y \times U \times W' \times U$:

$$\begin{pmatrix} \mathscr{L}''_{(y,u)}(y_k,u_k,p_k) & -e'(y_k,u_k)^* & 0 \\ -e'(y_k,u_k) & 0 & 0 \\ (\,0\; I-\chi_A\,) & 0 & -\alpha^{-1}\chi_A \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_p \\ \delta_\lambda \end{pmatrix}$$

$$= \begin{pmatrix} e_y(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e_u(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e(y_k,u_k) \\ -u_k + \max(0, u_k + \alpha^{-1}(\lambda_k - \beta)) + \min(0, u_k + \alpha^{-1}(\lambda_k + \beta)) \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$, $y_{k+1} = y_k + \delta_y$, $p_{k+1} = p_k + \delta_p$, $\lambda_{k+1} = \lambda_k + \delta_\lambda$ and $k = k+1$.
5: **until** Stopping criteria.

Example 6.1. We consider the following semilinear optimal control problem:

$$\begin{cases} \min\; J(y,u) = \dfrac{1}{2}\|y - z_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2} + \beta\|u\|_{L^1}, \\ \text{subject to:} \\ \qquad -\Delta y + y^3 = u \qquad \text{in } \Omega = (0,1)^2, \\ \qquad\qquad y = 0 \qquad \text{on } \Gamma. \end{cases}$$

The computed controls are depicted in Fig. 6.1 both when no sparsity term is included and when the L^1-norm weight β is set equal to 0.008. The parameter α takes, in both cases, the value 0.001. The sparse structure of the corresponding optimal control can be clearly identified from the plot on the right.

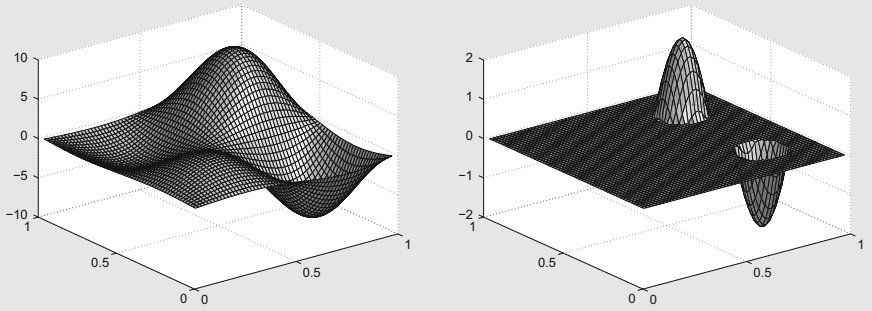

Fig. 6.1 Optimal control ($\beta = 0$) and sparse optimal control ($\beta = 0.008$)

Program: Sparse Optimization of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Regularization parameter: ');

[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('sin(2*pi*x).*sin(2*pi*y).*exp(2*x)/6','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

b=0.008;  %%%%% L1 weight %%%%%
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); y=sparse(n^2,1);
p=sparse(n^2,1); lam=sparse(n^2,1);
res=1; iter=0;

while res >= 1e-10
    iter=iter+1

    %%%%% Semismooth Newton step %%%%%
    Y=spdiags(y,0,n^2,n^2); P=spdiags(p,0,n^2,n^2);
    Act1=spdiags(spones(max(0,u+1/alpha*(lam-b))),0,n^2,n^2);
    Act2=spdiags(spones(min(0,u+1/alpha*(lam+b))),0,n^2,n^2);
    Act=Act1+Act2;

    A=[lap+3*Y.^2 -speye(n^2) sparse(n^2,n^2) sparse(n^2,n^2)
       -speye(n^2)+6*Y.*P sparse(n^2,n^2) lap+3*Y.^2 sparse(n^2,n^2)
       sparse(n^2,n^2) alpha*speye(n^2) speye(n^2) speye(n^2)
       sparse(n^2,n^2) speye(n^2)-Act sparse(n^2,n^2) -1/alpha*Act];

    F=[ -lap*y-y.^3+u
        -lap*p-3*Y.^2*p+y-z
        -p-alpha*u-lam
        -u+max(0,u+1/alpha*(lam-b))+min(0,u+1/alpha*(lam+b))];

    delta=A\F;

    uprev=u;    yprev=y;    pprev=p;
```

```
    y=y+delta(1:n^2);
    u=u+delta(n^2+1:2*n^2);
    p=p+delta(2*n^2+1:3*n^2);
    lam=lam+delta(3*n^2+1:4*n^2);
    res=l2norm(u-uprev)+l2norm(y-yprev)+l2norm(p-pprev)
end
```

6.2 Pointwise State Constraints

Although the action of the design variable u naturally imposes bounds on the state y, it is sometimes important to restrict the maximum or minimum pointwise value that the state variables can reach. Such is the case, for instance, of problems that involve heat transfer in solid materials, where the temperature has to remain below a critical melting point (see, e.g., [45]).

From a mathematical point of view, the presence of pointwise state constraints adds several difficulties to the treatment of the optimization problems. Analytically, different function spaces with low regularity functions have to be considered for the associated multipliers (which are just regular Borel measures). Due to this fact, also alternative approximation strategies have to be designed for the numerical solution of such problems.

A prototypical type of PDE-constrained optimization problem with pointwise state constraints can be formulated in the following manner:

$$\begin{cases} \min\ J(y,u), \\ \text{subject to:} \\ \qquad Ay = u, \\ \qquad y(x) \leq y_b \ \text{ a.e. in } \Omega, \end{cases} \tag{6.11}$$

where A stands for a linear elliptic operator with sufficiently regular coefficients, Ω is of class $\mathscr{C}^2$ Y is a reflexive Banach space and $J\colon Y \times L^2(\Omega) \longrightarrow \mathbb{R}$ is a tracking type functional given by:

$$J(y,u) = \frac{1}{2}\int_\Omega |y - z_d|^2\, dx + \frac{\alpha}{2}\|u\|^2_{L^2}.$$

In order to derive optimality conditions for problem (6.11), the following abstract Lagrange multiplier theorem may be used (for the proof we refer to [11]).

Theorem 6.2. *Let U, Y be Banach spaces and $K \subset Y$ a convex set with nonempty interior. Let $\bar{u} \in U$ be a local optimal solution of the problem*

$$\begin{cases} \min J(u), \\ \text{subject to: } G(u) \in K, \end{cases} \tag{6.12}$$

where $J : U \to \mathbb{R}$ and $G : U \to Y$ are Gâteaux differentiable mappings. If there exists $u_0 \in U$ such that

$$G(\bar{u}) + G'(\bar{u})(u_0 - \bar{u}) \in int(K),$$

then there exists a Lagrange multiplier $\mu \in Y'$ such that:

$$\langle J'(\bar{u}) + G'(\bar{u})^* \mu, u - \bar{u} \rangle_{U',U} \geq 0, \qquad \text{for all } u \in K, \tag{6.13}$$

$$\langle \mu, w - G(\bar{u}) \rangle_{Y',Y} \leq 0, \qquad \text{for all } w \in Y. \tag{6.14}$$

Since convex sets of the type $\{v \in L^p(\Omega) : v \leq y_b \text{ a.e. in } \Omega\}$ have empty interior (see, e.g., [56, p. 326]), a stronger topology is required in order to apply the previous theorem. A usual approach consists in considering the state $y = G(u)$ in the space of continuous functions, where the nonempty interior hypothesis is satisfied. However, this may be the case only if some extra regularity of the state can be obtained from the governing equation. This occurs, for instance, in elliptic problems under additional regularity of the coefficients and the domain Ω.

Theorem 6.3. *Let $\bar{u} \in L^2(\Omega)$ be a local optimal solution to (6.11) and $\bar{y} \in Y$ its associated state. Assume that $A \in \mathscr{L}(Y, L^2(\Omega))$ is bijective and that $Y \hookrightarrow C(\Omega)$ with dense and continuous injection. Then there exists $u_0 \in L^2(\Omega)$ such that $\bar{y} + w_{u_0 - \bar{u}} < y_b$ a.e. in Ω, where $Aw_{u_0-\bar{u}} = u_0 - \bar{u}$, and there exists $p \in L^2(\Omega)$ solution of the adjoint equation*

$$\int_\Omega p A w \, dx = \langle \varphi, w \rangle_{\mathscr{M}(\Omega), C(\Omega)}, \text{ for all } w \in Y, \tag{6.15}$$

for any $\varphi \in \mathscr{M}(\Omega)$, the space of regular Borel measures. Moreover, there exist $\mu \in \mathscr{M}(\Omega)$ and $\bar{p} \in L^2(\Omega)$ such that the following optimality system holds:

$$A\bar{y} = \bar{u}, \tag{6.16a}$$

$$\int_\Omega \bar{p} A w \, dx = \int_\Omega (\bar{y} - z_d) w \, dx + \langle \mu, w \rangle_{\mathscr{M}(\Omega), C(\Omega)} \qquad \text{for all } w \in Y, \tag{6.16b}$$

$$\bar{p} + \alpha \bar{u} = 0, \tag{6.16c}$$

$$\langle \mu, w - \bar{y} \rangle_{\mathscr{M}(\Omega), C(\Omega)} \leq 0, \qquad \text{for all } w \in Y. \tag{6.16d}$$

Due to the poor regularity of the multipliers involved in (6.16), a Moreau–Yosida regularization is frequently used for the numerical solution of the optimization problem. This approach consists in penalizing the pointwise state constraints by means of the C^1-function

$$\max(0, \hat{\lambda} + \gamma(y - y_b))^2, \text{ with some fixed } \hat{\lambda} \in L^2(\Omega),$$

yielding the following problem:

$$\begin{cases} \min\ J(y,u) + \frac{1}{2\gamma}\int_\Omega \max(0, \hat{\lambda} + \gamma(y - y_b))^2\, dx, \\ \text{subject to:} \\ \qquad Ay = u. \end{cases} \tag{6.17}$$

Existence of a solution to (6.17) can be argued in a similar manner as for the unconstrained problem. Moreover, a first-order optimality system may be derived using the techniques of Chap. 3.

Theorem 6.4. *Let $(\hat{u}, \hat{y})$ be a local optimal solution to (6.17) and let $\hat{\lambda} = 0$. Then there exists an adjoint state $p \in L^2(\Omega)$ such that*

$$A\hat{y} = \hat{u}, \tag{6.18a}$$
$$A^* p = \hat{y} - z_d + \propto, \tag{6.18b}$$
$$p + \alpha\hat{u} = 0, \tag{6.18c}$$
$$\propto = \max(0, \gamma(\hat{y} - y_b)). \tag{6.18d}$$

Proof. The existence of an adjoint state is obtained by following the lines of the proof of Theorem 3.3. In what follows let us introduce the variable

$$\propto := \max(0, \gamma(\hat{y} - y_b)) \in L^2(\Omega).$$

By computing the derivative of the reduced cost functional we obtain:

$$f'(\hat{u})h = \langle J_y(y(\hat{u}), \hat{u}), y'(\hat{u})h\rangle_{Y',Y} + \gamma \int_\Omega \max(0, \hat{y} - y_b) y'(\hat{u})h\, dx + J_u(y(\hat{u}), \hat{u})h.$$

which, using the adjoint equation (6.18b), implies that

$$\begin{aligned} f'(\hat{u})h &= \langle A^* p, y'(\hat{u})h)\rangle_{Y',Y} + J_u(y(\hat{u}), \hat{u})h \\ &= (p, Ay'(\hat{u})h)_{L^2} + J_u(y(\hat{u}), \hat{u})h \end{aligned}$$

Considering the linearized equation $Ay'(\hat{u})h = h$, we finally get that

$$f'(\hat{u})h = (p,h)_{L^2} + J_u(y(\hat{u}),\hat{u})h \tag{6.19}$$

and, therefore,

$$p + \alpha\hat{u} = 0 \text{ in } L^2(\Omega).$$

□

The solutions so obtained yield a sequence $\{(y_\gamma,u_\gamma)\}_{\gamma>0}$ that approximates the solution to (6.11) in the following sense.

Theorem 6.5. *The sequence $\{(y_\gamma,u_\gamma)\}_{\gamma>0}$ of solutions to (6.17) contains a subsequence which converges strongly in $Y\times L^2(\Omega)$ to an optimal solution $(\bar{y},\bar{u})$ of (6.11).*

Proof. Let $(\bar{y},\bar{u})\in Y\times U$ be a solution to (6.11). From the properties of the regularized cost functional we know that

$$J_\gamma(y_\gamma,u_\gamma) \le J_\gamma(\bar{y},\bar{u}) = J(\bar{y},\bar{u}). \tag{6.20}$$

Consequently, since $\alpha>0$, the sequence $\{u_\gamma\}_{\gamma>0}$ is uniformly bounded in $L^2(\Omega)$, which implies that $\{y_\gamma\}_{\gamma>0}$ is uniformly bounded in Y. Therefore, there exists a subsequence, denoted the same, such that $y_\gamma \rightharpoonup \hat{y}$ weakly in Y and $u_\gamma \rightharpoonup \hat{u}$ weakly in $L^2(\Omega)$.

Additionally, from (6.20) the term

$$\frac{1}{2\gamma}\left\|\max(0,\gamma(y_\gamma - y_b))\right\|^2_{L^2(\Omega)} \tag{6.21}$$

is uniformly bounded with respect to γ. Hence,

$$\lim_{\gamma\to\infty}\left\|\max(0,y_\gamma - y_b)\right\|_{L^2(\Omega)} = 0.$$

Applying Fatou's Lemma to the previous term we get that $\hat{y}\le y_b$. Considering additionally that

$$J(\hat{y},\hat{u}) \le \liminf_{\gamma\to\infty} J(y_\gamma,u_\gamma) \le \limsup_{\gamma\to\infty} J_\gamma(y_\gamma,u_\gamma) \le J(\bar{y},\bar{u}), \tag{6.22}$$

we get that $(\hat{y},\hat{u})$ is solution of (6.11). Subsequently, we denote the optimal pair by $(\bar{y},\bar{u})$.

To verify strong convergence, let us first note that, due to (6.22)

$$\lim_{\gamma\to\infty}\|y_\gamma - z_d\|^2_{L^2} + \alpha\|u_\gamma\|^2_{L^2} = \|\bar{y}-z_d\|^2_{L^2} + \alpha\|\bar{u}\|^2_{L^2}$$

and, hence, $u_\gamma \to \bar{u}$ strongly in $L^2(\Omega)$. From the state equations it can be verified that the difference $y_\gamma - \bar{y}$ satisfies the equation $A(y_\gamma - \bar{y}) = u_\gamma - \bar{u}$, which thanks to the boundedness of A^{-1} implies that $y_\gamma \to \bar{y}$ strongly in Y. □

For the numerical solution of (6.18) the difficulty arises from the last nonsmooth equation. Using again the generalized derivative of the *max* function given by (6.9), the semismooth Newton step is given by

$$\delta_\infty - \gamma\chi_{\{y>y_b\}}\delta_y = -\infty + \max(0, \gamma(y - y_b)),$$

and the complete algorithm can be formulated as follows.

Algorithm 9 Moreau-Yosida

1: Choose $(y_0, u_0, p_0, \infty_0) \in Y \times L^2(\Omega) \times W' \times L^2(\Omega)$.
2: **repeat**
3: Solve for $(\delta_y, \delta_u, \delta_p, \delta_\infty) \in Y \times L^2(\Omega) \times W' \times L^2(\Omega)$

$$\begin{pmatrix} \mathcal{L}''_{(y,u)}(y_k,u_k,p_k) & -e'(y_k,u_k)^* & 0 \\ -e'(y_k,u_k) & 0 & 0 \\ (-\gamma\chi_{\{y_k>y_b\}}\ 0) & 0 & I \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \end{pmatrix} \\ \delta_p \\ \delta_\infty \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} e_y(y_k,u_k)^* p_k - J_y(y_k,u_k) \\ e_u(y_k,u_k)^* p_k - J_y(y_k,u_k) \end{pmatrix} \\ -e(y_k,u_k) \\ -\infty_k + \max(0, \gamma(y_k - y_b)) \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$, $y_{k+1} = y_k + \delta_y$, $p_{k+1} = p_k + \delta_p$, $\infty_{k+1} = \infty_k + \delta_\infty$ and $k = k+1$.
5: **until** Stopping criteria.

Program: State-Constrained Optimal Control of a Semilinear Equation

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Tikhonov regularization parameter: ');
gama=1e4;

[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);
```

```
yb=0.2*ones(n^2,1);  %%%%% Upper bound %%%%%
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); y=sparse(n^2,1); p=sparse(n^2,1); mu=sparse(n^2,1);
res=1; iter=0;

while res >= 1e-3
    iter=iter+1

    %%%%% Semismooth Newton step %%%%%
    Y=spdiags(y,0,n^2,n^2); P=spdiags(p,0,n^2,n^2);
    Act=spdiags(spones(max(0,gama*(y-yb))),0,n^2,n^2);

    A=[ lap+3*Y.^2 -speye(n^2) sparse(n^2,n^2) sparse(n^2,n^2)
        -speye(n^2)+6*Y.*P sparse(n^2,n^2) lap+3*Y.^2 -speye(n^2)
        sparse(n^2,n^2) alpha*speye(n^2) speye(n^2) sparse(n^2,n^2)
        -gama*Act sparse(n^2,n^2) sparse(n^2,n^2) speye(n^2)];

    F=[ -lap*y-y.^3+u
        -lap*p-3*Y.^2*p+y-z+mu
        -p-alpha*u
        -mu+max(0,gama*(y-yb))];

    delta=A\F;

    uprev=u;    yprev=y;    pprev=p;
    y=y+delta(1:n^2);
    u=u+delta(n^2+1:2*n^2);
    p=p+delta(2*n^2+1:3*n^2);
    mu=mu+delta(3*n^2+1:4*n^2);
    res=l2norm(u-uprev)+l2norm(y-yprev)+l2norm(p-pprev)
end
```

6.3 Variational Inequality Constraints

Another type of nonsmooth optimization problems occurs when the constraints are given by so-called partial variational inequalities. An elliptic variational inequality problem has the following form: Find $y \in Y$ such that

$$a(y, v-y) + j(v) - j(y) \geq \langle f, v-y \rangle_{Y',Y}, \qquad \text{for all } v \in Y, \tag{6.23}$$

where Y, U are Hilbert function spaces defined on a bounded domain $\Omega \subset \mathbb{R}^N$, $a(\cdot,\cdot)$ is a continuous and coercive bilinear form, $j : Y \to \mathbb{R} \cup \{\infty\}$ is a convex nondifferentiable functional and $f \in Y'$.

Inequalities of this type arise in contact mechanics, elastoplasticity, viscoplastic fluid flow, among others (see, e.g., [18, 20]). If the convex functional $j(\cdot)$ corresponds to the indicator functional of a convex set, the variational inequality has special structural properties. Something similar occurs if the term has the form

$$j(y) = \int_S |Ky| ds,$$

with $S \subset \bar{\Omega}$ and $K \in \mathscr{L}(Y, (L^2(S))^m)$, for some $m \geq 1$. Both cases will be considered in the sequel.

The optimization of variational inequalities is closely related to the field of mathematical programming with equilibrium constraints (MPEC), which has received increasing interest in the past years, both in finite-dimensions and in function spaces [15, 27, 28, 44, 49]. Due to the nondifferentiable structure of the constraints, the characterization of solutions via optimality conditions becomes challenging, and the same extends to the numerical solution of such problems.

A general tracking type distributed optimization problem can be formulated as follows:

$$\begin{cases} \min\ J(y,u) = \frac{1}{2}\int_\Omega |y - z_d|^2\, dx + \frac{\alpha}{2}\|u\|_U^2, \\ \text{subject to:} \\ \qquad a(y, v-y) + j(v) - j(y) \geq \langle u, v-y\rangle_{Y',Y}, \qquad \text{for all } v \in Y. \end{cases} \tag{6.24}$$

For simplicity we restrict our attention to the cases where $U = L^2(\Omega)$ and assume that $Y \hookrightarrow L^2(\Omega) \hookrightarrow Y'$ with compact and continuous embeddings. Existence of a unique solution to (6.23) can be easily justified by the well-known Stampacchia's theorem, while existence of an optimal solution to (6.24) is shown in the following result.

Theorem 6.6. *There exists an optimal solution for problem* (6.24).

Proof. Since the cost functional is bounded from below, there exists a minimizing sequence $\{(y_n, u_n)\}$, i.e., $J(y_n, u_n) \to \inf_u J(y,u)$, where y_n stands for the unique solution to

$$a(y_n, v - y_n) + j(v) - j(y_n) \geq \langle u_n, v - y_n\rangle_{Y',Y}, \text{ for all } v \in Y. \tag{6.25}$$

From the structure of the cost functional it also follows that $\{u_n\}$ is bounded in U. Additionally, it follows from (6.25) that $\{y_n\}$ is bounded in Y. Therefore, there exists a

subsequence (denoted in the same way) such that

$$u_n \rightharpoonup \hat{u} \text{ weakly in } U \qquad \text{and} \qquad y_n \rightharpoonup \hat{y} \text{ weakly in } Y.$$

Due to the compact embedding $L^2(\Omega) \hookrightarrow Y'$ it then follows that

$$u_n \to \hat{u} \text{ strongly in } Y'.$$

From (6.25) we directly obtain that

$$a(y_n, y_n) - a(y_n, v) + j(y_n) - j(v) - \langle u_n, y_n - v \rangle_{Y',Y} \leq 0, \ \forall v \in Y.$$

Thanks to the convexity and continuity of $a(\cdot,\cdot)$ and $j(\cdot)$ we may take the limit inferior in the previous inequality and obtain that

$$a(\hat{y}, \hat{y}) - a(\hat{y}, v) + j(\hat{y}) - j(v) - \langle \hat{u}, \hat{y} - v \rangle_{Y',Y} \leq 0, \ \forall v \in Y, \tag{6.26}$$

which implies that $\hat{y}$ solves (6.23) with $\hat{u}$ on the right hand side.

Thanks to the weakly lower semicontinuity of the cost functional we finally obtain that

$$J(\hat{y}, \hat{u}) \leq \liminf_{n \to \infty} J(y(u_n), u_n) = \inf_u J(y(u), u),$$

which implies the result. □

6.3.1 *Inequalities of the First Kind*

If the non-differentiable term $j(\cdot)$ corresponds to the indicator functional of a convex set of the type $C := \{v \in Y : v \leq \psi \text{ a.e. in } \Omega\}$, with $\psi \in Y : \psi \geq 0$ a.e. in Ω, the variational inequality problem consists in finding $y \in C$ such that

$$a(y, v - y) \geq \langle u, v - y \rangle_{Y',Y}, \qquad \text{for all } v \in C. \tag{6.27}$$

These type of inequalities are commonly known as *obstacle type* inequalities. For this particular instance, additional properties can be investigated. For example, if the domain is of class $\mathscr{C}^2$, the state space $Y = H_0^1(\Omega)$ and the right hand side belongs to $U = L^2(\Omega)$, then there exists a unique solution $y \in H^2(\Omega) \cap H_0^1(\Omega)$ to (6.27) (see, e.g., [4]). Moreover, there exists a slack multiplier $\lambda \in L^2(\Omega)$ such that (6.27) can be equivalently writ-

ten as the complementarity problem:

$$a(y,v)+(\lambda,v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.28a}$$
$$y\le\psi \text{ a.e., } \lambda\ge 0 \text{ a.e.}, \tag{6.28b}$$
$$(\lambda,y-\psi)_{L^2}=0. \tag{6.28c}$$

The corresponding optimization problem can then be cast as a mathematical program with complementarity constraints. The problem reads as follows:

$$\begin{cases} \min\ J(y,u)=\frac{1}{2}\int_\Omega |y-z_d|^2\,dx+\frac{\alpha}{2}\|u\|^2_{L^2}, \\ \text{subject to:} \\ \quad a(y,v)+(\lambda,v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \\ \quad y\le\psi \text{ a.e., } \lambda\ge 0 \text{ a.e.}, \\ \quad (\lambda,y-\psi)_{L^2}=0. \end{cases} \tag{6.29}$$

Due to the nonsmooth nature of the constraints, different type of stationary points may be characterized for problem (6.29), in contrast to what happens in differentiable optimization, where a single stationarity concept suffices.

Definition 6.1. A point $\bar{u}\in U$ is called C(larke)-stationary for problem (6.29), if it satisfies the following system:

$$a(y,v)+(\lambda,v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.30a}$$
$$y\le\psi, \qquad \text{a.e. in } \Omega, \tag{6.30b}$$
$$\lambda\ge 0, \qquad \text{a.e. in } \Omega, \tag{6.30c}$$
$$(\lambda,y-\psi)_{L^2}=0, \tag{6.30d}$$
$$a(p,v)+\langle\xi,v\rangle_{Y',Y}=(y-z_d,v)_{L^2}, \qquad \text{for all } v\in Y \tag{6.30e}$$
$$p+\alpha u=0, \qquad \text{a.e. in } \Omega \tag{6.30f}$$
$$\langle\xi,p\rangle_{Y',Y}\ge 0, \tag{6.30g}$$
$$p=0, \qquad \text{a.e. in } \mathscr{I}:=\{x:\lambda>0\} \tag{6.30h}$$

and, additionally,

$$\langle\xi,\phi\rangle_{Y',Y}=0, \qquad \forall\phi\in Y:\phi=0 \text{ a.e. in } \{x:y=\psi\}.$$

A point $\bar{u}\in U$ is called strong stationary, if it satisfies (6.30) and

$$p\le 0 \qquad \text{a.e. in } B,$$

$$\langle \xi, \phi \rangle_{Y',Y} \leq 0, \qquad \forall \phi \in Y : \phi \geq 0 \text{ a.e. in } B \text{ and } \phi = 0 \text{ a.e. in } \mathscr{I},$$

where $B := \{x : y = \psi \wedge \lambda = 0\}$ stands for the biactive set.

The derivation of optimality conditions for this type of problems is actually a current topic of research. Regularization approaches as well as generalized differentiability properties (directional, conical) of the solution map or elements of set valued analysis have been explored with different outcomes, advantages, and disadvantages (see, e.g., [4, 5, 27, 28, 29, 42, 46, 50]).

Back to problem (6.29) and its numerical treatment, note that the last three complementarity relations can be formulated in reduced form as:

$$\lambda = \max(0, \lambda + c(y - \psi)), \text{ for any } c > 0. \tag{6.31}$$

This reformulation enables the development of new algorithmic ideas for handling the constraints. An immediate regularized version of (6.31) is obtained, for instance, if the multiplier inside the *max* function is replaced by a function $\bar{\lambda} \in L^2(\Omega)$ (possibly $\bar{\lambda} = 0$) and c is considered as a regularization parameter γ, which may tend to infinity. If, in addition, a local regularization of the *max* function is utilized in order to obtain a smooth solution operator, then, instead of the governing variational inequality, the following nonlinear PDE is obtained as constraint:

$$a(y, v) + (\max_\gamma(0, \gamma(y - \psi)), v)_{L^2} = (u, v)_{L^2}, \text{ for all } v \in Y, \tag{6.32}$$

where $\max_\gamma$ is a C^1-approximation of the max function given by

$$\max_\gamma(0, x) := \begin{cases} x & \text{if } x \geq \frac{1}{2\gamma}, \\ \frac{\gamma}{2}\left(x + \frac{1}{2\gamma}\right)^2 & \text{if } |x| \leq \frac{1}{2\gamma}, \\ 0 & \text{if } x \leq -\frac{1}{2\gamma}. \end{cases}$$

The resulting optimization problem reads as follows:

$$\begin{cases} \min J(y, u) = \frac{1}{2} \int_\Omega |y - z_d|^2 \, dx + \frac{\alpha}{2} \|u\|^2_{L^2}, \\ \text{subject to:} \\ \quad a(y, v) + (\max_\gamma(0, \gamma(y - \psi)), v)_{L^2} = (u, v)_{L^2}, \text{ for all } v \in Y. \end{cases} \tag{6.33}$$

An optimality condition for problem (6.33) may be obtained using the techniques of Chap. 3. The resulting optimality system is given as follows:

$$a(y,v)+\gamma(\max_\gamma(0,y-\psi),v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.34a}$$

$$a(p,v)+\gamma(\operatorname{sign}_\gamma(y-\psi)\,p,v)_{L^2}=(y-z_d,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.34b}$$

$$\alpha u+p=0, \tag{6.34c}$$

where

$$\operatorname{sign}_\gamma(x)=\begin{cases}1 & \text{if } x\geq \frac{1}{2\gamma},\\ \gamma\left(x+\frac{1}{2\gamma}\right) & \text{if } |x|\leq \frac{1}{2\gamma},\\ 0 & \text{if } x\leq -\frac{1}{2\gamma}.\end{cases}$$

In the next theorem convergence of the solutions of the regularized variational inequalities toward the solution of (6.27) is established.

Theorem 6.7. *Let y and y_γ be solutions to (6.27) and (6.32), respectively, both with $u\in L^2(\Omega)$ on the right hand side. Then*

$$y_\gamma\to y \text{ strongly in } Y \qquad \text{as } \gamma\to+\infty.$$

Proof. Let us define the primitive function

$$\Phi_\gamma(x)=\int_0^x \max_\gamma(0,s)\,ds,$$

and consider the energy minimization problem

$$\min_{y\in Y}\ \frac{1}{2}a(y,y)+\int_\Omega \frac{1}{\gamma}\Phi_\gamma(\gamma(y-\psi))\,dx-(u,y)_{L^2}. \tag{6.35}$$

Equation (6.32) is a necessary and sufficient optimality condition for (6.35). Therefore, y_γ also solves (6.35) and, from the optimality, we get that

$$\frac{1}{2}a(y_\gamma,y_\gamma)+\frac{1}{\gamma}\int_\Omega \Phi_\gamma(\gamma(y_\gamma-\psi))\,dx-(u,y_\gamma)\leq \frac{1}{2}a(\psi,\psi)-(u,\psi).$$

Since the function in (6.35) is radially unbounded (see (3.4)), we get that $\{y_\gamma\}$ is bounded in Y and, moreover,

$$\frac{\kappa}{2}\|y_\gamma\|_Y^2+\frac{1}{\gamma}\int_\Omega \Phi_\gamma(\gamma(y_\gamma-\psi))\,dx\leq C,$$

for some constants $\kappa>0$ and $C>0$. Therefore, there exists a subsequence, denoted the same, such that $y_\gamma\rightharpoonup\tilde{y}$ weakly in Y.

Since $0 \le \max(0,x) \le \max_\gamma(0,x)$, it follows that

$$0 \le \frac{1}{2\gamma}\int_\Omega |\max(0,\gamma(y_\gamma-\psi))|^2\,dx \le \frac{1}{\gamma}\int_\Omega \Phi_\gamma(\gamma(y_\gamma-\psi))\,dx \le C.$$

Consequently,

$$\|\max(0,y_\gamma-\psi)\|_{L^2}^2 \to 0 \qquad \text{as } \gamma\to\infty,$$

and, thanks to Fatou's Lemma, $|\max(0,\tilde{y}-\psi)| = 0$.

Let $\lambda_\gamma := \max_\gamma(0,\gamma(y_\gamma-\psi)) \ge 0$. It then follows that

$$(\lambda_\gamma, y_\gamma - y) = (\lambda_\gamma, y_\gamma-\psi) + (\lambda_\gamma, \overbrace{\psi-y}^{\ge 0}) \ge \frac{1}{\gamma}(\lambda_\gamma,\gamma(y_\gamma-\psi)) \ge 0. \tag{6.36}$$

Taking the difference between (6.32) and (6.28a), with the test function $v = y_\gamma - y$, leads to the equation

$$a(y_\gamma-y,y_\gamma-y) + (\lambda_\gamma-\lambda,y_\gamma-y)_{L^2} = 0. \tag{6.37}$$

Using (6.36), the coercivity of the bilinear form, the complementarity relations (6.28b)–(6.28c) and the feasibility of $\tilde{y}$, we get that

$$\begin{aligned} 0 \le \kappa\|y_\gamma-y\|_Y^2 &\le a(y_\gamma-y,y_\gamma-y) + (\lambda_\gamma,y_\gamma-y) \\ &= (\lambda,y_\gamma-\tilde{y})_{L^2} + (\lambda,\tilde{y}-\psi)_{L^2} + \overbrace{(\lambda,\psi-y)_{L^2}}^{=0} \\ &\le (\lambda,y_\gamma-\tilde{y})_{L^2}. \end{aligned}$$

Taking the limit as $\gamma\to\infty$ yields the result. □

As a consequence of the previous result, the regularized optimal solutions, obtained by solving (6.33), converge to the original solution of (6.29) in the following sense.

Theorem 6.8. *Every sequence $\{u_\gamma\}_{\gamma>0}$ of solutions to (6.33) contains a subsequence which converges strongly in U to an optimal solution $\bar{u}\in U$ of (6.29). Moreover, if (6.29) has a unique optimal solution, then the whole sequence converges strongly in U towards $\bar{u}$, as $\gamma\to\infty$.*

Proof. Let $\bar{u}$ denote an optimal solution for (6.29). From the structure of the cost functional we obtain that

$$J(y_\gamma,u_\gamma) \le J(0,0) \le C_0, \text{ for all } \gamma > 0, \tag{6.38}$$

and, therefore, the sequence $\{u_\gamma\}$ is uniformly bounded in U. Consequently, there exists a weakly convergent subsequence, which will be also denoted by $\{u_\gamma\}$.

Let $\hat{u}$ be a weak accumulation point of $\{u_\gamma\}$. Thanks to Theorem 6.7, $y_\gamma(u_\gamma) \to y(u_\gamma)$ strongly in Y. Additionally, considering (6.27) with $\hat{u}$ and u_γ on the right hand side, respectively, and adding both inequalities, we get that

$$a(y(\hat{u}) - y(u_\gamma), y(\hat{u}) - y(u_\gamma)) \leq (\hat{u} - u_\gamma, y(\hat{u}) - y(u_\gamma)),$$

which, thanks to the ellipticity of $a(\cdot,\cdot)$, implies that $y(u_\gamma) \to y(\hat{u})$ strongly in Y. Consequently,

$$y_\gamma(u_\gamma) \to y(\hat{u}) \text{ strongly in } Y.$$

Due to the weakly lower semicontinuity of the cost functional, we obtain that

$$J(y(\hat{u}), \hat{u}) \leq \liminf_{\gamma\to\infty} J(y_\gamma(u_\gamma), u_\gamma) \leq \liminf_{\gamma\to\infty} J(y_\gamma(\bar{u}), \bar{u}) = J(\bar{y}, \bar{u})$$

and, therefore, $(y(\hat{u}), \hat{u})$ is an optimal solution to (6.29). From the last inequality it also follows that

$$\lim_{\gamma\to\infty} \|u_\gamma\|_U^2 = \|\bar{u}\|_U^2,$$

which, together with the weak convergence $u_\gamma \rightharpoonup \bar{u}$, implies strong convergence in U. If, in addition, the optimal solution is unique, convergence of whole sequence takes place. □

Remark 6.1. By passing to the limit in (6.34) an optimality system of C-type is obtained for the limit point (see, e.g., [33]).

Since $\max_\gamma$ is continuously differentiable and sign_γ is a Newton differentiable function with derivative

$$\text{sign}'_\gamma(x) = \begin{cases} \gamma & \text{if } |x| \leq \frac{1}{2\gamma}, \\ 0 & \text{if not,} \end{cases}$$

a semismooth Newton method may be used for solving system (6.34), yielding the following adjoint equation update:

$$\begin{aligned} a(\delta_p, v) + \gamma(\text{sign}_\gamma(y-\psi)\,\delta_p, v)_{L^2(\Omega)} + \gamma^2 \int_{|y-\psi|\leq\frac{1}{2\gamma}} p\,\delta_y\,v\,dx - (\delta_y, v)_{L^2(\Omega)} \\ = -a(p, v) - \gamma(\text{sign}_\gamma(y-\psi)\,p, v)_{L^2(\Omega)} + (y - z_d, v)_{L^2(\Omega)}. \end{aligned} \tag{6.39}$$

The complete algorithm is given through the following steps:

Algorithm 10 SSN–Optimization with VI constraints (first kind)

1: Choose $(y_0, u_0, p_0) \in Y \times U \times W'$.
2: **repeat**
3: Solve for $(\delta_y, \delta_u, \delta_p) \in Y \times U \times Y$

$$\begin{pmatrix} A+\gamma \operatorname{sign}_\gamma(y_k-\psi) & -I & 0 \\ \gamma^2 \chi_{|y_k-\psi|\le \frac{1}{\gamma}}\, p - I & 0 & A+\gamma \operatorname{sign}_\gamma(y_k-\psi) \\ 0 & \alpha I & I \end{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \\ \delta_p \end{pmatrix} = - \begin{pmatrix} Ay_k + \gamma \max_\gamma(0, y_k-\psi) - u_k \\ Ap_k + \gamma \operatorname{sign}_\gamma(y_k-\psi)\, p_k - y_k + z_d \\ \alpha u_k + p_k \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$, $y_{k+1} = y_k + \delta_y$, $p_{k+1} = p_k + \delta_p$ and $k = k+1$.
5: **until** Stopping criteria.

Program: Optimal Control of an Obstacle Problem

```
clear all;
n=input('Mesh points: '); h=1/(n+1);
alpha=input('Tikhonov regularization parameter: ');
gama=1e3;

[x1,y1]=meshgrid(h:h:1-h,h:h:1-h);  %%%%% Coordinates %%%%%

%%%%% Desired state %%%%%
desiredstate=inline('x.*y','x','y');
z=feval(desiredstate,x1,y1); z=reshape(z,n^2,1);

yb=0.2*ones(n^2,1);  %%%%% Upper obstacle bound %%%%%
lap=matrices(n,h);  %%%%% Laplacian %%%%%

%%%%% Initialization %%%%%
u=sparse(n^2,1); y=sparse(n^2,1); p=sparse(n^2,1); mu=sparse(n^2,1);
res=1; iter=0;

while res >= 1e-3
    iter=iter+1

    %%%%% Semismooth Newton step %%%%%
```

```
    S=spdiags(signg(y-yb,gama),0,n^2,n^2); P=spdiags(p,0,n^2,n^2);
    Act=spdiags(spones(max(1/(2*gama)-abs(y-yb),0)),0,n^2,n^2);

    A=[ lap+gama*S -speye(n^2) sparse(n^2,n^2)
        -speye(n^2)+gama^2*Act*P sparse(n^2,n^2) lap+gama*S
        sparse(n^2,n^2) alpha*speye(n^2) speye(n^2)];

    F=[ -lap*y-gama*maxg(y-yb,gama)+u
        -lap*p-gama*signg(y-yb,gama).*p+y-z
        -p-alpha*u];

    delta=A\F;

    uprev=u;    yprev=y;    pprev=p;
    y=y+delta(1:n^2);
    u=u+delta(n^2+1:2*n^2);
    p=p+delta(2*n^2+1:3*n^2);
    res=l2norm(u-uprev)+l2norm(y-yprev)+l2norm(p-pprev)
end
```

6.3.2 *Inequalities of the Second Kind*

There are several problems where the variational inequality constraints do not have an obstacle-type structure like (6.27), but are rather characterized by a so-called threshold behavior. This means that a certain constitutive property is preserved until a quantity surpasses a limit, from which on a different qualitative behavior takes place. This is the case, for instance, of a body in frictional contact with a surface, where no tangential movement occurs until the external forces are large enough so that the friction threshold is surpassed and the displacement starts. Something similar occurs in elastoplasticity or viscoplastic fluids, to name a few application examples (see [20] and the references therein).

These inequalities are characterized by the presence of a non-differentiable term of the form

$$j(y) = \int_S |Ky| ds,$$

where $S \subset \bar{\Omega}$ and $K \in \mathscr{L}(Y, (L^2(S))^m)$, for some $m \geq 1$, yielding the following problem: Find $y \in Y$ such that

$$a(y, v-y) + g \int_S |Kv| ds - g \int_S |Ky| ds \geq \langle f, v-y \rangle_{Y',Y}, \qquad \text{for all } v \in Y, \tag{6.40}$$

where $g > 0$ stands for the threshold coefficient.

Similar to variational inequalities of the first kind, existence of a multiplier $q \in (L^2(S))^m$ can be proved by duality techniques (see, e.g., [20]) and an equivalent system to (6.40) is given by:

$$a(y,v) + \int_S qKv\,ds = \langle f, v\rangle_{Y',Y}, \qquad \text{for all } v \in Y, \tag{6.41a}$$

$$(q(x), Ky(x))_{\mathbb{R}^m} = g|Ky(x)|_{\mathbb{R}^m}, \qquad \text{a.e. in } S, \tag{6.41b}$$

$$|q(x)| \le g, \qquad \text{a.e. in } S. \tag{6.41c}$$

Note that, if $Ky(x) = 0$, no pointwise information about $q(x)$ is obtained from (6.41).

For simplicity, let us hereafter focus on the special case $S = \Omega$ and $K : Y \to L^2(\Omega)$ the canonical injection. A distributed type optimization problem may then be formulated as follows:

$$\begin{cases} \min\ J(y,u) = \frac{1}{2}\int_\Omega |y - z_d|^2\,dx + \frac{\alpha}{2}\|u\|^2_{L^2}, \\ \text{subject to:} \\ \qquad a(y,v) + (q,v)_{L^2} = (u,v)_{L^2}, & \text{for all } v \in Y, \\ \qquad q(x)y(x) = g|y(x)|, & \text{a.e. in } \Omega, \\ \qquad |q(x)| \le g, & \text{a.e. in } \Omega. \end{cases} \tag{6.42}$$

Based on the behavior of the solution and its multipliers on the biactive set, also different stationarity concepts arise in this case. The study of stationary points in this context is actually a matter of current research (see [15, 17]).

Definition 6.2. A point $\bar{u} \in L^2(\Omega)$ is called C(larke)-stationary for problem (6.42), if it satisfies the following system:

$$a(y,v) + (q,v)_{L^2} = (u,v)_{L^2}, \qquad \text{for all } v \in Y, \tag{6.43a}$$

$$q(x)y(x) = g|y(x)| \qquad \text{a.e. in } \Omega, \tag{6.43b}$$

$$|q(x)| \le g \qquad \text{a.e. in } \Omega, \tag{6.43c}$$

$$a(p,v) + \langle \xi, v\rangle_{Y',Y} = (y - z_d, v)_{L^2}, \qquad \text{for all } v \in Y \tag{6.43d}$$

$$\alpha u + p = 0 \qquad \text{a.e. in } \Omega \tag{6.43e}$$

$$p = 0 \qquad \text{a.e. in } \mathscr{I} := \{x : |q(x)| < g\}. \tag{6.43f}$$

and, additionally,

$$\langle \xi, p\rangle_{Y',Y} \ge 0, \qquad \langle \xi, y\rangle_{Y',Y} = 0. \tag{6.44}$$

A point $\bar{u} \in U$ is called strong stationary, if it satisfies (6.43) and

$$\langle \xi, v\rangle_{Y',Y} \ge 0, \qquad \forall v \in Y : v(x) = 0 \text{ where } |q(x)| < g \text{ and } v(x)q(x) \ge 0 \text{ a.e. in } B,$$

$$p(x)q(x) \geq 0 \qquad \text{a.e. in } B,$$

where $B := \{x : y = 0 \wedge |q| = g\}$ stands for the biactive set.

Since the biactive set is typically small, one is tempted to ignore its importance. The following simple example illustrates that optimal solutions may correspond to biactive points in a significant number of cases.

Example 6.2. For $u \in \mathbb{R}$, consider the finite-dimensional variational inequality

$$2y(v-y) + |v| - |y| \geq u(v-y), \text{ for all } v \in \mathbb{R},$$

and its energy formulation

$$\min_y j(y) = \{y^2 + |y| - uy\}. \tag{6.45}$$

Analyzing by cases, for $y \geq 0$ and $y \leq 0$, we get the following:

For $y \geq 0$: $j(y) = y^2 + (1-u)y$ and $j'(y) = 2y + (1-u)$. Therefore, the solution is given by

$$y = \begin{cases} \frac{1}{2}(u-1) & \text{if } u \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

For $y \leq 0$: $j(y) = y^2 - (1+u)y$ and $j'(y) = 2y - (1+u)$. Thus, the solution is given by

$$y = \begin{cases} \frac{1}{2}(u+1) & \text{if } u \leq -1, \\ 0 & \text{otherwise.} \end{cases}$$

Summarizing both cases, the solution to (6.45) is the following:

$$y = \begin{cases} \frac{1}{2}(u-1) & \text{if } u \geq 1, \\ 0 & \text{if } u \in [-1,1], \\ \frac{1}{2}(u+1) & \text{if } u \leq -1. \end{cases}$$

Considering the cost function

$$J(y,u) = \frac{1}{2}(y-1)^2 + \frac{\alpha}{2}u^2, \qquad \alpha \in (0,1),$$

and plotting the level curves, it can be observed that the minimum value will be attained either at $(0,0)$ or at some point $(y(\hat{u}),\hat{u}) \neq (0,1)$, depending on the value of α. Consequently, the optimal solution is not biactive.

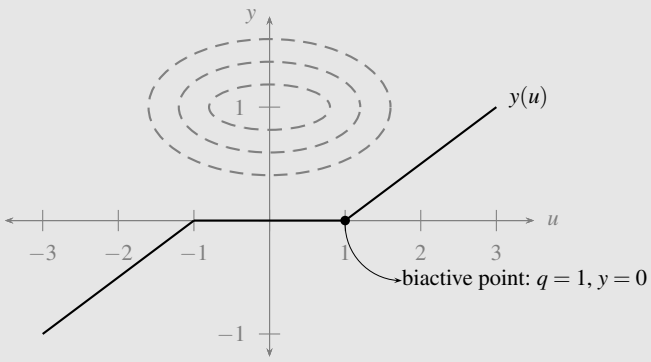

Fig. 6.2 Solution operator and contour lines of the cost function.

Considering a more general cost function:

$$J(y,u) = \frac{1}{2}(y-\xi_1)^2 + \frac{\alpha}{2}(u-\xi_2)^2, \qquad \alpha \in (0,1),$$

it can be verified that for $\xi_2 \leq -1$ and $\xi_1 \geq -2\alpha(1+\xi_2)$ the minimum is attained at $(\bar{y},\bar{u}) = (0,-1)$, which is a biactive point.

Note that equations (6.43b)–(6.43c) can also be formulated as the inclusion $q \in \partial g|y|$, where $\partial\phi$ stands for the subdifferential of ϕ. For solving the problem numerically, the subdifferential may be replaced by the following C^1-approximation of it:

$$h_\gamma(x) = \begin{cases} g\frac{x}{|x|} & \text{if } \gamma|x| \geq g + \frac{1}{2\gamma}, \\ \frac{x}{|x|}(g - \frac{\gamma}{2}(g-\gamma|x| + \frac{1}{2\gamma})^2) & \text{if } g - \frac{1}{2\gamma} \leq \gamma|x| \leq g + \frac{1}{2\gamma}, \\ \gamma x & \text{if } \gamma|x| \leq g - \frac{1}{2\gamma}, \end{cases} \tag{6.46}$$

for γ sufficiently large.

Proposition 6.1. *The regularizing function h_γ satisfies the following approximation property:*

$$\left| h_\gamma(x) - \frac{g\gamma x}{\max(g, \gamma|x|)} \right| \leq \frac{1}{\gamma}, \quad \text{for all } x \in \mathbb{R}. \tag{6.47}$$

Proof. If x is such that $\gamma|x| \leq g - \frac{1}{2\gamma}$ or $\gamma|x| \geq g + \frac{1}{2\gamma}$, the results is obvious. Let x be such that $g < \gamma|x| \leq g + \frac{1}{2\gamma}$. It then follows that $\max(g, \gamma|x|) = \gamma|x|$ and

$$\left| h_\gamma(x) - \frac{g\gamma x}{\max(g, \gamma|x|)} \right| \leq \left| \frac{\gamma}{2}(g - \gamma|x| + \frac{1}{2\gamma})^2 \right| \leq \frac{\gamma}{2}\left(\frac{1}{2\gamma}\right)^2 = \frac{1}{8\gamma}.$$

If x is such that $g - \frac{1}{2\gamma} \leq \gamma|x| \leq g$, then

$$\begin{aligned} \left| h_\gamma(x) - \frac{g\gamma x}{\max(g, \gamma|x|)} \right| &\leq \left| \left(g - \frac{\gamma}{2}(g - \gamma|x| + \frac{1}{2\gamma})^2 \right) \frac{x}{|x|} - \gamma x \right| \\ &\leq |g - \gamma|x|| + \left| \frac{\gamma}{2}(g - \gamma|x| + \frac{1}{2\gamma})^2 \right| \\ &\leq \frac{1}{2\gamma} + \frac{\gamma}{2}\left(\frac{1}{\gamma}\right)^2 = \frac{1}{\gamma}. \end{aligned}$$

□

By using the function h_γ, the following family of regularized equations is obtained:

$$a(y, v) + (q, v)_{L^2(\Omega)} = (u, v)_{L^2(\Omega)}, \qquad \text{for all } v \in Y, \tag{6.48a}$$

$$q = h_\gamma(y) \qquad \text{a.e. in } \Omega. \tag{6.48b}$$

Existence of a unique solution to (6.48) follows from the monotonicity of h_γ. In the next theorem, convergence of the regularized solutions toward the original one is shown.

Theorem 6.9. *Let y and y_γ be solutions to (6.41) and (6.48), respectively, both with $u \in L^2(\Omega)$ on the right hand side. Then*

$$y_\gamma \to y \text{ strongly in } Y \qquad \text{as } \gamma \to +\infty.$$

Proof. Let $\hat{y}$ denote the unique solution to the auxiliary problem:

$$a(\hat{y}, v) + (\hat{q}, v) = (u, v)_{L^2}, \text{ for all } v \in Y, \tag{6.49a}$$

$$\hat{q} = g\gamma \frac{\hat{y}}{\max(g, \gamma|\hat{y}|)}. \tag{6.49b}$$

Existence and uniqueness of a solution to (6.49) can be obtained by using Minty–Browder's theorem.

Taking the difference between the solutions to (6.41) and (6.48), and using the triangle inequality, we obtain that

$$\|y - y_\gamma\|_Y \leq \|y - \hat{y}\|_Y + \|\hat{y} - y_\gamma\|_Y. \tag{6.50}$$

For the first term on the right hand side of (6.50) we take the difference between (6.41a) and (6.49a), with $v = y - \hat{y}$. We obtain that

$$\kappa \|y - \hat{y}\|_Y^2 \leq (q - \hat{q}, \hat{y} - y), \text{for some } \kappa > 0. \tag{6.51}$$

Analyzing the last term pointwisely, we get the following:

On $\mathscr{A}_\gamma := \{x : \gamma|\hat{y}| \geq g\}$, thanks to (6.41b)–(6.41c):

$$\begin{aligned}(\hat{q}(x) - q(x))(y(x) - \hat{y}(x)) &= g\frac{\hat{y}(x)}{|\hat{y}(x)|}y(x) - q(x)y(x) - g|\hat{y}(x)| + q(x)\hat{y}(x) \\ &\leq g|y(x)| - g|y(x)| - g|\hat{y}(x)| + g|\hat{y}(x)| = 0.\end{aligned}$$

On $\mathscr{I}_\gamma := \Omega \backslash \mathscr{A}_\gamma = \{x : \gamma|\hat{y}| < g\}$, using (6.41b)–(6.41c) and the set's definition:

$$\begin{aligned}(\hat{q}(x) - q(x))(y(x) - \hat{y}(x)) &= \gamma\hat{y}(x)y(x) - \gamma|\hat{y}(x)|^2 - q(x)y(x) - q(x)\hat{y}(x) \\ &\leq (g - \gamma|\hat{y}(x)|)|\hat{y}(x)| \\ &\leq (g - \gamma|\hat{y}(x)|)\frac{g}{\gamma} \leq \frac{g^2}{\gamma}.\end{aligned}$$

Altogether we get that

$$(q - \hat{q}, \hat{y} - y)_{L^2} < \frac{g^2}{\gamma}|\Omega|, \tag{6.52}$$

and, consequently,

$$\|y - \hat{y}\|_Y \leq \frac{C_1}{\sqrt{\gamma}}, \qquad \text{for some constant } C_1 > 0. \tag{6.53}$$

For the second term on the right hand side of (6.50), by taking the difference between (6.48) and (6.49), we obtain that

$$a(y_\gamma - \hat{y}, y_\gamma - \hat{y}) + \big(h_\gamma(y_\gamma) - h_\gamma(\hat{y}), y_\gamma - \hat{y}\big)_{L^2} = -\left(h_\gamma(\hat{y}) - \frac{g\gamma\hat{y}}{\max(g, \gamma|\hat{y}|)}, y_\gamma - \hat{y}\right)_{L^2}.$$

Using the coercivity of $a(\cdot,\cdot)$ and the monotonicity of h_γ, we then get that

$$\kappa\|y_\gamma-\hat{y}\|_Y^2 \le -\left(h_\gamma(y_\gamma)-\frac{g\gamma\hat{y}}{\max(g,\gamma|\hat{y}|)},y_\gamma-\hat{y}\right)_{L^2}, \text{ for some } \kappa>0.$$

Thanks to (6.47), it follows that

$$\left|\left(h_\gamma(y_\gamma)-\frac{g\gamma\hat{y}}{\max(g,\gamma|\hat{y}|)},y_\gamma-\hat{y}\right)_{L^2}\right| \le \frac{1}{\gamma}\int_\Omega |y_\gamma-\hat{y}|\,dx,$$

which implies that

$$\|y_\gamma-\hat{y}\|_Y^2 \le \frac{C_2}{\gamma}\|y_\gamma-\hat{y}\|_Y, \tag{6.54}$$

for some $C_2>0$. From (6.53) and (6.54), the result is obtained. □

A family of regularized PDE-constrained optimization problems is also obtained by using (6.48). The problems read as follows:

$$\begin{cases} \min\ J(y,u)=\frac{1}{2}\int_\Omega |y-z_d|^2\,dx+\frac{\alpha}{2}\|u\|_{L^2}^2, \\ \text{subject to:} \\ \qquad a(y,v)+(q,v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \\ \qquad q=h_\gamma(y) \qquad \text{a.e. in } \Omega. \end{cases} \tag{6.55}$$

Existence of an optimal solution for each regularized problem can be argued by classical techniques. Moreover, by using the techniques of Chap. 3, the following optimality systems are obtained:

$$a(y,v)+(q,v)_{L^2}=(u,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.56a}$$
$$q=h_\gamma(y) \qquad \text{a.e. in } \Omega, \tag{6.56b}$$
$$a(p,v)+(\lambda_\gamma,v)_{L^2}=(y-z_d,v)_{L^2}, \qquad \text{for all } v\in Y, \tag{6.56c}$$
$$\lambda_\gamma=h_\gamma'(y)^*p \qquad \text{a.e. in } \Omega, \tag{6.56d}$$
$$\alpha u+p=0 \qquad \text{a.e. in } \Omega. \tag{6.56e}$$

Theorem 6.10. *Every sequence $\{u_\gamma\}_{\gamma>0}$ of solutions to (6.55) contains a subsequence which converges strongly in $L^2(\Omega)$ to an optimal solution $\bar{u}\in L^2(\Omega)$ of (6.42).*

Proof. The proof is similar to the one of Theorem 6.8, with $\bar{u}$ denoting an optimal solution for (6.42). From the structure of the cost functional it follows that $\{u_\gamma\}$ is uniformly bounded in $L^2(\Omega)$ and, therefore, there exists a weakly convergent subsequence $\{u_\gamma\}$ and a limit point $\hat{u}$ such that $u_\gamma \rightharpoonup \hat{u}$ weakly in $L^2(\Omega)$.

Thanks to Theorem 6.9, $y_\gamma(u_\gamma) \to y(u_\gamma)$ strongly in Y. Moreover, from the structure of the variational inequality, we get that

$$a(y(u_\gamma), y(\hat{u}) - y(u_\gamma)) + j(y(\hat{u})) - j(y(u_\gamma)) \geq (u_\gamma, y(\hat{u}) - y(u_\gamma))$$

and

$$a(y(\hat{u}), y(u_\gamma) - y(\hat{u})) + j(y(u_\gamma)) - j(y(\hat{u})) \geq (\hat{u}, y(u_\gamma) - y(\hat{u})).$$

Adding both inequalities and using the ellipticity of the bilinear form, we get that:

$$\|y(u_\gamma) - y(\hat{u})\|_Y \leq C\|u_\gamma - \hat{u}\|_{Y'},$$

for some constant $C > 0$. Altogether we proved that the regularized states $y_\gamma(u_\gamma) \to y(\hat{u})$ strongly in Y.

Thanks to the weakly lower semicontinuity of the cost functional, we obtain that $(y(\hat{u}), \hat{u})$ is an optimal solution to (6.42) and, also that,

$$\lim_{\gamma \to \infty} \|u_\gamma\|_U^2 = \|\bar{u}\|_U^2,$$

which implies strong convergence. □

Remark 6.2. By passing to the limit in the regularized systems (6.56), an optimality system of C-type is obtained for any accumulation point of $\{u_\gamma\}_{\gamma>0}$ (see [15]).

The derivative of $h_\gamma(x)$ is given by the function

$$h'_\gamma(x) = \begin{cases} 0 & \text{if } \gamma|x| \geq g + \frac{1}{2\gamma} \\ \gamma^2(g - \gamma|x| + \frac{1}{2\gamma}) & \text{if } g - \frac{1}{2\gamma} \leq \gamma|x| \leq g + \frac{1}{2\gamma} \\ \gamma & \text{if } \gamma|x| \leq g - \frac{1}{2\gamma}, \end{cases} \tag{6.57}$$

which is piecewise continuous and semismooth. By replacing the multipliers q_γ and λ_γ in the state and adjoint equations, respectively, a generalized Newton step for both equations is given by

$$a(\delta_y, v) + (h'_\gamma(y_k)\delta_y, v)_{L^2} - (\delta_u, v)_{L^2} = -a(y, v) - (h_\gamma(y_k), v)_{L^2} + (u, v)_{L^2}, \tag{6.58}$$

$$\begin{aligned} a(\delta_p, v) + (h'_\gamma(y_k)\,\delta_p, v)_{L^2} + (h''_\gamma(y_k)\delta_y\, p_k, v)_{L^2} - (\delta_y, v)_{L^2} \\ = -a(p, v) - (h'_\gamma(y_k)\,p, v)_{L^2} + (y - z_d, v)_{L^2}, \end{aligned} \tag{6.59}$$

where

$$h''_\gamma(x) := \begin{cases} -\gamma^3 \frac{x}{|x|} & \text{if } g - \frac{1}{2\gamma} \le \gamma|x| \le g + \frac{1}{2\gamma} \\ 0 & \text{elsewhere.} \end{cases} \tag{6.60}$$

The complete algorithm is given next.

Algorithm 11 SSN–Optimization with VI constraints (second kind)

1: Choose $(y_0, u_0, p_0) \in Y \times U \times W'$.
2: **repeat**
3: Solve for $(\delta_y, \delta_u, \delta_\phi) \in Y \times U \times W'$

$$\begin{pmatrix} A + h'_\gamma(y_k) & -I & 0 \\ h''_\gamma(y_k)\, p - I & 0 & A + h'_\gamma(y_k) \\ 0 & \alpha I & I \end{pmatrix} \begin{pmatrix} \delta_y \\ \delta_u \\ \delta_p \end{pmatrix} = - \begin{pmatrix} Ay_k + h_\gamma(y_k) - u_k \\ Ap_k + h'_\gamma(y_k)\, p_k - y_k + z_d \\ \alpha u_k + p_k \end{pmatrix}.$$

4: Set $u_{k+1} = u_k + \delta_u$, $y_{k+1} = y_k + \delta_y$, $p_{k+1} = p_k + \delta_p$ and $k = k + 1$.
5: **until** Stopping criteria.

Remark 6.3. If the operator K in (6.40) is not the canonical injection, globalization strategies may be needed for the semismooth Newton method to converge (see [16]).

References

1. E. L. Allgower, K. Boehmer, F. A. Potra and W. C. Rheinboldt. A mesh independence principle for operator equations and their discretization. *SIAM J. Numer. Anal.*, 23(1):160–169, 1986.
2. W. Alt. Mesh-Independence of the LagrangeNewton method for nonlinear optimal control problems and their discretizations. *Annals of Operations Research*, 101, 101–117, 2001.
3. K. Atkinson and W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer Verlag, 2009.
4. V. Barbu. *Analysis and Control of nonlinear infinite dimensional systems*. Academic Press, New York, 1993.
5. M. Bergounioux. Optimal control of problems governed by abstract elliptic variational inequalities with state constraints. *SIAM Journal on Control and Optimization*, 36(1):273–289, 1998.
6. M. Bergounioux. K Ito and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM Journal on Control and Optimization*, 37(4):1176–1194, 1999.
7. M. Bergounioux and F. Mignot. Optimal control of obstacle problems: existence of Lagrange multipliers. *ESAIM: Control, Optimisation and Calculus of Variations*, Vol. 5, 4570, 2000.
8. A. Borzi and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.
9. H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer-Verlag, 2011.
10. A. Carnarius, J. C. De los Reyes, B. Günther, F. Thiele, F. Tröltzsch, D. Wachsmuth. Numerical study of the optimization of separation control. *AIAA Paper 2007-0058*, 2007.
11. E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints, *SIAM Journal on Control and Optimization,* 31(4):993–1006, 1993.
12. P. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.
13. J.C. De los Reyes and K. Kunisch. A semi-smooth Newton method for control constrained boundary optimal control of the Navier-Stokes equations. *Nonlinear Analysis. Theory, Methods & Applications*, 62(7):1289–1316, 2005.
14. J.C. De los Reyes and F. Tröltzsch. Optimal control of the stationary Navier-Stokes equations with mixed control-state constraints. *SIAM Journal on Control and Optimization,* 46: 604–629, 2007.

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9

15. J. C. De los Reyes. Optimal control of a class of variational inequalities of the second kind, *SIAM Journal on Control and Optimization,* 49(4):1629–1648, 2011.
16. J.C. De los Reyes. Optimization of mixed variational inequalities arising in flows of viscoplastic materials. *Computational Optimization and Applications*, 52:757–784, 2012.
17. J.C. De los Reyes and C. Meyer. Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the 2nd kind, *ArXiv e-prints,* 1404.4787, 2014.
18. G. Duvaut and J.L. Lions, *Inequalities in mechanics and physics*, Springer-Verlag, Berlin, 1976.
19. L.C. Evans *Partial differential equations.* American Mathematical Society, Providence, RI, 2010.
20. R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer Series in Computational Physics, Springer-Verlag, 1984.
21. R. Glowinski, T.W. Pan, A. Kearsley and J. Periaux. Numerical simulation and optimal shape for viscous flow by a fictitious domain method, *International Journal for Numerical Methods in Fluids*, 20(8–9):695–711, 1995.
22. A. Griewank. The local convergence of Broyden-like methods on Lipschitzian problems in Hilbert spaces. *SIAM J. Numer. Anal.*, 24(3), 684–705, 1987.
23. C. Grossmann, H.-G. Roos and M. Stynes *Numerical Treatment of Partial Differential Equations.* Springer-Verlag, 2007.
24. M. Gunzburger, L. Hou and T. Svobodny. Boundary velocity control of incompressible flow with an application to viscous drag reduction. *SIAM Journal on Control and Optimization*, 30(1):167–181, 1992.
25. M. Gunzburger *Perspectives in Flow Control and Otimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.
26. S. B. Hazra and V. Schulz. Simultaneous Pseudo-Timestepping for Aerodynamic Shape Optimization Problems with State Constraints. *SIAM J. Sci. Comput.*, 28(3), 10781099, 2006.
27. R. Herzog, C. Meyer, and G. Wachsmuth. B-and strong stationarity for optimal control of static plasticity with hardening. *SIAM Journal on Optimization*, 23(1):321–352, 2013.
28. M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C-and strong stationarity and a path-following algorithm. *SIAM Journal on Optimization*, 20(2):868–902, 2009.
29. M. Hintermüller, B. Mordukhovich, and T. Surowiec. Several approaches for the derivation of stationarity conditions for elliptic mpecs with upper-level control constraints. , to appear.
30. M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich. *Optimization with PDE Constraints*. Springer Verlag, 2010.
31. W. W. Hu, *Approximation and Control of the Boussinesq Equations with Application to Control of Energy Efficient Building Systems*. Ph.D. Thesis, Department of Mathematics, Virginia Tech, 2012.
32. K. Ito and K. Kunisch. Augmented Lagrangian formulation of nonsmooth, convex optimization in Hilbert spaces. In *Lecture Notes in Pure and Applied Mathematics. Control of Partial Differential Equations and Applications, E. Casas (Ed.)*, Vol. 174, 107–117, Marcel Dekker, 1995.
33. K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Applied Mathematics and Optimization*, 41, 343364, 2000.
34. K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
35. J. Jahn. *Introduction to the theory of nonlinear optimization*. Springer Verlag, 2007.
36. E. Kalnay. *Atmospheric modeling, data assimilation, and predictability*. Cambridge university press, 2003.

37. C. T. Kelley. *Iterative Methods for Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.
38. C. T. Kelley and E. W. Sachs. Quasi-Newton methods and unconstrained optimal control problems. *SIAM J. Control and Optimization*, 25, 1503–1517, 1987.
39. C. T. Kelley and E. W. Sachs. A new proof of superlinear convergence for Broyden's method in Hilbert space. *SIAM J. Optim.*, 1, pp. 146–150, 1991.
40. C. T. Kelley and E. W. Sachs. Mesh independence of the gradient projection method for optimal control problems. *SIAM J. Optim.*, 30:2, pp. 477–493, 1992.
41. C. T. Kelley and E. W. Sachs. Solution of optimal control problems by a pointwise projected Newton method. *SIAM J. Optim.*, 33:6, pp. 1731–1757, 1995.
42. K. Kunisch and D. Wachsmuth. Sufficient optimality conditions and semi-smooth newton methods for optimal control of stationary variational inequalities. *ESAIM: Control, Optimisation and Calculus of Variations*, 180:520–547, 2012.
43. D. G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons Inc., New York, 1969.
44. Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, Cambridge, 1996.
45. C. Meyer, P. Philip and F. Troeltzsch. Optimal control of a semilinear PDE with nonlocal radiation interface conditions. *SIAM J. Control Optimization*, 45 (2006), 699–721.
46. F. Mignot and J.-P. Puel. Optimal control in some variational inequalities. *SIAM J. Control Optim.*, 22(3):466–476, 1984.
47. P. Neittanmaaki, J. Sprekels and D. Tiba. *Optimization of Elliptic Systems: Theory and Applications.* Springer Verlag 2010.
48. J. Nocedal and S. Wright. *Numerical Optimization.* Springer Verlag 1999.
49. J. V. Outrata. A generalized mathematical program with equilibrium constraints. *SIAM J. Control Optim.*, 38(5):1623–1638 (electronic), 2000.
50. J. Outrata, J. Jarušek, and J. Stará. On optimality conditions in control of elliptic variational inequalities. *Set-Valued and Variational Analysis*, 19(1):23–42, 2011.
51. E. Polak. An historical survey of computational methods in optimal control. *SIAM Review,* 15(2):553–584, 1973.
52. A. Quarteroni. *Numerical Models for Differential Problems.* Springer Verlag, 2012.
53. S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. MIT Press, 2012.
54. G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Computational Optimization and Applications,* 44:159–181, 2009.
55. J. Stoer. Two examples on the convergence of certain rank-2 minimization methods for quadratic functionals in Hilbert space. *Linear Algebra and its Applications*, 28, 217–222, 1979.
56. F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications.* American Mathematical Society, 2010.
57. M. Ulbrich. Constrained optimal control of Navier-Stokes flow by semismooth Newton methods. *Systems and Control Letters*, 48:297–311, 2003.
58. M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung.* Birkhäuser, 2012.

Index

J. C. De los Reyes, *Numerical PDE-Constrained Optimization*,
SpringerBriefs in Optimization, DOI 10.1007/978-3-319-13395-9