Lecture Notes in Geosystems Mathematics and Computing

Mathias Richter

Inverse Problems

Basics, Theory and Applications in Geophysics

Birkhäuser

Lecture Notes in Geosystems Mathematics and Computing

Birkhäuser

More information about this series at http://www.springer.com/series/15481

Mathias Richter

Inverse Problems

Basics, Theory and Applications in Geophysics

Birkhäuser

Mathias Richter
Fakultät für Elektrotechnik und
Informationstechnik
Universität der Bundeswehr München
Neubiberg, Germany

Lecture Notes in Geosystems Mathematics and Computing
ISBN 978-3-319-48383-2 ISBN 978-3-319-48384-9 (eBook)
DOI 10.1007/978-3-319-48384-9

Library of Congress Control Number: 2016960201

Printed on acid-free paper

This book is published under the trade name Birkhäuser (www.birkhauser-science.com)
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The term "inverse problem" has no acknowledged mathematical definition; its meaning relies on notions from physics. One assumes that there is a known mapping

$$T : \mathbb{U} \to \mathbb{W},$$

which models a physical law or a physical device. Here, $\mathbb{U}$ is a set of "causes" and $\mathbb{W}$ is a set of "effects." The computation of an effect $T(u)$ for a given cause u is called a **direct problem**. Finding a cause $u \in \mathbb{U}$, which entails a given effect $w \in \mathbb{W}$, is called an **inverse problem**. Solving an inverse problem thus means to ask for the solution of an equation $T(u) = w$.

Maybe a certain effect $w \in \mathbb{W}$ is desirable and one is looking for some $u \in \mathbb{U}$ to produce it. An inverse problem of this kind is called a **control problem**. In the following, it will be assumed that an effect is actually observed and that its cause has to be found. An inverse problem of this kind is called an **identification problem**. It arises, when an interesting physical quantity is not directly amenable to measurements, but can only be derived from observations of its effects. There are numerous examples of identification problems in science and engineering like:

- **Inverse gravimetry**: Given its mass density distribution (cause), the gravitational force (effect) exercised by a body can be computed (direct problem). The inverse problem is to derive the mass density distribution from measured gravitational forces. An application is the detection of oil or gas reservoirs in geological prospecting.
- **Transmission tomography**: Given the density distribution of tissue (cause), the intensity loss (effect) of an X-ray traversing it can be computed (direct problem). Inversely, one tries to find the tissue's density from measured intensity losses of X-rays traversing it. Applications exist in diagnostic radiology and nondestructive testing.
- **Elastography**: From known material properties (cause), the displacement field (effect) of an elastic body under external forces can be computed (direct

problem). The inverse problem is to derive material properties from observed displacements. An application is the medical imaging of soft tissue.
- **Seismic tomography**: Knowing the spatial distribution of mechanical characteristics (cause) of the earth's crust and mantle, seismic energy propagation (effect) from controlled sound sources or earthquakes can be computed (direct problem). Inversely, one seeks to find mechanical characteristics from records of seismic data. A geophysical application is to map the earth's interior.

Very often, the solution u^* of an inverse problem $T(u) = w$ depends in an extremely sensitive manner on w, two very similar effects having very different causes. This would not be a principal problem, if w was known exactly. However, in practice, one never knows w perfectly well but can only construct some approximation $\tilde{w}$ of it from a limited number of generally inexact measurements. Then the true cause $u^* \in \mathbb{U}$ defined by $T(u) = w$ cannot even be found to good approximation, no matter the available computing power. This is so because w is not known, and, by the aforementioned sensitivity, solving $T(u) = \tilde{w}$ would produce a completely wrong answer $\tilde{u}$, far away from u^*. There is no way to miraculously overcome this difficulty, unless we are lucky and have some *additional information* of the kind "the true cause u^* has property P." If this is the case, then we might replace the problem of solving $T(u) = \tilde{w}$ by the *different problem* of finding a cause $\tilde{u}$ within the set of all candidates from $\mathbb{U}$ *also having property* P, such that some distance between $T(\tilde{u})$ and $\tilde{w}$ is minimized. If the solution $\tilde{u}$ of this replacement problem depends less sensitively on the effect $\tilde{w}$ than the solution u^* of the original problem depends on w but converges to u^* if $\tilde{w}$ converges to w, then one speaks of a **regularization** of the original inverse problem.

Scope

Many excellent books on inverse problems have been written; we only mention [Kir96, EHN96], and [Isa06]. These books rely on functional analysis as an adequate mathematical tool for a unified approach to analysis and regularized solution of inverse problems. Functional analysis, however, is not easily accessible to non-mathematicians. It is the first goal of the present book to provide an access to inverse problems without requiring more mathematical knowledge than is taught in undergraduate math courses for scientists and engineers. From abstract analysis, we will only need the concept of functions as vectors. Function spaces are introduced informally in the course of the text, when needed. Additionally, a more detailed but still condensed introduction is given in Appendix B. We will not deal with regularization theory for operators. Instead, inverse problems will first be discretized and described approximately by systems of algebraic equations and only then will be regularized by setting up a replacement problem, which will always be a minimization problem. A second goal is to elaborate on the single steps to be taken when solving an inverse problem: discretization, regularization, and practical

solution of the regularized optimization problem. Rather than being as general as possible, we want to work out these steps for model problems from the fields of inverse gravimetry and seismic tomography. We will not delve into details of numerical algorithms, though, when high-quality software is readily available on which we can rely for computations. For the numerical examples in this book, the programming environment Matlab ([Mat14]) was used as well as were C programs from [PTVF92].

Content

We start in Chap. 1 by presenting four typical examples of inverse problems, already illustrating the sensitivity issue. We then formalize inverse problems as equations in vector spaces and also formalize their sensitivity as "ill-posedness." In the chapter's last two sections, we have a closer look on problems from inverse gravimetry and seismic tomography. We define specific model problems that we will tackle and solve in later chapters. Special attention will be payed to the question whether sought-after parameters (causes) can be uniquely identified from observed effects, at least in the hypothetical case of perfect observations.

All model problems introduced in Chap. 1 are posed in function spaces: effects and causes are functions. Chapter 2 is devoted to the discretization of such problems, that is, to the question of how equations in function spaces can approximately be turned into equations in finite-dimensional spaces. Discretization is a prerequisite for solving inverse problems on a computer. A first section on spline approximation describes how general functions can be represented approximately by a finite set of parameters. We then focus on linear inverse problems. For these, the least squares method, investigated in Sect. 2.2, is a generally applicable discretization method. Two alternatives, the collocation method and the method of Backus-Gilbert, are presented in Sects. 2.3 and 2.4 for special but important classes of inverse problems, known as Fredholm equations of the first kind. Two of our model problems belong to this class. Even more specific are convolutional equations, which play a significant role in inverse gravimetry but also in other fields like communications engineering. Convolutional equations can be discretized in the Fourier domain, which leads to very efficient solution algorithms. This is discussed in Sect. 2.5. In the nonlinear case, discretizations are tailored to the specific problem to be solved. We present discretizations of two nonlinear model problems from inverse gravimetry and seismic tomography in Sect. 2.6.

The last two chapters investigate possibilities for a regularized solution of inverse problems in finite-dimensional spaces (i.e., after discretization). In Chap. 3, we treat the linear case, where the discretized problem takes the form of a linear system of equations $Ax = b$. More generally, one considers the linear least squares problem of minimizing $\|b - Ax\|_2$, which is equivalent to solving $Ax = b$, if the latter *does* have a solution, and is still solvable, if the system $Ax = b$ no longer is. In the first two sections, the linear least squares problem is analyzed, with an emphasis on

quantifying the sensitivity of its solution with respect to the problem data, consisting of the vector b and of the matrix A. A measure of sensitivity can be given, known as condition number in numerical analysis, which allows to estimate the impact of data perturbations on the solution of a linear least squares problem. If the impact is too large, a meaningful result cannot be obtained, since data perturbations can never be completely avoided. The least squares problem then is not solvable practically. Regularization tries to set up a new problem, the solution of which is close to the actually desired one *and* which can be computed reliably. The general ideas behind regularization are outlined in Sect. 3.3. The most popular regularization technique, Tikhonov regularization, is explained and analyzed in Sect. 3.4, including a discussion of how a regularized solution can be computed efficiently. Tikhonov regularization requires the choice of a so-called regularization parameter. The proper choice of this parameter is much debated. In Sect. 3.5, we present only a single method to make it, the discrepancy principle, which is intuitively appealing and often successful in practice. For alternative choices, we refer only to the literature. Tikhonov regularization can also be used to solve problems derived by the Backus-Gilbert method or for convolutional equations transformed to the Fourier domain. These topics are treated in Sects. 3.7 and 3.8. Very interesting alternatives to Tikhonov regularization are iterative regularization methods, which are attractive for their computational efficiency. Two of these methods, the Landweber iteration and the conjugate gradient method, are described in Sects. 3.9 and 3.10. It will be found that the Landweber iteration can be improved by two measures which relate it back to Tikhonov regularization. One of these measures, a coordinate transform, can also be taken to improve the conjugate gradient method. Technical details about this coordinate transform are given in Sect. 3.6, which describes a transformation of Tikhonov regularization to some standard form. This is also of independent interest, since Tikhonov regularization in standard form is the easiest to analyze.

Regularization of nonlinear problems is studied in Chap. 4. In Sect. 4.1, nonlinear Tikhonov regularization is treated abstractly, whereas in Sect. 4.2, this method is applied to a model problem from nonlinear inverse gravimetry. Nonlinear Tikhonov regularization leads to an optimization problem, namely, a nonlinear least squares problem. Section 4.3 discusses various possibilities for solving nonlinear least squares problem numerically. All of these methods require the computation of gradients, which can mean a prohibitive computational effort, if not done carefully. The "adjoint method," presented in Sect. 4.4, can sometimes drastically reduce the numerical effort to obtain gradients. The adjoint method is presented in the context of a nonlinear problem from seismic tomography, which is then solved in Sect. 4.5. A final section presents one example of a nonlinear iterative regularization method, the "inexact Newton-CG method." The usefulness of this method is illustrated for a problem of inverse gravimetry.

Appendix A lists required material from linear algebra and derives the singular value decomposition of a matrix. Appendix B gives a condensed introduction into function spaces, i.e., into abstract vector spaces containing functions as elements. Appendix C gives an introduction into the subject of (multidimensional) Fourier transforms and their numerical implementation, including the case of

non-equidistant sample points. Finally, Appendix D contains a technical proof outsourced from Chap. 3, which shows the regularizing property of the conjugate gradient method applied to linear least squares problems.

Acknowledgments

I am much indebted to my editor, Dr. Clemens Heine, from Birkhäuser-Springer publishers for his promotion of this book. I am greatly thankful to my friend and colleague, Professor Stefan Schäffler, for his critical advice when reading my manuscript and for having supported my work for well over 20 years.

Munich, Germany Mathias Richter

Contents

Chapter 1
Characterization of Inverse Problems

In Sect. 1.1 we present four typical examples of inverse problems from physics and engineering and already give an idea of why these problems can be difficult to solve in practice. In all examples causes as well as effects are functions. Therefore, any useful notion of inverse problems as equations relating causes to effects must be general enough to include equations in function spaces, like differential or integral equations. Section 1.2 begins with an informal introduction of function spaces by way of an example, a more formal introduction is provided in Appendix B. Later in this section we indeed describe inverse problems as equations in vector spaces and define "ill-posedness", which is the primary concern when dealing with inverse problems. In Sects. 1.3 and 1.4, we formulate four model problems from inverse gravimetry and seismic tomography, which will be taken up in later chapters to illustrate the methods presented in this book.

1.1 Examples of Inverse Problems

Example 1.1 (Determination of growth rates) The initial value problem

$$w'(t) = \frac{dw(t)}{dt} = u(t) \cdot w(t), \quad w(t_0) = w_0 > 0, \quad t_0 \le t \le t_1, \quad t_0 < t_1, \tag{1.1}$$

is a simple and well known mathematical model for a growth process. Here, $w(t)$ might represent the population size of a colony of bacteria at time t. Then, w_0 represents the known initial population size and $u(t)$ represents a variable bacterial growth rate. For a given, continuous function $u : [t_0, t_1] \to \mathbb{R}$, (1.1) has a unique,

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9_1

continuously differentiable solution $w : [t_0, t_1] \to (0, \infty)$, to wit:

$$w(t) = w_0 \cdot e^{U(t)}, \quad U(t) = \int_{t_0}^{t} u(s)\, ds, \quad t_0 \le t \le t_1. \tag{1.2}$$

Thus, a cause u (the growth rate) entails an effect w (the population size). Formally, this can be described as a mapping $T : u \mapsto w$, parameterized by t_0, t_1, and w_0 (a mathematically correct definition of this mapping from one function space into another is deferred to Sect. 1.2). Inversely, for a given function w, one asks for a function u such that $T(u) = w$. This inverse problem also has a unique solution, explicitly given by:

$$u(t) = \frac{w'(t)}{w(t)} = \frac{d}{dt} \ln(w(t)), \quad t_0 \le t \le t_1. \tag{1.3}$$

Function u is the input of the direct problem ("the data") and function w is its result. For the inverse problem, function w is the input and function u is the result. In practice, we never know the respective inputs exactly, since this would mean to exactly know infinitely many values $u(t)$ (for the direct problem) or $w(t)$ (for the inverse problem). Rather, we only have a finite number of measurements, subject to measurement errors. From these measurements, we can only approximate the true input function.

For the direct problem, such unavoidable errors do not have serious consequences. To see this, assume that $\tilde{u}$ is an approximation of the true input u with

$$\max\{|u(t) - \tilde{u}(t)|;\ t_0 \le t \le t_1\} \le \varepsilon \quad \text{for some} \quad \varepsilon > 0.$$

Then, for the corresponding results $\tilde{w}$ and w we have

$$\max\{|w(t) - \tilde{w}(t)|;\ t_0 \le t \le t_1\} \le \varepsilon C$$

with some constant C.[1] This means that deviations of results can be kept under control if one cares to keep deviations of inputs small enough. This quality of the mapping from inputs to outputs is called **stability** or **robustness** of the direct

[1] From Theorem 12.V in [Wal98] we get

$$C = \frac{w_0}{\mu} \mathrm{e}^{(\mu+\varepsilon)(t_1-t_0)} (\mathrm{e}^{\mu(t_1-t_0)} - 1) \quad \text{where} \quad \mu := \max\{|u(t)|;\ t_0 \le t \le t_1\}.$$

problem. The inverse problem behaves very differently. For example, take

$$\text{input} \quad \left\{ \begin{array}{rcl} w : [t_0, t_1] & \to & \mathbb{R} \\ t & \mapsto & \mathrm{e}^{\sin(t)} \end{array} \right\} \quad \text{and result} \quad \left\{ \begin{array}{rcl} u : [t_0, t_1] & \to & \mathbb{R} \\ t & \mapsto & \cos(t) \end{array} \right\}$$

and consider the specific perturbed inputs

$$w_n : [t_0, t_1] \to \mathbb{R}, \quad t \mapsto w(t) \cdot \left(1 + \frac{1}{\sqrt{n}} \cos(nt)\right), \quad n \in \mathbb{N}, n \geq 2, \tag{1.4}$$

with $\max\{|w_n(t) - w(t)|;\ t_0 \leq t \leq t_1\} \to 0$ for $n \to \infty$. Inputs w_n lead to results

$$u_n : [t_0, t_1] \to \mathbb{R}, \quad t \mapsto u(t) - \frac{\sqrt{n} \sin(nt)}{1 + \frac{1}{\sqrt{n}} \cos(nt)}, \quad n \in \mathbb{N}, n \geq 2,$$

with

$$\max\{|u_n(t) - u(t)|;\ t_0 \leq t \leq t_1\} \to \infty \quad \text{for} \quad n \to \infty.$$

The better $w_n(t)$ approximates $w(t)$, the larger becomes the deviation of $u_n(t)$ from $u(t)$! The reason for this is the differentiation operation in (1.3). Whereas integration as in (1.2) is a smoothing operation (which, for example, will level out sharp function peaks), the inverse differentiation operation necessarily must roughen all function details which are smoothed by integration. But this will eventually also blow up perturbations of the true input function. As a consequence, the explicit formula (1.3), albeit mathematically correct, is practically useless. Figure 1.1 illustrates this example for $n = 1000$. The graphs of input w and result u are shown as solid fat lines on the left and on the right, respectively. The graphs of w_n und u_n oscillate wildly in the shaded regions between the dashed lines. ◇

The next example probably is the best known inverse problem.

Example 1.2 (Computerized tomography, CT) A plane section of a body is characterized by a density function $f : \mathbb{R}^2 \to \mathbb{R}$, $x \mapsto f(x)$, where we assume that

$$f(x) = 0 \text{ for } x \notin D := \{x \in \mathbb{R}^2;\ \|x\|_2 < 1\} \subseteq \mathbb{R}^2,$$

which can always be achieved by proper scaling. The value $f(x)$ models the tissue's density at position x. Knowing $f(x)$ at every position $x \in D$ means to know the body's interior. If we knew f, then we also could compute the intensity loss of an X-ray sent through the body section along a straight line L: by Beer's law, the ray's

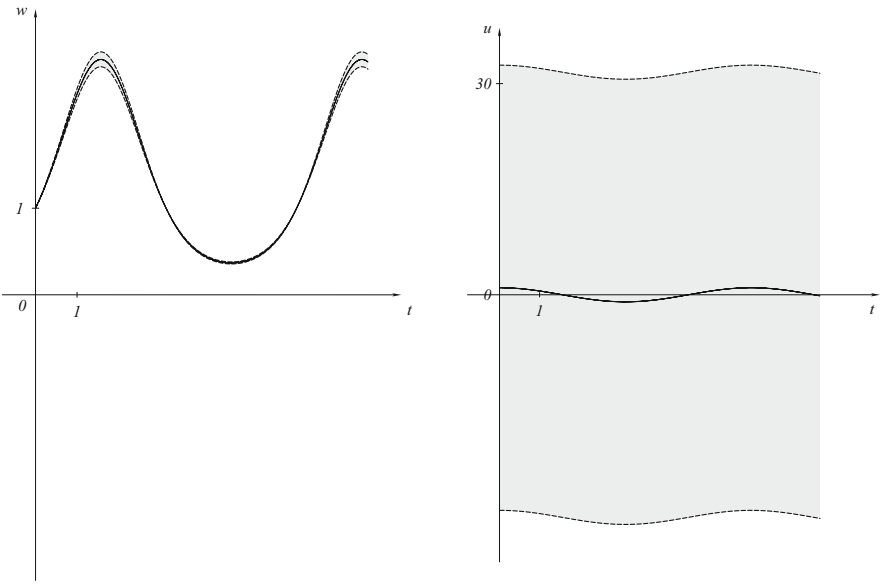

Fig. 1.1 Effects w, w_n and causes u, u_n in Example 1.1, $n = 1000$

intensity loss at position x is proportional to $f(x)$ and so the total loss corresponds to the line integral

$$\int_L f(x)\, ds,$$

see Fig. 1.2, left. According to this figure, f is assumed to be a piecewise constant function, equal function values being indicated by the same shading intensity with darker shades meaning denser tissue. We will rewrite the line integral. For a given angle φ let us define vectors

$$\theta := \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \quad \text{and} \quad \theta^{\perp} := \begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix}.$$

Since every straight line L through D can be parameterized in the form

$$L = \{s\theta + t\theta^{\perp};\ t \in \mathbb{R}\} \quad \text{for} \quad \varphi \in [0, \pi) \quad \text{and} \quad s \in (-1, 1),$$

the line integral becomes

$$Rf(\varphi, s) := R_\varphi f(s) := \int_{-\infty}^{\infty} f(s\theta + t\theta^{\perp})\, dt. \tag{1.5}$$

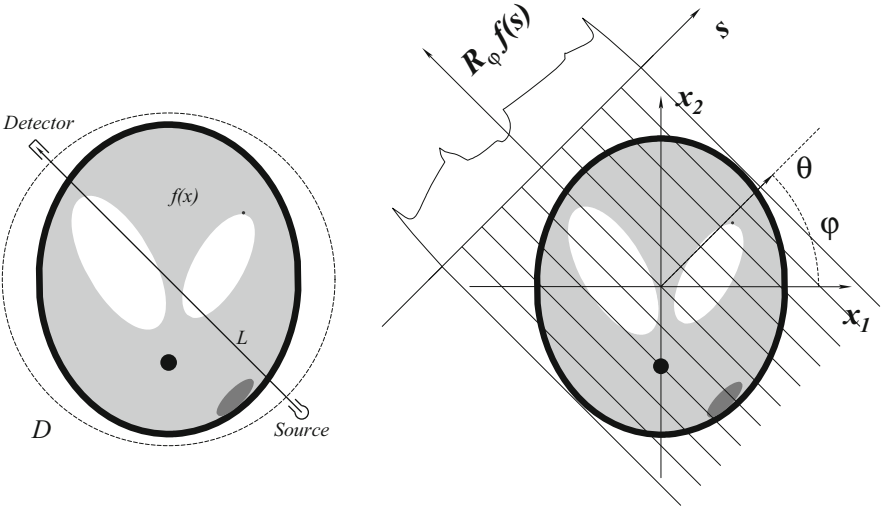

Fig. 1.2 Principle of CT

Since $f(x) = 0$ for $x \notin D$, this is only seemingly an improper integral. Figure 1.2 in its right part illustrates $R_\varphi f$ for a fixed angle φ as a function of s. Note that local maxima appear where the total intensity loss reaches a local maximum (depending on the density of the traversed tissue). The function

$$Rf : [0, \pi) \times (-1, 1) \to \mathbb{R}, \quad (\varphi, s) \mapsto Rf(\varphi, s),$$

is called **Radon transform** of f. The role of the physical law modelled by a mapping T is now played by the transform R, function f is the "cause" and Rf is the "effect". The inverse problem consists in solving the equation $Rf = g$ for an unknown f when g is known. An explicit inversion formula was found in 1917 by the Austrian mathematician Johann Radon:

$$f(x) = \frac{1}{2\pi^2} \int_0^\pi \int_{-1}^1 \frac{\frac{d}{ds} g(\varphi, s)}{x \cdot \theta - s} \, ds d\varphi, \quad x \in D. \tag{1.6}$$

The appearance of derivatives in (1.6) is a hint that f depends sensitively on g, as it was the case in Example 1.1. ◇

The next example has the form of an **integral equation**. Consider a relation $T(u) = w$ between cause and effect $u, w : [a, b] \to \mathbb{R}$, which is given in the form

$$\int_a^b k(s, t, u(t)) \, dt = w(s), \quad a \le s \le b, \tag{1.7}$$

with $k : [a,b]^2 \times \mathbb{R} \to \mathbb{R}$ a known, so-called **kernel function**. Equation (1.7) is called **Fredholm integral equation of the first kind**. A special case is the **linear Fredholm integral equation of the first kind** having the form

$$\int_a^b k(s,t)u(t)\,dt = w(s), \quad a \le s \le b, \tag{1.8}$$

where again $k : [a,b]^2 \to \mathbb{R}$ is given. If this kernel function has the special property

$$k(s,t) = 0 \quad \text{for} \quad t > s,$$

then (1.8) can be written in the form

$$\int_a^s k(s,t)u(t)\,dt = w(s), \quad a \le s \le b, \tag{1.9}$$

and in this case is called **Volterra integral equation of the first kind**. Another special case are kernel functions with the property

$$k(s,t) = k(s-t).$$

In this case, a linear Fredholm integral equation is called **convolutional equation** and can be written as

$$\int_a^b k(s-t)u(t)\,dt = w(s), \quad a \le s \le b. \tag{1.10}$$

One speaks of Fredholm or Volterra integral equations of the *second kind*, if the function u also appears outside the integral. An example is

$$u(s) + \lambda \int_a^s k(s,t)u(t)\,dt = w(s), \quad a \le s \le b, \quad \lambda \in \mathbb{R}.$$

Linear Fredholm equations of the first and second kind do have quite different properties. Technical details are given in [Eng97], Corollary 2.40. We only provide an informal explanation, as follows. If the kernel function k is "smooth" (continuous, for example) then the mapping $u \mapsto w$ defined by integral equations of the first kind also has a smoothing property. Thus, the solution of an integral equation, which means an inversion of integration, necessarily has to roughen the right hand side, thereby also amplifying errors of approximations $\tilde{w}$ to w. In fact the computation of

derivatives, which was found to be problematic in Example 1.1, can be interpreted as solving a Volterra integral equation:

$$u(t) = w'(t),\ w(t_0) = 0 \quad \Longleftrightarrow \quad w(t) = \int_{t_0}^{t} u(s)\, ds\ .$$

In contrast, integral equations of the second kind also contain an unsmoothed copy of u. The solution of such equations thus does not necessarily have to roughen a given right hand side. Fredholm equations are also defined for functions having multiple arguments. In this case, the integration interval $[a, b]$ generalizes to a compact subset of $\mathbb{R}^s$, $s \in \mathbb{N}$, see the following example.

Example 1.3 (Inverse gravimetry) Let $D \subset \mathbb{R}^3$ be a *bounded domain*, i.e. an open, connected set, and let $S := \overline{D}$ be its closure (consisting of all interior and boundary points of D). Let a body B occupy the compact region $S \subset \mathbb{R}^3$ in space and have a mass density given by a function $\rho : S \to \mathbb{R}$, $x \mapsto \rho(x) \geq 0$. The gravitational potential of B at any point $x \in \mathbb{R}^3 \setminus S$ is defined by the volume integral

$$V(x) = -G \int_S \frac{\rho(y)}{\|x - y\|_2}\, dy. \tag{1.11}$$

Here, $\|x - y\|_2$ is the Euclidean distance between x and y, G is the gravitational constant, and S and ρ are assumed to be regular enough to make (1.11) a properly defined (Lebesgue) integral. Equation (1.11) is a convolutional equation for functions of three independent variables. The gravitational force exercised by B on a unit mass located in $x \in \mathbb{R}^3 \setminus S$ is given by the negative gradient of the potential: $F(x) = -\nabla V(x)$.

It is unrealistic to assume that F can be measured everywhere outside B. We consider the following situation. Let $S \subset \Omega \subset \mathbb{R}^3$, where Ω is a *convex domain with sufficiently smooth boundary* (a ball or a half-space would do), let $\Gamma \subset \partial\Omega$ *contain an interior (with respect to* $\partial\Omega$*) point*, and assume that the magnitude of the gravitational force induced by B can be measured on Γ. We thus assume that

$$g : \Gamma \to \mathbb{R}, \quad x \mapsto \|\nabla V(x)\|_2 \tag{1.12}$$

can be observed. Then, (1.11) and (1.12) define a mapping of the form

$$T : \rho \mapsto g \tag{1.13}$$

(a precise definition of this mapping from one function space into another will only be given in Sect. 1.3). The density function ρ is interpreted as being the cause of the observed effect g. The inverse problem of gravimetry consists in finding ρ such that $T(\rho) = g$, where g is given. Figure 1.3 illustrates a two-dimensional plane section of a body B with constant mass density, containing an inclusion with different, but also

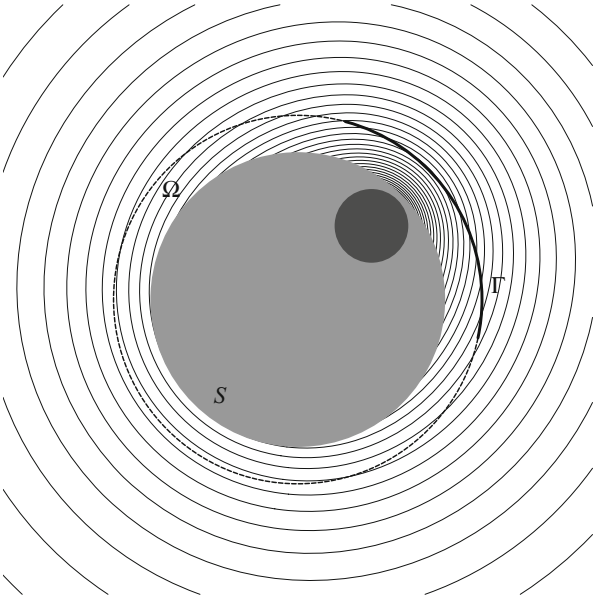

Fig. 1.3 Sections of equipotential surfaces of a body with inhomogenous mass density

constant mass density. The corresponding section of the boundary part Γ, where gravitational forces are measured, is drawn as a fat line. Sections of equipotential surfaces, where the potential V is constant, are also shown. The gradient $\nabla V(x)$ at a point x is orthogonal to the equipotential surface through x.

It can be shown, that under appropriate conditions on S, Ω, Γ, and ρ, two potentials V_1 and V_2 necessarily coincide everywhere on $\mathbb{R}^3 \setminus S$, if $\|\nabla V_1(x)\|_2 = \|\nabla V_2(x)\|_2$ for all $x \in \Gamma$ (see Lemma 2.1.1 in [Isa90]). In this sense it is sufficient to observe $\|\nabla V(x)\|_2$ for $x \in \Gamma$. It is also known, however, that even complete knowledge of V on $\mathbb{R}^3 \setminus S$ is *not enough* to uniquely determine ρ. To achieve uniqueness, restrictions beyond equation (1.11) have to be imposed on ρ. For example, within the set

$$\mathscr{L} := \{\rho : S \to [0, \infty);\ \rho \text{ solves (1.11) and } \int_S \rho(x)^2\, dx < \infty\}$$

there exists a unique mass density distribution ρ_h, such that

$$\int_S \rho_h(x)^2\, dx \le \int_S \rho(x)^2\, dx \quad \text{for all } \rho \in \mathscr{L}.$$

This function ρ_h is harmonic, i.e. the Laplace equation $\Delta \rho_h(x) = 0$ holds for all x in the interior of S, see, e.g., the survey article [MF08]. By the maximum principle

for harmonic functions, this means that ρ_h takes its maximal values at the surface of S, which might be undesirable. For example, the mass density distribution of the earth's interior does not have maximal values at the surface. There may be situations, however, where we are only interested in knowing any ρ satisfying (1.11) – for example if the real goal is to construct $V(x)$ for all $x \notin S$ from observed data (1.12). This could in principle be done by finding ρ_h and then using (1.11) to compute V.

Another situation where uniqueness holds, and which is of evident interest in geological prospecting, is the one depicted in Fig. 1.3, where a body of constant mass density is included in another body, also having constant mass density. One might think of the task to detect a subterranean ore deposit. In certain cases it is possible to determine the exact shape and location of the hidden body. This is the case, if both mass densities are known and if the inclusion is convex "along one dimension". A set $S \subset \mathbb{R}^3$ is x_j-convex for some $j \in \{1, 2, 3\}$ and thus "convex along one dimension", if its intersection with any straight line parallel to the x_j-axis is an interval (convex sets are x_1-convex, x_2-convex, and x_3-convex at the same time). This uniqueness result goes back to the Russian mathematician Novikov and the year 1938. Technical details are described in [Isa90], see for example Theorem 3.1.4. We will consider the problem of finding a hidden body when formulating a model problem for inverse gravimetry in Sect. 1.3. Of course, the assumption of constant mass densities can only be approximately true in reality. ◇

Example 1.4 (Seismic tomography) The goal of seismic tomography is to obtain information about the earth's subsurface material parameters. To this end, seismic waves are generated at or near the surface and the pressure or the velocity field of the reflected waves is measured, usually on a part of the earth's surface or in the ocean. Figure 1.4 schematically illustrates a marine seismic experiment. An acoustic source, towed behind a survey ship, creates waves traveling into the subsurface. In the simplified situation shown in Fig. 1.4, the medium is composed of plane homogenuous layers. Within each layer as well as in sea water, waves propagate at a constant speed. At the interface between two layers, a traveling wave is transmitted in part and reflected in part. Hydrophones record the pressure field of upcoming reflected waves below the sea surface. Note that despite the simplicity of the model, multiple reflections are considered.

Mathematically, the propagation of seismic waves in a domain $\Omega \subset \mathbb{R}^3$ is modelled by the **elastic wave equation**, see, e.g., [Sch07] or [Sym09]. We will exclusively use a simpler mathematical model, the **acoustic wave equation**, referring to [Sym09] for a discussion of the conditions under which this simplification is adequate. Acoustic waves consist of small localized material displacements, which propagate through the medium and lead to deviations from time-independent equilibrium pressure. The displacement at location $x \in \Omega$ and at time $t \geq 0$ is a vector $\mathbf{u}(x, t) \in \mathbb{R}^3$, defining a vector field $\mathbf{u} : \Omega \times [0, \infty) \to \mathbb{R}^3$. The **sound pressure**, which is the deviation from equilibrium pressure, at time t and location $x \in \Omega$ is a scalar $p(x, t) \in \mathbb{R}$, defining a scalar function $p : \Omega \times [0, \infty) \to \mathbb{R}$. Both are related by **Hooke's law**

$$p = -\kappa \nabla \cdot \mathbf{u} + S, \tag{1.14}$$

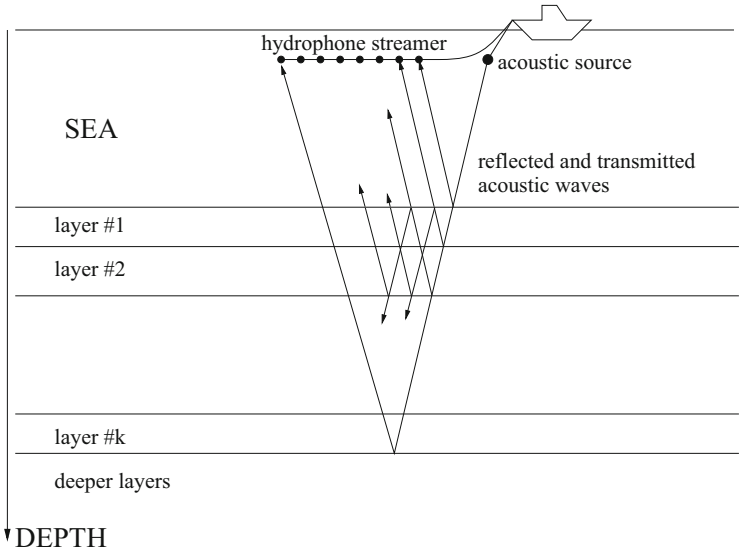

Fig. 1.4 Marine seismic experiment

where $\kappa = \kappa(x)$ is a material parameter, called **elasticity coefficient** or **bulk modulus** and where

$$\nabla \cdot \mathbf{u} := \frac{\partial \mathbf{u}}{\partial x_1} + \frac{\partial \mathbf{u}}{\partial x_2} + \frac{\partial \mathbf{u}}{\partial x_3} = \mathrm{div}(\mathbf{u})$$

is a convenient notation for the divergence of $\mathbf{u}$ *taken only with respect to* x. Further, $S = S(x,t)$ is a time and space dependent "source term". This could be a source injecting material into the medium and thus exciting the wave, like an air gun in the marine seismic experiment. A second relation between sound pressure and displacement is given by **Newton's law**

$$\nabla p = -\rho \frac{\partial^2 \mathbf{u}}{\partial t^2}, \tag{1.15}$$

where the gradient ∇p again is to be taken only with respect to the space variable x and where $\rho = \rho(x)$ is another material parameter, the mass density. In the acoustic model, knowing Ω, ρ and κ means to completely know the medium. From (1.14) and (1.15) one can immediately derive the (scalar) acoustic wave equation for p, reading

$$\frac{1}{\rho c^2}\frac{\partial^2 p}{\partial t^2} - \nabla \cdot \left(\frac{1}{\rho}\nabla p\right) = -\frac{1}{\kappa}\frac{\partial^2 S}{\partial t^2}, \tag{1.16}$$

where we have introduced the **wave velocity**

$$c = c(x) = \sqrt{\frac{\kappa(x)}{\rho(x)}}. \tag{1.17}$$

The value $c(x)$ is the acoustic wave's propagation speed at $x \in \Omega$. The differential equation (1.16) is complemented with initial conditions

$$p(x,0) = 0, \quad \frac{\partial p}{\partial t}(x,0) = 0, \quad x \in \Omega, \tag{1.18}$$

and boundary conditions

$$p(x,t) = 0, \quad x \in \partial\Omega, \quad t > 0. \tag{1.19}$$

Under appropriate assumptions concerning ρ, κ, Ω, and S, one can prove that a unique solution p of (1.16), (1.18), and (1.19) exists. This means that there is a mapping $F : (\rho, \kappa) \mapsto p$. Hydrophones measure the sound pressure $p(x,t)$ on a subset $M \subset \Omega \times [0,\infty)$, so we have another mapping

$$T : (\rho, \kappa) \mapsto p_{|M}$$

with $p_{|M}$ the restriction of p to the subset M. A precise definition of a variant of this mapping will be given in Sect. 1.4. The inverse problem takes the usual form of finding the cause (ρ, κ) responsible for an observed effect $p_{|M}$. This is an example of a parameter identification problem for a partial differential equation.

Similar to the situation in inverse gravimetry, it is known that T in general is not one-to-one, i.e. ρ and κ can not be determined from $p_{|M}$. In Sect. 1.4 we will consider a much simplified situation, where it is at least possible to determine the product $\rho \cdot \kappa$ as a function of the wave's travel time. ◇

1.2 Ill-Posed Problems

We present Hadamard's definition of ill-posedness, formulating it specifically for identification problems, i.e. in the context of equation solving. "Solving an equation" has to be understood in a broad sense, including equations in function spaces, like differential or integral equations. So the sets X and Y used in the subsequent definition may be function spaces and the operator T mapping subsets of X to subsets of Y may be a differential or an integral operator. Some material on function spaces is included in Appendix B. In the following paragraph we very briefly *and informally* introduce this concept by way of an important example.

Function Spaces

The s-dimensional space of real numbers is denoted by $\mathbb{R}^s$, $s \in \mathbb{N}$. A subset of $\mathbb{R}^s$ is called a **domain**, if it is open and connected, and it is called **compact** if it is closed and bounded. A **multi-index** α is a s-dimensional vector having non-negative integers as components, i.e. $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{N}_0^s$. The length of a multi-index α is defined by $|\alpha| = \sum_{i=1}^{s} \alpha_i$. Given a function $v : \mathbb{R}^s \to \mathbb{R}$, its partial derivatives of order $|\alpha|$ at x may be written as

$$D^\alpha v(x) = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_s^{\alpha_s}}(x).$$

For any $M \subset \mathbb{R}^s$ we write

$$C(M) = \{v : M \to \mathbb{R};\ v \text{ continuous}\}.$$

The set $C(M)$ is a **function space**. We write $v \in C(M)$ to express that v is a scalar-valued continuous function defined on M. Such a function v is considered a point in $C(M)$ in the same way as a triple $x = (x_1, x_2, x_3)$ of real numbers is considered a point in the three-dimensional Euclidean space $\mathbb{R}^3$. If v is bounded on M, then one can define its **maximum-norm**

$$\|v\|_{C(M)} := \sup\{|v(x)|;\ x \in M\}.$$

The norm $\|v\|_{C(M)}$ measures the distance of v to the origin (the zero function) in the space $C(M)$ as the (Euclidean) norm $\|x\|_2$ of a point $x \in \mathbb{R}^3$ measures its distance from the origin in $\mathbb{R}^3$. However, if v is not bounded on M, then $\|v\|_{C(M)}$ does not exist. This can not happen if M is compact, since continuous functions do take their maximum value on compact sets, so one can write

$$\|v\|_{C(M)} = \max\{|v(x)|;\ x \in M\}$$

in this case. Now let $D \subseteq \mathbb{R}^s$ be a domain. For any $k \in \mathbb{N}_0$ one defines

$$C^k(D) = \{v : D \to \mathbb{R};\ D^\alpha v \in C(D) \text{ for all multi-indices } \alpha \text{ of length } \leq k\},$$

where $C^0(D) := C(D)$. The set $C^k(D)$ is called the space of k times continuously differentiable functions (defined on D). It is another example of a function space. To define an analog of the maximum-norm, one must ensure the existence of maximal values. To this end, one requires that D is a *bounded* domain, meaning that its closure $\bar{D}$ (i.e. the union of D with its boundary points) is compact. One defines

$$C^k(\bar{D}) = \{v \in C^k(D);\ D^\alpha v \in C(\bar{D}) \text{ for all } |\alpha| \leq k\},$$

which is a proper subset of $C^k(D)$. By the compactness of $\bar{D}$ and the continuity of $D^\alpha v$ on $\bar{D}$,

$$\|v\|_{C^k(\bar{D})} := \sum_{|\alpha| \le k} \|D^\alpha v\|_{C(\bar{D})}$$

exists for all $v \in C^k(\bar{D})$ – it is called the maximum-norm on $C^k(\bar{D})$. $C^k(\bar{D})$ is yet another example of a function space.

The function spaces introduced above are (abstract) **vector spaces**, which have essential points in common with "standard" vector spaces like $\mathbb{R}^3$. Like vectors in $\mathbb{R}^3$, elements in any vector space can be added to each other. For example, with $v, w \in C^k(\bar{D})$, $v + w$ defined by $(v + w)(x) = v(x) + w(x)$ is again a function in $C^k(\bar{D})$. The other operation declared in vector spaces besides vector addition is scalar multiplication. For example, if $\lambda \in \mathbb{R}$ and $v \in C^k(\bar{D})$, then $\lambda \cdot v$ defined by $(\lambda \cdot v)(x) = \lambda \cdot v(x)$ is again a function in $C^k(\bar{D})$. Using addition and scalar multiplication, one can build the difference $v - w$ of two functions $v, w \in C^k(\bar{D})$. This can be used to measure the distance between v and w by taking the norm $\|v - w\|_{C^k(\bar{D})}$. In the same way the distance between two vectors $x, y \in \mathbb{R}^3$ is taken by computing the Euclidean norm $\|x - y\|_2$. The tuple $(C^k(\bar{D}), \| \bullet \|_{C^k(\bar{D})})$ consisting of the space $C^k(\bar{D})$ and the corresponding maximum-norm is called a **normed vector space** or **normed linear space**. See Appendix B for other examples of normed linear spaces $(X, \| \bullet \|_X)$. In the one-dimensional case, bounded domains are open, bounded intervals: $D = (t_0, t_1)$. In this case, we will write $C^k[t_0, t_1]$ instead of $C^k(\bar{D}) = C^k([t_0, t_1])$.

An **operator** is a mapping between vector spaces. For example, $T = D^\alpha$ is an operator, which maps a function $v \in C^k(D)$ to its partial derivative $D^\alpha v \in C(D)$. It thus is a mapping from the space $C^k(D)$ into the space $C(D)$. Another example would be

$$T : C^k(D) \to C(D), \quad u \mapsto \sum_{|\alpha| \le k} c_\alpha D^\alpha u,$$

where $c_\alpha \in C(D)$. T is a linear partial differential operator. Using this operator, a partial differential equation of the form

$$\sum_{|\alpha| \le k} c_\alpha(x) D^\alpha u(x) = w(x), \quad x \in D,$$

can be written concisely as an equation in function space, namely $T(u) = w$.

Definition and Practical Significance of Ill-Posedness

Definition 1.5 (Ill-posed inverse problems) Let $(X, \| \bullet \|_X)$ and $(Y, \| \bullet \|_Y)$ be normed linear spaces and let

$$T : \mathbb{U} \subseteq X \to \mathbb{W} \subseteq Y$$

be a given operator (function). Consider the inverse problem of solving the equation

$$T(u) = w, \quad u \in \mathbb{U}, \quad w \in \mathbb{W}, \tag{1.20}$$

for u, when w is given. This problem is called **well-posed (in the sense of Hadamard)**, or equivalently is called **properly posed**, if

(1) for every $w \in \mathbb{W}$ a solution $u \in \mathbb{U}$ does exist (**condition of existence**),
(2) the solution is unique (**condition of uniqueness**), and
(3) the inverse function $T^{-1} : \mathbb{W} \to \mathbb{U}$ is continuous (**condition of stability**).

Otherwise, problem (1.20) is called **ill-posed**, or equivalently is called **improperly posed**.

Properties (1) and (2) ensure the existence of an inverse function $T^{-1} : \mathbb{W} \to \mathbb{U}$. Continuity of T^{-1} means

$$\lim_{n\to\infty} \|w_n - w\|_Y = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} \|T^{-1}(w_n) - T^{-1}(w)\|_X = 0. \tag{1.21}$$

The significance of this stability condition lies in the following: the solution $u^* = T^{-1}(w)$ of $T(u) = w$ can be approximated arbitrarily well by the solution $\tilde{u} = T^{-1}(\tilde{w})$ of $T(u) = \tilde{w}$, if $\tilde{w} \in \mathbb{W}$ approximates w arbitrarily well. If the stability condition is violated, we can not hope that $\tilde{u}$ tells us anything about u^*, no matter how much effort we undertake to get a good approximation $\tilde{w}$ of w.

Example 1.6 Computing the inverse Fourier transform of a known square integrable function is a well-posed problem. Plancherel's Theorem (C.6) shows that the condition of stability is met. ◇

Example 1.7 (Determination of growth rates, part 2) In Example 1.1 the vector space $X = \mathbb{U} = C[t_0, t_1]$ had been equipped with the norm $\| \bullet \|_X := \| \bullet \|_{C[t_0,t_1]}$ and we had equipped $Y = C^1[t_0, t_1]$ with the same norm $\| \bullet \|_Y = \| \bullet \|_{C[t_0,t_1]}$. We also had used $\mathbb{W} := \{w \in Y;\ w(t) > 0 \text{ for } t_0 \le t \le t_1\}$. Direct and inverse problem are determined by the mapping

$$T : \mathbb{U} \to \mathbb{W}, \quad u \mapsto w, \quad w(t) = w_0 e^{U(t)},\ U(t) = \int_{t_0}^{t} u(s)\, ds.$$

Formula (1.3) shows that a unique solution of the inverse problem exists for every $w \in \mathbb{W}$. But the inverse operator T^{-1} is not continuous, as was already observed in Example 1.1: for the sequence $(w_n)_{n\in\mathbb{N}}$ defined in (1.4) and for $w(t) = \exp(\sin(t))$ we had seen

$$\lim_{n\to\infty} \|w_n - w\|_Y = 0, \quad \text{but} \quad \lim_{n\to\infty} \|T^{-1}(w_n) - T^{-1}(w)\|_X = \infty.$$

Consequently, the determination of growth rates is an ill-posed problem, since the condition of stability is violated. ◇

In practice, already the existence condition can be problematic, because one usually only disposes of an approximation $\tilde{w}$ of the true effect w and there is no guarantee that $\tilde{w} \in T(\mathbb{U})$. In the presence of modelling errors, it would even be possible that $w \notin T(\mathbb{U})$. This will happen if T is a simplified mathematical description of an existing physical law, such that not even the exact effect w can be described in the form $T(u) = w$. However, we will not consider this case. Rather, we will always assume $w \in T(\mathbb{U})$, tacitly adding an eventual modelling error to the data error $\tilde{w} - w$. Since we investigate identification problems, we can not concede the uniqueness condition. If it was violated, then it would not be possible to *identify* a specific cause u responsible for the effect w. In such a case we would either have to observe more or other kinds of effects or to somewhow restrict the set of possible causes in order to restore uniqueness. The latter was done in Example 1.3 for inverse gravimetry and will also be done for seismic tomography. Taking the uniqueness condition for granted, one could approximately solve an inverse problem by first finding a "best" approximation $\hat{w} \in T(\mathbb{U})$ of $\tilde{w}$, as defined, for example, by the projection theorem (see Theorem B.4), if $T(\mathbb{U})$ is closed and convex and if Y is a Hilbert space, and by then finding the unique element $\hat{u} \in \mathbb{U}$ with $T(\hat{u}) = \hat{w}$. Whether this approach produces a meaningful approximation $\hat{u}$ of the sought-after solution depends on the stability condition. Violation of the stability condition is the remaining problem and in fact is a major concern when dealing with identification problems.

Amendments to Definition 1.5

Definition 1.5 is academic and needs to be complemented by precise prescriptions concerning the appropriate choice of $\mathbb{U}$, $\mathbb{W}$, $\|\bullet\|_X$, and $\|\bullet\|_Y$, whenever a specific practical application is considered.

First of all, stability means continuity of the operator T^{-1}. Continuity, as defined in (1.21), depends on the chosen norms $\|\bullet\|_X$ and $\|\bullet\|_Y$. It is well possible that an operator is continuous with respect to one norm, but discontinuous with respect to another, see Example 1.8 below and also Proposition B.10. If $\|\bullet\|_X$ and $|\bullet|_X$ are two norms defined on a linear space X and if there is a constant $C > 0$ with

$$\|x\|_X \le C|x|_X \quad \text{for all } x \in X,$$

then $|\bullet|_X$ is called **stronger** than $\|\bullet\|_X$ and $\|\bullet\|_X$ is called **weaker** than $|\bullet|_X$, since convergence of a sequence $(x_n)_{n\in\mathbb{N}}$ with respect to $|\bullet|_X$ entails convergence with respect to $\|\bullet\|_X$, but not vice versa. Consequently, if one replaces the norm $\|\bullet\|_Y$ on Y by a stronger norm $|\bullet|_Y$, then there are less sequences $(w_n)_{n\in\mathbb{N}}$ converging to w and thus there are less sequences for which the implication in (1.21) must hold. The condition of stability imposed on T^{-1} is thus weakened and a discontinuous operator T^{-1} may artificially be turned into a continuous one. Likewise, if the norm $\|\bullet\|_X$ on X is replaced by a weaker norm, then the implication in (1.21) is easier to fulfill. Again, this weakens the stability condition imposed on T^{-1}.

Example 1.8 (Determination of growth rates, part 3) On the linear space $Y = C^1[t_0, t_1]$, we can replace the norm $\|\bullet\|_{C[t_0,t_1]}$ by the stronger one $\|\bullet\|_{C^1[t_0,t_1]}$. The sequence $(w_n)_{n\in\mathbb{N}}$ from (1.4) is no longer convergent with respect to $\|\bullet\|_{C^1[t_0,t_1]}$ and can no longer furnish an example for discontinuity of T^{-1}. Quite to the contrary, T^{-1} *is continuous* with respect to the norms $\|\bullet\|_{C^1[t_0,t_1]}$ on Y and $\|\bullet\|_{C[t_0,t_1]}$ on X, compare Proposition B.10. We have turned an ill-posed inverse problem into a well-posed one! ◇

Enforcing stability by changing norms is a mathematical trick which is not helpful in practice. In Example 1.1 we have access to $w(t)$, but *not* to $w'(t)$, the determination of which is the essence of the inverse problem. If we could somehow observe w' directly, the inverse problem would be trivial. Its difficulty only comes up since w' can not be observed and thus we can not hope to find some approximation $\tilde{w}$ being close to w with respect to the norm $\|\bullet\|_{C^1[t_0,t_1]}$. Instead, (noisy) observations might get us a function $\tilde{w}$ which approximates well the function values of w, but at the same time $\tilde{w}'$ might approximate badly the values of w'.

A *second* shortcoming of Definition 1.5 is its dependence on the choice of $\mathbb{U}$ and $\mathbb{W}$. These sets must be chosen appropriately for the application under consideration. For example, arguing that one is only interested in solving the problem $T(u) = w_0$ for a single, specific effect w_0, one could be tempted to define $\mathbb{W} = \{w_0\}$, a set containing only this single element. Continuity of T^{-1} then would become trivially true. But such an artificial choice of $\mathbb{W}$ does not reflect the fact that w_0 never is exactly known in practice. $\mathbb{W}$ will have to contain some neighbourhood of w_0, if the mathematical model of the inverse problem is to be meaningful. If one was in fact only interested in the solution for a specific right hand side $w_0 \in \mathbb{W}$, then the definition of well-posedness should be changed. One would still require existence and uniqueness of an element $u_0 \in \mathbb{U}$ such that $T(u_0) = w_0$, but the stability condition would have to be modified. One should demand the existence of $r > 0$ such that for any sequence $(u_n)_{n\in\mathbb{N}} \subseteq \mathbb{U}$ with $\|u_0 - u_n\|_X < r$ for all $n \in \mathbb{N}$, the following implication holds:

$$\|T(u_n) - T(u_0)\|_Y \overset{n\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad \|u_n - u_0\|_X \overset{n\to\infty}{\longrightarrow} 0.$$

In this case, the inverse problem is called **locally well-posed in** w_0.

A *third* objection to Definition 1.5 again concerns the stability condition. Continuity of the inverse T^{-1} just tells us that we can compute u to arbitrary precision, if w is known to arbitrary precision. In practice we only know w to some *finite* precision, i.e. we have at hands some $\tilde{w}$ and hopefully know some finite ε such that $\|\tilde{w} - w\|_Y \leq \varepsilon$. Then we would like to estimate how close $T^{-1}(\tilde{w})$ is to $T^{-1}(w)$, i.e. we would like to bound $\|T^{-1}(\tilde{w}) - T^{-1}(w)\|_X$. Such bounds will be given in Sect. 3.2, where we introduce so-called **condition numbers** as a quantitative measure of ill-posedness for linear least squares problems in finite dimensions. The same could be done for general operators in vector spaces by means of "linearization", but this would require us to introduce operator derivatives.

1.3 Model Problems for Inverse Gravimetry

Starting from Example 1.3, two model problems for inverse gravimetry will be formulated, which were taken from [Sam11]. The situation is illustrated in Fig. 1.5 (note the orientation of the x_3-axis!). Here, for known constants a and $0 < h < H$,

$$S = \{x \in \mathbb{R}^3;\ -a \leq x_1 \leq a,\ -a \leq x_2 \leq a,\ h \leq x_3 \leq H\}$$

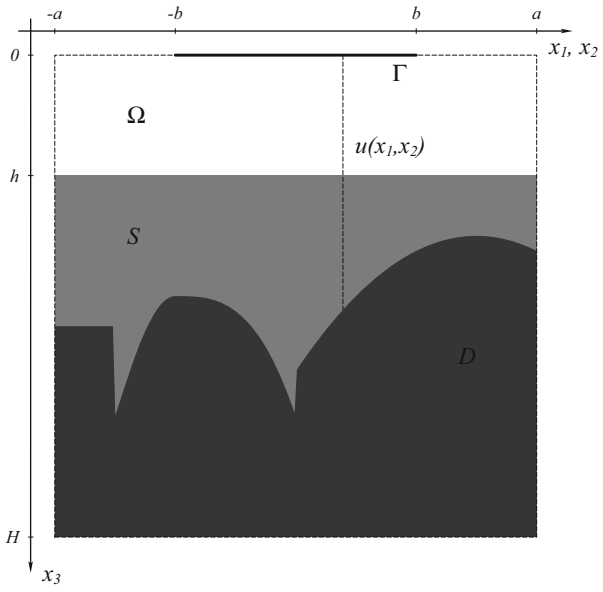

Fig. 1.5 A body S containing an inclusion D convex in x_3

is (the location of) a given body with known constant mass density c_S. S contains an inclusion D with *known*, constant density c_D and of (partially) unknown shape

$$D = \{x \in \mathbb{R}^3;\ -a \le x_1 \le a,\ -a \le x_2 \le a,\ h \le x_3 = u(x_1, x_2) \le H\}.$$

The shape of D is defined by a continuous function

$$u : [-a, a]^2 \to \mathbb{R}, \quad (x_1, x_2) \mapsto u(x_1, x_2),$$

which has to be determined. One might think of a planar approximation of an outer segment of the earth with D being the earth's mantle and $S \setminus D$ being the earth's crust, the task being to determine their interface – the so-called "Mohorovičić discontinuity".

According to Fig. 1.5, we assume that D is "convex in x_3", meaning that the intersection of D and every straight line parallel to the x_3-axis is an interval. In this situation, Theorem 3.1.4 from [Isa90] tells us that it is indeed possible to determine the unknown function u from knowledge of

$$\|\nabla V(x)\|_2, \quad x \in \Gamma = \{x \in \mathbb{R}^3;\ -b \le x_1 \le b,\ -b \le x_2 \le b,\ x_3 = 0\},$$

where $0 < b \le a$ is a known constant and where, omitting constants,

$$V(x) = \int_{-a}^{a}\int_{-a}^{a}\int_{h}^{u(t_1,t_2)} \frac{1}{\sqrt{(x_1-t_1)^2 + (x_2-t_2)^2 + (x_3-t_3)^2}}\, dt_3\, dt_2\, dt_1 \tag{1.22}$$

is the potential of $S \setminus D$. We assume that the first two components of $\nabla V(x)$ are of negligible size, so that $\|\nabla V(x)\|_2$, $x \in \Gamma$, is approximately determined by the vertical component $V_{x_3} = \partial V/\partial x_3$ of the gravitational force, measured at $x = (x_1, x_2, 0) \in \Gamma$. From the fundamental theorem of calculus[2] we get, differentiating under the integral sign:

$$\begin{aligned} V_{x_3}(x_1, x_2, 0) = &- \int_{-a}^{a}\int_{-a}^{a} \frac{1}{\sqrt{(x_1-t_1)^2 + (x_2-t_2)^2 + u(t_1,t_2)^2}}\, dt_2\, dt_1 \\ &+ \underbrace{\int_{-a}^{a}\int_{-a}^{a} \frac{1}{\sqrt{(x_1-t_1)^2 + (x_2-t_2)^2 + h^2}}\, dt_2\, dt_1}_{=:\, V_C(x_1,x_2)} . \end{aligned} \tag{1.23}$$

The values $V_{x_3}(x_1, x_2, 0)$ represent exact satellite measurements of the gravitational force, where the coordinate system is chosen such that the satellite's height is 0.

[2] $f(b) - f(a) = \int_a^b f'(x)\, dx$

Actually, one would observe the vertical component F_{x_3} of the joint gravitational force induced by $S \setminus D$ and D *together*. But since gravitational forces induced by two bodies simply add up, and since the constant mass densities c_D and c_S are known, one can easily find $V_{x_3}(x_1, x_2, 0)$, $(x_1, x_2) \in [-b, b]^2$, from the values $F_{x_3}(x_1, x_2, 0)$. Thus the left hand side of (1.23) can be considered a known function of x_1 and x_2 (although, in reality, we only dispose of a discrete set of noisy measurements). Likewise, the term V_C in (1.23) can be computed analytically and therefore also is a known function. We are led to our first model problem:

Problem 1.9 (Nonlinear inverse gravimetry) Solve the nonlinear Fredholm equation of the first kind

$$w(x_1, x_2) = \int_{-a}^{a} \int_{-a}^{a} k(x_1, x_2, t_1, t_2, u(t_1, t_2))\, dt_2\, dt_1 \tag{1.24}$$

for the unknown, continuous function $u : [-a, a]^2 \to [h, H]$. Here,

$$w(x_1, x_2) = V_C(x_1, x_2) - V_{x_3}(x_1, x_2, 0), \quad (x_1, x_2) \in [-b, b]^2, \tag{1.25}$$

is a known function (since V_{x_3} ("observation") as well as V_C defined in (1.23) are known). Further, the kernel function is defined by

$$k(x_1, x_2, t_1, t_2, u) := \frac{1}{\sqrt{(x_1 - t_1)^2 + (x_2 - t_2)^2 + u^2}} \tag{1.26}$$

for $x_1, x_2, t_1, t_2 \in \mathbb{R}$ and $0 < h \le u \le H$.

Remark Problem 1.9 is an inverse problem, which can formally be described using the notation of Definition 1.5. Namely, set $X = C([-a, a]^2)$ with norm $\| \bullet \|_X = \| \bullet \|_{C([-a,a]^2)}$ and set $Y = C([-b, b]^2)$ with norm $\| \bullet \|_Y = \| \bullet \|_{C([-b,b]^2)}$. Set

$$\mathbb{U} := \{u \in C([-a, a]^2);\ 0 < h \le u(x_1, x_2) \le H \text{ for } (x_1, x_2) \in [-a, a]^2\}, \tag{1.27}$$

set $\mathbb{W} = Y = C([-b, b]^2)$ and define the mapping T by

$$T : \mathbb{U} \to \mathbb{W}, \quad u \mapsto w, \quad w(x_1, x_2) = \int_{-a}^{a} \int_{-a}^{a} k(x_1, x_2, t_1, t_2, u(t_1, t_2))\, dt_2\, dt_1, \tag{1.28}$$

where the kernel function k is given by (1.26).

Next we will consider a special case, assuming the values $u(x_1, x_2)$ do not deviate much from a constant average value $u_0 > h > 0$, i.e

$$u(x_1, x_2) = u_0 + \Delta u(x_1, x_2), \quad (x_1, x_2) \in [-a, a]^2,$$

$\Delta u(x_1, x_2)$ being "small". Evidently, this is *not* true in the situation of Fig. 1.5, but may be a reasonable assumption in case u describes the Mohorovičić discontinuity. Then we might *linearize* (1.24) by replacing the integrand k using the first order approximation:

$$k(x_1, x_2, t_1, t_2, u_0 + \delta u) \doteq k(x_1, x_2, t_1, t_2, u_0) + \frac{\partial k(x_1, x_2, t_1, t_2, u_0)}{\partial u} \cdot \delta u \, .$$

The unknown function Δu will then be computed by solving the equation

$$\begin{aligned} w(x_1, x_2) = &\int_{-a}^{a}\int_{-a}^{a} \frac{1}{\sqrt{(x_1 - t_1)^2 + (x_2 - t_2)^2 + u_0^2}} \, dt_2 \, dt_1 \\ &- \int_{-a}^{a}\int_{-a}^{a} \frac{u_0}{\left((x_1 - t_1)^2 + (x_2 - t_2)^2 + u_0^2\right)^{3/2}} \Delta u(t_1, t_2) \, dt_2 \, dt_1, \end{aligned} \tag{1.29}$$

which *approximates* (1.24). The first integral on the right hand side of (1.29) can be computed analytically. We are thus lead to

Problem 1.10 (Linear inverse gravimetry) Solve the *linear* Fredholm equation of the first kind (the convolution equation)

$$f(x_1, x_2) = u_0 \int_{-a}^{a}\int_{-a}^{a} k(x_1 - t_1, x_2 - t_2) \Delta u(t_1, t_2) \, dt_2 \, dt_1, \tag{1.30}$$

for an unknown, continuous function $\Delta u : [-a, a]^2 \to \mathbb{R}$, where

$$k(r, s) := \left(r^2 + s^2 + u_0^2\right)^{-3/2}, \quad r, s \in \mathbb{R}, \quad u_0 > h > 0, \tag{1.31}$$

and where $f \in C([-b, b]^2)$ is known.

From (1.29) one actually gets

$$f(x_1, x_2) = \int_{-a}^{a} \int_{-a}^{a} \frac{1}{\sqrt{(x_1 - t_1)^2 + (x_2 - t_2)^2 + u_0^2}} \, dt_2 \, dt_1 - w(x_1, x_2) \tag{1.32}$$

for $(x_1, x_2) \in [-b, b]^2$ and for w from (1.24).

1.4 Model Problems for Seismic Tomography

Seismic tomography "in the acoustic approximation" and in the setting of marine experiments was introduced in Example 1.4. The ill-posedness of this problem in general is proven in [KR14]. In the following, we will consider *land* experiments. In this case, geophones are used to record the vertical component of particle velocity on the earth's surface (whereas hydrophones access sound pressure in the ocean). We introduce a drastic simplification by considering only *plane* acoustic waves propagating in a *horizontally layered* medium, which means that we actually only consider one-dimensional wave propagation. The results presented below were obtained by Alain Bamberger, Guy Chavent, and Patrick Lailly, see [BCL77] and [BCL79]. The following mathematical model is used

$$\rho(z) \frac{\partial^2 u(z,t)}{\partial t^2} - \frac{\partial}{\partial z} \left(\kappa(z) \frac{\partial u(z,t)}{\partial z} \right) = 0, \qquad z > 0, \ t > 0, \tag{1.33}$$

$$u(z,t) = 0, \quad \frac{\partial u(z,t)}{\partial t} = 0, \qquad z > 0, \ t = 0, \tag{1.34}$$

$$-\kappa(z) \frac{\partial u(z,t)}{\partial z} = g(t), \qquad z = 0, \ t > 0. \tag{1.35}$$

Here, $t \geq 0$ means time and $z \geq 0$ means the depth of an infinitely deep wave propagation medium. Equation (1.33) can be derived from (1.14) and (1.15) with $S = 0$ (no source in the interior of the propagation medium) and where $u = \mathbf{u}$ now is a scalar displacement field. The source term was renamed and moved into the (Neumann) boundary condition (1.35), which implies that waves are excited at the surface $z = 0$ of the medium only. From (1.14) one can see that $g(t)$ means the pressure applied at time t on the surface. Equation (1.34) defines initial conditions. The mass density ρ and the elasticity coefficient κ were already introduced in Example 1.4. Under appropriate regularity conditions on ρ, κ and g, the system of equations (1.33), (1.34), and (1.35) has a unique solution u, which is differentiable with respect to t. So we may set

$$Y_d(t) := \frac{\partial u}{\partial t}(0, t), \quad t \geq 0, \tag{1.36}$$

which is the velocity of particle displacement at the surface $z = 0$ and thus models the seismogram obtained, i.e. the observed effect of the seismic land experiment. In reality, of course, we will only dispose of a discrete set of noisy observations.

It is known that the mapping $(\rho, \kappa) \mapsto Y_d$ is *not one to one*, meaning that different pairs of parameter functions $(\rho, \kappa) \neq (\tilde{\rho}, \tilde{\kappa})$ can produce exactly the same seismic record Y_d. Therefore, even complete knowledge of Y_d is not sufficient to determine ρ and κ. To guarantee uniqueness, either more data have to be observed or the model has to be simplified. To go the second way, let us now define

$$\psi : [0, \infty) \to [0, \infty), \quad z \mapsto \psi(z) := \int_0^z \frac{ds}{c(s)} =: x, \tag{1.37}$$

with $c = c(z) = \sqrt{\kappa(z)/\rho(z)}$ the wave velocity, as introduced in (1.17). By definition, $\psi(z)$ means the time a wave moving at speed c needs to travel from the surface of the medium down to depth z. Therefore, ψ is called **travel-time transformation**. Since c may be assumed to be positive and continuous almost everywhere, ψ is monotonously increasing, differentiable almost everywhere and has the property $\psi(0) = 0$. It will be used to re-parameterize the solution u of (1.33), (1.34), and (1.35), setting

$$y(x, t) := u(z, t) \quad \text{with} \quad x = \psi(z).$$

One finds that the system (1.33), (1.34), and (1.35) transforms into

$$\sigma(x) \frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma(x) \frac{\partial y(x,t)}{\partial x}\right) = 0, \qquad x > 0,\ t > 0, \tag{1.38}$$

$$y(x, 0) = 0, \quad \frac{\partial y(x,0)}{\partial t} = 0, \qquad x > 0, \tag{1.39}$$

$$-\sigma(0) \frac{\partial y(0,t)}{\partial x} = g(t), \qquad t > 0, \tag{1.40}$$

with only a single parameter function remaining, the so called **acoustical impedance**

$$\sigma(x) = \sqrt{\rho(z)\kappa(z)}, \quad x = \psi(z)\ . \tag{1.41}$$

Because of $\psi(0) = 0$, one has $Y_d(t) = \partial u(0,t)/\partial t = \partial y(0,t)/\partial t$, meaning that the solutions u of (1.33), (1.34), and (1.35) and y of (1.38), (1.39), and (1.40) produce the same seismogram Y_d. Now assume that values $Y_d(t)$ are observed only for $0 < t < T_0$, where T_0 is a fixed point in time. According to (1.38), (1.39), and (1.40), waves are excited exclusively at the medium's surface $x = 0$ and any upcoming wave reaching the surface must be a reflection of a wave originally sent down from the surface. By equation (1.38) all waves travel at a constant speed $\sqrt{\sigma/\sigma} = 1$ and

consequently, no waves reflected from "depths" greater than $x = T_0/2$ can reach the surface and have influence on the seismogram during the time interval $[0, T_0]$. We may as well "cut off the medium" at $x = X_0 := T_0/2$ and switch to the following, third model:

$$\sigma(x)\frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial y(x,t)}{\partial x}\right) = 0, \qquad x \in (0, X_0), t \in (0, T_0), \tag{1.42}$$

$$y(x,0) = 0, \quad \frac{\partial y(x,0)}{\partial t} = 0, \qquad x \in (0, X_0), \tag{1.43}$$

$$-\sigma(0)\frac{\partial y(0,t)}{\partial x} = g(t), \quad y(X_0, t) = 0 \qquad t \in (0, T_0). \tag{1.44}$$

The new boundary condition introduced at $x = X_0$ produces artificial wave reflections from depth X_0, but these reflected waves do not have any influence on the seismogram $Y_d(t)$ for $0 < t < T_0$. The benefit of introducing the new condition is that it bounds the spatial domain. The "method of characteristics" can then be used to solve the wave equation, see Sect. 2.6.

In what follows, we make use of two function spaces other than $C^k(D)$ and $C^k(\bar{D})$ (which were introduced in Sect. 1.2). One is $L_2(a,b)$, $-\infty \le a < b \le \infty$, the space of **square integrable functions**. One may think of its elements as being functions $f : (a,b) \to \mathbb{R}$ which are regular enough such that

$$\|f\|_{L_2(a,b)} := \left(\int_a^b |f(t)|^2\, dt\right)^{1/2} \tag{1.45}$$

exists, although this notion of $L_2(a,b)$ is not entirely correct (compare Appendix B for a formal definition). If $-\infty < a < b < \infty$, then every $f \in C[a,b]$ belongs to $L_2(a,b)$. A function $f : [a,b] \to \mathbb{R}$ also belongs to $L_2(a,b)$, if it is continuous but at exceptional points $t_1, \ldots, t_n$, and if the one-sided limits

$$\lim_{t \downarrow t_i} f(t) = \lim_{t \to t_i +} f(t) \quad \text{and} \quad \lim_{t \uparrow t_i} f(t) = \lim_{t \to t_i -} f(t), \quad i = 1, \ldots, n,$$

exist. Here $t \uparrow t_i$ (or, equivalently, $t \to t_i -$) means $t \to t_i$ *and* $t \le t_i$. Step functions are examples of non-continuous elements of $L_2(a,b)$.

The other space needed below is the **Sobolev space** $H^1(a,b)$. Although it is again not entirely correct, one may think of its elements as being functions $f : (a,b) \to \mathbb{R}$, which are regular enough such that they can be differentiated everywhere but at exceptional points ("almost everywhere") and such that

$$\|f\|_{H^1(a,b)} := \left(\int_a^b |f(t)|^2\, dt + \int_a^b |f'(t)|^2\, dt\right)^{1/2} \tag{1.46}$$

exists (as a finite number). An example would be the function $f : (-1, 1) \to \mathbb{R}$, $t \mapsto |t|$, which can be differentiated everywhere except at $t = 0$. A formal definition of Sobolev spaces is given in Appendix B. If $-\infty < a < b < \infty$, then it can be shown that

$$H^1(a, b) \subset C[a, b]. \tag{1.47}$$

The spaces $L_2(a, b)$ and $H^1(a, b)$ contain many more elements than $C(a, b)$ and $C^1(a, b)$, respectively (which makes it possible to find a solution of a (differential or integral) equation in $H^1(a, b)$, say, if it can not be found in $C^1(a, b)$), and have an important property that $C(a, b)$ and $C^1(a, b)$ fail to have: they are Hilbert spaces – see Appendix B.

Returning to seismic tomography, let us now define the set

$$\mathscr{S} := \{\sigma \in H^1(0, X_0);\ 0 < \sigma_- \le \sigma(x) \le \sigma_+ < \infty\}, \tag{1.48}$$

where

$$X_0 := T_0/2, \tag{1.49}$$

and where σ_- and σ_+ are fixed constants. Note that for $\sigma \in H^1(0, X_0)$, the restriction $\sigma_- \le \sigma(x) \le \sigma_+$ makes sense because of (1.47). On the other hand, $\sigma \in H^1(0, X_0)$ does *not* imply differentiability of σ and therefore (1.42) has to be understood in some "generalized" or "weak" sense. Generalizing differential equations to a weak form is a well established technique, see, e.g., Part II of [Eva98]. We will only provide a few details about it – and only in a specific situation – in Sect. 2.6 (see also Remark 2 after the formulation of Problem 1.11). We state the following results from [BCL77], Théorème 4.2. For $\sigma \in \mathscr{S}$ and $g \in L_2(0, T_0)$:

(1) A unique solution y of (1.42), (1.43), and (1.44) – to be understood in a generalized sense – exists. The solution is differentiable in a weak sense with respect to t at $x = 0$, such that $Y_d \in L_2(0, T_0)$, where $Y_d(t) := \partial y(0, t)/\partial t$.
(2) The mapping

$$T : \mathscr{S} \to L_2(0, T_0), \quad \sigma \mapsto Y_d,$$

is injective, meaning that σ can be uniquely identified from Y_d.

The following model problem is tailored to this situation.

Problem 1.11 (Nonlinear seismic tomography) Let $T_0 > 0$ be a given point in time and let X_0 and $\mathscr{S}$ be defined as in (1.49) and (1.48). Let

(continued)

Problem 1.11 (continued)
$g \in L_2(0, T_0)$ and define the mapping

$$T : \mathscr{S} \to L_2(0, T_0), \quad \sigma \mapsto Y_d, \quad Y_d(t) := \frac{\partial y(0,t)}{\partial t}, \tag{1.50}$$

where y is the weak solution of (1.42), (1.43), and (1.44). Given Y_d, find $\sigma \in \mathscr{S}$ such that $T(\sigma) = Y_d$.

Remark 1 Problem 1.11 is an inverse problem, which can be described using the notation of Definition 1.5. Namely, set $X = H^1(0, X_0)$ with norm $\| \bullet \|_X = \| \bullet \|_{H^1(0,X_0)}$ given by

$$\|f\|^2_{H^1(0,X_0)} = \int_0^{X_0} |f(t)|^2 \, dt + \int_0^{X_0} |f'(t)|^2 \, dt$$

and set $Y = L_2(0, T_0)$ with norm $\| \bullet \|_Y = \| \bullet \|_{L_2(0,T_0)}$ given by

$$\|f\|^2_{L_2(0,T_0)} = \int_0^{T_0} |f(t)|^2 \, dt.$$

Set $\mathbb{U} = \mathscr{S} \subset X$ and $\mathbb{W} = L_2(0, T_0) = Y$ and define $T : \mathbb{U} \to \mathbb{W}$ by (1.50).

Remark 2 The cited results about the mapping T are important, since they show that σ can in fact be identified from the observed seismogram Y_d. It is not very satisfactory, however, to state a result about a "weak solution" without clearly saying what this shall precisely mean. On the other hand, in practice we will only consider discretized impedances σ. We defer it to Sect. 2.6 to give mathematically precise statements about weak solutions at least in this special case.

We finally consider a linearized version of the above model problem provided that, first, $g \in H^1(0, T_0)$ and that, second,

$$\sigma = \sigma_0(1 + f), \quad f \in H^1(0, X_0), \quad f(0) = 0, \qquad |f| \text{ "small"}, \tag{1.51}$$

where σ_0 is a *known and constant* acoustic impedance. Let y_0 be the unique solution of (1.38), (1.39), and (1.40) corresponding to the constant value $\sigma = \sigma_0$ and let y be the (generalized) solution corresponding to $\sigma = \sigma_0(1+f)$. Note that we dropped the artificial boundary condition at $x = X_0$ – it will not be needed to solve the linearized

problem. For $w := y - y_0$, one derives the following partial differential equation

$$\frac{\partial^2 w(x,t)}{\partial t^2} - \frac{\partial^2 w(x,t)}{\partial x^2} = -\left[f(x)\frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(f(x)\frac{\partial y(x,t)}{\partial x}\right)\right] \tag{1.52}$$

The **Born approximation** consists in replacing y on the right hand side of (1.52) by y_0, which is known, since σ_0 is known. Setting

$$h(x,t) := -\left[f(x)\frac{\partial^2 y_0(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(f(x)\frac{\partial y_0(x,t)}{\partial x}\right)\right], \tag{1.53}$$

an approximation v of w is determined as the unique solution of the partial differential equation

$$\frac{\partial^2 v(x,t)}{\partial t^2} - \frac{\partial^2 v(x,t)}{\partial x^2} = h(x,t), \quad x > 0, \quad t > 0, \tag{1.54}$$

with initial values

$$v(x,0) = 0, \quad \frac{\partial v(x,0)}{\partial t} = 0, \quad x > 0, \tag{1.55}$$

and boundary values

$$\frac{\partial v(0,t)}{\partial x} = 0, \quad t > 0. \tag{1.56}$$

The solution v of this system depends *linearly* on f. This means that if v_1 is the solution of (1.54), (1.55), and (1.56) for $f = f_1$ and v_2 is the solution for $f = f_2$, then $\alpha v_1 + \beta v_2$ is the solution for $f = \alpha f_1 + \beta f_2$. The inverse problem consists in determining f from the observation of

$$V_d(t) := \frac{\partial v(0,t)}{\partial t} = \frac{\partial y(0,t)}{\partial t} - \frac{\partial y_0(0,t)}{\partial t}, \quad t > 0. \tag{1.57}$$

We show that this problem takes the form of solving an integral equation. Integral equations, especially the so-called **Lippmann-Schwinger** integral equation, also play a prominent role in inverse scattering theory, see, e.g., [Kir96]. First, the solution y_0 of (1.38), (1.39), and (1.40) for constant $\sigma = \sigma_0$ is explicitly given as

$$y_0(x,t) = \frac{1}{\sigma_0} G(t-x), \quad \text{where} \quad G(z) := \int_0^z g(s)\, ds \tag{1.58}$$

and where g is continued by $g(t) := 0$ for $t \leq 0$ to become a function defined on the real line $\mathbb{R}$. From this, one immediately gets a formula for h in (1.53):

$$h(x,t) = -\frac{1}{\sigma_0} f'(x) g(t-x). \tag{1.59}$$

(More precisely, this equality holds but at the exceptional points, where f is not differentiable. The behaviour of f' at these exceptional points has no influence on the value of the following integrals). Second, the solution of (1.54), (1.55), and (1.56) (inhomogenous equation with initial values and homogenous Neumann boundary values) can be found by a standard procedure called "reflection method" (see [Eva98]). The solution reads

$$v(x,t) = \frac{1}{2} \int_0^t \int_{x-(t-s)}^{x+(t-s)} h(y,s)\, dy\, ds\,, \tag{1.60}$$

if $x > t$. In case $x \leq t$, the solution is given by

$$v(x,t) = \frac{1}{2} \int_0^t \int_0^{x+(t-s)} h(y,s)\, dy\, ds + \frac{1}{2} \int_0^t \int_0^{(t-s)-x} h(y,s)\, dy\, ds\,. \tag{1.61}$$

Third, an explicit formula for $\partial v(0,t)/\partial t$ can be derived from (1.61):

$$\frac{\partial v(0,t)}{\partial t} = \int_0^t h(t-s,s)\, ds \overset{(1.59)}{=} -\frac{1}{\sigma_0} \int_0^t g(s-(t-s)) f'(t-s)\, ds. \tag{1.62}$$

We finally note that f can be determined from knowledge of f', since $f(0) = 0$. The Born approximation thus leads to

Problem 1.12 (Linear seismic tomography) Let $T_0 > 0$, let $X_0 = T_0/2$, and let $0 < \sigma_- < \sigma_0 < \sigma_+$ be given constants. Let $g \in H^1(0, T_0)$ be extended by $g(t) := 0$ for $t \leq 0$. Let $f \in H^1(0, X_0)$ be an unknown function with $f(0) = 0$. Determine f from solving the equation

$$V_d(t) = -\frac{1}{\sigma_0} \int_0^t g(t-2s) f'(s)\, ds \tag{1.63}$$

for f', where $V_d \in H^1(0, T_0)$ is a known function, as defined in (1.57).

Note that $H^1(0, X_0) \subset C[0, X_0]$ (see Appendix B), so that requiring $f(0) = 0$ makes sense. Further, from (1.63) one gets

$$V_d(t) = -\frac{1}{2\sigma_0} \int_0^{2t} g(t-s) f'(s/2)\, ds = -\frac{1}{2\sigma_0} \int_0^{T_0} g(t-s) f'(s/2)\, ds \tag{1.64}$$

where the last equality holds because $g(t) = 0$ for $t \leq 0$. This shows that (1.63) is a linear Fredholm equation of the first kind and even a convolutional equation for the function $s \mapsto f'(s/2)$ (from which f can be derived).

Chapter 2
Discretization of Inverse Problems

Parameter identification problems can be formulated as equations

$$T(u) = w, \quad u \in \mathbb{U} \subseteq X, \quad w \in \mathbb{W} \subseteq Y.$$

In many interesting cases, X and Y are infinite-dimensional spaces of functions – this was so for all model problems presented in Chap. 1. Although we generally suppose that a unique solution $u^* \in \mathbb{U}$ exists, explicit formulae for its computation are only rarely available, so that one has to be satisfied with the construction of an approximate solution by numerical methods. In practice, not even the equation $T(u) = w$ itself is perfectly known, if w is a function. Rather, an approximation of w has to be constructed on the basis of a finite number of (inexact) measurements (observations). All numerical solution methods for inverse problems are based on **discretization**, by which we mean an approximate description and solution of the inverse problem $T(u) = w$ in spaces of *finite dimension*. To achieve this, we choose spaces X_n and Y_m of finite dimension, approximate w by an element $w_m \in Y_m$, approximate T by an operator $T_{n,m} : X_n \to Y_m$ and find $u_n \in X_n$ such that $T_{n,m}(u_n)$ approximates w_m. Then u_n will be considered an approximation of the exact solution u^*.

There are many ways to implement these general ideas and choosing a good discretization is not trivial. It not only decides how well the solution u^* can be approximated, but also shapes the resulting finite dimensional problem and determines, which practical solution methods are applicable and efficient. Necessarily, choosing a good discretization depends on the problem to be solved. We let ourselves be guided by the four model problems posed in Sects. 1.3 and 1.4 and only highlight some aspects of discretization. We will not expose sophisticated approaches like adaptive or multiscale approximation (refer to Chapter 3 from [Cha09]) or specific approximation methods for functions defined on spheres and balls (refer to [FNS10]). In Sect. 2.1 we present spaces of piecewise constant and of piecewise (bi-)linear functions as *candidates* for the choice of X_n and Y_m. These are

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9_2

very easy to handle, but can approximate well a large class of functions. In Sect. 2.2 we discuss the least squares method to find an approximant u_n of u^* in the special case where T is a *linear* mapping and where $\mathbb{U} = X$. Special attention will be payed to an analysis of the error $u^* - u_n$. In Sect. 2.3 the collocation method is presented, which, in the context of Fredholm integral equations, can be interpreted as a special case of the least squares method. Section 2.4 again focusses on linear mappings T implicitly defined by Fredholm integral equations and introduces the method of Backus and Gilbert as an approach to approximately invert such operators. Fourier transform methods are an efficient and therefore attractive solution method in the important case of linear problems defined by convolutional Fredholm equations. These methods are considered and analyzed in Sect. 2.5. Finally, in Sect. 2.6, two specific discretizations for the nonlinear model problems of gravimetry and seismic tomography from Sects. 1.3 and 1.4 are derived. All model problems will be reformulated in a discretized version.

2.1 Approximation of Functions

In this section, the approximation of univariate and bivariate functions (i.e. of functions having one or two arguments) by **spline functions** of low order is discussed. Splines are only one possible choice of candidates for the approximation of functions out of many others possible (including polynomials, Fourier sums, or wavelets). It is a practically very successful choice, since splines are quite easy to handle on a computer and at the same time can approximate well a large class of functions. Splines can be generalized to any space dimension $s \in \mathbb{N}$.

Approximation in One Space Dimension

Definition 2.1 (Univariate splines of orders 1 and 2) Let $a < b$ and let $a = t_1 < t_2 < \ldots < t_m = b$ be a partitioning of the interval $[a, b]$. A function $s : [a, b] \to \mathbb{R}$ having the property

$$s(t) = c_i \in \mathbb{R}, \quad t_i \le t < t_{i+1}, \quad i = 1, \ldots, m-1,$$

(and $s(t_m) = s(t_{m-1})$) is called a **(univariate) spline function of order 1**. If s is continuous and has the property

$$s(t) = a_i t + b_i, \quad a_i, b_i \in \mathbb{R}, \quad t_i \le t \le t_{i+1}, \quad i = 1, \ldots, m-1,$$

then it is called a **(univariate) spline function of order 2**. The numbers t_i are called **knots**. The set of all spline functions of order k defined by the knots $t_1, \ldots, t_m$ is denoted by $\mathcal{S}_k(t_1, \ldots, t_m)$.

Remark Splines of order 1 are step functions, splines of order 2 (also called "linear splines") are polygonal lines. More generally, linear splines could be allowed to have discontinuities at the knots. Also, one could define univariate splines composed of higher order polynomial pieces.

The set $\mathcal{S}_k(t_1, \ldots, t_m)$ is a vector space, since $\alpha_1 s_1 + \alpha_2 s_2 \in \mathcal{S}_k(t_1, \ldots, t_m)$, if $s_1, s_2 \in \mathcal{S}_k(t_1, \ldots, t_m)$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. The dimension of this space is

$$\dim \mathcal{S}_k(t_1, \ldots, t_m) = m + k - 2. \tag{2.1}$$

A basis of $\mathcal{S}_1(t_1, \ldots, t_m)$ is given by the functions

$$N_{j,1}(t) := \begin{cases} 1 \ , \ t_j \le t < t_{j+1} \\ 0 \ , \ \text{else} \end{cases} \quad , \quad j = 1, \ldots, m-1 \tag{2.2}$$

(with additional agreement: $N_{m-1,1}(t_m) := 1$). A basis of $\mathcal{S}_2(t_1, \ldots, t_m)$ is given by the "hat functions"

$$N_{j,2}(t) := \begin{cases} \dfrac{t - t_{j-1}}{t_j - t_{j-1}} \ , \ t \in [t_{j-1}, t_j] & (\text{if } j \ge 2) \\ \dfrac{t_{j+1} - t}{t_{j+1} - t_j} \ , \ t \in [t_j, t_{j+1}] & (\text{if } j \le m-1) \\ 0 \ , \qquad \text{else} \end{cases} \quad , \quad j = 1, \ldots, m, \tag{2.3}$$

having the property $N_{j,2}(t_j) = 1$ for $j = 1, \ldots, m$. These basis functions are called **B-splines** of order 1 or 2, respectively. Any spline function $s \in \mathcal{S}_k(t_1, \ldots, t_m)$ can uniquely be written as a linear combination of B-splines:

$$s(t) = \sum_{j=1}^{m+k-2} \alpha_j N_{j,k}(t), \quad a \le t \le b. \tag{2.4}$$

A convenient way to describe approximation by splines from $\mathcal{S}_k(t_1, \ldots, t_m)$ is the introduction of approximation operators. For example, one may set

$$I_1 : L_2(a, b) \to \mathcal{S}_1(t_1, \ldots, t_m), \quad f \mapsto I_1(f) := \sum_{j=1}^{m-1} \alpha_j N_{j,1}, \tag{2.5}$$

with coefficients α_j defined by

$$\alpha_j := \left(\frac{1}{t_{j+1} - t_j} \int_{t_j}^{t_{j+1}} f(t)\, dt \right), \quad j = 1, \ldots, m-1.$$

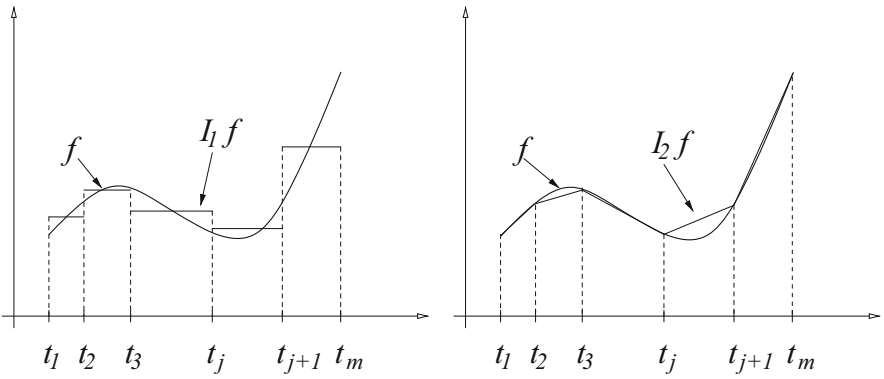

Fig. 2.1 Piecewise constant and linear approximation

I_1 thus is an operator, mapping any square integrable function f to a spline $I_1(f)$ of order 1, which is considered an approximant of f.[1] The approximant $I_1(f)$ has the same local mean values as f. Another example would be the **interpolation operator**

$$I_2 : C[a,b] \to \mathscr{S}_2(t_1,\ldots,t_m), \quad f \mapsto I_2(f) = \sum_{j=1}^{m} f(t_j) N_{j,2}. \tag{2.6}$$

which maps a continuous function f to its linear spline interpolant. The name interpolation operator is chosen because $s = I_2(f)$ evidently has the property

$$s(t_i) = f(t_i), \quad i = 1,\ldots,m. \tag{2.7}$$

Figure 2.1 illustrates the approximation schemes (2.5) and (2.6). The **approximation error** $f - I_k(f)$ in general can become arbitrarily large. Bounds can only be given under additional (smoothness) conditions on f. For example, if $f \in C^2[a,b]$, one can show that

$$\|f - I_2(f)\|_{C[a,b]} \le \frac{1}{8} h^2 \|f''\|_{C[a,b]} \quad \text{with} \quad h := \max_{i=1,\ldots,m-1} \{(t_{j+1} - t_j)\}, \tag{2.8}$$

see [dB90], p. 37. The bound given in (2.8) is no longer useful if f is not a C^2-function. A different bound, given in the following theorem, applies to a larger class of functions f than does (2.8), since $C[a,b] \supset H^1(a,b) \supset C^2[a,b]$. Here, the second inclusion is evident, and the first one – already stated as (1.47) – is a consequence of Sobolev's embedding theorem, see, for example, Theorem A.5 in [LT03].

[1] Of course, this is not the only possibility. Requiring $s(t_j) = f(t_j)$ to hold for $j = 1,\ldots,m-1$, would define a different spline approximant $s \in \mathscr{S}_1(t_1,\ldots,t_m)$ of $f \in C[a,b]$.

Theorem 2.2 (Approximation errors) *Let $k \in \{1, 2\}$, let $a = t_1 < \ldots < t_m = b$ and let $h := \max_{i=1,\ldots,m-1}\{t_{i+1} - t_i\}$. Let $I_k : C[a, b] \to \mathcal{S}_k(t_1, \ldots, t_m)$ be defined by (2.5) for $k = 1$ and by (2.6) for $k = 2$. Then I_k is continuous as a mapping from $(H^1(a, b), \| \bullet \|_{H^1(a,b)})$ to $(L_2(a, b), \| \bullet \|_{L_2(a,b)})$ or to $(H^1(a, b), \| \bullet \|_{H^1(a,b)})$ for $k = 1$ or $k = 2$, respectively, and the bound*

$$\|f - I_k(f)\|_{L_2(a,b)} \leq kh \, \|f\|_{H^1(a,b)} \tag{2.9}$$

on the approximation error holds.

We do not give a proof of this result, which is not trivial, but well known in approximation theory. From Theorem 2.2 it can be seen that $I_k(f) \in S_k(t_1, \ldots, t_m)$ converges to f with respect to the norm $\| \bullet \|_{L_2(a,b)}$, if $h := \max\{t_{i+1} - t_i\}$ tends to 0.

Approximation in Two Space Dimensions

The definition of a two-dimensional analogon of spline functions requires partitionings of two-dimensional sets $D \subset \mathbb{R}^2$. For the sake of simplicity we restrict ourselves to polygonal regions D.

Definition 2.3 (Triangulation, rectangular partitioning) Let $D := \overline{\Omega}$ be the closure of a bounded polygonal domain $\Omega \subset \mathbb{R}^2$. A **triangulation** (**rectangular partitioning**) of D is a set $\mathcal{T} = \{T_1, \ldots, T_m\}$ consisting of closed plane triangles (closed plane rectangles), which meet the following three conditions.

(1) $D = \bigcup_{i=1}^{m} T_i$.
(2) If $T_i \cap T_j$ contains only a single point, then this is a vertex of both, T_i and T_j.
(3) If $T_i \cap T_j$, $i \neq j$, contains more than a single point, then $T_i \cap T_j$ is a common edge of T_i and T_j.

Examples of triangulations and rectangular partitionings are shown in Fig. 2.2. Talking of a partitioning is not fully justified, since the elements T_i can never be mutually disjoint. A triangulation of a polygonal region D does always exist, a rectangular partitioning does exist for rectangles. Triangulations can be generalized to three (and four, ...) space dimensions using simplices. In the following, the word "partitioning" is used to designate either a triangulation or a rectangular partitioning.

We will use the notation $\mathcal{T}_h$ to designate any finite partitioning $\{T_1, \ldots, T_m\}$, when the maximal diameter of $T_i \in \mathcal{T}_h$ is equal to h. A family $\{\mathcal{T}_h\}$ of partitionings (consisting of partitionings for different values h) is called **quasi-uniform**, if there exists a constant $\kappa > 0$ such that every $T \in \mathcal{T}_h$ contains a ball of radius ρ_T with

$$\rho_T \geq h_T/\kappa, \quad h_T := \operatorname{diam}(T), \tag{2.10}$$

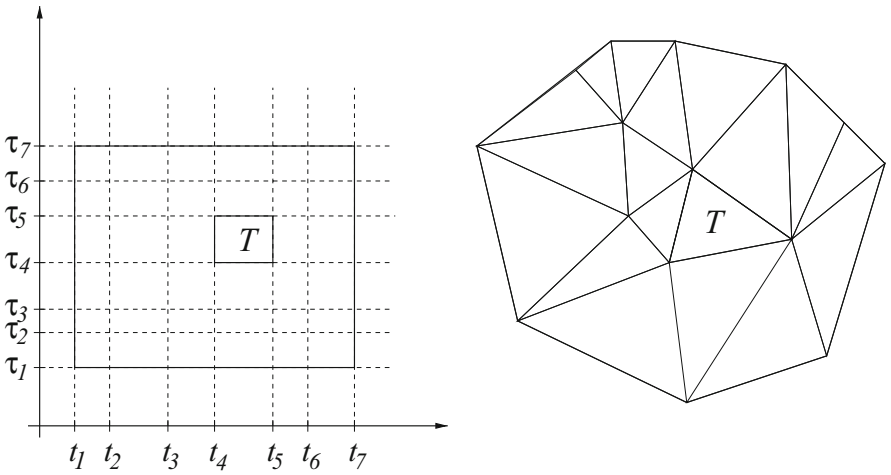

Fig. 2.2 Partitioning of a rectangle into *rectangles* and of a polygon into *triangles*

and it is called **uniform**, if there exists a constant $\kappa > 0$ such that

$$\rho_T \geq h/\kappa, \quad h = \max_{T \in \mathscr{T}_h} \{ \operatorname{diam}(T) \}. \tag{2.11}$$

Quasi-uniform partitionings can not contain arbitrarily narrow elements. In a uniform family of partitionings all elements $T \in \mathscr{T}_h$ shrink at the same rate, if $h \to 0$. Any uniform family is quasi-uniform, but not vice versa.

We will designate by $\mathscr{S}_1(\mathscr{T}_h)$ the set of all piecewise constant functions for a given partitioning $\mathscr{T}_h$. This means that $s \in \mathscr{S}_1(\mathscr{T}_h)$ shall take a constant value c_i in the interior of each $T_i \in \mathscr{T}_h$. We leave it open what values s shall take on edges.

If $\mathscr{T}_h$ is a *rectangular* partitioning of $D = [a, b] \times [c, d]$, then it is defined by $a = t_1 < \ldots < t_{m_1} = b$ and $c = \tau_1 < \ldots < \tau_{m_2} = d$, as shown in Fig. 2.2. We defined linear B-splines $N_{i,2} \in \mathscr{S}_2(t_1, \ldots, t_{m_1})$ and $N_{j,2} \in \mathscr{S}_2(\tau_1, \ldots, \tau_{m_2})$ in the last paragraph. From these one can build **bilinear B-splines**

$$N_{i,j,2} : [a, b] \times [c, d] \to \mathbb{R}, \quad (t, \tau) \mapsto N_{i,j,2}(t, \tau) := N_{i,2}(t) N_{j,2}(\tau).$$

Because of the appearance of terms $t \cdot \tau$, these functions are not piecewise linear. Figure 2.3 shows an example of a bilinear B-spline for the special case where $t_{i+1} - t_i = \tau_{j+1} - \tau_j = h$ for all i and j. The fat line shows the boundary of the support of another one. Let us define the space

$$\mathscr{S}_2^B(\mathscr{T}_h) := \operatorname{span}\{N_{i,j,2};\ i = 1, \ldots, m_1,\ j = 1, \ldots, m_2\}, \tag{2.12}$$

of **bilinear spline functions** with basis $\{N_{i,j,2};\ i = 1, \ldots, m_1,\ j = 1, \ldots, m_2\}$. All elements $s \in \mathscr{S}_2^B(\mathscr{T}_h)$ belong to $C(D)$, i.e. they are continuous functions. An

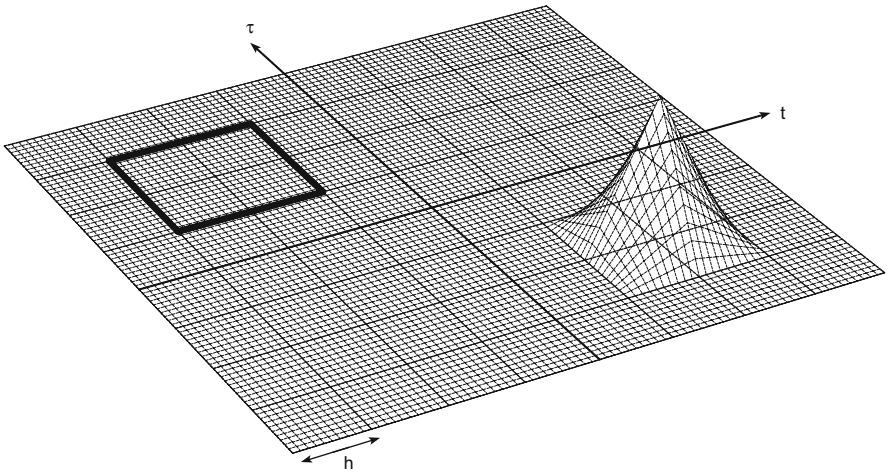

Fig. 2.3 Bilinear B-splines on an equidistant grid

approximation operator can be defined by

$$I_2^B : C(D) \to \mathscr{S}_2^B(\mathscr{T}_h), \quad f \mapsto \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} f(t_i, \tau_j) N_{i,j,2}. \tag{2.13}$$

I_2^B is in fact an interpolation operator, since $I_2^B(f)(t_i, \tau_j) = f(t_i, \tau_j)$ for all i and j.

As in the univariate case, the approximation error $f - I_2^B(f)$ can not be bounded unless some additional regularity of f is assumed, which will be expressed by differentiability conditions. For a domain $\Omega \subset \mathbb{R}^2$ and for $k \in \mathbb{N}_0$, let (informally and not fully correctly) $H^k(\Omega)$ be the Sobolev space of functions $f : \Omega \to \mathbb{R}$, for which

$$\|f\|_{H^k(\Omega)} := \left(\sum_{|\alpha| \le k} \int_\Omega |D^\alpha f(x)|^2 \, dx \right)^{1/2} \tag{2.14}$$

exists. All partial derivatives up to order k have to exist in a sense which gives a meaning to the above integrals and makes them have finite values. As in one dimension, this does not imply a pointwise existence of partial derivatives (as it is required for functions $f \in C^k(\Omega)$). The function

$$f : \Omega \to \mathbb{R} \cup \{\pm\infty\}, \quad (x, y) \mapsto \ln\left(\ln\left(\frac{2}{x^2 + y^2} \right) \right), \tag{2.15}$$

where $\Omega = \{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 < 1\}$, is an example of a $H^1(\Omega)$-function. This function has a pole at $(x, y) = (0, 0)$ and can not with full right be considered as

"defined pointwise". See Appendix B for a formal definition of $H^k(\Omega)$. If Ω is a bounded polygonal domain, then it can be shown that

$$H^2(\Omega) \subset C(D), \quad D := \overline{\Omega}, \tag{2.16}$$

implying that $f \in H^2(\Omega)$ must have point values defined for every $x_0 \in D$. In two space dimensions, this is *not true* for $k = 1$, as shown by (2.15). Note that $H^0(\Omega) = L_2(\Omega)$.

Let us return now to the operator I_2^B and the space $\mathcal{S}_2^B(\mathcal{T}_h)$. The latter is contained in $H^1(\Omega)$, as proved, e.g. in [Bra07], p. 62. The following error estimate is proved in [Bra07], p. 81.

Theorem 2.4 (Approximation error for bilinear spline interpolation) *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded rectangle with closure D. Let $f \in H^2(\Omega)$ and let $\{\mathcal{T}_h\}$ be a family of quasi-uniform rectangular partitionings of D. Then there exists a constant $c = c(\kappa)$ (dependent only on κ from (2.10)) such that*

$$\|f - I_2^B(f)\|_{L_2(\Omega)} \leq c \cdot h^2 \|f\|_{H^2(\Omega)}. \tag{2.17}$$

We finally consider a bounded polygonal domain $\Omega \subset \mathbb{R}^2$ with closure D and a *triangular* partitioning $\mathcal{T}_h$ of D. Let us define

$$\mathcal{S}_2(\mathcal{T}_h) := \{s \in C(D);\ s_{|T} \text{ is linear for all } T \in \mathcal{T}_h\}, \tag{2.18}$$

which is a space of piecewise linear bivariate functions. It can be shown that $\mathcal{S}_2(\mathcal{T}_h) \subset H^1(\Omega)$, see [Bra07], p. 62. Similarly to (2.13), one can define an interpolation operator $I_2 : C(D) \to \mathcal{S}_2(\mathcal{T}_h)$, and an error estimate of the form (2.17) also holds for I_2. But one can not define an interpolation operator on $H^1(\Omega)$, since $H^1(\Omega) \not\subset C(D)$, so that point values of $H^1(\Omega)$-functions are not defined. However, approximation operators $H^1(\Omega) \to \mathcal{S}_2(\mathcal{T}_h)$ *do* exist. The following theorem is proved in [Bra07], p. 83 ff.

Theorem 2.5 (Clément's operator) *Let $\Omega \in \mathbb{R}^2$ be a bounded polygonal domain with closure D. Let $\{\mathcal{T}_h\}$ be a family of quasi-uniform triangulations of D. Then there exists a mapping $I_h : H^1(\Omega) \to \mathcal{S}_2(\mathcal{T}_h)$ and a constant $c = c(\kappa)$ such that*

$$\|u - I_h u\|_{H^m(\Omega)} \leq c h^{1-m} \|u\|_{H^1(\Omega)}, \quad u \in H^1(\Omega), \quad m = 0, 1. \tag{2.19}$$

Theorem 2.5 tells us that any function $u \in H^1(\Omega)$ can be approximated arbitrarily well by a function $s \in \mathcal{S}_2(\mathcal{T}_h)$ with respect to the norm $\| \bullet \|_{L_2(\Omega)}$, if h is chosen small enough.

2.2 Discretization of Linear Problems by Least Squares Methods

In this section – which is based on Chapter 3 of [Kir96] – we present a discretization method which is applicable to a large class of *linear* inverse problems of the form $Tu = w$.[2] We make the following

Assumption 2.6 *Let* $(X, \|\bullet\|_X)$ *and* $(Y, \langle\bullet|\bullet\rangle_Y)$ *be real Hilbert spaces. Let* $T : X \to Y$ *be linear, continuous, and injective.*

Refer to Appendix B for the definitions of Hilbert spaces and of linear and continuous operators. The most important restriction here is the required linearity of the operator T. Linearity implies that T is defined on a linear space, which we have directly taken as X, not as a subspace. Considering subsets $\mathbb{U} \subset X$ as a constraint for a solution u^* of the inverse problem will lead to nonlinearity and is only considered in the context of nonlinear problems (see Sect. 2.6 and Chap. 4). It was already argued in Sect. 1.2, that injectivity of the map T is the one quality of well posed problems that can not be conceded when the goal is to identify parameters. Bijectivity of T is *not* assumed, however, nor do we assume continuity of the inverse of $T : X \to T(X)$. Requiring that X and Y are real spaces is not essential; the following could be generalized to complex vector spaces.

Description of the Method

Let $w \in T(X)$ and let u^* be the unique solution of $Tu = w$. Choose some d_n-dimensional subspace $X_n \subset X$:

$$X_n = \operatorname{span}\{\varphi_1, \ldots, \varphi_{d_n}\} \subset X, \quad \varphi_1, \ldots, \varphi_{d_n} \in X, \text{ linearly independent.} \tag{2.20}$$

The **least squares method** determines an approximant $u_n \in X_n$ of u^* by requiring that

$$\|Tu_n - w\|_Y \leq \|Tv - w\|_Y \quad \text{for all} \quad v \in X_n, \tag{2.21}$$

where $\|\bullet\|_Y$ is the norm induced by $\langle\bullet|\bullet\rangle_Y$. Solving this problem conceptually can be split into two steps.

- First, find the best approximation w_n of w in the d_n-dimensional subspace $Y_n := T(X_n)$ of Y with respect to the norm $\|\bullet\|_Y$. According to Theorem B.5, w_n is

[2] For *linear* operators, it is common usage to write Tu instead of $T(u)$.

determined by the system of n equations

$$\langle w_n | T\varphi_i \rangle_Y = \langle w | T\varphi_i \rangle_Y \quad \text{for} \quad i = 1, \ldots, d_n. \tag{2.22}$$

- Second, find the unique $u_n \in X_n$ with $Tu_n = w_n$. This works, since by assumption 2.6 the map $T : X_n \to Y_n$ is bijective.

Both steps can be assembled into a single one by making the ansatz

$$u_n = \sum_{j=1}^{d_n} x_j \varphi_j, \quad x_j \in \mathbb{R}. \tag{2.23}$$

The parameters x_j need to be determined. Inserting into (2.22) leads to

$$\sum_{j=1}^{d_n} x_j \langle T\varphi_j | T\varphi_i \rangle_Y = \langle w | T\varphi_i \rangle_Y, \quad i = 1, \ldots, d_n. \tag{2.24}$$

This system can be compactly written in matrix form:

$$Ax = b, \quad A_{i,j} = \langle T\varphi_j | T\varphi_i \rangle_Y, \quad b_i = \langle w | T\varphi_i \rangle_Y, \tag{2.25}$$

where $A \in \mathbb{R}^{d_n, d_n}$ is symmetric and positive definite and where $x, b \in \mathbb{R}^{d_n}$. By the properties of A, a unique solution of this system exists for any right hand side. It defines a unique $u_n \in X_n$, the **least squares solution** of $T(u) = w$.

The same idea works if one only knows an approximation $\tilde{w} \approx w$. This approximation can be any vector $\tilde{w} \in Y$, even a non-attainable one (which means that $\tilde{w} \notin T(X)$ is permissible), because any $\tilde{w} \in Y$ can be projected on $T(X_n)$. The only thing to change is to replace w in (2.24) by $\tilde{w}$. This leads to a system $Ax = \tilde{b}$ with inaccuracies shifted from $\tilde{w}$ to $\tilde{b}$. Another possible source of inaccuracy is the numerical evaluation of the scalar products $\langle w | T\varphi_i \rangle_Y$ (or rather $\langle \tilde{w} | T\varphi_i \rangle_Y$), for example if scalar products are defined by integrals, as it is the case for Sobolev spaces. Not distinguishing between measurement errors, approximation errors, and evaluation errors for scalar products, we will simply assume that we are given approximate discrete values

$$b_i^\delta \approx b_i = \langle w | T\varphi_i \rangle_Y, \quad i = 1, \ldots, d_n,$$

with the total data error bounded by

$$\| b^\delta - b \|_2 \leq \delta, \tag{2.26}$$

whence the notation b^δ. The number $\delta \geq 0$ is a numerical bound for the overall inaccuracy in the discrete data. A more sophisticated, but also much more

complicated approach would be to use stochastic error models, but we will not go into this. Let us define

$$x^\delta := A^{-1}b^\delta, \quad u_n^\delta := \sum_{j=1}^{d_n} x_j^\delta \varphi_j \tag{2.27}$$

for the solution of the least squares problem computed for inaccurate data. In practice, x^δ will not be computed by inverting A, but by solving the linear system of equations $Ax = b^\delta$. To summarize:

Least squares method for linear problem $Tu = w$.

- Choose an d_n-dimensional subspace $X_n \subset X$ with basis $\{\varphi_1, \ldots, \varphi_{d_n}\}$.
- Set up the matrix A with components $A_{i,j} = \langle T\varphi_j | T\varphi_i \rangle_Y$ as in (2.24). Set up the vector b^δ with components $b_i^\delta \approx \langle w | T\varphi_i \rangle_Y$ from available measurement values of w.
- Compute the solution x^δ of $Ax = b^\delta$ and get an approximant

$$u_n^\delta = \sum_{j=1}^{d_n} x_j^\delta \varphi_j$$

of u^* as in (2.27).

Application to Model Problem 1.12: Linear Seismic Tomography

To recall, for $T_0 > 0$ and $X_0 = T_0/2$, the task is to find $f \in H^1(0, X_0)$ from observing

$$V_d(t) = -\frac{1}{\sigma_0} \int_0^t g(t-2s) f'(s)\, ds = -\frac{1}{2\sigma_0} \int_0^{T_0} g(t-s) f'(s/2)\, ds \tag{2.28}$$

for $0 \le t \le T_0$, where σ_0 is a constant value, $g \in H^1(\mathbb{R})$ is a known function with $g(t) = 0$ for $t \le 0$ and where $f(0) = 0$. Since the determination of f from f' is straightforward, we simplify our task and only ask for f'. This means we seek the solution u^* of $Tu = w$, where

$$T : X \to Y, \quad u \mapsto w, \quad w(t) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t-2s) u(s)\, ds \tag{2.29}$$

with $X = L_2(0, X_0)$ and $Y = L_2(0, T_0)$, which are both Hilbert spaces, when equipped with the scalar products $\langle \bullet | \bullet \rangle_X = \langle \bullet | \bullet \rangle_{L_2(0,X_0)}$ and $\langle \bullet | \bullet \rangle_Y = \langle \bullet | \bullet \rangle_{L_2(0,T_0)}$, respectively. Since $g \in H^1(\mathbb{R})$ is continuous, so is $w = Tu$ for any $u \in X$. For a discretization we choose a parameter $n \in \mathbb{N}$ and define

$$h := X_0/n \quad \text{and} \quad \tau_j := jh,\ j = 0, \ldots, n.$$

As an $(n+1)$-dimensional subspace of X we use the spline space

$$X_{n+1} := \mathscr{S}_2(\tau_0, \ldots, \tau_n)$$

from Definition 2.1. A basis of X_{n+1} is given by the linear B-splines $\varphi_j = N_{j,2}$, $j = 0, \ldots, n$, see (2.3). $T\varphi_j$ is defined by the integral values

$$T\varphi_j(t) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t - 2s) N_{j,2}(s)\, ds$$

for $0 \le t \le T_0$, and evaluation of the scalar products

$$\langle T\varphi_i | T\varphi_j \rangle_Y = \int_0^{T_0} T\varphi_i(t) \cdot T\varphi_j(t)\, dt, \quad b_i = \langle w | T\varphi_i \rangle_Y = \int_0^{T_0} w(t) T\varphi_i(t)\, dt$$

means integration again. In the following example, values of all $T\varphi_j$ are computed (approximately) on a fine grid using numerical integration (by the trapezoidal rule). Based on these values, the scalar products $\langle T\varphi_i | T\varphi_j \rangle_Y$ are computed (approximately) using the trapezoidal rule again.

Example 2.7 As a kernel function we use the **Ricker pulse**

$$g(t) = aG(f_0 t - 1) \tag{2.30}$$

where a is the amplitude and f_0 is the "center frequency" of the pulse. The function G is proportional to the second derivative of the Gaussian function, namely:

$$G(\theta) = (1 - 2\pi^2\theta^2)\mathrm{e}^{-\pi^2\theta^2}, \quad \theta \in \mathbb{R}.$$

Note that $g(t) = 0$ is only approximately true for $t \le 0$. We set $T_0 = 1$, $a = 1$, $f_0 = 5$, and take w such that

$$u^* : \left[0, \frac{1}{2}\right] \to \mathbb{R}, \quad t \mapsto 2t(1 - 2t) \tag{2.31}$$

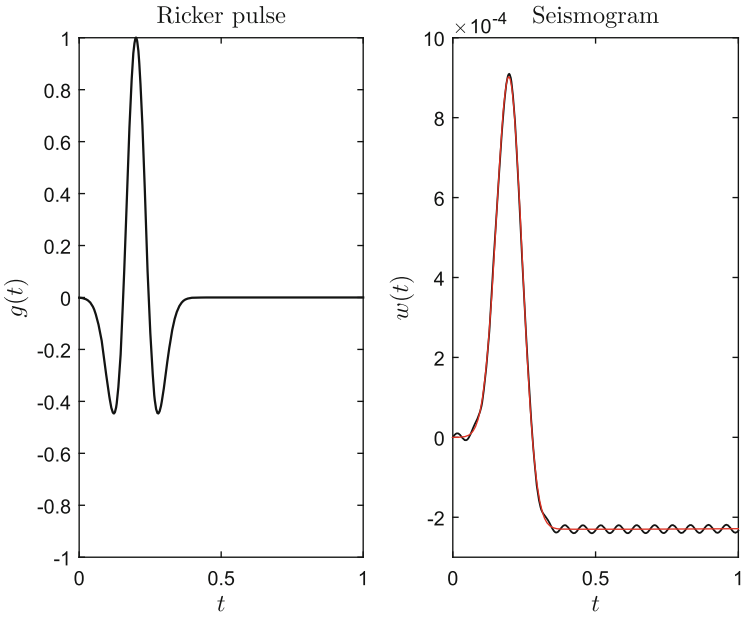

Fig. 2.4 Ricker pulse and exact and noisy seismogram

is the exact solution of $Tu = w$. Figure 2.4. shows the Ricker pulse function (to the left) and a noisy function w^δ (to the right, in black), which was constructed from the exact effect w (to the right, in red) by adding the function $10^{-5}\sin(100t)$. From w^δ values $b_i^\delta = \langle w^\delta | T\varphi_i \rangle_Y$ were computed and an approximation u_n^δ of u^* was constructed as in (2.27). Figure 2.5 shows reconstructions u_n^δ (in black) we got for various values of n, as compared to the exact solution u^* (in red). Figure 2.6 shows one more reconstruction (left) and the L_2-error $\|u^* - u_n^\delta\|_X$ for various values of n (right). Visibly, the reconstruction at first gets better with n growing and then gets worse, the result for $n = 25$ already being totally useless. This behaviour of discretized solutions for inverse problems is typical and will be explained below. Unluckily, we have no practical way to determine the optimal value of n (without knowing the solution of the inverse problem). ◇

Analysis of the Method

The following theorem gives an error estimate for least squares reconstructions.

Theorem 2.8 (Error estimates for least squares method) *Let Assumption 2.6 hold. Let X_n be a d_n-dimensional subspace of X with basis $\{\varphi_1, \ldots, \varphi_{d_n}\}$ and let R_n*

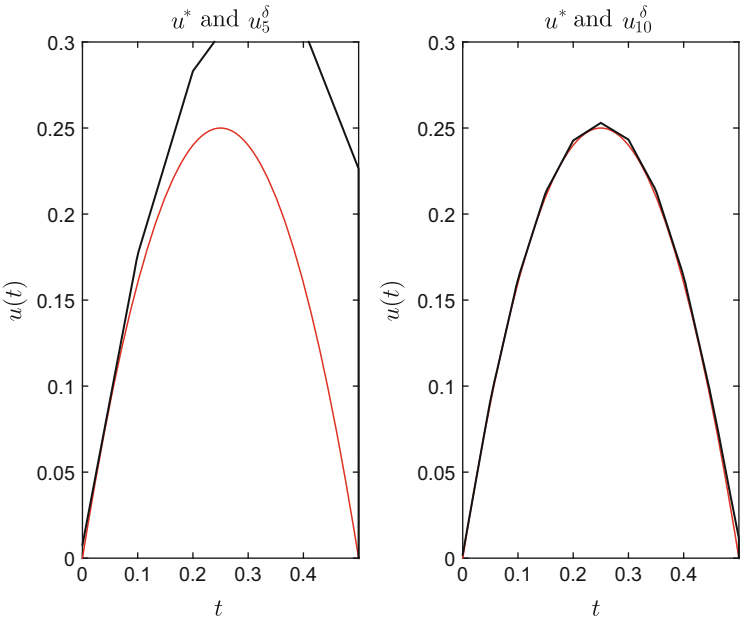

Fig. 2.5 Two reconstructions at different discretization levels

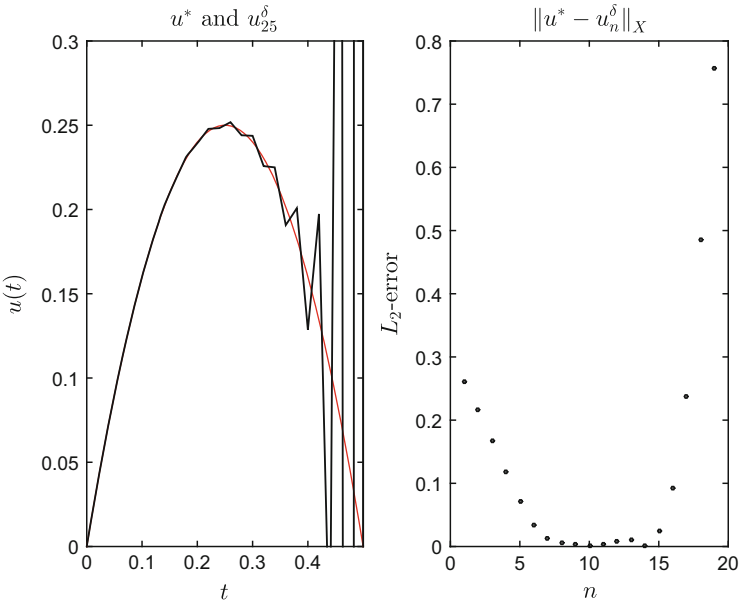

Fig. 2.6 Reconstruction errors as a function of n

be the linear "reconstruction operator"

$$R_n : Y \to X_n, \quad y \mapsto u_n = \sum_{j=1}^{d_n} x_j \varphi_j, \tag{2.32}$$

where $x = (x_1, \ldots, x_{d_n})^T$ *is the solution of the linear system*

$$Ax = b, \quad A_{i,j} = \langle T\varphi_j | T\varphi_i \rangle_Y, \quad b_i = \langle y | T\varphi_i \rangle_Y, \tag{2.33}$$

as in (2.25). Let $w \in T(X)$ *and let* $u^* \in X$ *be the corresponding unique solution of* $Tu = w$. *Let* $w^\delta \in Y$ *with* $\|w - w^\delta\|_Y \le \delta$ *for some* $\delta > 0$, *then*

$$\|u^* - R_n w^\delta\|_X \le \|R_n\|\delta + \|R_n T u^* - u^*\|_X. \tag{2.34}$$

Now let b^δ *be a perturbation of b, defined by* $b_i = \langle w | T\varphi_i \rangle_Y$, *such that* $\|b^\delta - b\|_2 \le \delta$. *Let* u_n^δ *be defined by* b^δ *according to (2.27). Then the following error estimates hold*

$$\|u^* - u_n^\delta\|_X \le \frac{a_n}{\sigma_n}\delta + \|R_n T u^* - u^*\|_X \tag{2.35}$$

$$\|u^* - u_n^\delta\|_X \le b_n \|R_n\|\delta + \|R_n T u^* - u^*\|_X \tag{2.36}$$

where

$$a_n := \max\left\{ \left\| \sum_{j=1}^{d_n} \rho_j \varphi_j \right\|_X ; \quad \sum_{j=1}^{d_n} |\rho_j|^2 = 1 \right\} \tag{2.37}$$

$$b_n := \max\left\{ \sqrt{\sum_{j=1}^{d_n} |\rho_j|^2}; \quad \left\| \sum_{j=1}^{d_n} \rho_j \cdot T(\varphi_j) \right\|_Y = 1 \right\} \tag{2.38}$$

and where σ_n *is the smallest singular value of the matrix A defined in (2.25).*

Remark $R_n w^\delta$ is computed in the same way as $R_n w$, but with $y := w^\delta$ instead of $y := w$ in (2.33). The first error estimate (2.34) concerns a situation where the true effect w is pertubed to become w^δ. For the second error estimates (2.35) and (2.36), nothing is assumed about w^δ, errors are directly attributed to the discrete data. This is the reason why an additional factor b_n appears in (2.36).

Proof The estimate (2.34) easily follows from the triangle inequality, using the identity $u^* - R_n w^\delta = u^* - R_n T u^* + R_n w - R_n w^\delta$. From the triangle inequality one also gets $\|u^* - u_n^\delta\|_X \le \|u_n^\delta - R_n w\|_X + \|R_n w - u^*\|_X$. Since $Tu^* = w$, this means that only $\|u_n^\delta - R_n w\|_X$ has to be estimated in (2.35) and (2.36). Using $R_n w = \sum_{j=1}^{d_n} x_j \varphi_j$,

one can write $u_n^\delta - R_n w = \sum_{j=1}^{d_n} (x_j^\delta - x_j)\varphi_j$ and estimate

$$\begin{aligned}\|u_n^\delta - R_n w\|_X &\le a_n \|x^\delta - x\|_2 = a_n \|A^{-1}(b^\delta - b)\|_2 \\ &\le a_n \|A^{-1}\|_2 \|b^\delta - b\|_2 \le \frac{a_n}{\sigma_n}\delta,\end{aligned}$$

with $\|A^{-1}\|_2 \le 1/\sigma_n$ by Theorem A.2. This shows (2.35). For the proof of (2.36), refer to [Kir96], p. 75. □

Note that R_n is defined on Y, not only on $T(X)$, so we can reconstruct approximative causes for noisy effects $\tilde{w} \notin T(X)$. The estimates of Theorem 2.8 show that the total reconstruction error can be bounded by the sum of two terms. The first term on the right hand sides of (2.35) and (2.36) tells us how much the discrete data error $b^\delta - b$ is amplified by the reconstruction and thus gives a measure for the robustness of the reconstruction. The factors a_n and b_n depend on the basis of X_n and can be made equal to 1 if $\{\varphi_1, \ldots, \varphi_{d_n}\}$ or $\{T\varphi_1, \ldots, T\varphi_{n_d}\}$ are chosen to be an orthonormal basis of X_n or $T(X_n)$, respectively. The second term $\|R_n T u^* - u^*\|_X$ is an estimate of the discretization error alone, disregarding data errors. It tells us how well R_n approximates the inverse of $T : X \to T(X)$. Natterer has investigated under which conditions $\|R_n T u^* - u^*\|_X$ can be made arbitrarily small, see [Nat77]. We cite one of his results from [Kir96], Theorem 3.10, omitting the proof.

Theorem 2.9 (Convergence of least squares method) *Let Assumption 2.6 hold. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of d_n-dimensional subspaces of X and let $(R_n)_{n\in\mathbb{N}}$ be a corresponding sequence of reconstruction operators as in Theorem 2.8. Define*

$$\gamma_n := max\{\|z_n\|_X;\ z_n \in X_n,\ \|Tz_n\|_Y = 1\}. \tag{2.39}$$

Beyond Assumption 2.6, assume that for every $x \in X$ and for every $n \in \mathbb{N}$ there exists an element $\tilde{x}_n \in X_n$ with

$$\|x - \tilde{x}_n\|_X \to 0 \quad for \quad n \to \infty \tag{2.40}$$

and also assume that there is a constant $c > 0$ such that

$$\min_{z_n \in X_n} \{\|x - z_n\|_X + \gamma_n \|T(x - z_n)\|_Y\} \le c\|x\|_X \quad for\ all\ x \in X\ . \tag{2.41}$$

Under these two conditions, the reconstruction is convergent, i.e.

$$\|R_n T u - u\|_X \to 0 \quad for\ all\ u \in X\ and\ for\ n \to \infty \tag{2.42}$$

and the operators R_n are bounded: $\|R_n\| \le \gamma_n$.

Some comments are in order.

- Equation (2.42) is called "pointwise convergence" of R_n to $T^{-1} : T(X) \to X$. As a prerequisite for this convergence to hold, we have, naturally, condition (2.40), *but we also need* the technical condition (2.41). There are in fact examples showing that (2.40) alone *is not sufficient for convergence*.
- Since every $z_n \in X_n$ can be written as $z_n = T^{-1}(y_n)$ for some $y_n \in T(X_n)$, we have

$$\gamma_n = \max\{\|T^{-1}(y_n)\|_X;\ y_n \in T(X_n),\ \|y_n\|_Y = 1\}. \tag{2.43}$$

  Therefore, for $y_n, z_n \in T(X_n)$, we can estimate

$$\|T^{-1}(y_n) - T^{-1}(z_n)\|_X \le \gamma_n \|y_n - z_n\|_Y$$

  and interpret γ_n as a "measure of stability" for the inverse of $T_{|X_n}$.
- The operators R_n evidently have the "projection property" $R_n T \tilde{x}_n = \tilde{x}_n$ for all $\tilde{x}_n \in X_n$. For any linear mapping $R_n : Y \to X_n$ with this quality, we have, by definition of the operator norm:

$$\begin{aligned}\|R_n\| &= \sup\left\{\frac{\|R_n y\|_X}{\|y\|_Y};\ y \in Y,\ y \neq 0\right\} \ge \sup\left\{\frac{\|R_n T x\|_X}{\|Tx\|_Y};\ x \in X,\ x \neq 0\right\} \\ &\ge \sup\left\{\frac{\|R_n T v\|_X}{\|Tv\|_Y};\ v \in X_n,\ v \neq 0\right\} = \sup\left\{\frac{\|v\|_X}{\|Tv\|_Y};\ v \in X_n,\ v \neq 0\right\} \\ &= \gamma_n. \end{aligned} \tag{2.44}$$

  This means that among all convergent reconstruction processes having the (desirable) projection property, the least squares method is the most robust one, i.e. the one which leads to the least amplification of data errors.
- If $T^{-1} : T(X) \to X$ is not continuous (the usual case for inverse problems) then

$$\gamma_n \longrightarrow \infty \quad \text{for} \quad n \to \infty. \tag{2.45}$$

  This can be seen as follows. If we had $\gamma_n \le C$ for all $n \in \mathbb{N}$, then $\|v\|_X \le C\|Tv\|_Y$ for all $v \in X_n$ and all n. Because of (2.40) we then would also have $\|x\|_X \le C\|Tx\|_Y$ for all $x \in X$ and therefore $\|T^{-1}(y)\|_X \le C\|y\|_Y$ for all $y \in T(X)$ in contradiction to the unboundedness of $T^{-1} : T(X) \to X$.

In (the usual) case $\gamma_n \to \infty$ for $n \to \infty$, (2.44) and (2.36) tell us that the total error $\|u^* - u_n^\delta\|_X$ in general can not become arbitrarily small for a finite value $\delta > 0$ and rather may blow up if n is chosen too big. This is illustrated in Fig. 2.7 and explains the error behaviour observed in Example 2.7.

It is not always easy to apply Theorem 2.9 in practical situations. Condition (2.40) evidently is fulfilled if $X = H^1(\Omega)$ for some bounded (polygonal) domain $\Omega \subset \mathbb{R}$ or $\Omega \subset \mathbb{R}^2$, if $n \sim \frac{1}{h}$, and if X_n is chosen as a space of spline functions $\mathcal{S}_2(\mathcal{T}_h)$ as in Definition 2.1 for the one-dimensional case or in (2.18)

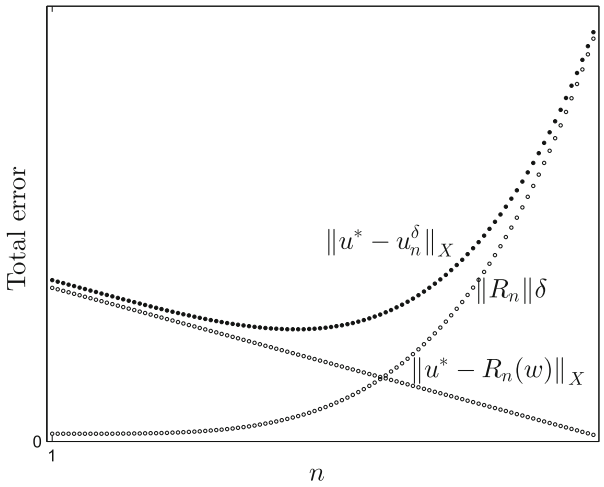

Fig. 2.7 Total reconstruction error; $T(u^*) = w$; δ, R_n and u_n^δ as in Theorem 2.8

for the two-dimensional case, compare Theorems 2.2 and 2.5. The difficulty comes from condition (2.41), which not only depends on X_n, but also on T itself. Natterer investigates the condition (2.41) in [Nat77], but his results can not be directly applied to the operator T from Example 2.7.

2.3 Discretization of Fredholm Equations by Collocation Methods

The least squares method of Sect. 2.2 does not yet tell us how to fully discretize a linear inverse problem, since it does not specify how to compute the scalar products $\langle w|T\varphi_i\rangle_Y$ required to set up the matrix equation $Ax = b$ in (2.25). In the present section – based again on Chapter 3 of [Kir96] – we will present a fully discrete method, assuming that function samples of w are available. Although collocation methods exist for other inverse problems too, we will only discuss them in the context of linear Fredholm integral equations of the first kind. Precisely, we assume the following.

Assumption 2.10 *Let $\mathbb{U} = X \subseteq L_2(a,b)$ be a real Hilbert space and let $k \in C([a,b]^2)$ be such that*

$$T : X \to C[a,b], \quad u \mapsto w,\ w(t) = Tu(t) = \int_a^b k(t,s)u(s)\,ds \tag{2.46}$$

is linear, continuous, and injective. Let $a \le t_1 < t_2 < \cdots < t_m \le b$ and assume that samples $w(t_i)$, $i = 1, \ldots, m$ are given.

Under Assumption 2.10, $Y := T(X)$ is a Hilbert space, when equipped with the scalar product

$$\langle y|z\rangle_Y := \langle T^{-1}(y)|T^{-1}(z)\rangle_{L_2(a,b)}. \tag{2.47}$$

Assumption 2.10 and the collocation method could be generalized to the case of multi-dimensional Fredholm equations.

Description of the Method

The numbers t_i are called **collocation points**. We set up the **collocation equations**

$$Tu(t_i) = w(t_i), \quad i = 1, \ldots, m, \tag{2.48}$$

as a substitute for equation $Tu = w$, which we actually want to solve. Choosing an n-dimensional subspace

$$X_n := \langle \varphi_1, \ldots, \varphi_n \rangle \subset X, \quad \varphi_1, \ldots, \varphi_n \text{ linearly independent},$$

as in Sect. 2.2, one can rate an approximant of u^* (the solution of $Tu = w$) as

$$u_n = \sum_{j=1}^{n} x_j \varphi_j \in X_n, \quad x_j \in \mathbb{R}. \tag{2.49}$$

Requiring (2.48) to hold, one obtains the linear system of equations

$$Tu_n(t_i) = \sum_{j=1}^{n} x_j \cdot T(\varphi_j)(t_i) = w(t_i), \quad i = 1 \ldots, m, \tag{2.50}$$

which can be written in matrix form with $A \in \mathbb{R}^{m,n}$ and $b \in \mathbb{R}^m$:

$$Ax = b, \quad A_{ij} = T\varphi_j(t_i), \quad b_i = w(t_i). \tag{2.51}$$

There is no guarantee that a unique solution of (2.51) exists. In case $n > m$, (2.51) most likely is underdetermined and will have an infinite number of solutions (the data set $\{w(t_1), \ldots, w(t_m)\}$ is not sufficient to determine a unique u_n). In case $n < m$, (2.51) most likely is overdetermined and has no solution at all. We therefore will replace (2.51) by the minimization problem

$$\text{minimize } \|b - Ax\|_2, \quad x \in \mathbb{R}^n, \tag{2.52}$$

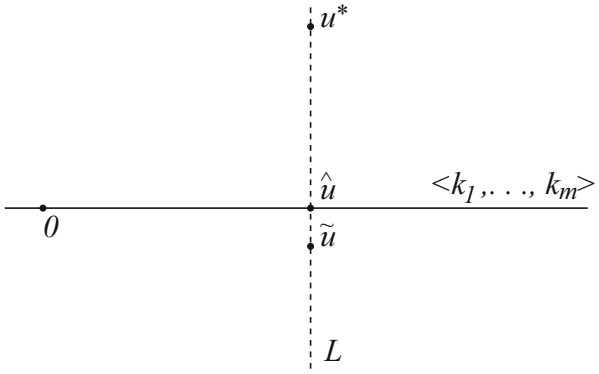

Fig. 2.8 Minimum norm solution of collocation equations

which always has a solution. If (2.51) does have a unique solution, then (2.52) will have the same unique solution. Minimization problems of the form (2.52) will be further discussed in the following chapter.

No use was made so far of the special form (2.46) of T. Note that for any function $u \in L_2(a,b)$

$$Tu(t_i) = \int_a^b k(t_i,s)u(s)\,ds = \langle k_i|u\rangle_{L_2(a,b)}, \quad k_i(s) := k(t_i,s),\ i = 1,\ldots,m. \tag{2.53}$$

Since $w = Tu^*$ is the exact solution of the inverse problem, the collocation equations (2.48) may equivalently be written in the form

$$\langle k_i|u^*\rangle_{L_2(a,b)} = \langle k_i|u\rangle_{L_2(a,b)} \quad \Longleftrightarrow \quad \langle k_i|u^* - u\rangle_{L_2(a,b)} = 0. \tag{2.54}$$

This situation is depicted in Fig. 2.8: The solid line symbolizes the linear subspace $\langle k_1,\ldots,k_m\rangle$ of $L_2(a,b)$, which is spanned by $k_1,\ldots,k_m$ and which, like every linear subspace, contains the zero element of $L_2(a,b)$. The dashed line symbolizes the affine subspace L of $L_2(a,b)$, which is the solution set of the collocation equations (2.48) and which, by virtue of (2.54), is orthogonal to $\langle k_1,\ldots,k_m\rangle$. As stated by Theorem B.5, there exists a unique **minimum norm solution** $\hat{u}$ of the collocation equations, i.e. there exists a unique square integrable function $\hat{u} \in L_2(a,b)$ characterized by

$$\|\hat{u}\|_{L_2(a,b)} = \min\{\|z\|_{L_2(a,b)};\ z \in L_2(a,b) \text{ satisfies (2.48)}\}. \tag{2.55}$$

This function $\hat{u}$ can be interpreted as the projection of u^* into the linear space spanned by $k_1,\ldots,k_m$. Any other function $\tilde{u} \in L$ (i.e. any other L_2-solution of the collocation equations) has larger norm (i.e. greater distance from 0). If $k_1,\ldots,k_m$

are linearly independent, we can find $\hat{u}$ by making the ansatz

$$\hat{u} = \sum_{j=1}^{m} x_j k_j \tag{2.56}$$

and solving the linear system (2.51) with $n = m$ and $\varphi_j = k_j, j = 1, \ldots, m$. Finally, using (2.47) we can write the collocation equations in the form

$$\langle Tk_i | Tu \rangle_Y = \langle Tk_i | w \rangle_Y, \quad i = 1, \ldots, m. \tag{2.57}$$

This directly compares to (2.22) if $n = m$, $X_n = \langle k_1, \ldots, k_m \rangle$, and $k_1, \ldots, k_m$ are linearly independent. In this special case, the collocation method applied to linear Fredholm integral equations coincides with the least squares method.

Application to Model Problem 1.12: Linear Seismic Tomography

We take up again the problem of linear seismic tomography in the simplified form presented in Sect. 2.2, where we had

$$T : X \to Y, \quad u \mapsto w, \quad w(t) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t - 2s) u(s)\, ds \tag{2.29}$$

with $X = L_2(0, X_0)$ and $Y = L_2(0, T_0)$, and where $T_0 > 0$ and $X_0 = T_0/2$ are fixed parameters. Since $g \in H^1(\mathbb{R})$ is continuous, an effect w always is a continuous function, which can be sampled. Let us therefore assume that $m \in \mathbb{N}$ samples

$$w(t_i) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t_i - 2s) u(s)\, ds, \quad 0 \le t_1 < \ldots < t_m \le T_0,$$

are given. As observed above, we could choose the functions defined by $k_i(s) = g(t_i - 2s)$ as a basis to construct an approximate solution of $T(u) = w$ and would end up with a least squares solution. But we can also choose a parameter $n \in \mathbb{N}$, define $h = X_0/n$ and $\tau_j := jh$, $j = 1, \ldots, n$, and work with the n-dimensional subspace

$$X_n := \mathscr{S}_2(\tau_1, \ldots, \tau_n) = \langle N_{1,2}, \ldots, N_{n,2} \rangle$$

of X, which is spanned by the linear B-splines $N_{j,2}$, $j = 1, \ldots, n$. Any spline $u_n \in X_n$ has the property $u_n(0) = 0$. The collocation equations take the form of system (2.51)

with matrix $A \in \mathbb{R}^{m,n}$ having components

$$A_{i,j} = -\frac{1}{\sigma_0} \int_0^{X_0} g(t_i - 2s) N_{j,2}(s)\, ds, \quad i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{2.58}$$

The system $Ax = b$ corresponding to (2.51) and (2.58) is much simpler to set up than the corresponding system for the least squares solution, since only a single integration has to be carried out per matrix component. But we can no longer guarantee that a solution of the system exists.

Example 2.11 Like in Example 2.7, we use the Ricker pulse defined in (2.30) as kernel function. Our choice of parameters is the same as in Example 2.7: $T_0 = 1$, $a = 1$, and $f_0 = 5$. Also, we choose w such that

$$u^* : \left[0, \frac{1}{2}\right] \to \mathbb{R}, \quad t \mapsto 2t(1 - 2t) \tag{2.31}$$

is the exact solution of $T(u) = w$. We fix the value $m = 20$ and use

$$t_i := i\frac{T_0}{m}, \quad i = 1, \ldots, m$$

as collocation points. Further, we let $n \in \mathbb{N}$ and define

$$h := \frac{1}{2n}, \quad \tau_i := hi, \quad i = 1, \ldots, n,$$

to define the spline space $X_n := \mathscr{S}_2(\tau_1, \ldots, \tau_n)$. With these values, the components of $A \in \mathbb{R}^{m,n}$ in (2.58) were computed by exact (and not by numerical) integration. A perturbed version w^δ of w was constructed just as in Example 2.7. It defines samples $b_i^\delta = w^\delta(t_i)$, $i = 1 \ldots, m$. A solution x^δ of (2.52) with b replaced by b^δ was determined, which defines the spline function $u_n^\delta(t) = \sum_{j=1}^n x_j^\delta N_{j,2}(t)$ approximating u^*. In Fig. 2.9 the best reconstruction result is shown (in black, as compared to the exact solution, shown in red), which was obtained for $n = 9$ and which is of comparable quality as the best reconstruction achieved by the least squares method. We also plot the reconstruction error versus the discretization level parameterized by n. The behaviour is quite similar to the one observed for the least squares method. Again, we have no practical way to determine the optimal parameter n. But choosing n to small or too large will produce an unacceptable result. ◇

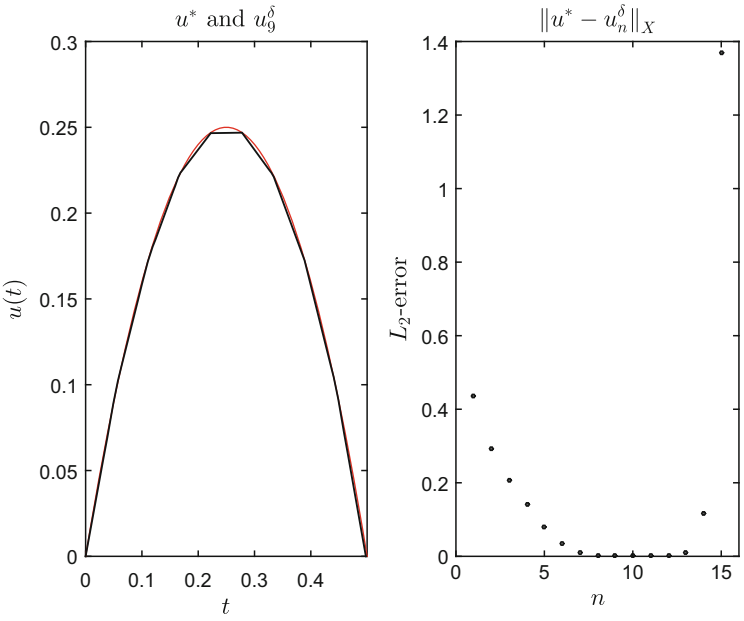

Fig. 2.9 Best collocation reconstruction and reconstruction error as a function of n

Analysis of the Method

For an analysis, especially concerning error estimates analoguous to Theorem 2.8, see Theorems 3.21 and 3.24 in [Kir96]. The analysis confirms a similar qualitative behaviour of the error as for the least squares method. This was observed in Example 2.11 and is shown in the error plot in Fig. 2.9.

2.4 The Backus-Gilbert Method and the Approximative Inverse

This method will be presented in the context of Fredholm integral equations of the first kind, following Section 3.6 of [Kir96]. Assumption 2.10 is made again. Using the functions $k_i \in C[a,b]$ introduced in (2.53), let us define the linear operator

$$K : L_2(a,b) \to \mathbb{R}^m, \quad u \mapsto (y_1, \ldots, y_m), \quad y_i := \int_a^b k_i(s)u(s)\,ds, \ i = 1, \ldots, m, \tag{2.59}$$

mapping a cause u to a finite number of samples of its effect w. The collocation equations (2.48) can be written compactly in form $Ku = y$, where $y = (w(t_1), \ldots, w(t_m))$. If all L_2-functions are admitted as candidates for a solution, an additional requirement is needed to make a solution unique. This could be a minimum norm requirement, as in (2.55). A different idea is to look for a linear operator of the form

$$S : \mathbb{R}^m \to L_2(a,b), \quad y \mapsto Sy, \quad Sy(t) = \sum_{j=1}^{m} y_j \psi_j(t), \tag{2.60}$$

which "approximately inverts" K such that – in a sense to be made precise

$$SKx \approx x \quad \textit{for all } x \in L_2(a,b). \tag{2.61}$$

Here, $\psi_1, \ldots, \psi_m$ are fixed functions to be defined pointwise (and not as functions in $L_2(a,b)$) such that (2.60) is meaningful. The (yet imprecise) "inversion requirement" (2.61) replaces the collocation equations. From the definitions of S and K we get

$$x_m(t) := SKx(t) = \sum_{j=1}^{m} \psi_j(t) \int_a^b k_j(s)x(s)\,ds = \int_a^b \underbrace{\left(\sum_{j=1}^{m} k_j(s)\psi_j(t) \right)}_{=:\,\varphi(s,t)} x(s)\,ds. \tag{2.62}$$

Depending on how the magnitude of the difference $SKx - x$ is measured, different answers will be given as for how to choose well the functions ψ_j. One idea is to require a pointwise approximation of x by SKx, i.e. to ask for

$$SKx(t) = x_m(t) \approx x(t). \tag{2.63}$$

Ideally, this should hold for all $t \in [a,b]$ and "for all" x. However, it certainly makes no sense to require (2.63) for all $x \in L_2(a,b)$, since point values are not defined for L_2-functions. A precise formulation of what (2.63) actually shall mean will only be given below. For the moment one might think of (2.63) as demanded for continuous functions x only. If (2.63) was to hold, then according to (2.62) a point value $x(t)$ should be produced from averaging the values $x(s)$ for $a \le s \le b$ – which is why φ is called an **averaging kernel**. It is clear that $SKx(t) \approx x(t)$ can be expected to hold in good approximation at some point t, if the averaging kernel is normalized such that $\int_a^b \varphi(s,t)\,ds = 1$, if it is sharply peaked at $s = t$ and if it is close to zero for $s \neq t$. The less $\varphi(\cdot,t)$ corresponds to such a function, the less $\int_a^b \varphi(s,t)x(s)\,ds$ can be expected to produce a good approximation $x(t)$ for all admissible functions x at the same time. Of course, since ψ_j, $j = 1, \ldots, m$, are functions of t, φ can not be independent of t: the averaging kernel changes with t.

Description of the Backus-Gilbert Method

The Backus-Gilbert Method does not determine *functions* ψ_j, but rather *values* $\psi_j(t)$, $j = 1, \ldots, m$, for some *fixed* parameter $t \in [a, b]$, such that $SKx(t) \approx x(t)$ *at this point* t. The process can then be repeated any number of times to determine values $\psi_j(\bar{t})$ for other parameters $\bar{t} \in [a, b]$ in order to make $SKx(\bar{t}) \approx x(\bar{t})$ hold. So let us now keep $t \in [a, b]$ fixed. To find $v_j := \psi_j(t), j = 1, \ldots, m$ (for fixed t), one solves the minimization problem

$$\text{minimize} \quad \int_a^b (s-t)^2 \varphi(s,t)^2 \, ds, \quad \text{where} \quad \varphi(s,t) = \sum_{j=1}^m v_j k_j(s), \tag{2.64}$$

subject to the linear constraint

$$\int_a^b \varphi(s,t) \, ds \;=\; \int_a^b \sum_{j=1}^m v_j k_j(s) \, ds \;=\; 1. \tag{2.65}$$

The normalization condition (2.65) excludes the trivial solution $v = 0$ of (2.64). Together, (2.64) and (2.65) are a mathematical formulation of the qualities desired for φ, as discussed after (2.62). This minimization problem replaces the vague approximate identity (2.63) and can be written in compact form as

$$\text{minimize} \quad v^T Q(t) v \quad \text{subject to} \quad v^T c = 1, \tag{2.66}$$

where $Q(t) \in \mathbb{R}^{m,m}$ and $c \in \mathbb{R}^m$ have components

$$\begin{aligned} Q_{ij}(t) &= \int_a^b (s-t)^2 k_i(s) k_j(s) \, ds, \quad i,j = 1, \ldots, m, \\ c_i &= \int_a^b k_i(s) \, ds, \quad i = 1, \ldots, m. \end{aligned} \tag{2.67}$$

If $c \neq 0$ and if the functions k_i are linearly independent, then the matrix $Q(t)$ is positive definite for every $t \in [a, b]$ and problem (2.66) does have a unique solution v, see Theorem 3.31 in [Kir96]. The solution can be computed analytically using the method of Lagrange multipliers. Since minimization of $v^T Q(t) v$ and $v^T Q(t) v/2$ is equivalent, $L(v, \mu) = \frac{1}{2} v^T Q(t) v + \mu v^T c$ can be used as a Lagrange function. Its gradient with respect to v is $Q(t)v + \mu c$. Setting the gradient to zero gives $v = -\mu Q(t)^{-1} c$. Multiplying this term with c^T from the left and making use of $c^T v = 1$ gives $\mu = -1/(c^T Q(t)^{-1} c)$. The solution then reads

$$v = \frac{Q(t)^{-1} c}{c^T Q(t)^{-1} c}, \tag{2.68}$$

but is *not* to be computed by inverting $Q(t)$, but by solving $Q(t)z = c$ for z. The above derivation also shows that v is determined by the linear system

$$\begin{pmatrix} Q(t) & c \\ c^T & 0 \end{pmatrix} \begin{pmatrix} v \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{m+1}. \tag{2.69}$$

Having determined v and knowing $y := (w(t_1), \ldots, w(t_m))$, one computes

$$u_m(t) := \sum_{j=1}^{m} y_j v_j = \int_a^b \varphi(s,t) u(s)\, ds \tag{2.70}$$

As already mentioned, the whole process can be repeated for any other $t \in [a,b]$. Every time, a new matrix $Q(t)$ has to be compiled and a new linear system of equations has to be solved. This means a high computational effort, but can theoretically produce an approximation u_m of u which is defined on the whole interval $[a,b]$. In [Kir96], Lemma 3.3, it is shown that the function u_m is analytic and therefore infinitely often differentiable.

Application to Model Problem 1.12: Linear Seismic Tomography

Let us consider again the problem of linear seismic tomography in the simplified form presented in Sect. 2.2, where we had

$$T : X \to Y, \quad u \mapsto w, \quad w(t) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t - 2s) u(s)\, ds \tag{2.29}$$

with $X = L_2(0, X_0)$ and $Y = L_2(0, T_0)$, and where $T_0 > 0$ and $X_0 = T_0/2$ are fixed parameters. As noted before, $g \in H^1(\mathbb{R})$ implies $T(X) \subset C[a,b]$. Assume that $m \in \mathbb{N}$ samples

$$y_i := w(t_i) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t_i - 2s) u(s)\, ds, \quad 0 < t_1 < \ldots < t_m < T_0$$

are given. According to (2.29) $k_i(s) = -g(t_i - 2s)/\sigma_0$, $i = 1, \ldots, m$. For fixed $t \in [a,b]$ one can compute the point value $u_m(t) = \sum_{j=1}^{m} v_j y_j$ of an approximation u_m of u^*, where $v \in \mathbb{R}^m$ is determined by (2.68).

Example 2.12 We take up Example 2.7, using the same parameters T_0, a, f_0 to define the Ricker pulse g and also using the same functions u^* and w. Let us define $y := (w(t_1), \ldots, w(t_m))$ (this computation was done by exact integration),

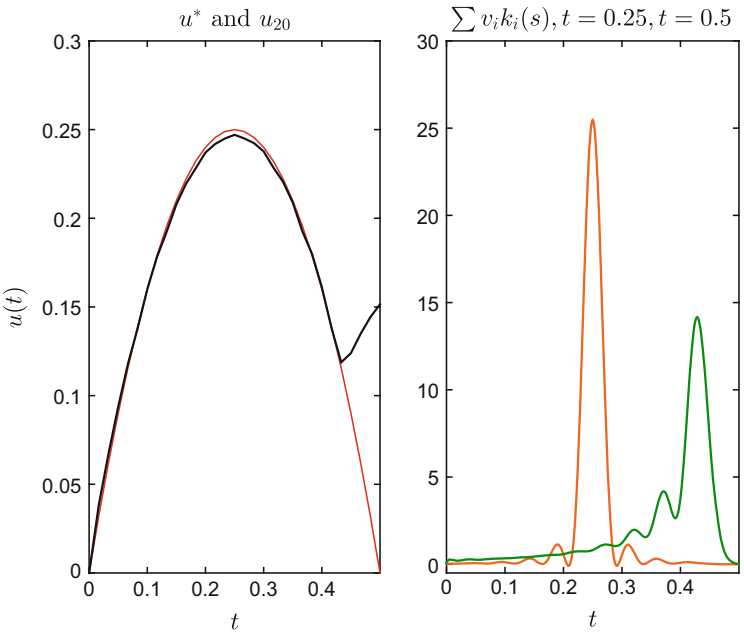

Fig. 2.10 Backus-Gilbert reconstruction and two averaging kernels

fix $m = 20$, and let $a = 0$, $b = X_0 = T_0/2$ and $k_i(s) = -g(t_i - 2s)/\sigma_0$, where $t_i = iT_0/m$, $i = 1, \ldots, m$. The exact solution u^* was approximated at $n = 30$ equidistant locations $\tau_j = jX_0/n$ by computing $Q(t) \in \mathbb{R}^{m,m}$ (numerical integration by the trapezoidal rule on a fine grid) and $c \in \mathbb{R}^m$ (exact integration) for every $t = \tau_j$, solving (2.69), and then computing $u_m(t) = \sum_j y_j v_j$. The graph of u_m is shown in the left part of Fig. 2.10 (in black) as compared to the graph of u^* (in red). The reconstruction is good except at the right border of the interval $[0, X_0]$. To see why this is so, note again from considering (2.70), that the averaging kernel $\varphi(s,t) = \sum_j v_j k_j(s)$ is required to be sharply peaked at $s = t$ and to be approximately zero elsewhere, if the reconstruction is to be good at $s = t$. But since all functions $k_j(s), j = 1, \ldots, m$, vanish at $t = X_0$, we can not expect that $\varphi(s,t)$ has this quality near $t = X_0$. As shown in Fig. 2.10 (to the right) the averaging kernel computed for $t = X_0/2 = 0.25$ (orange line) is peaked as and where it should be, but not so the averaging kernel for $t = X_0 = 0.5$ (green line). ◇

Analysis of the Method

Error estimates and convergence results for the Backus-Gilbert method can be found in Theorems 3.34 and 3.36 of [Kir96]. The approximation error $u^* - u_m$ can only be bounded if

(1) u^* has some degree of smoothness. In Theorem 3.34 of [Kir96], Lipschitz continuity is required or, alternatively, $u^* \in H^1(a, b)$.
(2) The space

$$X_m = \langle k_1, \ldots, k_n \rangle$$

spanned by the Fredholm kernel functions k_i must allow the construction of "good" averaging kernels. More precisely,

$$e_m(t) := \min \left\{ \int_a^b (s-t)^2 \psi(s)^2 \, ds; \ \psi \in X_m, \ \int_a^b \psi(s) \, ds = 1 \right\}$$

must be bounded (uniformly in t).

The difficulties observed in Example 2.12 can be explained by the fact that near $t \approx X_0$ the functions $k_1, \ldots, k_m$ can not be combined to an averaging kernel sharply peaked at t and therefore $e_m(t)$ becomes large for these values of t.

The Approximative Inverse

Assumption 2.10 is still made. The Backus-Gilbert method was already interpreted as a method to approximately invert the operator K from (2.59), but the term "Approximative Inverse" usually is meant to designate a different, although related method. We describe it following [Lou96], restricting Louis' much more general approach to Fredholm integral equations. The same ideas were previously applied in the context of computerized tomography, refer to the very accessible presentation in [Nie86]. To start with, define for a fixed parameter $t \in \mathbb{R}$ the function

$$e_\gamma(\cdot, t) : \mathbb{R} \to \mathbb{R}, \quad s \mapsto e_\gamma(s, t) = \begin{cases} 1/\gamma, & |s-t| \le \gamma/2 \\ 0, & \text{else} \end{cases}, \tag{2.71}$$

where $\gamma > 0$ is another parameter. The integral

$$\int_a^b e_\gamma(s, t) u(s) \, ds \quad = \quad \langle u | e_\gamma(\cdot, t) \rangle_{L_2(a,b)},$$

which can be computed for any function $u \in L_2(a,b)$, gives a mean value of u, locally averaged in a neighbourhood of t, the size of which is controlled by γ. It defines a continuous function

$$u_\gamma : [a,b] \to \mathbb{R}, \quad t \mapsto \langle u | e_\gamma(.,t) \rangle_{L_2(a,b)}$$

which can be considered a smoothed approximation of u. For this reason e_γ commonly is called a **mollifier**. Further let us introduce the operator

$$K^* : \mathbb{R}^m \to L_2(a,b), \quad y \mapsto K^* y, \quad K^* y(s) = \sum_{i=1}^{m} y_i k_i(s) \tag{2.72}$$

which is called **adjoint operator** with respect to K, since

$$\langle Ku | y \rangle \;=\; \langle u | K^* y \rangle_{L_2(a,b)} \quad \text{for all} \quad u \in L_2(a,b) \text{ and } y \in \mathbb{R}^m, \tag{2.73}$$

where $\langle \bullet | \bullet \rangle$ on the left hand side means the Euclidean scalar product. Assume one could find a solution $v_\gamma \in \mathbb{R}^m$ of the equation

$$K^* v \;=\; e_\gamma(\cdot,t), \quad v \in \mathbb{R}^m, \tag{2.74}$$

then, by virtue of (2.73), one would get

$$u_\gamma(t) = \langle u | e_\gamma(\cdot,t) \rangle_{L_2(a,b)} = \langle u | K^* v_\gamma \rangle_{L_2(a,b)} = \langle Ku | v_\gamma \rangle = \langle y | v_\gamma \rangle = \sum_{i=1}^{m} y_i v_{\gamma,i}$$

whenever $Ku = y$. This means that one could reconstruct a smooth approximation u_γ of an exact solution u^* of $Ku = y$, with $u_\gamma(t) \to u^*(t)$ for $\gamma \to 0$ and $t \in (a,b)$, if u^* is continuous. It is unlikely, however, that a solution of equation (2.74) exists, since the range of K^* is an at most m-dimensional subspace of $L_2(a,b)$. Alternatively, one determines a vector $v = v_\gamma \in \mathbb{R}^m$ such that $\|K^* v - e_\gamma(\cdot,t)\|_{L_2(a,b)}$ becomes minimal. If the functions $k_i \in L_2(a,b)$, $i = 1,\ldots,m$, are linearly independent, then it follows from Theorem B.5 (the projection theorem), that this optimization problem has a unique solution, which can be determined by solving the equations

$$\langle K^* v - e_\gamma(\cdot,t) | k_i \rangle_{L_2(a,b)} = 0, \quad i = 1,\ldots,m.$$

A concise formulation is

$$Av = b(t), \tag{2.75}$$

where $A \in \mathbb{R}^{m,m}$ is a positive definite matrix with components

$$a_{ij} = \int_a^b k_i(s) k_j(s)\, ds, \quad i,j = 1,\ldots,m, \tag{2.76}$$

independent of t, and where $b(t) \in \mathbb{R}^m$ has components

$$b_i(t) = \int_a^b e_\gamma(s,t) k_i(s)\, ds, \quad i = 1, \ldots, m. \tag{2.77}$$

Formally, one may define an operator

$$S_\gamma : \mathbb{R}^m \to L_2(a,b), \quad y \mapsto x_\gamma, \quad x_\gamma(t) = \sum_{i=1}^m y_i v_{\gamma,i}, \quad v_\gamma = A^{-1} b(t), \tag{2.78}$$

which is called **approximative inverse** of K. It can be used to compute an approximation

$$\tilde{u}_\gamma = S_\gamma y \approx u_\gamma \approx u^*$$

of a solution u^* of $Ku = y$. A prerequisite for A^{-1} to exist is the linear independence of the functions k_i. Technically, one has to choose $\gamma > 0$ and $t \in [a,b]$ and one then performs the following three steps.

(1) Compute the positive definite matrix $A \in \mathbb{R}^{m,m}$ defined in (2.76) and the vector $b(t) \in \mathbb{R}^m$ defined in (2.77).
(2) Find the solution $v_\gamma \in \mathbb{R}^m$ of the linear system (2.75).
(3) Compute $\tilde{u}_\gamma(t) = \sum_{i=1}^m y_i v_{\gamma,i} \approx u_\gamma(t)$.

These steps can be repeated for any other parameter t. This is very similar to the method of Backus and Gilbert, with two differences. First, one no longer tries to recover u itself, but rather a smoothed (or "mollified") approximation of it. Second, the parameter t *does not enter the matrix* A (which can be computed and factorized once and for all), but only into the right hand side $b(t)$ of (2.75). Thus, the computational burden is much reduced. The method can be generalized with respect to the choice of e_γ, see [Lou96].

Example 2.13 Linearized seismic tomography as in Example 2.12 is reconsidered, with all parameter values retained. Reconstructions were performed using the approximate inverse as in (2.78), with $\gamma = 0.05$. The matrix A and the right hand side $b(t)$ were computed by numerical integration (trapezoidal rule on a fine grid). Figure 2.11, to the left, shows the function u^* (in red), its mollified approximation u_γ (in green), and the achieved reconstruction $\tilde{u}_\gamma$ (in black). To see why the reconstruction quality diminishes at the right boundary, look at equation (2.74), which the solution v_γ of (2.75) has to satisfy approximately, if $\tilde{u}_\gamma$ is expected to approximate u_γ well. This means that $\sum_i v_{\gamma,i} k_i$ must be a good approximation of $e_\gamma(\cdot,t)$. In the right picture, we illustrate $e_\gamma(\cdot,t)$ (solid black line) and its (rather good) approximation $\sum_i v_{\gamma,i} k_i$ (orange line) for $t = 0.25$. For $t = 0.4667$, near the right border, $e_\gamma(\cdot,t)$ (dashed black line) and $\sum_i v_{\gamma,i} k_i$ (green line) differ grossly.

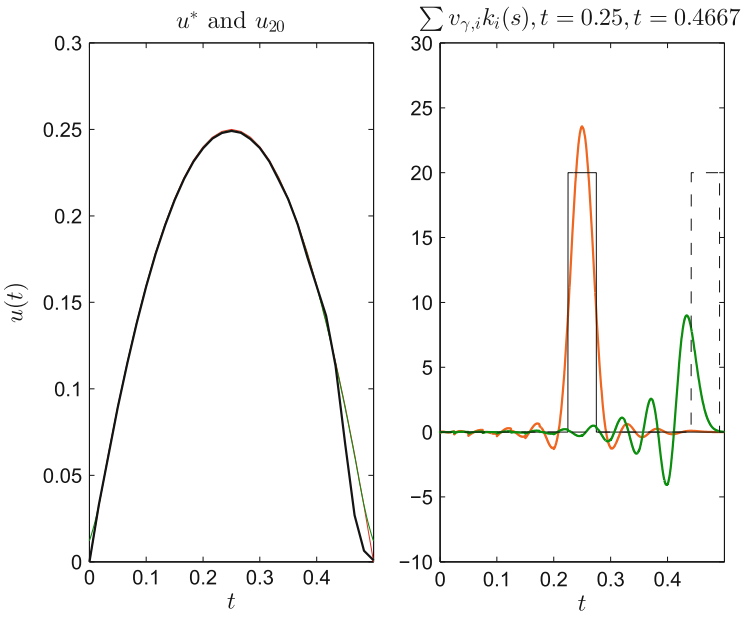

Fig. 2.11 Reconstruction by the approximate inverse and two averaging kernels

The difficulties caused by Fredholm kernel functions k_i vanishing near the right boundary, which already showed up for the Backus-Gilbert method, persist. ◇

2.5 Discrete Fourier Inversion of Convolutional Equations

In this section we make use of the definitions and results presented in Appendix C. Many linear inverse problems take the form of convolutional Fredholm equations as introduced – in the one-dimensional case – in (1.10). For a multi-dimensional example consider Problem 1.10 of inverse gravimetry. To recall, one is asked to solve an equation of the type

$$w(x_1, x_2) = \int_{-a}^{a} \int_{-a}^{a} k(x_1 - y_1, x_2 - y_2) u(y_1, y_2)\, dy_2\, dy_1 \tag{2.79}$$

where $k \in L_2(\mathbb{R}^2)$ is defined by

$$k(u, v) := \left(u^2 + v^2 + u_0^2\right)^{-3/2}, \quad u, v \in \mathbb{R}, \quad u_0 > 0, \tag{2.80}$$

The continuous function $w : \mathbb{R}^2 \to \mathbb{R}$ is known for $-b \le x_1, x_2 \le b$ and the solution function $u^* : [-a, a]^2 \to \mathbb{R}$ of (2.79) is continuous, too. Any $u \in C([-a, a]^2)$ can be extended to become a $L_2(\mathbb{R}^2)$-function by setting $u(x) := 0$ for $x \notin [-a, a]^2$. With this extension, the domain of integration in (2.79) formally can be extended to $\mathbb{R}^2$, such that (2.79) can be written as a two-dimensional convolution:

$$w(x_1, x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k(x_1 - y_1, x_2 - y_2) u(y_1, y_2)\, dy_2\, dy_1. \tag{2.81}$$

In the same way, convolutions are defined in the general case of s-variate L_2-functions (i.e. of functions having s arguments), see (C.7), where the usual notation

$$w = k * u, \quad u, k, w \in L_2(\mathbb{R}^s) \tag{2.82}$$

is used to express (2.81) (with $s = 2$). We already saw in (1.64), that the linearized Problem 1.12 of seismic tomography also can be cast in this form (with $s = 1$). From (C.8) one gets for the Fourier transform of w:

$$w(x) = (k * u)(x) \multimap\!\bullet\ \hat{k}(y)\hat{u}(y) \quad \text{for} \quad k(x) \multimap\!\bullet\ \hat{k}(y),\ u(x) \multimap\!\bullet\ \hat{u}(y).$$

Thus, under the condition that the Fourier transformed kernel function

$$\hat{k}(y) = \int_{\mathbb{R}^s} k(x) \mathrm{e}^{-2\pi i x \cdot y}\, dy$$

has no zeros, a principle possibility to solve (2.82) for u consists in performing the following three steps, which is what is meant by "Fourier inversion of convolutional equations":

1. Compute the Fourier transform $\hat{w}$ of w.
2. Compute $\hat{u} = \hat{w}/\hat{k}$.
3. Compute the inverse Fourier transform u of $\hat{u}$.

Fourier inversion is not possible for the linearized model problem of seismic tomography, since the (one-dimensional) Fourier transform of the Ricker pulse $k := g$ with g defined in (2.30) is given by

$$\hat{k}(y) = \frac{2y^2}{\sqrt{\pi} f_0^3} \mathrm{e}^{-y^2/f_0^2 - 2\pi i y/f_0} \tag{2.83}$$

and *does* have a zero at frequency $y = 0$. The situation is different for the linearized problem of inverse gravimetry: For k as defined in (2.80) one gets the two-dimensional Fourier transform

$$\hat{k}(y) = \hat{k}(y_1, y_2) = \frac{2\pi}{u_0} \mathrm{e}^{-2\pi u_0 \sqrt{y_1^2 + y_2^2}}, \tag{2.84}$$

which has no zeroes. Nevertheless, the division $\hat{w}/\hat{k}$ means to multiply high frequency components of w by huge values. This will have a desastrous effect when only a noisy approximation $\tilde{w}$ of w is known, because at high frequencies the spectrum of noisy data usually is dominated by noise. Noise will therefore blow up catastrophically when $\hat{\tilde{w}}/\hat{k}$ is computed instead of $\hat{w}/\hat{k}$. It is the task of regularization to cope with this problem, as detailed in the next chapter. For the moment we only ask how Fourier inversion can be discretized.

Description of the Method

With an intended application to the model problem of linearized gravimetry in mind, a discretization of Fourier inversion will only be discussed for bivariate functions. The one-dimensional case is briefly summarized in Appendix C, whereas a generalization to spaces of arbitrary dimension $s \in \mathbb{N}$, albeit possible, is omitted.

According to what was said after (2.79), we think of Fourier inversion (for the linearized gravimetry model problem) as a method to reconstruct a function

$$u \in C(\overline{Q}) \cap L_2(\mathbb{R}^2) \cap L_1(\mathbb{R}^2), \quad Q = (-a, a)^2, \quad a > 0,$$

vanishing outside $\overline{Q}$. Such a function can be approximated by a bilinear spline function. Let $N \in \mathbb{N}$ be an even number and let

$$h := \frac{2a}{N} \quad \text{and} \quad W := \left\{ \alpha \in \mathbb{Z}^2;\ -\frac{N}{2} \le \alpha_j < \frac{N}{2},\ j = 1, 2 \right\}. \tag{2.85}$$

For samples of u on an equidistant grid let us write

$$u_\alpha := u(x_\alpha), \quad \text{where} \quad x_\alpha = (h\alpha_1, h\alpha_2),\ \alpha \in W. \tag{2.86}$$

A bilinear B-spline adapted to the grid points x_α is defined by

$$\Phi(x) := B_2(x_1) \cdot B_2(x_2), \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{2.87}$$

based in turn on the univariate B-spline B_2 defined by

$$B_2(t) := \begin{cases} t + 1, & -1 \le t \le 0 \\ 1 - t, & 0 \le t \le 1 \\ 0, & \text{else} \end{cases},$$

as in (C.11). See also Sect. 2.1, where the different notation $N_{i,j,2}$ was used for bilinear B-splines with non-equidistant knots and see Fig. 2.3 for an illustration.

A bilinear spline interpolant of u is given by

$$u_N(x) := \sum_{\alpha \in W} u_\alpha \Phi(x/h - \alpha), \quad x \in \mathbb{R}^2. \tag{2.88}$$

This interpolant can be Fourier transformed *exactly*:

$$\widehat{u_N}(y_1, y_2) = h^2 \left(\frac{\sin(\pi h y_1)}{\pi h y_1} \right)^2 \left(\frac{\sin(\pi h y_2)}{\pi h y_2} \right)^2 \cdot \sum_{\alpha \in W} u_\alpha \mathrm{e}^{-2\pi i h(y_1\alpha_1 + y_2\alpha_2)}. \tag{2.89}$$

For $y = \beta/(2a)$, $\beta \in \mathbb{Z}^2$, one gets:

$$\widehat{u_N}\left(\frac{\beta}{2a}\right) = \sigma_\beta \cdot \underbrace{\left(\frac{1}{N}\right)^2 \sum_{\alpha \in W} u_\alpha \mathrm{e}^{-2\pi i(\alpha_1\beta_1 + \alpha_2\beta_2)/N}}_{=: U_\beta} \tag{2.90}$$

with data independent **attenuation factors**

$$\sigma_\beta := 4a^2 \left(\frac{\sin(\pi\beta_1/N)}{\pi\beta_1/N} \right)^2 \cdot \left(\frac{\sin(\pi\beta_2/N)}{\pi\beta_2/N} \right)^2, \quad \beta \in \mathbb{Z}^2, \tag{2.91}$$

and values U_β, which need only be computed for $\beta \in W$ because of their periodicity. In fact, for an arbitrary index $\beta \in \mathbb{Z}^2$, there exists a unique index $\gamma \in W$ such that $\beta = \gamma + N\alpha$ for some $\alpha \in \mathbb{Z}^2$, and $U_\beta = U_\gamma$. The computation of values U_β, $\beta \in W$, from values u_α, $\alpha \in W$, is called (two-dimensional) **discrete Fourier transform (DFT)**. Inversely, the values u_α can be computed from U_β by

$$u_\alpha = \sum_{\beta \in W} U_\beta \mathrm{e}^{+2\pi i(\alpha_1\beta_1 + \alpha_2\beta_2)/N}, \quad \alpha \in W, \tag{2.92}$$

which is the two-dimensional **inverse discrete Fourier transform (IDFT)**. We will use the notation

$$\{u_\alpha\}_{\alpha \in W} \circ\!\!-\!\!\bullet \{U_\beta\}_{\beta \in W} \quad \text{and} \quad \{U_\beta\}_{\beta \in W} \bullet\!\!-\!\!\circ \{u_\alpha\}_{\alpha \in W}$$

to express DFT and IDFT. If N is a power of two, the two-dimensional DFT and IDFT can both be computed with $\mathcal{O}(N^2 \log(N))$ arithmetical operations by the two-dimensional FFT algorithm. Refer to [PTVF92], pp. 521 ff. for a detailed description of this algorithm. In case one does not dispose of equidistant samples of u or in case one wants to compute non-equidistant samples of its Fourier transform $\hat{u}$, there also exist (more complex) efficient algorithms. Appendix C explains one approach for the one-dimensional case and gives a pointer to the literature for the multi-dimensional case.

Formula (2.90) can also be used "backwards", by which we mean the following: assume that one does not know $u \in C(\overline{Q}) \cap L_2(\mathbb{R}^2) \cap L_1(\mathbb{R}^2)$, but rather its Fourier transform $\hat{u}$ (which necessarily is a continuous function). Then an approximant u_N of u can be constructed by setting

$$U_\beta := \frac{\hat{u}\,(\beta/2a)}{\sigma_\beta}, \quad \beta \in W, \tag{2.93}$$

and by computing

$$\{U_\beta\}_{\beta\in W} \bullet\!\!-\!\!\circ \{u_\alpha\}_{\alpha\in W}, \quad u_N(x) = \sum_{\alpha\in W} u_\alpha \Phi(x/h - \alpha). \tag{2.94}$$

This amounts to adjusting u_N such that the Fourier interpolation conditions

$$\widehat{u_N}\left(\frac{\beta}{2a}\right) = \hat{u}\left(\frac{\beta}{2a}\right), \quad \beta \in W, \tag{2.95}$$

hold. We remark that because of $\sin(x)/x \geq 2/\pi$ for $|x| \leq \pi/2$, there is no danger of a zero division in (2.93). We propose the following method for an approximate solution of (2.82).

Approximate inversion of convolutional equation $w = k * u$.

- Compute via DFT:

$$\{w_\alpha\}_{\alpha\in W} \circ\!\!-\!\!\bullet \{W_\beta\}_{\beta\in W} \quad \text{where} \quad w_\alpha := w(x_\alpha),\ \alpha \in W. \tag{2.96}$$

- Compute

$$U_\beta := \frac{W_\beta}{\hat{k}(\beta/2a)}, \quad \beta \in W. \tag{2.97}$$

- Compute via IDFT:

$$\{U_\beta\}_{\beta\in W} \bullet\!\!-\!\!\circ \{u_\alpha\}_{\alpha\in W} \tag{2.98}$$

and take

$$u_N(x) = \sum_{\alpha\in W} u_\alpha \Phi(x/h - \alpha) \tag{2.99}$$

as an approximant of u.

The same method can be used in the one-dimensional case, using a one-dimensional grid $W = \{-N/2, \ldots, N/2 - 1\}$. The idea behind the first step is to take

$$w_N(x) = \sum_{\alpha \in W} w_\alpha \Phi(x/h - \alpha), \quad w_\alpha = w(x_\alpha),\ \alpha \in W, \tag{2.100}$$

as an approximant of w and make use of its Fourier transform values

$$\widehat{w_N}\left(\frac{\beta}{2a}\right) = \sigma_\beta W_\beta \approx \hat{w}\left(\frac{\beta}{2a}\right). \tag{2.101}$$

The idea behind the last two steps is to determine the coefficients u_α of an approximant

$$u_N(x) = \sum_{\alpha} u_\alpha \Phi(x/h - \alpha) \approx u(x)$$

such that

$$\hat{k}\left(\frac{\beta}{2a}\right)\hat{u}\left(\frac{\beta}{2a}\right) \approx \hat{k}\left(\frac{\beta}{2a}\right)\widehat{u_N}\left(\frac{\beta}{2a}\right) = \hat{k}\left(\frac{\beta}{2a}\right)\sigma_\beta U_\beta = \sigma_\beta W_\beta = \\ = \widehat{w_N}\left(\frac{\beta}{2a}\right) \approx \hat{w}\left(\frac{\beta}{2a}\right), \quad \beta \in W.$$

Application to Linearized Model Problem 1.10 of Inverse Gravimetry

It was already found that Fourier inversion is (mathematically) viable for Problem 1.10, since the Fourier transform of the kernel function has no zeros, see (2.84). To find an approximant u_N of u as in (2.99), one has to (approximately) know the values

$$\hat{u}\left(\frac{\beta}{2a}\right) = \left[\hat{k}\left(\frac{\beta}{2a}\right)\right]^{-1} \hat{w}\left(\frac{\beta}{2a}\right), \quad \beta \in W. \tag{2.102}$$

Fourier transform values of k can be computed analytically. Approximations of $\hat{w}(\beta/(2a))$ could be computed according to (2.96) if the values $w(x_\alpha)$, $\alpha \in W$, were known. For the sake of simplicity we will actually assume that they *are* known, thereby also presuming that $b = a$ in the setting of Problem 1.10. In practice, we can not rely on our good luck to have just the values $w(x_\alpha)$, $\alpha \in W$, available. It is still possible to efficiently compute $\hat{w}(\beta/(2a))$, $\beta \in W$, from non-equidistantly

sampled function values of w, even if their number is not equal to N^2. One such approach was developed by Fourmont, see [Fou99] and Appendix C.

The factors $1/\hat{k}(\beta/2a)$ are exponentially growing with $\|\beta\|_2$ and small errors in the computed values $\hat{w}(\beta/2a)$ will therefore be catastrophically amplified in (2.102). Without regularization, formula (2.102) is practically useless. For this reason numerical results for Fourier inversion will only be presented after introducing a regularized reconstruction in Sect. 3.8.

Analysis of the Method

A detailed investigation of the approximation errors associated with the discretized Fourier reconstruction process is given in Appendix C. It turns out that u can be approximated well by u_N if the Fourier interpolation condition (2.95) holds (the corresponding error is estimated in Lemma C.4) and $\hat{w}$ can also be approximated well by $\widehat{w_N}$ if w_N is chosen as a bilinear spline interpolant of w (the corresponding error is estimated in Lemma C.3). The essential problem with Fourier inversion is the division step (2.97): whenever k is a smooth, rapidly decaying function, then $\hat{k}$ also decays rapidly for large arguments; even small errors in the computed coefficients W_β are then grossly magnified. This difficulty will be addressed in Sect. 3.8.

2.6 Discretization of Nonlinear Model Problems

There are many ways to discretize nonlinear problems and there is no generally recommendable approach. (Admittedly, even for linear problems the methods presented in Sects. 2.2 and 2.3 could be judged oversimplified, since by Assumption 2.6 no constraints were considered, nor were adaptive or multiscale discretizations discussed, which are of great practical importance). In the present section specific discretizations for the specific model problems 1.9 and 1.11 from inverse gravimetry and seismic tomography will be considered.

Discretization of Model Problem 1.9: Inverse Gravimetry

The goal is to discretize a non-linear Fredholm equation of the first kind

$$w(x_1, x_2) = \int_{-a}^{a} \int_{-a}^{a} k(x_1, x_2, y_1, y_2, u(y_1, y_2))\, dy_2\, dy_1, \tag{2.103}$$

where the kernel function is defined by

$$k(x_1, x_2, y_1, y_2, z) := \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + z^2}} \tag{2.104}$$

for $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$ and $z > 0$ and where

$$\mathbb{U} := \left\{u : [-a, a]^2 \to [b_1, b_2];\ u \text{ continuous},\ 0 < b_1 < b_2\right\}$$

is the set of admissible causes. Here, the parameters h and H from Problem 1.9 were renamed into b_1 and b_2, respectively. The continuous function $w \in C(\mathbb{R}^2)$ is observed for $x \in [-b, b]^2$ and the corresponding exact solution of (2.103) will be denoted by u^*.

A very simple discretization scheme will be used, based on equidistant grid points

$$x_\alpha := h\alpha, \quad h := \frac{a}{n}, \quad \alpha \in G_n := \{(\alpha_1, \alpha_2) \in \mathbb{Z}^2;\ -n \le \alpha_j \le n,\ j = 1, 2\}, \tag{2.105}$$

for $n \in \mathbb{N}$ to be chosen. As an approximant of the unknown function u^* we use

$$u^*(x) \approx u_n(x) = \sum_{\alpha \in G_n} c_\alpha \Phi(x/h - \alpha) \tag{2.106}$$

where the parameters c_α are yet unknown and where the bilinear B-spline

$$\Phi(x) := B_2(x_1) \cdot B_2(x_2), \quad x = (x_1, x_2) \in \mathbb{R}^2 \tag{2.107}$$

is defined exactly as in (2.87). Any approximant u_n of the form (2.106) is continuous and the constraint $u_n \in \mathbb{U}$ immediately translates into **box constraints**

$$c_\alpha \in [b_1, b_2], \quad \alpha \in G_n, \tag{2.108}$$

for the parameters c_α. Substituting u_n for u, integral (2.103) can be evaluated approximately for any $x = (x_1, x_2)$ using the two-dimensional trapezoidal rule, *based on the same grid* as the approximant u_n:

$$w(x) \approx w_n(x) := \sum_{\alpha \in G_n} \omega_\alpha k(x, x_\alpha, c_\alpha) \tag{2.109}$$

with constant weights

$$\omega_\alpha := \begin{cases} h^2/4, & \text{if } |\alpha_1| = |\alpha_2| = n, \\ h^2/2, & \text{if } |\alpha_1| = n \text{ or } |\alpha_2| = n, |\alpha_1| + |\alpha_2| < 2n, \\ h^2, & \text{else.} \end{cases} \tag{2.110}$$

To arrive at (2.109), one makes use of the identity $u_n(x_\alpha) = c_\alpha$, $\alpha \in G_n$, which is the reason why the trapezoidal rule was based on the grid G_n. More sophisticated quadrature rules exist than the trapezoidal rule, but since we can not guarantee any smoothness of the integrand beyond continuity, there is no guarantee that these rules produce more accurate results. Approximating (2.103) by (2.109) is known as **Nyborg's method**. Suppose now in accordance with Problem 1.9, that the observation of w consists in sampling values

$$w(\hat{x}_\beta), \quad \hat{x}_\beta \in [-b, b]^2, \quad \beta \in B, \tag{2.111}$$

where the sample points $\hat{x}_\beta$ are pairwise different and where $B \subset \mathbb{Z}^2$ is a finite index set. This leads to the following nonlinear system of equations

$$y_\beta := w(\hat{x}_\beta) = \sum_{\alpha \in G_n} \omega_\alpha k(\hat{x}_\beta, x_\alpha, c_\alpha), \quad \beta \in B, \tag{2.112}$$

consisting of

$$M := |B| \quad \text{equations for} \quad N := (2n+1)^2 = |G_n| \quad \text{unknowns } c_\alpha. \tag{2.113}$$

Summarizing observations and unknowns as vectors[3]

$$y = (y_\beta;\ \beta \in B) \in \mathbb{R}^M, \quad c = (c_\alpha;\ \alpha \in G_n) \in \mathbb{R}^N, \tag{2.114}$$

and defining $C := [b_1, b_2]^N$ and a mapping

$$F : C \to \mathbb{R}^M, \quad c \mapsto F(c) = (F_\beta(c);\ \beta \in B)$$

by setting

$$F_\beta(c) = \sum_{\alpha \in G_n} \omega_\alpha k(\hat{x}_\beta, x_\alpha, c_\alpha), \quad \beta \in B, \tag{2.115}$$

this system can be written in condensed form as

$$F(c) = y, \quad c \in C. \tag{2.116}$$

As a consequence of discretization errors, even if exact values $c_\alpha^* = u^*(x_\alpha)$ were used, $w(\hat{x}_\beta) \approx F_\beta(c^*)$ would not be an exact equality. Therefore, a solution of

[3] Since two-dimensional index sets can be ordered in multiple ways, the following is not a mathematically precise definition. For example, the vector c could consist of elements $c_{(-n,-n)}, \ldots, c_{(-n,n)}, \ldots, c_{(n,-n)}, \ldots, c_{(n,n)}$ in that order ("rowwise ordering") or in any other order as well.

system (2.116) can not be expected to exist[4] and this system will be replaced by a minimization problem with box constraints

$$\text{minimize} \quad \frac{1}{2}\|y - F(c)\|_2^2, \quad c \in C = [b_1, b_2]^N. \tag{2.117}$$

This minimization problem will be investigated further and solved in Sects. 4.2 and 4.6.

Discretization of Model Problem 1.11: Seismic Tomography

Model Problem 1.11 was formulated in Sect. 1.4. To recall, for

$$\mathscr{S} = \left\{\sigma \in H^1(0, X_0);\ 0 < \sigma_- \le \sigma(x) \le \sigma_+ < \infty \ \text{ for } \ 0 \le x \le X_0\right\},$$

an operator $T : \mathscr{S} \to L_2(0, T_0)$ was defined, which maps an acoustic impedance $\sigma \in \mathscr{S}$ to an observed seismogram $Y_d \in L_2(a, b)$. The inverse problem to solve has the form of an equation

$$T(\sigma) = Y_d, \quad \sigma \in \mathscr{S}.$$

To formulate a discrete version, acoustic impedances $\sigma \in \mathscr{S}$ will be approximated by splines of order 1. Choose $n \in \mathbb{N}$, set

$$\Delta := X_0/(n+1) \qquad \text{and} \qquad x_k := k\Delta, \quad k = 0, \ldots, n+1, \tag{2.118}$$

and define

$$\begin{aligned} \mathscr{S}_n := \{\sigma : [0, X_0] \to \mathbb{R};\ \sigma_- \le \sigma(x) \le \sigma_+ \ \text{ for } \ 0 \le x \le X_0 \ \text{ and} \\ \sigma(x) \equiv \text{ const. for } x_k \le x < x_{k+1},\ k = 0, \ldots, n\} \end{aligned} \tag{2.119}$$

(with supplementary agreement $\sigma(X_0) = \sigma(x_{n+1}) := \sigma(x_n)$ for $\sigma \in \mathscr{S}_n$). $\mathscr{S}_n$ is a set of acoustic impedances having constant values on fixed travel time intervals of length Δ. It coincides with the spline space $\mathscr{S}_1(x_0, \ldots, x_{n+1})$ introduced in Sect. 2.1. Every $\sigma \in \mathscr{S}_n$ can uniquely be written in the form

$$\sigma(x) = \sum_{j=1}^{n+1} \sigma_j N_{j-1,1}(x), \quad x \in [0, X_0], \tag{2.120}$$

[4] Already the basic assumption made in Problem 1.9, namely that a body of constant mass density includes another body of a different, constant mass density, is an idealization and means a modelling error.

where $N_{j-1,1}$ are the B-splines of order 1 defined in (2.2). Because of (2.120), every $\sigma \in \mathscr{S}_n$ can be identified with a vector $(\sigma_1, \sigma_2, \ldots, \sigma_{n+1}) \in \mathbb{R}^{n+1}$. It can also be interpreted as a spline of order 1 interpolating an exact impedance $\sigma_e \in \mathscr{S}$ at $x_0, \ldots, x_n$. Any $\sigma_e \in \mathscr{S}$ can be approximated arbitrarily well (with respect to the norm $\| \bullet \|_{L_2(0,X_0)}$) by its interpolant $\sigma \in \mathscr{S}_n$, if n is chosen large enough (compare Theorem 2.2). However,

$$\mathscr{S}_n \not\subset \mathscr{S}.$$

In fact, a function $\sigma \in \mathscr{S}_n$ possibly is not even continuous and can not be admitted offhand as a parameter in the wave equation (1.42). We *do admit* it nevertheless, but can no longer require the equation

$$\sigma(x)\frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial y(x,t)}{\partial x}\right) = 0, \quad x \in (0, X_0), t \in (0, T_0), \tag{1.42}$$

to hold literally. To give this equation a meaning for $\sigma \in \mathscr{S}_n$, let us define domains

$$D_k := \{(x,t);\ x_{k-1} < x < x_k,\ 0 < t < T_0\}, \quad k = 1, \ldots, n+1.$$

For $\sigma \in \mathscr{S}_n$ and for $(x,t) \in D_k$, (1.42) simplifies to

$$\frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial^2 y(x,t)}{\partial x^2} = 0, \quad (x,t) \in D_k, \tag{2.121}$$

with general solution

$$y(x,t) = F(x-t) + G(x+t) \tag{2.122}$$

where F and G are "arbitrary" univariate functions, see, e.g. Section 2.4 in [Eva98]. If F and G are differentiable twice, so is y, as required by equation (2.121). Formula (2.122) can also be used as a *definition* of y for $F, G \in C^1(\mathbb{R})$, although in this case (2.121) does not hold (second derivatives are not defined for y). The function y will nevertheless be called a **weak solution** of (2.121), if it is defined by (2.122) with $F, G \in C^1(\mathbb{R})$. We call y a **weak solution of (1.42)**, if it is a weak solution of (2.121) on every domain D_k and if moreover the continuity relations

$$\lim_{x \to x_k +} \frac{\partial y(x,t)}{\partial t} = \lim_{x \to x_k -} \frac{\partial y(x,t)}{\partial t} \tag{2.123}$$

and

$$\lim_{x \to x_k +} \frac{\sigma(x)\partial y(x,t)}{\partial x} = \lim_{x \to x_k -} \frac{\sigma(x)\partial y(x,t)}{\partial x} \tag{2.124}$$

hold for $k = 1, \ldots, n$. Equation (2.123) means that there are no shifts in particle velocity accross layer interfaces and equation (2.124) means that there are no shifts in sound pressure.

Remark It would even be possible to require F and G to be H^1-functions only, but then the partial derivatives of y_k would exist only almost everywhere and more care would have to be taken to formulate equations (2.123) and (2.124).

Theorem 2.14 (Seismic record for media with piecewise constant impedance) *Under the conditions $\sigma \in \mathscr{S}_n$ and $g \in C[0, T_0]$, $g(0) = 0$, the system (1.42), (1.43), and (1.44) has a unique weak solution, whose trace $\partial y(0,t)/\partial t$ belongs to $C[0, T_0]$. The map*

$$T : C[0, T_0] \to C[0, T_0], \quad g \mapsto Y_d \quad \text{with} \quad Y_d(t) = \frac{\partial y(0,t)}{\partial t} \tag{2.125}$$

is explicitly given by

$$Tg(t) = \frac{1}{\sigma_1} \sum_{k=0}^{n} \lambda_k g(t - t_k) \tag{2.126}$$

where $t_k := 2k\Delta$, where g is continued by 0 for $t < 0$, and where the coefficients λ_k are computed as follows. First define

$$r_0 := 0, \quad r_k := \frac{\sigma_k - \sigma_{k+1}}{\sigma_k + \sigma_{k+1}}, \ k = 1, \ldots, n. \tag{2.127}$$

Then compute state variables $u_{i,j}$ and $v_{i,j}$ by the linear recursions

$$\begin{aligned} u_{i,j} &= (1 - r_{i-j})u_{i,j-1} + r_{i-j}v_{i-1,j} \\ v_{i,j} &= -r_{i-j}u_{i,j-1} + (1 + r_{i-j})v_{i-1,j} \end{aligned} \qquad i = 1, \ldots, n, \ j = 0, \ldots, i \tag{2.128}$$

with boundary and initial conditions

$$\begin{aligned} u_{i,-1} &= 0 \\ v_{i-1,i} &= u_{i,i} \end{aligned} \quad i = 1, \ldots, n, \qquad \begin{aligned} u_{0,0} &= 0 \\ v_{0,0} &= 2 \end{aligned}. \tag{2.129}$$

Finally set

$$\lambda_i = \frac{1}{2}(u_{i,i} + v_{i,i}), \quad i = 0, 1, \ldots, n. \tag{2.130}$$

Proof Details are given in [BCL77], but since this reference is not easily accessible, we outline a proof below.

Part I We first collect well known results for the one-dimensional wave equation, compare, e.g., Section 2.4 of [Eva98]. Since σ is piecewise constant, within each region $D_k := \{(x,t);\ x_{k-1} < x < x_k,\ 0 < t < T_0\}$ one can reformulate (1.42) as

$$\sigma_k \frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial}{\partial x}\left(\sigma_k \frac{\partial y(x,t)}{\partial x}\right) = 0 \quad \Longleftrightarrow \quad \frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial^2 y(x,t)}{\partial x^2} = 0\,.$$

Any two times differentiable solution y_k of this equation defines the differentiable functions

$$\begin{aligned} u_k : D_k \to \mathbb{R}, \quad u_k(x,t) &= \frac{\partial y_k(x,t)}{\partial t} + \frac{\partial y_k(x,t)}{\partial x}, \\ v_k : D_k \to \mathbb{R}, \quad v_k(x,t) &= \frac{\partial y_k(x,t)}{\partial t} - \frac{\partial y_k(x,t)}{\partial x}, \end{aligned} \tag{2.131}$$

which are called *upgoing* and *downgoing* waves, since (using (1.42))

$$\begin{aligned} \frac{d}{dt} u_k(x(t),t) = 0 \quad \text{for} \quad \frac{dx(t)}{dt} = -1, \\ \frac{d}{dt} v_k(x(t),t) = 0 \quad \text{for} \quad \frac{dx(t)}{dt} = +1, \end{aligned} \tag{2.132}$$

meaning that u_k and v_k are constant on straight lines $(x(t),t)$ with slopes -1 and $+1$, respectively. These straight lines are called characteristics. Adding and subtracting the equalities in (2.131), one gets

$$\frac{\partial y_k}{\partial t} = \frac{1}{2}(u_k + v_k) \qquad \text{and} \qquad \frac{\partial y_k}{\partial x} = \frac{1}{2}(u_k - v_k). \tag{2.133}$$

Conversely, if u_k and v_k solve (2.132) and y_k satisfies (2.133), then y_k is a solution of (1.42). It is well known that the general solution y_k of (1.42) on D_k is given as

$$y_k(x,t) = F(x-t) + G(x+t), \tag{2.134}$$

where F and G are arbitrary functions, which have to be chosen two times differentiable, if y_k shall be differentiable twice. As stated above, y_k is called a weak solution if $F, G \in C^1(\mathbb{R})$. In this case, u_k and v_k as defined by (2.131) are continuous, square integrable functions on D_k.

Next we look at (1.42), (1.43), and (1.44), restricting arguments to $(x,t) \in [0,\Delta) \times [0,\Delta)$. From (2.134) and from initial and boundary conditions (everything at rest for $t = 0$, wave excitation starts at $x = 0$) one sees that the solution of (1.42) in this case has the form $y_1(x,t) = F(x-t)$. From this and (1.44) one gets $\partial y_1(0,t)/\partial x = F'(-t) = -g(t)/\sigma_1$, which determines F'. We also can deduce $\partial y_1(0,t)/\partial t = -F'(-t) = g(t)/\sigma_1$ and thus we get from (2.131)

$$v_1(0,t) = \frac{2}{\sigma_1} g(t) \quad \text{and} \quad u_1(0,t) = 0, \quad 0 \le t < \Delta. \tag{2.135}$$

The surface excitation propagates down at velocity 1, which means that $v_1(x,t) = 2g(t-x)/\sigma_1$ for $0 \le t < \Delta$ and $0 \le x < \Delta$. Arrived at depth $x = \Delta$, v_1 is in part transmitted to the next deeper layer and in part reflected back into the direction of the surface. We look at transmission and reflection now.

Assume one knows solutions y_k of (1.42) on D_k and y_{k+1} (of (1.42) on D_{k+1}). Using (2.131) we rewrite the continuity conditions for velocity and pressure accross the interface $x = x_k := k\Delta$ of layers k and $k+1$ as

$$\begin{aligned} \frac{u_k(x_k,t) + v_k(x_k,t)}{2} &= \frac{u_{k+1}(x_k,t) + v_{k+1}(x_k,t)}{2} \\ \sigma_k \frac{u_k(x_k,t) - v_k(x_k,t)}{2} &= \sigma_{k+1} \frac{u_{k+1}(x_k,t) - v_{k+1}(x_k,t)}{2} \end{aligned} \tag{2.136}$$

This system of equations can be solved for $u_k(x_k,t)$ and $v_{k+1}(x_k,t)$:

$$\begin{aligned} u_k(x_k,t) &= (1-r_k)u_{k+1}(x_k,t) + r_k v_k(x_k,t) \\ v_{k+1}(x_k,t) &= -r_k u_{k+1}(x_k,t) + (1+r_k)v_k(x_k,t), \end{aligned} \tag{2.137}$$

Part II We have seen that until time $t = \Delta$, everything is at rest except for a single downgoing wave $v_1(x,t) = 2g(t-x)/\sigma_1$ which is then partially reflected and partially transmitted at the interface $x = \Delta$ according to the rules given in (2.137) (where we have to set $u_2 = 0$, since nothing can yet come up from D_2). We can thus compute u_1 and v_2 from (2.137) for the time $\Delta < t < 2\Delta$. At time $t = 2\Delta$, the reflected wave reaches the surface and contributes to Y_d in addition to the ongoing surface excitement. At the same time, the transmitted wave reaches the interface $x = 2\Delta$ and is reflected and transmitted in turn. In Fig. 2.12 the relevant characteristics for one-dimensional wave propagation are shown in the $x-t$-plane. *Since within each layer waves propagate at constant velocity 1 or* -1*, all reflections and transmissions at all interfaces happen synchronously at time instances* $k\Delta$. Continuing the above consideration "in timesteps Δ", one arrives at formulae (2.128) and (2.129), with the following interpretation

(1) Index pairs (i,j) correspond to coordinates where characteristics in the $x-t$-plane intersect and where transmission and reflection occur. The index difference $i-j = k$ corresponds to the interface between layers k and $k+1$ (with an "air layer" 0 above the surface) and $r_{i-j} = r_k$ is the corresponding reflection coefficient as in (2.137). Further, $u_{i,j}$ relates to the wave sent up from layer $i-j$ into layer $i-j-1$ (from node (i,j) to node $(i,j+1)$) at time $(i+j)\Delta$ and $v_{i,j}$ relates to the wave sent down from layer $i-j$ into layer $i-j+1$ (from node (i,j) to node $(i+1,j)$) at time $(i+j)\Delta$.

(2) Since air has an acoustic impedance of (nearly) 0, waves coming up to the surface are completely reflected without transmission. This is expressed by setting $r_0 = 0$ in (2.128) and by setting $v_{i-1,j} = u_{i,i}$ in (2.129). Thus $r_0 = 0$ is not a reflection coefficient.

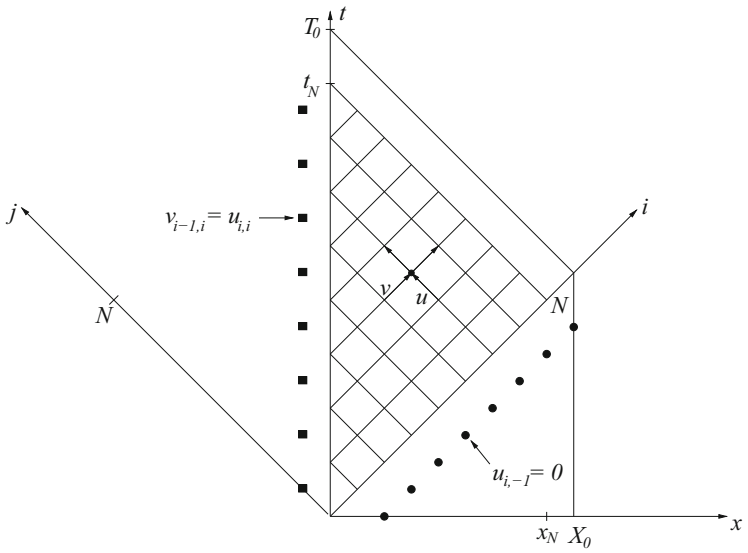

Fig. 2.12 Reflection and transmission of waves

(3) Because everything is at rest at $t = 0$ and waves are only excited at the surface, all upgoing waves must be reflections of downgoing waves. This is expressed by the boundary condition $u_{i,-1} = 0$ in (2.129).
(4) The initial conditions $u_{0,0} = 0$ and $v_{0,0} = 2$ correspond to the factors 0 and 2 in (2.135).

The trace Y_d is computed by

$$Y_d(t) = \frac{\partial y(0,t)}{\partial t} = \frac{1}{2}\left(u(0,t) + v(0,t)\right)$$

where u and v are the functions whose restrictions to D_k were called u_k and v_k, respectively. It was already shown that $v(0,t) = 2g(t)/\sigma_1$ and $u(0,t) = 0$ for $0 \le t < t_1 = 2\Delta$, i.e. we have (2.126) with $\lambda_0 = 1$. At time instances $t = t_k$, $k = 1, \ldots, n$, new reflected copies of this wave arrive at the surface and have to be added to the record. This shows (2.126) with λ_k as defined in (2.130). □

The first factors λ_k read

$$\lambda_0 = 1, \quad \lambda_1 = 2r_1, \quad \lambda_2 = 2r_2(1 - r_1^2) + 2r_1^2, \tag{2.138}$$

as can be verified by direct calculation via the formulae stated in Theorem 2.14. Knowing g and Y_d it is principally possible to determine the coefficients λ_k – this will be discussed in a moment. Once the values λ_k are known, and if we additionally know σ_1, we can determine the other impedance values.

Theorem 2.15 *The mapping*

$$G : \mathbb{R}^n \to \mathbb{R}^n, \quad (r_1, \ldots, r_n) \mapsto (\lambda_1, \ldots, \lambda_n), \tag{2.139}$$

which is defined by (2.128), (2.129), *and* (2.130) *is one to one. Consequently,* $\sigma_2, \ldots, \sigma_{n+1}$ *can be retrieved from* $\lambda_1, \ldots, \lambda_n$ *via* (2.127). *Thus,*

$$H : [\sigma_-, \sigma_+]^n \to \mathbb{R}^n, \quad (\sigma_2, \ldots, \sigma_{n+1}) \mapsto (\lambda_1, \ldots, \lambda_n), \tag{2.140}$$

defined by a given value $\sigma_1 \in [\sigma_-, \sigma_+]$ *and by* (2.127), (2.128), (2.129), *and* (2.130) *also is one to one.*

Proof It is not difficult to show

$$\lambda_i = 2r_i \prod_{k=1}^{i-1}(1 - r_k^2) + P_i(r_1, \ldots, r_{i-1}), \quad i = 1, \ldots, n, \tag{2.141}$$

where $P_i(r_1, \ldots, r_{i-1})$ are polynomials in $r_1, \ldots, r_{i-1}$, see [BCL79], page 29. Since $0 < |r_k| < 1$ for $k = 1, \ldots, n$, the statement follows. □

To set up a fully discrete model, we assume that g is sampled equidistantly and set

$$g_k := g(t_k), \quad t_k := 2k\Delta, \quad k = 0, \ldots, n+1 \tag{2.142}$$

We assume as well, that samples of the seismogram Y_d are taken at the same time instances. Thus, from (2.126) we get the relation

$$Y_k := Y_d(t_k) = \frac{1}{\sigma_1} \sum_{\ell=0}^{n} \lambda_\ell g_{k-\ell}, \quad k = 0, \ldots, n+1, \tag{2.143}$$

where $g_{k-\ell} := 0$ for $\ell > k$. By a proper choice of the time origin we can achieve

$$g_0 = 0 \quad \text{and} \quad g_1 \neq 0 \tag{2.144}$$

to hold. Then, the first members of the sequence $(Y_k)_k$ read

$$\begin{aligned} Y_0 &= 0 \\ Y_1 &= g_1/\sigma_1 \\ Y_2 &= (g_2 + \lambda_1 g_1)/\sigma_1 \\ Y_3 &= (g_3 + \lambda_1 g_2 + \lambda_2 g_1)/\sigma_1 \\ &\vdots \quad \vdots \end{aligned} \tag{2.145}$$

It is easy to see that (2.143) can be uniquely solved for $\lambda_1, \ldots, \lambda_n$ if the values g_k and Y_k are known for $k = 2, \ldots, n+1$. All together, the mapping

$$F : [\sigma_-, \sigma_+]^n \to \mathbb{R}^n, \quad (\sigma_2, \ldots, \sigma_{n+1}) \mapsto (Y_2, \ldots, Y_{n+1}), \tag{2.146}$$

defined by

$$(\sigma_2, \ldots, \sigma_{n+1}) \overset{(2.127)}{\mapsto} (r_1, \ldots, r_n) \overset{(2.128)\text{–}(2.130)}{\mapsto} (\lambda_2, \ldots, \lambda_{n+1}) \overset{(2.143)}{\mapsto} (Y_2, \ldots, Y_{n+1})$$

with σ_1, $g_1, \ldots, g_{n+1}$ known, is one to one. The discrete version of Problem 1.11 thus takes the form: Solve the nonlinear equation

$$F(\sigma) = Y$$

for $\sigma = (\sigma_2, \ldots, \sigma_{n+1})$, where $Y = (Y_2, \ldots, Y_{n+1})$ are known samples of the observed seismogram as defined in (2.143).

It is theoretically possible to explicitly invert the mapping F, see formulae (2.145), (2.141), and (2.127). However, already the solution of (2.143) ("deconvolution") for $\lambda_1, \ldots, \lambda_n$ is very unstable if $|g_1|$ is small, as it is the case for the Ricker pulse. Also, the values λ_k depend quite sensitively on r_k (see (3.36) in [BCL79]). Thus, as the observed seismogram Y_d always contains noise, explicit inversion of F becomes practically unfeasible. A regularized inversion will be considered in Sect. 4.5.

Chapter 3
Regularization of Linear Inverse Problems

The discretization of linear identification problems led to linear systems of algebraic equations. In case a solution does not exist, these can be replaced by linear least squares problems, as already done in Sect. 2.3. We will give a detailed sensitivity analysis of linear least squares problem and introduce their **condition number** as a quantitative measure of ill-posedness. If the condition of a problem is too bad, it can not be solved practically. We introduce and discuss the concept of **regularization** which formally consists in replacing a badly conditioned problem by a better conditioned one. The latter can be solved reliably, but whether its solution is of any relevance depends on how the replacement problem was constructed. We will especially consider Tikhonov regularization and iterative regularization methods. The following discussion is restricted to the Euclidean space over the field of real numbers $\mathbb{K} = \mathbb{R}$, but could easily be extended to complex spaces.

3.1 Linear Least Squares Problems

Solving a system of linear equations

$$Ax = b \quad \text{with} \quad A \in \mathbb{R}^{m,n},\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m,\ m \geq n, \tag{3.1}$$

is an inverse problem: b represents an effect, x represents a cause and "the physical law" is represented by the function $T : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$. The case $m < n$ is excluded in (3.1), because injectivity of the mapping T ("identifiability of a cause") requires $\operatorname{rank}(A) = n$, which is only possible if $m \geq n$. If $m > n$, the system of equations is called **overdetermined**. Inaccuracies in the components of A or b lead to contradictions between the equations of an overdetermined system, such that no

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9_3

exact solution x of (3.1) exists: the **residual**

$$r(x) := b - Ax$$

does not vanish for any $x \in \mathbb{R}^n$. As a substitute for a solution, one can look for a vector x minimizing the residual, for example with respect to the Euclidean norm:

$$\text{Find } \hat{x} \text{ such that} \quad \|r(\hat{x})\|_2 \leq \|r(x)\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{3.2}$$

(3.2) is called **linear least squares problem**. Solving (3.2) means to minimize

$$f(x) = \|r(x)\|_2^2 = r(x)^T r(x) = x^T A^T A x - 2x^T A^T b + b^T b.$$

The gradient of f at x is $\nabla f(x) = 2A^T Ax - 2A^T b$ and since $\nabla f(\hat{x}) = 0$ is a necessary condition for $\hat{x}$ to be a minimizer of f, one gets the so called **normal equations**

$$A^T A\hat{x} = A^T b \quad \Longleftrightarrow \quad A^T \hat{r} = 0 \quad \text{with} \quad \hat{r} := r(\hat{x}). \tag{3.3}$$

These conditions are not only necessary, but also sufficient for $\hat{x}$ to be a minimizer. To see this, take any $x \in \mathbb{R}^n$ with according residual $r(x) = \hat{r} + A(\hat{x} - x)$. Consequently

$$\|r(x)\|_2^2 = \hat{r}^T \hat{r} + \underbrace{2(\hat{x} - x)^T A^T \hat{r}}_{=0} + (\hat{x} - x)^T A^T A(\hat{x} - x) \geq \|\hat{r}\|_2^2,$$

where $\|A(\hat{x} - x)\|_2 = 0 \Leftrightarrow A(\hat{x} - x) = 0 \Leftrightarrow r(x) = \hat{r}$. This proves

Theorem 3.1 (Existence and uniqueness of a solution of the least squares problem) *A necessary and sufficient condition for $\hat{x}$ to be a solution of (3.2) is the compliance with the normal equations (3.3). A unique minimizer $\hat{x}$ exists if and only if all columns of A are linearly independent, i.e. if rank$(A) = n$. The residual $\hat{r}$ is always unique.*

The "full rank condition" rank$(A) = n$ is equivalent to the injectivity of T. As mentioned repeatedly, it would be impossible to identify causes responsible for observed effects, if this condition was violated. Therefore, rank$(A) = n$ will usually be required in the following.

The normal equations have a geometrical interpretation. Since

$$A = \left(\begin{array}{c|c|c|c} & & & \\ a_1 & a_2 & \cdots & a_n \\ & & & \end{array} \right), \quad \text{all } a_j \in \mathbb{R}^m \quad \Longrightarrow \quad A^T \hat{r} = \begin{pmatrix} a_1^T \hat{r} \\ \vdots \\ a_n^T \hat{r} \end{pmatrix},$$

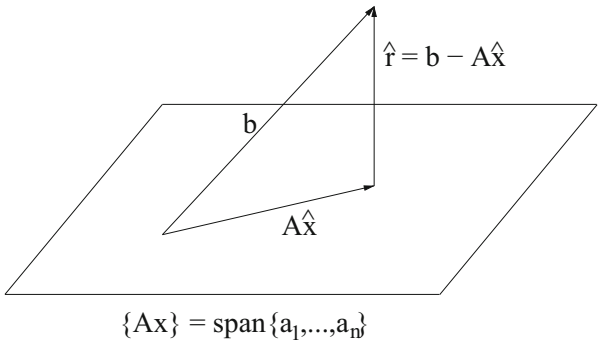

Fig. 3.1 Solution of a linear least squares problem

(3.3) means $\hat{r} \perp \mathscr{R}_A$, i.e. the residual $\hat{r}$ is orthogonal to $\mathscr{R}_A = \text{span}\{a_1, \ldots, a_n\} = \{Ax;\ x \in \mathbb{R}^n\}$, see Fig. 3.1.

Example 3.2 (Straight line of best fit) Assume there is a causal dependence

$$T : \mathbb{R} \to \mathbb{R}, \quad t \mapsto T(t) = \alpha + \beta(t - 4),$$

with unknown parameters $\alpha, \beta \in \mathbb{R}$. Assume further we have the following measurements (taken from [DB74], Example 5.7.3): This corresponds to an overde-

t	1	3	4	6	7
$T(t)$	−2.1	−0.9	−0.6	0.6	0.9

termined system of linear equations for $x = (\alpha, \beta)^T$:

$$\underbrace{\begin{pmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 0 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}}_{=:A} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \underbrace{\begin{pmatrix} -2.1 \\ -0.9 \\ -0.6 \\ 0.6 \\ 0.9 \end{pmatrix}}_{=:b}.$$

This system has no solution, since inexact measurements led to deviations from the true values $T(t)$. The normal equations in this example read

$$\begin{pmatrix} 5 & 1 \\ 1 & 23 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -2.1 \\ 11.1 \end{pmatrix}$$

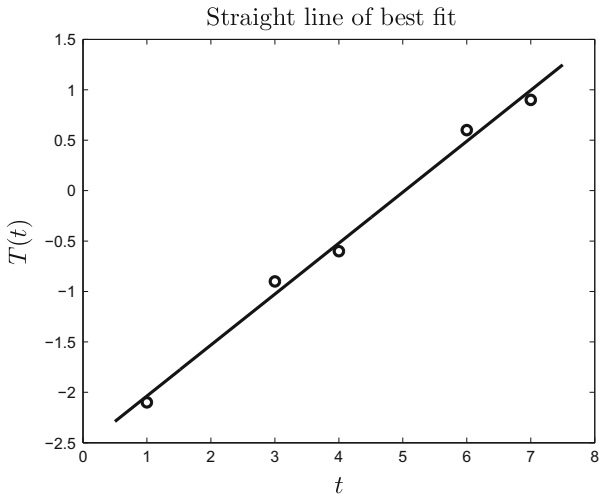

Fig. 3.2 Straight line of best fit

an have a unique solution $\alpha \approx -0.52105$ and $\beta \approx 0.50526$. Figure 3.2 shows the graph of the reconstructed function T, called **straight line of best fit** to the measurements. The measurement values are also marked in the figure. ◇

Variants of the minimization problem (3.2) can be formulated using norms other than $\|\cdot\|_2$. For example one could consider minimization of $\|r(x)\|_1 = |r_1(x)| + \ldots + |r_n(x)|$ or $\|r(x)\|_\infty = \max\{|r_1(x)|, \ldots, |r_n(x)|\}$. The solution of these problems is much more complicated, however, than the minimization of $\|r(x)\|_2$. Under certain conditions concerning the measurement errors, using the Euclidean norm can also be justified by stochastic arguments (the Gauß-Markov Theorem).

3.2 Sensitivity Analysis of Linear Least Squares Problems

Under the assumption that $A \in \mathbb{R}^{m,n}$ has full rank $n \leq m$, the linear least squares problem (3.2) has a unique solution $\hat{x} = (A^T A)^{-1} A^T b$. Since linear mappings between finite-dimensional spaces are always continuous, (3.2) is a well-posed problem according to Definition 1.5. We will now investigate, *how* sensitively the solution of (3.2) depends on A and b. As a technical tool, we will use the singular value decomposition of A, see (A.1) in Appendix A.

The matrix A and the right hand side b are input values of the linear least squares problem. The according result $\hat{x}$ is characterized by the inequality

$$\|b - A\hat{x}\|_2 \leq \|b - Ax\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{3.4}$$

If the input data are changed towards a matrix $A + \delta A$ and a right hand side $b + \delta b$, the minimizer changes too, to become $\tilde{x}$. It is characterized by the inequality

$$\|(b + \delta b) - (A + \delta A)\tilde{x}\|_2 \leq \|(b + \delta b) - (A + \delta A)x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{3.5}$$

The following two theorems give bounds of the absolute and relative difference of $\hat{x}$ and $\tilde{x}$ in dependence of absolute and relative differences of input values, respectively.

Theorem 3.3 (Absolute sensitivity of the linear least squares problem) *Let $A, \delta A \in \mathbb{R}^{m,n}$ and let $b, \delta b \in \mathbb{R}^m$, $m \geq n$. Let A have singular values $\sigma_1 \geq \ldots \geq \sigma_n > 0$ and thus have full rank. Let $\hat{x}$ be the unique solution of (3.2), characterized by (3.4), having residual $\hat{r} := b - A\hat{x}$. Under the condition*

$$\eta := \frac{\|\delta A\|_2}{\sigma_n} < 1, \quad \textit{i.e.} \quad \|\delta A\|_2 < \sigma_n,$$

there exists a unique $\tilde{x} \in \mathbb{R}^n$ with property (3.5). For $\delta x := \tilde{x} - \hat{x}$,

$$\|\delta x\|_2 \leq \frac{1}{\sigma_n(1-\eta)} \cdot (\|\delta b\|_2 + \|\delta A\|_2 \|\hat{x}\|_2) + \frac{1}{\sigma_n^2(1-\eta)^2} \cdot \|\delta A\|_2 \cdot \|\hat{r}\|_2 \tag{3.6}$$

holds.

Proof The matrix $A + \delta A$ has full rank, since

$$\begin{aligned}(A + \delta A)x = 0 \Rightarrow Ax = -\delta A x \Rightarrow \sigma_n \|x\|_2 \leq \|Ax\|_2 \leq \|\delta A\|_2 \|x\|_2 \\ \Rightarrow \underbrace{(\sigma_n - \|\delta A\|_2)}_{>0} \|x\|_2 \leq 0,\end{aligned}$$

which is only possible in case $x = 0$. According to Theorem 3.1, $\tilde{x}$ is uniquely determined.

Setting $\tilde{A} = A + \delta A$ and $\tilde{b} = b + \delta b$, one concludes from Theorem 3.1 that $\tilde{x} = M\tilde{A}^T\tilde{b}$, where $M := (\tilde{A}^T\tilde{A})^{-1}$ (normal equations). Consequently, we have

$$\begin{aligned}\delta x = \tilde{x} - \hat{x} &= M\tilde{A}^T\tilde{b} - \hat{x} = M\tilde{A}^T(\tilde{b} - \tilde{A}\hat{x}) \\ &= M\tilde{A}^T(b - A\hat{x}) + M\tilde{A}^T(\delta b - \delta A\hat{x}) \\ &= M(\delta A)^T\hat{r} + M\tilde{A}^T(\delta b - \delta A\hat{x}) \quad [(A + \delta A)^T\hat{r} = (\delta A)^T\hat{r} \text{ from (3.3)}]\end{aligned}$$

and from this we get

$$\|\delta x\|_2 \leq \|M\|_2 \|\delta A\|_2 \|\hat{r}\|_2 + \|M\tilde{A}^T\|_2 (\|\delta b\|_2 + \|\delta A\|_2 \|\hat{x}\|_2).$$

Everything is settled if $\|M\|_2 \leq 1/(\sigma_n^2(1-\eta)^2)$ and $\|M\tilde{A}^T\|_2 \leq 1/(\sigma_n(1-\eta))$. From an SVD $\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ we get

$$M = \tilde{V}(\tilde{\Sigma}^T\tilde{\Sigma})^{-1}\tilde{V}^T = \tilde{V}\mathrm{diag}\left(1/\tilde{\sigma}_1^2, \ldots, 1/\tilde{\sigma}_n^2\right)\tilde{V}^T$$

and from this we get $M\tilde{A}^T = \tilde{V}\tilde{\Sigma}^+\tilde{U}^T$, where $\tilde{\Sigma}^+ = \mathrm{diag}(1/\tilde{\sigma}_1, \ldots, 1/\tilde{\sigma}_n) \in \mathbb{R}^{n,m}$. We conclude $\|M\|_2 = 1/\tilde{\sigma}_n^2$ as well as $\|M\tilde{A}^T\|_2 = 1/\tilde{\sigma}_n$. From Theorem A.3 it follows $|\sigma_n - \tilde{\sigma}_n| \leq \|\delta A\|_2$, so that $\tilde{\sigma}_n \geq \sigma_n - \|\delta A\|_2 = \sigma_n(1-\eta)$ and thus $\|M\tilde{A}^T\|_2 \leq 1/(\sigma_n(1-\eta))$. An analoguous argument holds for $\|M\|_2$. □

The bound given for $\|\delta x\|_2$ in the above theorem is "nearly sharp": [Bjö96], p. 29, gives an example where it is approximately attained.

Theorem 3.4 (Relative sensitivity of the linear least squares problem) *For $m \geq n$, let $A, \delta A \in \mathbb{R}^{m,n}$ and $b, \delta b \in \mathbb{R}^m$. Let A have singular values $\sigma_1 \geq \ldots \geq \sigma_n > 0$ and let*

$$\kappa_2(A) := \frac{\sigma_1}{\sigma_n}\,. \tag{3.7}$$

Let $\hat{x} \neq 0$ be the solution of (3.2), characterized by (3.4) and having residual $\hat{r} := b - A\hat{x}$. Further assume that for some $\varepsilon > 0$

$$\|\delta A\|_2 \leq \varepsilon\|A\|_2, \quad \|\delta b\|_2 \leq \varepsilon\|b\|_2, \quad \textit{and} \quad \kappa_2(A)\varepsilon < 1 \tag{3.8}$$

holds. Then (3.5) uniquely characterizes $\tilde{x} \in \mathbb{R}^n$ and for $\delta x := \tilde{x} - \hat{x}$

$$\frac{\|\delta x\|_2}{\|\hat{x}\|_2} \leq \frac{\kappa_2(A)\varepsilon}{1-\kappa_2(A)\varepsilon}\left(2 + \left(\frac{\kappa_2(A)}{1-\kappa_2(A)\varepsilon} + 1\right)\frac{\|\hat{r}\|_2}{\|A\|_2\|\hat{x}\|_2}\right) \tag{3.9}$$

holds.

Proof From an SVD of A one gets $\sigma_1 = \|A\|_2$. Therefore, $\|\delta A\|_2 \leq \varepsilon\|A\|_2$ means that $\eta = \|\delta A\|_2/\sigma_n \leq \kappa_2(A)\varepsilon$. Under the condition $\kappa_2(A)\varepsilon < 1$ the condition $\eta < 1$ from Theorem 3.3 is fulfilled a fortiori and the estimate (3.6) remains valid, if η is replaced by $\kappa_2(A)\varepsilon$. Division by $\|\hat{x}\|_2$ shows

$$\frac{\|\delta x\|_2}{\|\hat{x}\|_2} \leq \frac{1}{\sigma_n(1-\kappa_2(A)\varepsilon)}\left(\frac{\|\delta b\|_2}{\|\hat{x}\|_2} + \|\delta A\|_2\right) + \frac{1}{(\sigma_n(1-\kappa_2(A)\varepsilon))^2}\cdot\frac{\|\delta A\|_2\|\hat{r}\|_2}{\|\hat{x}\|_2}\,.$$

Because of $\|\delta A\|_2 \leq \varepsilon\|A\|_2$, $\|\delta b\|_2 \leq \varepsilon\|b\|_2$, and $\|A\|_2/\sigma_n = \kappa_2(A)$, we get

$$\frac{\|\delta x\|_2}{\|\hat{x}\|_2} \leq \frac{\kappa_2(A)\varepsilon}{1-\kappa_2(A)\varepsilon}\left(\frac{\|b\|_2}{\|A\|_2\|\hat{x}\|_2} + 1\right) + \frac{\kappa_2(A)^2\varepsilon}{(1-\kappa_2(A)\varepsilon)^2}\cdot\frac{\|\hat{r}\|_2}{\|A\|_2\|\hat{x}\|_2}\,.$$

(3.9) finally follows from $\|b\|_2 \leq \|b - A\hat{x}\|_2 + \|A\hat{x}\|_2 \leq \|\hat{r}\|_2 + \|A\|_2\|\hat{x}\|_2$. □

Theorems 3.3 and 3.4 give a measure of how sensitively the solution of the least squares problem depends on perturbations of the input data A and b in the worst case. Evidently, $\delta x \to 0$ for $\delta A \to 0$ and $\delta b \to 0$, so we can state once more that according to Definition 1.5, the linear least squares problem formally is well-posed. In practice, however, we can not hope that ε tends to zero. It rather remains a finite, positive value, the size of which is determined by the limited precision of measurement devices. It is important to note that neither the input error size ε nor the output error size $\|\delta x\|_2/\|\hat{x}\|_2$ have anything to do with the finite precision of computer arithmetic. These errors would even show up if computations could be carried out exactly.

Both theorems contain statements about linear systems of equations as a special case: if A is nonsingular, the solution $\hat{x}$ of (3.1) is unique and coincides with the solution of (3.2). Theorems 3.3 and 3.4 can then be used with $\hat{r} = 0$. It turns out that the number $\kappa_2(A)\varepsilon$ is decisive for the worst case error in the solution of $Ax = b$. This number clearly has to be much smaller than 1 to guarantee a meaningful solution. If $\varepsilon \ll \kappa_2(A)$, one has $\kappa_2(A)/(1 - \kappa_2(A)\varepsilon) \approx \kappa_2(A)$. Then, $\kappa_2(A)$ can be interpreted as a factor, by which relative input errors of size ε may be amplified in the result. For linear least squares problems, an additional (approximate) amplification factor $\kappa_2(A)^2\|\hat{r}\|_2$ appears on the right hand side of (3.9), meaning that in this case the worst case error also depends on b. If b is close to $\mathscr{R}_A$, i.e. if the system $Ax = b$ is nearly consistent ("nearly solvable"), then $\|\hat{r}\|_2$ is small. Otherwise, the solution of a least squares problem may depend more sensitively on A and b than does the solution of a linear system of equations.

An inverse problem of form (3.1) or (3.2) is called **well conditioned**, if small perturbations in the data A and b do only lead to small perturbations in the solution. It is called **badly conditioned**, if small perturbations in the data A and b may lead to large perturbations in the solution. The **condition number** of the inverse problem is a factor by which relative perturbations in the input data A or b might be amplified in the result *in the worst case*. The condition number of (3.1) is approximately given by $\kappa_2(A)$, which is called **condition number of matrix** A. The condition number of (3.2) is approximately given by the maximum of $\kappa_2(A)$ and $\kappa_2(A)^2\|\hat{r}\|_2/(\|A\|_2\|\hat{x}\|_2)$.

Example 3.5 (Condition numbers for the collocation method) Reconsider Example 2.11 from Sect. 2.3. The linearized model problem of seismic tomography was discretized by the collocation method. This led to an overdetermined system $Ax = b$, which was replaced by the corresponding least squares problem. In this case we suppose the system $Ax = b$ to be nearly consistent. It is to be expected, therefore, that $\kappa_2(A)\varepsilon$ is a bound for the relative error in the least squares solution (disregarding computing errors due to finite precision computer arithmetic). In Fig. 3.3, the condition number $\kappa_2(A)$ is shown (on a logarithmic scale) for matrices

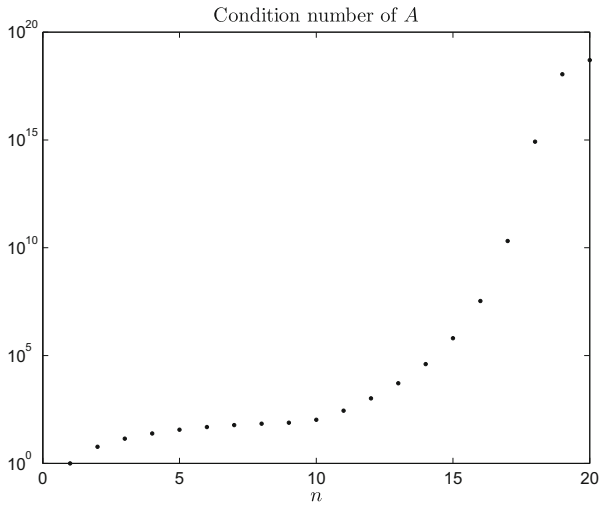

Fig. 3.3 Condition numbers $\kappa_2(A)$ as a function of n

A resulting from various discretization levels n. Visibly, $\kappa_2(A)$ is growing quickly with n. For a computer with "double precision arithmetic",[1] an input error size $\varepsilon \geq 10^{-16}$ has to be assumed, which would result from the very conversion of measurement data to machine numbers – even in the purely hypothetical case of exact measurements. According to Fig. 3.3, a discretization level $n > 18$ may then already lead to a meaningless result. In Example 2.11, we actually had $\varepsilon \approx 10^{-5}$ (compare definition of w^δ and w). Thus, even a condition number $\kappa_2(A) > 10^5$ will have a desastrous effect and a discretization level $n \geq 15$ should therefore be avoided. ◇

Matrices with Deficient Rank

In case of a rank deficient matrix A, the smallest singular value vanishes: $\sigma_n = 0$. One may then formally set

$$\kappa_2(A) = \sigma_1/0 = \infty$$

and call problems (3.1) or (3.2) "infinitely badly conditioned". This is in accordance with the fact that no unique solution exists anymore. Uniqueness can be re-

[1] Such computers use 53 bits to represent the mantissae of floating point numbers. Then, the rounding error committed when converting a real to a machine number is bounded by $2^{-53} \approx 10^{-16}$.

established if an *additional requirement* is added.[2] In case of (3.2) we might ask for the solution with smallest Euclidean norm. This means that (3.2) is changed to become a *new problem*, namely:

$$\text{Find } \hat{x} \text{ such that } \|\hat{x}\|_2 \leq \|x\|_2 \text{ for all } x \in M := \arg\min\{\|b - Ax\|_2\}, \tag{3.10}$$

where $\arg\min\{\|b - Ax\|_2\}$ is the set of all vectors solving (3.2). A solution of (3.10) is called **minimum norm solution** of the least squares problem (3.2). In case $\text{rank}(A) = n$ the unique solution $\hat{x}$ of (3.2) necessarily also is the unique solution of (3.10). The following theorem shows that (3.10) has a unique solution for any matrix A. It also shows that (3.10) may be better conditioned than (3.2).

Theorem 3.6 (Linear least squares problem for general matrices) *Let $A \in \mathbb{R}^{m,n}$ have rank $r \leq \min\{m, n\}$ and have an SVD*

$$A = U\Sigma V^T = \underbrace{(}_{}\underbrace{U_1}_{r}\ \underbrace{U_2}_{m-r}\underbrace{)}_{} \underbrace{\begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}}_{\begin{matrix} r \\ m-r \end{matrix}} \underbrace{\begin{matrix} \\ \end{matrix}}_{\begin{matrix} r & n-r \end{matrix}} \underbrace{(}_{}\underbrace{V_1}_{r}\ \underbrace{V_2}_{n-r}\underbrace{)^T}_{}$$

$$= U_1 \Sigma_1 V_1^T .$$

(a) All solutions of (3.2) have a representation of the form

$$x = V_1 \Sigma_1^{-1} U_1^T b + V_2 z, \quad z \in \mathbb{R}^{n-r}. \tag{3.11}$$

(b) Among all solutions of (3.2) there is a unique one having minimal Euclidean norm, i.e. there is a unique solution of (3.10). This solution is given by

$$x = V_1 \Sigma_1^{-1} U_1^T b .$$

Its norm is bounded by $\|x\|_2 \leq \|b\|_2/\sigma_r$.

(c) If the right hand side b is perturbed to become $b + \delta b$, the according (unique) solution of (3.10) becomes $x + \delta x$ and

$$\|\delta x\|_2 \leq \frac{\|\delta b\|_2}{\sigma_r}$$

holds.

[2] The same kind of idea was used to force the inverse gravimetry problem into having a unique solution.

Proof Part (a):

$$\|b - Ax\|_2^2 = \left\| U^T b - \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} U_1 \Sigma_1 V_1^T x \right\|_2^2 = \left\| \begin{pmatrix} U_1^T b - \Sigma_1 V_1^T x \\ U_2^T b \end{pmatrix} \right\|_2^2$$
$$= \|U_1^T b - \Sigma_1 V_1^T x\|_2^2 + \|U_2^T b\|_2^2$$

becomes minimal if and only if

$$\Sigma_1 V_1^T x = U_1^T b \quad \Longleftrightarrow \quad x = V_1 \Sigma_1^{-1} U_1^T b + V_2 z$$

for arbitrary $z \in \mathbb{R}^{n-r}$, since

$$\mathcal{N}_{V_1^T} = \mathcal{R}_{V_1}^{\perp} = \{V_2 z;\ z \in \mathbb{R}^{n-r}\}.$$

Part (b): Since V_1 and V_2 are orthogonal, from (3.11) and Pythagoras' Theorem we get

$$\|x\|_2^2 = \|V_1 \Sigma^{-1} U_1^T b\|_2^2 + \|V_2 z\|_2^2.$$

The right hand side becomes minimal if and only if $V_2 z = 0$, i.e. if and only if $z = 0$. For $z = 0$ we get

$$\|x\|_2^2 = \|V_1 \Sigma_1^{-1} U_1^T b\|_2^2 = \left\| \begin{pmatrix} u_1^T b/\sigma_1 \\ \vdots \\ u_r^T b/\sigma_r \end{pmatrix} \right\|_2^2 \le \frac{1}{\sigma_r^2} \sum_{j=1}^{r} |u_j^T b|^2 \le \frac{\|b\|_2^2}{\sigma_r^2}.$$

Part (c): Replacing b by $b + \delta b$ in part (b) we get the minimum norm solution $x + \delta x = V_1 \Sigma_1^{-1} U_1^T (b + \delta b)$. Thus $\delta x = V_1 \Sigma_1^{-1} U_1^T \delta b$ and we can directly use the last estimate from part (b) to end the proof. □

The sensitivity of the minimum norm solution with respect to perturbations of b is determined by the size of the smallest non vanishing singular value σ_r of A. More generally, one can also consider perturbations of the elements of A. Bounds for $\|\delta x\|_2$ do exist also in this general case, which are analoguous to the one given in Theorem 3.3 with σ_n replaced by σ_r, see, e.g., Theorem 1.4.6 in [Bjö96].

Definition 3.7 (Pseudoinverse) Using the notation of Theorem 3.6 define

$$A^+ := V_1 \Sigma_1^{-1} U_1^T.$$

This matrix is called the **pseudoinverse** of A.

According to Theorem 3.6, the unique minimum norm solution of (3.2) can formally be written as

$$x = A^+ b.$$

In case $\operatorname{rank}(A) = n$, a unique solution of (3.2) exists and $A^+ = (A^T A)^{-1} A^T$. In case $\operatorname{rank}(A) = n$ *and* $m = n$, a unique solution of (3.1) exists and $A^+ = A^{-1}$.

Example 3.8 Let $\varepsilon > 0$ and

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \Longrightarrow \quad A^+ = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}, \quad A^+ b = \begin{pmatrix} b_1 \\ b_2/\varepsilon \end{pmatrix}.$$

When (3.2) or (3.10) are solved for these data, errors in component b_2 of the right hand side are amplified by a factor ε^{-1}. Since $\kappa_2(A) = \varepsilon^{-1}$, this is in accordance with Theorem 3.4. The condition of problem (3.2) becomes arbitrarily bad with $\varepsilon \to 0$. Consider, on the other hand

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{=U} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{=\Sigma} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T}_{=V^T} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T \quad \Longrightarrow \quad B^+ = B.$$

Matrix B has rank 1 with smallest non vanishing singular value $\sigma_1 = 1$. The solution $B^+ b = (b_1, 0)^T$ of problem (3.10) is perfectly well conditioned with respect to the data b. ◇

Theorem 3.6 and the above example show that the presence of small singular values can be more annoying than a deficient rank.

Generalized Least Squares Problems

Even if the matrix A is rank deficient, the minimum norm problem (3.10) still has a unique solution. In this paragraph, a more general additional requirement is added to the least squares problem than the minimum norm criterion. This will be of use in the context of regularization. To start with, a characterization of the pseudoinverse will be given, which differs from Definition 3.7.

Let $S \subset \mathbb{R}^n$ be a linear subspace of $\mathbb{R}^n$ with orthonormal basis $\{s_1, \ldots, s_k\}$. Define the matrices

$$C := (s_1 | \cdots | s_k) \in \mathbb{R}^{n,k}, \quad P_S := CC^T, \quad \text{and} \quad P_{S^\perp} := I_n - CC^T. \tag{3.12}$$

It can easily be seen that P_S defines a mapping $\mathbb{R}^n \to S$, $y \mapsto P_S y$, where $\|y - P_S y\|_2 \leq \|y - x\|_2$ for all vectors $x \in S$. For this reason, P_S is called an

orthogonal projector onto the subspace S of $\mathbb{R}^n$. Likewise, $P_{S^\perp}$ is an orthogonal projector onto the orthogonal complement $S^\perp$ of S. Any vector $x \in \mathbb{R}^n$ has a unique decomposition $x = u + v$ with $u = P_S x \in S$ and $v = P_{S^\perp} x \in S^\perp$. From Theorem 3.6 and Definition 3.7 it can readily be seen that

$$\begin{aligned} P_{\mathscr{R}_A} &= AA^+ = U_1 U_1^T,\ P_{\mathscr{R}_A^\perp} = I_m - AA^+ = U_2 U_2^T \\ P_{\mathscr{N}_A^\perp} &= A^+A = V_1 V_1^T,\ P_{\mathscr{N}_A} = I_n - A^+A = V_2 V_2^T. \end{aligned} \tag{3.13}$$

Using these relations, the pseudoinverse can be interpreted as follows.

Lemma 3.9 *Under the conditions of Theorem 3.6, one has*

$$A^+ = B \circ P_{\mathscr{R}_A},$$

where B is the inverse of the bijective mapping

$$\tilde{A} = A_{|\mathscr{N}_A^\perp} : \mathscr{N}_A^\perp \to \mathscr{R}_A.$$

Further on, the pseudoinverse A^+ is the unique matrix $X \in \mathbb{R}^{n,m}$ having the following four properties

$$\begin{aligned} AXA &= A,\ (AX)^T = AX, \\ XAX &= X,\ (XA)^T = XA, \end{aligned} \tag{3.14}$$

*which are called **Moore-Penrose axioms**.*

Proof Reusing the notation from Theorem 3.6, any $y \in \mathbb{R}^m$ can uniquely be written in the form $y = U_1 z_1 + U_2 z_2$, where $z_1 \in \mathbb{R}^r$ and $z_2 \in \mathbb{R}^{m-r}$. One therefore gets $A^+ y = V_1 \Sigma_1^{-1} z_1$. On the other hand, one also has

$$B \circ P_{\mathscr{R}_A} y = A^{-1}(U_1 z_1) = V_1 \Sigma_1^{-1} z_1,$$

since $V_1 \Sigma_1^{-1} z_1 \in \mathscr{N}_A^\perp$ and $A(V_1 \Sigma_1^{-1} z_1) = U_1 \Sigma_1 V_1^T V_1 \Sigma_1^{-1} z_1 = U_1 z_1$. This proves the first part of the lemma. Properties (3.14) follow immediately from Theorem 3.6 and Definition 3.7 for $X = A^+$. Knowing that $X = A^+$ satisfies (3.14), assume that X_1 and X_2 both satisfy (3.14). Then,

$$AX_1 = (AX_1)^T = X_1^T A^T = X_1^T (AX_2 A)^T = (AX_1)^T (AX_2)^T = AX_1 AX_2 = AX_2.$$

In the same way, $X_1 A = X_2 A$. Therefore

$$X_1 = X_1 AX_1 = X_1 AX_2 = X_2 AX_2 = X_2,$$

meaning that *only* $X = A^+$ can satisfy (3.14): the Moore-Penrose axioms *determine* A^+. □

Theorem 3.10 (Generalized linear least squares problem) *Let all conditions of Theorem 3.6 hold and let $L \in \mathbb{R}^{p,n}$. There exists a unique solution of*

$$\min_{x \in M} \|Lx\|_2, \qquad M := \arg\min\{\|b - Ax\|_2\} \tag{3.15}$$

if and only if

$$\mathcal{N}_A \cap \mathcal{N}_L = \{0\}. \tag{3.16}$$

If (3.16) holds, this minimizer is given by

$$x = (I_n - (LP_{\mathcal{N}_A})^+ L)A^+ b. \tag{3.17}$$

Proof The necessity of (3.16) for the uniqueness of a minimizer is clear, since for every $z \in \mathcal{N}_A \cap \mathcal{N}_L$ and for all $x \in \mathbb{R}^n$ one has $L(x+z) = Lx$ and $A(x+z)-b = Ax-b$. So let now (3.16) hold. From Theorem 3.6 it is known that all elements of M have the form

$$x = A^+ b + P_{\mathcal{N}_A} z, \quad z \in \mathbb{R}^n,$$

so the minimization problem (3.15) can be written in the form

$$\min_{z \in \mathbb{R}^n} \|LP_{\mathcal{N}_A} z + LA^+ b\|_2$$

which is a standard linear least squares problem. Again from Theorem 3.6 one concludes that the general solution of *this* problem has the form

$$z = -(LP_{\mathcal{N}_A})^+ LA^+ b + (I_n - (LP_{\mathcal{N}_A})^+ LP_{\mathcal{N}_A})y, \quad y \in \mathbb{R}^n, \tag{3.18}$$

where (3.13) was used to express the projection onto the nullspace of $LP_{\mathcal{N}_A}$. Thus the general solution of (3.15) has the form

$$x = A^+ b + P_{\mathcal{N}_A} z, \quad z \quad \text{from (3.18)}. \tag{3.19}$$

Observe now that

$$LP_{\mathcal{N}_A} x = 0 \iff \underbrace{P_{\mathcal{N}_A} x}_{\in \mathcal{N}_A} \in \mathcal{N}_L \overset{\mathcal{N}_A \cap \mathcal{N}_L = \{0\}}{\iff} P_{\mathcal{N}_A} x = 0 \iff x \in \mathcal{N}_A^{\perp}, \tag{3.20}$$

such that, by (3.13), $(I_n - (LP_{\mathcal{N}_A})^+ LP_{\mathcal{N}_A})y \in \mathcal{N}_A^{\perp}$ for all $y \in \mathbb{R}^n$. This shows that

$$P_{\mathcal{N}_A}(I_n - (LP_{\mathcal{N}_A})^+ LP_{\mathcal{N}_A})y = 0$$

for all $y \in \mathbb{R}^n$. Using this and (3.18) in (3.19), we have proven that there is a unique solution

$$x = A^+b - P_{\mathcal{N}_A}(LP_{\mathcal{N}_A})^+LA^+b = (I_n - P_{\mathcal{N}_A}(LP_{\mathcal{N}_A})^+L)A^+b$$

of (3.15). From this, (3.17) follows, if $P_{\mathcal{N}_A}(LP_{\mathcal{N}_A})^+ = (LP_{\mathcal{N}_A})^+$. The latter identity can be verified as follows

$$\mathcal{R}_{(LP_{\mathcal{N}_A})^+} \overset{\text{Lemma (3.9)}}{=} \mathcal{N}_{LP_{\mathcal{N}_A}}^{\perp} \overset{(3.20)}{=} \left(\mathcal{N}_A^{\perp}\right)^{\perp} = \mathcal{N}_A$$

and this ends the proof. □

The above theorem gives rise to the following definition

Definition 3.11 (Weighted pseudoinverse) Let all conditions of Theorem 3.6 hold, let $L \in \mathbb{R}^{p,n}$ and let (3.16) hold. Then the matrix

$$A_L^+ := (I_n - (LP_{\mathcal{N}_A})^+L)A^+ \in \mathbb{R}^{n,m}$$

is called L**-weighted pseudoinverse** of A. Analogously, the matrix

$$L_A^+ := (I_n - (AP_{\mathcal{N}_L})^+A)L^+ \in \mathbb{R}^{n,p}$$

is called A**-weighted pseudoinverse** of L.

Whereas A^+ was introduced as a mapping relating $b \in \mathbb{R}^m$ to the unique least squares solution of $Ax = b$ having minimal norm, A_L^+ is defined as the mapping relating b to the unique least squares solution of $Ax = b$ which minimizes the norm of Lx. But the L-weighted pseudoinverse of A has another interpretation similar to the one given for A^+ in Lemma 3.9. To see this, introduce the scalar product on $\mathbb{R}^n$ defined by

$$\langle x|\bar{x}\rangle_* := \langle Ax|A\bar{x}\rangle + \langle Lx|L\bar{x}\rangle, \quad x, \bar{x} \in \mathbb{R}^n \tag{3.21}$$

with standard Euclidean scalar products $\langle \bullet|\bullet\rangle$ on the right hand side. Note that (3.16) is needed for positive definiteness to hold. Let us write

$$x \perp_* y = 0 \quad :\Longleftrightarrow \quad \langle x|y\rangle_* = 0.$$

The orthogonal complement of $\mathcal{N}_A$ with respect to $\langle \bullet|\bullet\rangle_*$ is given by

$$\mathcal{N}_A^{\perp *} = \{x \in \mathbb{R}^n;\ L^TLx \perp \mathcal{N}_A\}, \tag{3.22}$$

which can be verified easily:

$$\begin{aligned} x \in \mathscr{N}_A^{\perp *} &\iff \langle x|\bar{x}\rangle_* = \langle Ax|A\bar{x}\rangle + \langle Lx|L\bar{x}\rangle = 0 \text{ for all } \bar{x} \in \mathscr{N}_A \\ &\iff \langle Lx|L\bar{x}\rangle = \langle L^T Lx|\bar{x}\rangle = 0 \text{ for all } \bar{x} \in \mathscr{N}_A. \end{aligned}$$

Lemma 3.12 *Under the conditions of Definition 3.11, one has*

$$A_L^+ = B_* \circ P_{\mathscr{R}_A},$$

where B_* is the inverse of the bijective mapping

$$A_* = A_{|\mathscr{N}_A^{\perp *}} : \mathscr{N}_A^{\perp *} \to \mathscr{R}_A.$$

Proof It has to be shown that for every $b \in \mathbb{R}^m$, $B_*(P_{\mathscr{R}_A} b)$ is a least squares solution of $Ax = b$ and minimizes $\|Lx\|_2$ among all least squares solutions. Let $x_0 := B_*(P_{\mathscr{R}_A} b)$ and let $x_1 := A^+ b$. Then

$$A(x_0 - x_1) = AA^{-1}(P_{\mathscr{R}_A} b) - AA^{-1}(P_{\mathscr{R}_A} b) = 0,$$

showing that $x_0 - x_1 \in \mathscr{N}_A$. This shows that x_0 is a least squares solution, as is x_1. Next let x_2 be any least squares solution. Then

$$\begin{aligned} \|Lx_2\|_2^2 &= \|Lx_0 + L(x_2 - x_0)\|_2^2 = \|Lx_0\|_2^2 + \|L(x_2 - x_0)\|_2^2 + 2(x_2 - x_0)^T L^T Lx_0 \\ &= \|Lx_0\|_2^2 + \|L(x_2 - x_0)\|_2^2, \end{aligned}$$

since $x_2 - x_0 \in \mathscr{N}_A$, since $x_0 = B_*(P_{\mathscr{R}_A} b) \in \mathscr{N}_A^{\perp *}$, and since (3.22) holds. The term $\|L(x_2 - x_0)\|_2^2$ vanishes if and only if $x_2 - x_0 \in \mathscr{N}_A \cap \mathscr{N}_L \overset{(3.16)}{=} \{0\}$. Thus x_0 is the only least squares solution minimizing the norm of Lx. □

3.3 The Concept of Regularization

To explain what is meant by "regularization", we first turn back to parameter identification problems in their general form. Later, specific regularization methods will be presented for discretized (finite-dimensional) problems, but the origin of these must be kept in mind to find useful regularizations.

Parameter identification in the linear case means to solve an equation of the form $Tu = w$, where $T : X \to \mathbb{W} \subseteq Y$ is linear and bijective and $(X, \| \bullet \|_X)$ and $(Y, \| \bullet \|_Y)$ are normed spaces. A unique solution u^* of this problem is assumed to exist, but usually two difficulties impede its practical computation.

1. Only an approximation $w^\delta \in Y$ of w is known. At best one can bound the data error $w - w^\delta$ by $\|w - w^\delta\|_Y \le \delta$ for some $\delta > 0$, but one can generally *not* assume $w^\delta \in \mathbb{W}$.
2. The inverse $T^{-1} : \mathbb{W} \to X$ possibly is not continuous. Then, even if one knew a good approximation $\tilde{w} \in \mathbb{W}$ of w, $T^{-1}(\tilde{w})$ could be far away from u^*.

Regularization is an attempt to bypass both difficulties in order to find a good approximation of u^*. Before giving the usual, abstract definition of regularization, we tentatively define it to be a set $\{R_t;\ t \in I\}$ of *continuous* operators $R_t : Y \to X$, such that

$$\|R_t(w) - T^{-1}(w)\|_X \xrightarrow{t \to 0} 0 \quad \text{for all} \quad w \in \mathbb{W}. \tag{3.23}$$

The index set I may be any (possibly uncountable) set $I \subseteq \mathbb{R}_0^+$ having 0 as a limit point, so that (3.23) makes sense. If all operators R_t are linear, the regularization is called linear. Note two essential points:

1. The operators R_t are defined on the whole space Y and thus can be applied to any $w^\delta \in Y$, even if $w^\delta \notin \mathbb{W}$.
2. The operators R_t are continuous approximants of the possibly discontinuous operator $T^{-1} : \mathbb{W} \to X$, converging pointwise to T^{-1} on $\mathbb{W}$.

The next example shows that discretization already means regularization (in the sense of (3.23)).

Example 3.13 (Regularization by the least squares discretization method) This method was introduced in Sect. 2.2. We chose a subspace $X_n = \langle \varphi_1, \dots, \varphi_{d_n} \rangle \subset X$ and defined an approximation $u_n = \sum_{j=1}^{d_n} x_j \varphi_j$ of the true solution u^*. The approximation was formalized by operators

$$R_n : Y \to X_n \subseteq X, \quad y \mapsto u_n,$$

defined in (2.32), see Theorem 2.8. These operators are continuous for all $n \in \mathbb{N}$. Let us rename R_n into $R_{1/n}$, $n \in \mathbb{N}$. Theorem 2.9 states that under conditions (2.40) and (2.41) convergence $\|R_{1/n}(w) - T^{-1}(w)\|_X \to 0$ holds for $n \to \infty$ and for all $w \in T(X) = \mathbb{W}$. Thus the operator set $\{R_t;\ t \in I\}$ with $I = \{1/n;\ n \in \mathbb{N}\}$ is a regularization in the sense of (3.23). ◇

The above tentative definition is not sufficient, since perturbed data are not considered. Assume some $w^\delta \in Y$ (possibly not contained in $\mathbb{W}$) was given and assume knowledge of $\delta > 0$ such that $\|w^\delta - w\|_Y \le \delta$. If the regularization is linear, then one can estimate, making use of $Tu^* = w \in \mathbb{W}$,

$$\begin{aligned} \|R_t w^\delta - u^*\|_X &\le \|R_t(w^\delta - w)\|_X + \|R_t w - u^*\|_X \\ &\le \|R_t\|\delta + \|R_t w - u^*\|_X, \end{aligned} \tag{3.24}$$

which corresponds to estimate (2.34) from Theorem 2.8. The total reconstruction error is bounded by the sum of two terms, with the second term $\|R_t w - u^*\|_X = \|R_t w - T^{-1}(w)\|_X$ tending to zero for $t \to 0$, if (3.23) holds. But at the same time the operator norm $\|R_t\|$ grows beyond all bounds for $t \to 0$, whenever T^{-1} is not continuous.[3] This was shown in the special case of regularization by discretization (by the linear least squares method) in (2.44) and (2.45). For a finite value $\delta > 0$ and for given data w^δ, one would like to choose an optimal parameter t^* such that $\|R_{t^*}(w^\delta) - u^*\|_X$ becomes as small as possible. In the special case of Example 3.13 (regularization by discretization) this would mean to select an optimal dicretization level, which is not possible in practice, since u^* is not known. For general operators R_t, an optimal choice $t = t^*$ is even less possible. Anyway some rule $\rho(\delta, w^\delta) = t$ is required in order to choose a parameter t in dependence of given data w^δ and given data error magnitude δ. Since an optimal parameter selection is out of reach, one more modestly demands the total error $\|R_{\rho(\delta,w^\delta)}(w^\delta) - T^{-1}(w)\|_X$ to tend to zero for all $w \in \mathbb{W}$, if $\delta \to 0$. This convergence shall be uniform for all w^δ in the neighbourhood of w. These considerations lead to the following definition.

Definition 3.14 (Regularization) Let $(X, \| \bullet \|_X)$ and $(Y, \| \bullet \|_Y)$ be normed linear spaces and let $T : X \to \mathbb{W} \subseteq Y$ be linear, continuous, and bijective. Let $I \subset \mathbb{R}_0^+$ be an index set having 0 as a limit point and let $\{R_t;\ t \in I\}$ be a set of continuous operators $R_t : Y \to X$. If a mapping $\rho : \mathbb{R}^+ \times Y \to \mathbb{R}_0^+$ exists such that for all $w \in \mathbb{W}$

$$\sup\{\rho(\delta, w^\delta);\ w^\delta \in Y, \|w - w^\delta\|_Y \le \delta\} \xrightarrow{\delta \to 0} 0 \tag{3.25}$$

and

$$\sup\{\|R_{\rho(\delta,w^\delta)}(w^\delta) - T^{-1}(w)\|_X;\ w^\delta \in Y,\ \|w - w^\delta\|_Y \le \delta\} \xrightarrow{\delta \to 0} 0, \tag{3.26}$$

then the pair $(\{R_t;\ t \in I\}, \rho)$ is called **regularization** of T^{-1}. If all R_t are linear, the regularization is called linear. Each number $\rho(\delta, w^\delta)$ is called a **regularization parameter**. The mapping ρ is called **parameter choice**. If a parameter choice does only depend on δ but not on the data w^δ, then it is called a **parameter choice a priori**. Otherwise it is called a **parameter choice a posteriori**.

From (3.26) it follows that

$$\|R_{\rho(\delta,w)}(w) - T^{-1}(w)\|_X \xrightarrow{\delta \to 0} 0 \quad \text{for all} \quad w \in \mathbb{W},$$

[3] Assume, on the contrary, that $\|R_t\| \le C$ for some constant C and for $t \to 0$. Then, $\|R_t y\|_Y \le C\|y\|_Y$ for all $y \in T(X)$ and $t \to 0$. Since $R_t y \to T^{-1}y$, it follows that $\|T^{-1}y\|_Y \le C\|y\|_Y$ for all $y \in T(X)$, a contradiction to the unboundedness of T^{-1}.

meaning that the above definition generalizes the previous, tentative one. From (3.24) it can be seen that for linear regularizations with (3.23) an a priori parameter choice $\rho = \rho(\delta)$ is admissible if

$$\rho(\delta) \overset{\delta\to 0}{\longrightarrow} 0 \quad \text{and} \quad \|R_{\rho(\delta)}\|\delta \overset{\delta\to 0}{\longrightarrow} 0. \tag{3.27}$$

Example 3.15 (Regularization by the least squares discretization method) Let us look again at discretization by the least squares method. With the renaming introduced in Example 3.15, the continuous discretization operators form a set $\{R_t;\ t \in I\}$, $I = \{1/n;\ n \in \mathbb{N}\}$. From Theorem 2.9 it follows that under conditions (2.40) and (2.41) the convergence property $\|R_t w - T^{-1}(w)\|_X \to 0$ holds for $t \to 0$. From Theorem 2.8, estimate (2.34), it can be seen that a parameter choice satisfying (3.27) makes discretization by the least squares method a regularization in the sense of Definition 3.14 (if all conditions (2.40), (2.41), and (3.27) hold). ◇

Discretization of a parameter identification problem $Tu = w$ is a kind of a regularization. This could be verified in the above examples for the least squares method, but it is also true for the collocation method. In both cases, discretization leads to a system of equations $Ax = b$ or, more generally, to a least squares problem

$$\text{minimize} \quad \|b - Ax\|_2 \quad \text{for} \quad A \in \mathbb{R}^{m,n} \text{ and } b \in \mathbb{R}^m, \tag{3.28}$$

where in practice only an approximation b^δ of b is known. But (3.28) is a parameter identification problem by itself. In case $\text{rank}(A) = n$, it has a unique solution $x^* = A^+ b$, the pseudoinverse A^+ now playing the role of T^{-1}. The mapping $b \mapsto A^+ b$ formally is continuous (as a linear mapping between finite-dimensional spaces), but A may be badly conditioned, so that $A^+ b^\delta$ may be far away from x^*, even if b^δ is close to b. As confirmed by the examples of Chap. 2, a bad condition of A results from choosing too fine a discretization – which is likely to happen, since we have no way of choosing an optimal one. So there may well be a need to regularize (3.28) – not in the sense of approximating a discontinuous operator by a continuous one (as in Definition 3.14), but in the sense of approximating a badly conditioned operator (matrix) by a better conditioned one. This can not be achieved by discretization again. Also, there is no point in just replacing A^+ by a better conditioned operator (matrix) M, if Mb^δ is not close to x^*. We can only hope to to make Mb^δ be close to x^*, *if we have some additional information about* x^* beyond the knowledge $x^* = A^+ b$, which can not be exploited, since b is not known.

"Additional information" can take various forms and we do not have a formalism to capture them all. We assume the following.

Assumption 3.16 (Additional information) *Let Assumption 2.6 hold and let* u^* *be the solution of the corresponding inverse problem* $Tu = w$. *Let* $X_0 \subseteq X$ *be a subspace of* X *and let* $\|\bullet\|_0$ *be a semi-norm on* X_0. *Assume that the "additional*

information"

$$u^* \in X_0 \quad \textit{and} \quad \|u^*\|_0 \leq S, \tag{3.29}$$

is available about u^*.

Remark Often, it is required that $\|\bullet\|_0$ is a *norm*, not only a semi-norm (see Appendix B for the difference between both), and that, moreover, $\|\bullet\|_0$ is stronger than the norm $\|\bullet\|_X$ on X, i.e. that

$$\|u\|_X \leq C\|u\|_0 \quad \text{for all } u \in X_0 \tag{3.30}$$

holds, compare Definition 1.18 in [Kir96]. These stronger requirements on $\|\bullet\|_0$ are needed to develop a theory of "worst case errors" and "regularizations of optimal order", i.e. regularizations for which the convergence (3.26) is of highest possible order, compare Lemma 1.19, Theorem 1.21, and Theorems 2.9 and 2.12 in [Kir96]. We will not go into these abstract concepts, because we are rather interested in achieving a small error

$$\|R_{\rho(\delta,w^\delta)}(w^\delta) - T^{-1}(w)\|_X$$

for a finite value $\delta > 0$ than achieving optimal convergence order for $\delta \to 0$. Also, "optimal order of convergence" is not a helpful concept in the finite-dimensional setting, where the inversion operator A^+ is continuous, see the highlighted comments after Theorem 3.26. Requiring $\|\bullet\|_0$ to be only a semi-norm will prove practical in the examples of the following section.

Example 3.17 (Fredholm integral equation) Assume we want to solve a linear Fredholm integral equation of the first kind. This problem is defined by a mapping

$$T : L_2(a,b) \to L_2(a,b), \quad u \mapsto w, \quad w(t) = \int_a^b k(s,t)u(s)\,ds.$$

Here, $X = L_2(a,b)$ with norm $\|\bullet\|_X = \|\bullet\|_{L_2(a,b)}$. Let $X_0 := H^1(a,b) \subset L_2(a,b)$ with norm $\|\bullet\|_0 = \|\bullet\|_{H^1(a,b)}$ (in this case, even (3.30) holds). Functions contained in X_0 are distinguished from L_2-functions by some guaranteed extra smoothness. Additional information $u^* \in X_0$ means that we have an a priori knowledge about the smoothness of u^*. ◇

If additional information about u^* as in (3.29) is known, it is natural to require (3.29) to hold also for a discretized approximant. This will be described in the following Problem 3.18, which is not yet a concrete problem, but rather a guideline to set up one.

Problem 3.18 (Regularized linear least squares problem) Let Assumption 3.16 hold, especially let $Tu = w$ be an identification problem with unique exact solution u^*, about which we have the additional information

$$u^* \in X_0 \quad \text{and} \quad \|u^*\|_0 \le S$$

as in (3.29). As with standard discretization, we choose a n-dimensional subspace $X_n \subset X$, but now under the additional requirement

$$\varphi_j \in X_0, \quad j = 1, \ldots, n, \tag{3.31}$$

for the basis vectors φ_j of X_n. Approximants $u_n \in X_n \subset X_0$ of u^* are rated in the form

$$u_n = \sum_{j=1}^{n} x_j \varphi_j, \quad x_j \in \mathbb{R},\ j = 1, \ldots, n,$$

as in Sects. 2.2 and 2.3 (for simplicity, we let $d_n = n$ here). The coefficients x_j are to be found by solving a system $Ax = b$ defined in (2.25) for the least squares method and in (2.51) for the collocation method. It will be assumed that $\operatorname{rank}(A) = n$. Let only perturbed data $b^\delta \in \mathbb{R}^m$ be given such that $\|b - b^\delta\|_2 \le \delta$ for some $\delta > 0$. Based on these perturbed data compute an approximation $u_n^\delta = \sum_{j=1}^n x_j^\delta \varphi_j$ of u^*, with coefficients x_j^δ defined as a solution of the problem

$$\text{minimize} \quad \|b^\delta - Ax\|_2 \quad \text{under the constraint} \quad \left\| \sum_{j=1}^{n} x_j \varphi_j \right\|_0 \le S. \tag{3.32}$$

An investigation of why (3.32) is indeed a regularization of (3.28) will be postponed to Sect. 3.4. The following example shows, how adding information works in practice.

Example 3.19 (Numerical differentiation) Let $w \in C^1[a,b]$ with $w(a) = 0$. The derivative $u^* = w'$ of w is the solution of the Volterra integral equation $Tu = w$ defined by the mapping

$$T : C[a,b] \to C^1[a,b], \quad u \mapsto Tu = w, \quad w(t) = \int_a^t u(s)\, ds, \quad a \le t \le b.$$

We have already seen that $Tu = w$ is ill-posed, if the norm $\| \bullet \|_{C[a,b]}$ is used for both, $X = C[a,b]$ and $Y = C^1[a,b]$. Assume we have the following additional

information: $w \in H^2(a,b)$ and $\|w''\|_{L_2(a,b)} \le S$. Consequently, $u^* \in H^1(a,b)$ and $\|(u^*)'\|_{L_2(a,b)} \le S$. The normed linear space $(X, \|\bullet\|_X)$ given by $X = C[a,b]$ and $\|\bullet\|_X = \|\bullet\|_{C[a,b]}$ contains $X_0 := H^1(a,b)$ as a subspace. X_0 can be equipped with the semi-norm

$$\|\bullet\|_0 : X_0 \to \mathbb{R}, \quad x \mapsto \|x\|_0 := \|x'\|_{L_2(a,b)}.$$

(All constant functions x are contained in X_0 and have $\|x\|_0 = 0$ even if $x \neq 0$. Therefore, $\|\bullet\|_0$ is a semi-norm only). Thus, we have an additional information of the form

$$u^* \in X_0 \quad \text{and} \quad \|u^*\|_0 = \|(u^*)'\|_{L_2(a,b)} \le S,$$

which corresponds to (3.29). The inverse problem will be discretized by the collocation method. To find a spline approximant u_n of u^*, let $n \in \mathbb{N}$ with $n \ge 2$, let $h := (b-a)/(n-1)$ and let $t_i := a+(i-1)h$, $i = 1, \ldots, n$. Every $u_n \in \mathscr{S}_2(t_1, \ldots, t_n)$ can be written as

$$u_n = \sum_{j=1}^{n} x_j N_{j,2} \quad \Longrightarrow \quad \|u_n\|_0^2 = \sum_{j=1}^{n-1} h \left(\frac{x_{j+1} - x_j}{h} \right)^2 = \frac{1}{h} \|Lx\|_2^2 \tag{3.33}$$

where L is the matrix defined by

$$L = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n-1,n} . \tag{3.34}$$

By construction, $u_n \in H^1(a,b)$. The semi-norm condition for u_n directly translates into a semi-norm condition for the coefficient vector x, see (3.33). To make computations easy, assume that samples of $w(t) = \int_a^t u(s)\, ds$ are given at t_i, $i = 2, \ldots, n$, and additionally at $t_{i-1/2} := t_i - h/2$, $i = 2, \ldots, n$. Then, (2.51) takes the form of the following equations:

$$\begin{aligned} w(t_i) &= h \left(\frac{x_1}{2} + x_2 + \ldots + x_{i-1} + \frac{x_i}{2} \right), \quad i = 2, \ldots, n, \\ w(t_{1.5}) &= \frac{3}{8} x_1 + \frac{1}{8} x_2, \quad \text{and} \\ w(t_{i-1/2}) &= h \left(\frac{x_1}{2} + x_2 + \ldots + x_{i-2} + \frac{7x_{i-1}}{8} + \frac{x_i}{8} \right), \quad i = 3, \ldots, n. \end{aligned}$$

Setting

$$\beta := \begin{pmatrix} w(t_{1.5}) \\ w(t_2) \\ w(t_{2.5}) \\ w(t_3) \\ \vdots \\ w(t_{n-0.5}) \\ w(t_n) \end{pmatrix} \text{ and } A := h \begin{pmatrix} 0.375 & 0.125 & & & & & \\ 0.5 & 0.5 & & & & & \\ 0.5 & 0.875 & 0.125 & & & & \\ 0.5 & 1 & 0.5 & & & & \\ \vdots & \vdots & \ddots & \ddots & & & \\ 0.5 & 1 & \cdots & 1 & 0.875 & 0.125 \\ 0.5 & 1 & \cdots & 1 & 1 & 0.5 \end{pmatrix} \in \mathbb{R}^{2n-2,n}$$

and assuming that only perturbed data β^δ are available instead of β, the minimization problem (3.32) takes the form

$$\text{minimize } \|\beta^\delta - Ax\|_2 \text{ under the constraint } \|Lx\|_2 \le \sqrt{h}S. \tag{3.35}$$

For a numerical example let $a = 0$, $b = 1$ and $u^*(t) = t(1-t) - 1/6$ such that $\|(u^*)'\|_{L_2(a,b)} = 1/\sqrt{3} =: S$. Choose $n = 101$. In Fig. 3.4 (left) samples β of the exact effect w corresponding to u^* are shown (in red) as well as perturbed values β^δ (in black), which were obtained from β by adding to each component a random number drawn from a normal distribution with mean value 0 and standard deviation 10^{-3}. To the right we show the exact solution u^* (in red) as well as the approximation $u_n^\delta = \sum_{j=1}^n x_j^\delta N_{j,2}$ obtained from minimizing $\|\beta^\delta - Ax\|_2$ (in

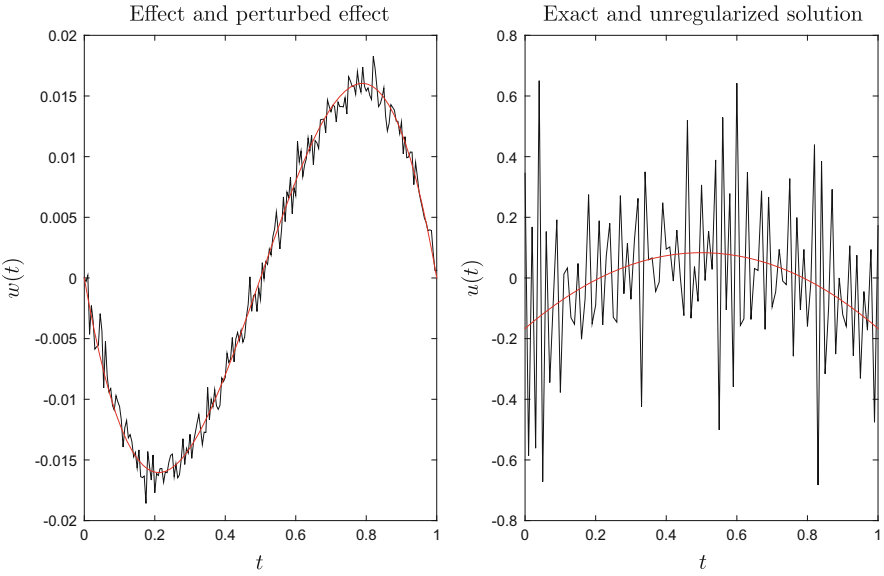

Fig. 3.4 Unregularized numerical differentiation

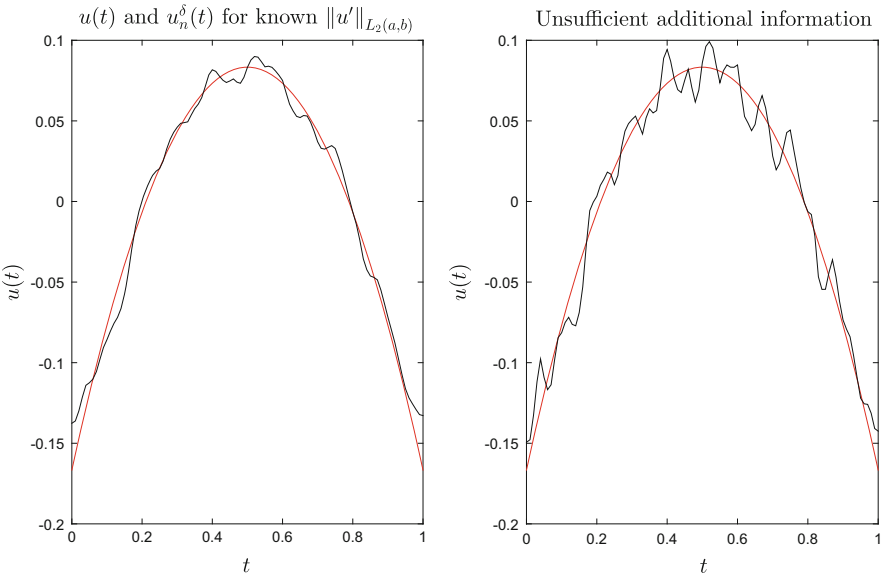

Fig. 3.5 Numerical differentiation regularized by using additional information

black). In Fig. 3.5 (left) we show (in black) the approximation u_n^δ obtained from minimizing $\|\beta^\delta - Ax\|_2$ under the constraint $\|Lx\|_2 \leq \sqrt{h}S$, as suggested by (3.35). For comparison, the exact solution is again shown in red. A practical solution method for the constrained minimization problem will only be discussed in Sect. 3.4. To the right we show an approximation u_n^δ obtained by using the less sharp constraint $\|(u_n^\delta)'\|_2 \leq 1 \cdot \sqrt{h}$. This shows that if the available information on u^* is of unsufficient quality, then the reconstruction quality might suffer accordingly. ◇

3.4 Tikhonov Regularization

Motivated by Sect. 3.3 and Example 3.19, we will investigate

Problem 3.20 (Linear least squares problem with quadratic constraints) Let

$$A \in \mathbb{R}^{m,n}, \quad L \in \mathbb{R}^{p,n}, \quad b \in \mathbb{R}^m, \quad \text{and} \quad S > 0\,. \tag{3.36}$$

Solve the constrained minimization problem

$$\text{minimize } \|b - Ax\|_2 \text{ under the constraint } \|Lx\|_2 \leq S. \tag{3.37}$$

An equivalent formulation of (3.36) is

$$\text{min.} \quad f(x) := \|b - Ax\|_2^2 \quad \text{such that} \quad h(x) := \|Lx\|_2^2 - S^2 \leq 0. \tag{3.38}$$

From the formulation of Problem 3.18 it becomes clear, how Problem 3.20 relates to a linear parameter identification problem $Tu = w$, discretized by the method of least squares or by the collocation method.

Replacing the linear least squares problem (3.28) by (3.37) is called **Tikhonov regularization** and goes back to the Russian mathematician Tikhonov and the American mathematician Phillips. An often considered case is $L = I_n \in \mathbb{R}^{n,n}$ (unity matrix of dimension $n \times n$), the so-called **standard case**. Matrices $L \neq I_n$ are used to formulate constraints on the discrete version of derivatives, as with L defined in (3.34) in Example 3.19. Likewise one could choose

$$L = \begin{pmatrix} -1 & 2 & -1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & -1 & 2 & -1 \end{pmatrix} \in \mathbb{R}^{n-2,n}, \tag{3.39}$$

to formulate a constraint on second derivatives. Note that $x \mapsto \|Lx\|_2$ only is a semi-norm if $p < n$, since in that case $\|Lx\|_2 = 0$ is possible for $x \neq 0$. This is why we only required $\| \bullet \|_0$ to be a semi-norm in Assumption 3.16. The (quadratic) function f in (3.38) is called **objective function**, the set $N := \{x \in \mathbb{R}^n;\ h(x) \leq 0\}$ is called **feasible region**. The objective function f is continuous and convex and N is convex, closed, and nonempty (since $0 \in N$). This ensures that a solution of (3.37) always exists. The constraint $h(x) \leq 0$ is called **binding**, if

$$h(\hat{x}) > 0 \quad \text{for all} \quad \hat{x} \in M := \{x \in \mathbb{R}^n;\ f(x) = \min\}. \tag{3.40}$$

In this case, no element of M can be a solution of (3.37). Put differently, if x^* *is* a solution of (3.37), then $f(\hat{x}) < f(x^*)$ for all $\hat{x} \in M$. Because of the convexity of f and the convexity and closedness of N, this does mean that x^* necessarily lies on the boundary of N whenever (3.40) is valid.

Theorem 3.21 (Linear least squares problem with quadratic constraints) *Let A, L, b and S be as in (3.36) and let f and h be defined as in (3.38). Also assume that*

$$\mathit{rank} \begin{pmatrix} A \\ L \end{pmatrix} = n. \tag{3.41}$$

Then the linear system of equations

$$(A^T A + \lambda L^T L)x = A^T b \tag{3.42}$$

has a unique solution x_λ for every $\lambda > 0$. If (3.40) holds, i.e. if the constraint of the minimization problem (3.37) is binding, then there exists a unique $\lambda > 0$ such that

$$\|Lx_\lambda\|_2 = S \tag{3.43}$$

and the corresponding x_λ is the unique solution of (3.37).

Remarks Condition (3.41) can equivalently be written in the form

$$\mathcal{N}_A \cap \mathcal{N}_L = \{0\}. \tag{3.44}$$

We usually require $\operatorname{rank}(A) = n$, in which case condition (3.41) is automatically fulfilled. Then there exists a unique minimizer $\hat{x}$ of $\|b - Ax\|_2$. Either $h(\hat{x}) \leq 0$, such that $\hat{x}$ also is the unique minimizer of (3.37), or $h(\hat{x}) > 0$, meaning that (3.40) holds, since $\hat{x}$ is the only element of M.

Proof Let $\lambda > 0$. When (3.41) holds, then

$$\operatorname{rank} \begin{pmatrix} A \\ tL \end{pmatrix} = n \quad \text{for} \quad t = \sqrt{\lambda} > 0.$$

According to Theorem 3.1 a unique solution of the linear least squares problems

$$\min_{x \in \mathbb{R}^n} \left\{ \left\| \begin{pmatrix} A \\ \sqrt{\lambda} L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2 \right\}, \quad \lambda > 0, \tag{3.45}$$

exists, which is defined as the unique solution of the corresponding normal equations, namely (3.42). From now on let (3.40) hold. As already seen, this implies that any solution of (3.37) necessarily is located on the boundary of N. Problem (3.37) then is equivalent to

$$\text{minimize} \quad f(x) \quad \text{under the constraint} \quad h(x) = 0.$$

It is easy to verify that for the gradient ∇h of h

$$\nabla h(x) = 0 \quad \Longleftrightarrow \quad L^T L x = 0 \quad \Longleftrightarrow \quad \|Lx\|_2 = 0$$

holds. From this one concludes that problem (3.38) is **regular**, i.e. the gradient of the constraint function does not vanish at an optimal point. In fact, if x^* is an optimal point, then $h(x^*) = \|Lx^*\|_2^2 - S^2 = 0$ and from $S > 0$ follows $\|Lx^*\|_2 \neq 0$, thus $\nabla h(x^*) \neq 0$. The well known theorem on Lagrange multipliers – applicable since (3.38) is regular – tells us a scalar λ exists for which

$$0 = \nabla f(x^*) + \lambda \nabla h(x^*) = 2(A^T A + \lambda L^T L)x^* - 2A^T b, \tag{3.46}$$

where x^* still is a solution of (3.38). It is known the gradient of a function points into the direction of ascending function values. Therefore $\nabla h(x^*)$ must point outwards of N, x^* being a boundary point. But then λ can not be negative: if it was, then we would conclude from (3.46) that moving from x^* into the (feasible) interior of N, values of f would diminish, a contradiction to the optimality of x^*. If λ would vanish, then (3.46) would tell us $\nabla f(x^*) = 0$, meaning that x^* would be a minimizer of (the convex function) f, which would contradict (3.40). The only remaining possibility is $\lambda > 0$. But in this case, as we have already seen, (3.46) has a unique solution $x^* = x_\lambda$, for which $h(x_\lambda) = 0$, i.e. for which (3.43) must hold. It remains to show that λ is unique. To do so, define $J, E : \mathbb{R}^+ \to \mathbb{R}_0^+$ by setting

$$J(\lambda) := f(x_\lambda) = \|b - Ax_\lambda\|_2^2 \quad \text{and} \quad E(\lambda) := \|Lx_\lambda\|_2^2. \tag{3.47}$$

The proof is accomplished if we can show that E is strictly monotonic decreasing. Since $f, h : \mathbb{R}^n \to \mathbb{R}$ both are convex, so is $f + \lambda h : \mathbb{R}^n \to \mathbb{R}$ for every $\lambda > 0$. This function is minimized where its gradient $\nabla f(x) + \lambda \nabla h(x)$ – compare (3.46) – vanishes. As we have seen, for every $\lambda > 0$ a unique $x = x_\lambda$ exists where this happens. Consequently, for two values $0 < \lambda_1 < \lambda_2$ we must have

$$J(\lambda_1) + \lambda_1 E(\lambda_1) < J(\lambda_2) + \lambda_1 E(\lambda_2)$$

as well as

$$J(\lambda_2) + \lambda_2 E(\lambda_2) < J(\lambda_1) + \lambda_2 E(\lambda_1).$$

Adding both inequalities shows

$$(\lambda_1 - \lambda_2)(E(\lambda_1) - E(\lambda_2)) < 0,$$

from which $E(\lambda_1) > E(\lambda_2)$ follows. Likewise one can show that J is strictly monotonic increasing (divide the first inequality by λ_1, the second by λ_2 and add them afterwards). □

The proof shows how Problem 3.20 can be solved in principle, at least in case $\operatorname{rank}(A) = n$ (in this case condition (3.40) is easily checked).

Solution method for Problem 3.20

(1) Compute the unique minimizer x_0 of $\|b - Ax\|_2$ and check, whether $\|Lx_0\|_2 \leq S$. If this is the case, x_0 is the sought-after solution. Otherwise:

(2) Using some numerical method, for example Newton's method, solve $E(\lambda) = 0$, knowing that $E : (0, \infty) \to \mathbb{R}$, $\lambda \mapsto \|Lx_\lambda\|_2^2 - S^2$, is strictly

(continued)

monotonic decreasing. Each evaluation of E requires the computation of a solution x_λ of the *unconstrained* linear least squares problem (3.45), which can be found by solving the normal equations (3.42). The zero λ^* of E determines the solution $x^* = x_{\lambda^*}$ of (3.37).

Before Problem 3.20 is analyzed and solved, we turn to a closely related, "symmetric" problem, which will come into play in Sect. 3.5 and which will be relevant, if additional information of the form $\|Lx\|_2 \leq S$ is not available.

Problem 3.22 (Linear least squares problem with quadratic constraints) Let

$$A \in \mathbb{R}^{m,n}, \quad L \in \mathbb{R}^{p,n}, \quad b \in \mathbb{R}^m, \quad \text{and} \quad \delta > 0\,. \tag{3.48}$$

Solve the constrained minimization problem

$$\text{minimize } \|Lx\|_2 \text{ under the constraint } \|Ax - b\|_2 \leq \delta. \tag{3.49}$$

An equivalent formulation of (3.48) is

$$\text{min.} \quad f(x) := \|Lx\|_2^2 \quad \text{such that} \quad h(x) := \|Ax - b\|_2^2 - \delta^2 \leq 0. \tag{3.50}$$

Theorem 3.23 (Linear least squares problems with quadratic constraints) *Let A, L, b and δ be as in (3.48) and let f and h be defined as in (3.50). Let $\mathcal{N}_A \cap \mathcal{N}_L = \{0\}$ and let $x_0 \in \mathbb{R}^n$ be a minimizer of $\|Ax - b\|_2$ (without constraint). Assume that*

$$\|Ax_0 - b\|_2 < \delta < \|b\|_2. \tag{3.51}$$

The linear system of equations

$$(A^T A + \lambda L^T L)x = A^T b \tag{3.52}$$

has a unique solution x_λ for every $\lambda > 0$. If the constraint of $h(x) \leq 0$ is binding, then there exists a unique $\lambda > 0$ such that

$$\|Ax_\lambda - b\|_2 = \delta \tag{3.53}$$

and the corresponding x_λ is the unique solution of (3.49).

Remark The first inequality in (3.51) is needed for regularity of Problem 3.22, see the proof below. The second inequality in (3.51) excludes the trivial solution $x = 0$ and is necessary for the constraint to be binding.

Proof The existence of a solution x^* of Problem 3.22 is proved as for Problem 3.20. The existence of a unique solution of (3.52) follows from $\mathcal{N}_A \cap \mathcal{N}_L = \{0\}$ as in Theorem 3.21. From now on, let the constraint be binding, i.e. let $h(x) > 0$ for any minimizer of $\|Lx\|_2$ (for any $x \in \mathcal{N}_L$). In this case, any solution x^* of Problem 3.22 necessarily lies on the boundary of the feasible region, such that Problem 3.22 can be replaced equivalently by

$$\text{minimize} \quad f(x) \quad \text{under the constraint} \quad h(x) = 0. \tag{3.54}$$

For a solution x^* of (3.54) one must have $h(x^*) = 0$. One easily verifies

$$\nabla h(x^*) = 0 \quad \Longleftrightarrow \quad A^T A x^* = A^T b,$$

which is equivalent to x^* being a minimizer of $h(x)$. But in this case $h(x^*) = h(x_0) < 0$ according to (3.51). Consequently, $\nabla h(x^*) \neq 0$ for a solution of (3.54), which is the regularity condition. From the Lagrange multiplier theorem one deduces the existence of $\mu \in \mathbb{R}$ such that

$$\nabla f(x^*) + \mu \nabla h(x^*) = 0. \tag{3.55}$$

As in the proof of Theorem 3.21, it can be shown that necessarily $\mu > 0$. But then (3.55) is equivalent to (3.52), with $\lambda = 1/\mu$. The rest of the proof is equal to that of Theorem 3.21. □

Analysis of Problem 3.20 and Its Practical Solution

An analysis of (3.37) often is based on the "Generalized Singular Value Decomposition (GSVD)". This is not necessary. The following alternative approach, which goes back to [Rei67], has the advantage that a very efficient practical method to compute the solution $x_{\lambda *}$ of (3.37) can directly be derived from it. *The analysis could be carried out under condition* (3.41)*, but to make things easier, we will more specifically assume that* $\operatorname{rank}(A) = n \leq m$. In this case the normal equations (3.42) already have a unique solution x_0 for $\lambda = 0$. If $h(x_0) \leq 0$, we are done, since then x_0 is the solution of (3.37). So let $h(x_0) > 0$ from now on.

Since $\operatorname{rank}(A) = n$, the matrix $A^T A$ is positive definite. In addition, $L^T L$ is positive semidefinite. According to (A.3), a non-singular matrix $V = (v_1 | v_2 | \ldots | v_n) \in \mathbb{R}^{n,n}$ exists such that

$$V^T A^T A V = \operatorname{diag}(1, \ldots, 1) \quad \text{and} \quad V^T L^T L V = \operatorname{diag}(\kappa_1, \ldots, \kappa_n), \tag{3.56}$$

where $\kappa_i \geq 0$ for all i. This can be put differently by saying

$$v_i^T A^T A v_j = \begin{cases} 1, & i = j \\ 0, & \text{else} \end{cases} \quad \text{and} \quad v_i^T L^T L v_j = \begin{cases} \kappa_i, & i = j \\ 0, & \text{else} \end{cases} .$$

Choose the numbering such that

$$\kappa_1 \geq \ldots \geq \kappa_r > 0 \quad \text{and} \quad \kappa_{r+1} = \ldots = \kappa_n = 0 \quad \text{where} \quad r := \operatorname{rank}(L).$$

From (3.56) one gets

$$A^T A v_i = \frac{1}{\kappa_i} L^T L v_i, \quad i = 1, \ldots, r.$$

For fixed $\lambda \geq 0$ the solution x_λ of (3.42) can be written in the form $x_\lambda = \sum_{i=1}^n \tau_i v_i$. Substituting this into (3.42) one gets

$$\sum_{i=1}^n \tau_i (A^T A v_i + \lambda L^T L v_i) = \sum_{i=1}^r \tau_i \left(\frac{1}{\kappa_i} + \lambda \right) L^T L v_i + \sum_{i=r+1}^n \tau_i A^T A v_i = A^T b.$$

To determine the unknowns τ_i, multiply these identities from the left by $v_1^T, \ldots, v_n^T$ to get

$$x_\lambda = \sum_{i=1}^n \left(\frac{\gamma_i}{1 + \lambda \kappa_i} \right) v_i, \quad \gamma_i = v_i^T A^T b, \tag{3.57}$$

which is valid for $\lambda \geq 0$. One also gets

$$E(\lambda) = \|L x_\lambda\|_2^2 = \sum_{i=1}^n \left(\frac{\gamma_i}{1 + \lambda \kappa_i} \right)^2 \kappa_i \, . \tag{3.58}$$

This shows that under the conditions $\operatorname{rank}(A) = n$ and $E(0) = \|L x_0\|_2^2 > S^2$, the function $E : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is not only positive and strictly monotonic decreasing from $E(0) > S^2$ to 0, but that it is also convex. As a consequence, if we use Newton's method to find the solution λ^* of $E(\lambda) - S^2 = 0$ and if we start it with $\lambda_0 = 0$ (or with some positive λ_0 to the left of λ^*), then it will produce a sequence $(\lambda_k)_{k \in \mathbb{N}_0}$ converging monotonously to λ^*. Using the abbreviations $E_k := E(\lambda_k)$ and $E_k' := E'(\lambda_k)$, the sequence $(\lambda_k)_{k \in \mathbb{N}_0}$ is recursively defined by Newton's method:

$$\lambda_{k+1} = \lambda_k - \frac{E(\lambda_k) - S^2}{E'(\lambda_k)} = \lambda_k - \frac{E_k - S^2}{E_k'}, \quad k = 0, 1, 2, \ldots \tag{3.59}$$

and requires the computation of derivatives $E'(\lambda)$:

$$E'(\lambda) = \frac{d}{d\lambda}\left(x_\lambda^T L^T L x_\lambda\right) = 2x_\lambda^T L^T L x'_\lambda, \quad x'_\lambda = \frac{d}{d\lambda} x_\lambda.$$

Implicit differentiation of identity (3.42) with respect to λ shows

$$L^T L x_\lambda + (A^T A + \lambda L^T L) x'_\lambda = 0,$$

so we get the explicit formula

$$E'(\lambda)/2 = -x_\lambda^T L^T L (A^T A + \lambda L^T L)^{-1} L^T L x_\lambda. \tag{3.60}$$

Note that if we compute x_λ from (3.42) via a Cholesky factorization $A^T A + \lambda L^T L = R^T R$, then

$$E'(\lambda)/2 = -\|R^{-T} L^T L x_\lambda\|_2^2$$

only needs the computation of $z = R^{-T} L^T L x_\lambda$, which can be done very cheaply by solving $R^T z = L^T L x_\lambda$. The need to compute derivatives often is a disadvantage of Newton's method, but in the present case this computation means nearly no effort at all.

From [Rei67] we take the recommendation to replace $E(\lambda) - S^2 = 0$ by the equivalent equation

$$G(\lambda) - \frac{1}{S} = 0, \quad G(\lambda) = \frac{1}{\sqrt{E(\lambda)}}, \tag{3.61}$$

and use Newton's method to solve the latter. Since E is positive and strictly monotonic decreasing, G is positive and strictly monotonic increasing. From $E(0) > S^2$ and $E(\lambda) \to 0$ for $\lambda \to \infty$ we get $G(0) < S^{-1}$ and $G(\lambda) \to \infty$ for $\lambda \to \infty$, so there is a unique solution of (3.61). Using $G'(\lambda) = -(1/2)E(\lambda)^{-3/2} E'(\lambda)$ (which can be computed as efficiently as $E'(\lambda)$), Newton's method applied to (3.61) will produce a sequence

$$\lambda_{k+1} = \lambda_k + 2\frac{E_k^{3/2}}{E'_k} \cdot (E_k^{-1/2} - S^{-1}), \quad k = 0, 1, 2, \ldots \tag{3.62}$$

(the same abbreviations are used as for (3.59)). The recursion will be started again at $\lambda_0 = 0$. Since G is concave (which can be shown with some effort), (3.62) again is a monotonously converging sequence. The advantage of using (3.62) instead of (3.59) lies in its faster convergence. To see this, consider the ratio of the respective increments $(\lambda_{k+1} - \lambda_k)$ which both methods take when started at a

common reference point λ_k:

$$\nu_k := 2\frac{E_k^{3/2}(E_k^{-1/2} - S^{-1})}{E_k'} \cdot \frac{-E_k'}{E_k - S^2} = \frac{2}{q_k + \sqrt{q_k}}$$

with $q_k := S^2/E_k$. Since the Newton iteration is started at $\lambda_0 = 0$, to the left of the solution λ^* of $E(\lambda) - S^2$, we have $0 < q_k < 1$. Consequently, $\nu_k > 1$. In a computer implementation, the iteration will be stopped as soon as $\lambda_{k+1} \leq \lambda_k$ is observed, since this can only happen due to finite precision arithmetic. Before the iteration is started, the condition $\|Lx_0\|_2 > S$ has to be checked.

As an example for regularization as in Problem 3.20 we have already considered numerical differentiation (Example 3.19). It is a good idea to use additional information as formulated in Assumption 3.16, but unluckily, such kind of information often is not available. In Example 3.19 it turned out that if $\|(u^*)'\|_{L_2(a,b)}$ is not exactly known, then the reconstruction quality of u^* suffers. We are now looking for an alternative approach to regularization.

Generalization of Tikhonov Regularization

When information as in Assumption 3.16 is not available, heuristics come into play. Our starting point for going into this is a modification of Problem 3.20.

Problem 3.24 (Linear least squares problem with Lagrange parameter) Let

$$A \in \mathbb{R}^{m,n}, \quad L \in \mathbb{R}^{p,n}, \quad \text{and} \quad b \in \mathbb{R}^m. \tag{3.63}$$

Let $\operatorname{rank}(A) = n$. Find the minimizer $x_\lambda \in \mathbb{R}^n$ of

$$\min_x \left\{ \|b - Ax\|_2^2 + \lambda \|Lx\|_2^2 \right\} = \min_x \left\{ \left\| \begin{pmatrix} A \\ \sqrt{\lambda} L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2 \right\}, \tag{3.64}$$

where the parameter $\lambda \geq 0$ has to be chosen.

We know that (3.64) has a unique minimizer x_λ for each $\lambda \geq 0$. If λ is chosen as the Lagrange multiplier corresponding to the optimization Problem 3.20, then we will get the same result as before. By choosing $\lambda \in [0, \infty)$ freely, the minimizer x_λ can be shaped to some extent. To see how this works, remember the functions

$J, E : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ defined by

$$J(\lambda) = \|b - Ax_\lambda\|_2^2 \quad \text{and} \quad E(\lambda) = \|Lx_\lambda\|_2^2 \tag{3.65}$$

in (3.47). We found that J is strictly monotonic increasing whereas E is strictly monotonic decreasing with $E(0) = \|Lx_0\|_2^2$ and $\lim_{\lambda\to\infty} E(\lambda) = 0$. Evidently the parameter λ is weighting the respective "importance" one is attributing to the minimization of J and E. In the limit $\lambda \to 0$ we will get the minimizer x_0 of $\|b - Ax\|_2$. Among all x_λ, $\lambda \geq 0$, this is the one which satisfies best the system of equations $Ax = b$ and thus is "closest to the data" b. Therefore, $J(\lambda)$ can be considered an inverse measure of the data fit of x_λ. In the other limit case $\lambda \to \infty$ we can derive the identity

$$x_\infty = \sum_{i=r+1}^{n} \frac{v_i^T A^T b}{v_i^T A^T A v_i} v_i \in \mathcal{N}_L = \langle v_{r+1}, \ldots, v_n \rangle$$

from (3.57). This is the unique minimizer of $\|b - Ax\|_2$ in the subspace $\mathcal{N}_L$.[4] In Example 3.19, function $u_n = \sum_{j=1}^{n} x_{\infty,j} N_{j,2}$ is the constant function, which fits best the data. Since the matrices L used so far are related to (discretized) derivatives of functions $u_n = \sum x_{\lambda,j}\varphi_j$, one may consider $E(\lambda)$ an inverse measure of smoothness. The parameter λ thus controls the balance x_λ holds between fidelity to the data and smoothness of a solution. In contrast to the objective way in which additional information is used according to Assumption 3.16, a free choice of λ corresponds to a more subjective belief of how smooth u_n should be versus how much it should be close to the data. Choosing λ subjectively can be quite dangerous and drive an approximation u_n in a wrong direction, if one does not really know that the exact solution u^* actually *has* some degree of smoothness. Still, we could replace (3.64) by the more general minimization problem

$$\min_{\alpha\in D} \left\{ \|b - Ax\|_2^2 + \lambda G(x) \right\}, \tag{3.66}$$

where G is some positive function incorporating a priori information about u^*. The term $G(x)$ should measure (inversely), to what extent $u_n = \sum_j x_j\varphi_j$ has a quality that we know u^* also has. Conversely, the value $G(x)$ should become large, whenever $u_n = \sum_j x_j\varphi_j$ does not have this desired quality. For this reason, G is called a **penalty function**. For example, if we know that $u_n = \sum_{j=1}^{n} x_j^*\varphi_j$ is somehow close to u^*, we

[4] To see this, compute the gradient of the convex function $h(\tau_{r+1}, \ldots, \tau_n) = \|b - A(\sum_{i=r+1}^{n} \tau_i v_i)\|_2^2$ considering $v_i^T A^T A v_j = 0$ for $i \neq j$.

could choose

$$G(x) := \|x - x^*\|_2^2. \tag{3.67}$$

If we know that $u^* : [a, b] \to \mathbb{R}$ is not oscillating, and if $u_n = \sum x_j \varphi_j$ is a polygonal line with vertices at ordinates x_j, then we could choose

$$G(x) := \sum_{j=1}^{n-1} \left| x_{j+1} - x_j \right|, \tag{3.68}$$

which measures the total variation of u_n. Choosing a subset $D \subseteq \mathbb{R}^n$ in (3.66) is another way to incorporate information about u^*. For example, if we know that u^* is positive, monotone, or convex, then D should be chosen such that for $x \in D$ a corresponding approximant $u_n = \sum x_j \varphi_j$ also is positive, monotone, or convex, respectively. We will come back to generalized Tikhonov regularization of the form (3.66) in Sect. 4.5.

Before we discuss a method to choose λ in Problem 3.24, we show that (3.64) is a regularized version of the linear least squares problem. To make the analysis easier, we specialize on the ("norm") case $L = I_n \in \mathbb{R}^{n,n}$. This is not an important restriction, though, since any problem of the form (3.64) can be transformed to a standard problem with $L = I$, see Sect. 3.6.

Theorem 3.25 (Approximation power of Tikhonov regularization) *Let $A \in \mathbb{R}^{m,n}$ with rank$(A) = n \leq m$ and singular values $\sigma_1 \geq \ldots \geq \sigma_n > 0$. Let $L = I_n$, $b, b^\delta \in \mathbb{R}^m$ and $\|b - b^\delta\|_2 \leq \delta$ for $\delta > 0$. Let $\lambda \geq 0$ and denote by*

$\hat{x}$ *the unique solution of $\|b - Ax\|_2 = \min!$ (exact data) and by*
x_λ *the unique solution of (3.64) for b replaced by b^δ (perturbed data).*

Then the following estimates hold

(1) $\|\hat{x} - x_0\|_2 \leq \dfrac{\delta}{\sigma_n}$ *and*

(2) $\|\hat{x} - x_\lambda\|_2 \leq \dfrac{\sqrt{\lambda}}{2\sigma_n}\|\hat{x}\|_2 + \dfrac{\delta}{2\sqrt{\lambda}}$ *for $\lambda > 0$.*

From (2) one concludes that if λ is chosen such that for $\delta \to 0$

$$\lambda \to 0 \quad \text{and} \quad \frac{\delta^2}{\lambda} \to 0,$$

then (3.64) is a regularization of the linear least squares problem according to Definition 3.14.

Proof Part (1) was already proven in Theorem 3.6. Part (2): Using the SVD $A = U\Sigma V^T$ we get

$$V^T A^T A V = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) \quad \text{and} \quad V^T I_n V = I_n$$

as a special case of (3.56). Observing $v_i^T A^T = \sigma_i u_i^T$, (3.57) becomes

$$x_\lambda = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda} \cdot (u_i^T b^\delta) \cdot v_i \,. \tag{3.69}$$

For each $\lambda \geq 0$ we can define an operator

$$A_\lambda^+ : \mathbb{R}^m \to \mathbb{R}^n, \quad y \mapsto A_\lambda^+ y := \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda} \cdot (u_i^T y) \cdot v_i,$$

such that $x_\lambda = A_\lambda^+ b^\delta$. Since $\hat{x} = A^+ b$, we get

$$\|\hat{x} - x_\lambda\|_2 \leq \|A^+ b - A_\lambda^+ b\|_2 + \|A_\lambda^+ b - A_\lambda^+ b^\delta\|_2. \tag{3.70}$$

Both terms on the right hand side will be estimated separately. Since $A^+ = A_0^+$ we get

$$A^+ b - A_\lambda^+ b = \sum_{i=1}^{n} \frac{\lambda}{\sigma_i^2 + \lambda} \frac{1}{\sigma_i} (u_i^T b) v_i.$$

It is easy to check that

$$\frac{\sigma^2}{\sigma^2 + \lambda} \leq \frac{\sigma}{2\sqrt{\lambda}} \quad \text{for} \quad \sigma, \lambda > 0, \tag{3.71}$$

from which $\lambda/(\sigma_i^2 + \lambda) \leq \sqrt{\lambda}/(2\sigma_n)$ follows, such that

$$\|A^+ b - A_\lambda^+ b\|_2^2 \leq \frac{\lambda}{4\sigma_n^2} \cdot \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left| u_i^T b \right|^2 = \frac{\lambda}{4\sigma_n^2} \|A^+ b\|_2^2.$$

This shows that the first term on the right hand side of (3.70) has $(\sqrt{\lambda}/(2\sigma_n)) \cdot \|\hat{x}\|_2$ as an upper bound. To estimate the second term, (3.71) is reused:

$$\|A_\lambda^+ b - A_\lambda^+ b^\delta\|_2^2 = \sum_{i=1}^{n} \underbrace{\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda} \cdot \frac{1}{\sigma_i} \right)^2}_{\leq 1/(4\lambda)} \left| u_i^T (b - b^\delta) \right|^2 \leq \frac{1}{4\lambda} \|b - b^\delta\|_2^2.$$

By the way, this gives an upper bound

$$\|A_\lambda^+\|_2 \le \frac{1}{2\sqrt{\lambda}} \tag{3.72}$$

for the spectral norm of the operator A_λ^+, which regularizes A^+. □

3.5 Discrepancy Principle

We have to set the parameter λ in Problem 3.24. This can be done according to the following rule known as **Morozov's discrepancy principle**. It is assumed that only an approximation b^δ to the data b is at hand, and that its error can be bounded in the form $\|b - b^\delta\|_2 \le \delta$ with $\delta > 0$ known.

Morozov's discrepancy principle for choosing the regularization parameter λ in (3.64), with b replaced by b^δ.

- Choose $\lambda = \infty$, if

$$\|b^\delta - Ax_\infty\|_2 \le \delta. \tag{3.73}$$

- Choose $\lambda = 0$, if

$$\|b^\delta - Ax_0\|_2 > \delta. \tag{3.74}$$

- Otherwise choose the unique value λ, for which

$$\|b^\delta - Ax_\lambda\|_2 = \delta. \tag{3.75}$$

Note that $\lambda = \lambda(\delta, b^\delta)$ is chosen depending on the error level δ as well as on the data b^δ, so this is a parameter choice a posteriori. The following are the two basic ideas behind the discrepancy principle.

1. One does not strive to achieve a better data fit than $\|b^\delta - Ax_\lambda\|_2 = \delta$, since b^δ itself does not approximate the exact data b better.
2. Since $J(\lambda) = \|b^\delta - Ax_\lambda\|_2$ is strictly monotonic increasing, the largest possible λ is chosen such that $\|b^\delta - Ax_\lambda\|_2 \le \delta$. (In the limiting cases (3.73) and (3.74) all, or no, x_λ satisfy $\|b^\delta - Ax_\lambda\|_2 \le \delta$).

From Theorem 3.23 it can be seen that *Tikhonov regularization with parameter choice according to the discrepancy principle solves Problem 3.22, if a solution exists. An a priori information of the form* $\|Lx\|_2 \leq S$ *about the true solution, as it was used in Problem 3.20, is replaced by an a priori information of the form* $\|b - b^\delta\|_2 \leq \delta$ *about the data error.* The discrepancy principle has to be accompanied by two warnings. *First*, and most importantly, choosing λ as large as possible seems justified *if one knows that* u^* *has the smoothness property, which is inversely measured by* E, but it can also be misleading, if u^* does not have this property. We can not make a totally subjective choice of E and then hope to miraculously detect u^* from the data b^δ alone. For example, if u^* is sharply peaked (the location of the peak might be an important information), but E is chosen such that large derivatives are penalized, then regularization will carry us away from the sought-after solution. *Second*, assuming that $u_n^* = \sum x_j^* \varphi_j$ is an optimal approximant of the exact solution u^* (according to some norm), then it may happen that $\|b^\delta - Ax^*\|_2 < \delta$. In this case, the discrepancy principle will choose too large a parameter λ. The discrepancy principle is then said to "over-regularize".

By the strict monotonocity of J, uniqueness of λ with (3.75) is guaranteed. Note that with $\hat{x} = A^+ b$ we have

$$\|b^\delta - Ax_0\|_2 \leq \|b^\delta - A\hat{x}\|_2 \leq \|b^\delta - b\|_2 + \|b - A\hat{x}\|_2, \tag{3.76}$$

where the second term on the right hand side should tend to 0 with the discretization getting finer and finer. So if we encounter case (3.74) for some finite value δ, this is a hint that the discretization was chosen too coarse. In this case we can only find some x with $\|b^\delta - Ax\|_2 \leq \delta$ if δ is an upper bound for data error *plus* discretization error. The following theorem formally shows that the discrepancy principle makes Tikhonov regularization a regularization according to Definition 3.14.

Theorem 3.26 (Tikhonov regularization and discrepancy principle) *Let* $A \in \mathbb{R}^{m,n}$ *with* $\text{rank}(A) = n \leq m$ *and with singular values* $\sigma_1 \geq \ldots \geq \sigma_n > 0$ *and let* $L \in \mathbb{R}^{p,n}$. *Let* $b, b^\delta \in \mathbb{R}^m$. *Let* $\lambda \geq 0$ *and denote by*

$\hat{x}$ *the unique solution of* $\|b - Ax\|_2 = \min!$ *(exact data) and by*
x_λ *the unique solution of* (3.64) *for* b *replaced by* b^δ *(perturbed data).*

Assume that for some known $\delta > 0$

$$\|b - b^\delta\|_2 + \|b - A\hat{x}\|_2 \leq \delta. \tag{3.77}$$

Determine λ according to the discrepancy principle. Then case (3.74) can not occur and the estimate

$$\|\hat{x} - x_\lambda\|_2 \leq 2\frac{\delta}{\sigma_n} \tag{3.78}$$

holds.

Proof Because of (3.76) and (3.77), case (3.74) can not occur. The discrepancy principle will therefore choose $0 < \lambda \leq \infty$ and a corresponding x_λ. For the residual $r_\lambda = b^\delta - Ax_\lambda$ one has $\|r_\lambda\|_2 \leq \delta$. From

$$\begin{aligned} A^T r_\lambda = A^T b^\delta - A^T A x_\lambda = \underbrace{A^T b}_{= A^T A\hat{x}} - A^T(b - b^\delta) - A^T A x_\lambda \\ = A^T A(\hat{x} - x_\lambda) - A^T(b - b^\delta) \end{aligned}$$

we get

$$\hat{x} - x_\lambda = (A^T A)^{-1} A^T r_\lambda + (A^T A)^{-1} A^T (b - b^\delta).$$

From $\|r_\lambda\|_2 \leq \delta$, from $\|b - b^\delta\|_2 \leq \delta$, and from $\|(A^T A)^{-1} A^T\|_2 \leq \sigma_n^{-1}$ (using an SVD of A), the estimate (3.78) follows. □

In functional analytic regularization theory for operator inversion, regularization methods (including parameter choices) as in Definition 3.14 are compared by their "order of convergence". This is the rate at which the error defined in (3.26) tends to zero for $\delta \to 0$ (which does not tell much about the actual error magnitude achieved for some given, finite value $\delta > 0$). Best possible rates can be achieved under specific assumptions on the exact solution u^* of $Tu = w$. In the infinite-dimensional case, one can generally not hope that this error goes to zero faster than $\mathcal{O}(\delta^\mu)$, where $0 < \mu < 1$ is a parameter depending on u^*, see Section 3.2 of [EHN96]. In the finite-dimensional case however, where the "operator" A^+ is continuous, part (c) of Theorem 3.6 shows $\|A^+ b - x_0\|_2 \leq \delta/\sigma_n$ for the *unregularized* solution $x_0 = A^+ b^\delta$, which is an even better bound than the one given for $\|A^+ b - x_\lambda\|_2$ in (3.78). Theorem 3.26 therefore can not justify the need for regularization of finite-dimensional problems (which are well-posed in the sense of Definition 1.5) and just shows formally, that the combination of Tikhonov regularization and discrepancy principles fulfills the requirements of Definition 3.14. The benefit of regularizing a finite-dimensional least squares problem can rather be seen from (the proof of) Theorem 3.25:

(continued)

comparing the formulae for x_0 and (3.69) for x_λ, namely

$$x_0 = \sum_{i=1}^{n} \frac{1}{\sigma_i} \cdot (u_i^T b^\delta) \cdot v_i \quad \text{and} \quad x_\lambda = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda} \cdot (u_i^T b^\delta) \cdot v_i,$$

one notes that the error-amplifying effect of small singular values is moderated by choosing a regularization parameter $\lambda > 0$.

In an implementation of the discrepancy principle one first has to check for the cases (3.73) and (3.74). If these can be excluded, then there exists a unique $\lambda = \lambda^*$ fulfilling (3.75). This is the unique solution of the equation

$$J(\lambda) - \delta^2 = 0,$$

which can be found by Newton's method. The required derivatives can be found analogously to (3.60) from

$$J'(\lambda) = -\lambda E'(\lambda).$$

However, the function $J - \delta^2$ is not convex. It is advantageous to consider the equivalent equation

$$I(\lambda) := J\left(\frac{1}{\lambda}\right) - \delta^2 = 0 \tag{3.79}$$

instead, as will be seen now. Since J is strictly monotonic increasing, I is strictly monotonic decreasing. We compute

$$I'(\lambda) = \frac{1}{\lambda^3} E'\left(\frac{1}{\lambda}\right) \quad \Longrightarrow \quad I''(\lambda) = -\frac{3}{\lambda^4} E'\left(\frac{1}{\lambda}\right) - \frac{1}{\lambda^5} E''\left(\frac{1}{\lambda}\right).$$

From (3.58) one gets, with $r = \operatorname{rank}(L)$

$$E'(\lambda) = \sum_{i=1}^{r} \frac{-2\kappa_i^2 \gamma_i^2}{(1+\lambda\kappa_i)^3} \quad \Rightarrow \quad -\frac{3}{\lambda^4} E'\left(\frac{1}{\lambda}\right) = \sum_{i=1}^{r} \frac{6\kappa_i^2 \gamma_i^2}{\lambda(\lambda+\kappa_i)^3}$$

$$E''(\lambda) = \sum_{i=1}^{r} \frac{+6\kappa_i^3 \gamma_i^2}{(1+\lambda\kappa_i)^4} \quad \Rightarrow \quad -\frac{1}{\lambda^5} E''\left(\frac{1}{\lambda}\right) = \sum_{i=1}^{r} \frac{-6\kappa_i^3 \gamma_i^2}{\lambda(\lambda+\kappa_i)^4}$$

and from this

$$I''(\lambda) = \sum_{i=1}^{r} \underbrace{\frac{6\kappa_i^2 \gamma_i^2}{\lambda(\lambda+\kappa_i)^3}}_{>0} \underbrace{\left[1 - \frac{\kappa_i}{\lambda+\kappa_i}\right]}_{>0} > 0.$$

This shows that the strictly monotonic decreasing function $I : (0, \infty) \to \mathbb{R}$, which has a unique zero under condition (3.75), also is strictly convex. Newton's method will therefore produce a sequence

$$\lambda_{k+1} = \lambda_k - \frac{I(\lambda_k)}{I'(\lambda_k)}, \quad I'(\lambda) = -\frac{2}{\lambda^3} x_{1/\lambda}^T L^T L (A^T A + \frac{1}{\lambda} L^T L)^{-1} L^T L x_{1/\lambda},$$

converging monotonously to this zero, when started with a small, positive value λ_0. The iteration will be stopped as soon as $\lambda_{k+1} \leq \lambda_k$, which can only happen due to finite precision arithmetic.

Random Data Perturbations

Often data perturbations are modelled by random variables and can not be bounded in the form $\|b - b^\delta\|_2 \leq \delta$. We will only consider a very special case, where each component b_i^δ is considered a realization of a random variable B_i with normal distribution having mean value b_i and standard deviation δb_i. Moreover, $B_1, \ldots, B_m$ are assumed to be stochastically independent. Then,

$$Z_i := \frac{B_i - b_i}{\delta b_i}, \quad i = 1, \ldots, m,$$

are random variables with standard normal distribution (mean value 0 and standard deviation 1) and therefore

$$X := \sum_{i=1}^{m} Z_i^2$$

is a random variable with χ^2 distribution and m degrees of freedom. Consequently, X has mean value m and standard deviation $\sqrt{2m}$. Setting

$$W := \begin{pmatrix} 1/\delta b_1 & 0 & \ldots & 0 \\ 0 & 1/\delta b_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & 1/\delta b_m \end{pmatrix},$$

the number

$$\|W(b^\delta - b)\|_2^2 = \sum_{i=1}^{m} \left(\frac{b_i^\delta - b_i}{\delta b_i} \right)^2$$

can be considered a realization of the random variable X. In this case, we compute x_λ as minimizer of

$$\|W(b^\delta - Ax)\|_2^2 + \lambda \|Lx\|_2^2 \tag{3.80}$$

instead of solving (3.64), thereby determining λ according to the discrepancy principle with $\delta = \sqrt{m}$. This means we find λ such that the corresponding solution x_λ of (3.80) has the property

$$\|W(b^\delta - Ax_\lambda)\|_2 = \sqrt{m}\,. \tag{3.81}$$

To do so, we need to know or estimate the standard deviation δb_i associated with each measurement.

Example 3.27 (Numerical differentiation) Let us reconsider Example 3.19. Discrete measurement values $b \in \mathbb{R}^m$ are perturbed by adding to each component a random number drawn from a normal distribution with mean value 0 and standard deviation $\sigma = 10^{-3}$. The term $\|b - b^\delta\|_2^2$ then has an expected value of $m\sigma^2$, see above. We apply the discrepancy principle with $\delta := \sqrt{m}\sigma \approx \|b - b^\delta\|_2$. The results for two randomly generated data sets are shown in Fig. 3.6 (exact u^* in red, reconstruction u_n in black). The parameter λ was determined numerically by finding a zero of function I defined in (3.79), using Newton's method. The quality of the reconstruction evidently depends on the data b^δ. The reconstructed functions u_n flatten at the boundaries of their domains of definition, where the reconstruction error becomes largest. On the one hand, this is due to using $\|Lx\|_2$ as an inverse measure of smoothness, because this measure penalizes functions having steep slopes. On the other hand it is inherently due to the boundary position: the difference in the derivatives of u_n and u^* would finally lead to increasing differences in function values and consequently to increasing data errors $Ax - b^\delta$, which would be inhibited by our minimization, but this effect can only show up in the long run and not at the boundary. ◇

Example 3.28 (Linearized seismic tomography) Let us apply Tikhonov regularization and discrepancy principle to Problem 1.12 of linearized seismic tomography, discretized by the collocation method as detailed in Sect. 2.3. We repeat the main points. We have to solve the convolution equation

$$w(t) = -\frac{1}{\sigma_0} \int_0^{x_0} g(t - 2s)u(s)\, ds, \tag{2.29}$$

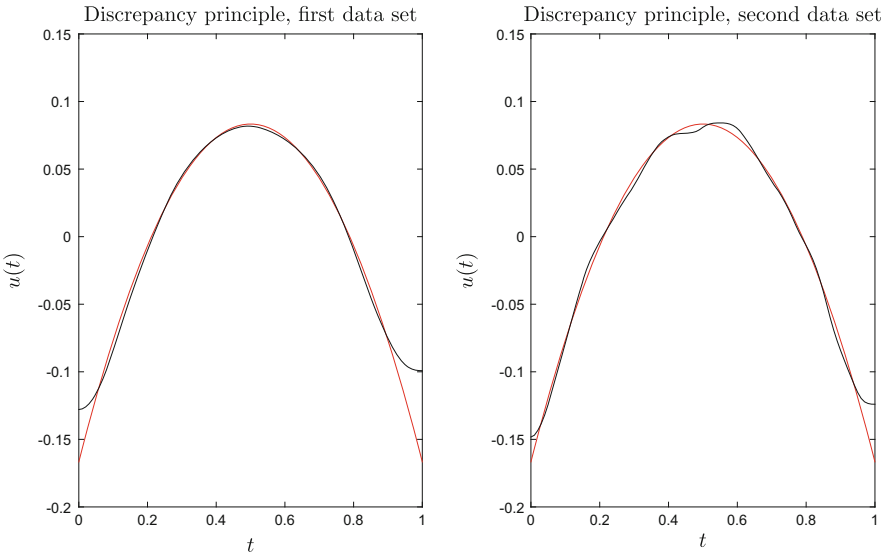

Fig. 3.6 Tikhonov regularization and discrepancy principle

where we choose $X_0 = T_0/2 = 1/2$, $\sigma_0 = 1$, and g as the Ricker pulse function defined in (2.30) in Example 2.7 (same parameters a and f_0). The function w is chosen such that the quadratic polynomial

$$u^* : \left[0, \frac{1}{2}\right] \to \mathbb{R}, \quad t \mapsto 2t(1-2t) \tag{2.31}$$

is the exact solution of (2.29). We assume that for $m \in \mathbb{N}$ collocation points $t_i := iT_0/m$, $i = 1, \ldots, m$,

$$\text{measurements} \quad b_i^\delta \quad \text{of the true sample values} \quad b_i := w(t_i) \tag{3.82}$$

are available. We choose $n \in \mathbb{N}$ and define $h := X_0/n$ and $\tau_j := jh$, $j = 1, \ldots, n$. An approximant u_n of u^* is sought in the space

$$X_n := \mathscr{S}_2(\tau_1, \ldots, \tau_n) = \langle N_{1,2}, \ldots, N_{n,2} \rangle,$$

i.e. having the form

$$u_n = \sum_{j=1}^{n} x_j \varphi_j, \quad \varphi_j = N_{j,2},\ j = 1, \ldots, n.$$

This leads to a system of linear equations $Ax = b^\delta$ with $A \in \mathbb{R}^{m,n}$ having components

$$A_{i,j} = -\frac{1}{\sigma_0} \int_0^{X_0} g(t_i - 2s) N_{j,2}(s)\, ds, \quad i = 1, \ldots, m,\ j = 1, \ldots, n, \tag{2.58}$$

which were computed by exact integration. The values b_i^δ were simulated by incrementing b_i by random numbers drawn from a normal distribution with zero mean and standard deviation $\sigma = 10^{-5}$, leading to $\|b - b^\delta\|_2 \approx \sqrt{m}\sigma =: \delta$. We found x^δ and a corresponding approximant

$$u_n^\delta = \sum_{j=1}^{n} x_j^\delta N_{j,2}$$

by solving the regularized linear least squares problem (3.64), determining λ according to the discrepancy principle with $\delta = \sqrt{m}\sigma$. In Fig. 3.7 we show the results (u^* in red, u_n^δ in black) obtained for $n = m = 100$ and for two choices of L, namely $L = I_n$ (identity matrix) and $L = L_2$ as in (3.39) ("second derivatives"). The second choice is the better one in this case and demonstrates the potential of regularization. It also shows that care must be taken concerning a good choice of the regularization term, since this will drive the reconstruction into a specific

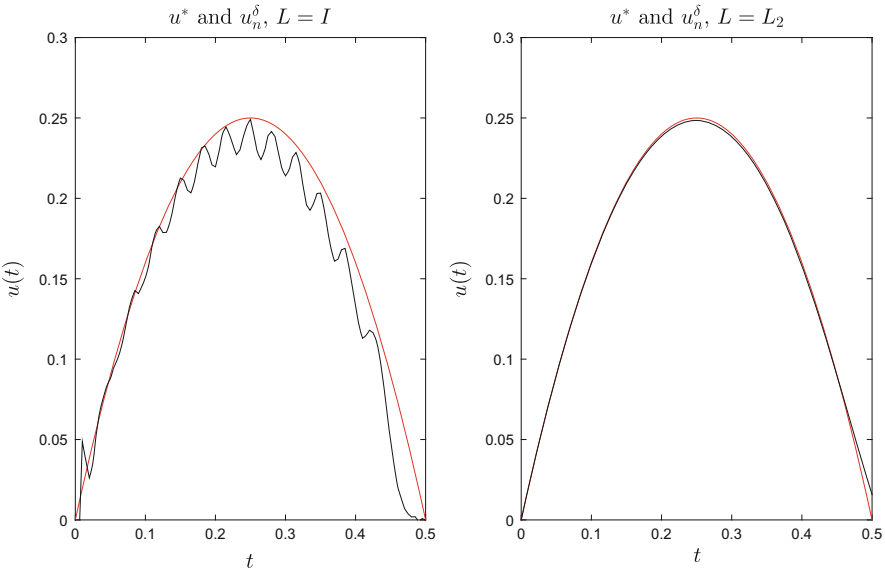

Fig. 3.7 Regularized linear seismic tomography, collocation method

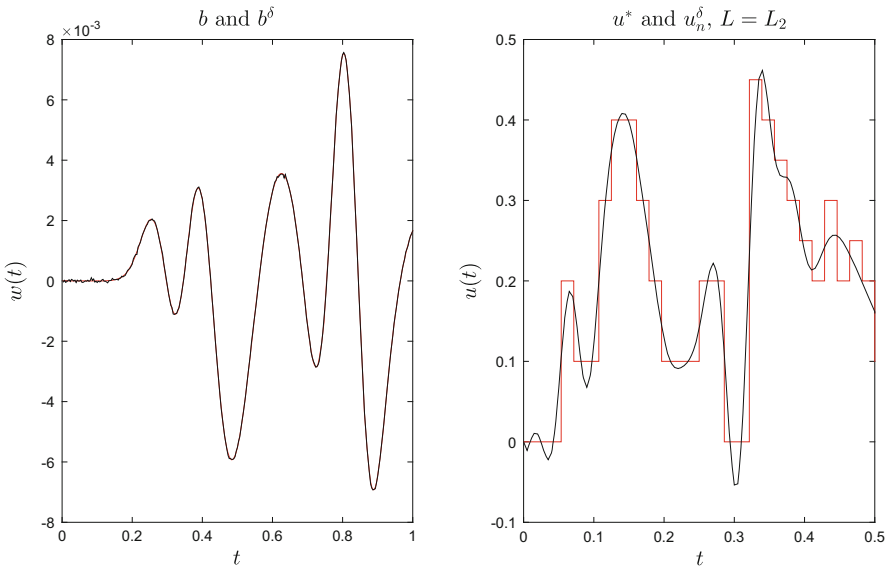

Fig. 3.8 Regularized linear seismic tomography, collocation method, $\sigma = 10^{-2}$

direction. Computing an unregularized solution is completely out of question here, the condition number of A being way too large ($\kappa_2(A) > 10^{20}$). ◇

Example 3.29 (Linearized seismic tomography) Example 3.28 is repeated, but this time with a (much less smooth) true solution $u^* : [0, \frac{1}{2}] \to \mathbb{R}$, defined as the step function shown in red in the right parts of Figs. 3.8 and 3.9. Samples b_i were perturbed by adding realizations of independent normally distributed random variables with zero mean and standard deviation σ. The reconstruction was performed using Tikhonov regularization with the regularization term based on $L = L_2$ defined by (3.39) as in Example 3.28. The regularization parameter was determined by the discrepancy principle. Figure 3.8 shows the exact and perturbed seismogram (in red and in black, respectively, in the left picture) and the exact and the reconstructed solution (in red and in black, respectively, in the right picture) for $\sigma = 10^{-2}$. Figure 3.9 shows the same for $\sigma = 10^{-1}$. In the presence of jumps in the true solution function, it is often recommended to base regularization on the total variation as an inverse smoothness measure and not on the norm of second derivatives, as formulated in (3.66) and (3.68). Since this leads to a nonlinear optimization problem, we will only come back to this choice in the next chapter. ◇

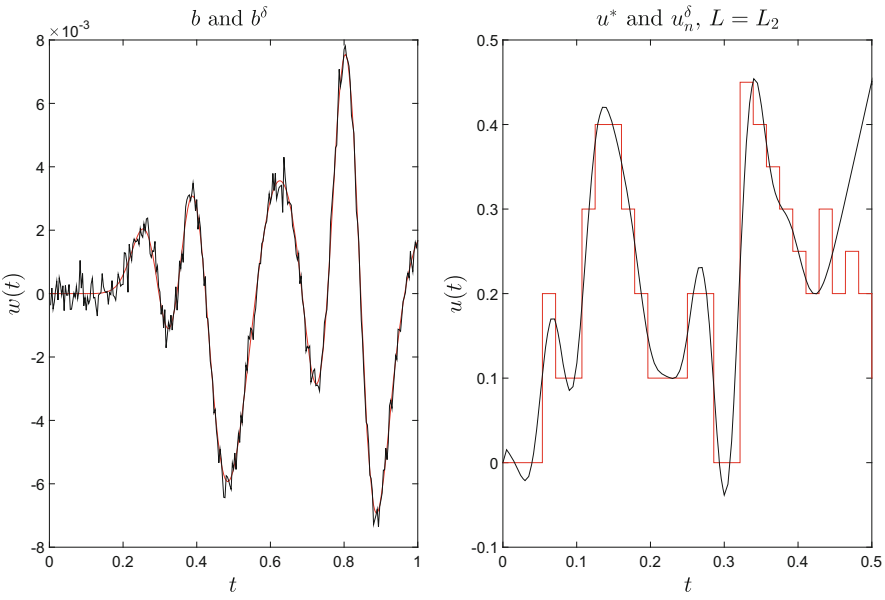

Fig. 3.9 Regularized linear seismic tomography, collocation method, $\sigma = 10^{-1}$

Other Heuristics to Determine Regularization Parameters

There are many other heuristics and recommendations of how to choose a regularization parameter λ in (3.64). Examples include the **generalized cross validation** ([GHW79]) or the **L-curve criterion** ([Han92]). For each of these criteria one can find examples where they deliver better results than their competitors, but no one is best in all cases.

3.6 Reduction of Least Squares Regularization to Standard Form

Tikhonov regularization as considered in Problem 3.24 with parameter λ determined by Morozov's discrepancy principle was seen to solve the constrained minimization problem formulated in Problem 3.22:

$$\min_{x \in M} \|Lx\|_2, \quad M := \{x \in \mathbb{R}^n;\ \|Ax - b\|_2 \leq \delta\}, \tag{3.83}$$

where $\delta > 0$. Under the assumptions of Theorem 3.23, especially if the constraint $\|Ax - b\|_2 \leq \delta$ is binding, a unique solution of (3.83) exists. It is determined as the unique solution x_λ of $(A^T A + \lambda L^T L)x = A^T b^\delta$ with $\|b^\delta - Ax_\lambda\|_2 = \delta$. As shown by

[Eld82], (3.83) can be transformed into an equivalent problem with $L = I_n$. In the following theorem, the weighted pseudoinverse L_A^+ is used, which was introduced in Definition 3.11.

Theorem 3.30 *Let $A \in \mathbb{R}^{m,n}$ and $L \in \mathbb{R}^{p,n}$ with $p \le n \le m$ and*

$$\mathcal{N}_A \cap \mathcal{N}_L = \{0\}. \tag{3.84}$$

Define

$$\begin{aligned} B &= AL_A^+, & P &= P_{\mathcal{N}_L} = I_n - L^+L, \\ Q &= I_m - (AP)(AP)^+, & c &= Qb, \end{aligned} \tag{3.85}$$

and assume that

$$\|(I_m - AA^+)b\|_2 \le \delta < \|Qb\|_2 \tag{3.86}$$

holds. Then problem (3.83) has a unique solution of the form

$$x^* = L_A^+\hat{x} + (AP)^+b, \tag{3.87}$$

where $\hat{x} \in \mathbb{R}^p$ is the unique solution of the constrained least squares problem

$$\min_{z \in N} \|z\|_2, \quad N := \{z \in \mathbb{R}^p;\ \|Bz - c\|_2 = \delta\}. \tag{3.88}$$

Conversely, if $x^* \in \mathbb{R}^n$ is the unique solution of (3.83), then $z = Lx^*$ is the unique solution of (3.88).

Proof Any vector $x \in \mathbb{R}^n$ can be decomposed into its projections on $\mathcal{N}_L^\perp$ and $\mathcal{N}_L$, namely

$$x = L^+Lx + Px = L^+z + Px, \quad \text{with } z := Lx. \tag{3.89}$$

Q as in (3.85) is the orthogonal projector from $\mathbb{R}^m$ onto $\mathcal{R}_{AP}^\perp$ and $\tilde{Q} := I_m - Q$ is the orthogonal projector onto $\mathcal{R}_{AP}$. From Pythagoras' theorem one gets

$$\|Ax - b\|_2^2 = \underbrace{\|Q(Ax - b)\|_2^2}_{=:\,\tau_1} + \underbrace{\|\tilde{Q}(Ax - b)\|_2^2}_{=:\,\tau_2}.$$

Since $QAP = AP - AP(AP)^+AP = 0$ (by virtue of the Moore-Penrose axioms), one concludes from (3.89) and from Definition 3.11 that

$$\begin{aligned} \tau_1 &= \|QAL^+z + QAPx - Qb\|_2^2 \\ &= \|A(I_n - P(AP)^+A)L^+z - Qb\|_2^2 = \|AL_A^+z - Qb\|_2^2 \end{aligned} \tag{3.90}$$

(using the identity $P(AP)^+ = (AP)^+$, which follows similarly as the symmetric identity $P_{\mathcal{N}_A}(LP_{\mathcal{N}_A})^+ = (LP_{\mathcal{N}_A})^+$ at the end of the proof of Theorem 3.10). Since $\tilde{Q}AP = (AP)(AP)^+AP = AP$ (by virtue of the Moore-Penrose axioms) and since $P = P^2$, one concludes from (3.89) that

$$\tau_2 = \|AP(Px) - \tilde{Q}(b - AL^+z)\|_2^2. \tag{3.91}$$

Further on

$$\|Lx\|_2^2 \overset{(3.89)}{=} \|LL^+Lx + LPx\|_2 = \|z\|_2,$$

where the Moore-Penrose axioms were used once more. Observe from formulae (3.90) and (3.91), that τ_1 depends on the component L^+z of x alone but not on Px, whereas τ_2 depends on both orthogonal components L^+z and Px of x. We have found that problem (3.83) is equivalent to

$$\min_{z\in\mathbb{R}^p} \|z\|_2 \quad \text{subject to} \quad \tau_1 = \|Bz - c\|_2^2 \le \delta^2 - \tau_2, \tag{3.92}$$

where equivalence precisely means the following. If x is a solution of (3.83), then $z = Lx$ is a solution of (3.92), the constraint being determined also by Px. If, conversely, z and Px are chosen such that $\|z\|_2$ becomes minimal under the constraints in (3.92), then $x = L^+z + Px$ is a solution of (3.83). Because of (3.84),

$$Px = (AP)^+(b - AL^+z) \tag{3.93}$$

is the unique choice of Px that leads to $\tau_2 = 0$ (verified by direct computation). Since $\min \|z\|_2$ is decreasing with the right hand side of the inequality constraint in (3.92) growing, one can state that (3.83) is equivalent to (3.92) with $\tau_2 = 0$.

Since $\|AA^+b - b\|_2 \le \|Ax - b\|_2$ for all $x \in \mathbb{R}^n$, requiring $\|(I_m - AA^+)b\|_2 \le \delta$ is necessary for a solution of (3.83) to exist at all. Requiring $\delta < \|Qb\|$ means that the constraint in (3.83) is binding, since for every minimizer of $\|Lx\|_2$, i.e. for every $x \in \mathcal{N}_L$, one has

$$\min_{x\in\mathcal{N}_L}\{\|Ax - b\|_2\} = \min_{x\in\mathbb{R}^n}\{\|APx - b\|_2\} = \|AP(AP)^+b - b\|_2 = \|Qb\|_2.$$

Thus from (3.86) it follows by Theorem 3.23, that a unique solution x^* of (3.83) exists, which is located at the boundary of the feasible region, i.e. $\|Ax^* - b\|_2 = \delta$ holds. Since x^* is unique, so must be the solution of (3.92) with $\tau_2 = 0$ and with the constraint in (3.92) replaced by $\|Bz - c\|_2 = \delta$.

From (3.89), from (3.93), and from Definition 3.11, one concludes that

$$x^* = L^+z + (AP)^+(b - AL^+z) = (I_n - (AP)^+A)L^+z + (AP)^+b = L_A^+z + (AP)^+z,$$

where z is the solution of (3.88). □

By virtue of Theorem 3.30, the solution of (3.83) can be found by solving problem (3.88), which is in standard form. This requires computation of L_A^+, AL_A^+, $(AP)^+$, and Q. In [Eld77], a method is described to achieve this, if one assumes

$$\operatorname{rank}(L) = p, \tag{3.94}$$

i.e. that matrix L has full rank. If $\operatorname{rank}(L) = p = n$, we get $L^+ = L^{-1}$ and $P = 0$, such that

$$L_A^+ = L^{-1}, \quad B = AL^{-1}, \quad (AP)^+ = 0, \quad \text{and} \quad I_m - (AP)(AP)^+ = I_m \tag{3.95}$$

in Theorem 3.30. If $\operatorname{rank}(L) = p < n$, one starts with a QR-decomposition

$$L^T = V \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} p \\ \} n\text{-}p \end{matrix}, \qquad V = (\underbrace{V_1}_{p} \underbrace{V_2}_{n-p}) \in \mathbb{R}^{n,n}, \tag{3.96}$$

where V is orthogonal and R is upper triangular and non-singular because of (3.94). One then computes a QR-decomposition of AV_2:

$$AV_2 = Q \begin{pmatrix} U \\ 0 \end{pmatrix} \begin{matrix} \} n\text{-}p \\ \} m\text{-}n\text{+}p \end{matrix}, \qquad Q = (\underbrace{Q_1}_{n-p} \underbrace{Q_2}_{m-n+p}) \in \mathbb{R}^{m,m}, \tag{3.97}$$

where Q is orthogonal (this is not the same matrix Q as in Theorem 3.30) and $U \in \mathbb{R}^{n-p,n-p}$ is upper triangular and non-singular, because otherwise AV_2 could not have full rank $n-p$. (To see that AV_2 has full rank, assume $AV_2x = 0$ for some x. Then, $V_2x \in \mathcal{N}_A$, but at the same time $V_2x \in \mathcal{N}_L$ by (3.96). This implies $V_2x = 0$, since $\mathcal{N}_A \cap \mathcal{N}_L = \{0\}$, and then implies $x = 0$, since V_2 has full rank).

Lemma 3.31 *Let $A \in \mathbb{R}^{m,n}$ and $L \in \mathbb{R}^{p,n}$ with $p < n \leq m$ and let (3.84) and (3.94) hold. Let P be defined according to (3.85), and let V, R, Q, and U be computed according to (3.96) and (3.97). Then*

$$\begin{aligned} L_A^+ &= (V_1 - V_2U^{-1}Q_1^TAV_1)R^{-T}, \\ AL_A^+ &= Q_2Q_2^TAV_1R^{-T}, \\ (AP)^+ &= V_2U^{-1}Q_1^T, \text{ and} \\ I_m - (AP)(AP)^+ &= Q_2Q_2^T. \end{aligned}$$

If $\operatorname{rank}(L) = p = n$, then the above matrices are given by (3.95).

Proof From Definition 3.11 one knows that

$$L_A^+ = (I_n - (AP)^+A)L^+, \quad P = P_{\mathcal{N}_L}.$$

We compute the individual terms in this expression. First, from (3.96) one deduces

$$L = R^T V_1^T \quad \Longrightarrow \quad L^+ = V_1 R^{-T}$$

by the Moore-Penrose axioms. Second, $P = V_2 V_2^T$ and $(AP)^+ = (AV_2 V_2^T)^+ = V_2 (AV_2)^+$, again easily verified by the Moore-Penrose axioms. Third,

$$AV_2 = Q_1 U \quad \Longrightarrow \quad (AV_2)^+ = U^{-1} Q_1^T,$$

once more by the Moore-Penrose axioms. This proves the first and, by the way, the third equality. Further, $(AP)(AP)^+ = Q_1 Q_1^T$, which shows the last equality. Concerning the second equality, observe that

$$AL_A^+ = (A - AV_2 U^{-1} Q_1^T A) V_1 R^{-T} = (I_m - Q_1 Q_1^T) AV_1 R^{-T}$$

and that $I_m - Q_1 Q_1^T = Q_2 Q_2^T$. □

In practice, L often is a sparse, banded matrix, like in (3.34) and (3.39). Moreover, $n-p$ usually is a small number. Then, (3.96) and (3.97) can be computed efficiently. The decompositions (3.96) and (3.97) can also be used to transform (3.64) for any $\lambda > 0$ into standard form, as formulated in the following

Theorem 3.32 *Let $A \in \mathbb{R}^{m,n}$ and $L \in \mathbb{R}^{p,n}$ with $p < n \leq m$ and let (3.84) and (3.94) hold. Let P be defined according to (3.85), and let $V = (V_1, V_2)$, R, $Q = (Q_1, Q_2)$, and U be computed according to (3.96) and (3.97). Let $\lambda > 0$ be arbitrary and let x_λ be the unique minimizer of*

$$\min \left\{ \|b - Ax\|_2^2 + \lambda \|Lx\|_2^2;\ x \in \mathbb{R}^n \right\}. \tag{3.98}$$

Then $z_\lambda = Lx_\lambda$ is the unique minimizer of

$$\min \left\{ \|\tilde{b} - \tilde{A}z\|_2^2 + \lambda \|z\|_2^2;\ z \in \mathbb{R}^p \right\}, \tag{3.99}$$

where

$$\tilde{A} := Q_2^T A V_1 R^{-T} \quad \text{and} \quad \tilde{b} = Q_2^T b. \tag{3.100}$$

Conversely, if z_λ is the unique solution of (3.99), then

$$x_\lambda = (V_1 - V_2 U^{-1} Q_1^T A V_1) R^{-T} z_\lambda + V_2 U^{-1} Q_1^T b \tag{3.101}$$

is the unique solution of (3.98). If $\text{rank}(L) = p = n$, the same equivalences hold, but with $\tilde{A} = AL^{-1}$ and $\tilde{b} = b$ in (3.100) and with $x_\lambda = L^{-1} z_\lambda$ in (3.101).

Proof The proof is taken from [Eld77]. Only the non-trivial case $p < n$ is considered. Problem (3.98) is equivalent to minimizing $\|r\|_2$, where

$$r = \begin{pmatrix} A \\ \sqrt{\lambda}L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

With (3.96) and the change of variables

$$x = Vy = V_1 y_1 + V_2 y_2, \tag{3.102}$$

the vector r becomes

$$r = \begin{pmatrix} AV_1 & AV_2 \\ \sqrt{\lambda}R^T & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

The first m components of r are multiplied by Q^T from (3.97). This defines a vector

$$\tilde{r} = \begin{pmatrix} \tilde{r}_1 \\ \tilde{r}_2 \\ \tilde{r}_3 \end{pmatrix} = \begin{pmatrix} Q_1^T AV_1 & U \\ Q_2^T AV_1 & 0 \\ \sqrt{\lambda}R^T & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} Q_1^T b \\ Q_2^T b \\ 0 \end{pmatrix}.$$

with $\|r\|_2 = \|\tilde{r}\|_2$. The components $\tilde{r}_2$ and $\tilde{r}_3$ are independent of y_2, which can always be chosen such that $\tilde{r}_1 = 0$. This means that (3.98) is equivalent to

$$\min \left\{ \left\| \begin{pmatrix} Q_2^T AV_1 \\ \sqrt{\lambda}R^T \end{pmatrix} y_1 - \begin{pmatrix} Q_2^T b \\ 0 \end{pmatrix} \right\|_2 ;\ y_1 \in \mathbb{R}^p \right\} \tag{3.103}$$

and

$$y_2 = U^{-1} Q_1^T (b - AV_1 y_1). \tag{3.104}$$

By one more change of variables

$$z = R^T y_1, \tag{3.105}$$

(3.103) takes the standard form

$$\min\{\|\tilde{A}z - \tilde{b}\|_2^2 + \lambda \|z\|_2^2;\ z \in \mathbb{R}^p\}$$

with $\tilde{A}$ and $\tilde{b}$ defined in (3.100). □

From Lemma 3.31 the following relations between the matrices B and $\tilde{A}$ and between the vectors c and $\tilde{b}$ defined in Theorems 3.30 and 3.32 are immediate:

$$B = Q_2 \tilde{A} \quad \text{and} \quad c = Q_2 \tilde{b}.$$

Since the columns of Q_2 are orthonormal, from Pythagoras' theorem it follows that

$$\|Bz - c\|_2 = \|\tilde{A}z - \tilde{b}\|_2 \quad \text{for all} \quad z \in \mathbb{R}^p,$$

so we may deliberately replace B by $\tilde{A}$ and c by $\tilde{b}$ in Theorem 3.30. Lemma 3.31 also shows that the transform (3.101) defined in Theorem 3.32 can be written in the form $x_\lambda = L_A^+ z_\lambda + (AP)^+ b$, as in Theorem 3.30.

Summary

Let all conditions of Theorems 3.30 and 3.32 hold and assume moreover, that $\operatorname{rank}(A) = n$, such that a unique minimizer $\hat{x}$ of $\|Ax - b\|_2$ exists. In this case, one also has $\operatorname{rank}(\tilde{A}) = p$,[5] such that a unique minimizer $\hat{z}$ of $\|\tilde{A}z - \tilde{b}\|_2$ exists.

It was seen that the function $x \mapsto z = Lx$ maps the solution of (3.98) to the solution of (3.99) *and also* maps the solution of (3.83) to the solution of (3.88). Conversely, for any $\lambda > 0$, the affine function

$$F_b : \mathbb{R}^p \to \mathbb{R}^n, \quad z \mapsto x = L_A^+ z + (AP)^+ b$$

maps the unique solution of (3.99) to the unique solution of (3.98). Since $\operatorname{rank}(A) = n$, the latter must also hold in the limit for $\lambda \to 0$, i.e. the unique minimizer $\hat{z}$ of $\|\tilde{A}z - \tilde{b}\|_2$ is mapped to the unique minimizer $\hat{x} = F_b(\hat{z})$ of $\|Ax - b\|_2$. *The same function* F_b also maps the unique solution of (3.88) to the unique solution of (3.83).

3.7 Regularization of the Backus-Gilbert Method

The method of Backus-Gilbert was introduced in Sect. 2.4 in the context of solving a Fredholm integral equation

$$w(t) = \int_a^b k(t,s)u(s)\,ds$$

under Assumption 2.10, when samples $y_i = w(t_i)$, $i = 1, \ldots, m$ are available. Technically, a parameter t has to be fixed, the matrix $Q(t)$ and the vector c defined

[5] Assume that $\tilde{A}x = Q_2^T A V_1 R^{-T} x = 0$ for some x. This means that $AV_1R^{-T}x$ belongs to the space spanned by the columns of Q_1, which is the space spanned by the columns of AV_2 [see (3.97)]. Therefore, there is some y such that $AV_1R^{-T}x = AV_2y$. Since A has full rank, this means $V_1R^{-T}x = V_2y$, but the spaces spanned by the columns of V_1 and V_2 are orthogonal. This means that $V_1R^{-T}x = 0$, and since V_1 and R^{-T} have full rank, this implies $x = 0$.

in (2.67) have to be computed, the vector $v \in \mathbb{R}^m$ (dependent on t) has to be found by solving the system (2.69), and a pointwise approximation $u_m(t) = \sum_{j=1}^m y_j v_j$ of the solution u^* at t has to computed according to (2.70). The process has to be repeated for every t, where a reconstruction is desired. We have already seen in Sect. 2.4, that the method will fail, if the averaging kernel

$$\varphi(s,t) = \sum_{i=1}^m v_i k_i(s) \quad \text{with} \quad k_i(s) = k(t_i, s)$$

fails to be a function (of s) sharply peaked at t and approximately zero elsewhere. Another concern is the ill-conditioning of $Q(t)$ which will show up for growing values of m and which requires an adequate regularization when solving (2.69). To describe a possible regularization, let us replace the linear system (2.69) by the equivalent least squares problem

$$\|Ax - b\|_2^2 = \min!, \quad A = A(t) = \begin{pmatrix} Q(t) & c \\ c^T & 0 \end{pmatrix}, \quad x = \begin{pmatrix} v \\ \mu \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{3.106}$$

with $A(t) \in \mathbb{R}^{m+1,m+1}$ and $x, b \in \mathbb{R}^{m+1}$ (b has m components equaling 0). To derive a regularization term, let us assume that instead of $y_i = w(t_i)$ we will observe perturbed samples

$$\tilde{y}_i = y_i + n_i = w(t_i) + n_i, \quad i = 1, \ldots, m,$$

where n_i are realizations of normally distributed random variables N_i with mean 0 and variance $\sigma_{ii}^2 > 0$. Double indices are used because we will designate the covariance of N_i and N_j by $\sigma_{ij}^2 \geq 0$. Based on perturbed data $\tilde{y}_i$, the reconstructed value becomes

$$\tilde{u}_m(t) := \sum_{i=1}^m v_i \tilde{y}_i$$

and can be interpreted as realization of a random variable with variance

$$v^T S v, \quad S := (\sigma_{ij}^2)_{i,j=1,\ldots,m} \in \mathbb{R}^{m,m}. \tag{3.107}$$

The so-called **covariance matrix** S is positive definite (it is a diagonal matrix if the random variables N_i are uncorrelated) and the variance $v^T S v$ is a measure *not* of the error $\tilde{u}_m(t) - u^*(t)$, but of how much the reconstructed value $\tilde{u}_m(t)$ is subject to change, if measurement data are perturbed. It can therefore be considered an inverse measure of the **stability** of the reconstruction. In contrast, the value $v^T Q v$ is an inverse measure of the "peakedness" of the averaging kernel $\varphi(s,t)$ at t and thus assesses the resolution achieved by the Backus-Gilbert reconstruction. One

approach to a regularized method of Backus-Gilbert consists in minimizing

$$v^T Q(t)v + \lambda v^T S v \text{ under the constraint } v^T c = 1, \tag{3.108}$$

for some suitable parameter λ. An explicit solution of (3.108) can be computed as in (2.68) with Q replaced by $Q + \lambda S$. An alternative approach is Tikhonov regularization of the linear system $Ax = b$ from (3.106). To this end, determine a Cholesky factorization $S = L_1^T L_1$, where L_1 is a lower triangular matrix – in case of uncorrelated random variables N_i, this is just a diagonal matrix with entries σ_{ii} on the diagonal. Then define

$$L := \begin{pmatrix} L_1 & 0 \end{pmatrix} \in \mathbb{R}^{m,m+1} \quad \Longrightarrow \quad \|Lx\|_2^2 = v^T S v \quad \text{for} \quad x = \begin{pmatrix} v \\ \mu \end{pmatrix}.$$

Finally define a regularized vector v as the first m components of a solution of

$$\min \left\{ \|b - Ax\|_2^2 + \lambda \|Lx\|_2^2;\ x \in \mathbb{R}^{m+1} \right\}. \tag{3.109}$$

To determine the parameter λ, one could devise some bound Δ and require $\|Lx\|_2 \leq \Delta$. This would mean that in fact we have to solve a constrained optimization problem like Problem 3.20:

$$\min_x \left\{ \|b - Ax\|_2^2 \right\} \quad \text{under the constraint} \quad \|Lx\|_2 \leq \Delta, \tag{3.110}$$

which was discussed in Sect. 3.4. Note that it is well possible to choose $\Delta = \Delta(t)$ in dependance of t.

Example 3.33 (Linearized seismic tomography) We take up linearized seismic tomography as in Example 2.12, but now we choose $m = 40$ sampling points and use perturbed values $\tilde{y}_i$ obtained from exact measurement values $y_i = w(t_i)$, $i = 1, \ldots, m$, by adding realizations n_i of independent normal random variables with zero mean and standard deviation $\sigma = 10^{-5}$. In Fig. 3.10 we show the results obtained by the regularized Backus-Gilbert method (3.109) for a fixed value $\lambda = 10^{-6}$ (u^* in red, u_m in black). We also show the measures for the achieved resolution and stability. Regularization can compensate the ill-conditioning of the matrix $Q(t)$ (without regularization, u_m would nowhere be an acceptable approximation of u^*). At the right boundary, the approximation quality remains poor, however. As explained in Example 2.12, the low resolution achieved by the kernel functions k_i at the right boundary is responsible for this. ◇

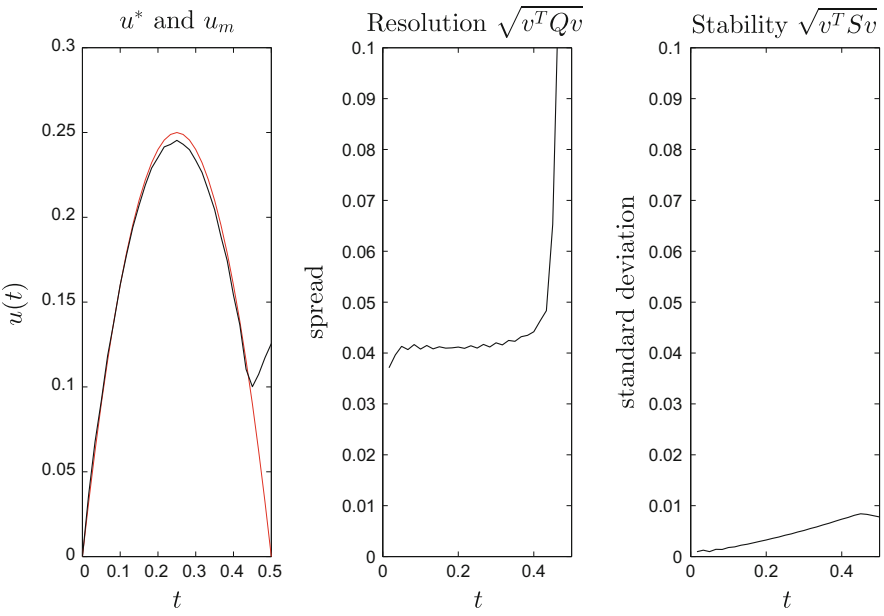

Fig. 3.10 Regularized linear seismic tomography, Backus-Gilbert method

3.8 Regularization of Fourier Inversion

A discrete Fourier inversion of convolutional, s-variate Fredholm equations

$$w = k * u, \quad u, k, w \in L_2(\mathbb{R}^s). \tag{2.82}$$

with exact solution u^* was presented in Sect. 2.5. Reusing the notation from this section, the method can be summarized in one line:

$$\{w_\alpha = w(x_\alpha)\}_\alpha \circ\!\!-\!\!\bullet \{W_\beta\}_\beta, \quad \left\{U_\beta = \frac{W_\beta}{\hat{k}(\beta/2a)}\right\}_\beta, \quad \{U_\beta\}_\beta \bullet\!\!-\!\!\circ \{u_\alpha\}_\alpha, \tag{3.111}$$

where $u_\alpha = u_N(x_\alpha) \approx u(x_\alpha)$ and where $\alpha, \beta \in W := \{-N/2, \ldots, N/2 - 1\}^s$ for an even integer $N \geq 2$. With an application to the linearized problem of inverse gravimetry in mind, we will focus on the bivariate case $s = 2$ from now on. Let us define

$$K_\beta := \hat{k}\left(\frac{\beta}{2a}\right), \quad \beta \in \mathbb{Z}^2.$$

The smoother the function k, the faster the values $|K_\beta|$ will decay to 0 for $|\beta| \to \infty$. In practice, we will not dispose of exact measurement values (samples) $w_\alpha = w(x_\alpha)$, $\alpha \in W$, but will have to work with perturbed data w_α^δ, $\alpha \in W$, assuming that for some $\delta > 0$

$$\sum_{\alpha \in W} \left| w_\alpha - w_\alpha^\delta \right|^2 = \delta^2. \tag{3.112}$$

Accordingly, the first step in (3.111) will produce perturbed values W_β^δ instead of W_β. Then, for $|K_\beta|$ close to 0, $U_\beta^\delta = W_\beta^\delta / K_\beta$ will contain a huge error $U_\beta^\delta - U_\beta$, such that carrying out the last step of (3.111) will produce a useless result. Tikhonov regularization suggests to replace the division step in (3.111) and determine U_β^δ by minimizing

$$\sum_{\beta \in W} \left| W_\beta^\delta - U_\beta^\delta K_\beta \right|^2 + \lambda \sum_{\beta \in W} \left| U_\beta^\delta \right|^2$$

for some value $\lambda \geq 0$ yet to be determined. The minimization can be carried out for each index β separately. A short calculation shows

$$\begin{aligned} &\left| W_\beta^\delta - U_\beta^\delta K_k \right|^2 + \lambda \left| U_\beta^\delta \right|^2 = \\ &= \left(\lambda + \left| K_\beta \right|^2 \right) \left| U_\beta^\delta - \frac{W_\beta^\delta \overline{K_\beta}}{\lambda + \left| K_\beta \right|^2} \right|^2 + \left| W_\beta^\delta \right|^2 - \frac{\left| W_\beta^\delta \overline{K_\beta} \right|^2}{\lambda + \left| K_\beta \right|^2} \end{aligned}$$

and this term becomes minimial for the choice

$$U_\beta^\delta = \frac{W_\beta^\delta \overline{K_\beta}}{\lambda + \left| K_\beta \right|^2}, \quad \beta \in W. \tag{3.113}$$

Note that for $\lambda = 0$ ("no regularization"), this formula becomes $U_\beta^\delta = W_\beta^\delta / K_\beta$, which is exactly the division step from (3.111). Choosing a value $\lambda > 0$ one can avoid division by frequencies close to zero. A discrete inverse Fourier transform of U_β^δ (the last step of (3.111)) will give coefficients u_α^δ, $\alpha \in W$, of a B-spline aproximant u^δ of u^*.[6] One can insert u^δ into (2.82) and sample *its* output – this gives the samples of a "simulated effect" $\tilde{w}_\alpha^\delta$. The discrepancy principle then suggests to

[6]The values u_α^δ in general are complex numbers. To derive a real valued approximant of u^*, we just take the real parts of u_α^δ.

choose a maximal value λ, such that

$$\sum_{\alpha\in W} |w_\alpha^\delta - \tilde{w}_\alpha^\delta|^2 \le \delta^2.$$

To apply this criterion directly in the frequency domain, observe that, starting from (3.112), one gets (with $h = 2a/N$):

$$\begin{aligned}
\delta^2 &= \frac{1}{h^2}\cdot h^2 \cdot \sum_{\alpha\in W} \left|w_\alpha - w_\alpha^\delta\right|^2 \overset{(2.100)}{\approx} \frac{1}{h^2}\int_{\mathbb{R}^2} \left|w_N(x) - w_N^\delta(x)\right|^2 dx \\
&\overset{(C.6)}{=} \frac{1}{h^2}\int_{\mathbb{R}^2} \left|\widehat{w_N}(y) - \widehat{w_N^\delta}(y)\right|^2 dy \approx \frac{1}{h^2}\cdot\frac{1}{(2a)^2}\cdot\sum_{\beta\in W}\left|\widehat{w_N}\left(\frac{\beta}{2a}\right) - \widehat{w_N^\delta}\left(\frac{\beta}{2a}\right)\right|^2 \\
&\overset{(2.101)}{=} \frac{N^2}{(2a)^4}\sum_{\beta\in W}\sigma_\beta^2\left|W_\beta - W_\beta^\delta\right|^2. \qquad (3.114)
\end{aligned}$$

Typically, errors contain significant high frequency components, which are neglected in the last finite sum. Thus it is to be expected that (3.114) somewhat underestimates the error. From (3.114), $\lambda \ge 0$ will be determined such that

$$\begin{aligned}
S(\lambda) &:= N^2\sum_{\beta\in W}\left(\frac{\sigma_\beta}{4a^2}\right)^2\left|U_\beta^\delta K_\beta - W_\beta^\delta\right|^2 \\
&\overset{(3.113)}{=} N^2\sum_{\beta\in W}\left(\frac{\sigma_\beta}{4a^2}\right)^2\left|W_\beta^\delta\right|^2\left(\frac{\lambda}{\lambda + |K_\beta|^2}\right)^2 = \delta^2.
\end{aligned}$$

This nonlinear equation could be solved using Newton's method. It is preferable to set $\mu := 1/\lambda$ and determine a zero of the function $T : (0,\infty) \to \mathbb{R}$ defined by

$$T(\mu) := S(\lambda) = N^2\sum_{\beta\in W}\left(\frac{\sigma_\beta}{4a^2}\right)^2\left|W_\beta^\delta\right|^2\left(\frac{1}{1+\mu|K_\beta|^2}\right)^2, \qquad (3.115)$$

since this function is monotonic decreasing and convex, as can be checked by computing its first and second derivative (compare [Hof99], p. 141). Consequently, Newton's method applied to the equation $T(\mu) = \delta^2$ will converge monotonously, if

$$0 < \delta^2 < N^2\sum_{\beta\in W}\left(\frac{\sigma_\beta}{4a^2}\right)^2\left|W_\beta^\delta\right|^2 \approx \sum_{\alpha\in W}|w_\alpha|^2.$$

Application to Linear Inverse Gravimetry

Problem 1.10 concerning linearized inverse gravimetry was formulated in Sect. 1.3. The goal is to find the Mohorovičić discontinuity modeled by a function $\Delta u : [-a,a]^2 \to \mathbb{R}$, which is implicitly determined as the solution of a convolutional equation (1.30). Continuing Δu by zero values to become a function $u : \mathbb{R}^2 \to \mathbb{R}$ and setting $w := f/u_0$, equation (1.30) can be written as a convolutional equation of the form (2.82), namely

$$w(x) = \int_{\mathbb{R}^2} k(x-y)u(y)\,dy, \quad x \in [-b,b]^2. \tag{3.116}$$

The function k is known, see (1.31), and can be Fourier transformed analytically:

$$\hat{k}(y_1,y_2) = \frac{2\pi}{u_0} \mathrm{e}^{-2\pi u_0 \sqrt{y_1^2+y_2^2}}. \tag{2.84}$$

Example 3.34 (Linear inverse gravimetry) Let $a = b = 4$ and $N = 64$. An effect w was simulated for the step function u shown in Fig. 3.11 and sampled at $x_\alpha = (h\alpha_1, h\alpha_2)$, $\alpha \in W$, $h = 2a/N$. To the samples $w_\alpha := w(x_\alpha)$ we added realizations of independent Gaussian random variables with zero mean and standard deviation σ, leading to perturbed data w_α^δ, $\alpha \in W$, and a value $\delta = N\sigma$ in (3.112). A regularized Fourier reconstruction as described by (3.111) and (3.113) was carried out, with the regularization parameter λ determined according to Morozov's discrepancy principle, see (3.114) and (3.115). Figure 3.12 shows the result we

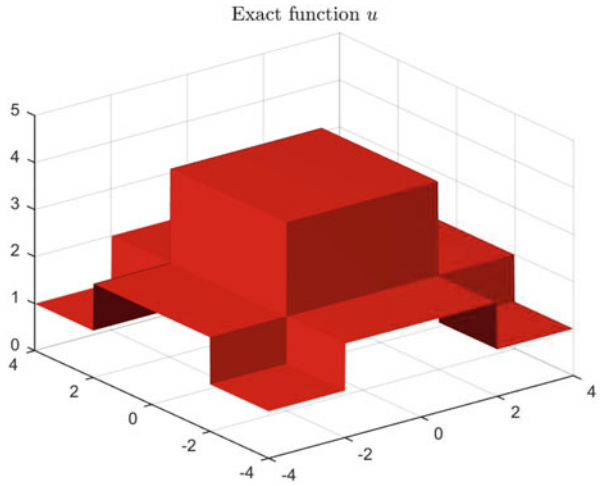

Fig. 3.11 Step function taking values 1, 2, and 4

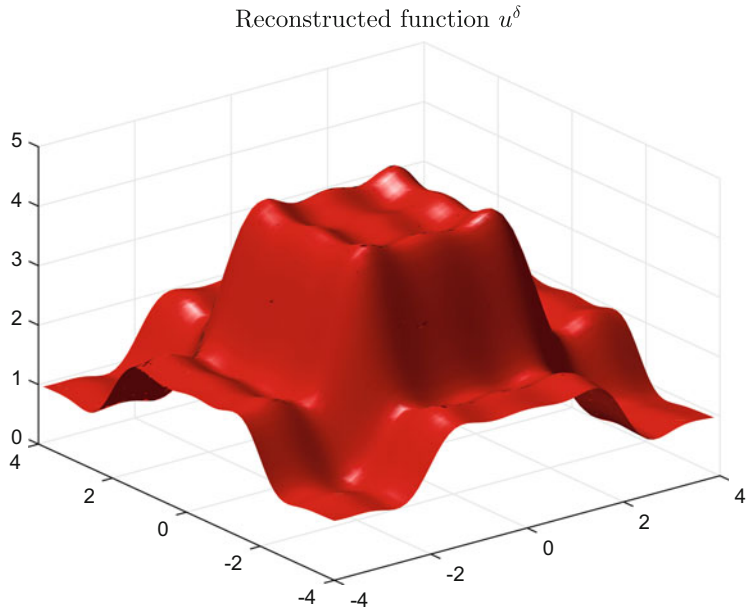

Fig. 3.12 Reconstructed step function, $\sigma = 10^{-2}$

obtained for $\sigma = 10^{-2}$ and Fig. 3.13 shows the result we obtained for $\sigma = 10^{-5}$, illustrating the convergence of Fourier reconstruction. ◇

It should be mentioned that alternative approaches exist for Fourier reconstruction. A famous example is **Wiener filtering**, which is advocated in [Sam11]. For a readable introduction into Wiener filtering, see Section 13.3 of [PTVF92]. Wiener filtering relies on a stochastic modelling of data perturbations and requires more knowledge about the noise than we assumed in (3.112).

3.9 Landweber Iteration and the Curve of Steepest Descent

Assuming that the matrix $A \in \mathbb{R}^{m,n}$ has full rank n, the least squares problem

$$\min_{x \in \mathbb{R}^n} F(x), \quad F(x) := \frac{1}{2}\|b - Ax\|_2^2, \tag{3.117}$$

has a unique solution $\hat{x}$, which can (theoretically) be found by solving the normal equations $A^T Ax = A^T b$, or by using a QR decomposition or an SVD of A. When only an approximation $b^\delta \approx b$ is available and when A is badly conditioned, the system $A^T Ax = A^T b^\delta$ must be replaced by a regularized variant, see Sects. 3.4 and 3.5. Alternatively, an iterative optimization algorithm can be used to solve (3.117)

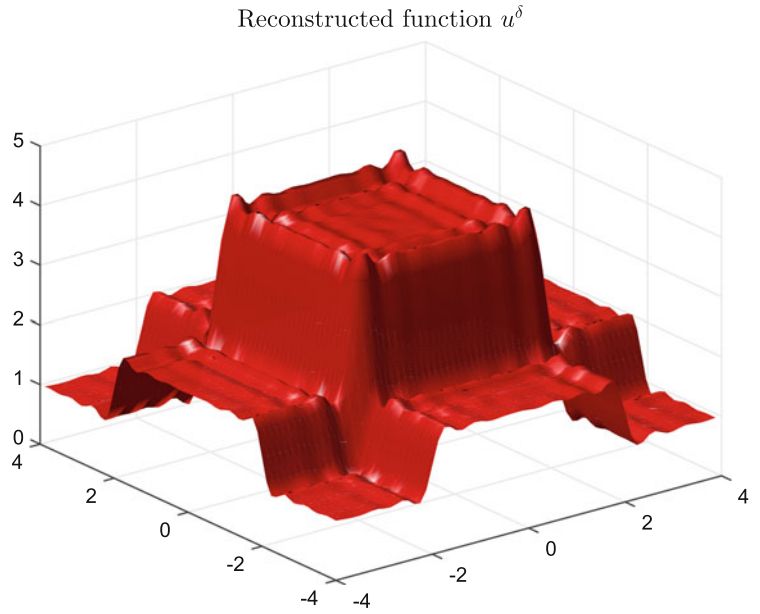

Fig. 3.13 Reconstructed step function, $\sigma = 10^{-5}$

directly. If only perturbed data b^δ are available, such an algorithm would search for the minimizer $\bar{x}$ of

$$\min_{x\in\mathbb{R}^n} F^\delta(x), \quad F^\delta(x) := \frac{1}{2}\|b^\delta - Ax\|_2^2. \tag{3.118}$$

If A is badly conditioned, $\bar{x}$ may again be far away from $\hat{x}$. It will be seen that the iteration has to be stopped prematurely in order to obtain a regularized approximation of $\hat{x}$.

The simplest optimization method for (3.118) is the **method of steepest descent**, which is called **Landweber iteration** in the context of linear least squares problems. It is started with some initial guess $x^0 \in \mathbb{R}^n$ of $\bar{x}$, for example $x^0 = 0$. Then, one iteratively computes

$$x^{k+1} := x^k - s_k \nabla F^\delta(x^k), \quad k = 0, 1, 2, \ldots,$$

where $s_k > 0$ is a **step size** to be chosen. If the step size is chosen to be constant, i.e. $s_k = s$ for all k, then this iteration, applied to (3.118), becomes

$$x^{k+1} = x^k - sA^T(Ax^k - b^\delta) = (I - sA^TA)x^k + sA^Tb^\delta. \tag{3.119}$$

For the choice $x^0 = 0$ one gets an explicit formula for the kth iterate:

$$x^k = s\sum_{j=0}^{k-1}(I - sA^TA)^jA^Tb^\delta, \quad k = 1, 2, \ldots \tag{3.120}$$

This can be written as $x^k = R_k b^\delta$, where $(R_k)_{k\in\mathbb{N}}$ is the sequence of matrices

$$R_k = s\sum_{j=0}^{k-1}(I - sA^TA)^jA^T \quad \in \quad \mathbb{R}^{m,n}. \tag{3.121}$$

Using the reduced SVD $A = \hat{U}\hat{\Sigma}V^T$ (see (A.2)), a short calculation shows that

$$R_k y = \sum_{i=1}^{n}\frac{[1 - (1 - s\sigma_i^2)^k]}{\sigma_i}\cdot(u_i^T y)\cdot v_i, \quad y \in \mathbb{R}^m. \tag{3.122}$$

Here, u_i and v_i, $i = 1, \ldots, n$, are the columns of $\hat{U}$ and V, respectively, and σ_i, $i = 1, \ldots, n$, are the singular values of A. The vector $x^k = R_k b^\delta$ found after k steps of the Landweber iteration applied to (3.118) has to be compared with the vector one is actually interested in, namely

$$\hat{x} = A^+ b = \sum_{i=1}^{n}\frac{1}{\sigma_i}\cdot(u_i^T b)\cdot v_i. \tag{3.123}$$

A comparison with (3.122) shows, that R_k is an approximation of the matrix (operator) A^+, the factors $1/\sigma_i$ being modified by multiplication with attenuation factors $q(k, \sigma_i) = 1 - (1 - s\sigma_i^2)^k$. Under the condition

$$0 < s \le \frac{1}{\sigma_1^2} = \frac{1}{\|A\|_2^2} \tag{3.124}$$

one gets $q(k, \sigma) \to 1$ for $k \to \infty$ and for $0 < \sigma \le \sigma_1 = \|A\|_2$. Consequently,

$$\|R_k y - A^+ y\|_2 \to 0 \quad \text{for} \quad k \to \infty \quad \text{and for all} \quad y \in \mathbb{R}^m,$$

if (3.124) holds. This shows that $(R_k)_{k\in\mathbb{N}}$ is converging to A^+, but it does not yet show the regularizing properties of the Landweber iteration. From (3.124) one deduces $0 < q(k, \sigma) \le 1$ for all $k \in \mathbb{N}$ and for $0 < \sigma \le \sigma_1$. From Bernoulli's inequality one gets

$$1 - (1 - s\sigma^2)^k \le \sqrt{1 - (1 - s\sigma^2)^k} \le \sqrt{1 - (1 - ks\sigma^2)} = \sqrt{ks}\sigma,$$

and if this estimate is used in (3.122), one sees $\|R_k\|_2 \leq \sqrt{ks}$. Therefore,

$$\begin{aligned}\|R_k b^\delta - A^+ b\|_2 &\leq \|R_k b^\delta - R_k b\|_2 + \|R_k b - A^+ b\|_2 \\ &\leq \sqrt{ks}\|b^\delta - b\|_2 + \|R_k b - A^+ b\|_2. \end{aligned} \tag{3.125}$$

This estimate evidently is the analog of (3.24). The second term on the right hand side converges to 0 for $k \to \infty$. The total error $R_k b^\delta - A^+ b$ only converges to 0, if $\|b^\delta - b\|_2 \leq \delta \to 0$ and if $k\delta^2 \to 0$ for $\delta \to 0$ and $k \to \infty$. The iteration index k plays the role of a regularization parameter: the larger it is, the better R_k approximates A^+, but the more data perturbations $b^\delta - b$ are amplified. The following theorem, which is the analog of Theorem 3.26, suggests a choice for k.

Theorem 3.35 (Stopping rule for the Landweber iteration) *Let $A \in \mathbb{R}^{m,n}$, $m \geq n$, with rank$(A) = n$. Let $b, b^\delta \in \mathbb{R}^m$. Denote by*

$\hat{x} = A^+ b$ *the unique minimizer of* $\|b - Ax\|_2 = \min!$ *(exact data) and by*
$x^k = R_k b^\delta$ *the kth iterate produced by* (3.120) *(perturbed data; $R_0 := 0$).*

The step size restriction (3.124) *applies. Assume that for some known $\delta > 0$*

$$\|b - b^\delta\|_2 + \|b - A\hat{x}\|_2 \leq \delta. \tag{3.126}$$

Assume that $\|b^\delta\|_2 > \tau\delta$ for a fixed parameter $\tau > 1$. Then:

(1) There exists a smallest integer $k \in \mathbb{N}$ (depending on b^δ), such that

$$\|b^\delta - AR_k b^\delta\|_2 \leq \tau\delta < \|b^\delta - AR_{k-1} b^\delta\|_2. \tag{3.127}$$

(2) With k as in (1), an estimate

$$\|\hat{x} - x^k\|_2 \leq \frac{\tau + 1}{\sigma_n}\delta \tag{3.128}$$

holds, where σ_n is the minimal singular value of A.

Remarks In case $\|b^\delta\|_2 \leq \tau\delta$, $x^0 = 0$ will be declared a regularized solution. The estimate (3.128) holds in this case. The comparison of (3.122) and (3.123) shows that regularization of A^+ by the Landweber iteration relies on an approximation of the function $\sigma \mapsto 1/\sigma$, $\sigma > 0$, by the polynomial $[1 - (1 - s\sigma)^k]/\sigma$. Variants of the Landweber iteration can be obtained by choosing other polynomial approximations.

Proof The proof of part (1) is modelled after [Kir96], p. 54 ff. Some modifications are needed, since we do not assume that the system $Ax = b$ has a solution. As already seen, the Landweber iteration produces iterates $x^k = R_k b^\delta$, $k = 1, 2, \ldots$,

where R_k is the operator defined by

$$R_k y = \sum_{i=1}^{n} \frac{[1-(1-s\sigma_i^2)^k]}{\sigma_i} \cdot (u_i^T y) \cdot v_i, \quad y \in \mathbb{R}^m, \tag{3.122}$$

which can be derived from a SVD $A = U\Sigma V^T$ of A. The column vectors $u_i \in \mathbb{R}^m$ of U build an orthonormal basis of $\mathbb{R}^m$, and the column vectors $v_i \in \mathbb{R}^n$ of V build an orthonormal basis of $\mathbb{R}^n$. The values $\sigma_1 \geq \cdots \geq \sigma_n > 0$ are the singular values of A. Since $Av_i = \sigma_i u_i$, $i = 1, \ldots, n$, one gets

$$AR_k y = \sum_{i=1}^{n} [1-(1-s\sigma_i^2)^k](u_i^T y)u_i \quad \text{for all } y \in \mathbb{R}^m. \tag{3.129}$$

Since $y = \sum_{i=1}^{m}(u_i^T y)u_i$ for all $y \in \mathbb{R}^m$, one further gets

$$\|AR_k y - y\|_2^2 = \sum_{i=1}^{n}(1-s\sigma_i^2)^{2k}(u_i^T y)^2 + \sum_{i=n+1}^{m}(u_i^T y)^2, \tag{3.130}$$

where the orthomormality of $\{u_1, \ldots, u_m\} \subset \mathbb{R}^m$ was used. By the step size restriction (3.124), all factors $(1-s\sigma_i^2)^{2k}$ are bounded by 1, thus (3.130) shows

$$\|AR_k - I_m\|_2 \leq 1. \tag{3.131}$$

Set $\hat{b} := A\hat{x} \in \mathscr{R}_A \subset \mathbb{R}^m$, such that $\hat{b} = \sum_{i=1}^{n}(u_i^T \hat{b})u_i$ ($\hat{b}$ has no components in directions $u_{n+1}, \ldots, u_m$). From

$$AR_k b^\delta - b^\delta = (AR_k - I_m)(b^\delta - \hat{b}) + AR_k\hat{b} - \hat{b}$$

and from (3.131), one gets

$$\|AR_k b^\delta - b^\delta\|_2 \leq 1 \cdot \|b^\delta - \hat{b}\|_2 + \|AR_k\hat{b} - \hat{b}\|_2. \tag{3.132}$$

The first term on the right hand side is bounded by δ by assumption (3.126). For the second term we get

$$\|AR_k\hat{b} - \hat{b}\|_2^2 = \sum_{i=1}^{n}(1-s\sigma_i^2)^{2k}(u_i^T\hat{b})^2, \tag{3.133}$$

similarly as in (3.130), but with the last summands missing because of the special choice of $\hat{b}$. From (3.124) one concludes that $\|AR_k\hat{b} - \hat{b}\|_2 \to 0$ for $k \to \infty$, such that $\|AR_k b^\delta - b^\delta\|_2 \leq \tau\delta$ for any $\tau > 1$, if only k is chosen large enough. This

proves part (1). Concerning part (2), set $r^k = b^\delta - Ax^k = b^\delta - AR_k b^\delta$. From

$$\begin{aligned} A^T r^k = A^T b^\delta - A^T A x^k &= \underbrace{A^T b}_{= A^T A \hat{x}} - A^T(b - b^\delta) - A^T A x^k \\ &= A^T A(\hat{x} - x^k) - A^T(b - b^\delta), \end{aligned}$$

we get

$$\hat{x} - x^k = (A^T A)^{-1} A^T r^k + (A^T A)^{-1} A^T (b - b^\delta).$$

From $\|r^k\|_2 \le \tau\delta$, from $\|b - b^\delta\|_2 \le \delta$, and from $\|(A^T A)^{-1} A^T\|_2 \le \sigma_n^{-1}$ (using an SVD of A), the estimate (3.128) follows. □

It is instructive to reconsider the Landweber iteration from a different perspective. The method of steepest descent for finding the minimum of $F^\delta(x) = \frac{1}{2}\|b^\delta - Ax\|_2^2$, when started at $x^0 = 0$, has a continuous analogon, namely the initial value problem

$$x'(t) = -\nabla F^\delta(x(t)) = A^T b^\delta - A^T A x(t), \quad x(0) = 0. \tag{3.134}$$

The solution of (3.134) is called **curve of steepest descent**. Solving (3.134) numerically by the explicit Euler method with constant step size s produces approximations

$$x^k \approx x(t_k), \quad t_k = k \cdot s, \quad k \in \mathbb{N}_0,$$

which are exactly the same as the ones defined by the Landweber iteration (3.119). But the curve of steepest descent can also be found analytically. To see this, use the SVD $A = U\Sigma V^T$ and the transformation $y(t) := V^T x(t)$, which decouples (3.134) into n separate initial value problems

$$y_i'(t) + \sigma_i^2 y_i(t) = \sigma_i (u_i^T b^\delta), \quad y_i(0) = 0, \quad i = 1, \ldots, n,$$

for the components of y, which can be solved easily. Transforming the solutions back gives the explicit solution formula

$$x(t) = \sum_{i=1}^{n} \frac{1 - \exp(-\sigma_i^2 t)}{\sigma_i} \cdot (u_i^T b^\delta) \cdot v_i. \tag{3.135}$$

Evidently, $x(t) \to \bar{x} = A^+ b^\delta$ for $t \to \infty$, compare (3.123). For finite values $t = T$, the vector $x(T)$ defined by (3.135) is a regularized approximation of $\hat{x} = A^+ b$, as formally proved, e.g., in [Hof99], p. 153. Now if some eigenvalues σ_i^2 of $A^T A$

are much larger than others,[7] then some components of $x(t)$ converge much faster than others, as can immediately be seen from (3.135). In this case the differential equation in (3.134) is called **stiff**. It is well known in numerical analysis that solving stiff equations by explicit numerical methods requires small step sizes s for the approximate values $x^k \approx x(k \cdot s)$ to converge to $\bar{x}$ for $k \to \infty$. A step size restriction consequently reappears for the Landweber iteration in the form $s \leq 1/\|A\|_2^2$ and might lead to a very slow convergence of this method. *Implicit* methods are more adequate for the numerical integration of stiff differential equations. The implicit Euler method with step sizes $s_k = t_{k+1} - t_k$, $k \in \mathbb{N}_0$, determines approximations $x^k \approx x(t_k)$, $k \in \mathbb{N}_0$, from the equations

$$x^{k+1} = x^k - s_k \nabla F^\delta(x^{k+1}) = x^k + s_k(A^T b^\delta - A^T A x^{k+1}), \quad x^0 = 0.$$

These identities can be reformulated to become

$$(I + s_k A^T A)x^{k+1} = x^k + s_k A^T b^\delta, \quad x^0 = 0, \tag{3.136}$$

or as well in the form

$$\left(A^T A + \frac{1}{s_k} I\right)(x^{k+1} - x^k) = A^T b^\delta - A^T A x^k, \quad x^0 = 0. \tag{3.137}$$

There is a close relation between (3.137) and Newton's method for the minimization of $F^\delta(x)$. Since the Hessian of F^δ is given by $\nabla^2 F^\delta(x) = A^T A$, Newton's method determines iteration updates by solving

$$\begin{aligned} \nabla^2 F^\delta(x^k)(x^{k+1} - x^k) &= -\nabla F^\delta(x^k) \\ &\iff A^T A(x^{k+1} - x^k) = A^T b^\delta - A^T A x^k. \end{aligned} \tag{3.138}$$

Of course, when started with $x^0 = 0$, Newton's method means nothing else than a solution of the normal equations and will converge in a single step. But this will be different in the nonlinear case to be considered in the next chapter. Comparing (3.138) to (3.137), one sees that solving the initial value problem (3.134) by the implicit Euler method exactly corresponds to a *regularized version of the Newton method*. This kind of regularization is known as **trust region method** in optimization. The equivalence of trust region methods, regularization, and numerical solution of the initial value problem (3.134) is described in [Sch12]. The iteration (3.137) can be analyzed easily for constant step sizes $s_k = s$. From (3.136)

[7] This is just the case when A is badly conditioned and thus the usual case for inverse problems.

one gets the explicit formula

$$x^k = \sum_{j=1}^{k} s(I + sA^TA)^{-j}A^Tb^\delta, \quad k \in \mathbb{N}. \tag{3.139}$$

This can be written as $x^k = R_k b^\delta$, where $(R_k)_{k\in\mathbb{N}}$ is the sequence of matrices

$$R_k = \sum_{j=1}^{k} s(I + sA^TA)^{-j}A^T \quad \in \quad \mathbb{R}^{m,n}. \tag{3.140}$$

Using an SVD $A = U\Sigma V^T$, one finds that

$$R_k y = \sum_{i=1}^{n} \frac{1}{\sigma_i} \underbrace{\left(1 - \frac{1}{(1+s\sigma_i^2)^k}\right)}_{=:\, q(k,\sigma_i)} \cdot (u_i^T y) \cdot v_i, \quad y \in \mathbb{R}^m. \tag{3.141}$$

Now, $q(k,\sigma) \overset{k\to\infty}{\longrightarrow} 1$ and thus $x^k \to \bar{x}$ independently of $s > 0$ und $\sigma > 0$. *The step size restriction required for the Landweber iteration is no longer needed.* From Bernoulli's inequality one further gets

$$\left(1 - \frac{s\sigma^2}{1+s\sigma^2}\right)^k \geq 1 - \frac{ks\sigma^2}{1+s\sigma^2}$$

for all $s, \sigma > 0$ and $k \in \mathbb{N}$, leading to

$$|q(k,\sigma)| \leq \sqrt{q(k,\sigma)} = \sqrt{1 - \left(1 - \frac{s\sigma^2}{1+s\sigma^2}\right)^k} \leq \sqrt{\frac{ks\sigma^2}{1+s\sigma^2}} \leq \sqrt{ks}\sigma.$$

From this estimate one can deduce that an estimate like (3.125) also holds for the iterates x^k determined by (3.137). Again, the iteration index k plays the role of a regularization parameter.

Theorem 3.36 (Stopping rule for the implicit Euler method) *Let $A \in \mathbb{R}^{m,n}$ with $m \geq n$ and $\operatorname{rank}(A) = n$. Let $b, b^\delta \in \mathbb{R}^m$. Denote by*

$\hat{x} = A^+ b$ *the unique minimizer of $\|b - Ax\|_2 = \min!$ (exact data) and by*
$x^k = R_k b^\delta$ *the kth iterate produced by (3.141) (perturbed data; $R_0 := 0$).*

Assume that for some known $\delta > 0$

$$\|b - b^\delta\|_2 + \|b - A\hat{x}\|_2 \leq \delta. \tag{3.142}$$

Assume that $\|b^\delta\|_2 > \tau\delta$ *for a fixed parameter* $\tau > 1$. *Then:*

(1) There exists a smallest integer $k \in \mathbb{N}$ *(depending on* b^δ*), such that*

$$\|b^\delta - AR_k b^\delta\|_2 \leq \tau\delta < \|b^\delta - AR_{k-1} b^\delta\|_2. \tag{3.143}$$

(2) With k as in (1), an estimate

$$\|\hat{x} - x^k\|_2 \leq \frac{\tau + 1}{\sigma_n}\delta \tag{3.144}$$

holds, where σ_n *is the smallest singular value of A.*

Proof The kth iterate can be written in form $x^k = R_k b^\delta$, where $R_k \in \mathbb{R}^{n,m}$ is the matrix given by

$$R_k y = \sum_{i=1}^{n} \frac{1}{\sigma_i}\left(1 - \frac{1}{(1 + s\sigma_i^2)^k}\right) \cdot (u_i^T y) \cdot v_i, \quad y \in \mathbb{R}^m. \tag{3.145}$$

compare (3.141). This representation can be derived from a SVD $A = U\Sigma V^T$ of A, as in (A.1). The column vectors $u_i \in \mathbb{R}^m$ of U build an orthonormal basis of $\mathbb{R}^m$, and the column vectors $v_i \in \mathbb{R}^n$ of V build an orthonormal basis of $\mathbb{R}^n$. The values $\sigma_1 \geq \cdots \geq \sigma_n > 0$ are the singular values of A. Since $Av_i = \sigma_i u_i$, $i = 1, \ldots, n$, one gets

$$AR_k y = \sum_{i=1}^{n}\left(1 - \frac{1}{(1 + s\sigma_i^2)^k}\right)(u_i^T y)u_i \quad \text{for all } y \in \mathbb{R}^m. \tag{3.146}$$

Since $y = \sum_{i=1}^{m}(u_i^T y)u_i$ for all $y \in \mathbb{R}^m$, one further gets

$$\|AR_k y - y\|_2^2 = \sum_{i=1}^{n} \frac{1}{(1 + s\sigma_i^2)^{2k}}(u_i^T y)^2 + \sum_{i=n+1}^{m}(u_i^T y)^2, \tag{3.147}$$

where the orthomormality of $\{u_1, \ldots, u_m\} \subset \mathbb{R}^m$ was used. Hence,

$$\|AR_k - I_m\|_2 \leq 1. \tag{3.148}$$

Set $\hat{b} := A\hat{x} \in \mathscr{R}_A \subset \mathbb{R}^m$, such that $\hat{b} = \sum_{i=1}^{n}(u_i^T \hat{b})u_i$ ($\hat{b}$ has no components in directions $u_{n+1}, \ldots, u_m$). From

$$AR_k b^\delta - b^\delta = (AR_k - I_m)(b^\delta - \hat{b}) + AR_k\hat{b} - \hat{b}$$

and from (3.148), one gets

$$\|AR_k b^\delta - b^\delta\|_2 \leq 1 \cdot \|b^\delta - \hat{b}\|_2 + \|AR_k\hat{b} - \hat{b}\|_2. \tag{3.149}$$

The first term on the right hand side is bounded by δ by assumption (3.142). For the second term we get

$$\|AR_k\hat{b} - \hat{b}\|_2^2 = \sum_{i=1}^{n} \frac{(u_i^T\hat{b})^2}{(1 + s\sigma_i^2)^{2k}}, \tag{3.150}$$

similarly as in (3.147), but with the last summands missing because of the special choice of $\hat{b}$. Evidently $\|AR_k\hat{b} - \hat{b}\|_2 \to 0$ for $k \to \infty$, such that $\|AR_k b^\delta - b^\delta\|_2 \leq \tau\delta$ for any $\tau > 1$, if only k is chosen large enough. This proves part (1). The proof of part (2) is exactly the same as for Theorem 3.35. □

In practice, an automatic step size control has to be provided for the implicit Euler method (3.137). Let $\tilde{x}(t)$ be the solution of the differential equation $x'(t) = -\nabla F^\delta(x(t))$ *with initial value* $\tilde{x}(t_k) = x^k$. Then $\|x^{k+1} - \tilde{x}(t_k + s_k)\|_\infty$ is the magnitude of the error when taking a single step from t_k to t_{k+1}, neglecting the error already contained in x^k. It is the goal of a step size control to choose s_k as large as possible while keeping the error magnitude under control by requiring that

$$\frac{\|x^{k+1} - \tilde{x}(t_k + s_k)\|_\infty}{s_k} \leq \varepsilon \tag{3.151}$$

for some chosen value ε (the left hand side of (3.151) is commonly known as local discretization error). For the (implicit) Euler method, it is known that

$$\frac{x^{k+1} - \tilde{x}(t_k + s_k)}{s_k} \doteq \tau_k \cdot s_k, \tag{3.152}$$

for some constant vector τ_k, independent of s_k. The dotted equality sign in (3.152) means equality up to terms of order s_k^2 or higher. Thus, (3.151) can be achieved by choosing s_k small enough, but we can not say offhand *how* small, since τ_k is not known. One therefore at first computes x^{k+1} according to (3.137) for some tentative value $s_k = s$. Further, one computes an approximation $\tilde{x}^{k+1}$ by taking *two* subsequent steps of size $s/2$. Using (3.152) it can be shown that the corresponding error is given by

$$\frac{\tilde{x}^{k+1} - \tilde{x}(t_k + s)}{s} \doteq \tau_k \cdot \frac{s}{2}. \tag{3.153}$$

Comparing (3.152) (for $s_k = s$) and (3.153), one gets

$$\Delta := \frac{\|x^{k+1} - \tilde{x}^{k+1}\|_\infty}{s} \doteq \|\tau_k\|_\infty \cdot \frac{s}{2},$$

meaning that Δ can serve as an estimator for the magnitude of the error in (3.153). Moreover, this error scales with s, such that taking a step of rescaled size $s \cdot \frac{\varepsilon}{\Delta}$ should produce an error of size ε, the maximum allowed according to (3.151). Following common advice, we will choose a somewhat smaller new step size $0.9 \cdot s \cdot \frac{\varepsilon}{\Delta}$ and also take care to avoid extreme step size changes. We arrive at

Implicit Euler method with step size control for least squares problems
Let all conditions of Theorem 3.36 hold. Choose $\varepsilon \in (0,1), s > 0$ and $x^0 \in \mathbb{R}^m$ (e.g. $x^0 = 0$). Set $k = 0$.

Step 1: Compute $r^k := b^\delta - Ax^k$ and $c^k := A^T r^k$. If $\|r^k\|_2 \le \tau\delta$, then stop. Otherwise, go to Step 2.

Step 2: If $A^TA + \frac{1}{s}I$ is so badly conditioned that a numerical Cholesky factorization fails, then set $s := \frac{s}{2}$ and go to Step 2. Otherwise go to Step 3.

Step 3: Get the solution p of $(A^TA + \frac{1}{s}I)x = c^k$ [via Cholesky factorization], set $\eta^s := x^k + p$, and goto Step 4.

Step 4: Get the solution p of $(A^TA + \frac{2}{s}I)x = c^k$ and set $\eta^{s/2} := x^k + p$. Compute $\tilde{r} = b^\delta - A\eta^{s/2}$ and $\tilde{c} = A^T\tilde{r}$. Get the solution q of $(A^TA + \frac{2}{s}I)x = \tilde{c}$ and set $\tilde{\eta}^s = \eta^{s/2} + q$.

Step 5: Compute $\Delta := \|\eta^s - \tilde{\eta}^s\|_\infty / s$. If $\Delta < 0.1 \cdot \varepsilon$, set $f := 10$, else set $f := \varepsilon/\Delta$. If $f < 0.1$, set $f = 0.1$. Set $s := s \cdot f \cdot 0.9$. If $\Delta \le \varepsilon$, set $x^{k+1} := \tilde{\eta}^s$, set $k = k+1$, and goto Step 1. Otherwise, go to Step 3.

Example 3.37 (Linear seismic tomography) We reconsider Example 3.28 of linearized seismic tomography, setting up A and b from collocation equations with $m = n = 100$. Perturbed data b^δ are simulated by adding realizations of independent normally distributed random variables with zero mean and standard deviation $\sigma = 10^{-5}$ to the components of b. Thus, we set $\delta := \sqrt{m}\sigma$. Figure 3.14 shows the result we achieved for the choice $\varepsilon = 10^{-2}$ and $\tau = 1$ (right) and also shows how well the minimization of $F(x) = \frac{1}{2}\|b^\delta - Ax\|_2$ proceeded with the iteration index k. The implicit Euler method stopped with $k = 10$. The result – of comparable quality as the original Tikhonov regularization for $L = I_n$ – is not very satisfactory. ◇

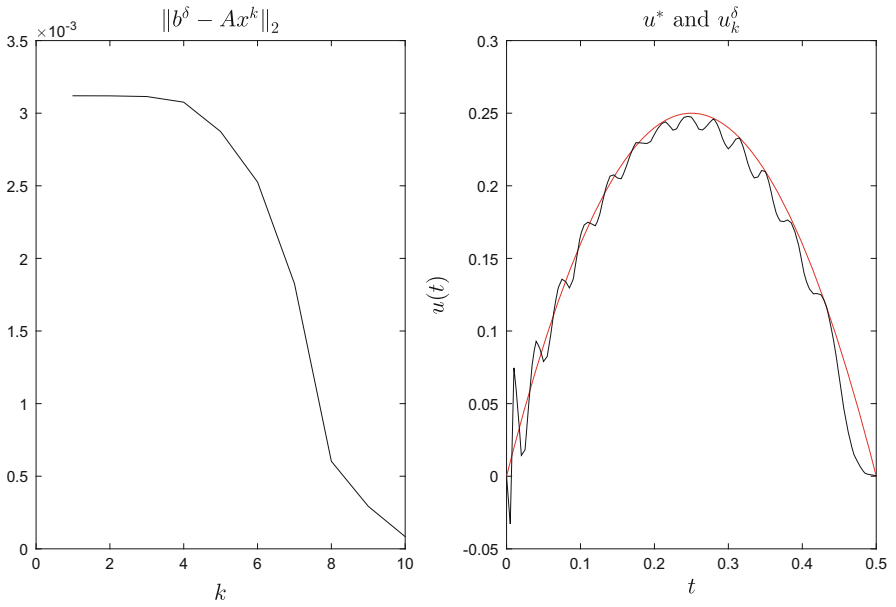

Fig. 3.14 Implicit Euler method, $\varepsilon = 10^{-2}$ and $\tau = 1$

Let us have one more look at formula (3.137), which we may rewrite in the form

$$\left(A^T A + \frac{1}{s_k} I\right) u^k = A^T r^k, \quad x^0 = 0, \tag{3.154}$$

if we set

$$u^k := x^{k+1} - x^k \quad \text{and} \quad r^k := b^\delta - Ax^k.$$

A solution u^k of (3.154) can equivalently be characterized as the minimizer of

$$\|Au - r^k\|_2^2 + \frac{1}{s_k}\|u\|_2^2, \quad u \in \mathbb{R}^n,$$

which means that the above algorithm implicitly uses the Euclidean norm $\|u\|_2$ as regularity measure for computing iteration updates. More generally, an update u^k could be computed as the minimizer of

$$\|Au - r^k\|_2^2 + \frac{1}{s_k}\|Lu\|_2^2, \quad u \in \mathbb{R}^n,$$

but this will not be done by the implicit Euler method. However, we can force the implicit Euler method into behaving as if $\|L \bullet \|_2$ was used as a regularity measure

by transforming the general Tikhonov regularization into a standard one. According to Theorem 3.32, we will do the following. First, compute

$$\tilde{A} \quad \text{and} \quad \tilde{b^\delta} \quad \text{as in} \quad (3.100) \quad [b \text{ replaced by } b^\delta]. \tag{3.155}$$

Second, apply the implicit Euler method to the minimization of

$$\|\tilde{A}z - \tilde{b^\delta}\|_2, \quad z \in \mathbb{R}^p. \tag{3.156}$$

Finally, transform the obtained solution z^δ back to

$$x^\delta = (V_1 - V_2 U^{-1} Q_1^T A V_1) R^{-T} z^\delta + V_2 U^{-1} Q_1^T b^\delta, \quad \text{as in (3.101)}. \tag{3.157}$$

This will be tested in the following example.

Example 3.38 (Linear seismic tomography) We solve the problem from 3.37 again, keeping all previous data, but including a coordinate transform. This means we follow the steps (3.155), (3.156), and (3.157), where we choose $L = L_2 \in \mathbb{R}^{p,n}$, $p = n-2$, as in (3.39) ("second derivatives"). Figure 3.15 shows the result achieved, which is much better than in the previous example. ◇

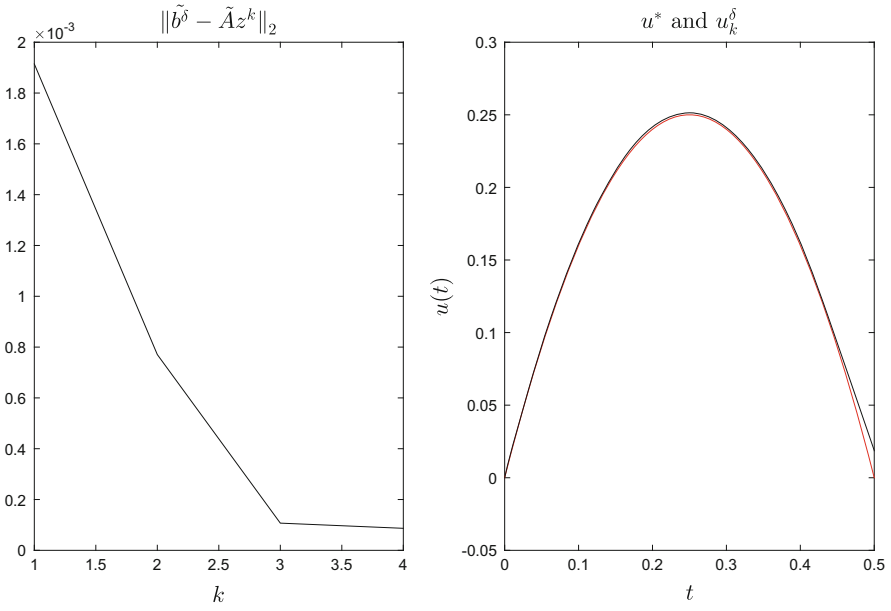

Fig. 3.15 Implicit Euler method with coordinate transform, $\varepsilon = 10^{-2}$ and $\tau = 1$

The above example shows that by using the implicit Euler iteration, one can only seemingly avoid the choice of a regularization term. The latter implicitly reappears in form of choosing a suitable coordinate transform. The same observation will be made for the other iterative methods to be presented below.

3.10 The Conjugate Gradient Method

The conjugate gradient method is an iterative algorithm to solve

$$\min_{x\in\mathbb{R}^n} F^\delta(x), \quad F^\delta(x) := \frac{1}{2}\|b^\delta - Ax\|_2^2, \tag{3.118}$$

where $A \in \mathbb{R}^{m,n}$, $m \geq n$, is a matrix of full rank n. Stopped prematurely, it will produce a regularized approximation of the unique minimizer $\hat{x}$ of $\|b - Ax\|_2$, as did the algorithms presented in Sect. 3.9.

Conjugate Gradient Method for Linear Systems of Equations

The conjugate gradient method will at first be described as an iterative method to solve a linear system of equations

$$Ax = b, \quad A \in \mathbb{R}^{n,n} \text{ symmetric and positive definite}, \quad b \in \mathbb{R}^n, \tag{3.158}$$

having the unique solution $x^* \in \mathbb{R}^n$. In the following paragraph, this iteration will be applied to the normal equations of (3.118), i.e. we will replace the matrix A in (3.158) by A^TA (which is symmetric and positive definite for any matrix $A \in \mathbb{R}^{m,n}$ of full rank) and the vector b by A^Tb^δ.

Since A is symmetric and positive definite, the mapping

$$\|\bullet\|_A : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \|x\|_A := \sqrt{x^*Ax} \tag{3.159}$$

defines a norm on $\mathbb{R}^n$, the so-called A**-norm**. The A-norm is induced by a scalar product on $\mathbb{R}^n$ given by

$$\langle x|y\rangle_A := x^TAy = \langle x|Ay\rangle = \langle Ax|y\rangle, \quad x, y \in \mathbb{R}^n.$$

We will further need the so-called **Krylov spaces**

$$\mathscr{K}_k := \langle b, Ab, \ldots, A^{k-1}b\rangle, \quad k = 1, \ldots, n.$$

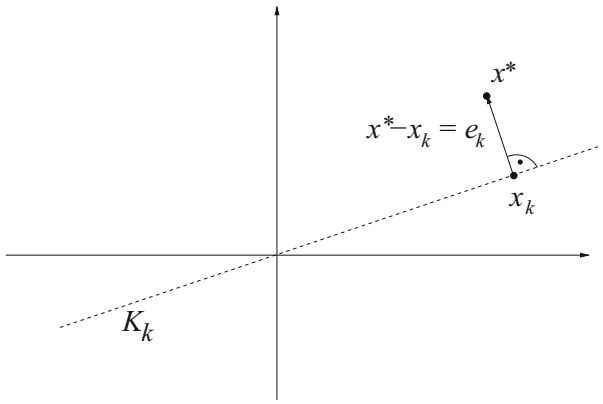

Fig. 3.16 *A*-orthogonal projection onto Krylov space

There is a unique best approximation $x_k \in \mathscr{K}_k$ of x^*, i.e. a unique x_k such that

$$\|x_k - x^*\|_A \leq \|x - x^*\|_A \quad \text{for all} \quad x \in \mathscr{K}_k \tag{3.160}$$

This solution is characterized by

$$\langle x^* - x_k | x \rangle_A = 0 \quad \text{for all} \quad x \in \mathscr{K}_k, \tag{3.161}$$

see Theorem B.5, meaning that x_k is the A-orthogonal projection of x^* onto $\mathscr{K}_k$, as illustrated in Fig. 3.16. Two vectors $x, y \in \mathbb{R}^n$ are called A-orthogonal or **conjugate** (with respect to A), iff $\langle x|y\rangle_A = 0$. In this case we write $x \perp_A y$, whereas $x \perp y$ designates orthogonality with respect to the Euclidean scalar product.

Assume we knew k *nonzero*, mutually conjugate vectors $p_1, \ldots, p_k \in \mathscr{K}_k$, i.e. $\langle p_i|p_j\rangle_A = 0$ for $i \neq j$. Such vectors necessarily are linearly independent (implying $\dim(\mathscr{K}_k) = k$), since $\alpha_1 p_1 + \ldots + \alpha_k p_k = 0$ implies

$$0 = \langle p_j | \sum_{i=1}^{k} \alpha_i p_i \rangle_A = \sum_{i=1}^{k} \alpha_i \langle p_j | p_i \rangle_A = \alpha_j \|p_j\|_A^2 \ ,$$

and therefore $\alpha_j = 0$ for every j. Consequently, $\{p_1, \ldots, p_k\}$ is a basis of $\mathscr{K}_k$:

$$\mathscr{K}_k = \langle b, Ab, \ldots, A^{k-1}b \rangle = \langle p_1, \ldots, p_k \rangle. \tag{3.162}$$

Then, the unknown solution x_k of (3.161) can be rated as a linear combination of conjugate vectors: $x_k = \alpha_1 p_1 + \ldots + \alpha_k p_k$. From the definition of $\langle \bullet | \bullet \rangle_A$ we get an equivalent formulation of (3.161), namely

$$\langle x^* - x_k | p_j \rangle_A = 0 \quad \Longleftrightarrow \quad \langle Ax^* - Ax_k | p_j \rangle = 0, \quad j = 1, \ldots, k, \tag{3.163}$$

where the Euclidean scalar product is used on the right hand side. Knowing $Ax^* = b$ and substituting the ansatz $x_k = \sum_i \alpha_i p_i$ one gets

$$\langle b|p_j\rangle = \langle Ax_k|p_j\rangle = \alpha_j \langle p_j|p_j\rangle_A, \quad j = 1,\dots,k, \tag{3.164}$$

which can be used to compute the coefficients α_j. Setting $r_k := b - Ax_k = A(x^* - x_k)$, the solution x_k of (3.160) can be written in the form

$$x_k = \sum_{j=1}^{k} \alpha_j \cdot p_j\,, \quad \alpha_j = \frac{\langle b|p_j\rangle}{\langle Ap_j|p_j\rangle} = \frac{\langle x^*|p_j\rangle_A}{\langle p_j|p_j\rangle_A}. \tag{3.165}$$

It remains to find conjugate vectors. This could be achieved by Gram-Schmidt orthogonalization, but there is a cheaper way based on the following

Lemma 3.39 *Let $A \in \mathbb{R}^{n,n}$ be symmetric and positive definite, let $b \in \mathbb{R}^n$ and let $\mathcal{K}_k := \langle b, Ab, \dots, A^{k-1}b\rangle$ for $k = 1,\dots,n$. Let $x_k \in \mathcal{K}_k$ be the unique solution of (3.160) for $k = 1,\dots,n$. Let $r_0 := b$ and let $r_k := b - Ax_k$, $k = 1,\dots,n$. If $r_j \neq 0$ for $j = 0,\dots,k$, then*

$$\mathcal{K}_{k+1} = \langle r_0,\dots,r_k\rangle \tag{3.166}$$

and

$$\langle r_i|r_j\rangle = 0 \quad \text{for} \quad i \neq j, \quad i,j = 0,\dots,k, \tag{3.167}$$

implying $\dim(\mathcal{K}_{k+1}) = k+1$.

Proof The proof is by induction on k. For $k = 0$ one has by definition $\mathcal{K}_1 = \langle b\rangle = \langle r_0\rangle$, which shows (3.166), whereas (3.167) is void. Assume the statement to be true for $k-1$, i.e. $\mathcal{K}_k = \langle r_0,\dots,r_{k-1}\rangle$ and $\langle r_i|r_j\rangle = 0$ for $i \neq j$, $i,j = 0,\dots,k-1$. Although we have not yet come up with its explicit construction, the existence of an A-orthogonal basis $\{p_1,\dots,p_k\}$ of $\mathcal{K}_k$ is guaranteed, so x_k can be written in the form (3.165). On the one hand, $r_k := b - Ax_k = Ax^* - Ax_k \in \mathcal{K}_{k+1}$ by definition of $\mathcal{K}_{k+1}$ and since $x_k \in \mathcal{K}_k$. On the other hand, $r_k \perp \mathcal{K}_k = \langle p_1,\dots,p_k\rangle$ by (3.163). Therefore, either $r_k = 0$ or (3.166) and (3.167) hold. □

Conjugate directions p_k will be determined iteratively, starting with $p_1 = b$. Assume $p_1,\dots,p_k$ to be already determined and $x_k \in \mathcal{K}_k$ to be computed by (3.165). Let $r_k = b - Ax_k$. If $r_k = 0$, then $x_k = x^*$ and we are done – a solution of (3.158) is found. Otherwise $\mathcal{K}_{k+1} \ni r_k \perp \mathcal{K}_k = \langle p_1,\dots,p_k\rangle$, so that $\mathcal{K}_{k+1} = \langle p_1,\dots,p_k,r_k\rangle$. Now p_{k+1} can be determined by a Gram-Schmidt orthogonalization step, projecting r_k onto $\mathcal{K}_{k+1}$:

$$p_{k+1} = r_k - \sum_{j=1}^{k} \frac{\langle r_k|p_j\rangle_A}{\langle p_j|p_j\rangle_A} p_j = r_k + \beta_{k+1}p_k, \quad \beta_{k+1} := -\frac{\langle r_k|p_k\rangle_A}{\langle p_k|p_k\rangle_A}\,, \tag{3.168}$$

where (3.166) and (3.167) were used. The conjugate gradient iteration is defined by formulas (3.165) and (3.168), but the computation of α_k and β_{k+1} can still be simplified. Knowing $x_{k-1} \in \mathscr{K}_{k-1} \perp_A p_k$, (3.168) and $p_{k-1} \in \mathscr{K}_{k-1} \perp r_{k-1}$, one finds

$$\langle x^*|p_k\rangle_A = \langle x^* - x_{k-1}|p_k\rangle_A = \langle r_{k-1}|p_k\rangle = \langle r_{k-1}|r_{k-1}\rangle$$

and derives

$$\alpha_k = \frac{\langle r_{k-1}|r_{k-1}\rangle}{\langle p_k|p_k\rangle_A}. \tag{3.169}$$

Further on $r_k = b - Ax_k = r_{k-1} - \alpha_k A p_k$ and since $r_k \perp r_{k-1}$, one gets

$$-\alpha_k \langle r_k|p_k\rangle_A = \langle r_k| - \alpha_k A p_k\rangle = \langle r_k|r_k - r_{k-1}\rangle = \langle r_k|r_k\rangle.$$

This can be substituted together with (3.169) into (3.168) to give

$$\beta_{k+1} = \frac{\langle r_k|r_k\rangle}{\langle r_{k-1}|r_{k-1}\rangle}.$$

This gives the CG iteration of Hestenes and Stiefel:

The conjugate gradient (CG) iteration

$p_1 = r_0 = b$ and $x_0 = 0$

for $k = 1, 2, 3, \ldots$

$v := Ap_k$

$\alpha_k = \dfrac{\langle r_{k-1}|r_{k-1}\rangle}{\langle p_k|v\rangle}$

$x_k = x_{k-1} + \alpha_k p_k$

$r_k = r_{k-1} - \alpha_k v$

$\beta_{k+1} = \dfrac{\langle r_k|r_k\rangle}{\langle r_{k-1}|r_{k-1}\rangle}$

$p_{k+1} = r_k + \beta_{k+1} p_k$

The iteration surely has to be stopped as soon as $r_k = 0$, since this means $x_k = x^*$. To avoid near zero division, it should rather be stopped as soon as $\|r_k\|_2$ gets too small, see also Theorem 3.41 below.

For a convergence analysis of the CG iteration, we follow [TB97], p. 298. Since $x_0 = 0$, for any $y \in \mathscr{K}_k = \langle b, Ab, \ldots, A^{k-1}b\rangle$ we have

$$x^* - y \;=\; x^* - x_0 - \sum_{j=0}^{k-1} \gamma_j A^j b = x^* - x_0 - \sum_{j=0}^{k-1} \gamma_j A^j A(x^* - x_0)$$

$$=: I \cdot (x^* - x_0) - \sum_{j=1}^{k} \gamma_{j-1} A^j (x^* - x_0).$$

The above is the error when approximating the solution x^* of (3.158) by some $y \in \mathscr{K}_k$. This error can be written in the form

$$y \in \mathscr{K}_k \quad \Longrightarrow \quad x^* - y = p(A)(x^* - x_0)\ , \tag{3.170}$$

where p is a normalized polynomial of degree $\leq k$:

$$p \in P_k^0 := \{p \text{ polynomial of degree } \leq k;\ p(0) = 1\}.$$

Knowing that x_k is the best approximation of x^* from $\mathscr{K}_k$ with respect to the norm $\|\cdot\|_A$, one derives the optimality relation

$$\|x^* - x_k\|_A = \min_{p \in P_k^0} \|p(A)(x^* - x_0)\|_A, \tag{3.171}$$

relating the CG iteration to polynomial best approximation. Since A is symmetric and positive definite, it has n positive eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n > 0$ with corresponding eigenvalues $v_1, \ldots, v_n$ forming an orthonormal basis of $\mathbb{R}^n$. Consequently, there are scalars $\gamma_1, \ldots, \gamma_n$ such that

$$x^* - x_0 = \sum_{j=1}^{n} \gamma_j v_j \quad \Longrightarrow \quad p(A)(x^* - x_0) = \sum_{j=1}^{n} \gamma_j p(\lambda_j) v_j\ .$$

But then

$$\|x^* - x_0\|_A^2 = \langle x^* - x_0 | A(x^* - x_0)\rangle = \Big\langle \sum \gamma_j v_j | \sum \gamma_j \lambda_j v_j \Big\rangle = \sum |\gamma_j|^2 \lambda_j$$

and – for all $p \in P_k^0$:

$$\|p(A)(x^* - x_0)\|_A^2 = \Big\langle \sum \gamma_j p(\lambda_j) v_j | \sum \gamma_j p(\lambda_j) \lambda_j v_j \Big\rangle \leq \max_{\lambda \in \sigma(A)} |p(\lambda)|^2 \sum |\gamma_j|^2 \lambda_j\ ,$$

where $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$ is called the **spectrum** of A. This proves the following

Theorem 3.40 (Convergence of CG iteration)

$$\|x^* - x_k\|_A \leq \inf_{p \in P_k^0} \max_{\lambda \in \sigma(A)} |p(\lambda)| \cdot \|x^* - x_0\|_A \,. \tag{3.172}$$

From Theorem 3.40 one can conclude, that a clustering of eigenvalues is beneficial for the CG iteration. If A has only $k \leq n$ *distinct* eigenvalues, convergence will occur after k steps, since a polynomial $p \in P_k^0$ exists having these k eigenvalues as its zeros (since A is positive definite, 0 can not be an eigenvalue). By a special choice of p in (3.172) one can show that

$$\|x^* - x_k\|_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k \|x^* - x_0\|_A, \tag{3.173}$$

where $\kappa_2(A)$ is the spectral condition number of A, see [TB97], Theorem 38.5.

Conjugate Gradient Method for Linear Least Squares Problems

To minimize $\|b^\delta - Ax\|_2$, where $A \in \mathbb{R}^{m,n}$ has full rank n, the CG iteration can directly be applied to the corresponding normal equations $A^T Ax = A^T b^\delta$. Denote the unique solution of these equations by $\bar{x}$, whereas $\hat{x}$ denotes the unique solution of $A^T Ax = A^T b$ (unperturbed data). One easily derives the following algorithm from the original CG iteration, which is known as **CGNE (CG iteration applied to normal equations)** algorithm.

CGNE: CG iteration applied to $A^T Ax = A^T b^\delta$

$x_0 = 0, r_0 = b^\delta$ and $p_1 = A^T r_0$

for $k = 1, 2, 3, \ldots$

$$\alpha_k = \frac{\langle A^T r_{k-1} | A^T r_{k-1} \rangle}{\langle Ap_k | Ap_k \rangle}$$

$$x_k = x_{k-1} + \alpha_k p_k$$

$$r_k = r_{k-1} - \alpha_k A p_k$$

$$\beta_{k+1} = \frac{\langle A^T r_k | A^T r_k \rangle}{\langle A^T r_{k-1} | A^T r_{k-1} \rangle}$$

$$p_{k+1} = A^T r_k + \beta_{k+1} p_k$$

Here, $r_k := b^\delta - Ax_k$, as in the original CG iteration. However, the residual corresponding to the normal equations rather is

$$\rho_k := A^T b^\delta - A^T A x_k = A^T r_k. \tag{3.174}$$

For example, Lemma 3.39 in the context of CGNE applies to the vectors ρ_k, *not* to r_k! In analogy to (3.160), unless stopped by the occurrence of some $\rho_j = A^T r_j = 0$, CGNE determines x_k as the minimizer of

$$\begin{aligned} \|x - \bar{x}\|^2_{A^T A} &= (x - \bar{x})^T A^T A (x - \bar{x}) \\ &= x^T A^T A x - 2x^T A^T A \bar{x} + (\bar{x})^T A^T A \bar{x} \\ &\overset{A^T A \bar{x} = A^T b^\delta}{=} x^T A^T A x - 2x^T A^T b^\delta + \left[(\bar{x})^T A^T \bar{x}\right] \end{aligned}$$

over the Krylov space

$$\mathcal{K}_k = \langle A^T b^\delta, (A^T A) A^T b^\delta, \ldots, (A^T A)^{k-1} A^T b^\delta \rangle. \tag{3.175}$$

The term in brackets is constant, so that omitting it does not change the minimizer. Also, the minimizer is not changed if the constant term $(b^\delta)^T b^\delta$ is added. Therefore, the kth step of the CGNE iteration determines x_k as the minimizer of

$$x^T A^T A x - 2x^T A^T b^\delta + (b^\delta)^T b^\delta = \|b^\delta - Ax\|_2^2, \quad x \in \mathcal{K}_k. \tag{3.176}$$

Since b^δ is only a perturbed version of b, it generally makes no sense to exactly solve $A^T A x = A^T b^\delta$. Rather, one will stop the iteration as soon as an iterate x_k is determined such that $\|b^\delta - Ax_k\|_2 \le \delta$, where δ is a bound on the error $b - b^\delta$ in the data. The following theorem tells us that this is a feasible strategy making CGNE a regularization method for ill-posed inverse problems.

Theorem 3.41 (Stopping rule and convergence property of CGNE) *Let $A \in \mathbb{R}^{m,n}$, $m \ge n$, with rank$(A) = n$. Let $b, b^\delta \in \mathbb{R}^m$. Denote by*

$\hat{x} = A^+ b$	*the unique minimizer of $\|b - Ax\|_2 = \min!$ (exact data) and by*
x_k	*the kth iterate produced by CGNE (perturbed data b^δ).*

Assume that for some known $\delta > 0$

$$\|b - b^\delta\|_2 + \|b - A\hat{x}\|_2 \le \delta < \|b^\delta\|_2. \tag{3.177}$$

Then the following holds:

(1) There exists a smallest integer $k \in \mathbb{N}$ (depending on b^δ), such that

$$\|b^\delta - Ax_k\|_2 \leq \delta < \|b^\delta - Ax_{k-1}\|_2. \tag{3.178}$$

(2) With k as in (1),

$$\|\hat{x} - x_k\|_2 \leq 2\frac{\delta}{\sigma_n}, \tag{3.179}$$

with σ_n the smallest singular value of A.

Remarks In case $\|b^\delta\|_2 \leq \delta$, $x_0 := 0$ will be declared a solution. In this case $\|b^\delta - Ax_0\|_2 \leq \delta$ holds. *The bound given in* (3.179) *is worse than the comparable one from Theorem 3.6 for the unregularized solution A^+b^δ and does not help to understand how CGNE can have a regularizing effect when solving a finite-dimensional least squares problem. To see this, refer to the technical proof presented in Appendix D.*

Proof By (3.177), one knows $\|b^\delta - Ax_0\|_2 > \delta$, so the iteration can not stop with $k = 0$. CGNE will produce the minimizer $\bar{x}$ of $\|b^\delta - Ax\|_2$ after maximally n steps. Since

$$\|b^\delta - A\bar{x}\|_2 \leq \|b^\delta - A\hat{x}\|_2 \overset{(3.177)}{\leq} \delta,$$

this means that (3.178) will certainly be fulfilled after n steps of CGNE. This proves part (1). Assume now that k is determined such that (1) holds. With $r_k = b^\delta - Ax_k$ one gets

$$\begin{aligned} A^T r_k = A^T b^\delta - A^T A x_k &= \underbrace{A^T b}_{= A^T A\hat{x}} + A^T(b^\delta - b) - A^T A x_k \\ &= A^T A(\hat{x} - x_k) + A^T(b^\delta - b), \end{aligned}$$

showing that

$$\hat{x} - x_k = (A^T A)^{-1} A^T r_k - (A^T A)^{-1} A^T (b^\delta - b).$$

By (3.178), $\|r_k\|_2 \leq \delta$ and by (3.177), $\|b^\delta - b\|_2 \leq \delta$. Since $\|(A^T A)^{-1} A^T\|_2 \leq 1/\sigma_n$, as can be seen using an SVD of A, the estimate (3.179) follows. □

From Theorem 3.40 it is known that a clustering of eigenvalues of $A^T A$ will be beneficial for the progress of the CGNE method. If the matrix A and thus $A^T A$ is ill-conditioned, as it is expected to be for inverse problems, eigenvalues will cluster at 0, which is an advantage for the CG iteration. For further improvement of convergence,

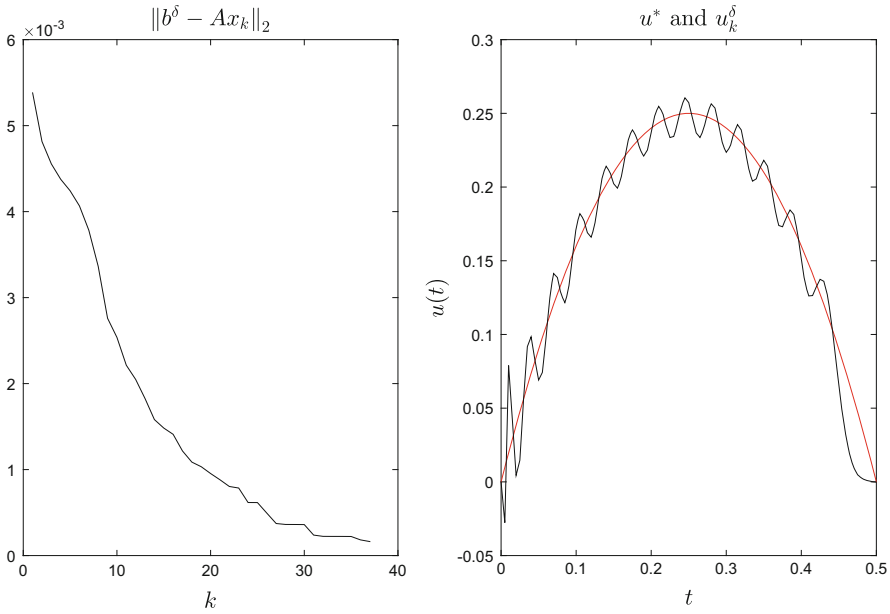

Fig. 3.17 CGNE iteration for linearized seismic tomography

one can try to shape the rest of the spectrum by replacing the system $A^TAx = A^Tb^\delta$ by an equivalent system

$$X^{-1}A^TAX^{-T}X^Tx = X^{-1}A^Tb^\delta, \quad X \in \mathbb{R}^{n,n} \text{ to be chosen.}$$

This is known as **preconditioning**. We will not deal with preconditioning, refer to Lecture 40 of [TB97] for a short introduction and pointers to the literature.

Example 3.42 (Linear seismic tomography) Example 3.37 is repeated with the same data, but using CGNE to solve the normal equations of the least squares problem. Figure 3.17 shows the result achieved, which is of comparable quality as the one obtained for the implicit Euler method presented in the last section. CGNE stopped after 36 iterations in this example. Note that one iteration step of CGNE is much cheaper than one iteration step of the implicit Euler method with step size control considered in Sect. 3.9, which requires at least the solution of three regularized systems of normal equations. To achieve regularization with respect to the semi-norm $x \mapsto \|Lx\|_2$, with $L = L_2 \in \mathbb{R}^{p,n}$, $p = n - 2$, as in (3.39) ("second derivatives"), we followed the steps (3.155), (3.156), and (3.157), as in Sect. 3.9 (but used CGNE to minimize (3.156)). Figure 3.18 shows the result obtained in this case, where additionally the parameter m was increased from 100 to 300. This

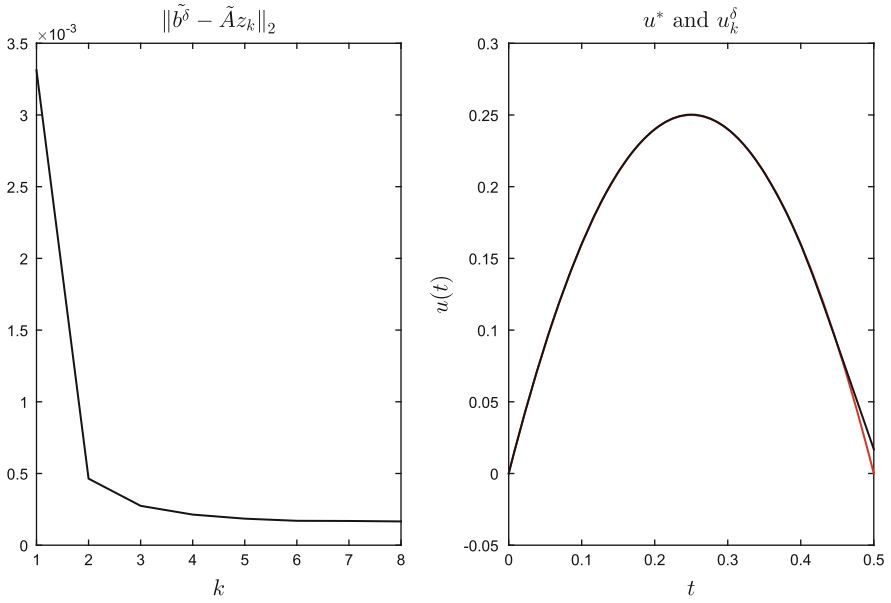

Fig. 3.18 CGNE iteration for linearized seismic tomography, coordinate transform

demonstrates the efficiency of CGNE to solve ill-conditioned linear problems. We also tested CGNE (with the transformations (3.155), (3.156), and (3.157) for $L = L_2$ from (3.39)) for the reconstruction of the step function from Example 3.29 and obtained results of comparable accuracy as in that example. ◇

Chapter 4
Regularization of Nonlinear Inverse Problems

Nonlinear inverse problems are much more difficult to solve than linear ones and the corresponding theory is far less developed. Each particular problem may demand a specific regularization. Nevertheless, one can formulate and analyze basic versions of nonlinear Tikhonov regularization and nonlinear iterative methods, which will be done in Sects. 4.1 and 4.6, respectively. These basic versions serve as starting points for regularizations adapted to our nonlinear model problems (Sects. 4.2, 4.5, and 4.6). Nonlinear Tikhonov regularization leads to nonlinear least squares problems. Sections 4.3 and 4.4 present algorithms to solve these numerically.

4.1 Tikhonov Regularization of Nonlinear Problems

The following theoretical results are mostly taken from [EKN89]. We consider a nonlinear system of equations

$$F(x) = y, \quad F : D \subseteq \mathbb{R}^n \to \mathbb{R}^m, \quad x \in D, \quad y \in \mathbb{R}^m, \tag{4.1}$$

where $D \subseteq \mathbb{R}^n$ is *closed* and *convex* and where F is a continuous mapping. Problem (4.1) results from discretizing a nonlinear parameter identification problem, as demonstrated in Sect. 2.6 for examples from inverse gravimetry and seismic tomography. In these examples, D was a multi-dimensional interval ("a box"). *We assume that* (4.1) *has at least one solution* $\hat{x}$. This is a real restriction: even if the original parameter identification problem is always required to have a (unique) solution, the situation is different for its finite-dimensional counterpart, when discretization errors come into play. On the other hand, data errors are always

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9_4

present. This means that we only dispose of some approximation $y^\delta \approx y$ with

$$\|y - y^\delta\|_2 \leq \delta, \tag{4.2}$$

where $\delta > 0$ is assumed to be known. We will not assume that $F(x) = y^\delta$ has a solution and therefore replace (4.1) by the **nonlinear least squares problem**

$$\text{minimize} \quad \|F(x) - y^\delta\|_2, \quad x \in D.^{1} \tag{4.3}$$

But even (4.3) will only have a solution under appropriate conditions concerning D and F. If a solution exists, it is not necessarily unique. Finally, if a unique solution exists, it may be far away from $\hat{x}$ even if $\delta > 0$ is small – this will be the case if (4.3) is badly conditioned, i.e. if a solution of (4.3) depends very sensitively on y^δ. Motivated by the linear case, instead of trying to solve the minimization problem (4.3), one might ask for some $\tilde{x} \in D$ which only fulfills $\|F(\tilde{x}) - y^\delta\|_2 \leq \delta$ but additionally has some other desirable quality which guarantees its being close to $\hat{x}$. One possibility to achieve this is by considering the following nonlinear analogon of Tikhonov regularization:

$$\text{minimize} \quad Z_\lambda(x) := \|F(x) - y^\delta\|_2^2 + \lambda \|Lx\|_2^2, \quad x \in D, \tag{4.4}$$

where $L \in \mathbb{R}^{p,n}$ and $\lambda \geq 0$ have to be chosen appropriately. Often, Lx will correspond to some discretized derivative of the function represented by the vector x. In the following analysis, only the case $L = I_n \in \mathbb{R}^{n,n}$ (identity matrix) will be investigated, but (4.4) will also be slightly generalized: choose $x^* \in D$ and

$$\text{minimize} \quad T_\lambda(x) := \|F(x) - y^\delta\|_2^2 + \lambda \|x - x^*\|_2^2, \quad x \in D. \tag{4.5}$$

This means that one is looking for some x close to x^*, which at the same time fulfills well the equality $F(x) = y^\delta$. The choice of x^* is crucial: only in case $x^* \approx \hat{x}$ one can expect a meaningful result, but this choice requires an approximate knowledge of the solution $\hat{x}$ of $F(x) = y$. Such knowledge may be very difficult to get, but is also required from a practical perspective by the restricted convergence properties of efficient numerical solvers for (4.5) (and for (4.4) as well) – see the remarks after Theorem 4.3. It will now be shown that the regularized variant (4.5) of (4.3) always has a solution.

[1] If there is no $\hat{x} \in D$ with $F(\hat{x}) = y$ due to discretization errors, then hopefully there will at least exist some $\bar{y} \approx y$ such that a solution $\hat{x} \in D$ of $F(x) = \bar{y}$ exists. If we know $\delta > 0$ such that $\|\bar{y} - y^\delta\|_2 \leq \delta$, we can tacitly add discretization to measurement errors, interpreting $\bar{y}$ as "true" right hand side.

Theorem 4.1 (Existence of a minimizer) *Let $\lambda > 0$, let $D \subseteq \mathbb{R}^n$ be closed and let $F : D \to \mathbb{R}^m$ be continuous. Then the function T_λ defined in (4.5) always has a minimizer $x_\lambda \in D$.*

Proof Since $T_\lambda(x) \geq 0$ for all $x \in D$, the infimum $\mu := \inf\{T_\lambda(x);\ x \in D\}$ exists. Thus for every $n \in \mathbb{N}$ there is an element $x_n \in D$ such that $T_\lambda(x_n) \leq \mu + 1/n$. This means the sequences $(F(x_n))_{n\in\mathbb{N}}$ and $(x_n)_{n\in\mathbb{N}}$ are bounded. By the theorem of Bolzano-Weierstraß there is a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that $x_{n_k} \to \bar{x} \in D$ (D is closed) and $F(x_{n_k}) \to \bar{y} = F(\bar{x})$ (F is continuous). From the continuity of T_λ one concludes $T_\lambda(\bar{x}) = \lim_{k\to\infty} T_\lambda(x_{n_k}) = \mu$, meaning that $x_\lambda = \bar{x}$ is a minimizer. □

The minimizer x_λ of T_λ is stable in the following sense. Let $(y^{\delta_n})_{n\in\mathbb{N}} \subset Y$ be a sequence with $y^{\delta_n} \to y^\delta$ and let $(x_n)_{n\in\mathbb{N}} \subset D$ be a corresponding sequence of minimizers, i.e. x_n minimizes

$$T_{\lambda,n} : D \to \mathbb{R}, \quad x \mapsto T_{\lambda,n}(x) := \|F(x) - y^{\delta_n}\|_2^2 + \lambda\|x - x^*\|_2^2.$$

Then conclude from

$$T_{\lambda,n}(x_n) \leq \left(\|F(x_\lambda) - y^\delta\|_2 + \|y^\delta - y^{\delta_n}\|_2\right)^2 + \lambda\|x_\lambda - x^*\|_2^2 \to T_\lambda(x_\lambda),$$

that $(T_{\lambda,n}(x_n))_{n\in\mathbb{N}}$ is a bounded sequence. As in the proof of Theorem 4.1, this means that $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence. One also concludes that every convergent subsequence of $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer of T_λ. If x_λ is unique, then one even gets $x_n \to x_\lambda$, meaning that in this case the minimization of T_λ is a properly posed problem in the sense of Hadamard (Definition 1.5).

It was assumed that $F(x) = y$ has a solution, so that $\{x \in D;\ F(x) = y\}$ is not empty. Since F is continuous, this set is closed and using the same arguments as in the proof of Theorem 4.1 one concludes that an element

$$\hat{x} \in D \text{ with } F(\hat{x}) = y \text{ and } \|\hat{x} - x^*\|_2 = \min\{\|x - x^*\|_2;\ x \in D,\ F(x) = y\} \tag{4.6}$$

exists. Any such (possibly not unique) $\hat{x}$ is called **x^*-minimum-norm solution** of the equation $F(x) = y$. The following theorem gives conditions for λ under which (4.5) defines a regularization for the computation of an x^*-minimum-norm solution of $F(x) = y$.

Theorem 4.2 (Regularization of x^*-minimum-norm solutions) *Let D be closed and let F be continuous. Let $(\delta_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to zero. For every δ_n choose a parameter $\lambda_n = \lambda_n(\delta_n) > 0$ such that*

$$\lambda_n \to 0 \quad \text{and} \quad \frac{\delta_n^2}{\lambda_n} \to 0 \quad \text{for } n \to \infty. \tag{4.7}$$

For every δ_n let $y^{\delta_n} \in \mathbb{R}^m$ be such that $\|y - y^{\delta_n}\|_2 \le \delta_n$ and let x_n be a minimizer of

$$T_{\lambda_n}(x) := \|F(x) - y^{\delta_n}\|_2^2 + \lambda_n \|x - x^*\|_2^2.$$

Then the sequence $(x_n)_{n\in\mathbb{N}}$ contains a subsequence converging to a x^*-minimum-norm solution $\hat{x}$ of $F(x) = y$. If there is a unique x^*-minimum-norm solution, then $x_n \to \hat{x}$ for $n \to \infty$.

Proof Being a minimizer of T_{λ_n}, the vector x_n fulfills the inequality

$$T_{\lambda_n}(x_n) \le T_{\lambda_n}(\hat{x}) \le \delta_n^2 + \lambda_n \|\hat{x} - x^*\|_2^2 \tag{4.8}$$

(here we made use of $\|F(\hat{x}) - y^{\delta_n}\|_2 = \|y - y^{\delta_n}\|_2 \le \delta_n$, where $\hat{x}$ is any fixed x^*-minimum-norm solution). The right hand side of (4.8) converges to 0, therefore $\|F(x_n) - y\|_2 \le \|F(x_n) - y^{\delta_n}\|_2 + \|y - y^{\delta_n}\|_2 \to 0$ and this means $F(x_n) \to y$. Division of (4.8) by λ_n shows

$$\frac{1}{\lambda_n}\|F(x_n) - y^{\delta_n}\|_2^2 + \|x_n - x^*\|_2^2 \le \frac{\delta_n^2}{\lambda_n} + \|\hat{x} - x^*\|_2^2 \to \|\hat{x} - x^*\|_2^2.$$

Consequently, $(x_n)_{n\in\mathbb{N}}$ is bounded with $\limsup_{n\to\infty} \|x_n - x^*\|_2 \le \|\hat{x} - x^*\|_2$. By the theorem of Bolzano-Weierstraß there exists a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ with $x_{n_k} \to \bar{x} \in D$. Conclude $F(\bar{x}) = y$ from the continuity of F. In addition,

$$\|\bar{x} - x^*\|_2 = \lim_{k\to\infty} \|x_{n_k} - x^*\|_2 \le \limsup_{n\to\infty} \|x_n - x^*\|_2 \le \|\hat{x} - x^*\|_2,$$

meaning that $\bar{x}$ is itself a x^*-minimum-norm solution. It was just shown that any sequence $(x_n)_{n\in\mathbb{N}}$ of minimizers has at least one limit point, which necessarily is a x^*-minimum-norm solution. So when there is exactly one x^*-minimum-norm solution, then the bounded sequence $(x_n)_{n\in\mathbb{N}}$ has exactly one limit point and therefore is convergent. □

The following theorem makes a statement about the convergence rate of the regularized solution in case the regularization parameter is chosen according to a variant of the discrepancy principle. To state this theorem, let us first recall a well known formula for the linearization error of a vector valued smooth function. Let $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ (U open) be two times continuously differentiable. Let $x_0, h \in \mathbb{R}^n$ be such that x_0, $x_0 + h$, and the whole straight line connecting them is contained in U. Then

$$F(x_0 + h) = F(x_0) + F'(x_0)h + r(x_0, h),$$

where $r(x_0, h)$ is vector valued with components

$$r_i(x_0, h) = \sum_{j,k=1}^{n} \left(\int_0^1 \frac{\partial^2 F_i}{\partial x_j \partial x_k}(x_0 + th)(1-t)\, dt \right) h_j h_k, \quad i = 1, \ldots, m. \tag{4.9}$$

Here, h_j, h_k are the components of h.

Theorem 4.3 (Rate of convergence of regularization) *Let $\lambda > 0$, let $D \subseteq U \subseteq \mathbb{R}^n$ with U open and D closed and convex, let $F : U \to \mathbb{R}^m$ be two times continuously differentiable, and let x_0 be a x^*-minimum-norm solution of equation $F(x) = y$. Let $y^\delta \in \mathbb{R}^m$ with $\|y - y^\delta\|_2 \leq \delta$ for some $\delta > 0$. Let x_λ be a minimizer of (4.5). Assume that there exists some $w \in \mathbb{R}^m$ such that*

$$x_0 - x^* = F'(x_0)^T w. \tag{4.10}$$

Set $h := x_\lambda - x_0$, define $r(x_0, h)$ as in (4.9) and assume that for w as in (4.10) one has

$$2|w^T r(x_0, h)| \leq \varrho \|h\|_2^2, \quad \varrho < 1. \tag{4.11}$$

If for some constants $C_1, C_2 > 0$ or $1 \leq \tau_1 \leq \tau_2$ we have

$$C_1\delta \leq \lambda \leq C_2\delta \quad \text{or} \quad \tau_1\delta \leq \|F(x_\lambda) - y^\delta\|_2 \leq \tau_2\delta, \tag{4.12}$$

respectively, then

$$\|x_\lambda - x_0\|_2 \leq C\sqrt{\delta} \tag{4.13}$$

with some constant C. Only a single x^*-minimum-norm solution x_0 can fulfil conditions (4.10) and (4.11).

Proof According to Theorem 4.1 a minimizer x_λ of T_λ exists. For any such minimizer one has

$$T_\lambda(x_\lambda) = \|F(x_\lambda) - y^\delta\|_2^2 + \lambda\|x_\lambda - x^*\|_2^2 \leq T_\lambda(x_0) \leq \delta^2 + \lambda\|x_0 - x^*\|_2^2,$$

because of $F(x_0) = y$ and $\|y - y^\delta\|_2 \leq \delta$. Consequently,

$$\begin{aligned}
&\|F(x_\lambda) - y^\delta\|_2^2 + \lambda\|x_\lambda - x_0\|_2^2 \\
&= \|F(x_\lambda) - y^\delta\|_2^2 + \lambda\left(\|x_\lambda - x^*\|_2^2 + \|x_\lambda - x_0\|_2^2 - \|x_\lambda - x^*\|_2^2\right) \\
&\leq \delta^2 + \lambda\left(\|x_0 - x^*\|_2^2 + \|x_\lambda - x_0\|_2^2 - \|x_\lambda - x^*\|_2^2\right) \\
&= \delta^2 + 2\lambda\left(x_0 - x^*\right)^T (x_0 - x_\lambda) = \delta^2 + 2\lambda w^T \left(F'(x_0)(x_0 - x_\lambda)\right),
\end{aligned}$$

where (4.10) was used to derive the last identity. Using $F(x_0) = y$ one gets

$$F'(x_0)(x_0 - x_\lambda) = \left(y - y^\delta\right) + \left(y^\delta - F(x_\lambda)\right) + \left(F(x_\lambda) - F(x_0) - F'(x_0)(x_\lambda - x_0)\right).$$

Moreover, $F(x_\lambda) - F(x_0) - F'(x_0)(x_\lambda - x_0) = r(x_0, h)$ and using (4.11) one derives from the above inequality

$$\|F(x_\lambda) - y^\delta\|_2^2 + \lambda\|x_\lambda - x_0\|_2^2 \le \delta^2 + 2\lambda\delta\|w\|_2 + 2\lambda\|w\|_2\|F(x_\lambda) - y^\delta\|_2 + \lambda\varrho\|x_\lambda - x_0\|_2^2.$$

This shows the inequality

$$\begin{aligned}\|F(x_\lambda) - y^\delta\|_2^2 + \lambda(1-\varrho)\|x_\lambda - x_0\|_2^2 \\ \le \delta^2 + 2\lambda\delta\|w\|_2 + 2\lambda\|w\|_2\|F(x_\lambda) - y^\delta\|_2,\end{aligned} \tag{4.14}$$

which can also be written in the form

$$\left(\|F(x_\lambda) - y^\delta\|_2 - \lambda\|w\|_2\right)^2 + \lambda(1-\varrho)\|x_\lambda - x_0\|_2^2 \le (\delta + \lambda\|w\|_2)^2. \tag{4.15}$$

Because of $\varrho < 1$, this inequality is true a fortiori, if the first summand on the left hand side is omitted, leading to

$$\|x_\lambda - x_0\|_2 \le \frac{\delta + \lambda\|w\|_2}{\sqrt{\lambda}\cdot\sqrt{1-\varrho}} \le \frac{\delta + C_2\delta\|w\|_2}{\sqrt{C_1}\sqrt{\delta}\cdot\sqrt{1-\varrho}} = C\sqrt{\delta},$$

whenever $C_1\delta \le \lambda \le C_2\delta$, proving (4.13). Otherwise, if $\tau_1\delta \le \|F(x_\lambda) - y^\delta\|_2 \le \tau_2\delta$, then from (4.14) one gets

$$\tau_1^2\delta^2 + \lambda(1-\varrho)\|x_\lambda - x_0\|_2^2 \le \delta^2 + 2\lambda\delta\|w\|_2 + 2\lambda\|w\|_2\tau_2\delta.$$

Using $(1 - \tau_1^2) \le 0$ this shows

$$\lambda(1-\varrho)\|x_\lambda - x_0\|_2^2 \le 2\lambda\|w\|_2(1+\tau_2)\delta$$

and so leads again to the estimate (4.13). Since $C\sqrt{\delta}$ gets arbitrarily small for $\delta \to 0$, a x^*-minimum-norm solution fulfilling the conditions of the theorem necessarily is unique (although multiple x^*-minimum-norm solutions may exist). □

Remarks Condition (4.10) is called **source condition**. It will certainly hold if the functional matrix $F'(x_0)$ has full rank n, which we reasonably expect in the setting of parameter identification problems. The source condition corresponds to some "abstract smoothness" condition for the solution x_0 and thus represents a kind of a priori knowledge about x_0, see for example Section 1.3 in [Kir96] or Section 3.2 in

[EHN96]. Condition (4.11) is fulfilled if either the second derivatives of F are small enough (meaning the function F is only "weakly nonlinear") or if x^* is close to x_0, meaning that x^* must be chosen close to a solution of $F(x) = y$ (one must have a good a priori knowledge of a desired solution). Both requirements are restrictive, but the latter usually is indispensable also from a practical point of view, since in general T_λ is non-convex and possesses stationary points, which are not minimizers. But efficient numerical methods to minimize T_λ are iterative and can only be guaranteed to converge to a stationary point next to where they are started. They will thus fail to detect a global minimizer unless started close enough to one. A final remark concerns the choice of the regularization parameter. It may well happen that T_λ has no unique minimizer. In that case the set

$$M_\lambda := \{\hat{x} \in D;\ T_\lambda(\hat{x}) \leq T_\lambda(x) \text{ for all } x \in D\}$$

contains more than one element. Nevertheless, T_λ takes the same value for all elements $x_\lambda \in M_\lambda$, thus the mapping $\tau : (0, \infty) \to \mathbb{R}_0^+$, $\lambda \mapsto T_\lambda(x_\lambda)$, is in fact a function of λ. Using arguments similar to the ones in the proof of Theorem 4.1, it is not difficult to show that this mapping is continuous. However, the term

$$J(x_\lambda) := \|F(x_\lambda) - y^\delta\|_2 = T_\lambda(x_\lambda) - \lambda\|x_\lambda - x^*\|_2,$$

in general is *not* a function of λ, possibly taking different values for different elements $x_\lambda \in M_\lambda$. Using the same technique as in the proof of Theorem 3.21 it is still possible to show

$$J(x_{\lambda_1}) \leq J(x_{\lambda_2}) \quad \text{for} \quad 0 < \lambda_1 < \lambda_2, \quad x_{\lambda_1} \in M_{\lambda_1}, \quad x_{\lambda_2} \in M_{\lambda_2}. \tag{4.16}$$

Since $J(x_\lambda)$ is not a continuous function of λ (unless T_λ has unique minimizers), one can not guarantee that $J(x_\lambda)$ takes any specific value and the discrepancy principle must be formulated with some care, as in the second alternative considered in (4.12). Still, additional conditions are required to guarantee the existence of some minimizer x_λ such that $\tau_1\delta \leq \|F(x_\lambda) - y^\delta\|_2 \leq \tau_2\delta$. This problem is considered in [Ram02].

4.2 Tikhonov Regularization for Nonlinear Inverse Gravimetry

Problem 1.9 concerning nonlinear inverse gravimetry was discretized in Sect. 2.6. For convenience, we repeat the main points. One has to solve a nonlinear Fredholm equation of the first kind

$$w(x_1, x_2) = \int_{-a}^{a}\int_{-a}^{a} k(x_1, x_2, t_1, t_2, u(t_1, t_2))\, dt_2\, dt_1, \tag{4.17}$$

where the kernel function is defined by

$$k(x_1, x_2, y_1, y_2, z) := \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + z^2}} \tag{4.18}$$

for $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$ and $z > 0$. A unique solution exists, which is a continuous function $\hat{u} : [-a, a]^2 \to [b_1, b_2]$, $0 < b_1 < b_2$. A discrete version of (4.17) is the nonlinear system of equations

$$F(c) = y, \quad c \in C. \tag{4.19}$$

Here, $y \in \mathbb{R}^M$ is a vector with components[2]

$$y_\beta = w(\hat{x}_\beta), \quad \hat{x}_\beta \in [-b, b]^2, \quad \beta \in B,$$

for an index set B comprising M elements. The components c_α of vectors $c \in \mathbb{R}^N$ are interpreted as samples of $u \in C([-a, a]^2)$:

$$c_\alpha = u(x_\alpha) \quad \text{with} \quad x_\alpha = h\alpha, \quad h = \frac{a}{n}, \quad \alpha \in G_n,$$

where

$$G_n = \left\{(\alpha_1, \alpha_2) \in \mathbb{Z}^2;\ -n \le \alpha_j \le n\right\}, \quad |G_n| = N = (2n + 1)^2.$$

Finally, $C = [b_1, b_2]^N$ is an N-dimensional interval and the values

$$F_\beta(c) = \sum_{\alpha \in G_n} \omega_\alpha \left(\|\hat{x}_\beta - x_\alpha\|_2^2 + c_\alpha^2 \right)^{-1/2}, \quad \beta \in B, \tag{4.20}$$

which compose the vector $F(c)$, are the exact integrals (4.17) for

- u the bilinear spline interpolant of (x_α, c_α), $\alpha \in G_n$,
- $(x_1, x_2) = \hat{x}_\beta$, and
- constant factors ω_α defined in (2.110).

[2]To arrange multi-indexed components into a vector, a certain index ordering has to be fixed. In case of rectangular index grids, we will always use a row-wise ordering. For example, the indices $\alpha \in G_n$ for the grid G_n defined below, will always be ordered to form the sequence

$$(-n, -n), \ldots, (-n, n), \quad \ldots, \quad (n, -n), \ldots, (n, n).$$

System (4.19) is to be replaced by the minimization problem

$$\text{minimize} \quad \frac{1}{2}\|y - F(c)\|_2^2, \quad c \in C, \tag{4.21}$$

as discussed in Sect. 2.6. Let us define a vector $\hat{c} \in \mathbb{R}^N$ by

$$\hat{c}_\alpha := \hat{u}(x_\alpha), \quad \alpha \in G_n, \tag{4.22}$$

with $\hat{u}$ the solution of (4.17) for given w. Deviating from above, let us further *define*

$$y := F(\hat{c}),$$

i.e. y will not be the discretized true effect, but a (discretized) effect simulated by $\hat{c}$. This way, the system (4.19) is artificially made to have a solution (namely $\hat{c}$). Perturbed data values

$$y_\beta^\delta \approx y_\beta, \quad \beta \in B,$$

thus contain measurement *and* discretization errors. If we require $\|y^\delta - y\|_2 \le \delta$, then δ has to be an upper bound for both errors together, unless it can be argued that discretization errors are negligible as compared to measurement errors. If a discretization method is convergent (as it should be), the latter will be achievable by choosing a discretization fine enough. To reconstruct an approximation $\tilde{c}$ of $\hat{c}$ from inexact data y^δ, Tikhonov regularization will be used. We experimented with the following objective function

$$Z_\lambda(c) := \|F(c) - y^\delta\|_2^2 + \lambda \left(\rho \|Lc\|_2^2 + \|c - c^*\|_2^2\right), \quad c \in C. \tag{4.23}$$

Here, assuming c contains the values c_α, $\alpha \in G_n$, ordered row-wise

$$c_{(-n,-n)}, \ldots, c_{(-n,n)}, \quad \ldots, \quad c_{(n,-n)}, \ldots, c_{(n,n)},$$

and with

$$A_n := \begin{pmatrix} -1 & 4 & -1 & & \\ & -1 & 4 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 4 & -1 \end{pmatrix} \quad \text{and} \quad B_n := \begin{pmatrix} 0 & -1 & 0 & & \\ & 0 & -1 & 0 & \\ & & \ddots & \ddots & \ddots \\ & & & 0 & -1 & 0 \end{pmatrix}$$

both being $(2n-1) \times (2n+1)$-matrices, set

$$L := \begin{pmatrix} B_n & A_n & B_n & & \\ & B_n & A_n & B_n & \\ & & \ddots & \ddots & \ddots \\ & & & B_n & A_n & B_n \end{pmatrix} \in \mathbb{R}^{(2n-1)^2,(2n+1)^2},$$

such that Lc corresponds to a discrete version of the (negative) Laplace operator $-\Delta$ applied to u and evaluated at the interior grid points x_α. The term $\|Lc\|_2^2$ inversely measures some kind of smoothness of a reconstruction u at the interior grid points. The second regularization term $\|c - c^*\|_2^2$ inversely measures closeness of c (discrete representative of reconstruction u) to some reference c^* (discrete representative of preselected u^*), corresponding to x^* in Sect. 4.1. The weighting factor ρ will be chosen such that both regularization terms $\|Lc\|_2^2$ and $\|c - c^*\|_2^2$ approximately have the same weight (this depends on c and on c^*).

Example 4.4 Let $a = 4$, let $b = 1$, and let $n = 16$, such that $N = 1089$. Let

$$\hat{u}(x) = \left[1 + \frac{1}{10}\cos\left(\frac{\pi x_1}{a}\right)\right]\left[1 + \frac{1}{10}\sin\left(\frac{\pi x_2}{a}\right)\right], \quad x = (x_1, x_2) \in [-a, a]^2$$

be the function which we want to reconstruct and let its samples be given by

$$\hat{c}_\alpha = \hat{u}(h\alpha), \quad \alpha \in G_n, \quad h = a/n.$$

For $m \in \mathbb{N}$, let

$$\hat{h} := b/m, \quad B := \{(\beta_1, \beta_2) \in \mathbb{Z}^2;\ -m \le \beta_j < m\}, \quad \hat{x}_\beta := \hat{h}\beta,\ \beta \in B.$$

Set $m = 64$, such that $M = 16384$, and let

$$y_\beta := F_\beta(c), \quad \beta \in B.$$

These values are samples of an artificial effect *w simulated* by using the bilinear spline interpolant of (x_α, c_α), $\alpha \in G_n$, as function u in (4.17). In contrast to y_β, let

$$w_\beta := \int_{-a}^{a}\int_{-a}^{a} k(\hat{x}_\beta, y, \hat{u}(y))\, dy_2\, dy_1, \quad y = (y_1, y_2),$$

be the samples of the true effect caused by $\hat{u}$. We found

$$\frac{1}{M}\sum_{\beta \in B} |y_\beta - w_\beta|^2 \approx 4 \cdot 10^{-5} \tag{4.24}$$

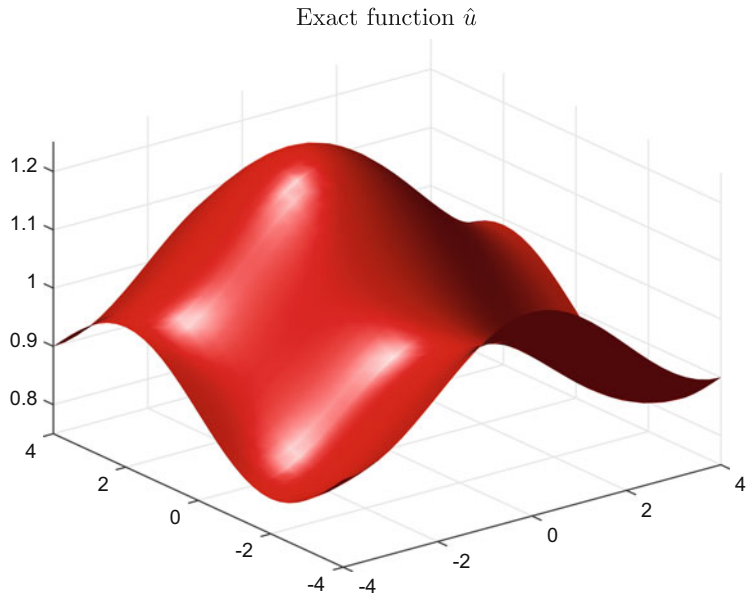

Fig. 4.1 Exact function $\hat{u}$ to be reconstructed

for the mean square discretization error. Values y_β^δ were defined by adding to y_β realizations of independent random variables with normal distribution, having mean value 0 and standard deviation $\sigma = 10^{-2}$. The corresponding mean square error

$$\frac{1}{M}\sum_{\beta\in B}|y_\beta^\delta - y_\beta|^2 \approx \sigma^2 = 10^{-4}$$

is slightly dominating the discretization error (4.24). In Fig. 4.1, the exact function $\hat{u}$ is shown. In Fig. 4.2, the simulated noisy observation is shown, and in Fig. 4.3, the reconstructed function u^δ is shown, which is defined as the bilinear spline interpolant of $(x_\alpha, c_\alpha^\delta)$, $\alpha \in G_n$, where the values c_α^δ were found by numerically determining a minimizer of (4.23), with $\lambda = 10^{-3}$, $\rho = 400$, and $c_\alpha^* = 1$ for all $\alpha \in G_n$. For this choice we observed

$$\rho\|Lc^\delta\|_2^2 \approx 4.3, \quad \|c^\delta - c^*\|_2^2 \approx 8.0, \quad \|F(c^\delta) - y^\delta\|_2^2 \approx 1.636 \approx m^2\sigma^{-4}.$$

The next section will go into some technical details of how to determine a solution of nonlinear least squares problems. In the present example, we used Matlab's optimization function `lsqnonlin`, see also the next section. Actually, the constraints $c_\alpha \in [b_1, b_2]$ were simply ignored in this example. The vector $c = c^*$ with all components equal to 1 was chosen as a start value for `lsqnonlin`, making the iteration stop after 6 iterations. Finally, in Fig. 4.4, the largest 180 singular values

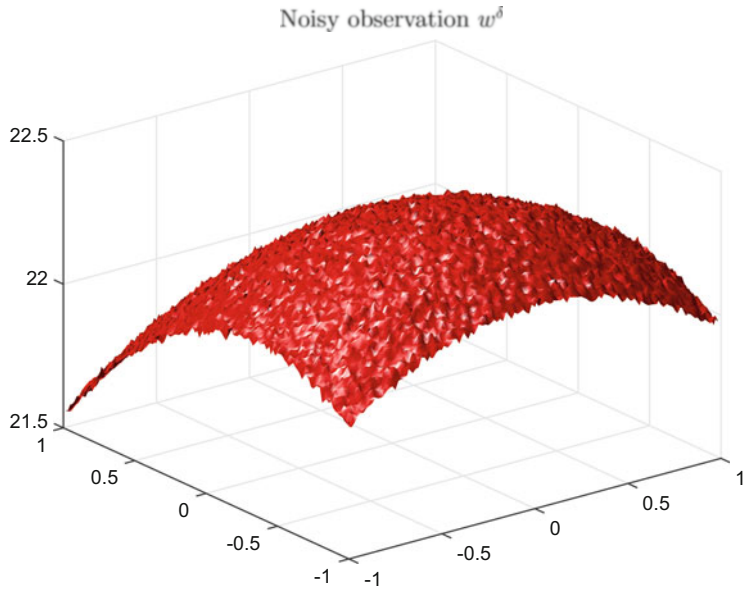

Fig. 4.2 Noisy measurement of gravitational forces

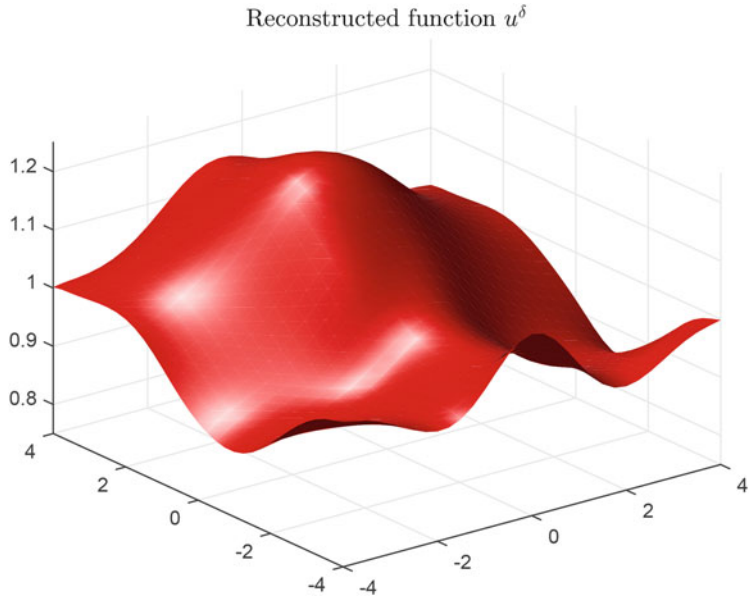

Fig. 4.3 Reconstructed approximation u^δ of $\hat{u}$ from noisy data, $\sigma = 10^{-2}$

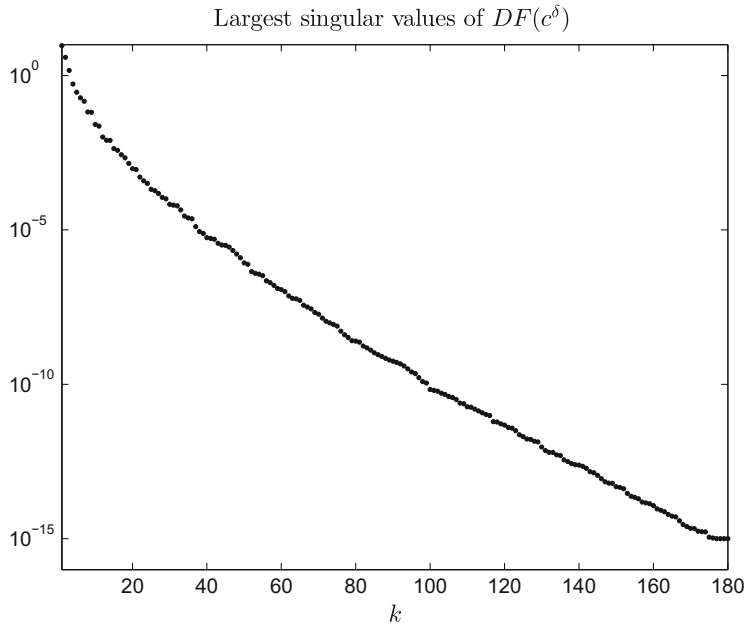

Fig. 4.4 Singular values of Jacobian at optimal point

of the Jacobian $DF(c^\delta) = F'(c^\delta)$ are shown. This tells how badly conditioned this problem is. ◇

4.3 Nonlinear Least Squares Problems

In this section practical methods are considered to solve nonlinear least squares problems of the form

$$\text{minimize} \quad Z(x) := \frac{1}{2}\|F(x) - y\|_2^2, \quad x \in D, \tag{4.25}$$

where $D \subseteq U \subseteq \mathbb{R}^n$ with U open and D closed and convex and where $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is assumed to be two times continuously differentiable. Nonlinear Tikhonov regularization as in (4.4) and (4.5) leads to optimization problems of this kind with

$$Z(x) = \left\| \begin{pmatrix} F(x) \\ \sqrt{\lambda} Lx \end{pmatrix} - \begin{pmatrix} y^\delta \\ 0 \end{pmatrix} \right\|_2^2 \quad \text{or} \quad Z(x) = \left\| \begin{pmatrix} F(x) \\ \sqrt{\lambda} x \end{pmatrix} - \begin{pmatrix} y^\delta \\ \sqrt{\lambda} x^* \end{pmatrix} \right\|_2^2,$$

respectively. In the latter cases an exact global minimizer is searched. Alternatively, one can directly tackle the original, unregularized problem (4.25) and stop an iterative solver prematurely, *before* it has found a minimizer. As in the linear case, this will produce a regularized solution. A specific method will be investigated in Sect. 4.6.

Problem (4.25) can be quite difficult to solve. In the following only some basic algorithmic ideas will be presented. Only the unconstrained case $D = \mathbb{R}^n$ will be treated. The outlined methods can be generalized to work for $D \subset \mathbb{R}^n$, although this requires some effort. Consult [BCL99] for the important case where D is a box (a multi-dimensional interval). We will also not deal with a very serious limitation of most actual solvers, namely that they only detect stationary points of Z, and not global minima. In Chapter 3 of [Cha09], Chavent proposes a way to overcome this limitation for certain classes of "weakly nonlinear" problems.

One idea to solve (4.25) for $D = \mathbb{R}^n$ is to use **Newton's method**. Gradient and Hessian of Z are easily computed

$$\begin{aligned} \nabla Z(x) &= F'(x)^T(F(x) - y), \\ \nabla^2 Z(x) &= F'(x)^T F'(x) + \sum_{i=1}^{m} (F_i(x) - y_i)\nabla^2 F_i(x). \end{aligned} \tag{4.26}$$

Here, F_i means the ith component of the vector field F, $\nabla^2 F_i$ is the corresponding Hessian and $F'(x)$ is the Jacobian of F evaluated at x. Newton's method tries to find some $\hat{x}$ with $\nabla Z(\hat{x}) = 0$. Beginning at a **start value** $x^0 \in D$, the original Newton method successively computes approximations x^i of $\hat{x}$, defined by the rule

$$x^{i+1} := x^i + s \quad \text{where} \quad \nabla^2 Z(x^i)s = -\nabla Z(x^i), \quad i = 0, 1, 2, \ldots$$

This means that s is a zero of the linearization

$$\nabla Z(x^i + s) \overset{\bullet}{=} \nabla Z(x^i) + \nabla^2 Z(x^i)s.$$

One obtains a so-called Quasi-Newton method if the Hessian $\nabla^2 Z(x)$ is approximated by some other matrix. The choice

$$\nabla^2 Z(x) \approx F'(x)^T F'(x)$$

suggests itself. This is a reasonable approximation if it can be assumed that $F(x) = y$ is a "nearly consistent" system of equations, unsolvable only because of perturbed data y. It is to be expected that in such a situation (for x close to a solution $\hat{x}$) the contributions of the Hessians $\nabla^2 F_i(x)$ to $\nabla^2 Z(x)$ in (4.26) are damped by small values $|F_i(x) - y_i|$. Using the approximation $\nabla^2 Z(x) \approx F'(x)^T F'(x)$, one gets the **Gauß-Newton method**. Starting from an approximation x^i of $\hat{x}$ it computes a next approximation x^{i+1} according to the following rules:

Gauß-Newton step

(a) Compute $b := F(x^i) - y$ and $J := F'(x^i)$.
(b) Solve $J^T b + J^T J s = 0$ for s.
(c) Set $x^{i+1} := x^i + s$.

The same method results if F is linearized at x^i:

$$F(x^i + s) - y \overset{\bullet}{=} F(x^i) - y + F'(x^i)s = b + Js$$

and if then the *linear* least squares problem

$$\text{minimize} \quad \|b + Js\|_2, \quad s \in \mathbb{R}^n,$$

is solved. The vector s is called **search direction**. In case $\text{rank}(J) = n$, the matrix $J^T J$ is positive definite and so is $(J^T J)^{-1}$. Consequently

$$-\nabla Z(x^i)^T s = -b^T J s = b^T J (J^T J)^{-1} J^T b > 0$$

(unless $b = 0$, which means that $F(x^i) = y$). Consequently, s points into a direction of decreasing values of Z, as does $-\nabla Z(x^i)$, the negative gradient.[3] But this does not yet mean $Z(x^{i+1}) < Z(x^i)$, since the step might go into a good direction, but take us too far across a valley and uphill again on the opposite side. To guarantee a descent, part (c) of the above rule needs to be modified by a so-called **step size control**:

$$\textit{Do not set} \quad x^{i+1} = x^i + s, \quad \textit{but rather} \quad x^{i+1} = x^i + \mu s,$$

where $\mu \in (0, 1]$ is chosen such that $Z(x^{i+1}) < Z(x^i)$. Such a value μ always exists and there are a number of well-know algorithms to find one, see [LY08], Section 8.5.

A different idea is to keep (c), but to replace (b) in the above rule by defining s as the solution of the optimization problem

$$\text{minimize} \quad \|b + Js\|_2 \quad \text{under the constraint} \quad \|s\|_2 \le \Delta,$$

where Δ is a parameter to be chosen. The corresponding algorithm is known as **Levenberg-Marquardt method** and is also called a **trust region method**. The constraint defines a ball of radius Δ, centered at x^i, where the linearization is trusted to be a good approximation of F. More generally, one can consider the optimization

[3] This is generally true even if $\text{rank}(J) = n$ does not hold, provided only that $\nabla Z(x^i) \neq 0$, see [Bjö96], p. 343.

problem

$$\text{minimize} \quad \|b + Js\|_2 \quad \text{under the constraint} \quad \|Ds\|_2 \leq \Delta \tag{4.27}$$

with some positive definite (diagonal) matrix D. This modification means to distort the ball where the linearization $F(x^i + s) - y \stackrel{\bullet}{=} b + Js$ is trusted to allow for different change rates of F in different directions. The parameters D and Δ have to be changed as the iteration proceeds – details are described in [Mor78]. Using the method of Lagrange multipliers, (4.27) is seen to be equivalent to the linear system

$$(J^T J + \lambda D^T D)s = -J^T b \tag{4.28}$$

where λ is the Lagrange parameter corresponding to Δ. From (4.28) it can be seen, that a search direction s will be found which is different from the one in the Gauß-Newton method. The parameter λ has to be found iteratively as described in Sect. 3.4. This means a considerable arithmetical effort, if the dimension n is high. To cut down the costs, variants of (4.27) have been developed, which look for a solution s of (4.27) only in a carefully chosen, two-dimensional subspace of $\mathbb{R}^n$. This is described in [BCL99].

Matlab provides a function `lsqnonlin` implementing the algorithm described in [BCL99]. It is possible to include box constraints, i.e. to solve (4.25) with

$$D = \{x \in \mathbb{R}^n;\ \ell_i \leq x_i \leq r_i,\ i = 1, \ldots, n\},$$

where $-\infty \leq \ell_i < r_i \leq \infty$ for $i = 1, \ldots, n$. As mentioned, function `lsqnonlin` was used in Example 4.4. When using `lsqnonlin` (or any other gradient based minimization routine), derivatives have to be computed. In Example 4.4, the Jacobian J required in (4.27) can easily be computed from (4.20):

$$\frac{\partial F_\beta(c)}{\partial c_\alpha} = -\omega_\alpha c_\alpha \left(\|\hat{x}_\beta - x_\alpha\|_2^2 + c_\alpha^2 \right)^{-3/2}.$$

In other cases, gradient information may be much more difficult to obtain. This is the case for the mapping F defined in (2.146), which describes the nonlinear problem of seismic tomography. Derivatives could still be approximated by finite differences as in

$$\frac{\partial Z(x)}{\partial x_j} \approx \frac{Z(x + he^j) - Z(x)}{h},$$

where e^j is the jth canonical unit vector in $\mathbb{R}^n$. But besides being imprecise, finite differencing can mean a huge numerical effort for large n and therefore often is prohibitive. In the following section, an efficient way to compute gradient information for certain kinds of mappings (like F defined in (2.146)) is described, known as the **adjoint method**.

4.4 Computation of Derivatives by the Adjoint Method

The adjoint method will be presented following Chapter 2 of [Cha09], but restricted here to the finite-dimensional case. The method is not new – Chavent already used it in 1974, see [Cha74].

The task is to compute the gradient ∇Z of the objective function Z defined in (4.25), or the Jacobian $F'(x)$ of the corresponding function $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$. It will be assumed that a function evaluation $v = F(x)$ can be decomposed into two steps. *First*, given $x \in D$, solve an equation

$$e(x, z) = 0 \quad \text{for} \quad z \in \mathbb{R}^p \tag{4.29}$$

and *second*, compute

$$v = M(z) \quad [= F(x)]. \tag{4.30}$$

Here, $e(x, z) = 0$ is called **state equation**, defined by a mapping $e : D \times \mathbb{R}^p \to \mathbb{R}^p$, and $M : \mathbb{R}^p \to \mathbb{R}^m$ is an **observation operator**.

Example 4.5 In Problem 1.11 of nonlinear seismic tomography, as presented in its discrete version in Sect. 2.6, we have to deal with a function

$$F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n, \quad \sigma \mapsto Y,$$

where $D = [\sigma_-, \sigma_+]^n$, and where $\sigma := (\sigma_2, \ldots, \sigma_{n+1}) \in D$ plays the role of the parameter vector x – see (2.146). We now show that an evaluation of $F(\sigma)$ can be decomposed into two steps according to (4.29) and (4.30). Recall that σ_1 is assumed to be known and therefore is not considered a sought-after parameter. From the one to one relations

$$r_0 := 0, \quad r_k := \frac{\sigma_k - \sigma_{k+1}}{\sigma_k + \sigma_{k+1}}, \quad k = 1, \ldots, n, \tag{2.127}$$

between $(r_1, \ldots, r_n)$ and $(\sigma_2, \ldots, \sigma_{n+1})$, it is easily seen that the vector $r = (r_1, \ldots, r_n)$ can equally well serve as a model parameter. Corresponding to $r \in \mathbb{R}^n$ (and thus to $\sigma \in \mathbb{R}^n$), there is a vector of $p = (n + 1)(n + 2) - 2$ state variables

$$z = (u_{1,0}, u_{1,1}, \ldots, u_{n,0}, \ldots, u_{n,n}, v_{1,0}, v_{1,1}, \ldots, v_{n,0}, \ldots, v_{n,n}) \in \mathbb{R}^p,$$

defined by the linear equations

$$\begin{aligned} u_{i,j} &= (1 - r_{i-j})u_{i,j-1} + r_{i-j}v_{i-1,j} \\ v_{i,j} &= -r_{i-j}u_{i,j-1} + (1 + r_{i-j})v_{i-1,j} \end{aligned} \qquad i = 1, \ldots, n,\ j = 0, \ldots, i, \tag{2.128}$$

and by the boundary and initial conditions

$$\begin{array}{ll} u_{i,-1} = 0 \\ v_{i-1,i} = u_{i,i} \end{array} \quad i = 1, \ldots, n, \qquad \begin{array}{l} u_{0,0} = 0 \\ v_{0,0} = 2 \end{array}. \tag{2.129}$$

Note that the state variables can be computed in the order given in (2.128) (outer loop i, inner loop j, first $u_{i,j}$, then $v_{i,j}$), since $v_{i-1,i}$ is not needed for the computation of $u_{i,i}$ because of $r_{i-i} = r_0 = 0$. With $u_{i,i}$ computed and $v_{i-1,i} := u_{i,i}$ set according to (2.129), (2.128) can be continued with the computation of $v_{i,i}$. Equations (2.127), (2.128), and (2.129) evidently uniquely define all needed values $u_{i,j}$ and $v_{i,j}$ for $i = 1, \ldots, n$ and $j = 0, \ldots, i$, and can be written as a system of equations of the form $e(x, z) = 0$ for $x = \sigma$. Finally, the mapping M in this example is given as the concatenation of the two mappings $z \mapsto \lambda$ defined in (2.130) and $\lambda \mapsto Y$ defined in (2.145), where $Y := (Y_2, \ldots, Y_{n+1})$, namely

$$M : \mathbb{R}^p \to \mathbb{R}^n, \quad (u_{1,0}, \ldots, v_{n,n}) \mapsto \frac{1}{\sigma_1} \begin{pmatrix} g_2 + u_{1,1} g_1 \\ g_3 + u_{1,1} g_2 + u_{2,2} g_1 \\ \vdots \\ g_k + \sum_{i=1}^{n} u_{i,i} g_{k-i} \\ \vdots \end{pmatrix}. \tag{4.31}$$

The constant values g_j are defined in (2.142) with $g_1 \neq 0$ and $g_{k-i} = 0$ for $i \geq k$. $\diamond$

Assumption 4.6 *Concerning e and M in (4.29) and (4.30), it will always be required that*

- *Equation (4.29) has a unique solution $z \in \mathbb{R}^p$ for every $x \in D$,*
- *$e : D \times \mathbb{R}^p \to \mathbb{R}^p$ is continuously differentiable in all points of $\mathring{D} \times \mathbb{R}^p$, where $\mathring{D}$ is the set of all interior points of D,*
- *the matrix $\partial e(x, z)/\partial z \in \mathbb{R}^{p,p}$ is invertible for all $x \in \mathring{D}$, and that*
- *the mapping M is continuously differentiable.*

The first requirement means that (4.29) defines a mapping $D \mapsto \mathbb{R}^p$, $x \mapsto z = z(x)$. From the next two requirements it follows (by virtue of the implicit function theorem) that this mapping is continuously differentiable with Jacobian

$$z'(x) = -\left[\frac{\partial e(x, z)}{\partial z}\right]^{-1} \frac{\partial e(x, z)}{\partial x} \in \mathbb{R}^{p,n}$$

for all $x \in \mathring{D}$. It is easy to check that all requirements are fulfilled in Example 4.5. Now assume further that a function

$$G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, \quad G \text{ continuously differentiable},$$

is given and that the gradient of $x \mapsto G(x, F(x))$ with F from (4.25) shall be computed. This gradient will be written as ∇G for short and its computation is of interest for special choices of G. If one chooses

$$G(x, v) := \frac{1}{2}\|v - y\|_2^2, \tag{4.32}$$

(the function values are actually independent of x in this case), then

$$G(x, F(x)) = Z(x) \text{ for } x \in D, \text{ with } Z \text{ from (4.25)}.$$

In this case $\nabla G = \nabla Z$, the gradient of Z as in (4.26). If, on the other hand, one chooses

$$G(x, v) = v^T w, \quad w \in \mathbb{R}^m \quad \text{(fixed)}, \tag{4.33}$$

then

$$G(x, F(x)) = F(x)^T w$$

and $\nabla G = F'(x)^T w$. In the special case $w = e^j$, with e^j the jth unit vector from $\mathbb{R}^m$, this would mean to compute the jth column of $F'(x)^T$, which is the jth row of $F'(x)$.

Theorem 4.7 (Computation of derivatives by the adjoint method) *Let* (4.29) *and* (4.30) *define a decomposition of the mapping* $F : D \to \mathbb{R}^m$ *and let Assumption* 4.6 *hold. Further let* $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *be continuously differentiable and define*

$$L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}, \quad L(x, z, \lambda) := G(x, M(z)) + e(x, z)^T \lambda. \tag{4.34}$$

Then the mapping $D \to \mathbb{R}$, $x \mapsto G(x, F(x))$, *is continuously differentiable on* $\mathring{D}$, *and its gradient* ∇G *is given by*

$$\nabla G = \frac{\partial L}{\partial x}(x, z, \lambda) = \frac{\partial G}{\partial x}(x, M(z)) + \frac{\partial e}{\partial x}(x, z)^T \lambda, \tag{4.35}$$

where z *is the unique solution of the state equation* (4.29) *and where* $\lambda \in \mathbb{R}^p$ *is the unique solution of the so-called* ***adjoint state equation***

$$\frac{\partial L}{\partial z}(x, z, \lambda) = 0. \tag{4.36}$$

Proof From Assumption 4.6 it follows, as already noted above, that $e(x, z) = 0$ defines a mapping $D \to \mathbb{R}^p$, $x \mapsto z = z(x)$, which is continuously differentiable on $\mathring{D}$. One therefore has $F(x) = M(z(x))$ and $G(x, F(x)) = G(x, M(z(x)))$, the latter being a continuously differentiable function on $\mathring{D}$, as follows from the corresponding

assumptions on M and G. Furthermore, for $z = z(x)$, one has $e(x, z) = 0$ and the definition of L from (4.34) reduces to

$$L(x, z(x), \lambda) = G(x, F(x)) \quad \text{for all} \quad x \in D,\ \lambda \in \mathbb{R}^p.$$

Differentiating this identity with respect to x, one gets

$$\nabla G = \frac{\partial L}{\partial x}(x, z, \lambda) + z'(x)^T \cdot \frac{\partial L}{\partial z}(x, z, \lambda), \quad x \in \mathring{D}.$$

This shows that (4.35) is true, if (4.36) holds. It remains to show that for any $x \in \mathring{D}$ and $z = z(x)$, there is a unique $\lambda \in \mathbb{R}^p$ solving (4.36). Differentiating, one obtains

$$\begin{aligned} \frac{\partial L}{\partial z}(x, z, \lambda) &\overset{(4.34)}{=} \frac{d}{dz}\left[G(x, M(z)) + e(x, z)^T \lambda\right] \\ &= M'(z)^T \cdot \frac{\partial G}{\partial v}(x, M(z)) + \frac{\partial e}{\partial z}(x, z)^T \lambda. \end{aligned}$$

Here, $\partial G/\partial v$ means differentiation of G with respect to its second argument. By assumption (4.6), the matrix $\partial e(x, z)/\partial z$ is invertible and thus (4.36) can indeed be uniquely solved for λ. □

Remark 1 The function L is the Lagrangian to be used for finding an extremum of $G(x, M(z))$ under the constraint $e(x, z) = 0$, which is the same as finding an extremum of $G(x, F(x))$.

Remark 2 It was shown above, using the definition (4.33) for G, how the adjoint method can be used to compute the Jacobian $F'(x)$ row by row. A much more straightforward way to compute this matrix column by column is by direct differentiation

$$\frac{\partial F}{\partial x_j} = M'(z)\frac{\partial z}{\partial x_j}. \tag{4.37}$$

Further, implicit differentiation of $e(x, z) = 0$ leads to

$$\frac{\partial e}{\partial x_j}(x, z) + \frac{\partial e}{\partial z}(x, z) \cdot \frac{\partial z}{\partial x_j} = 0. \tag{4.38}$$

This is a linear system of equations, which can be solved for the vector $\partial z/\partial x_j$, to be used afterwards in (4.37) to compute $\partial F/\partial x_j$, the jth column of $F'(x)$. To get all columns of $F'(x)$, one has to solve n systems of equations of the form (4.38), which is the main effort with this approach. This effort must also be spent if one is only interested in getting the gradient $\nabla Z = F'(x)^T(F(x) - y)$. On the other hand, ∇Z can be computed by the adjoint method for the choice (4.32) of G by solving only a single adjoint state equation (4.36).

Remark 3 Equation (4.36) can equivalently be stated in "variational form", setting all possible directional derivatives to zero:

$$\left[\frac{\partial L}{\partial z}(x, z, \lambda)\right]^T \delta z = 0 \quad \text{for all} \quad \delta z \in \mathbb{R}^p. \tag{4.39}$$

Directional derivatives are also known under the name of "first variations", whence the name "variational form".

Example 4.8 For our discretized model problem of nonlinear seismic tomography, the forward problem is given by the mapping

$$F : D = [\sigma_-, \sigma_+]^n \subset \mathbb{R}^n \to \mathbb{R}^n$$

defined in Sect. 2.6 as concatenation of the mappings

$$\underbrace{(\sigma_2, \ldots, \sigma_{n+1})}_{=\sigma} \overset{(2.127)}{\mapsto} \underbrace{(r_1, \ldots, r_n)}_{=r} \overset{(2.128)-(2.130)}{\mapsto} \underbrace{(\lambda_2, \ldots, \lambda_{n+1})}_{=\lambda} \overset{(2.143)}{\mapsto} \underbrace{(Y_2, \ldots, Y_{n+1})}_{=Y}.$$

Below, we will continue to use the same indices as in Sect. 2.6, such that σ_2 means the *first* component of vector σ and equally for λ and Y. The Jacobian of F has components

$$\frac{\partial Y_k}{\partial \sigma_\ell} = \frac{\partial Y_k}{\partial r_{\ell-1}} \cdot \frac{\partial r_{\ell-1}}{\partial \sigma_\ell} + \frac{\partial Y_k}{\partial r_\ell} \cdot \frac{\partial r_\ell}{\partial \sigma_\ell}, \quad k, \ell = 2, \ldots, n+1, \tag{4.40}$$

where we made use of (2.127), which directly gets us

$$\frac{\partial r_{\ell-1}}{\partial \sigma_\ell} = \frac{-2\sigma_{\ell-1}}{(\sigma_{\ell-1} + \sigma_\ell)^2}, \ \ell = 2, \ldots, n+1, \qquad \frac{\partial r_\ell}{\partial \sigma_\ell} = \frac{2\sigma_{\ell+1}}{(\sigma_\ell + \sigma_{\ell+1})^2}, \ \ell = 1, \ldots, n. \tag{4.41}$$

Moreover, $\partial r_i / \partial \sigma_\ell = 0$ for all indices i not listed in (4.41). It remains to compute derivatives $\partial Y_k / \partial r_\ell$, $k = 2, \ldots, n+1$, $\ell = 1, \ldots, n$, which will be done using the adjoint method. Preparations were done in Example 4.5. We will use $x = r$ (not $x = \sigma$) as parameter vector. The state variable $z \in \mathbb{R}^p$ has $p = (n+1)(n+2) - 2$ components $u_{1,0}, u_{1,1}, \ldots, u_{n,0}, \ldots, u_{n,n}$ and $v_{1,0}, v_{1,1}, \ldots, v_{n,0}, \ldots, v_{n,n}$. The state equation $e(x, z) = 0$ is given by (2.128) and (2.129) and the mapping M is given by (4.31).

To facilitate identification with the formulae introduced in Sect. 2.6, we will continue to use indices $k = 2, \ldots, n+1$ below (Y has first component Y_2, for example). To compute the kth row of $\partial Y / \partial r$, we then have to define $G(x, v) = v^T e^{k-1}$, with e^{k-1} the $(k-1)$th unit vector of $\mathbb{R}^n$, as suggested by (4.33), and set up

the Lagrange function (see remarks after this expression!)

$$L(x,z,\lambda) = M(z)^T e^{k-1} + e(x,z)^T\lambda = \left(g_k + \sum_{i=1}^{n} u_{i,i} g_{k-i}\right)/\sigma_1 +$$
$$+ \sum_{i=1}^{n}\sum_{j=0}^{i-1}\left[u_{i,j} - (1-r_{i-j})u_{i,j-1} - r_{i-j}v_{i-1,j}\right]p_{i,j} + \sum_{i=1}^{n}\left[u_{i,i} - u_{i,i-1}\right]p_{i,i}$$
$$+ \sum_{i=1}^{n}\sum_{j=0}^{i-1}\left[v_{i,j} + r_{i-j}u_{i,j-1} - (1+r_{i-j})v_{i-1,j}\right]q_{i,j} + \sum_{i=1}^{n}\left[v_{i,i} - u_{i,i}\right]q_{i,i}.$$

Here, the components of z are called $u_{i,j}$ and $v_{i,j}$ as in Example 4.5. Further, the components of the Lagrange parameter λ are called $p_{i,j}$ and $q_{i,j}$. Finally, the boundary condition $v_{i-1,i} = u_{i,i}$ was eliminated by splitting (2.128) into parts for $j < i$ and $j = i$. Now the adjoint equation (4.36) is set up in its variational form (4.39):

$$\frac{\partial L}{\partial z}\cdot\delta z = \sum_{i=1}^{n}\delta u_{i,i}\, g_{k-i}/\sigma_1 +$$
$$+ \sum_{i=1}^{n}\sum_{j=0}^{i-1}\left[\delta u_{i,j} - (1-r_{i-j})\,\delta u_{i,j-1} - r_{i-j}\,\delta v_{i-1,j}\right]p_{i,j} + \sum_{i=1}^{n}\left[\delta u_{i,i} - \delta u_{i,i-1}\right]p_{i,i}$$
$$+ \sum_{i=1}^{n}\sum_{j=0}^{i-1}\left[\delta v_{i,j} + r_{i-j}\,\delta u_{i,j-1} - (1+r_{i-j})\,\delta v_{i-1,j}\right]q_{i,j} + \sum_{i=1}^{n}\left[\delta v_{i,i} - \delta u_{i,i}\right]q_{i,i},$$

where $\delta u_{i,-1} := 0$ and $\delta v_{0,0} := 0$. Reordering gives

$$\frac{\partial L}{\partial z}\cdot\delta z = \sum_{i=1}^{n}\delta u_{i,i}\left[g_{k-i}/\sigma_1 + p_{i,i} - q_{i,i}\right] +$$
$$+ \sum_{i=1}^{n-1}\delta v_{i,i}\left[q_{i,i} - r_1 p_{i+1,i} - (1+r_1)q_{i+1,i}\right] + \delta v_{n,n}\, q_{n,n}$$
$$+ \sum_{i=1}^{n}\sum_{j=0}^{i-1}\delta u_{i,j}\left[p_{i,j} - (1-r_{i-j-1})p_{i,j+1} + r_{i-j-1}q_{i,j+1}\right]$$
$$+ \sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\delta v_{i,j}\left[q_{i,j} - r_{i+1-j}p_{i+1,j} - (1+r_{i+1-j})q_{i+1,j}\right] + \sum_{j=0}^{n-1}\delta v_{n,j}\, q_{n,i}.$$

Requiring $(\partial L/\partial z)\delta z = 0$ for all possible δz therefore means to require that the recursive relations

$$\begin{aligned} q_{i,j} &= r_{i-j+1}p_{i+1,j} + (1 + r_{i-j+1})q_{i+1,j}, \quad i = 1,\dots,n-1,\ j = 0,\dots,i, \\ p_{i,j} &= (1 - r_{i-j-1})p_{i,j+1} - r_{i-j-1}q_{i,j+1}, \quad i = 1,\dots,n,\ \ j = 0,\dots,i-1, \end{aligned} \tag{4.42}$$

must hold together with the boundary conditions

$$\begin{aligned} q_{n,j} &= 0, & j &= 0,\dots,n, \\ p_{i,i} - q_{i,i} &= -g_{k-i}/\sigma_1, & i &= 1,\dots,n. \end{aligned} \tag{4.43}$$

A solution of these equations is possible by computation in the order $i = n,\dots,1$ (outer loop), $j = i,\dots,0$ (inner loop). This defines the components of λ. According to (4.35), the kth row (index $k = 2$ for the first row, as above!) of the Jacobian is given by

$$\nabla G = \frac{L}{\partial x}(x,z,\lambda) = \frac{\partial}{\partial x}\left(M(z)^T e^{k-1}\right) + \frac{\partial e}{\partial x}(x,z)^T\lambda.$$

The first term is equal to zero because $M(z)^T e^{k-1}$ does not depend on x. The second vector has ℓth component ($\ell = 2$ for first component!)

$$\begin{aligned} \frac{\partial Y_k}{\partial r_\ell} &= \frac{\partial e}{\partial r_\ell}^T \lambda = \sum_{\substack{i=1 \\ i-j=\ell}}^{n} \sum_{j=0}^{i-1} \left[(u_{i,j-1} - v_{i-1,j})p_{i,j} + (u_{i,j-1} - v_{i-1,j})q_{i,j}\right] \\ &= \sum_{\substack{i=1 \\ i-j=\ell}}^{n} \sum_{j=0}^{i-1} (u_{i,j-1} - v_{i-1,j})(p_{i,j} + q_{i,j}). \end{aligned} \tag{4.44}$$

The only place where the index k enters is (4.43). ◇

4.5 Tikhonov Regularization for Nonlinear Seismic Tomography

Problem 1.11 of nonlinear seismic tomography was discretized in Sect. 2.6. The forward mapping $F : D = [\sigma_-, \sigma_+]^n \subset \mathbb{R}^n \to \mathbb{R}^n$ defining this problem as a nonlinear system of equations $F(\sigma) = Y$ to be solved was re-considered in Example 4.8 of Sect. 4.4, where it was shown how the Jacobian $F'(\sigma)$ of F at σ can be computed. Only perturbed data $Y^\delta \approx Y$ are given and the system of equations

will be replaced by the more general minimization problem

$$\min_{\sigma \in D} \frac{1}{2} \|F(\sigma) - Y^\delta\|_2^2, \tag{4.45}$$

which always has a solution. Problem (4.45) will be regularized by replacing it in turn by the following minimization problem

$$\min_{\sigma \in D} \frac{1}{2} \|F(\sigma) - Y^\delta\|_2^2 + \frac{\alpha}{2} G(\sigma), \tag{4.46}$$

where

$$G(\sigma) = \sum_{i=1}^{n} |\sigma_{i+1} - \sigma_i|. \tag{4.47}$$

Here, as in Sect. 2.6, we assume that σ_1 is a known acoustic impedance to be treated as a fixed parameter, whereas $\sigma = (\sigma_2, \ldots, \sigma_{n+1})$. The term $G(\sigma)$ is known as the **total variation** of the piecewise constant acoustic impedance function $\tilde{\sigma} \in \mathscr{S}_n$ defined by the $(n+1)$-dimensional vector $(\sigma_1, \ldots, \sigma_{n+1})$, see (2.119) and (2.120). Optimization problems of the form (4.46) were mentioned as a generalization of standard Tikhonov regularization in Sect. 3.4, see (3.66). Total variation as a regularization term was introduced for "image denoising" by Rudin, Osher, and Fatemi in [ROF92], but was already recommended in the present case by Bamberger, Chavent, and Lailly in [BCL77]. In contrast to regularization terms measuring smoothness by taking first or second derivatives, total variation can be defined for step functions. Total variation as a regularization term is useful, if a function shall be reconstructed which is known not be oscillatory, but which may contain jumps. A disadvantage of (4.47) is its non-differentiability, which complicates the solution of (4.46). Either minimization algorithms have to be used which can deal with non-differentiable objective functions or one uses an approximation of G by a smooth functions. We opted for the latter approach and set

$$G_\beta : D \to \mathbb{R}^n, \quad \sigma \mapsto \begin{pmatrix} \sqrt[4]{(\sigma_2 - \sigma_1)^2 + \beta} \\ \sqrt[4]{(\sigma_3 - \sigma_2)^2 + \beta} \\ \vdots \\ \sqrt[4]{(\sigma_{n+1} - \sigma_n)^2 + \beta} \end{pmatrix} \quad \text{for} \quad \beta > 0 \quad \text{"small"}, \tag{4.48}$$

such that

$$G(\sigma) \approx \|G_\beta(\sigma)\|_2^2.$$

In the following examples, $\beta = 10^{-12}$ was used. The minimization problem (4.46) can now be approximated by a standard nonlinear least squares problem, namely

$$\min_{\sigma \in D} \frac{1}{2}\|F(\sigma) - Y^\delta\|_2^2 + \frac{\alpha}{2}\|G_\beta(\sigma)\|_2^2 \quad = \quad \min_{\sigma \in D} \frac{1}{2}\left\| \begin{pmatrix} F(\sigma) - Y^\delta \\ \sqrt{\alpha} G_\beta(\sigma) \end{pmatrix} \right\|_2^2. \tag{4.49}$$

For the practical solution of this minimization problem in all the following examples, Matlab's function `lsqnonlin` was used. Since we never encountered a case where a negative impedance value σ_i was computed, the optimization in fact was carried out over all of $\mathbb{R}^n$, no constraint of the form $\sigma \in D$ was used. Function `lsqnonlin` in all examples was provided with constant start values $\sigma_2 = \ldots = \sigma_{n+1} = \sigma_1$, the latter assumed to be known, as required above.

Example 4.9 Let $n = 300$, let $T = 2.24$, and let $\hat{\sigma} \in \mathscr{S}_n$ be the piecewise constant acoustic impedance shown in Fig. 4.5 as a black line. From these impedance data, an exact (up to rounding errors) seismic record $Y = (Y_2, \ldots, Y_{n+1})$ was computed, solving the corresponding hyperbolic equation by the method of characteristics, as described in Sect. 2.6, Theorem 2.14. The same Ricker pulse function g with $a = 1$ and $f_0 = 5$ as in Example 2.7 of Sect. 2.2 was used to do so, see (2.30). The corresponding seismogram is shown in Fig. 4.6 as a black line. We first attempted to reconstruct $\hat{\sigma}$ from these exact data Y, setting $\alpha = 0$ in (4.49), i.e. using *no regularization*. The resulting impedance function $\tilde{\sigma}$ is shown in Fig. 4.5 as a red line. Then, the reconstructed $\tilde{\sigma}$ was used to exactly simulate *its* seismic record $\tilde{Y}$. The corresponding error $Y - \tilde{Y}$ is shown in Fig. 4.6 as a red line. The error is not distinguishable from zero, which means that the very different impedances $\hat{\sigma}$ and

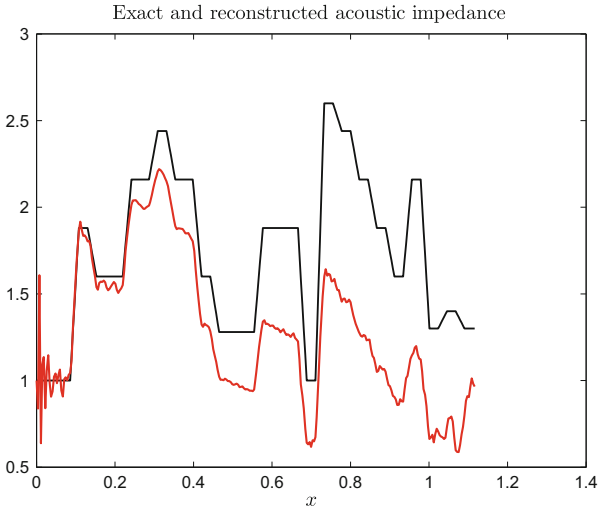

Fig. 4.5 Exact and reconstructed acoustic impedance, no noise, no regularization

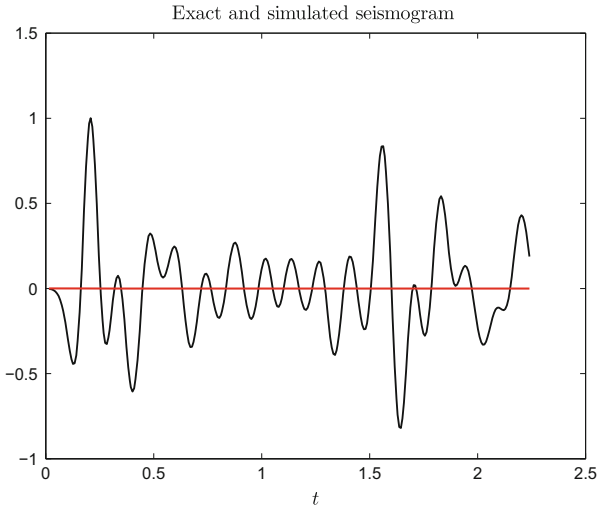

Fig. 4.6 Exact and reproduced seismic record, no noise, no regularization

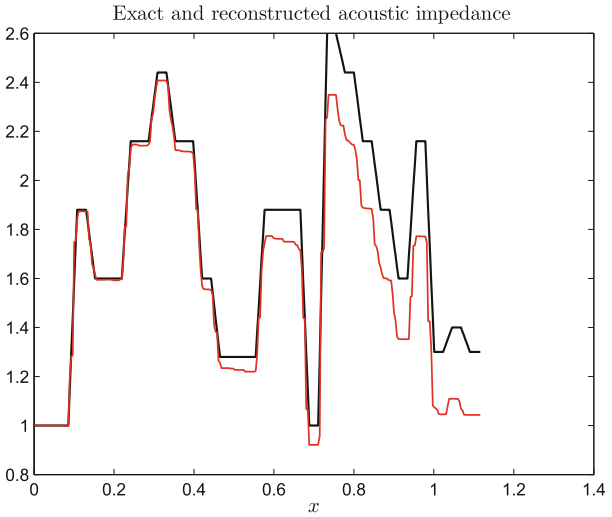

Fig. 4.7 Exact and reproduced impedance, no noise, $\alpha = 2 \cdot 10^{-7}$

$\tilde{\sigma}$ do in fact produce nearly the same seismogram. Even in case of (nearly, up to rounding errors) exact data, some regularization is needed to recover $\hat{\sigma}$ better. From Fig. 4.5 it becomes clear that using regularization based on total variation is a good idea for small values x (near the surface), where $\tilde{\sigma}$ is heavily oscillating. But it is not clear how such regularization can help in order to overcome the "deviation in mean value" observed for larger values x, where $\tilde{\sigma}$ in fact is not oscillatory. Setting $\alpha = 2 \cdot 10^{-7}$ in (4.49), we obtained a reconstruction σ_α of $\hat{\sigma}$ shown as a red line in Fig. 4.7 ($\hat{\sigma}$ still in black). It seems as if the choice of a small regularization parameter

is sufficient to suppress the heavy oscillations previously observed for $\tilde{\sigma}$ and small values x. The mean value deviation, however, persists for growing values x. I did not find an overall satisfactory regularized reconstruction of $\hat{\sigma}$. One can not choose α according to the discrepancy principle in the absence of noise. ◇

In [Lai80] it is conjectured that the "mean value deviation" observed in the previous example is caused by the fact that the Ricker pulse g defined in (2.30) has a zero mean value.

Example 4.10 The previous example is repeated with one difference in the setup. Namely, instead of the Ricker pulse function g defined in (2.30) (with $a = 1$ and $f_0 = 5$), we used a "modified Ricker pulse" defined by

$$\tilde{g} : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \tilde{g}(t) := \begin{cases} g(t), & \text{if } g(t) \geq 0 \\ 2g(t), & \text{if } g(t) < 0 \end{cases} \tag{4.50}$$

This function is similar in shape to the pulse function used in [BCL77], its mean value is no longer zero. However, for $\alpha = 0$ (no regularization) we still observed a "mean value deviation" as in Example 4.9, albeit less pronounced. We continue to use exact seismogram data Y. Experimenting with different regularization parameters, we found for $\alpha = 5 \cdot 10^{-4}$ a reconstruction σ_α shown in Fig. 4.8 as a red line. Now, the overall reconstruction quality is rather satisfactory, In Fig. 4.9, the corresponding residual $(F(\sigma_\alpha) - Y, \sqrt{\alpha} G_\beta(\sigma_\alpha))$ is shown. The first n components of this vector are $F(\sigma_\alpha) - Y$ and express the data fidelity and the last n components

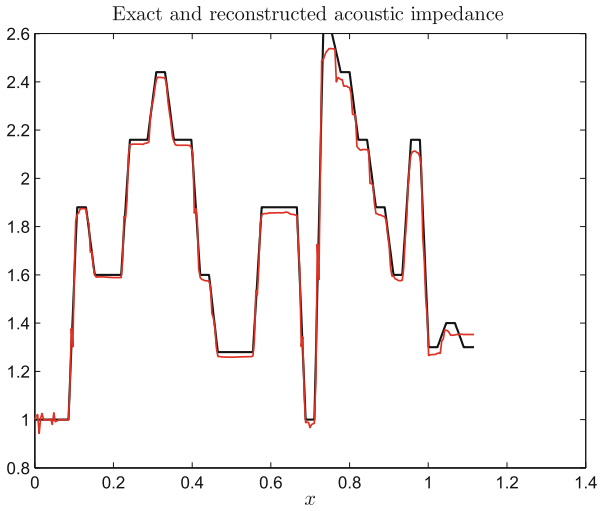

Fig. 4.8 Exact and reproduced impedance, no noise, $\alpha = 5 \cdot 10^{-4}$

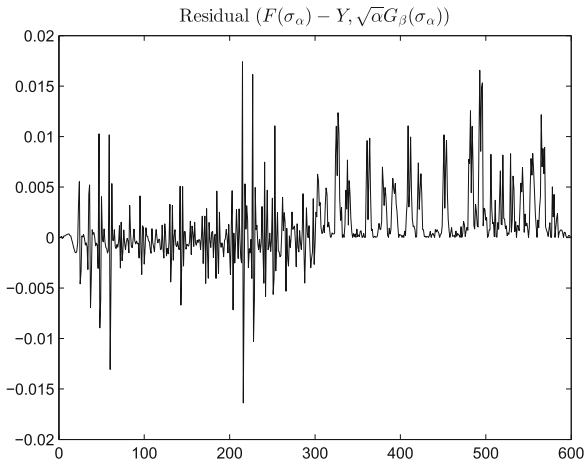

Fig. 4.9 Residual vector, no noise, $\alpha = 5 \cdot 10^{-4}$

express the regularity of σ_α. All components are of about the same size, which can serve as a justification for the choice of the regularization parameter: the latter balances the influence of both optimization criteria. ◇

We still have to test reconstructions from noisy seismograms.

Example 4.11 The setup of Example 4.10 is kept, including the acoustic impedance function $\hat{\sigma}$ shown in Fig. 4.5 (black line) and also including the modified Ricker pulse $\tilde{g}$ defined in (4.50) based on (2.30) with $a = 1$ and $f_0 = 5$. Again let Y be the exact seismogram computed from $\hat{\sigma}$, this time using $\tilde{g}$. A noisy seismogram Y^δ was computed by adding realizations of independent Gaussian random variables with zero mean and standard deviation 10^{-2} to the components of Y, leading to an expected value $\|Y - Y^\delta\|_2 \approx \sqrt{n} \cdot 10^{-2} =: \delta$. With σ_α being defined as the solution of (4.49) for fixed $\alpha \geq 0$, the discrepancy principle suggests to choose the largest possible value α such that $\|F(\sigma_\alpha) - Y^\delta\|_2 \leq \delta$. In the present example ($n = 300$), this meant to choose $\alpha = 7 \cdot 10^{-3}$. In Fig. 4.10, the reconstructed σ_α for this value of α is shown as a red line. The relative shape of the exact impedance is well captured, but again a mean value deviation can be observed. In [BCL77], the authors report on a successful correction of this deviation by incorporation of a known mean value of $\hat{\sigma}$ into the regularized reconstruction. I do not pursue this, not knowing whether such an information on $\hat{\sigma}$ is likely to be available a priori.

Increasing the noise must diminish the quality of the reconstruction. For example, adding realizations of independent Gaussian random variables with zero mean and standard deviation of $5 \cdot 10^{-2}$ leads to an expected error $\|Y - Y^\delta\|_2$ five times as large as before. In Fig. 4.11 the exact and a perturbed seismogram Y and Y^δ are shown for this case. According to the discrepancy principle, $\alpha = 8 \cdot 10^{-2}$ has to be chosen now as a regularization parameter, leading to the result shown in Fig. 4.12. Here, at

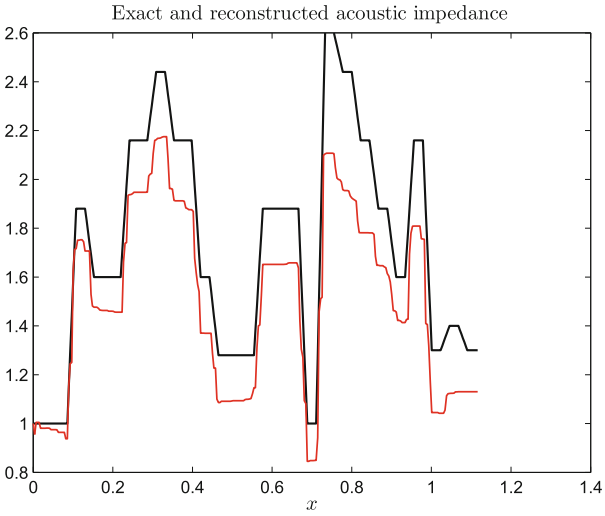

Fig. 4.10 Exact and reproduced impedance, noise standard deviation 10^{-2}, $\alpha = 7 \cdot 10^{-3}$

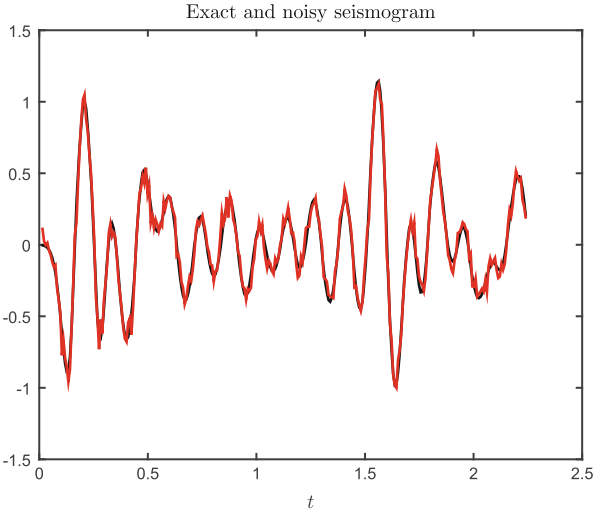

Fig. 4.11 Exact and noisy seismogram, noise standard deviation $5 \cdot 10^{-2}$

least the location of layer boundaries, where the impedance changes considerably, can still be reconstructed. ◇

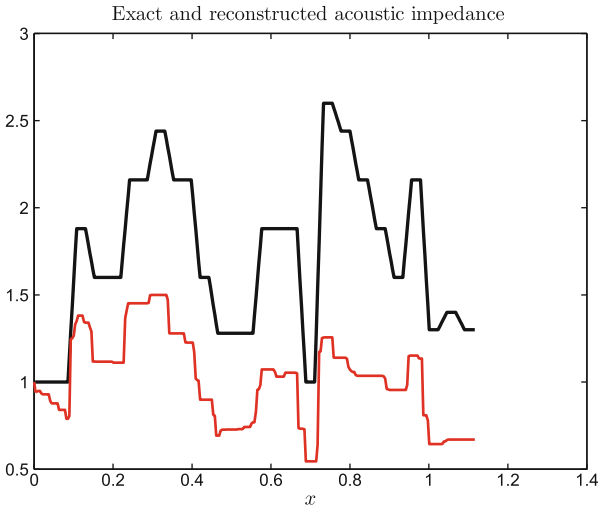

Fig. 4.12 Exact and reconstructed impedance, noise standard deviation $5 \cdot 10^{-2}$, $\alpha = 8 \cdot 10^{-2}$

4.6 Iterative Regularization

Iterative regularization of a nonlinear system of equations $F(x) = y$ means to stop an iterative solver prematurely, i.e. before it converges. This was investigated for linear systems $Ax = b$ in Sects. 3.9 and 3.10, but the same idea can be used in the nonlinear case. For example one can generalize the Landweber iteration to nonlinear equations, see Chapter 11 of [EHN96]. In the following, we rather consider algorithms based on Newton's method. The results presented below are taken from [Han97a] and [Han97b].

Starting point is again the nonlinear system of equations

$$F(x) = y, \quad F : D \subseteq \mathbb{R}^n \to \mathbb{R}^m, \quad x \in D, \quad y \in \mathbb{R}^m, \tag{4.51}$$

where $D \subseteq \mathbb{R}^n$ is *closed* and *convex* and where $F \in C^2(D)$. Below, we will simply take $D = \mathbb{R}^n$, not considering any constraints, which would be possible, but would complicate matters. We assume that a solution $\hat{x}$ of (4.51) exists, but that we only dispose of some approximation $y^\delta \approx y$ with

$$\|y - y^\delta\|_2 \leq \delta, \tag{4.52}$$

for $\delta > 0$ known. With $\hat{x}$ a solution of (4.51) and x_n some approximation of it, one has

$$F(\hat{x}) - F(x_n) = F'(x_n)(\hat{x} - x_n) + R(\hat{x}; x_n), \tag{4.53}$$

where $R(\hat{x}; x_n)$ is the Taylor remainder. Adding y^δ on both sides of (4.53) and using $F(\hat{x}) = y$, one gets

$$F'(x_n)(\hat{x} - x_n) = \underbrace{y^\delta - F(x_n)}_{=: \tilde{y}_n} + y - y^\delta - R(\hat{x}; x_n) =: y_n \tag{4.54}$$

If y_n was known, then an ideal update $z = \hat{x} - x_n$ of x_n (which directly leads to $\hat{x}$) could be computed as a solution of the linear system of equations

$$A_n x = y_n, \quad A_n := F'(x_n). \tag{4.55}$$

However, only the approximation $\tilde{y}_n$ of y_n is known with

$$\|y_n - \tilde{y}_n\|_2 \leq \delta + \|R(\hat{x}; x_n)\|_2. \tag{4.56}$$

One could try to solve $A_n x = \tilde{y}_n$ instead of $A_n x = y_n$. This would give a non-ideal update $\tilde{z}$ and a next, hopefully improved, approximation $x_{n+1} = x_n + \tilde{z}$ of $\hat{x}$. According to the discrepancy principle, the iterative procedure should be stopped as soon as $\|y^\delta - F(x_n)\|_2 \leq \tau\delta$, where $\tau \geq 1$ is a chosen parameter. However, if the solution $\hat{x}$ of the nonlinear problem $F(x) = y$ is badly conditioned, then chances are high that so is the matrix A_n, and if this is the case, then a solution $\tilde{z}$ of $A_n x = \tilde{y}_n$ may be far away from a solution z of $A_n x = y_n$. Therefore, it is advisable to use a regularization method for solving $A_n x = \tilde{y}_n$. Hanke ([Han97b]) proposes to use the powerful CG method applied to the normal equations

$$A_n^T A_n x = A_n^T \tilde{y}_n,$$

which is known as CGNE method. The CGNE method as an iterative equation solver was investigated in Sect. 3.10. In the present case it is used as an *inner* iteration within the *outer* Newton iteration and will produce a sequence $z_0, z_1, z_2, \ldots$ of approximations to the solution z of $A_n x = y_n$. It should be stopped according to the discrepancy principle as soon as

$$\|y^\delta - A_n z_k\|_2 \leq \|y_n - \tilde{y}_n\|_2 \tag{4.57}$$

holds for an index $k \in \mathbb{N}_0$, see Theorem 3.41. However, not even the upper bound (4.56) for the right hand side of (4.57) is known, unless we make assumptions on the Taylor remainder term in (4.53). Following [Han97b], it will be required that for a certain ball $B_r(\hat{x}) \subset D = \mathbb{R}^n$ around the solution $\hat{x}$ of (4.51), there exists a constant $C > 0$ such that

$$\|F(x) - F(\tilde{x}) - F'(\tilde{x})(x - \tilde{x})\|_2 \leq C\|x - \tilde{x}\|_2 \|F(x) - F(\tilde{x})\|_2 \tag{4.58}$$

for all $x, \tilde{x} \in B_r(\hat{x})$. The left hand side of (4.58) is the magnitude of the linearization error $R(x;\tilde{x})$ for F, as in (4.53). Since $F \in C^2(D)$, it is clear that the linearization error can be bounded in the form

$$\|F(x) - F(\tilde{x}) - F'(\tilde{x})(x-\tilde{x})\|_2 \le C\|x-\tilde{x}\|_2^2,$$

but (4.58) additionally requires the linearization error to be controlled by the size of the nonlinear residual $F(x) - F(\tilde{x})$. This can be interpreted as a restrictive assumption on F, which is not allowed to be "too nonlinear" near $\hat{x}$. Using (4.58), the requirement (4.57) can be replaced by

$$\|y^\delta - A_n z_k\|_2 \le \delta + C\|\hat{x} - x_n\|_2 \|y - F(x_n)\|_2,$$

which still is not directly usable, since $\hat{x}$ and y are unknown. But as long as the outer Newton iteration is not terminated by the discrepancy principle, one has $\delta < \|y^\delta - F(x_n)\|_2$ and a sufficiently large fraction of $\|y^\delta - F(x_n)\|_2$ can serve as an upper bound for $\delta + C\|\hat{x} - x_n\|_2\|y - F(x_n)\|_2$, if x_n is close enough to $\hat{x}$ (how close it has to be depends on the constant C). We take the following algorithm from [Han97b], which is based on CGNE from Sect. 3.10.

Inexact Newton-CG method for minimizing $\|F(x) - y^\delta\|_2$

choose parameters $\tau \ge 1$ and $0 < \rho < 1$
choose starting point $x_0 \in \mathbb{R}^n$ and set $n = 0$
while $\|y^\delta - F(x_n)\|_2 > \tau\delta$ % outer (Newton) iteration
  $b = y^\delta - F(x_n), A = F'(x_n)$
  $z_0 = 0, w_0 = b, r_0 = b, k = 0$
  repeat % inner (CGNE) iteration: $A^T A z = A^T b$
    $d_k = A^T w_k$
    $\alpha_k = \|A^T r_k\|_2^2 / \|A d_k\|_2^2$
    $z_{k+1} = z_k + \alpha_k d_k$
    $r_{k+1} = r_k - \alpha_k A d_k$
    $\beta_k = \|A^T r_{k+1}\|_2^2 / \|A^T r_k\|_2^2$
    $w_{k+1} = r_{k+1} + \beta_k w_k$
    $k = k + 1$
  until $\|r_k\|_2 < \rho\|b\|_2$ % end inner iteration

(continued)

$x_{n+1} = x_n + z_k$
$n = n + 1$
end % end outer iteration

The method is called an **inexact Newton method**, because the linear "update equation" $Az = b$ is not solved exactly, but only approximately (in order to obtain a regularized solution). In [Han97b] it is proved that under the conditions

$$\tau\rho^2 > 2 \quad \text{and} \quad x_0 \text{ close enough to } \hat{x} \tag{4.59}$$

the above algorithm works and terminates after a finite number of iterations. It produces a regularized solution x^δ, i.e. x^δ converges to $\hat{x}$ for $\delta \to 0$. In practice, one would like to choose τ close to 1 in order to obtain a good data fit, but (4.59) requires $\tau > 2/\rho^2 > 2$. For example, the combination $\rho = 0.9$ and $\tau = 2.5$ would be in accordance with (4.59). A further general restriction is the requirement (4.58), as already mentioned.

Application to Nonlinear Gravimetry

A discretized variant of the model problem of nonlinear gravimetry was summarized in Sect. 4.2, the corresponding nonlinear function F was defined in (4.20). I have not been successful in proving that (4.58) holds for this function F – but used the inexact Newton-CG method nevertheless. The parameters $\tau = 1.01$ and $\rho = 0.99$ were always kept fixed in the following, against the requirement $\tau\rho^2 > 2$.

Example 4.12 The numerical values from Example 4.4 were retained, especially we used $a = 4$ and $b = 1$ and the same exact function $\hat{u}$ to be reconstructed. The values $c_\alpha = \hat{u}(x_\alpha)$, $\alpha \in G_n$, define values $y_\beta = F_\beta(c)$, $\beta \in B$, as in Example 4.4. To these values y_β we added random numbers from a normal distribution with mean value 0 and standard deviation $\sigma = 10^{-2}$, resulting in perturbed samples y_β^δ. The inexact Newton-CG method was straightforwardly applied to the system of equations $F(c) = y^\delta$. This gave a regularized solution c^δ, which defines a bilinear spline function u^δ shown in Fig. 4.13. Visibly, u^δ is only a poor approximation of $\hat{u}$ (the latter was shown in Fig. 4.1). ◇

Even if the inexact Newton-CG method formally produces a regularized reconstruction u^δ, which converges to $\hat{u}$ for $\delta \to 0$, this does not mean that the reconstruction is good for a finite value of δ. If only a single data set is given, containing a finite error, then the careful choice of a regularization term matters more than convergence of the method. Now for the inexact Newton-CG method regularization

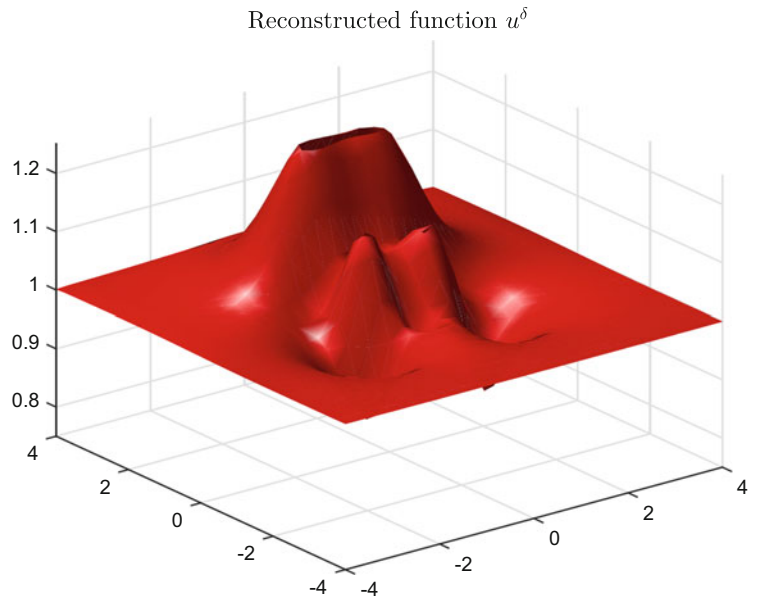

Fig. 4.13 Reconstruction of $\hat{u}$, straightforward application of inexact Newton-CG method

is achieved by prematurely stopping the inner CG iteration before it converges. It was observed in Sect. 3.10 that this implicitly corresponds to using a minimum norm regularization criterion, related to approximating a solution of $Ax = b$ by minimizing $\|Ax - b\|_2^2 + \lambda \|x\|_2^2$. But there is no obvious physical justification of why this regularization term is a good one. One could as well compute x as a minimizer of $\|Ax - b\|_2^2 + \lambda \|Lx\|_2^2$. In case L is invertible, the coordinate transform

$$Lx = z \quad \Longleftrightarrow \quad x = L^{-1}z \tag{4.60}$$

leads to $\|Ax - b\|_2^2 + \lambda \|Lx\|_2^2 = \|Bz - b\|_2^2 + \lambda \|z\|_2^2$ with matrix $B = AL^{-1}$. This means that applying CGNE to the transformed problem and then transforming back the obtained solution makes CGNE mimic a regularization not with respect to $\|x\|_2$, but with respect to $\|Lx\|_2$.

Example 4.13 All the settings of Example 4.12 are retained, but the inexact Newton-CG method will now be applied to a transformed problem, i.e. to problem $\tilde{F}(z) := F(L^{-1}z) = y^\delta$ where $Lc = z$ for an invertible matrix L. The obtained solution z will then be transformed back. It remains to choose L. The following choice is based on the discrete Laplace operator as in Sect. 4.2. Let

$$u_{i,j} := u(x_\alpha) = c_\alpha \text{ for } \alpha = (i,j) \in G_n,$$

where $u : [-a, a]^2 \to \mathbb{R}$. As usual, one approximates by finite differencing:

$$\Delta u(x_\alpha) \approx \frac{1}{h^2}\left[c_{i-1,j} + c_{i+1,j} - 4c_{i,j} + c_{i,j-1} + c_{i,j+1}\right], \quad \text{if } \alpha = (i,j), \tag{4.61}$$

which requires evaluation of u at five grid points, which are located to the west (index $(i-1,j)$), to the east (index $(i+1,j)$), to the south (index $(i,j-1)$), and to the north (index $(i,j+1)$) with respect to the central point with index (i,j). A perfectly regular vector c is supposed to fulfill

$$c_{i-1,j} + c_{i+1,j} - 4c_{i,j} + c_{i,j-1} + c_{i,j+1} = 0 \tag{4.62}$$

at all interior grid points. At boundary grid points, (4.62) makes no sense: if, e.g., $i = -n$ and $-n < j < n$, then $c_{i-1,j}$ is not defined. To deal with boundary grid points, additional boundary conditions have to be introduced. One possibility, much in use for elliptic partial differential equations, is **Robin's boundary condition**. For a function $u : D \to \mathbb{R}$ defined on some region $D \subset \mathbb{R}^s$ it requires that

$$\nabla_n u(x) + \alpha u(x) = 0 \quad \text{for} \quad x \in \partial D, \tag{4.63}$$

where n is the outer normal vector at $x \in \partial D$ and where $\nabla_n u(x) = (\nabla u(x))^T n$ is the exterior normal derivative of u at $x \in \partial D$. If the parameter $\alpha \in \mathbb{R}$ is chosen equal to zero, (4.63) becomes Neumann's boundary condition. If $D = [-a,a]^2$ is a rectangle, and x is located on the left boundary, then n points westwards. At a grid point indexed $\alpha = (i,j)$ with $i = -n$ and $-n < j < n$, (4.63) can formally be discretized by finite differencing, leading to

$$\frac{c_{i-1,j} - c_{i+1,j}}{2h} + \alpha c_{i,j} = 0. \tag{4.64}$$

Taking this equation as a *definition* of $c_{i-1,j}$ and inserting into (4.62) leads to

$$2c_{i+1,j} - (4 + 2\alpha h)c_{i,j} + c_{i,j-1} + c_{i,j+1} = 0. \tag{4.65}$$

For reasons of symmetry, which will become apparent in a moment, this equation is scaled by $\frac{1}{2}$, leading to

$$c_{i+1,j} - (2 + \alpha h)c_{i,j} + \frac{1}{2}c_{i,j-1} + \frac{1}{2}c_{i,j+1} = 0. \tag{4.66}$$

The same approach can be used to get equations at the right, the lower, and the upper boundary. There are still four corner points to consider, e.g. the lower left corner point with index $(i,j) = (-n,-n)$. Here, neither $c_{i-1,j}$ nor $c_{i,j-1}$ are defined in (4.62). But now, instead of (4.64), we get two equations

$$\frac{c_{i-1,j} - c_{i+1,j}}{2h} + \alpha c_{i,j} = 0 \text{ and}$$

$$\frac{c_{i,j-1} - c_{i,j+1}}{2h} + \alpha c_{i,j} = 0,$$

corresponding to two normal vectors, one pointing to the west and the other pointing to the south. These two equations are used to define $c_{i-1,j}$ and $c_{i,j-1}$. Inserting into (4.62) leads to

$$2c_{i+1,j} - (4 + 4\alpha h)c_{i,j} + 2c_{i,j+1} = 0.$$

After scaling by $\frac{1}{4}$, this becomes

$$\frac{1}{2}c_{i+1,j} - (1 + \alpha h)c_{i,j} + \frac{1}{2}c_{i,j+1} = 0. \tag{4.67}$$

The equations (4.62), (4.66), and (4.67) (with obvious modifications at the other boundary and corner points) can consistently be written in matrix form $Lc = 0$, where c is derived from $\{c_\alpha,\ \alpha \in G_n\}$ by rowwise ordering and where L is defined as follows. Set

$$A_n := \begin{pmatrix} -(2+\alpha h) & 1 & & & \\ 1 & -4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & 1 & -(2+\alpha h) \end{pmatrix} \quad \text{and} \quad B_n := \begin{pmatrix} \frac{1}{2} & 0 & & & \\ 0 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & 1 & 0 \\ & & & 0 & \frac{1}{2} \end{pmatrix}$$

both being $(2n+1) \times (2n+1)$-matrices, and define

$$\tilde{L} := \begin{pmatrix} \frac{1}{2}A_n - \alpha h I_{2n+1} & B_n & & & \\ B_n & A_n & B_n & & \\ & \ddots & \ddots & \ddots & \\ & & B_n & A_n & B_n \\ & & & B_n & \frac{1}{2}A_n - \alpha h I_{2n+1} \end{pmatrix} \in \mathbb{R}^{(2n+1)^2,(2n+1)^2},$$

which is a symmetric matrix. Finally, L is defined by adding $\alpha h/2$ to the diagonal of L at the positions of the four corner points. In case $\alpha = 0$, matrix L becomes singular, its kernel being spanned by the vector $e = (1, 1, \ldots, 1)$. In case $\alpha \neq 0$, this matrix can be used to define a coordinate transform as in (4.60). Applying the inexact Newton-CG method to the transformed problem means to implicitly use

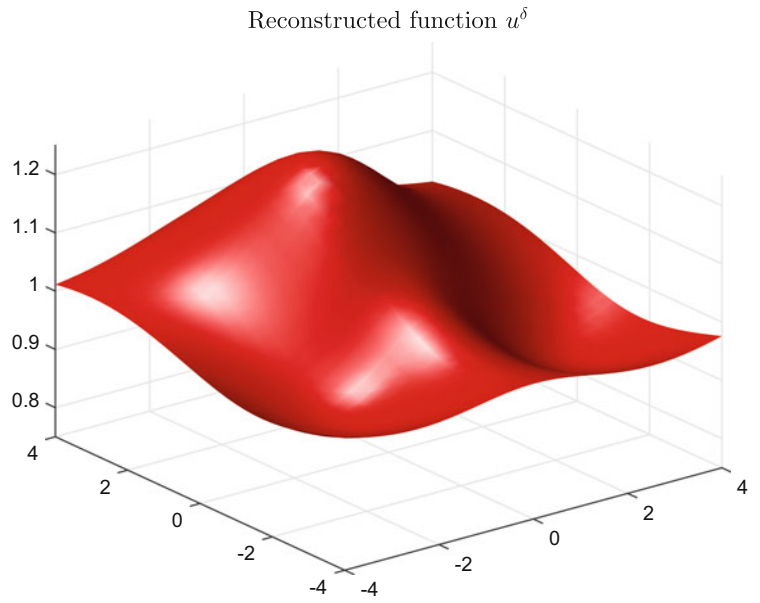

Fig. 4.14 Reconstruction of $\hat{u}$, inexact Newton-CG method applied to transformed problem

$\|Lc\|_2$ as an inverse measure for the regularity of c and the function u it defines. In Fig. 4.14 the result is shown we obtained for $\alpha = 1$ and which means a great improvement over the result from Example 4.12. ◇

From the above example one can learn that the inexact Newton-CG method has potential to achieve good reconstructions. It does not require the full Jacobian $F'(z)$ to be computed, but only relies on the computation of matrix times vector products of the form $F'(z) \cdot v$ and $F'(z)^T \cdot v$ in the inner iteration, so this method can be quite efficient. On the negative side we note the restrictive choices of τ and ρ and the restrictive condition (4.58) on the nonlinearity of F, which were required to prove convergence of the method. However, all these restrictions were ignored or not checked in the present example.

We finally repeat an observation already made for linear problems in Sects. 3.9 and 3.10. Working with iterative methods one can only seemingly avoid the choice of a regularity measure. In fact, an inverse regularity measure, which appears in Tikhonov regularization as a penalty term, also influences iterative methods via coordinate transforms. It has become apparent that the success of an iterative method depends on the proper choice of this transform. How to choose it well, or, equivalently, how to choose well a regularity measure, depends on whether one can successfully capture a priori information about the sought-after solution accordingly.

Appendix A
Results from Linear Algebra

We assume the reader is familiar with the mathematical concepts of real and complex vector spaces (equivalently called linear spaces) and with the definitions of "linear dependence", "linear independence", "dimension", "subspace", and "basis". After choice of some basis, every n-dimensional real or complex vector space can be identified with $\mathbb{R}^n$ or $\mathbb{C}^n$, respectively, where

$$x \in \mathbb{R}^n \text{ or } \mathbb{C}^n \quad \Longleftrightarrow \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{all } x_i \in \mathbb{R} \text{ or } \mathbb{C}, \text{respectively.}$$

We will also write $x = (x_1, \ldots, x_n)^T$. The superindex T means **transposed** and makes a row vector out of a column vector and vice versa. For the subspace of $\mathbb{R}^n$ ($\mathbb{C}^n$) spanned by k vectors $b_1, \ldots, b_k \in \mathbb{R}^n$ (or $\mathbb{C}^n$) we use the notation

$$\langle b_1, \ldots, b_k \rangle := \text{span}\{b_1, \ldots, b_k\} := \{\lambda_1 b_1 + \ldots + \lambda_k b_k;\ \lambda_1, \ldots, \lambda_k \in \mathbb{R}\ (\mathbb{C})\}\,.$$

A **matrix** is defined by its components

$$A \in \mathbb{R}^{m,n} \text{ or } \mathbb{C}^{m,n} \quad \Longleftrightarrow \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \text{all } a_{ij} \in \mathbb{R} \text{ or } \mathbb{C}, \text{resp.},$$

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9

or equally by its columns

$$A \in \mathbb{R}^{m,n} \text{ or } \mathbb{C}^{m,n} \quad \Longleftrightarrow \quad A = \left(\begin{array}{c|c|c|c} a_1 & a_2 & \cdots & a_n \end{array} \right), \quad \text{all } a_j \in \mathbb{R}^m \text{ or } \mathbb{C}^m, \text{resp.}$$

The rules for multiplying matrices should be known as well as the fact that a matrix $A \in \mathbb{R}^{m,n}$ defines a linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$. Conversely, every linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ can be represented in the form $f(x) = Ax$. The same equivalence holds for complex valued matrices $A \in \mathbb{C}^{m,n}$ and linear mappings $A : \mathbb{C}^n \to \mathbb{C}^m$. It is assumed the reader knows what is meant by the inverse of a matrix, by its determinant and by its rank. The columns of a given matrix $A \in \mathbb{R}^{m,n}$ ($A \in \mathbb{C}^{m,n}$) span the linear space

$$\mathscr{R}_A := \{Ax = \sum_{j=1}^{n} x_j a_j;\ x_j \in \mathbb{R}\ (\mathbb{C})\} \subseteq \mathbb{R}^m\ (\mathbb{C}^m)$$

of dimension rank(A). The kernel or null space of A is defined as

$$\mathscr{N}_A := \{x \in \mathbb{R}^n\ (\mathbb{C}^n);\ Ax = 0\} \subseteq \mathbb{R}^n\ (\mathbb{C}^n)$$

and has dimension $n - \text{rank}(A)$.

The columns of the unity matrix $I_n \in \mathbb{C}^{n,n}$ are designated by $e_1, \ldots, e_n$ and are called **canonical unity vectors** (of $\mathbb{R}^n$ and $\mathbb{C}^n$). A matrix $A \in \mathbb{C}^{m,n}$ with components a_{ij} has a **transposed** $A^T \in \mathbb{C}^{n,m}$ with components

$$(A^T)_{ij} := a_{ji}, \quad i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

and an **adjoint** or **hermitian conjugate** $A^* \in \mathbb{C}^{n,m}$ with components

$$(A^*)_{ij} := \overline{a_{ji}}, \quad i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

where $\overline{z}$ is the conjugate complex number of $z \in \mathbb{C}$. For a real number $z \in \mathbb{R}$ we have $\overline{z} = z$ and therefore $A^* = A^T$ in case of a real valued matrix A. We have $(AB)^T = B^T A^T$, $(AB)^* = B^* A^*$, and $(A^{-1})^* = (A^*)^{-1} =: A^{-*}$, whenever the inverse A^{-1} of A exists. In case $A = A^T$ the matrix A is called **symmetric** and in case $A = A^*$ it is called **hermitian** or **self-adjoint**. For $x, y \in \mathbb{C}^n$, we define the **Euclidean scalar product** (Euclidean inner product) by

$$\langle x|y\rangle := \overline{x^* y} = \sum_{i=1}^{n} x_i \overline{y_i}.$$

Here, the row vector $x^* = (\overline{x_1}, \ldots, \overline{x_n})$ is the adjoint of the column vector x and x^*y is a matrix product. The scalar product for real vectors is defined in the same way, but the overline has no effect in this case: $\langle x|y\rangle = \sum_{i=1}^{n} x_i y_i$ for $x, y \in \mathbb{R}^n$. Vectors $x, y \in \mathbb{C}^n$ are called **orthogonal**, if $x^*y = 0$. In this case, we write $x \perp y$. A *set* of vectors $b_1, \ldots, b_k \in \mathbb{C}^n$ is called **orthonormal**, if $b_i^* b_j = 0$ for $i \neq j$ and $b_i^* b_i = 1$ for all i. If additionally $k = n$, then $\{b_1, \ldots, b_n\}$ ist called an **orthonormal basis (ONB)** of $\mathbb{C}^n$. A matrix $V \in \mathbb{C}^{n,n}$ is called **unitary** (in case $A \in \mathbb{R}^{n,n}$ also: **orthogonal**), if its columns are an orthonormal basis of $\mathbb{C}^n$ ($\mathbb{R}^n$). This is equivalent to the identities

$$V^*V = I_n \quad \Longleftrightarrow \quad V^{-1} = V^*.$$

A matrix $A \in \mathbb{R}^{n,n}$ or $A \in \mathbb{C}^{n,n}$ is said to have an **eigenvalue** $\lambda \in \mathbb{C}$ (possibly complex valued even if the matrix is real valued!) and corresponding **eigenvector** $v \in \mathbb{C}^n$, if

$$Av = \lambda v \quad \text{and} \quad v \neq 0.$$

If A is hermitian, all eigenvalues are real valued and there exists an orthonormal basis $\{v_1, \ldots, v_n\} \subset \mathbb{C}^n$ of eigenvectors. Thus

$$Av_i = \lambda_i v_i,\ i = 1, \ldots, n \quad \Longleftrightarrow \quad AV = V\Lambda \quad \Longleftrightarrow \quad V^*AV = \Lambda,$$

where $V = (v_1|\cdots|v_n)$ (colums are eigenvectors) and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. In case A is real valued, V is real valued, too.

A matrix $A \in \mathbb{C}^{n,n}$ is called **positive definite**, if it is hermitian *and* $x^*Ax > 0$ for all $x \in \mathbb{C}^n \setminus \{0\}$. It is called **positive semidefinite**, if it is hermitian *and* $x^*Ax \geq 0$ for all $x \in \mathbb{C}^n$. A matrix $A \in \mathbb{C}^{n,n}$ is positive definite, if and only if it is hermitian and all its eigenvalues are positive and it is positive semidefinite, if and only if it is hermitian and has no negative eigenvalue. A is positive definite, if and only if there exists a nonsingular upper triangular matrix $R \in \mathbb{C}^{n,n}$ such that

$$A = R^*R.$$

This is called the **Cholesky factorization** of A. The matrix R can be chosen real valued if A is real valued.

The Singular Value Decomposition (SVD)

Let $m \geq n$ and $A \in \mathbb{C}^{m,n}$ with $\operatorname{rank}(A) = r$. Then $A^*A \in \mathbb{C}^{n,n}$ is positive semidefinite. Let $\sigma_1^2 \geq \ldots \geq \sigma_r^2 > 0$ and $\sigma_{r+1}^2 = \ldots = \sigma_n^2 = 0$ be its eigenvalues and let $v_1, \ldots, v_n$ be the corresponding orthonormal eigenvectors:

$$A^*Av_k = \sigma_k^2 v_k, \quad k = 1, \ldots, n.$$

Then $u_k := Av_k/\sigma_k \in \mathbb{C}^m$, $k = 1, \ldots, r$, are eigenvectors of AA^*, since $AA^*u_k = AA^*Av_k/\sigma_k = A\sigma_k v_k = \sigma_k^2 u_k$. The vectors u_k are also orthonormal:

$$u_i^* u_k = v_i^* A^* A v_k/(\sigma_i \sigma_k) = v_i^* v_k \sigma_k/\sigma_i = \delta_{i,k}.$$

Here, we make use of the Kronecker symbol defined by $\delta_{i,k} := 0$ for $i \neq k$ and $\delta_{i,i} := 1$. The set $\{u_1, \ldots, u_r\}$ is completed with $m - r$ orthonormal vectors $u_{r+1}, \ldots, u_m \in \mathbb{C}^m$ which span the $(m - r)$-dimensional nullspace $\mathcal{N}_{A^*}$:

$$A^* u_k = 0, \quad k = r + 1, \ldots, m,$$

and which are the remaining eigenvectors of AA^*. For $i \leq r < k$ we get $u_i^* u_k = v_i^* A^* u_k/\sigma_i = v_i^* 0/\sigma_i = 0$, so that $U := (u_1|\cdots|u_m) \in \mathbb{C}^{m,m}$ is a unitary matrix as is $V := (v_1|\cdots|v_n) \in \mathbb{C}^{n,n}$. From the definitions of u_k and v_k we get $Av_k = \sigma_k u_k$ for $k = 1, \ldots, r$ and $Av_k = 0$ for $k = r + 1, \ldots, n$. Together

$$AV = U\Sigma \quad \Longleftrightarrow \quad A = U\Sigma V^* \quad \text{with} \quad \Sigma_{i,j} = \sigma_i \delta_{i,j}. \tag{A.1}$$

One can drop the last $m - n$ columns of matrix U and the last $m - n$ rows of Σ to get

$$A = \hat{U}\hat{\Sigma}V^*, \quad \hat{U} := (u_1|\cdots|u_n) \in \mathbb{C}^{m,n}, \quad \hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n,n} \tag{A.2}$$

instead of (A.1). If $m < n$, a factorization (A.1) of A^* can be derived as above. Afterwards one takes the hermitian conjugate of the result. This shows

Theorem A.1 (Singular value decomposition (SVD)) *Let $A \in \mathbb{C}^{m,n}$ have rank r. Then there exist unitary matrices $U \in \mathbb{C}^{m,m}$ and $V \in \mathbb{C}^{n,n}$ and a matrix $\Sigma \in \mathbb{R}^{m,n}$ with components $\Sigma_{i,j} = \sigma_i \delta_{i,j}$ and*

$$\sigma_1 \geq \ldots \geq \sigma_r > 0, \quad \sigma_{r+1} = \ldots = \sigma_{\min\{m,n\}} = 0$$

such that

$$A = U\Sigma V^*.$$

*This factorization is called **singular value decomposition (SVD)** and the numbers $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$ are called **singular values** of A. For $m \geq n$, a factorization (A.2) exists, which is called **reduced SVD**.*

Using appropriate coordinate tranforms $y = U\eta$ in $\mathbb{C}^m$ and $x = V\xi$ in $\mathbb{C}^n$ any linear map $\mathbb{C}^n \to \mathbb{C}^m$, $x \mapsto y = Ax$, thus splits into r one-dimensional maps $\eta_i = \sigma_i \xi_i$ and $m - r$ trivial maps $\eta_i = 0$. For $A \in \mathbb{R}^{m,n}$, U and V can be chosen real orthonormal matrices. For numerical computation of the SVD, one must never explicitly form any of the matrices A^*A or AA^*, since this would lead to a numerically unstable

algorithm. Instead, specialized algorithms have to be used like the method of Golub and Reinsch, see, e.g., [Dem97], Section 5.4.

A pair of matrices $A, B \in \mathbb{C}^{n,n}$ has **generalized eigenvalue** $\lambda \in \mathbb{C}$ and corresponding **generalized eigenvector** $v \in \mathbb{C}^n$, if

$$Av = \lambda Bv \quad \text{and} \quad v \neq 0.$$

Now let A be positive semidefinite and let B be positive definite. Using the factorization $B = R^*R$ and the transformation $Rv = w$ we can equivalently reformulate the generalized eigevalue problem as an ordinary one:

$$R^{-*}AR^{-1}w = \lambda w, \quad w \neq 0,$$

where $R^{-*}AR^{-1}$ is a positive semidefinite matrix. Thus, there is an orthonormal basis $\{w_1, \ldots, w_n\}$ of eigenvectors corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n \geq 0$ of $R^{-*}AR^{-1}$. Defining the orthogonal matrix $W := (w_1|\cdots|w_n)$ and the nonsingular matrix $V := R^{-1}W$ we get

$$V^*BV = W^*R^{-*}R^*RR^{-1}W = W^*W = I_n$$

and we also get

$$V^*AV = W^*(R^{-*}AR^{-1})W = W^*W\text{diag}(\lambda_1, \ldots, \lambda_n) = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

To summarize: If $A \in \mathbb{C}^{n,n}$ is positive semidefinite and $B \in \mathbb{C}^{n,n}$ is positive definite, then there is a nonsingular matrix $V \in \mathbb{C}^{n,n}$ such that

$$V^*AV = \text{diag}(\lambda_1, \ldots, \lambda_n), \ \lambda_1, \ldots, \lambda_n \geq 0, \quad \text{and} \quad V^*BV = I_n. \tag{A.3}$$

We define the **Euclidean norm** on $\mathbb{R}^n$ ($\mathbb{C}^n$) by

$$\|x\| = \|x\|_2 := \sqrt{|x_1|^2 + \ldots + |x_n|^2} = \sqrt{x^*x}.$$

The **inequality of Cauchy-Schwarz** reads:

$$|x^*y| \leq \|x\|_2\|y\|_2.$$

By **Pythagoras' Theorem** we have

$$\|b_1 + \ldots + b_k\|_2^2 = \|b_1\|_2^2 + \ldots + \|b_k\|_2^2,$$

if $b_1, \ldots, b_k \in \mathbb{C}^n$ are mutually orthogonal. If $V \in \mathbb{C}^{n,n}$ is a unitary matrix, then

$$\|Vx\|_2^2 = x^*V^*Vx = x^*x = \|x\|_2^2 \quad \text{for all} \quad x \in \mathbb{C}^n.$$

We define the **spectral norm** of a matrix $A \in \mathbb{C}^{m,n}$ by

$$\|A\|_2 := \max\left\{\frac{\|Ax\|_2}{\|x\|_2};\ x \in \mathbb{C}^n \setminus \{0\}\right\} = \max\{\|Ax\|_2;\ \|x\|_2 = 1\}$$

(analogous definition in the real valued case). The spectral norm has the following properties (which must hold for every norm):

$$\|A\|_2 = 0 \Longleftrightarrow A = 0, \quad \|\lambda A\|_2 = |\lambda|\|A\|_2, \quad \text{and } \|A + B\|_2 \leq \|A\|_2 + \|B\|_2$$

for $A, B \in \mathbb{C}^{m,n}$ and $\lambda \in \mathbb{C}$. Additionally, it is **consistent** with the Euclidean norm – from which it is induced – and it also is consistent with itself. This means that

$$\|Ax\|_2 \leq \|A\|_2\|x\|_2 \quad \text{and} \quad \|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$$

for all $x \in \mathbb{C}^n$, $A \in \mathbb{C}^{m,n}$ and $B \in \mathbb{C}^{n,k}$. If $V \in \mathbb{C}^{n,n}$ is unitary, then $\|V\|_2 = 1$. If $U \in \mathbb{C}^{m,m}$ also is unitary, then we have:

$$\|A\|_2 = \|UA\|_2 = \|AV\|_2 = \|UAV\|_2$$

for every $A \in \mathbb{C}^{m,n}$. Norms and singular values are closely related. The following theorem is proven, e.g., in Lecture 5 of [TB97].

Theorem A.2 *Let $A \in \mathbb{C}^{m,n}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$. Then*

$$\|A\|_2 = \sigma_1.$$

In case $m = n$, A is invertible if and only if $\sigma_n > 0$. In this case

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}.$$

Let $\mathbb{M}_k$ be the set of matrices from $\mathbb{C}^{m,n}$ having rank lower than k ($\mathbb{M}_1$ contains only the nullmatrix). Then, for $k = 1, \ldots, \min\{m, n\}$

$$\min\{\|A - X\|_2;\ X \in \mathbb{M}_k\} \quad = \quad \sigma_k. \tag{A.4}$$

From Eq. (A.4) we see that $\sigma_{\min\{m,n\}} \leq \varepsilon$ is a warning that within the ε-vicinity of A there are matrices with deficient rank. Being a discontinuous function of the matrix components, the rank of a matrix is almost impossible to compute numerically, at least when less than $\min\{m, n\}$. Singular values, in contrast, are stable (see Theorem A.3 below) and can be computed reliably. Thus, computing the smallest singular values of a matrix answers best the question for its rank.

Theorem A.3 (Sensitivity of singular values) *Let $A, \delta A \in \mathbb{C}^{m,n}$. Let $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$ be the singular values of A and let $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_{\min\{m,n\}} \geq 0$ be the singular values of $A + \delta A$. Then*

$$|\sigma_i - \tilde{\sigma}_i| \leq \|\delta A\|_2, \quad i = 1, \ldots, n.$$

The upper bound is sharp.

Proof See, e.g., [Dem97], p. 198. □

Appendix B
Function Spaces

In all inverse problems treated in this book one is asking for an unknown *function*. For an abstract formulation of inverse problems as equations in vector spaces – as in Sect. 1.2 – we interpret functions as vectors. This does not help much with the practical solution of inverse problems. However, it is an adequate language to formulate inverse problems and to characterize many of the difficulties one encounters with their solution. It can also help to develop intuition, like in Example B.6 below, where approximation by Fourier sums is described as a projection into a function space. Below we will use the notation $\mathbb{K}$ to mean either $\mathbb{R}$ or $\mathbb{C}$, when it is not necessary nor desirable to be more specific.

Let $\emptyset \neq \Omega \subset \mathbb{R}^s$ and let

$$\mathscr{F}(\Omega, \mathbb{K}) := \{f : \Omega \to \mathbb{K}\}$$

be the set of all $\mathbb{K}$-valued functions defined on Ω. We can define the addition ("superposition") $f+g$ of two functions $f, g \in \mathscr{F}(\Omega, \mathbb{K})$ by $(f+g)(t) := f(t)+g(t)$ for all $t \in \Omega$. Note that $f + g$ is a function, whereas $f(t) + g(t)$ is the sum of two numbers $f(t), g(t) \in \mathbb{K}$. Likewise we can define a scalar multiplication $\lambda \cdot f$ for $f \in \mathscr{F}(\Omega, \mathbb{K})$ and $\lambda \in \mathbb{K}$ by $(\lambda \cdot f)(t) := \lambda f(t)$. The zero function

$$0 : \Omega \to \mathbb{K}, \quad t \mapsto 0(t) := 0,$$

will not be distinguished notationally from the number $0 \in \mathbb{K}$. The additive inverse of $f \in \mathscr{F}(\Omega, \mathbb{K})$ is given by $-f$, defined by $(-f)(t) := -f(t)$ for all $t \in \Omega$. We have $f+g = g+f$ for $f, g \in \mathscr{F}(\Omega, \mathbb{K})$, since $f(t)+g(t) = g(t)+f(t) \in \mathbb{K}$ for all $t \in \Omega$. In the same way one verifies that all associative, commutative, and distributive laws hold in $\mathscr{F}(\Omega, \mathbb{K})$. In short, $\mathscr{F}(\Omega, \mathbb{K})$ is a vector space (equivalently called linear space), the vectors being functions, vector addition being function superposition and scalar multiplication being defined as scaling of function values. The sine function $\sin : \mathbb{R} \to \mathbb{R}$ is an element or "point" in the space $\mathscr{F}(\mathbb{R}, \mathbb{R})$ and we write $\sin \in$

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9

$\mathscr{F}(\mathbb{R},\mathbb{R})$ in the same way as we write $(1,1,1)^T \in \mathbb{R}^3$ for a point in the Euclidean space $\mathbb{R}^3$.

Vector spaces become more interesting as soon as one defines norms, scalar products and operators. In the following, we will do so for special subspaces of $\mathscr{F}(\Omega,\mathbb{K})$. A nonempty set $Y \subseteq \mathscr{F}(\Omega,\mathbb{K})$ is a **subspace** of $\mathscr{F}(\Omega,\mathbb{K})$, if $f+g \in Y$ and $\lambda \cdot f \in Y$ for all $f,g \in Y$ and all $\lambda \in \mathbb{K}$. For example, the set of all continuous functions $f : \Omega \to \mathbb{K}$ is a subspace of $\mathscr{F}(\Omega,\mathbb{K})$, since the sum of two continuous functions is a continuous function and since the scalar multiple of a continuous function also remains continuous.

Normed Spaces, Inner Product Spaces, Convergence

Let X be a linear space over the field $\mathbb{K}$ ($\mathbb{K}=\mathbb{R}$ or $\mathbb{K}=\mathbb{C}$). A mapping

$$\| \bullet \| : X \to [0,\infty), \quad x \mapsto \|x\|,$$

is called a **norm** on X, if it has following properties:

(1) $\|x\| = 0 \iff x = 0$,
(2) $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbb{K}$ and $x \in X$, and
(3) $\|x+y\| \le \|x\| + \|y\|$ for all $x,y \in X$ (**triangle inequality**).

If $\| \bullet \|$ is a norm on X, then $(X, \| \bullet \|)$ is called a **normed space**.

If property (1) is replaced by the weaker condition

$$\|x\| = 0 \impliedby x = 0,$$

then $\| \bullet \|$ is called **semi-norm**. In this case, $\|x\| = 0$ can happen even if $x \neq 0$, see Example B.2 below.

A **scalar product** or **inner product** on X is a mapping

$$\langle \bullet | \bullet \rangle : X \times X \to \mathbb{K}, \quad (x,y) \mapsto \langle x|y\rangle,$$

when the following holds for all $x,y,z \in X$ and $\lambda \in \mathbb{K}$:

(1) $\langle x+y|z\rangle = \langle x|z\rangle + \langle y|z\rangle$,
(2) $\langle \lambda x|y\rangle = \lambda\langle x|y\rangle$,
(3) $\langle x|y\rangle = \overline{\langle y|x\rangle}$, and
(4) $\langle x|x\rangle > 0$ for $x \neq 0$.

In (3), $\overline{\langle y|x\rangle}$ means the complex conjugate of the number $\langle y|x\rangle \in \mathbb{K}$. In real spaces ($\mathbb{K}=\mathbb{R}$) this has no effect since $\overline{z} = z$ for $z \in \mathbb{R}$. In this case, the inner product is linear in both arguments. If $\mathbb{K}=\mathbb{C}$, linearity in the second argument does no longer hold, since $\langle x|\lambda y\rangle = \overline{\lambda}\langle x|y\rangle$. When $\langle \bullet|\bullet\rangle$ is a scalar product on X, then $(X, \langle \bullet|\bullet\rangle)$ is called **pre-Hilbert space** or **inner product space**.

In every inner product space $(X, \langle \bullet | \bullet \rangle)$ the scalar product **induces** a norm:

$$\|x\| := \sqrt{\langle x|x\rangle} \quad \text{for all} \quad x \in X,$$

such that the **Cauchy-Schwarz inequality**

$$|\langle x|y\rangle| \leq \|x\|\|y\| \quad \text{for all} \quad x, y \in X$$

holds. A sequence $(x_n)_{n\in\mathbb{N}} \subset X$ in a normed space $(X, \| \bullet \|)$ is said to **converge** to $x \in X$, if $\lim_{n\to\infty} \|x_n - x\| = 0$. Often one uses the shorthand notation $x_n \to x$, but has to be aware that convergence does not only depend on the sequence $(x_n)_{n\in\mathbb{N}}$ itself, but also on the norm used. A sequence $(x_n)_{n\in\mathbb{N}_0} \subseteq X$ is called a **Cauchy sequence**, if for every $\varepsilon > 0$ there is a $N \in \mathbb{N}$ such that

$$\|x_n - x_m\| < \varepsilon \quad \text{for all} \quad n, m \geq N.$$

Every sequence convergent in X is a Cauchy sequence. If the converse is also true, the normed space $(X, \| \bullet \|)$ is called **complete**. A complete normed space is called a **Banach space**. An inner product space, which is complete with respect to the induced norm is called a **Hilbert space**. Every normed space $(X, \| \bullet \|_X)$ can be **completed**, i.e. one can find a Banach space $(Y, \| \bullet \|_Y)$ containing it in such a way that $\|x\|_X = \|x\|_Y$ for all $x \in X$.[1]

Example B.1 (The spaces C^k) For $\emptyset \neq U \subseteq \mathbb{R}^s$, $s \in \mathbb{N}$, we denote by $C(U)$ the linear space of ($\mathbb{K}$-valued) continuous functions $v : U \to \mathbb{K}$. When $U = K$ is a **compact**, i.e. a closed and bounded subset of $\mathbb{R}^s$ and if $v \in C(U)$, then $|v|$ will take its maximum and its minimum on K. A norm can then be defined on $C(K)$ by setting

$$\|v\|_{C(K)} := \max\{|v(x)|;\ x \in K\}. \tag{B.1}$$

This is the so-called **maximum-norm**. The space $C(K)$ is known to be complete, so the limit function of a Cauchy sequence $(v_n)_{n\in\mathbb{N}} \subset C(K)$ is continuous. Convergence with respect to the norm (B.1) is called **uniform convergence**.

Now let $\Omega \subseteq \mathbb{R}^s$ be a nonempty open set. For a function $v : \Omega \to \mathbb{K}$ and a **multi-index** $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{N}_0^s$ we set

$$D^\alpha v(x) := \frac{\partial^{\alpha_1} \cdots \partial^{\alpha_s}}{\partial x_1^{\alpha_1} \cdots \partial x_s^{\alpha_s}} v(x), \quad x \in \Omega,$$

[1] The actual construction of a completion is rather abstract and defines Y using equivalence classes of Cauchy sequences in X. An element $x \in X$ is identified with (the equivalence class of) a stationary Cauchy sequence $(x, x, x, \ldots)$ and it is only in this sense that X is contained in Y.

whenever this partial derivative exists. We define the space of k times continuously differentiable, $\mathbb{K}$-valued functions on Ω as

$$C^k(\Omega) := \{v : \Omega \to \mathbb{K};\ D^\alpha v \in C(\Omega) \text{ for } \alpha \in \mathbb{N}_0^s,\ |\alpha| \le k\}\,. \tag{B.2}$$

Here and in the following we use the notation $|\alpha| := \alpha_1 + \ldots + \alpha_s$ for the multi-index $\alpha \in \mathbb{N}_0^s$. A special case of the above definition is $C^0(\Omega) = C(\Omega)$. The set Ω was chosen to be open, because then it does not contain its boundary and we don't run into difficulties when considering differential quotients $D^\alpha v(x)$.[2] On the other hand, continuous functions on open sets don't necessarily take their maximal values. Therefore we consider the **closure** of a set $\Omega \subset \mathbb{R}^s$, defined by

$$\overline{\Omega} := \{x \in \mathbb{R}^s;\ x = \lim_{n\to\infty} x_n \text{ for a sequence } (x_n)_{n\in\mathbb{N}} \subset \Omega\}.$$

$\overline{\Omega}$ is the set of all limit points of convergent sequences with elements in Ω and thus is a closed set. For any function $u \in C(\Omega)$ we will write $u \in C(\overline{\Omega})$, when u has a continuous extension to $\overline{\Omega}$, which then necessarily is unique and will also be notated by u. If $\Omega \subset \mathbb{R}^s$ is open and bounded, then its closure $\overline{\Omega}$ is compact and we can equip the set

$$C^k(\overline{\Omega}) := \left\{v \in C^k(\Omega);\ D^\alpha v \in C(\overline{\Omega}) \text{ for all } \alpha \in \mathbb{N}_0^s,\ |\alpha| \le k\right\} \tag{B.3}$$

with a norm defined by

$$\|v\|_{C^k(\overline{\Omega})} := \sum_{|\alpha|\le k} \|D^\alpha v\|_{C(\overline{\Omega})}, \quad v \in C^k(\overline{\Omega}). \tag{B.4}$$

It can be shown that $C^k(\overline{\Omega})$ together with the norm defined in (B.4) is a complete linear space.

Let $U \subseteq \mathbb{R}^s$ be any nonempty set. We define the **support** of $f : U \to \mathbb{K}$ to be the set

$$\operatorname{supp}(f) := \overline{\{x \in U;\ f(x) = 0\}}.$$

The support of a function f is always closed. If it is also bounded, then f is said to have **compact support**. For $\Omega \subseteq \mathbb{R}^s$ nonempty and open we introduce the notation

$$C_0^k(\Omega) := \{v \in C^k(\Omega);\ \operatorname{supp}(v) \text{ is compact}\}. \tag{B.5}$$

[2] Derivatives are defined as limits of differential quotients. But boundary points can not be approached from all directions, so some partial derivatives may not be defined there.

If $v \in C^k(\Omega)$ has compact support $S \subset \Omega$, then the support of all derivatives $D^\alpha v$ necessarily also is compact and contained in S. Then, (B.4) can be used as a norm on $C_0^k(\Omega)$ and makes this a complete linear space.

Sometimes we need

$$C^\infty(\Omega) = \bigcap_{k\in\mathbb{N}_0} C^k(\Omega) \quad \text{and} \quad C_0^\infty(\Omega) = \{v \in C^\infty(\Omega);\ \operatorname{supp}(v) \text{ is compact}\},$$

the elements of these sets being called **infinitely often differentiable functions (with compact support)**. We also write

$$C^k(\Omega, \mathbb{K}^m) := \{f = (f_1, \ldots, f_m); f_j \in C^k(\Omega) \text{ for } j = 1, \ldots, m\},$$

for the linear space of vector valued functions on Ω with k times continuously differentiable component functions. In the one-dimensional case, we will write $C^k[a, b]$ instead of $C^k([a, b])$ and $C^k(a, b)$ instead of $C^k((a, b))$. ◇

Example B.2 (The space L_2) Let $\Omega \in \mathbb{R}^s$ be a **domain**, i.e. an open and connected set.[3] If Ω is a "simple" bounded geometric object like a polytope or a ball and if $v \in C(\overline{\Omega})$, then it is clear what is meant by the integral

$$I_\Omega(v) = \int_\Omega v(x)\, dx.$$

But this integral can even be given a meaning as a so-called **Lebesgue integral** for any open set Ω and any nonnegative, "measurable" function $v : \Omega \to \mathbb{R}$. We will not give a definition of what is meant by "measurability". Non measurable functions are so exotic that we do not have to consider them – we simply claim that all functions needed for our purposes are measurable. Also, we do not go into Lebesgue integration theory, since whenever we actually want to carry out an integration for a specific function v and on a specific domain Ω, both will be regular enough to make the Lebesgue integral coincide with the Riemann integral. A function $v : \Omega \to \mathbb{R}$ (which may also take negative values) is called **integrable**, if $I_\Omega(|v|) < \infty$. In this case one sets

$$I_\Omega(v) = \int_\Omega v(x)\, dx = \int_\Omega v^+(x)\, dx - \int_\Omega v^-(x)\, dx$$

where

$$v = v^+ - v^-, \quad v^+(x) = \begin{cases} v(x), & \text{if } v(x) \geq 0 \\ 0, & \text{else} \end{cases}, \quad v^-(x) = \begin{cases} 0, & \text{if } v(x) \geq 0 \\ -v(x), & \text{else} \end{cases}.$$

[3] "Connected" means, that Ω can not be written as the union of two disjoint open sets.

A complex valued function $w = u + iv : \Omega \to \mathbb{C}$ with real and imaginary part $u, v : \Omega \to \mathbb{R}$ will also be called integrable if $I_\Omega(|v|) < \infty$. Its integral is a complex value defined by $I_\Omega(w) = I_\Omega(u) + iI_\Omega(v)$ with i being the imaginary unit. A subset $N \subset \Omega$ will be called a **nullset**, if it has volume 0.[4] Two functions $v_1, v_2 : \Omega \to \mathbb{R}$ which are equal except on a nullset are called **equal almost everywhere (a.e.)** For example, the function $v : \mathbb{R} \to \mathbb{R}$ with $v(x) = 1$ for $x = 0$ and $v(x) = 0$ for $x \neq 0$ is equal a.e. to the null function. Two integrable functions v_1 and v_2 being equal a.e. will produce the same integral value $I_\Omega(v_1) = I_\Omega(v_2)$. Let us now define

$$\|v\|_{L_2(\Omega)} := \left(\int_\Omega |v(x)|^2 \, dx \right)^{1/2} \tag{B.6}$$

(which may become infinite without further restrictions on v) and, provisionally,

$$L_2(\Omega) := \{v : \Omega \to \mathbb{K};\ \|v\|_{L_2(\Omega)} < \infty\}. \tag{B.7}$$

To the latter definition, we add a supplementary agreement: two functions $v_1, v_2 \in L_2(\Omega)$ shall be *identified*, whenever they are equal almost everywhere.[5] Only with this supplementary agreement the implication $(\|v\|_{L_2(\Omega)} = 0) \Rightarrow (v = 0)$ becomes true and $(L_2(\Omega), \| \bullet \|_{L_2(\Omega)})$ becomes a normed space, even a Banach space and even a Hilbert space with associated inner product

$$\langle u|v\rangle_{L_2(\Omega)} := \int_\Omega u(x)\overline{v(x)} \, dx. \tag{B.8}$$

Here, $\overline{v(x)}$ means again the complex conjugate value of $v(x)$. For a real valued function $v : \Omega \to \mathbb{R}$, we have $\overline{v(x)} = v(x)$. The drawback of the above supplementary agreement is, that it becomes meaningless to speak of the value $v(x_0)$, $x_0 \in \Omega$, of a function $v \in L_2(\Omega)$, since v has to be identified with any other function agreeing with it except at x_0. The triangle inequality in $(L_2(\Omega), \langle \bullet | \bullet \rangle_{L_2(\Omega)})$ is called **Minkowski's inequality** and the Cauchy-Schwarz inequality is called

[4] This is a rather careless statement, since one can not reasonably define a volume for *any* set $N \subseteq \mathbb{R}^s$. Defining a volume $|I| := (b_1 - a_1) \cdot \ldots \cdot (b_s - a_s)$ for an s-dimensional interval $I = (a_1, b_1) \times \ldots \times (a_s, b_s)$ with $a_j \leq b_j$ for all j, a precise statement would be: $N \subset \Omega$ is a nullset, if for every $\varepsilon > 0$ there exists a sequence $(I_j)_{j=1,2,\ldots}$ of s-dimensional intervals such that

$$N \subset \bigcup_{j=1}^{\infty} I_j \quad \text{and} \quad \sum_{j=1}^{\infty} |I_j| \leq \varepsilon.$$

[5] In a mathematically clean way, the elements of $L_2(\Omega)$ therefore have to be defined not as functions, but as equivalence classes of functions being equal a.e.

Hölder's inequality. In the one-dimensional case, we will write $L_2(a,b)$ instead of $L_2((a,b))$. ◇

Example B.3 (The Sobolev spaces H^k) Let $\Omega \subset \mathbb{R}^s$ be a domain with boundary $\partial\Omega$ smooth enough to allow an application of the divergence theorem needed below and let $k \in \mathbb{N}_0$. For $v \in C^1(\overline{\Omega})$ integration by parts (the divergence theorem) shows that

$$\int_\Omega \frac{\partial v}{\partial x_j}(x)\varphi(x)\,dx = -\int_\Omega v(x)\frac{\partial \varphi}{\partial x_j}(x)\,dx \quad \text{for all} \quad \varphi \in C_0^\infty(\Omega).$$

More generally, if $v \in C^k(\overline{\Omega})$, then repeated integration by parts shows that

$$\int_\Omega D^\alpha v(x)\varphi(x)\,dx = (-1)^{|\alpha|}\int_\Omega v(x)D^\alpha\varphi(x)\,dx \quad \text{for all} \quad \varphi \in C_0^\infty(\Omega)$$

for any multi-index $\alpha \in \mathbb{N}_0^s$ with $|\alpha| \le k$. For $v \in L_2(\Omega)$ derivatives $D^\alpha v$ can not be defined in the usual way as differential quotients. However, motivated by the above formula, whenever a function $u \in L_2(\Omega)$ exists such that

$$\int_\Omega u(x)\varphi(x)\,dx = (-1)^{|\alpha|}\int_\Omega v(x)D^\alpha\varphi(x)\,dx \quad \text{for all} \quad \varphi \in C_0^\infty(\Omega), \tag{B.9}$$

then u will be called **weak derivative** or **generalized derivative** of v and we will write $u = D^\alpha v \in L_2(\Omega)$. If a weak derivative of v exists, then it is uniquely defined by (B.9). Also, if $u \in C^k(\overline{\Omega})$, then its weak derivative coincides with its derivative in the usual sense.[6] For example, the function $v : (-1,1) \to \mathbb{R}$, $x \mapsto |x|$, which belongs to $C[-1,1] \subset L_2(-1,1)$, is not differentiable. It is weakly differentiable, though, the weak derivative being (identifiable with) the Heaviside function

$$u(x) = \begin{cases} 1, & \text{for } x \in (0,1), \\ 0, & \text{for } x = 0, \\ -1, & \text{for } x \in (-1,0) \end{cases}.$$

The Heaviside function in turn is *not even weakly differentiable*, since its weak derivative would have to coincide a.e. with the zero function. Then the left hand side of (B.9) would be zero for every $\varphi \in C_0^\infty(\Omega)$, but not so the right hand side.

We now define for $k \in \mathbb{N}$ the **Sobolev space**

$$H^k(\Omega) := \{v \in L_2(\Omega);\ D^\alpha v \in L_2(\Omega) \text{ for } |\alpha| \le k\}. \tag{B.10}$$

[6] "Coincides" is to be understood as equality of $L_2(\Omega)$-functions, not as pointwise equality.

It can be equipped with the scalar product

$$\langle u|v\rangle_{H^k(\Omega)} := \sum_{|\alpha|\le k} \langle D^\alpha u|D^\alpha v\rangle_{L_2(\Omega)}, \quad u, v \in H^k(\Omega), \tag{B.11}$$

which induces the **Sobolev norm**

$$\|u\|_{H^k(\Omega)} = \sqrt{\sum_{|\alpha|\le k} \|D^\alpha u\|^2_{L_2(\Omega)}}, \quad u \in H^k(\Omega).$$

It can be shown that $(H^k(\Omega), \langle\bullet|\bullet\rangle_{H^k(\Omega)})$ is a Hilbert space. In the one-dimensional case, we will write $H^k(a, b)$ instead of $H^k((a, b))$. ◇

Convexity, Best Approximation

Let $(X, \langle\bullet|\bullet\rangle)$ be an inner product space over the scalar field $\mathbb{K}$. Two vectors $x, y \in X$ are called **orthogonal**, if $\langle x|y\rangle = 0$. For real inner product spaces we define an angle $\alpha \in [0, \pi]$ between two vectors $x, y \in X \setminus \{0\}$ by requiring

$$\cos\alpha = \frac{\langle x|y\rangle}{\|x\|\,\|y\|},$$

where $\|\bullet\|$ is the norm induced by the scalar product. Orthogonality then means $\alpha = \pi/2$, as it should. A subset $C \subset X$ of a vector space is called **convex**, if the following implication holds

$$x, y \in C \text{ and } \lambda \in \mathbb{R},\ 0 \le \lambda \le 1 \quad \Longrightarrow \quad (1-\lambda)x + \lambda y \in C.$$

Convexity is essential for the existence of a unique solution of an important optimization problem, stated in the following **projection theorem**. A proof can be found in (all) text books on functional analysis.

Theorem B.4 (Projection theorem) *Let $(X, \langle\bullet|\bullet\rangle)$ be a Hilbert space over the scalar field $\mathbb{K}$ with induced norm $\|\bullet\|$ and let $C \subset X$ be a nonempty, closed, and convex subset. Then for every $x \in X$ there exists a unique element $y \in C$ such that*

$$\|x - y\| \le \|x - z\| \quad \textit{for all} \quad z \in C.$$

The vector y is uniquely characterized by

$$Re\langle x - y|z - y\rangle \le 0 \quad \textit{for all} \quad z \in C. \tag{B.12}$$

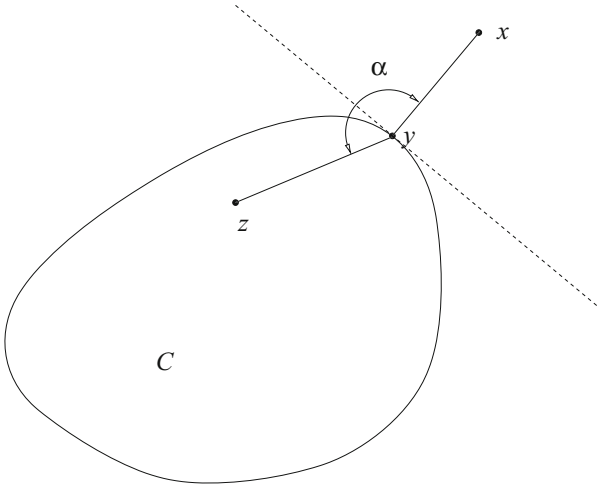

Fig. B.1 Geometric interpretation of the projection theorem

In the above theorem, $\mathrm{Re}(z)$ is the real part of a complex number $z \in \mathbb{C}$. In case $z \in \mathbb{R}$ one has $\mathrm{Re}(z) = z$. For real Hilbert spaces, (B.12) means the angle between $x - y$ and all vectors $z - y$ is obtuse or equal to $\pi/2$ which immediately leads to a geometric interpretation of the projection theorem, as illustrated in Fig. B.1. A finite-dimensional subspace

$$X_n = \mathrm{span}\{\hat{x}_1, \ldots, \hat{x}_n\} \subset X, \quad \hat{x}_1, \ldots, \hat{x}_n \text{ linearly independent},$$

is a nonempty, closed, and convex set. In this case, the projection theorem reads:

Theorem B.5 (Projection theorem, finite dimensional subspaces) *Let $(X, \langle \bullet | \bullet \rangle)$ be a Hilbert space with induced norm $\| \bullet \|$ and let $X_n \subset X$ be an n-dimensional subspace with basis $\{\hat{x}_1, \ldots, \hat{x}_n\}$. Then for every $x \in X$ there is a unique $x_n \in X_n$ such that*

$$\|x - x_n\| \leq \|x - v\| \quad \textit{for all} \quad v \in X_n.$$

The vector x_n is uniquely characterized by the equations

$$\langle x - x_n | \hat{x}_i \rangle = 0 \quad \textit{for} \quad i = 1, \ldots, n. \tag{B.13}$$

By Eq. (B.13), the residual $x - x_n$ is orthogonal to all basis vectors of X_n and thus to all vectors in X_n. Thus we get x_n by an orthogonal projection of x onto X_n – exactly the same will be done when solving the least squares problem, see Fig. 3.1. Actually, completeness of X is not needed in the projection theorem, it suffices that

C is a complete set. This is always the case in the situation of Theorem B.5 with $C = X_n$. The following example makes use of this generalization.

Example B.6 Let $\mathbb{K} = \mathbb{C}$ and let $X = C[0, 1]$ be the space of continuous, complex valued functions on $[0, 1]$. We define a scalar product:

$$\langle f|g\rangle := \int_0^1 f(t)\overline{g(t)}\,dt, \quad f, g \in X,$$

which makes $(X, \langle\bullet|\bullet\rangle)$ a pre-Hilbert space. The functions

$$e_k : [0, 1] \to \mathbb{C}, \quad t \mapsto \mathrm{e}^{ikt}, \quad k \in \mathbb{Z},$$

(with i the imaginary unit) are pairwise orthogonal and even **orthonormal**, since $\|e_k\| = \sqrt{\langle e_k|e_k\rangle} = 1$. By Theorem B.5, for every $f \in X$ there is a unique

$$f_n \in \mathbb{T}_n := \left\{ p = \sum_{k=-n}^{n} c_k e_k;\ c_k \in \mathbb{C} \right\}$$

satisfying $\|f - f_n\| \le \|f - p\|$ for all $p \in \mathbb{T}_n$. To compute f_n, we make the ansatz

$$f_n(t) = \sum_{k=-n}^{n} c_k(f)\mathrm{e}^{ikt}$$

with unknown coefficients $c_k(f)$ depending on f. From (B.13) and by orthonormality, we get:

$$\langle f_n|e_k\rangle = \sum_{j=-n}^{n} c_j(f)\langle e_j|e_k\rangle = c_k(f) = \langle f|e_k\rangle = \int_0^1 f(t)\mathrm{e}^{-ikt}\,dt.$$

The coefficients $c_k(f)$ are called **Fourier coefficients** of f and f_n is called n-th Fourier polynomial of f. ◇

Operators

Let X and Y be linear spaces over $\mathbb{K}$. A mapping $T : D \subseteq X \to Y$ is called **operator**. An operator $T : D \subseteq X \to Y$ is called **linear**, if D is a linear subspace of X and if

$$T(x + y) = T(x) + T(y) \quad \text{and} \quad T(\lambda x) = \lambda T(x)$$

hold for all $x, y \in D$ and for all $\lambda \in \mathbb{K}$. If T is linear, one usually writes Tx instead of $T(x)$.

Example B.7 The mapping

$$I : C[a,b] \to C^1[a,b], \quad x \mapsto y, \quad y(s) = \int_a^s x(t)\,dt, \quad a \le s \le b,$$

defining an antiderivative for every $x \in C[a,b]$, is a linear operator. Differentiation, as defined by

$$D : C^1[a,b] \to C[a,b], \quad x \mapsto y, \quad y(s) = x'(s), \quad a \le s \le b,$$

is another linear operator. ◇

Example B.8 The solution of the inverse problem from Example 1.1 is formally given by the "solution operator"

$$L : \{w \in C^1[t_0,t_1];\ w(t) > 0, t_0 \le t \le t_1\} \to C[t_0,t_1], \quad w \mapsto L(w) := w'/w,$$

which is *not* linear. ◇

Let $(X, \|\bullet\|_X)$ und $(Y, \|\bullet\|_Y)$ be normed spaces over $\mathbb{K}$. An operator $T : D \subseteq X \to Y$ is called **continuous** in $x_0 \in D$, if the following implication holds for any sequence $(x_n)_{n\in\mathbb{N}} \subseteq D$:

$$\lim_{n\to\infty} \|x_n - x_0\|_X = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} \|T(x_n) - T(x_0)\|_Y = 0. \tag{B.14}$$

T is called **continuous on** D, if it is continuous in every $x_0 \in D$.

A linear operator $T : X \to Y$ is called **bounded**, if there is a constant C such that $\|Tx\|_Y \le C\|x\|_X$ holds for all $x \in X$. For every linear bounded operator $T : X \to Y$ we can define its **operator norm**:

$$\|T\| := \sup_{x\in X\setminus\{0\}} \frac{\|Tx\|_Y}{\|x\|_X} < \infty.$$

$\|T\|$ depends on both, $\|\bullet\|_X$ and $\|\bullet\|_Y$, but this is not made explicit notationally. For a bounded operator we evidently have $\|Tx\|_Y \le \|T\| \cdot \|x\|_X$ for all $x \ne 0$. Thus, every linear bounded operator is continuous (everywhere). We even have

$$T \text{ is continuous} \quad \Longleftrightarrow \quad T \text{ is bounded}$$

when T is linear. A proof can be found in any textbook on functional analysis.

Example B.9 Consider $(X = C[a,b], \| \bullet \|_{C[a,b]})$ and $(Y = C^1[a,b], \| \bullet \|_{C[a,b]})$. The integral operator $I : X \to Y$ from Example B.7 is bounded:

$$\|Ix\|_{C[a,b]} = \max_{a \le s \le b} \left\{ \left| \int_a^s x(t)\, dt \right| \right\} \le (b-a) \max_{a \le s \le b} \{|x(t)|\} = (b-a)\|x\|_{C[a,b]},$$

and thus $\|I\|_{C[a,b]} \le (b-a)$ (one could even show $\|I\|_{C[a,b]} = (b-a)$). Consequently, I is continuous. ◇

Continuity of an operator essentially depends on the chosen norms !

Proposition B.10 (Continuity and discontinuity of differentiation) *With respect to the norms* $\| \bullet \|_X = \| \bullet \|_{C[a,b]}$ *on* $X = C^1[a,b]$ *and* $\| \bullet \|_Y = \| \bullet \|_{C[a,b]}$ *on* $Y = C[a,b]$*, differentiation*

$$D : C^1[a,b] \to C[a,b], \quad x \mapsto Dx := x',$$

is a discontinuous operator. With respect to $\| \bullet \|_X = \| \bullet \|_{C^1[a,b]}$ *on* X *and* $\| \bullet \|_Y$ *as above, the same operator is continuous.*

Proof Take the sequence $(x_n)_{n \in \mathbb{N}} \subset C^1[a,b]$ of functions

$$x_n : [a,b] \to \mathbb{R}, \quad t \mapsto x_n(t) = \frac{1}{\sqrt{n}} \sin(nt)$$

with derivatives $Dx_n(t) = (x_n)'(t) = \sqrt{n}\cos(nt)$. This sequence converges to the zero function $x = 0$ with respect to the norm $\| \bullet \|_{C[a,b]}$, since $\|x_n - 0\|_{C[a,b]} \le 1/\sqrt{n} \to 0$. However, $\|Dx_n - Dx\|_{C[a,b]} = \sqrt{n} \to \infty$ for $n \to \infty$, showing that D is not continuous with respect to this norm on X. On the other hand, for any sequence $(x_n)_{n \in \mathbb{N}} \subset C^1[a,b]$ and $x_0 \in C^1[a,b]$:

$$\begin{aligned} \|x_n - x_0\|_{C^1[a,b]} &:= \|x_n - x_0\|_{C[a,b]} + \|Dx_n - Dx_0\|_{C[a,b]} \overset{n\to\infty}{\longrightarrow} 0 \\ &\Longrightarrow \|Dx_n - Dx_0\|_{C[a,b]} \overset{n\to\infty}{\longrightarrow} 0, \end{aligned}$$

which proves continuity with respect to the other norm on X. □

A (linear) operator $T : X \to \mathbb{R}$ is called a (linear) **functional**. A norm on $\mathbb{R}$ is given by the absolute value $| \bullet |$. $\mathbb{R}$ will always be equipped with this norm.

Appendix C
The Fourier Transform

We define the space of s-variate Lebesgue integrable functions

$$L_1(\mathbb{R}^s) := \{v : \mathbb{R}^s \to \mathbb{C};\ \int_{\mathbb{R}^s} |v(x)|\, dx < \infty\}, \tag{C.1}$$

adding two supplementary agreements, just as in Appendix B. First, it is always assumed that functions v are "measurable", non-measurable functions being so exotic that we do not have to consider them. Second, two functions $v_1, v_2 : \mathbb{R}^s \to \mathbb{C}$ being equal almost everywhere and thus having the same integral value, are to be identified. The set $L_1(\mathbb{R}^s)$ does of course also include real valued functions. For every function $f \in L_1(\mathbb{R}^s)$ we define its **Fourier transform** as the function $\mathscr{F}f$ given by

$$(\mathscr{F}f)(y) := \hat{f}(y) := \int_{\mathbb{R}^s} f(x) \mathrm{e}^{-2\pi i x\cdot y}\, dx, \quad y \in \mathbb{R}^s. \tag{C.2}$$

Here, $x \cdot y = \sum_{j=1}^{s} x_j y_j$ and i is the imaginary unit. The Fourier transform in general is a complex valued function, even if $f \in L_1(\mathbb{R}^s)$ is real valued. The **inverse Fourier transform** is defined by

$$(\overline{\mathscr{F}}f)(x) := \int_{\mathbb{R}^s} f(y) \mathrm{e}^{+2\pi i x\cdot y}\, dy, \quad x \in \mathbb{R}^s. \tag{C.3}$$

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9

In case $f \in L_1(\mathbb{R}^s)$ *and* $\hat{f} \in L_1(\mathbb{R}^s)$, the function f is related to its Fourier transform by the **Fourier inversion formula** $f \overset{\text{a.e.}}{=} \overline{\mathscr{F}}\mathscr{F}f$, i.e.

$$f(x) \overset{\text{a.e.}}{=} \int_{\mathbb{R}^s} \hat{f}(y) \mathrm{e}^{2\pi i x \cdot y}\, dy, \quad x \in \mathbb{R}^s. \tag{C.4}$$

If $f \in L_1(\mathbb{R}^s) \cap L_2(\mathbb{R}^s)$ then it can be shown that

$$\hat{f} \in L_2(\mathbb{R}^s) \quad \text{and} \quad \|f\|_{L_2(\mathbb{R}^s)} = \|\hat{f}\|_{L_2(\mathbb{R}^s)} \tag{C.5}$$

($L_2(\mathbb{R}^s)$ and $\| \bullet \|_{L_2(\mathbb{R}^s)}$ are introduced in Appendix B). This can be used to define the Fourier transform on the space $L_2(\mathbb{R}^s)$: for every function $f \in L_2(\mathbb{R}^s)$ there exists a sequence of functions $f_n \in C_0^\infty(\mathbb{R}^s) \subset L_1(\mathbb{R}^s) \cap L_2(\mathbb{R}^s)$ such that $\|f_n - f\|_{L_2(\mathbb{R}^s)} \to 0$ (which is not trivial to see, but well known). By (C.5), the sequence $(\hat{f}_n)_{n\in\mathbb{N}}$ of Fourier transforms is a Cauchy sequence in the complete space $L_2(\mathbb{R}^s)$ and thus converges to a function $g \in L_2(\mathbb{R}^s)$. Reusing the notation from (C.2), one sets $\hat{f} := g$ and calls $\hat{f}$ the Fourier transform of f. By construction, $\hat{f} \in L_2(\mathbb{R}^s)$. Using $f_n = \overline{\mathscr{F}}\hat{f}_n$, one sees that the inversion formula (C.4) holds for f and $\hat{f}$. We use the notation

$$f(x) \circ\!\!-\!\!\bullet\, \hat{f}(y)$$

when $f \in L_2(\mathbb{R}^s)$. In this case, $\hat{f} \in L_2(\mathbb{R}^s)$ is guaranteed and (C.4) holds.

The Fourier transform is a linear, continuous operator on $L_2(\mathbb{R}^s)$ and we have

$$\|\hat{f}\|_{L_2(\mathbb{R}^s)} = \|f\|_{L_2(\mathbb{R}^s)} \quad \text{for} \quad f(x) \circ\!\!-\!\!\bullet\, \hat{f}(y), \tag{C.6}$$

which is known as **Plancherel's identity**. The **convolution** of $f, g \in L_2(\mathbb{R}^s)$ is the function $f * g \in L_2(\mathbb{R}^s)$ defined by

$$(f * g)(x) := \int_{\mathbb{R}^s} f(x-y) g(y)\, dy \tag{C.7}$$

The following relation between the convolution of two $L_2(\mathbb{R}^s)$-functions and their Fourier transform is the basis of efficient methods to solve Fredholm equations of the convolutional type:

$$(f * g)(x) \circ\!\!-\!\!\bullet\, \hat{f}(y)\hat{g}(y) \quad \text{for} \quad f(x) \circ\!\!-\!\!\bullet\, \hat{f}(y),\ g(x) \circ\!\!-\!\!\bullet\, \hat{g}(y). \tag{C.8}$$

The following example shows how convolution can be used to model multipath propagation in signal transmission.

Example C.1 If an analog signal (i.e. a function of time) $u \in C_0(\mathbb{R})$ is transmitted, e.g. in mobile communication, multipath propagation leads to interferences.

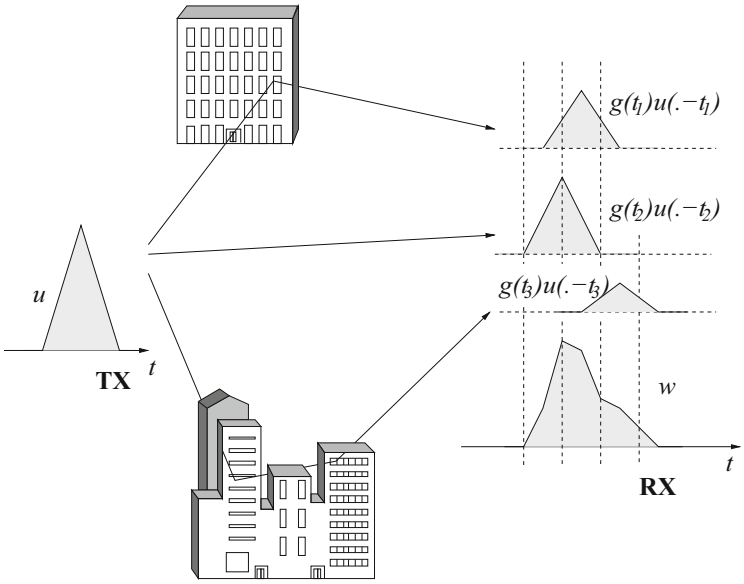

Fig. C.1 Interferences for multipath propagation

Figure C.1 shows an example. The mathematical model for multipath propagation is the following:

$$w(s) = \int_0^\ell g(t)u(s-t)\,dt. \tag{C.9}$$

Here,

- w is the received signal.
- $u(.-t)$ is the transmitted signal, delayed by t units (seconds). The delay corresponds to a propagation time from transmitter (TX) to receiver (RX).
- $g(t)$ is a damping factor corresponding to a loss of signal power. The function $g : [0, \ell] \to \mathbb{R}$ models the communication channel.
- ℓ is the channel length. Signals delayed by more than ℓ seconds are so weak, that they are neglected.

Setting $g(t) := 0$ for $t \notin [0, \ell]$, function g is made a member of $L_2(\mathbb{R})$, as is u. Then, (C.9) can be replaced by

$$w(s) = \int_{-\infty}^{\infty} g(t)u(s-t)\,dt,$$

which is a convolution of g and u. The task of an **equalizer** is to find out u, when w and g are known. This can be done by taking the Fourier transform on both sides of the convolution equation, which, according to (C.8), leads to

$$\hat{w}(y) = \hat{g}(y)\hat{u}(y), \quad y \in \mathbb{R}.$$

If $\hat{g}(y) \neq 0$ for all $y \in \mathbb{R}$, one can get u by applying the inverse Fourier transform to $\hat{u} = \hat{w}/\hat{g}$. In practice, some regularization has to be added to make this solution method work, see Sect. 3.8. ◇

One-Dimensional Discrete Fourier Transform

Practical computation of Fourier transforms can only be done approximately by means of discretization. We consider the Fourier transform of a function $u \in C_0(\mathbb{R})$ with $\operatorname{supp}(u) \subset [-a, a]$ for some $a > 0$, assuming moreover that for some even $N \in \mathbb{N}$, equidistant sample values

$$u_j := u(t_j), \quad t_j := jh, \quad h := \frac{2a}{N}, \quad j \in W := \left\{-\frac{N}{2}, \ldots, \frac{N}{2} - 1\right\} \tag{C.10}$$

of u are given. Using the linear B-Spline ("hat function")

$$B_2(t) := \begin{cases} t+1, & -1 \le t \le 0 \\ 1-t, & 0 \le t \le 1 \\ 0, & \text{else} \end{cases}, \tag{C.11}$$

(compare to (2.3), where the different notation $N_{j,2}$ was used) an approximation of u is given by

$$u_N(t) := \sum_{j=-N/2}^{N/2-1} u_j B_2(t/h - j). \tag{C.12}$$

This polygonal line, which interpolates u at the sample points, can be Fourier transformed *exactly*:

$$\widehat{u_N}(y) = \underbrace{\left(\frac{\sin(\pi h y)}{\pi h y}\right)^2}_{=:\,\sigma(y)} \cdot \underbrace{\left(h \sum_{j=-N/2}^{N/2-1} u_j \mathrm{e}^{-2\pi i j h y}\right)}_{=:\,U(y)}. \tag{C.13}$$

The so called **attenuation factors** $\sigma(y)$ are data independent and determine the decay rate of $\widehat{u_N}(y)$ to zero for $|y| \to \infty$, whilst $U(y)$ is periodic with period $N/2a$. If the exact continuity class of u is known, by the appropriate choice of an interpolation scheme $u \mapsto u_I$ (not necessarily piecewise linear interpolation) one can achieve that $\hat{u}$ and $\widehat{u_I}$ have the same rate of decay at infinity, see [Gau72]. If one only knows M *non*-eqidistant samples

$$u(t_j), \quad -a \le t_0 < \ldots < t_{M-1} \le a,$$

it is still possible to approximate $\hat{u}$ by taking the Fourier transform of some spline-approximant, but one can no longer split off attenuation factors like in (C.13). Alternatively, one can directly approximate the integral defining the Fourier transform, e.g. using the trapezoidal rule. This leads to

$$\hat{u}(y) \approx \sum_{j=0}^{M-1} u_j \mathrm{e}^{-2\pi i y t_j}, \tag{C.14}$$

with

$$u_j = \begin{cases} \frac{1}{2}u(t_0)(t_1 - t_0), & j = 0 \\ \frac{1}{2}u(t_j)(t_{j+1} - t_{j-1}), & j = 1, \ldots, M-2 \\ \frac{1}{2}u(t_{M-1})(t_{M-1} - t_{M-2}), & j = M-1 \end{cases}.$$

An efficient evaluation of formula (C.13) is best possible for the special choice $y = k/(2a)$, $k \in \mathbb{Z}$, in which case both, samples of the function u *and* of its Fourier transform $\widehat{u_N}$ are equidistant. We get

$$\widehat{u_N}\left(\frac{k}{2a}\right) = 2a \cdot \underbrace{\left(\frac{\sin(\pi k/N)}{\pi k/N}\right)^2}_{=:\,\sigma_k} \cdot \underbrace{\left(\frac{1}{N}\sum_{j=-N/2}^{N/2-1} u_j \mathrm{e}^{-2\pi i j k/N}\right)}_{=:\,U_k}, \quad k \in \mathbb{Z}. \tag{C.15}$$

Note that because of their periodicity, only U_k, $k = -N/2, \ldots, N/2-1$, need to be computed to get all values $\widehat{u_N}(k/2a)$, $k \in \mathbb{Z}$. By direct calculation one can show that

$$\begin{aligned} U_k &= \frac{1}{N}\sum_{j=-N/2}^{N/2-1} u_j \mathrm{e}^{-2\pi i j k/N}, \quad k = -\frac{N}{2}, \ldots, \frac{N}{2} - 1, \\ &\Longleftrightarrow \quad u_j = \sum_{k=-N/2}^{N/2-1} U_k \mathrm{e}^{2\pi i j k/N}, \quad j = -\frac{N}{2}, \ldots, \frac{N}{2} - 1. \end{aligned} \tag{C.16}$$

Computing the above values U_k from u_j is called **discrete Fourier transform (DFT)** and computing the values u_j from U_k is called **inverse discrete Fourier transform (IDFT)**. The famous **FFT algorithm** – of which many variants do exist – is an efficient implementation of DFT und IDFT. In case N is a power of 2, it requires only $\mathcal{O}(N\log(N))$ instead of N^2 arithmetical operations. For a description of this algorithm see [PTVF92], p. 504 ff.

Discrete Fourier Transform for Non-equidistant Samples

We come back to the evaluation of (C.13) at arbitrary frequencies and to the evaluation of (C.14), cases in which the FFT algorithm can not be directly used. One way to make it come into play nevertheless was proposed in [Fou99] and is based on the following

Lemma C.2 *Let $p > 1$ be such that $pN \in \mathbb{N}$ and let $1 \le \alpha < 2 - 1/p$. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an even, piecewise continuously differentiable function being continuously differentiable on $[-N/2, N/2]$ and having the properties*

$$supp(\Phi) \subset [-\alpha pN/2, \alpha pN/2] \quad and \quad \Phi(x) > 0 \text{ for } |x| \le N/2.$$

Then

$$\mathrm{e}^{-2\pi ij\xi/N} = \frac{1}{pN}\frac{1}{\Phi(j)}\sum_{\ell\in\mathbb{Z}} \hat{\Phi}\left(\frac{\xi - \ell/p}{N}\right)\mathrm{e}^{-2\pi ij\ell/(pN)} \tag{C.17}$$

for all $j \in \mathbb{Z}$ with $|j| \le N/2$.

Proof Consider

$$g : \mathbb{R} \to \mathbb{C}, \quad x \mapsto g(x) := \sum_{\ell\in\mathbb{Z}} \Phi(x + pN\ell)\mathrm{e}^{-2\pi i(x+pN\ell)\xi/N}.$$

By the assumptions on Φ and α, we have

$$g(j) = \Phi(j)\mathrm{e}^{-2\pi ij\xi/N} \quad \text{for} \quad |j| \le N/2.$$

Since g is periodic with period pN, it can be expressed as a Fourier series

$$g(x) = \sum_{\ell\in\mathbb{Z}} c_\ell \mathrm{e}^{-2\pi ix\ell/(pN)}$$

(equality valid for $|x| \le N/2$, convergence of the series everywhere) with Fourier coefficients

$$\begin{aligned} c_\ell &= \frac{1}{pN} \int_{-pN/2}^{pN/2} g(s) \mathrm{e}^{2\pi i \ell s/(pN)}\, ds \\ &= \frac{1}{pN} \int_{-pN/2}^{pN/2} \left(\sum_{k\in\mathbb{Z}} \Phi(s+pNk) \mathrm{e}^{-2\pi i (s+pNk)\xi/N} \right) \mathrm{e}^{2\pi i \ell s/(pN)}\, ds \\ &= \frac{1}{pN} \int_{\mathbb{R}} \Phi(s) \mathrm{e}^{-2\pi i s\xi/N} \mathrm{e}^{2\pi i \ell s/(pN)}\, ds \\ &= \frac{1}{pN} \int_{\mathbb{R}} \Phi(s) \mathrm{e}^{-2\pi i s(\xi-\ell/p)/N}\, ds = \frac{1}{pN} \hat{\Phi}\left(\frac{\xi - \ell/p}{N} \right), \quad \ell \in \mathbb{Z}, \end{aligned}$$

from which the result follows. □

Lemma C.2 can be used in (C.13) to compute approximations of the Fourier transform of an equidistantly sampled function at arbitrary frequencies, i.e. to compute

$$\hat{u}_k := U\left(\frac{\xi_k}{2a}\right) = \frac{2a}{N} \sum_{j=-N/2}^{N/2-1} u_j \mathrm{e}^{-2\pi i j \xi_k/N}, \quad k = 0, \ldots, M-1. \tag{C.18}$$

Note that $N = M$ is *not* required. Making use of (C.17) and rearranging leads to

$$\hat{u}_k = \frac{2a}{N} \sum_{\ell\in\mathbb{Z}} \hat{\Phi}\left(\frac{\xi_k - \ell/p}{N} \right) \cdot \frac{1}{pN} \sum_{j=-N/2}^{N/2} \frac{u_j}{\Phi(j)} \mathrm{e}^{-2\pi i j \ell/(pN)}. \tag{C.19}$$

This double sum can be computed approximately in three steps:

(1) Define values

$$\tilde{u}_j := \begin{cases} u_j/\Phi(j), & |j| \le N/2 \\ 0, & N/2 < |j| \le pN/2 \end{cases}.$$

(2) Compute values

$$\tilde{U}_\ell = \frac{1}{pN} \sum_{j=-pN/2}^{pN/2-1} \tilde{u}_j \mathrm{e}^{-2\pi i j \ell/(pN)}, \quad \ell \in \mathbb{Z}.$$

Since $(\tilde{U}_\ell)_{\ell\in\mathbb{Z}}$ is periodic, only pN values are to be computed. This can be done efficiently using an FFT of length pN.

(3) For $k = 0, \ldots, M-1$, approximately compute $\hat{u}_k$ as

$$\hat{u}_k \approx \frac{2a}{N} \sum_{\ell=\ell_0-K+1}^{\ell_0+K} \hat{\Phi}\left(\frac{\xi_k - \ell/p}{N}\right) \tilde{U}_\ell,$$

where ℓ_0 is the largest integer smaller than or equal to $p\xi_k$ and where $K \in \mathbb{N}$ is a constant.

For reasons of efficiency, K must be a small number. But for small values of K we can only expect the approximation to be good if Φ is chosen such that $|\hat{\Phi}(\nu)|$ decays rapidly for ν away from zero. In Lemma C.2, the requirement $\mathrm{supp}(\Phi) \subset [-\alpha pN/2, \alpha pN/2]$ instead of $\mathrm{supp}(\Phi) \subset [-N/2, N/2]$ was chosen having in mind that the "broader" Φ, the more "peaked" its Fourier transform $\hat{\Phi}$ can be. In [Fou99] the following choice is proposed: $p = 2$, $\alpha = 1.49$, K between 3 and 6 and

$$\Phi(x) := \begin{cases} I_0\left(\frac{\pi K}{\gamma}\sqrt{\gamma^2 - x^2}\right), & |x| \le \gamma := \alpha pN/2 \\ 0, & |x| > \gamma \end{cases},$$

where I_0 is the modified Bessel function of order 0:

$$I_0(x) = \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{k!(k+1)!}.$$

The Fourier transform of Φ is

$$\hat{\Phi}(y) = \frac{2\gamma}{\pi} \frac{\sinh(\pi\sqrt{K^2 - (2\gamma y)^2})}{\sqrt{K^2 - (2\gamma y)^2}}.$$

Lemma C.2 can also be used in (C.14) to compute

$$\hat{u}\left(\frac{k}{2a}\right) \approx \hat{u}_k := \sum_{j=0}^{M-1} u_j \mathrm{e}^{-2\pi i k t_j/(2a)}, \quad k = -N/2, \ldots, N/2-1. \tag{C.20}$$

These values approximate equidistant samples of the Fourier transform of a function which itself is sampled non-equidistantly. We define

$$\tau_j := \frac{N}{2a} t_j \quad \Longrightarrow \quad \tau_j \in [-N/2, N/2], \quad j = 0, \ldots, M-1,$$

and rewrite

$$\hat{u}_k = \sum_{j=0}^{M-1} u_j \mathrm{e}^{-2\pi i k \tau_j / N}.$$

Inserting (C.17) with $\xi = \tau_j$ and j replaced by k leads to

$$\hat{u}_k = \frac{1}{\Phi(k)} \frac{1}{pN} \sum_{\ell \in \mathbb{Z}} \mathrm{e}^{-2\pi i k \ell/(pN)} \sum_{j=0}^{M-1} u_j \hat{\Phi}\left(\frac{\tau_j - \ell/p}{N}\right). \tag{C.21}$$

The ℓ-sum is computed approximately by summing over $\ell \in \{-pN/2, \ldots, pN/2-1\}$ only. For each such ℓ, the j-sum is computed approximately by summing over those j with $|p\tau_j - \ell| \le K$. Then the ℓ-sum is computed by an FFT of length pN, the results of which are scaled by factors $1/\Phi(k)$.

The ideas presented above can be generalized to the computation of two- and higher dimensional Fourier transforms for non-equidistant samples, see [Fou99]. Fourmont also gives error estimates.

Error Estimates for Fourier Inversion in Sect. 2.5

In this paragraph we include two lemmas containing technical details showing that Fourier inversion of convolution equations is a convergent discretization.

Lemma C.3 (Fourier domain error of spline interpolant) *Let $s \in \{1, 2\}$, let $a > 0$, let $Q = (-a, a)^s$ and let N, h, and W be as in (2.85) for $s = 2$ and as in (C.10) for $s = 1$. Let $w \in H^2(\mathbb{R}^s)$ with Fourier transform $\hat{w}$. Assume the decay condition*

$$|w(x)| + |\widehat{w}(x)| \le C(1 + \|x\|_2)^{-s-\varepsilon} \quad \text{for all} \quad x \in \mathbb{R}^s$$

holds for some constants $C, \varepsilon > 0$. Let $w_N \in C_0(\mathbb{R}^s)$ be defined as in (2.100). Then there is a constant $C > 0$ depending on w such that

$$\left| \hat{w}\left(\frac{\beta}{2a}\right) - \widehat{w_N}\left(\frac{\beta}{2a}\right) \right| \le Ch^2 + h^2 \sum_{\alpha \notin W} |w(\alpha h)|. \tag{C.22}$$

Remark The first term on the right hand side of (C.22) tends to zero for $h \to 0$. The second term is a Riemann sum converging to

$$\int_{\mathbb{R}^s \setminus Q} |w(x)|\, dx$$

and *does not vanish* when the discretization level is increased, unless the support of w is contained in Q (in this case $\hat{w}$ would be called a "band limited" function in signal processing). Otherwise, the total error (C.22) can be made arbitrarily small only by choosing N *and* Q large enough.

Proof In the following proof some formulae are explicitly written only for $s = 2$ in order to keep the notation simple. Consider the function g defined by

$$g(y) := \sum_{\alpha \in \mathbb{Z}^s} \hat{w}\left(y - \frac{\alpha}{h}\right). \tag{C.23}$$

Our assumptions about w are sufficient for w to be continuous, have a continuous Fourier transform and make the sum (C.23) converge absolutely everywhere by the decay condition. Function g is periodic (in all coordinate directions) with period $1/h$ and can be developed into a Fourier series:

$$g(y) = \sum_{\beta \in \mathbb{Z}^s} g_\beta \mathrm{e}^{-2\pi i h \beta \cdot y}, \tag{C.24}$$

A short calculation shows that

$$g_\beta := h^s \int_{[-1/(2h),1/(2h)]^s} g(y) \mathrm{e}^{+2\pi i h \beta \cdot y}\, dy = h^s w(\beta h), \quad \beta \in \mathbb{Z}^s. \tag{C.25}$$

From (C.25) one sees that because of the assumed decay condition for w the sum $\sum_{\beta \in \mathbb{Z}^s} |g_\beta|$ converges. Therefore, the sum in (C.24) converges uniformly and pointwise equality holds. Equating (C.23) with (C.24) we get the so-called **generalized Poisson summation formula**

$$\sum_{\alpha \in \mathbb{Z}^s} \hat{w}\left(y - \frac{\alpha}{h}\right) = h^s \sum_{\alpha \in \mathbb{Z}^s} w(\alpha h) \mathrm{e}^{-2\pi i h \alpha \cdot y}. \tag{C.26}$$

Next we introduce the approximant

$$w_\infty(x) := \sum_{\alpha \in \mathbb{Z}^s} w_\alpha \Phi(x/h - \alpha), \quad w_\alpha := w(\alpha h),\ \alpha \in \mathbb{Z}^s,$$

of w, which is Fourier transformed to

$$\widehat{w_\infty}(y) = \sigma(y) W(y),$$

where (for $s = 2$)

$$\sigma(y) = \left(\frac{\sin(\pi h y_1)}{\pi h y_1}\right)^2 \cdot \left(\frac{\sin(\pi h y_2)}{\pi h y_2}\right)^2 \quad \text{and} \quad W(y) = h^2 \sum_{\alpha \in \mathbb{Z}^2} w_\alpha \mathrm{e}^{-2\pi i h \alpha \cdot y}.$$

Comparing with (C.26), one sees that $\widehat{w_\infty}(y) = \sigma(y)g(y)$. Decomposing $g = \hat{w} + \hat{\rho}$ with

$$\hat{\rho}(y) = \sum_{\alpha \neq 0} \hat{w}\left(y - \frac{\alpha}{h}\right),$$

we find – using (2.89) – that

$$\begin{aligned}\hat{w}(y) - \widehat{w_N}(y) &= \hat{w}(y) - \widehat{w_\infty}(y) + \widehat{w_\infty}(y) - \widehat{w_N}(y) \\ &= (1 - \sigma(y))\hat{w}(y) - \sigma(y)\hat{\rho}(y) + \sigma(y)h^2 \sum_{\alpha \notin W} w_\alpha \mathrm{e}^{-2\pi i h\alpha \cdot y} \qquad \text{(C.27)}\end{aligned}$$

We will estimate the size of this difference for $y = \beta/2a$, $\beta \in W$. To do so, we need two auxiliary results. The first one is the estimate

$$|\hat{w}(y)| \leq \frac{C}{1 + \|y\|_2^2}, \qquad \text{(C.28)}$$

which immediately follows from the decay condition. The second one is the expansion

$$\frac{\sin(\pi x)}{\pi x} = \sum_{n=0}^{\infty} \frac{(-\pi^2 x^2)^n}{(2n+1)!} = 1 - \frac{\pi^2 x^2}{3!} + R_x, \qquad \text{(C.29)}$$

which is valid for all $x \in \mathbb{R}$ and where $|R_x| \leq \pi^4 |x|^4/5!$ according to Leibniz' criterion. From (C.29) one derives for $s = 2$ the estimate

$$|1 - \sigma(\beta/2a)| \leq A_1 \frac{\|\beta\|_2^2}{N^2} + A_2 \frac{\|\beta\|_2^4}{N^4} + A_3 \frac{\|\beta\|_2^6}{N^6} + \ldots + A_8 \frac{\|\beta\|_2^{16}}{N^{16}} \qquad \text{(C.30)}$$

for some constants $A_1, \ldots, A_8$ (similar estimate in case $s = 1$). Also, from (C.28) we get $|\hat{w}(\beta/2a)| \leq A/(4a^2 + \|\beta\|_2^2)$ with some constant A. Since $\|\beta\|_2/N \leq 1$ for $\beta \in W$, this means that the first summand on the right hand side of (C.27) can be estimated for $\beta \in W$:

$$\left| \left[1 - \sigma\left(\frac{\beta}{2a}\right) \right] \hat{w}\left(\frac{\beta}{2a}\right) \right| \leq D_1 h^2 \qquad \text{(C.31)}$$

for some constant D_1. Also, one deduces from (C.28), that

$$\left| \hat{w}\left(\frac{\beta}{2a} - \frac{\alpha}{h}\right) \right| \leq h^2 \frac{C}{\|\beta/N - \alpha\|_2^2} \quad \text{for} \quad \beta \in W, \alpha \neq 0.$$

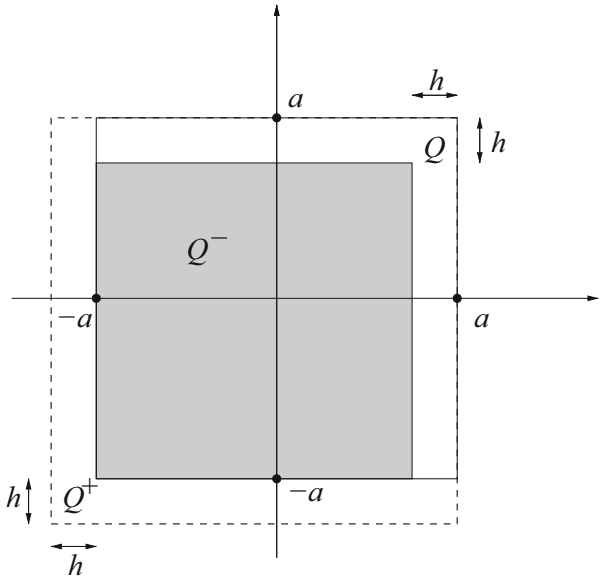

Fig. C.2 *Rectangle* Q, support Q^+ of u_N, and "domain of approximation" Q^-

Together with $|\sigma(y)| \leq 1$ this shows that the second term on the right hand side of (C.27) can be estimated by

$$\left|\sigma\left(\frac{\beta}{2a}\right)\hat{\rho}\left(\frac{\beta}{2a}\right)\right| \leq D_2 h^2$$

for some constant D_2. The last term on the right hand side of (C.27) evidently can be estimated by

$$h^2 \sum_{\alpha \notin W} |w(\alpha h)|$$

and this ends the proof. □

In Fig. C.2 we show Q together with $Q^+ = (-a-h, a)^s$, which is the support of u_N defined by (2.99). Function u_N can only be expected to be a good approximant of u on the domain $Q^- = (-a, a-h)^s$ contained in Q.

Lemma C.4 (Error from Fourier interpolation) *Let* $s \in \{1, 2\}$, *let* $a > 0$, *let* $Q = (-a, a)^s$, *let* $Q^- := (-a, a-h)^s$ *and let* $u \in H^s(Q)$ *be extended by zero values to become a function* $u \in L_2(\mathbb{R}^s)$ *with integrable Fourier transform* $\hat{u}$. *Let* $h = 2a/N$, $N \in \mathbb{N}$ *an even number, and determine* $u_N \in C_0(\mathbb{R}^s)$ *via* (2.93) *and* (2.94) *for* $s = 2$ *or via* (C.12), (C.15) *and* (C.16) *for* $s = 1$. *Then there is a constant* $C > 0$ *such that*

$$\|u - u_n\|_{L_2(Q^-)} \leq Ch\|u\|_{H^1(Q)}. \tag{C.32}$$

Proof For $u \in L_2(\mathbb{R}^s) \cap H^s(Q)$ define the function $g : \mathbb{R}^s \to \mathbb{R}$,

$$g(x) = \sum_{\alpha \in \mathbb{Z}^s} u(x - 2a\alpha), \tag{C.33}$$

which is periodic with period $2a$ and can be developed into a Fourier series, namely

$$g(x) = \sum_{\beta \in \mathbb{Z}^s} g_\beta \mathrm{e}^{+2\pi i \beta \cdot x/(2a)}, \quad x \in Q. \tag{C.34}$$

A short calculation shows that

$$g_\beta := \left(\frac{1}{2a}\right)^s \int_Q g(x) \mathrm{e}^{-2\pi i \beta \cdot x/(2a)}\, dx = \left(\frac{1}{2a}\right)^s \hat{u}\left(\frac{\beta}{2a}\right), \quad \beta \in \mathbb{Z}^s. \tag{C.35}$$

Equating (C.33) with (C.34) we get a(nother) generalized Poisson summation formula

$$\sum_{\alpha \in \mathbb{Z}^s} u\,(x - 2a\alpha) = \left(\frac{1}{2a}\right)^s \sum_{\beta \in \mathbb{Z}^s} \hat{u}\left(\frac{\beta}{2a}\right) \mathrm{e}^{+2\pi i \beta \cdot x/(2a)}. \tag{C.36}$$

Identity (C.36) will be used for $u - u_N$ instead of u. For arguments $x \in Q^-$ all summands on the left hand side then vanish except for $\alpha = 0$ (see Fig. C.2). We therefore get from (C.36)

$$\begin{aligned} \|u - u_N\|_{L_2(Q^-)} &\le \left(\frac{1}{2a}\right)^s \left\| \sum_{\beta \in \mathbb{Z}^s} \left(\hat{u}\left(\frac{\beta}{2a}\right) - \hat{u}_N\left(\frac{\beta}{2a}\right)\right) \mathrm{e}^{2\pi i \beta \cdot \bullet/(2a)} \right\|_{L_2(Q^-)} \\ &\le \left(\frac{1}{2a}\right)^s \left\| \sum_{\beta \in \mathbb{Z}^s} \left(\hat{u}\left(\frac{\beta}{2a}\right) - \hat{u}_N\left(\frac{\beta}{2a}\right)\right) \mathrm{e}^{2\pi i \beta \cdot \bullet/(2a)} \right\|_{L_2(Q)}. \end{aligned} \tag{C.37}$$

Because of the orthonormality of the functions $x \mapsto \mathrm{e}^{2\pi i \beta \cdot x/(2a)}$ with respect to the scalar product $\langle \bullet | \bullet \rangle_{L_2(Q)}$, we get from (C.37)

$$\|u - u_N\|^2_{L_2(Q^-)} \le S_1 + S_2 + S_3, \quad \text{where} \tag{C.38}$$

$$S_1 := (2a)^{-2s} \sum_{\beta \in W} \left| \hat{u}\left(\frac{\beta}{2a}\right) - \widehat{u_N}\left(\frac{\beta}{2a}\right)\right|^2,$$

$$S_2 := (2a)^{-2s} \sum_{\beta \notin W} \left| \hat{u}\left(\frac{\beta}{2a}\right) \right|^2, \text{ and}$$

$$S_3 := (2a)^{-2s} \sum_{\beta \notin W} \left| \widehat{u_N}\left(\frac{\beta}{2a}\right) \right|^2.$$

From the interpolation conditions (2.95) we conclude $S_1 = 0$. Differentiating (C.36) we get

$$\frac{\partial}{\partial x_j} u(x) = (2a)^{-s} \sum_{\beta \in \mathbb{Z}^s} 2\pi i \cdot \frac{\beta_j}{2a} \cdot \hat{u}\left(\frac{\beta}{2a}\right) \mathrm{e}^{2\pi i \beta \cdot x/(2a)}, \quad x \in Q,\ j = 1, \ldots, s.$$

From this, we get

$$\|u\|^2_{H^1(Q)} = \sum_{|\alpha| \le 1} \|D^\alpha u\|^2_{L_2(Q)} = (2a)^{-2s} \sum_{\beta \in \mathbb{Z}^s} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\beta\|_2^2\right) \left| \hat{u}\left(\frac{\beta}{2a}\right) \right|^2. \tag{C.39}$$

Now observe that $\|\beta\|_2^2 \ge (N/2)^2 = (a/h)^2$ for $\beta \notin W$. Consequently,

$$1 + \left[\frac{\pi}{a}\right]^2 \|\beta\|_2^2 \ge 1 + \left[\frac{\pi}{a}\right]^2 \left(\frac{a}{h}\right)^2 \ge \frac{1}{h^2} \quad \text{for} \quad \beta \notin W. \tag{C.40}$$

Therefore we can estimate

$$S_2 \le h^2 (2a)^{-2s} \sum_{\beta \notin W} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\beta\|_2^2\right) \left| \hat{u}\left(\frac{\beta}{2a}\right) \right|^2 \overset{(C.39)}{\le} h^2 \|u\|^2_{H^1(Q)}.$$

It remains to estimate S_3. For notational simplicity this will only be done for $s = 2$. The case $s = 1$ can be treated analogously. Reusing (C.40) we at first get

$$S_3 \le h^2 (2a)^{-4} \sum_{\beta \notin W} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\beta\|_2^2\right) \left| \widehat{u_N}\left(\frac{\beta}{2a}\right) \right|^2. \tag{C.41}$$

Every index $\beta \notin W$ can uniquely be written in the form $\beta = \alpha + N\gamma$, where $\alpha \in W$ and $\gamma \in \mathbb{Z}^2$, $\gamma \neq 0$. From (2.90) and (2.91) we know that[1]

$$\widehat{u_N}(\beta/(2a)) = \sigma_\beta U_\beta = \sigma_{\alpha + N\gamma} U_\alpha$$

[1] In case $s = 1$ one rather uses (C.15).

because of the discrete Fourier coefficients' periodicity. Rewriting the right hand side of (C.41), we get

$$S_3 \leq h^2(2a)^{-4} \sum_{\alpha\in W} \sum_{\gamma\neq 0} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\alpha + N\gamma\|_2^2\right) \left|\sigma_{\alpha+N\gamma} U_\alpha\right|^2 . \tag{C.42}$$

It is easy to see that $\sigma_{\alpha+N\gamma} = 0$ for $(\alpha_1 = 0$ and $\gamma_1 \neq 0)$ or $(\alpha_2 = 0$ and $\gamma_2 \neq 0)$. The corresponding terms on the right hand side of (C.42) vanish. For the non-vanishing terms we get

$$\sigma_{\alpha+N\gamma} = \sigma_\alpha \cdot \frac{1}{(1 + N\gamma_1/\alpha_1)^2(1 + N\gamma_2/\alpha_2)^2},$$

thereby setting $0/0 := 0$ in case $\alpha_1 = \gamma_1 = 0$ or $\alpha_2 = \gamma_2 = 0$. Thus we find for all pairs (α, γ) such that $\sigma_{\alpha+N\gamma}$ does not vanish and also using $0/0 := 0$:

$$\begin{aligned}
&\left(1 + \left[\frac{\pi}{a}\right]^2 \|\alpha + N\gamma\|_2^2\right) |\sigma_{\alpha+N\gamma}|^2 \\
&= |\sigma_\alpha|^2 \left(\frac{1 + \left[\frac{\pi}{a}\right]^2 \left(\alpha_1^2(1 + \gamma_1 N/\alpha_1)^2 + \alpha_2^2(1 + \gamma_2 N/\alpha_2)^2\right)}{(1 + N\gamma_1/\alpha_1)^4(1 + N\gamma_2/\alpha_2)^4}\right) \\
&\leq |\sigma_\alpha|^2 \left[1 + \left(\frac{\pi}{a}\right)^2 \|\alpha\|_2^2\right] \frac{1}{(1 + N\gamma_1/\alpha_1)^2(1 + N\gamma_2/\alpha_2)^2} \\
&\leq |\sigma_\alpha|^2 \left[1 + \left(\frac{\pi}{a}\right)^2\right] \|\alpha\|_2^2 \frac{1}{(1 + 2\gamma_1)^2(1 + 2\gamma_2)^2},
\end{aligned}$$

the final estimate being true without any restriction for $\alpha \in W$ and $\gamma \in \mathbb{Z}^2 \setminus \{0\}$. Inserting into (C.42) we arrive at

$$S_3 \leq h^2(2a)^{-4} \sum_{\alpha\in W} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\alpha\|_2^2\right) |\sigma_\alpha U_\alpha|^2 \left(\sum_{\gamma\in\mathbb{Z}^2} \frac{1}{(1 + 2\gamma_1)^2(1 + 2\gamma_2)^2}\right)$$

Since the sum over γ converges and since $\sigma_\alpha U_\alpha = \widehat{u_N}(\alpha/(2a)) = \hat{u}(\alpha/(2a))$ for all $\alpha \in W$, we finally get

$$S_3 \leq Ch^2(2a)^{-4} \sum_{\alpha\in W} \left(1 + \left[\frac{\pi}{a}\right]^2 \|\alpha\|_2^2\right) \left|\hat{u}\left(\frac{\alpha}{2a}\right)\right|^2 \leq Ch^2 \|u\|_{H^1(Q)}^2$$

and this ends the proof. □

Lemma C.4 assumes a hypothetic knowledge of exact Fourier transform values $\hat{u}(\beta/(2a))$, $\beta \in W$. In reality we only know approximate values

$$\frac{\widehat{w_N}\,(\beta/(2a))}{\hat{k}\,(\beta/(2a))} \approx \hat{u}\left(\frac{\beta}{2a}\right) = \frac{\hat{w}\,(\beta/(2a))}{\hat{k}\,(\beta/(2a))}, \quad \beta \in W.$$

Then the sum S_1 in (C.38) does no longer vanish. Rather, approximation errors $\hat{w}(\beta/(2a)) - \widehat{w_N}(\beta/(2a))$, magnified by the division through $\hat{k}(\beta/(2a))$, contribute to the overall error.

Appendix D
Regularization Property of CGNE

Theorem 3.41 showed that CGNE formally (i.e. in the sense of Definition 3.14) is a regularization method for the least squares problem. It did not show, however, that stopping CGNE after a number of iterations can produce a result closer to $\hat{x} = A^+ b$ than is the solution $\bar{x} = A^+ b^\delta$ of the least squares problem for perturbed data – compare Theorem 3.6. Theorem 3.41 therefore can not justify using CGNE for practical regularization of (finite-dimensional) least squares problems. The following proof of the regularizing property of CGNE is modelled after Section 4.3.3 of [Lou89].

We have already seen in the proof of Theorem 3.41, that (3.177) guarantees a stop of CGNE according to the criterion (3.178) after k iterations, with $0 < k \leq n$. Let now k be fixed accordingly. If $x_k = \bar{x}$ (i.e. CGNE has minimized $\|b^\delta - Ax\|_2$ exactly after k steps), then $\hat{x} - x_k = (A^T A)^{-1} A^T (b - b^\delta)$ and $\|\hat{x} - x_k\|_2 \leq \|(A^T A)^{-1} A^T\|_2 \|b - b^\delta\|_2 \leq \delta/\sigma_n$ with σ_n being the smallest singular value of A. This confirms estimate (3.179) from Theorem 3.41, but no regularization beyond this can be found. In the following, we rather assume that CGNE will be stopped by (3.178) for some index $k < n$, but $x_k \neq \bar{x}$ and that therefore

$$\rho_0, \ldots, \rho_k \neq 0. \tag{D.1}$$

Every $x \in \mathscr{K}_k$ (see (3.175)) can be written in the form $x = p(A^T A) A^T b^\delta$, where $p \in P_{k-1}$ is a polynomial of degree $k-1$ or less, uniquely defined by x. Using an SVD $A = U \Sigma V^T$ (the columns of U are $u_1, \ldots, u_m \in \mathbb{R}^m$, the columns of V are $v_1, \ldots, v_n \in \mathbb{R}^n$, and $\sigma_1 \geq \cdots \geq \sigma_n > 0$ are the singular values of A) and a reduced SVD $A = \hat{U} \hat{\Sigma} V^T$ of A (omitting the last $m - n$ columns of U and rows of Σ), a

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9

short calculation shows

$$\begin{aligned}\|b^\delta - Ax\|_2^2 &= \|b^\delta - Ap(A^TA)A^Tb^\delta\|_2^2 \\ &= \sum_{i=1}^{n} \left(1 - \sigma_i^2 p(\sigma_i^2)\right)^2 (u_i^T b^\delta)^2 + \sum_{i=n+1}^{m} (u_i^T b^\delta)^2 \end{aligned} \tag{D.2}$$

for every $x \in \mathscr{K}_k$, where $p \in P_{k-1}$ is defined by x. CGNE produces the vector $x = x_k \in \mathscr{K}_k$ minimizing $\|b^\delta - Ax\|_2$. The corresponding polynomial from P_{k-1}, which minimizes (D.2) and *which depends on* b^δ, will be designated by p_{k-1}. Thus,

$$x_k = p_{k-1}(A^TA)A^Tb^\delta = \sum_{i=1}^{n} \sigma_i p_{k-1}(\sigma_i^2)(u_i^T b^\delta)v_i, \tag{D.3}$$

as compared to the unregularized solution

$$\bar{x} = A^+ b^\delta = \sum_{i=1}^{n} \frac{1}{\sigma_i}(u_i^T b^\delta)v_i. \tag{D.4}$$

Now let

$$P_k^0 := \{q \in P_k;\ q(0) = 1\}.$$

Every polynomial $q \in P_k^0$ can uniquely be written in the form $q(t) = 1 - tp(t)$ for some $p \in P_{k-1}$. Especially set $q_k(t) := 1 - tp_{k-1}(t)$, with $p_{k-1} \in P_{k-1}$ as above. With $q_k \in P_k^0$ defined this way, the residual corresponding to x_k can be expressed in the form

$$\rho_k = A^Tb^\delta - A^TAx_k = A^Tb^\delta - A^TAp_{k-1}(A^TA)A^Tb^\delta = q_k(A^TA)A^Tb^\delta.$$

From this one computes the Euclidean scalar products

$$\begin{aligned}\langle \rho_j | \rho_\ell \rangle &= \left[q_j(A^TA)A^Tb^\delta\right]^T \left[q_\ell(A^TA)A^Tb^\delta\right] \\ &= \sum_{i=1}^{n} \sigma_i^2 q_j(\sigma_i^2) q_\ell(\sigma_i^2)(u_i^T b^\delta)^2 \quad \text{for } 1 \le j, \ell \le k, \end{aligned} \tag{D.5}$$

where we have used the SVD of A. From Lemma 3.39 one knows that $\langle \rho_j | \rho_\ell \rangle = 0$ for $j \neq \ell$ and this shows that the polynomials $q_1, \ldots, q_k \in P_k^0$ are orthogonal with respect to the scalar product defined on P_k by

$$\langle q | \tilde{q} \rangle_0 := \sum_{i=1}^{n} \sigma_i^2 q(\sigma_i^2)\tilde{q}(\sigma_i^2)(u_i^T b^\delta)^2, \quad q, \tilde{q} \in P_k, \tag{D.6}$$

which depends on b^δ.[1] From $\hat{x} = \sum(1/\sigma_i)(u_i^T b)v_i$ and from (D.3) one derives

$$\begin{aligned} \|\hat{x} - x_k\|_2^2 &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[(u_i^T b) - \sigma_i^2 p_{k-1}(\sigma_i^2)(u_i^T b^\delta)\right]^2 \qquad \text{(D.10)} \\ &= \sum_{\sigma_i \geq \tau} \cdots + \sum_{\sigma_i < \tau} \cdots , \end{aligned}$$

where τ is a freely chosen positive number, used to split the sum (D.10) into two parts, which will be bounded independently of each other. Concerning the first sum,

[1] It is evident that $\langle \bullet | \bullet \rangle_0$ is a symmetric bilinear form on P_k. Positive definiteness is established if

$$\langle q|q\rangle_0 = \sum_{i=1}^{n} \sigma_i^2 \left[q(\sigma_i^2)\right]^2 (u_i^T b^\delta)^2 > 0 \quad \text{for} \quad q \in P_k \setminus \{0\} \qquad \text{(D.7)}$$

can be shown. Since $\sigma_i > 0$ for $i = 1, \ldots, n$,

$$\langle q|q\rangle_0 = 0 \quad \Longleftrightarrow \quad q(\sigma_i^2)(u_i^T b^\delta) = 0, \quad i = 1, \ldots, n. \qquad \text{(D.8)}$$

Assume now that σ_{k_j}, $j = 1, \ldots s$, are *s pairwise different* singular values of A with multiplicities $\mu_1, \ldots, \mu_s$, respectively. Numbering multiple singular values multiple times, we denote the above by $\sigma_{i_1}, \ldots \sigma_{i_\ell}$, $\ell = \mu_1 + \ldots + \mu_s$. Let u_{i_j}, $j = 1, \ldots, \ell$ be the corresponding columns from $U \in \mathbb{R}^{m,m}$. Assume further that

$$b^\delta \in \langle u_{i_1}, \ldots, u_{i_\ell}, u_{n+1}, \ldots, u_m \rangle.$$

In this case, repeating the argument that led to (D.2), one gets for $x = p(A^T A)A^T b^\delta \in \mathscr{K}_k$

$$b^\delta - Ax = \sum_{j=1}^{\ell} \underbrace{(1 - \sigma_{i_j}^2 p(\sigma_{i_j}^2))}_{q(\sigma_{i_j}^2)} (u_{i_j}^T b^\delta) u_{i_j} + \sum_{i=n+1}^{m} (u_i^T b^\delta) u_i. \qquad \text{(D.9)}$$

Now if $s \leq k$, then one could choose

$$q_s(t) := \prod_{j=1}^{s} \left(1 - \frac{t}{\sigma_{k_j}^2}\right) \in P_s^0 \subseteq P_k^0$$

to achieve $b^\delta - Ax_s \in \langle u_{n+1}, \ldots, u_m \rangle = \mathscr{R}_A^\perp$ for the vector $x_s \in \mathscr{K}_s$ corresponding to q_s. But this would mean $A^T(b^\delta - Ax_s) = 0$ and we would have $x_s = \bar{x}$. On the other hand, from (D.9) it can immediately be seen that the polynomial q_s just defined minimizes $\|b^\delta - Ax\|_2$ over $\mathscr{K}_s \subset \mathscr{K}_k$ and this means that x_s is exactly the vector produced by CGNE after s steps: CGNE would stop after $s \leq k$ steps in contradiction to (D.1). If CGNE does *not* stop after k iterations this means that at least $k + 1$ terms from $(u_i^T b^\delta)$, $i = 1, \ldots, n$, belonging to *different* singular values σ_i must not vanish. But then in fact (D.8) can not happen for $q \neq 0$, since there is no polynomial $q \in P_k \setminus \{0\}$ having $k + 1$ different zeros.

begin with

$$\sum_{\sigma_i \geq \tau} \cdots \leq \tau^{-2} \sum_{\sigma_i \geq \tau} \left[u_i^T (b - b^\delta) + (1 - \sigma_i^2 p_{k-1}(\sigma_i^2)) u_i^T b^\delta \right]^2 .$$

Making use of $(a+b)^2 \leq 2a^2 + 2b^2$, of (3.177), and of

$$\begin{aligned} \sum_{\sigma_i \geq \tau} (1 - \sigma_i^2 p_{k-1}(\sigma_i^2))^2 (u_i^T b^\delta)^2 &\leq \sum_{i=1}^{n} (1 - \sigma_i^2 p_{k-1}(\sigma_i^2))^2 (u_i^T b^\delta)^2 \\ &\overset{(D.2)}{\leq} \|b^\delta - Ax_k\|_2^2 \overset{(3.178)}{\leq} \delta^2, \end{aligned}$$

this leads to

$$\sum_{\sigma_i \geq \tau} \cdots \leq 2\tau^{-2} \|b - b^\delta\|_2^2 + 2\tau^{-2}\delta^2 \leq 4\tau^{-2}\delta^2. \tag{D.11}$$

The second sum on the right hand side of (D.10) equals zero for $\tau \leq \sigma_n$. In this case one ends up with the same (qualitative) result as in Theorem 3.41. But one may as well choose $\tau > \sigma_n$, in which case the second sum does not vanish. An upper bound for this sum can be split into two summands:

$$\sum_{\sigma_i < \tau} \cdots \leq 2 \underbrace{\sum_{\sigma_i < \tau} \frac{(u_i^T b)^2}{\sigma_i^2}}_{=: S_1} + 2 \underbrace{\sum_{\sigma_i < \tau} \sigma_i^2 \left[p_{k-1}(\sigma_i^2) \right]^2 (u_i^T b^\delta)^2}_{=: S_2} \tag{D.12}$$

From $\hat{x} = A^+ b = \sum (1/\sigma_i)(u_i^T b) v_i$ one gets

$$S_1 \leq \frac{\tau^2}{\sigma_n^2} \sum_{\sigma_i < \tau} \frac{(u_i^T b)^2}{\sigma_i^2} \leq \frac{\tau^2}{\sigma_n^2} \|\hat{x}\|_2^2. \tag{D.13}$$

We also get

$$\begin{aligned} S_2 &= \sum_{\sigma_i < \tau} \sigma_i^2 \left[p_{k-1}(\sigma_i^2) \right]^2 (u_i^T b + u_i^T (b^\delta - b))^2 \\ &\leq 2 \sum_{\sigma_i < \tau} \left[\sigma_i p_{k-1}(\sigma_i^2) \right]^2 (u_i^T b)^2 + 2 \sum_{\sigma_i < \tau} \left[\sigma_i p_{k-1}(\sigma_i^2) \right]^2 (u_i^T b^\delta - u_i^T b)^2 \\ &\leq 2\tau^2 \left[Q_k(\tau) \right]^2 \sum_{\sigma_i < \tau} (u_i^T b)^2 + 2\tau^2 \left[Q_k(\tau) \right]^2 \sum_{\sigma_i < \tau} (u_i^T (b^\delta - b))^2, \end{aligned}$$

where the following abbreviation was used

$$Q_k(\tau) := \max_{0\le t\le\tau} |p_{k-1}(t^2)|. \tag{D.14}$$

Making use of the orthogonality of the polynomials $q_1, \ldots, q_k$, from which it follows that q_k has k distinct real zeros

$$0 < t_{k,1} < t_{k,2} < \cdots < t_{k,k} < \|A\|_2^2, \tag{D.15}$$

it is proven in Lemma 4.3.15 of [Lou89] that

$$Q_k(\tau) = \max_{0\le t\le\tau} |p_{k-1}(t^2)| = p_{k-1}(0) = \sum_{j=1}^{k} \frac{1}{t_{k,j}} \ge \frac{1}{t_{k,1}}.$$

Consequently

$$\tau^2 \left[Q_k(\tau)\right]^2 \le 1 \quad \text{if} \quad \tau \le [p_{k-1}(0)]^{-1} \le t_{k,1}, \tag{D.16}$$

and the above estimate can be continued as follows

$$S_2 \le 2 \sum_{\sigma_i<\tau} \sigma_i^4 \frac{(u_i^T b)^2}{\sigma_i^4} + 2\|b^\delta - b\|_2^2 \le 2\tau^4 \frac{\|\hat{x}\|_2^2}{\sigma_n^2} + 2\delta^2. \tag{D.17}$$

Putting (D.11), (D.12), (D.13) and (D.16) together, one arrives at

$$\|\hat{x} - x_k\|_2^2 \le 4\delta^2(1 + \tau^{-2}) + 2\frac{\|\hat{x}\|_2^2}{\sigma_n^2}(\tau^2 + 2\tau^4).$$

For $\tau \le \min\{1, [p_{k-1}(0)]^{-1}\}$ one further estimates

$$\|\hat{x} - x_k\|_2^2 \le 8\left(\tau^{-2}\delta^2 + \tau^2 \frac{\|\hat{x}\|_2^2}{\sigma_n^2}\right). \tag{D.18}$$

This estimate shows that the choice of a (not too) small value τ may counterbalance the negative effect of a small singular value σ_n. Setting for example

$$\tau = \theta \cdot \sqrt{\frac{\sigma_n \cdot \delta}{\|\hat{x}\|_2}}, \quad \theta \text{ such that } \tau \le \min\{1, [p_{k-1}(0)]^{-1}\},$$

one derives the estimate

$$\|\hat{x} - x_k\|_2^2 \leq \left(8\theta^2 + 8\theta^{-2}\right) \frac{\|\hat{x}\|_2}{\sigma_n} \delta$$

or, equivalently,

$$\|\hat{x} - x_k\|_2 \leq C \sqrt{\frac{\|\hat{x}\|_2}{\sigma_n}} \cdot \sqrt{\delta}, \quad C = \sqrt{8\theta^2 + 8\theta^{-2}}. \tag{D.19}$$

As to the regularizing effect of CGNE, the polynomial $q_k(t) = 1 - tp_{k-1}(t)$ has k distinct zeroes $0 < t_{k,1} < \cdots < t_{k,k} < \|A\|_2^2 = \sigma_1^2$, see (D.15). This means that p_{k-1} interpolates the function $t \mapsto 1/t$ at these zeroes. Consequently, the polynomial $tp_{k-1}(t^2)$ also interpolates $t \mapsto 1/t$ at $t_{k,1}, \ldots, t_{k,k}$, but has a zero at $t = 0$, where $t \mapsto 1/t$ has a pole. In view of the two representations (D.3) and (D.4) this explains how CGNE achieves a regularization of least squares problems.

References

[BCL77] A. Bamberger, G. Chavent, and P. Lailly. Etude mathématique et numérique d'un problème inverse pour l'équation des ondes à une dimension. *Rapport LABORIA nr. 226, IRIA*, 1977.

[BCL79] A. Bamberger, G. Chavent, and P. Lailly. About the Stability of the Inverse Problem in 1-D Wave Equations—Application to the Interpretation of Seismic Profiles. *Appl. Math. Optim.*, 5:1–47, 1979.

[BCL99] M. A. Branch, T. F. Coleman, and Y. Li. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.*, 21(1):1–23, 1999.

[Bjö96] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

[Bra07] D. Braess. *Finite Elements, Third Edition*. Cambridge University Press, 2007.

[Cha74] G. Chavent. Identification of Functional Parameters in Partial Differential Equations. In R. E. Goodson and M. Polis, editors, *Identification of Parameters in Distributed Systems*, pages 31–48. The American Society of Mechanical Engineering, 1974.

[Cha09] G. Chavent. *Nonlinear Least Squares for Inverse Problems*. Springer, 2009.

[DB74] G. Dahlquist and Å. Björck. *Numerical Methods*. Prentice Hall, 1974.

[dB90] C. de Boor. *Splinefunktionen*. Birkhäuser, 1990.

[Dem97] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

[EKN89] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5:523–540, 1989.

[Eld77] L. Eldén. Algorithms for the Regulariation of Ill-conditioned Least Squares Problems. *BIT*, 17:134–145, 1977.

[Eld82] L. Eldén. A weigthed pseudoinverse, generalized singular values, and constrained least squares problems. *BIT*, 22:487–502, 1982.

[Eng97] H. W. Engl. *Integralgleichungen*. Springer, 1997.

[Eva98] L. C. Evans. *Partial Differential Equations*. AMS, 1998.

[FNS10] W. Freeden, M. Z. Nashed, and T. Sonar. *Handbook of Geomathematics*. Springer, 2010.

[Fou99] K. Fourmont. *Schnelle Fourier-Transformation bei nichtäquidistanten Gittern und tomographische Anwendungen*. Ph. D. Thesis, Universität Münster, Germany, 1999.

[Gau72] W. Gautschi. Attenuation Factors in Practical Fourier Analysis. *Numer. Math.*, 18:373–400, 1972.

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9

[GHW79] G. H. Golub, M. Heath, and G. Wahba. Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21:215–224, 1979.

[Han92] P. C. Hansen. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Rev.*, 34:561–580, 1992.

[Han97a] M. Hanke. A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Problems*, 13:79–95, 1997.

[Han97b] M. Hanke. Regularizing properties of a truncated Newton-CG algorithm for nonlinear inverse problems. *Numer. Funct. Anal. and Optimiz.*, 18:971–993, 1997.

[Hof99] B. Hofmann. *Mathematik inverser Probleme*. Teubner, 1999.

[Isa90] V. Isakov. *Inverse Source Problems*. AMS, 1990.

[Isa06] V. Isakov. *Inverse Problems for Partial Differential Equations, 2nd ed.* Springer, 2006.

[Kir96] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Springer, 1996.

[KR14] A. Kirsch and A. Rieder. Seismic tomography is locally ill-posed. *Inverse Problems*, 30:125001, 2014.

[Lai80] P. Lailly. The inverse problem in 1-D reflection seismics. In R. Cassinis, editor, *The Solution of the Inverse Problem in Geophysical Interpretation*, pages 103–140. Plenum Press, 1980.

[Lou89] A. K. Louis. *Inverse und schlecht gestellte Probleme*. Teubner, 1989.

[Lou96] A.K. Louis. Approximate inverse for linear and some nonlinear problems. *Inverse Problems*, 12:175–190, 1996.

[LT03] S. Larsson and V. Thomée. *Partial Differential Equations with Numerical Methods*. Springer, 2003.

[LY08] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. Springer, 2008.

[Mat14] Matlab. *Release 2014b*. The MathWorks Inc., Natick, Massachusetts, U.S.A., 2014.

[MF08] V. Michel and A. S. Fokas. A unified approach to various techniques for the non-uniqueness of the inverse gravimetric problem and wavelet-based methods. *Inverse Problems*, 24:1–23, 2008.

[Mor78] J. J. Moré. The Levenberg-Marquardt Algorithm: Implementation and Theory. In G. A. Watson, editor, *Numerical Analysis. Proceedings Biennial Conference Dundee 1977, Lecture Notes in Mathematics*, volume 630, pages 105–116. Springer, 1978.

[Nat77] F. Natterer. Regularisierung schlecht gestellter Probleme durch Projektionsverfahren. *Numer. Math.*, 28:329–341, 1977.

[Nie86] Y. Nievergelt. Elementary Inversion of Radon's Transform. *SIAM Review*, 28:79–84, 1986.

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. B. Flannery. *Numerical Recipes in C, 2nd edition*. Cambridge University Press, 1992.

[Ram02] R. Ramlau. Morozov's discrepancy principle for Tikhonov regularization of nonlinear operators. *Numer. Funct. Anal. Optimization*, 23:147–172, 2002.

[Rei67] C. H. Reinsch. Smoothing by Spline Functions. *Numer. Math.*, 10:177–183, 1967.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[Sam11] D. Sampietro. GOCE Exploitation for Moho Modeling and Applications. In L. Ouwehand, editor, *Proc. of 4th International GOCE User Workshop, Munich, Germany*. ESA Communications, 2011.

[Sch07] G. T. Schuster. *Basics of Seismic Wave Theory*. http://utam.gg.utah.edu/tomo06/06_seg/basicseisbook.pdf, 2007.

[Sch12] S. Schäffler. *Global Optimization. A Stochastic Approach.* Springer, 2012.

[Sym09] W. W. Symes. The seismic reflection inverse problem. *Inverse Problems*, 25:123008, 2009.

[TB97] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[Wal98] W. Walter. *Ordinary Differential Equations*. Springer, 1998.

Index

M. Richter, *Inverse Problems*, Lecture Notes in Geosystems Mathematics and Computing, DOI 10.1007/978-3-319-48384-9