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 \( T_{n,m} : X_n \rightarrow Y_m \) 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
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 focuses 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

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

(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

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

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 $\mathcal{S}_k(t_1, \ldots, t_m)$. 
2.1 Approximation of Functions

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

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

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

\[
N_{j,1}(t) := \begin{cases} 
1, & t_j \leq t < t_{j+1}, \quad j = 1, \ldots, m - 1 \\
0, & \text{else}
\end{cases}
\]

(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”

\[
N_{j,2}(t) := \begin{cases} 
\frac{t - t_{j-1}}{t_j - t_{j-1}}, & t \in [t_{j-1}, t_j] \quad (\text{if } j \geq 2) \\
\frac{t_{j+1} - t}{t_{j+1} - t_j}, & t \in [t_j, t_{j+1}] \quad (\text{if } j \leq m - 1), \quad j = 1, \ldots, m, \\
0, & \text{else}
\end{cases}
\]

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:

\[
s(t) = \sum_{j=1}^{m+k-2} \alpha_j N_{j,k}(t), \quad a \leq t \leq b.
\]

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

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

with coefficients \( \alpha_j \) defined by

\[
\alpha_j := \left( \frac{1}{t_{j+1} - t_j} \int_{t_j}^{t_{j+1}} f(t) \, dt \right), \quad j = 1, \ldots, m - 1.
\]
$I_1$ 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 $f$.\footnote{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 $s \in \mathcal{S}_1(t_1, \ldots, t_m)$ of $f \in C[a,b]$.} The approximant $I_1(f)$ has the same local mean values as $f$. Another example would be the interpolation operator

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

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

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

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

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

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].
Theorem 2.2 (Approximation errors) Let \( k \in \{1, 2\} \), let \( a = t_1 < \ldots < t_m = b \) and let \( h := \max_{i=1,...,m-1} \{t_{i+1} - t_i\} \). 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 \((H^1(a, b), \| \cdot \|_{H^1(a, b)})\) to \((L_2(a, b), \| \cdot \|_{L_2(a, b)})\) or to \((H^1(a, b), \| \cdot \|_{H^1(a, b)})\) for \( k = 1 \) or \( k = 2 \), respectively, and the bound
\[
\| f - I_k(f) \|_{L_2(a, b)} \leq kh \| f \|_{H^1(a, b)}
\]
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 \( \| \cdot \|_{L_2(a, b)} \), 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 \( D \subseteq \mathbb{R}^2 \). 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 \( \Omega \subseteq \mathbb{R}^2 \). 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, \ldots) 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 \( \rho_T \) with
\[
\rho_T \geq h_T / \kappa, \quad h_T := \text{diam}(T),
\]
(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

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

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

We will designate by $S^1_T$ the set of all piecewise constant functions for a given partitioning $T$. This means that $s \in S^1_T$ shall take a constant value $c_i$ in the interior of each $T_i \in \mathcal{T}_h$. We leave it open what values $s$ shall take on edges.

If $\mathcal{T}_h$ is a rectangular partitioning of $D = [a, b] \times [c, d]$, then it is defined by

$$a = t_1 < \ldots < t_{m_1} = b \text{ and } c = \tau_1 < \ldots < \tau_{m_2} = d,$$

as shown in Fig. 2.2. We defined linear B-splines $N_{i,2} \in S^2(t_1, \ldots, t_{m_1})$ and $N_{j,2} \in S^2(\tau_1, \ldots, \tau_{m_2})$ 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

$$S^B_2(\mathcal{T}_h) := \text{span}\{N_{i,j,2}; \; i = 1, \ldots, m_1, \; j = 1, \ldots, m_2\}, \quad (2.12)$$

of **bilinear spline functions** with basis $\{N_{i,j,2}; \; i = 1, \ldots, m_1, \; j = 1, \ldots, m_2\}$. All elements $s \in S^B_2(\mathcal{T}_h)$ belong to $C(D)$, i.e. they are continuous functions.
approximation operator can be defined by

\[ I^B_2 : C(D) \rightarrow \mathcal{S}^B_D, \quad f \mapsto \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} f(t_i, \tau_j) N_{i,j,2}. \]  

(2.13)

\[ I^B_2 \] is in fact an interpolation operator, since \( I^B_2(f)(t_i, \tau_j) = f(t_i, \tau_j) \) for all \( i \) and \( j \).

As in the univariate case, the approximation error \( f - I^B_2(f) \) 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 \( f : \Omega \rightarrow \mathbb{R} \), for which

\[ \|f\|_{H^k(\Omega)} := \left( \sum_{|\alpha| \leq k} \int_{\Omega} |D^\alpha f(x)|^2 \, dx \right)^{1/2} \]  

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

\[ f : \Omega \rightarrow \mathbb{R} \cup \{ \pm \infty \}, \quad (x, y) \mapsto \ln \left( \ln \left( \frac{2}{x^2 + y^2} \right) \right), \]  

(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 $\Omega$ is a bounded polygonal domain, then it can be shown that

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

(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 $I^B_2$ and the space $\mathcal{S}^B_2(\mathcal{T}_h)$. 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 $\kappa$ from (2.10)) such that

$$\|f - I^B_2(f)\|_{L^2(\Omega)} \leq c \cdot h^2 \|f\|_{H^2(\Omega)}.$$  

(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

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

(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 $I_2 : C(D) \rightarrow \mathcal{S}_2(\mathcal{T}_h)$, and an error estimate of the form (2.17) also holds for $I_2$. But one cannot 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 $H^1(\Omega) \rightarrow \mathcal{S}_2(\mathcal{T}_h)$ do exist. The following theorem is proved in [Bra07], p. 83 ff.

**Theorem 2.5 (Clément’s operator)** Let $\Omega \subset \mathbb{R}^2$ 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 $I_h : H^1(\Omega) \rightarrow \mathcal{S}_2(\mathcal{T}_h)$ and a constant $c = c(\kappa)$ such that

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

(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 $\| \cdot \|_{L^2(\Omega)}$, 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 \((X, \| \cdot \|_X)\) and \((Y, (\cdot | \cdot)_Y)\) 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 \( U \subset X \) 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 \( X_n \subset X \):

\[
X_n = \text{span}\{\varphi_1, \ldots, \varphi_{d_n}\} \subset X, \quad \varphi_1, \ldots, \varphi_{d_n} \in X, \text{ linearly independent.} \quad (2.20)
\]

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

\[
\| Tu_n - w \|_Y \leq \| Tv - w \|_Y \quad \text{for all } \ v \in X_n, \quad (2.21)
\]

where \( \| \cdot \|_Y \) is the norm induced by \((\cdot | \cdot)_Y\). 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 \( \| \cdot \|_Y \). 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:

\[
Ax = b, \quad A_{ij} = \langle T \varphi_j | T \varphi_i \rangle_Y, \quad b_i = \langle w | T \varphi_i \rangle_Y, \tag{2.25}
\]

where \( A \in \mathbb{R}^{d_n \times d_n} \) 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 \( \tilde{w} \not\in T(X) \) 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 \( w \) 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 \( b^\delta \). 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

\[ x^\delta := A^{-1}b^\delta, \quad u^\delta_n := \sum_{j=1}^{d_n} x_j^\delta \varphi_j \]  

(2.27)

for the solution of the least squares problem computed for inaccurate data. In practice, \( x^\delta \) 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 \( A_{ij} = (T\varphi_j | T\varphi_i)_Y \) as in (2.24). Set up the vector \( b^\delta \) with components \( b_j^\delta \approx (w | T\varphi_i)_Y \) from available measurement values of \( w \).
- Compute the solution \( x^\delta \) of \( Ax = b^\delta \) and get an approximant

\[ u^\delta_n = \sum_{j=1}^{d_n} x_j^\delta \varphi_j \]

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

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

(2.28)

for \( 0 \leq t \leq T_0 \), where \( \sigma_0 \) is a constant value, \( g \in H^1(\mathbb{R}) \) is a known function with \( g(t) = 0 \) for \( t \leq 0 \) 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

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

(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 \( (\bullet|\bullet)_X = (\bullet|\bullet)_{L_2(0, X_0)} \) and \( (\bullet|\bullet)_Y = (\bullet|\bullet)_{L_2(0, T_0)} \), 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

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

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

\[
    X_{n+1} := \mathcal{S}_2(\tau_0, \ldots, \tau_n)
\]

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 \( 0 \leq t \leq T_0 \), 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**

\[
    g(t) = aG(f_0 t - 1)
\]

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:

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

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

\[
    u^* : [0, \frac{1}{2}] \rightarrow \mathbb{R}, \quad t \mapsto 2t(1 - 2t)
\]
is the exact solution of $Tu = w$. Figure 2.4 shows the Ricker pulse function (to the left) and a noisy function $w$ (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 $w$ values $b_i = \langle w | T\phi_i \rangle_Y$ were computed and an approximation $u_n$ of $u^*$ was constructed as in (2.27). Figure 2.5 shows reconstructions $u_n$ (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 $\| u^* - u_n \|_X$ 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”

\[ R_n : Y \rightarrow X_n, \quad y \mapsto u_n = \sum_{j=1}^{d_n} x_j \varphi_j, \]  

(2.32)

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

\[ Ax = b, \quad A_{i,j} = (T \varphi_j | T \varphi_i)_Y, \quad b_i = (y | T \varphi_i)_Y, \]  

(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 \( \| w - w^\delta \|_Y \leq \delta \) for some \( \delta > 0 \), then

\[ \| u^* - R_n w^\delta \|_X \leq \| R_n \| \delta + \| R_n Tu^* - u^* \|_X. \]  

(2.34)

Now let \( b^\delta \) be a perturbation of \( b \), defined by \( b_i = (w | T \varphi_i)_Y \), such that \( \| b^\delta - b \|_2 \leq \delta \). Let \( u^\delta_n \) be defined by \( b^\delta \) according to (2.27). Then the following error estimates hold

\[ \| u^* - u^\delta_n \|_X \leq \frac{a_n}{\sigma_n} \delta + \| R_n Tu^* - u^* \|_X \]  

(2.35)

\[ \| u^* - u^\delta_n \|_X \leq b_n \| R_n \| \delta + \| R_n Tu^* - u^* \|_X \]  

(2.36)

where

\[ a_n := \max \left\{ \left( \sum_{j=1}^{d_n} \rho_j \varphi_j \right) \left( \sum_{j=1}^{d_n} |\rho_j|^2 = 1 \right) \right\} \]  

(2.37)

\[ b_n := \max \left\{ \left( \sum_{j=1}^{d_n} |\rho_j|^2 \right) \left( \sum_{j=1}^{d_n} \rho_j \cdot T(\varphi_j) \right) = 1 \right\} \]  

(2.38)

and where \( \sigma_n \) 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 perturbed to become \( w^\delta \). For the second error estimates (2.35) and (2.36), nothing is assumed about \( w^\delta \), 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 \( u^* - R_n w^\delta = u^* - R_n Tu^* + R_n w - R_n w^\delta \). From the triangle inequality one also gets \( \| u^* - u^\delta_n \|_X \leq \| u^\delta_n - R_n w \|_X + \| R_n w - u^* \|_X \). Since \( Tu^* = w \), this means that only \( \| u^\delta_n - R_n w \|_X \) 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

$$\|u_n^\delta - R_n w\|_X \leq a_n \|x^\delta - x\|_2 = a_n \|A^{-1}(b^\delta - b)\|_2,$$

$$\leq a_n \|A^{-1}\|_2 \|b^\delta - b\|_2 \leq \frac{a_n}{\sigma_n} \delta,$$

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

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 $\|R_n Tu^* - u^*\|_X$ 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 $\|R_n Tu^* - u^*\|_X$ 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

$$\gamma_n := \max\{\|z_n\|_X; z_n \in X_n, \|Tz_n\|_Y = 1\}.$$  \hspace{1cm} (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

$$\|x - \tilde{x}_n\|_X \to 0 \text{ for } n \to \infty$$  \hspace{1cm} (2.40)

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

$$\min_{z_n \in X_n} \{\|x - z_n\|_X + \gamma_n \|T(x - z_n)\|_Y\} \leq c \|x\|_X \text{ for all } x \in X.$$  \hspace{1cm} (2.41)

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

$$\|R_n Tu - u\|_X \to 0 \text{ for all } u \in X \text{ and for } n \to \infty$$  \hspace{1cm} (2.42)

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

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

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

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

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

and interpret $\gamma_n$ 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:

$$\|R_n\| = \sup\left\{\frac{\|R_n y\|_X}{\|y\|_Y}; y \in Y, y \neq 0\right\} \geq \sup\left\{\frac{\|R_n T x\|_X}{\|T x\|_Y}; x \in X, x \neq 0\right\} \geq \sup\left\{\frac{\|R_n T v\|_X}{\|T v\|_Y}; v \in X_n, v \neq 0\right\} = \gamma_n. \quad (2.44)$$

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

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

This can be seen as follows. If we had $\gamma_n \leq C$ for all $n \in \mathbb{N}$, then $\|v\|_X \leq C \|Tv\|_Y$ for all $v \in X_n$ and all $n$. Because of (2.40) we then would also have $\|x\|_X \leq C \|Tx\|_Y$ for all $x \in X$ and therefore $\|T^{-1}(y)\|_X \leq C \|y\|_Y$ 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 $\|u^* - \bar{u}_n\|_X$ 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 $\mathcal{S}_r(\mathcal{T}_h)$ as in Definition 2.1 for the one-dimensional case or in (2.18).
Fig. 2.7 Total reconstruction error; $T(u^*) = w; \delta, R_n$, and $u_n^\delta$ 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 $\langle w | T \varphi \rangle_Y$ 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

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

(2.46)
is linear, continuous, and injective. Let \( a \leq t_1 < t_2 < \cdots < t_m \leq b \) 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

\[
\langle y|z \rangle_Y := \langle T^{-1}(y)|T^{-1}(z) \rangle_{L_2(a,b)},
\]

(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**

\[
Tu(t_i) = w(t_i), \quad i = 1, \ldots, m,
\]

(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

\[
u_n = \sum_{j=1}^{n} x_j \varphi_j \in X_n, \quad x_j \in \mathbb{R}.
\]

(2.49)

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

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

(2.50)

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

\[
Ax = b, \quad A_{ij} = T(\varphi_j)(t_i), \quad b_i = w(t_i).
\]

(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

\[
\text{minimize } \| b - Ax \|_2, \quad x \in \mathbb{R}^n.
\]

(2.52)
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 $u \in L^2(a,b)$

$$Tu(t_i) = \int_a^b k(t_i,s)u(s)\,ds = \langle k_i|u\rangle_{L^2(a,b)}, \quad k_i(s) := k(t_i, s), \quad i = 1, \ldots, m.$$  

(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

$$\langle k_i|u^*\rangle_{L^2(a,b)} = \langle k_i|u\rangle_{L^2(a,b)} \iff \langle k_i|u^* - u\rangle_{L^2(a,b)} = 0.$$  

(2.54)

This situation is depicted in Fig. 2.8: The solid line symbolizes the linear subspace \{k_1, \ldots, k_m\} 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 \{k_1, \ldots, k_m\}. 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

$$\|\hat{u}\|_{L^2(a,b)} = \min\{\|z\|_{L^2(a,b)}; \quad z \in L^2(a,b) \text{ satisfies } (2.48)\}.$$  

(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

$$\hat{u} = \sum_{j=1}^{m} x_j k_j$$  \hspace{1cm} (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

$$\langle Tk_i | Tu \rangle_Y = \langle Tk_i | w \rangle_Y, \hspace{0.5cm} i = 1, \ldots, m.$$  \hspace{1cm} (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

$$T : X \to Y, \hspace{0.5cm} u \mapsto w, \hspace{0.5cm} w(t) = -\frac{1}{\sigma_0} \int_0^{x_0} g(t - 2s)u(s) \, ds$$  \hspace{1cm} (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

$$w(t_i) = -\frac{1}{\sigma_0} \int_0^{x_0} g(t_i - 2s)u(s) \, ds, \hspace{0.5cm} 0 \leq t_1 < \ldots < t_m \leq T_0.$$  

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 $\tau_j := jh, \hspace{0.5cm} j = 1, \ldots, n$, and work with the $n$-dimensional subspace

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

of $X$, which is spanned by the linear B-splines $N_{j,2}, \hspace{0.5cm} j = 1, \ldots, n$. 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

$$A_{i,j} = -\frac{1}{\sigma_0} \int_0^{x_0} g(t_i - 2s)N_{j,2}(s) \, ds, \quad i = 1, \ldots, m, \ j = 1, \ldots, n. \quad (2.58)$$

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

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

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

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

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

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

to define the spline space $X_n := \mathcal{S}_2(\tau_1, \ldots, \tau_n)$. 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 $w^\delta$ 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 $x^\delta$ of (2.52) with $b$ replaced by $b^\delta$ was determined, which defines the spline function $u_i^\delta(t) = \sum_{j=1}^n x_j^\delta N_{j,2}(t)$ 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.
Analysis of the Method

For an analysis, especially concerning error estimates analogous 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

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

(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

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

which “approximately inverts” \( K \) such that – in a sense to be made precise

\[
SKx \approx x \quad \text{for all } x \in L_2(a, b).
\]

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

\[
x_m(t) := SKx(t) = \sum_{j=1}^m \psi_j(t) \int_a^b k_j(s)x(s) \, ds = \int_a^b \left( \sum_{j=1}^m k_j(s) \psi_j(t) \right)x(s) \, ds.
\]

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 \( \psi_j \). One idea is to require a pointwise approximation of \( x \) by \( SKx \), i.e. to ask for

\[
SKx(t) = x_m(t) \approx x(t).
\]

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 \( a \leq s \leq b \) – which is why \( \varphi \) 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 \( \psi_j, j = 1, \ldots, m \), are functions of \( t \), \( \varphi \) can not be independent of \( t \): the averaging kernel changes with \( t \).
2.4 The Backus-Gilbert Method and the Approximative Inverse

**Description of the Backus-Gilbert Method**

The Backus-Gilbert Method does not determine functions \( \psi_j \), 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 \( \psi_j(\tilde{t}) \) for other parameters \( \tilde{t} \in [a, b] \) in order to make \( SKx(\tilde{t}) \approx x(\tilde{t}) \) 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

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

subject to the linear constraint

\[
\int_a^b \varphi(s, t) \, ds = \int_a^b \sum_{j=1}^m v_j k_j(s) \, ds = 1. \quad (2.65)
\]

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 \( \varphi \), as discussed after (2.62). This minimization problem replaces the vague approximate identity (2.63) and can be written in compact form as

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

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

\[
Q_{ij}(t) = \int_a^b (s - t)^2 k_i(s)k_j(s) \, ds, \quad i, j = 1, \ldots, m, \quad (2.67)
\]

\[
c_i = \int_a^b k_i(s) \, ds, \quad i = 1, \ldots, m.
\]

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 \( v^T Q(t)v + v^T Q(t)v/2 \) 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

\[
v = \frac{Q(t)^{-1} c}{c^T Q(t)^{-1} c}, \quad (2.68)
\]
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
\[
\begin{pmatrix} Q(t) & c \\ c^T & 0 \end{pmatrix} \begin{pmatrix} v \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{m+1}.
\] (2.69)

Having determined \( v \) and knowing \( y := (w(t_1), \ldots, w(t_m)) \), one computes
\[
u_m(t) := \sum_{j=1}^{m} y_j v_j = \int_a^b \varphi(s, t)u(s) \, ds
\] (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
\[
T : X \rightarrow Y, \quad u \mapsto w, \quad w(t) = -\frac{1}{\sigma_0} \int_0^{X_0} g(t - 2s)u(s) \, ds
\] (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),
2.4 The Backus-Gilbert Method and the Approximative Inverse

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 \( t_j = jX_0/n \) 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 \( t = t_j \), 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

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

where \( \gamma > 0 \) is another parameter. The integral

\[
\int_a^b e_\gamma(s, t) u(s) \, ds = \langle u | e_\gamma(\cdot, t) \rangle_{L_2(a,b)},
\]
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 \( \gamma \). It defines a continuous function

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

which can be considered a smoothed approximation of \( u \). For this reason \( e_\gamma \) commonly is called a \textit{mollifier}. Further let us introduce the operator

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

which is called \textit{adjoint operator} with respect to \( K \), since

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

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

\[
K^*v = e_\gamma(\cdot, t), \quad v \in \mathbb{R}^m,
\]

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 \( u_\gamma \) 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

\[
Av = b(t),
\]

where \( A \in \mathbb{R}^{m \times m} \) is a positive definite matrix with components

\[
a_{ij} = \int_{a}^{b} k_i(s)k_j(s) \, ds, \quad i, j = 1, \ldots, m,
\]
independent of $t$, and where $b(t) \in \mathbb{R}^m$ has components

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

Formally, one may define an operator

$$S_{\gamma} : \mathbb{R}^m \rightarrow L_2(a, b), \quad y \mapsto x_{\gamma}, \quad x_{\gamma}(t) = \sum_{i=1}^m y_i v_{\gamma,i}, \quad v_{\gamma} = A^{-1} b(t),$$  \hfill (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 $K u = y$. 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 $e_{\gamma}$, 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 $\tilde{u}_{\gamma}$ (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 $v_{\gamma}$ of (2.75) has to satisfy approximately, if $\tilde{u}_{\gamma}$ is expected to approximate $u_{\gamma}$ 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

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

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

$$k(u, v) := \left( u^2 + v^2 + u_0^2 \right)^{-3/2}, \quad u, v \in \mathbb{R}, \quad u_0 > 0,$$
The continuous function \( w : \mathbb{R}^2 \to \mathbb{R} \) is known for \(-b \leq x_1, x_2 \leq b\) 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 \( u(\mathbf{x}) := 0 \) for \( \mathbf{x} \notin [-a, a]^2 \). 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 \( w \):

\[
w(x) = (k * u)(x) \leftrightarrow \hat{k}(y)\hat{u}(y) \quad \text{for} \quad k(x) \leftrightarrow \hat{k}(y), \ u(x) \leftrightarrow \hat{u}(y).
\]

Thus, under the condition that the Fourier transformed kernel function

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

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

\[
\hat{k}(y) = \frac{2y^2}{\sqrt{\pi f_0^3}}e^{-y^2/f_0^2} - 2\pi iy/f_0 \tag{2.83}
\]

and \emph{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

\[
\hat{k}(y) = \hat{k}(y_1, y_2) = \frac{2\pi}{u_0} e^{-2\pi u_0 \sqrt{y_1^2 + y_2^2}}, \tag{2.84}
\]
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 disastrous effect when only a noisy approximation $\hat{w}$ 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 $\hat{w}/\hat{k}$ 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

$$u \in C(Q) \cap L^2(\mathbb{R}^2) \cap L^1(\mathbb{R}^2), \quad Q = (-\alpha, \alpha)^2, \quad \alpha > 0,$$

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

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

For samples of $u$ on an equidistant grid let us write

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

A bilinear B-spline adapted to the grid points $x_{\alpha}$ is defined by

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

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

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

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:

\[
\hat{u}_N(y_1, y_2) = h^2 \left( \frac{\sin(\pi h y_1)}{\pi h y_1} \right)^2 \left( \frac{\sin(\pi h y_2)}{\pi h y_2} \right)^2 \sum_{\alpha \in W} u_{\alpha} e^{-2\pi i h (y_1 \alpha_1 + y_2 \alpha_2)}. \tag{2.89}\]

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

\[
\hat{u}_N \left( \frac{\beta}{2a} \right) = \sigma_{\beta} \cdot \left( \frac{1}{N} \right)^2 \sum_{\alpha \in W} u_{\alpha} e^{-2\pi i (\alpha_1 \beta_1 + \alpha_2 \beta_2)/N} =: U_{\beta} \tag{2.90}\]

with data independent attenuation factors

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

and values \( U_{\beta} \), 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 \( U_{\beta}, \beta \in W \), from values \( u_{\alpha}, \alpha \in W \), is called (two-dimensional) discrete Fourier transform (DFT). Inversely, the values \( u_{\alpha} \) can be computed from \( U_{\beta} \) by

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

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

\[
\{u_{\alpha}\}_{\alpha \in W} \circlearrowleft \{U_{\beta}\}_{\beta \in W} \quad \text{and} \quad \{U_{\beta}\}_{\beta \in W} \circlearrowright \{u_{\alpha}\}_{\alpha \in W}
\]

to express DFT and IDFT. If \( N \) is a power of two, the two-dimensional DFT and IDFT can both be computed with \( \Theta(N^2 \log(N)) \) 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 \( u \in C(\mathcal{D}) \cap L_2(\mathbb{R}^2) \cap L_1(\mathbb{R}^2) \), 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

\[
U_\beta := \frac{\hat{u}(\beta/2a)}{\sigma_\beta}, \quad \beta \in W.
\]

(2.93)

and by computing

\[
\{U_\beta\}_{\beta \in W} \rightarrow \{u_\alpha\}_{\alpha \in W}, \quad u_N(x) = \sum_{\alpha \in W} u_\alpha \Phi(x/h - \alpha).
\]

(2.94)

This amounts to adjusting \( u_N \) such that the Fourier interpolation conditions

\[
\hat{u}_N \left( \frac{\beta}{2a} \right) = \hat{u} \left( \frac{\beta}{2a} \right), \quad \beta \in W.
\]

(2.95)

hold. We remark that because of \( \sin(\lambda)/\lambda = 1/\Lambda - 2/\Lambda^2 \) for \( |\lambda| \leq \pi/2 \), 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** \( w = k \ast u \).

- Compute via DFT:

\[
\{w_\alpha\}_{\alpha \in W} \rightarrow \{W_\beta\}_{\beta \in W} \quad \text{where} \quad w_\alpha := w(x_\alpha), \quad \alpha \in W.
\]

(2.96)

- Compute

\[
U_\beta := \frac{W_\beta}{k(\beta/2a)}, \quad \beta \in W.
\]

(2.97)

- Compute via IDFT:

\[
\{U_\beta\}_{\beta \in W} \rightarrow \{u_\alpha\}_{\alpha \in W}
\]

(2.98)

and take

\[
u_N(x) = \sum_{\alpha \in W} u_\alpha \Phi(x/h - \alpha)
\]

(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

\[
    w_N(x) = \sum_{\alpha \in W} w_\alpha \Phi(x/h - \alpha), \quad w_\alpha = w(x_\alpha), \quad \alpha \in W, \tag{2.100}
\]
as an approximant of \( w \) and make use of its Fourier transform values

\[
    \hat{w}_N \left( \frac{\beta}{2a} \right) = \sigma_\beta W_\beta \approx \hat{w} \left( \frac{\beta}{2a} \right). \tag{2.101}
\]

The idea behind the last two steps is to determine the coefficients \( u_\alpha \) of an approximant

\[
    u_N(x) = \sum_{\alpha} u_\alpha \Phi(x/h - \alpha) \approx u(x)
\]
such that

\[
    \hat{k} \left( \frac{\beta}{2a} \right) \hat{u} \left( \frac{\beta}{2a} \right) \approx \hat{k} \left( \frac{\beta}{2a} \right) \hat{u}_N \left( \frac{\beta}{2a} \right) = \hat{k} \left( \frac{\beta}{2a} \right) \sigma_\beta U_\beta = \sigma_\beta W_\beta =
\]

\[
    \approx \hat{w}_N \left( \frac{\beta}{2a} \right) = \hat{w} \left( \frac{\beta}{2a} \right), \quad \beta \in W.
\]

**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 $1/k(\beta/2a)$ 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 $\hat{w}_N$ 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 $W_\beta$ 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

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

(2.103)
where the kernel function is defined by
\[
k(x_1, x_2, y_1, y_2, z) := \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + z^2}}
\]  
(2.104)
for \((x_1, x_2), (y_1, y_2) \in \mathbb{R}^2\) and \(z > 0\) and where

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

\[
x_\alpha := h\alpha, \quad h := \frac{a}{n}, \quad \alpha \in G_n := \{(\alpha_1, \alpha_2) \in \mathbb{Z}^2; \ -n \leq \alpha_j \leq n, \ j = 1, 2\},
\]
for \(n \in \mathbb{N}\) to be chosen. As an approximant of the unknown function \(u^*\) we use

\[
u^*(x) \approx u_n(x) = \sum_{\alpha \in G_n} c_\alpha \Phi(x/h - \alpha)
\]
(2.106)
where the parameters \(c_\alpha\) are yet unknown and where the bilinear B-spline

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

\[
c_\alpha \in [b_1, b_2], \quad \alpha \in G_n,
\]
(2.108)
for the parameters \(c_\alpha\). 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 \(u_n\):

\[
w(x) \approx w_n(x) := \sum_{\alpha \in G_n} \omega_\alpha k(x, x_\alpha, c_\alpha)
\]
(2.109)
with constant weights

\[
\omega_\alpha := \begin{cases} 
  h^2/4, & \text{if } |\alpha_1| = |\alpha_2| = n, \\
  h^2/2, & \text{if } |\alpha_1| = n \text{ or } |\alpha_2| = n, \ |\alpha_1| + |\alpha_2| < 2n, \\
  h^2, & \text{else}.
\end{cases}
\]
(2.110)
2.6 Discretization of Nonlinear Model Problems

To arrive at (2.109), one makes use of the identity $u_n(x) = c_\alpha$, $\alpha \in G_n$, 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 cannot 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

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

(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

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

(2.112)

consisting of

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

(2.113)

Summarizing observations and unknowns as vectors

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

(2.114)

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

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

by setting

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

(2.115)

this system can be written in condensed form as

$$F(c) = y, \quad c \in C.$$

(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

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

(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

\[ \mathcal{S} = \{ \sigma \in H^1(0, X_0); \; 0 < \sigma_- \leq \sigma(x) \leq \sigma_+ < \infty \; \text{for} \; 0 \leq x \leq X_0 \}, \]

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

\[ T(\sigma) = Y_d, \quad \sigma \in \mathcal{S}. \]

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

\[ \Delta := X_0/(n + 1) \quad \text{and} \quad x_k := k\Delta, \quad k = 0, \ldots, n + 1, \]  

(2.118)

and define

\[ \mathcal{S}_n := \{ \sigma : [0, X_0] \to \mathbb{R}; \; \sigma_- \leq \sigma(x) \leq \sigma_+ \; \text{for} \; 0 \leq x \leq X_0 \; \text{and} \; \sigma(x) \equiv \text{const. for} \; x_k \leq x < x_{k+1}, \; k = 0, \ldots, n \} \]  

(2.119)

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

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

(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 \( \sigma \in \mathcal{S}_n \) 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 \( \sigma_e \in \mathcal{S} \) at \( x_0, \ldots, x_n \). Any \( \sigma_e \in \mathcal{S} \) can be approximated arbitrarily well (with respect to the norm \( \| \bullet \|_{L^2(0,x_0)} \)) by its interpolant \( \sigma \in \mathcal{S}_n \), if \( n \) is chosen large enough (compare Theorem 2.2). However,

\[
\mathcal{S}_n \subsetneq \mathcal{S}.
\]

In fact, a function \( \sigma \in \mathcal{S}_n \) 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

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

(1.42)

to hold literally. To give this equation a meaning for \( \sigma \in \mathcal{S}_n \), 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 \( \sigma \in \mathcal{S}_n \) and for \((x,t) \in D_k\), (1.42) simplifies to

\[
\frac{\partial^2 y(x,t)}{\partial t^2} - \frac{\partial^2 y(x,t)}{\partial x^2} = 0, \quad (x,t) \in D_k,
\]

(2.121)

with general solution

\[
y(x,t) = F(x - t) + G(x + t)
\]

(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

\[
\lim_{x \to x_k^+} \frac{\partial y(x,t)}{\partial t} = \lim_{x \to x_k^-} \frac{\partial y(x,t)}{\partial t}
\]

(2.123)

and

\[
\lim_{x \to x_k^+} \frac{\sigma(x) \partial y(x,t)}{\partial x} = \lim_{x \to x_k^-} \frac{\sigma(x) \partial y(x,t)}{\partial x}
\]

(2.124)
hold for \( k = 1, \ldots, n \). Equation (2.123) means that there are no shifts in particle velocity across 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 \( \sigma \in \mathcal{S}_n \) 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

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

(2.125)

is explicitly given by

\[
T_g(t) = \frac{1}{\sigma_1} \sum_{k=0}^{n} \lambda_k g(t - t_k)
\]

(2.126)

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

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

(2.127)

Then compute state variables \( u_{ij} \) and \( v_{ij} \) by the linear recursions

\[
u_{ij} = (1 - r_{i-j})u_{i,j-1} + r_{i-j}v_{i-1,j} \quad i = 1, \ldots, n, \quad j = 0, \ldots, i
\]

(2.128)

with boundary and initial conditions

\[
u_{i,-1} = 0 \quad u_{0,0} = 0 \quad v_{i-1,i} = u_{i,i} \quad v_{0,0} = 2.
\]

(2.129)

Finally set

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

(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 $\sigma$ 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

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

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

$$u_k : D_k \to \mathbb{R}, \ u_k(x, t) = \frac{\partial y_k(x, t)}{\partial t} + \frac{\partial y_k(x, t)}{\partial x},$$

$$v_k : D_k \to \mathbb{R}, \ v_k(x, t) = \frac{\partial y_k(x, t)}{\partial t} - \frac{\partial y_k(x, t)}{\partial x}, \quad (2.131)$$

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

$$\frac{d}{dt} u_k(x(t), t) = 0 \quad \text{for} \quad \frac{dx(t)}{dt} = -1,$$

$$\frac{d}{dt} v_k(x(t), t) = 0 \quad \text{for} \quad \frac{dx(t)}{dt} = +1, \quad (2.132)$$

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

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

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

$$y_k(x, t) = F(x - t) + G(x + t), \quad (2.134)$$

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 $\frac{\partial y_1(0, t)}{\partial x} = F'(-t) = -g(t)/\sigma_1$, which determines $F'$. We also can deduce $\frac{\partial y_1(0, t)}{\partial t} = -F'(-t) = g(t)/\sigma_1$ and thus we get from (2.131)

$$v_1(0, t) = \frac{2}{\sigma_1} g(t) \quad \text{and} \quad u_1(0, t) = 0, \quad 0 \leq t < \Delta. \quad (2.135)$$
The surface excitation propagates down at velocity 1, which means that \( v_1(x, t) = 2g(t - x) / \sigma_1 \) for \( 0 \leq t < \Delta \) and \( 0 \leq x < \Delta \). 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 across the interface \( x = x_k := k\Delta \) of layers \( k \) and \( k + 1 \) as

\[
\frac{u_k(x_k, t) + v_k(x_k, t)}{2} = \frac{u_{k+1}(x_k, t) + v_{k+1}(x_k, t)}{2} \quad \text{and} \quad \frac{u_k(x_k, t) - v_k(x_k, t)}{2} = \frac{u_{k+1}(x_k, t) - v_{k+1}(x_k, t)}{2} \tag{2.136}
\]

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

\[
u_k(x_k, t) = (1 - r_k)u_{k+1}(x_k, t) + r_k v_k(x_k, t) \]
\[
v_{k+1}(x_k, t) = -r_k u_{k+1}(x_k, t) + (1 + r_k) v_k(x_k, t). \tag{2.137}
\]

**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 excitation. 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 \( \Delta \)”, 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 \( v_{i-1,j} = u_{i,j} \) in (2.129). Thus \( r_0 = 0 \) is not a reflection coefficient.
(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

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

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 $0 \leq t < t_1 = 2\Delta$, 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 $\lambda_k$ as defined in (2.130).

The first factors $\lambda_k$ read

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

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 $\lambda_k$ – this will be discussed in a moment. Once the values $\lambda_k$ are known, and if we additionally know $\sigma_1$, we can determine the other impedance values.
Theorem 2.15  The mapping
\[ G : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad (r_1, \ldots, r_n) \mapsto (\lambda_1, \ldots, \lambda_n), \]  
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,
\[ H : [\sigma_-, \sigma_+]^n \rightarrow \mathbb{R}^n, \quad (\sigma_2, \ldots, \sigma_{n+1}) \mapsto (\lambda_1, \ldots, \lambda_n), \]
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
\[ \lambda_i = 2r_i \prod_{k=1}^{i-1} (1 - r_k^2) + P_i(r_1, \ldots, r_{i-1}), \quad i = 1, \ldots, n, \]  
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
\[ g_k := g(t_k), \quad t_k := 2k\Delta, \quad k = 0, \ldots, n + 1 \]  
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
\[ Y_k := Y_d(t_k) = \frac{1}{\sigma_1} \sum_{\ell=0}^{n} \lambda_{\ell} g_{k-\ell}, \quad k = 0, \ldots, n + 1, \]  
where \( g_{k-\ell} := 0 \) for \( \ell > k \). By a proper choice of the time origin we can achieve
\[ g_0 = 0 \quad \text{and} \quad g_1 \neq 0 \]  
to hold. Then, the first members of the sequence \( (Y_k)_k \) read
\[ Y_0 = 0, \quad Y_1 = g_1/\sigma_1, \quad Y_2 = (g_2 + \lambda_1 g_1)/\sigma_1, \quad Y_3 = (g_3 + \lambda_1 g_2 + \lambda_2 g_1)/\sigma_1 \]  
\[ \vdots \]
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

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

(2.146)
defined by

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

with \( \sigma_1, g_1, \ldots, g_{n+1} \) 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 \( \lambda_k \) 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.
Inverse Problems
Basics, Theory and Applications in Geophysics
Richter, M.
2016, XII, 240 p. 52 illus., 32 illus. in color., Softcover
ISBN: 978-3-319-48383-2
A product of Birkhäuser Basel