Ill-posedness of linear problems

Inverse problems typically involve the estimation of certain quantities based on indirect measurements of these quantities. The inversion process is often ill-posed in the sense that noise in the data gives rise to significant errors in the estimate. In this chapter we introduce the concept of ill-posedness and analyze the solvability and ill-posedness of linear discrete equations. Our analysis is focused on a classical example in atmospheric remote sensing, namely the temperature retrieval by nadir sounding. In a continuous setting, this retrieval problem is modeled by a Fredholm integral equation of the first kind, which is the prototype of ill-posed problems.

2.1 An illustrative example

To explain the difficulties associated with the solution of linear inverse problems, we consider measuring a temperature profile by nadir sounding. Spaceborne remote sensing of atmospheric temperature uses absorption features of gases with well-known and constant mixing ratios as for instance, the CO$_2$ bands at 15 and 4.3 $\mu$m in the infrared or the O$_2$ lines at 60 GHz in the microwave regime. In a plane-parallel atmosphere and under the assumption that the background contribution from the surface can be neglected, the diffuse radiance at the top of the atmosphere $z = z_{\text{max}}$ and in a direction with zero scan angle is given by the Schwarzschild equation

$$I (\nu) = \int_0^{z_{\text{max}}} B (\nu, T (z)) \frac{\partial T}{\partial z} (\nu, z) \, dz.$$  

In the microwave region, the Rayleigh–Jeans approximation allows the following representation of the Planck function

$$B (\nu, T (z)) = 2ck_B \nu^2 T (z).$$

As a result and when neglecting the temperature dependence of the transmission, we obtain

$$I (\nu) = \int_0^{z_{\text{max}}} k (\nu, z) T (z) \, dz,$$

(2.1)
with
\[ k(\nu, z) = 2c_k \nu^2 \frac{\partial \mathcal{T}}{\partial z}(\nu, z) \] (2.2)
being the kernel function.

Equation (2.1), which we rewrite in the generic form \( y = I \) and \( x = T \)
\[ y(\nu) = \int_0^{z_{\text{max}}} k(\nu, z) x(z) \, dz, \quad \nu \in [\nu_{\text{min}}, \nu_{\text{max}}], \] (2.3)
is a Fredholm integral equation of the first kind and represents the mathematical model of a continuous problem. To formulate our problem in a general setting, we consider the space of real-valued, square integrable functions on the interval \([a, b]\), denoted by \( L^2([a, b]) \). Actually, \( L^2([a, b]) \) is a Hilbert space under the inner product \( \langle u, v \rangle = \int_a^b u(t) v(t) \, dt \) and the induced norm \( \|u\| = \sqrt{\int_a^b u(t)^2 \, dt} \). Assuming that \( y \) belongs to the Hilbert space \( Y = L^2([\nu_{\text{min}}, \nu_{\text{max}}]) \) and \( x \) belongs to the Hilbert space \( X = L^2([0, z_{\text{max}}]) \), we introduce the linear operator \( K : X \to Y \) by the relation
\[ Kx = \int_0^{z_{\text{max}}} k(\cdot, z) x(z) \, dz. \]
The integral equation (2.3) can then be written as
\[ Kx = y, \] (2.4)
and since (cf. (2.2)) \( k \in L^2([\nu_{\text{min}}, \nu_{\text{max}}] \times [0, z_{\text{max}}]) \), the linear operator \( K \) is bounded and compact (the image of any bounded sequence of functions in \( L^2 \) has at least one converging subsequence).

A spectral instrument cannot measure a continuous signal and the data is a collection of discrete observations. More specifically, the radiances
\[ y(\nu_i) = I(\nu_i), \]
with \( \{\nu_i\}_{i=1}^m \) being an equidistant set of points in the spectral interval \([\nu_{\text{min}}, \nu_{\text{max}}]\), represent the data, and the equation to be solved takes the form
\[ y(\nu_i) = \int_0^{z_{\text{max}}} k(\nu_i, z) x(z) \, dz, \quad i = 1, \ldots, m. \] (2.5)
The semi-discrete equation (2.5) is a mathematical model for discrete observations of a physical process and can be expressed in compact form as
\[ K_m x = y_m. \] (2.6)
The data \( y_m \) is a vector with entries \( [y_m]_i = y(\nu_i), \quad i = 1, \ldots, m \), and \( K_m \) is a linear operator acting between the Hilbert space \( X \) and the finite-dimensional Euclidean space \( \mathbb{R}^m \),
\[ [K_m x]_i = (Kx)(\nu_i) = \int_0^{z_{\text{max}}} k(\nu_i, z) x(z) \, dz. \]
The discretization approach which transforms the continuous equation (2.3) into the semi-discrete equation (2.5) is known as the collocation method. It should be pointed out that the collocation method can be regarded as a projection method with delta functions as basis functions.

For a complete discretization, we consider the subspace $X_n \subset X$ with the (not necessarily orthonormal) basis $\{\Phi_{nj}\}_{j=1}^n$ and define the approximation or the interpolant $x_n \in X_n$ of $x$ as the solution of the equation

$$K_m x_n = y_m. \quad (2.7)$$

Representing $x_n$ as a linear combination of basis functions,

$$x_n = \sum_{j=1}^n \xi_j \Phi_{nj},$$

we obtain the system of equations

$$y(\nu_i) = \sum_{j=1}^n \left[ \int_0^{z_{\text{max}}} k(\nu_i, z) \Phi_{nj}(z) \, dz \right] \xi_j, \quad i = 1, \ldots, m. \quad (2.8)$$

In matrix form, (2.8) can be written as

$$K_{mn} x_n = y_m, \quad (2.9)$$

where $x_n = [\xi_1, \ldots, \xi_n]^T$ is the coordinate vector and the matrix $K_{mn}$, with entries

$$[K_{mn}]_{ij} = [K_m \Phi_{nj}]_i = (K \Phi_{nj})(\nu_i) = \int_0^{z_{\text{max}}} k(\nu_i, z) \Phi_{nj}(z) \, dz,$$

is a linear map between the finite-dimensional Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^m$. The discretization approach which transforms the continuous equation (2.3) into the discrete equation (2.8) is called a projection method.

Let us make some comments on the choice of the set of basis functions $\{\Phi_{nj}\}$ for representing the state parameter $x$.

1. If $\{z_j\}_{j=0,n}$ is a discretization grid of the altitude interval $[0, z_{\text{max}}]$ with $z_0 = 0$ and $z_n = z_{\text{max}}$, we may choose the piecewise constant functions

$$P_{nj}(z) = \begin{cases} 1, & z_{j-1} \leq z < z_j, \\ 0, & \text{otherwise}, \end{cases} \quad j = 1, \ldots, n$$

as basis functions. Using the orthogonality relations $\langle P_{ni}, P_{nj} \rangle = 0$, $i \neq j$, and $\|P_{nj}\|^2 = z_j - z_{j-1}$, we obtain

$$\xi_j = \frac{1}{z_j - z_{j-1}} \langle x_n, P_{nj} \rangle = \frac{1}{z_j - z_{j-1}} \int_{z_{j-1}}^{z_j} x_n(z) \, dz$$

for $j = 1, \ldots, n$. Thus, the entries of the coordinate vector are the mean values of the atmospheric profile over each altitude interval (layer).
(2) For the discretization grid \( \{ z_j \}_{j=0}^{n+1} \) with \( z_0 = z_1 = 0 \) and \( z_n = z_{n+1} = z_{\text{max}} \), the piecewise linear functions (or hat functions),

\[ H_{nj}(z) = \begin{cases} 
\frac{z - z_{j-1}}{(z_j - z_{j-1})}, & z_{j-1} < z \leq z_j, \\
\frac{(z_{j+1} - z)}{(z_j + 1 - z_j)}, & z_j \leq z < z_{j+1}, \\
0, & \text{otherwise},
\end{cases} \]

for \( j = 1, \ldots, n \), can also be chosen as basis functions. Since

\[ H_{nj}(z_i) = \begin{cases} 
1, & i = j, \\
0, & i \neq j,
\end{cases} \]

for \( i, j = 1, \ldots, n \), it follows that \( \xi_j = x_n(z_j) \) for \( j = 1, \ldots, n \), and we conclude that the entries of the coordinate vector are the values of the atmospheric profile at each grid point (level).

(3) For a smoother and more accurate approximation, we have to use a piecewise polynomial approximation with higher-order pieces than broken lines. The most popular choice is the B-spline interpolation (de Boor, 2001). In this case, for the discretization grid \( \{ z_j \}_{j=1}^{n} \) with \( z_1 = 0 \) and \( z_n = z_{\text{max}} \), \( x_n \) is expressed as

\[ x_n(z) = \sum_{j=1}^{n} \xi_j B_{nkj}(z), \]

where \( B_{nkj} \) are the B-splines of order \( k \). Note that \( B_{nkj} \) is a piecewise polynomial of degree of at most \( k - 1 \), and that the \( B_{nkj}, j = 1, \ldots, n \), are locally linear independent. A well-conditioned basis of B-splines can be obtained with the recursion formulas

\[ B_{n1j}(z) = \begin{cases} 
1, & z_j \leq z < z_{j+1} \\
0, & \text{otherwise},
\end{cases} \]

\[ B_{nkj}(z) = \frac{z - t_j}{t_{j+k-1} - t_j} B_{nk-1j}(z) + \frac{t_{j+k} - z}{t_{j+k} - t_{j+1}} B_{nk-1j+1}(z), \quad k \geq 2, \]

where

\[ t_1 \leq t_2 \leq \ldots \leq t_{n+k} \]

are the knots at which the polynomials are tied together by the continuity conditions. In many problems, where extrapolation beyond \( z = 0 \) and \( z = z_{\text{max}} \) is not anticipated, it is a common practice to set

\[ t_1 = t_2 = \ldots = t_k = 0 \]

and

\[ t_{n+1} = t_{n+2} = \ldots = t_{n+k} = z_{\text{max}}. \]

The second-order B-spline \( B_{n2j} \) coincides with the hat functions, and for this reason, \( B_{n2j} \) is also called a linear B-spline.
2.2 Concept of ill-posedness

The mathematical formulation of inverse problems leads to equations that typically are ill-posed. According to Hadamard, the equation

\[ Kx = y, \quad (2.10) \]

with \( K \) being a linear operator acting from the Hilbert space \( X \) into the Hilbert space \( Y \), is called well-posed provided (Engl et al., 2000; Rieder, 2003; Vogel, 2002)

1. for any \( y \in Y \), a solution \( x \) exists;
2. the solution \( x \) is unique;
3. the solution is stable with respect to perturbations in \( y \), in the sense that if \( Kx_0 = y_0 \) and \( Kx = y \), then \( x \to x_0 \) whenever \( y \to y_0 \).

Equivalently, equation (2.10) is called well-posed if the operator \( K \) is bijective and the inverse operator \( K^{-1} \) is continuous. As equation (2.10) is a mathematical model of a continuous problem, the term ‘well-posed’ is also used when referring to the underlying problem. If one of Hadamard’s conditions is violated, the problem is called ill-posed.

Denoting by

\[ \mathcal{R}(K) = \{ Kx/x \in X \} \]

the range space of \( K \) and by

\[ \mathcal{N}(K) = \{ x \in X/Kx = 0 \} \]

the null space of \( K \), it is apparent that (Kress, 1999)

1. if \( K \) is not surjective (\( \mathcal{R}(K) \neq Y \)), then equation (2.10) is not solvable for all \( y \in Y \) (non-existence);
2. if \( K \) is not injective (\( \mathcal{N}(K) \neq \varnothing \)), then equation (2.10) may have more than one solution (non-uniqueness);
3. if \( K^{-1} \) exists but is not continuous, then the solution \( x \) of equation (2.10) does not depend continuously on the data \( y \) (instability).

Non-existence can occur in practice because the forward model is approximate or because the data contains noise. Non-uniqueness is introduced by the need for discretization and is a peculiarity of the so-called rank deficient problems, characterized by a matrix \( K_{mn} \) with a non-trivial null space. In particular, state vectors \( x_0 \) that lie in the null space of \( K_{mn} \) solve the equation \( K_{mn}x_0 = 0 \), and by superposition, any linear combination of these null-space solutions can be added to a particular solution and does not change the fit to the data. Violation of the third Hadamard condition creates serious numerical problems because small errors in the data space can be dramatically amplified in the state space.

When a continuous ill-posed problem is discretized, the underlying discrete problem inherits this ill-posedness and we say that we are dealing with ‘a discrete ill-posed problem’ (Hanke and Hansen, 1993). The ill-posedness of a discrete linear problem, written in the form of a linear system of equations, is reflected by a huge condition number of the coefficient matrix. In this regard, the term ‘ill-conditioned system of equations’ is also used to describe instability. To stabilize the inversion process we may impose additional constraints that bias the solution, a process that is generally referred to as regularization.
2.3 Analysis of linear discrete equations

The Fredholm integral equation $Kx = y$ is severely ill-posed, when $K$ is a compact operator with an infinite-dimensional range space (Engl et al., 2000). This means that the inverse operator $K^{-1}$ is unbounded and that the third Hadamard condition is not fulfilled. An analysis of continuous ill-posed problems regarding their solvability and ill-posedness is given in Appendix A; here we pay attention to the discrete case.

From a strictly mathematical point of view, the discrete equation (2.9), written in the more familiar form as

$$Kx = y,$$  \hspace{1cm} (2.11)

is well-posed, as any nonsingular matrix automatically has a continuous inverse. However, in terms of condition numbers, the fact that a continuous problem is ill-posed means that the condition number of its finite-dimensional approximation grows with the quality of the approximation (Hanke and Hansen, 1993). Increasing the degree of discretization, i.e., increasing the approximation accuracy of the operator, will cause a huge condition number of the matrix and a dramatic amplification of rounding errors. As a result, the approximate solution becomes less and less reliable.

2.3.1 Singular value decomposition

In order to demonstrate the ill-posed nature of the discrete equation (2.11) we first introduce the concept of singular value decomposition of a matrix.

For an $m \times n$ matrix $K$, the matrix $K^T K$ is symmetric and positive semidefinite, and as a result, the eigenvalues of $K^T K$ are real and non-negative. The non-negative square roots of these eigenvalues are called the singular values of $K$. If rank $(K) = r$, the matrix $K$ has exactly $r$ positive singular values counted according to their geometric multiplicity. To simplify our exposition we suppose that these singular values are simple and throughout this book, the claim rank $(A) = r$ tacitly assumes that the matrix $A$ has exactly $r$ positive and distinct singular values. Note that a symmetric matrix $A$ is said to be positive definite if $x^T Ax > 0$ for all $x \neq 0$, and positive semidefinite if $x^T Ax \geq 0$. All eigenvalues of a symmetric and positive definite matrix are positive real numbers. Also note that the rank of a matrix $A$ is the maximal number of linearly independent column vectors of $A$ (column rank), or the maximal number of linearly independent row vectors of $A$ (row rank).

If $K$ is of rank $r$, and $\{\sigma_i\}_{i=1}^{r,n}$ denotes the set of singular values appearing in decreasing order, $\sigma_1 > \sigma_2 > \ldots > \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$, then there exist the orthonormal sets $\{v_i\}_{i=1}^{r,n} \subseteq \mathbb{R}^n$ and $\{u_i\}_{i=1}^{r,m} \subseteq \mathbb{R}^m$ such that

$$K v_i = \sigma_i u_i, \hspace{0.5cm} K^T u_i = \sigma_i v_i, \hspace{0.5cm} i = 1, \ldots, r,$$  \hspace{1cm} (2.12)

and

$$K v_i = 0, \hspace{0.5cm} i = r+1, \ldots, n, \hspace{0.5cm} K^T u_i = 0, \hspace{0.5cm} i = r+1, \ldots, m.$$  \hspace{1cm} (2.13)

Each system $(\sigma_i; v_i, u_i)$ with these properties is called a singular system of $K$. The sets $\{v_i\}_{i=1}^{r}$ and $\{v_i\}_{i=r+1}^{n}$ are orthonormal bases of $\mathcal{N}(K)^\perp$ and $\mathcal{N}(K)$, respectively,
Analysis of linear discrete equations

Sect. 2.3

\[ \mathcal{N}(K)^\perp = \text{span} \{ v_i \}_{i=1}^r, \quad \mathcal{N}(K) = \text{span} \{ v_i \}_{i=r+1}^n, \]  
while \{ u_i \}_{i=1}^r \text{ and } \{ u_i \}_{i=r+1}^m \text{ are orthonormal bases of } \mathcal{R}(K) \text{ and } \mathcal{R}(K)^\perp, \text{ respectively.} \]

\[ \mathcal{R}(K) = \text{span} \{ u_i \}_{i=1}^r, \quad \mathcal{R}(K)^\perp = \text{span} \{ u_i \}_{i=r+1}^m. \]

In (2.14) and (2.15), \( \mathcal{N}(K)^\perp \) and \( \mathcal{R}(K)^\perp \) are the orthogonal complements of \( \mathcal{N}(K) \) and \( \mathcal{R}(K) \), respectively, and we have the representations \( \mathbb{R}^n = \mathcal{N}(K) \oplus \mathcal{N}(K)^\perp \) and \( \mathbb{R}^m = \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp \), where the notation ‘\( \oplus \)’ stands for the direct sum of two sets, \( \mathcal{A} \oplus \mathcal{B} = \{ x + y / x \in \mathcal{A}, y \in \mathcal{B} \} \). The condition number of the matrix \( K \) is defined as the ratio of the largest to the smallest singular value, that is, \( \kappa(K) = \sigma_1/\sigma_r \).

Equations (2.12)–(2.13) can be written in matrix form as

\[ K = U \Sigma V^T, \]
where \( U = [u_1, \ldots, u_m] \) and \( V = [v_1, \ldots, v_n] \) are orthogonal (or orthonormal) \( m \times m \) and \( n \times n \) matrices, respectively, and \( \Sigma \) is an \( m \times n \) matrix of the form

\[ \Sigma = \begin{bmatrix} \text{diag} (\sigma_i)_{r \times r} & 0 \\ 0 & 0 \end{bmatrix}, \]
with \( \text{diag} (\sigma_i)_{r \times r} \) being an \( r \times r \) diagonal matrix. The representation (2.16) is called the singular value decomposition (SVD) of the matrix \( K \).

A positive definite matrix \( A \) is nonsingular and its singular value decomposition coincides with its spectral decomposition, that is, \( A = V \Sigma V^T \). Throughout this book the discussion of positive definite matrices is restricted to symmetric matrices, although a general real matrix is positive definite if and only if its symmetric part \( (1/2)(A + A^T) \) is positive definite, or equivalently, if and only if its symmetric part has all positive eigenvalues. Hence, when we say that \( A \) is positive definite, in fact, we mean that \( A \) is symmetric and positive definite. Positive definite matrices are important in statistics essentially because the covariance matrix of a random vector is always positive definite, and conversely, any positive definite matrix is the covariance matrix of some random vector (in fact, of infinitely many). For \( A = V \Sigma V^T \), the square root of \( A \) is taken as \( A^{1/2} = V \Sigma^{1/2} V^T \), and evidently \( A^{1/2} \) is symmetric. However, a positive definite matrix has many non-symmetric square roots, among which the one obtained by the Cholesky factorization \( A = L^T L \), where \( L \) is upper triangular, is of particular interest.

2.3.2 Solvability and ill-posedness

Let \( K \) be an \( m \times n \) matrix of rank \( n \) with the singular system \( \{ (\sigma_i; v_i, u_i) \} \). The assumption \( \text{rank}(K) = n \) yields \( \mathcal{N}(K) = \emptyset \), which, in turn, implies that the linear operator \( K \) is injective.

The solvability of equation (2.11) is stated by the following result: the linear equation (2.11) is solvable if and only if \( y \in \mathcal{R}(K) \), and the unique solution is given by

\[ x^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (u_i^T y) v_i. \]
To prove this result we first assume that $x^\dagger$ is a solution of (2.11), i.e., $Kx^\dagger = y$. If $y_0 \in \mathcal{N}(K^T)$, we see that

$$y^T y_0 = x^\dagger T K^T y_0 = 0,$$

and since $\mathcal{R}(K) = \mathcal{N}(K^T)^\perp$, the necessity of condition $y \in \mathcal{R}(K)$ follows. Conversely, let $y \in \mathcal{R}(K)$. Then, $y$ can be expressed in terms of the orthonormal basis $\{u_i\}_{i=1}^n$ of $\mathcal{R}(K)$ as follows:

$$y = \sum_{i=1}^n (u_i^T y) u_i.$$

For $x^\dagger$ defined by (2.17), we have (cf. (2.12))

$$Kx^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (u_i^T y) K v_i = \sum_{i=1}^n (u_i^T y) u_i,$$

and we deduce that $Kx^\dagger = y$. Finally, the uniqueness of $x^\dagger$ follows from the injectivity of $K$.

Equation (2.17) defines a linear operator $K^\dagger : \mathbb{R}^m \to \mathbb{R}^n$ by the relation

$$K^\dagger y = \sum_{i=1}^n \frac{1}{\sigma_i} (u_i^T y) v_i, \quad y \in \mathbb{R}^m,$$

which also allows a representation by an $n \times m$ matrix,

$$K^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} v_i u_i^T.$$n

This operator or matrix, which maps $y \in \mathcal{R}(K)$ into the solution $x^\dagger$ of equation (2.11), that is,

$$x^\dagger = K^\dagger y,$$

is called the generalized inverse. By convention, the data vector $y$ which belongs to the range space of $K$, will be referred to as the exact data vector, and $x^\dagger = K^\dagger y$ will be called the exact solution.

In practice, the exact data is not known precisely and only the noisy data is available. The noisy data vector $y^\delta$ is a perturbation of the exact data vector $y$, and we have the representation

$$y^\delta = y + \delta,$$

where $\delta$ is the instrumental noise. In general, $y^\delta \in \mathbb{R}^m$, and there is no guarantee that $y^\delta \in \mathcal{R}(K)$. As a result, the linear equation is not solvable for arbitrary noisy data and we need another concept of solution, namely the least squares solution. For the noisy data vector

$$y^\delta = \sum_{i=1}^m (u_i^T y^\delta) u_i,$$
the least squares solution of the linear equation (2.11) is defined by

$$x^\delta = \sum_{i=1}^{n} \frac{1}{\sigma_i} \left( u_i^T y^\delta \right) v_i.$$  
(2.20)

The least squares solution can be characterized as follows:

1. the image of $x^\delta$ under $K$ is the projection of $y^\delta$ onto $\mathcal{R}(K)$, that is,

$$Kx^\delta = P_{\mathcal{R}(K)}y^\delta = \sum_{i=1}^{n} \left( u_i^T y^\delta \right) u_i;$$

2. $x^\delta$ has the optimality property

$$x^\delta = \arg \min_x \|y^\delta - Kx\|;$$

3. $x^\delta$ solves the normal equation

$$K^T K x = K^T y^\delta.$$  

Assertion (1) follows from (2.20) in conjunction with (2.12). Considering (2), we see that

$$\|y^\delta - Kx^\delta\| = \|y^\delta - P_{\mathcal{R}(K)}y^\delta\| = \min_{y \in \mathcal{R}(K)} \|y^\delta - y\| = \min_x \|y^\delta - Kx\|$$

and the conclusion is apparent. For proving (3), we use (2.19) and (2.20) to obtain

$$y^\delta - Kx^\delta = \sum_{i=n+1}^{m} (u_i^T y^\delta) u_i.$$ 

Thus, $y^\delta - Kx^\delta \in \mathcal{R}(K)^\perp = \mathcal{N}(K^T)$; this gives $K^T (y^\delta - Kx^\delta) = 0$ and the proof is complete.

By virtue of (2.18) and (2.20), the least squares solution can be expressed as

$$x^\delta = K^\dagger y^\delta,$$

and since $x^\delta$ solves the normal equation, the SVD of $K$ yields the factorization

$$K^\dagger = (K^T K)^{-1} K^T = V \Sigma^\dagger U^T,$$  
(2.21)

with

$$\Sigma^\dagger = \begin{bmatrix} \text{diag} \left( \frac{1}{\sigma_i} \right)_{n \times n} & 0 \end{bmatrix}.$$  

Note that for rank $(K) = r < n$, $x^\delta$ defined by

$$x^\delta = \sum_{i=1}^{r} \frac{1}{\sigma_i} \left( u_i^T y^\delta \right) v_i,$$
is an element of $\mathcal{N}(K)^\perp = \text{span } \{v_i\}_{i=1}^r$ and represents the unique least squares solution of equation (2.11) in $\mathcal{N}(K)^\perp$. If $x_0$ is an arbitrary vector in $\mathcal{N}(K)$, then

$$K(x^\delta + x_0) = P_{\mathcal{R}(K)}y^\delta,$$

and $x^\delta + x_0$ is a least squares solution of equation (2.11) in $\mathbb{R}^n$. Using the orthogonality relation $x_0^T x^\delta = 0$, we observe from

$$\|x^\delta + x_0\|^2 = \|x^\delta\|^2 + 2x_0^T x^\delta + \|x_0\|^2 = \|x^\delta\|^2 + \|x_0\|^2,$$

that $x^\delta$ represents the least squares minimal norm solution of equation (2.11).

For discrete ill-posed problems, the following features of the singular values and vectors are relevant (Hansen, 1998):

1. as the dimension of $K$ increases, the number of small singular values also increases;
2. as the singular values $\sigma_i$ decrease, the corresponding singular vectors $u_i$ and $v_i$ have more sign changes in their components.

As a consequence of the oscillatory behavior of the high-order singular vectors, the norm of the least squares solution becomes extremely large and $x^\delta$ is not a reliable approximation of $x^\dagger$. To be more concrete, we choose a large singular-value index $i^*$ and consider a perturbation of the exact data vector $y$ in the direction of the singular vector $u_{i^*}$,

$$y^\delta = y + \Delta u_{i^*},$$

with $\Delta = \|y^\delta - y\|$ being the noise level. The least squares solution is then given by

$$x^\delta = x^\dagger + \frac{\Delta}{\sigma_{i^*}} v_{i^*},$$

and the ratio

$$\frac{\|x^\delta - x^\dagger\|}{\|y^\delta - y\|} = \frac{1}{\sigma_{i^*}}$$

is very large if $\sigma_{i^*}$ is very small (Figure 2.1). In this context, any naive approach which tries to compute $x^\delta$ by using (2.20) will usually return a useless result with extremely large norm.

The instability of an ill-conditioned linear system of equations depends on the decay rate of the singular values. In this sense, we say that a discrete problem is mildly ill-posed if $\sigma_i = O(i^{-\beta})$ for some positive $\beta$, and severely ill-posed if $\sigma_i = O(e^{-i})$.

### 2.3.3 Numerical example

The difficulties associated with the solution of ill-posed problems will be demonstrated by considering the temperature nadir sounding problem (2.1).

As water absorption is dominant for frequencies below 40 GHz and above 120 GHz, we assume a single oxygen line at position $\tilde{\nu}_{02}$ and ignore other absorbers completely.
Neglecting the temperature-dependence of line strength and pressure broadening, and assuming an observer at infinity gives the absorption optical depth (omitting the gas index)

\[
\tau_{abs}(\nu, z) = \frac{1}{\pi} \int_{z}^{\infty} \frac{S_{\gamma L0} n(z)}{(\nu - \hat{\nu})^2 + \left[ \frac{\gamma_{L0} p(z)}{p_{ref}} \right]^2} \frac{p(z)}{p_{ref}} \, dz. \tag{2.22}
\]

The volume mixing ratio \(q\) of \(O_2\) is constant with altitude, i.e., \(q = 0.21\), and the number density depends on altitude through pressure and temperature,

\[
n(z) = \frac{q}{k_B T(z)} p(z). \tag{2.23}
\]

Taking into account that pressure varies approximately exponentially with altitude (see (1.1)), assuming \(p_{ref} = p_0\) and ignoring the altitude dependence of the temperature in (2.23) (\(T\) varies between 200 and 300 K), the integral in (2.22) can be evaluated analytically. The result is

\[
\tau_{abs}(\nu, z) = a \log \left( \frac{(\nu - \hat{\nu})^2 + \gamma_{L0}^2 \exp \left( -\frac{2z}{H} \right)}{(\nu - \hat{\nu})^2} \right),
\]

and the transmission is then given by

\[
T(\nu, z) = \exp \left( -\tau_{abs}(\nu, z) \right) = \left[ \frac{(\nu - \hat{\nu})^2}{(\nu - \hat{\nu})^2 + \gamma_{L0}^2 \exp \left( -\frac{2z}{H} \right)} \right]^a,
\]

with

\[
a = \frac{qSP_0 \bar{H}}{2\pi k_B T_{\gamma L0}} = \frac{qSP_0}{2\pi \gamma_{L0} mg}.
\]
Fig. 2.2. Transmission $T$ (left) and weighting function $\partial T / \partial z$ (right) for a temperature nadir sounding model with exponent $a = 1.0$, line position $\nu = 2.0 \, \text{cm}^{-1}$ and $\delta \nu = \nu - \hat{\nu} = 10^{-6}, \ldots, 10^{-2} \, \text{cm}^{-1}$.

Chosing $S = 1.51 \cdot 10^{-25} \, \text{cm}^{-1} / (\text{molec} \cdot \text{cm}^2)$, $\gamma_{L0} = 0.1 \, \text{cm}^{-1}$, $m = 5 \cdot 10^{-23} \, \text{g}$, and $p_0 = 10^6 \, \text{g cm}^{-1}s^{-2}$, we find that $a \approx 1$. The transmission $T$ and the weighting function, defined by $\partial T / \partial z$, are depicted in Figure 2.2.

Now, passing from the vertical coordinate $z$ to the non-dimensional coordinate

$$\zeta = \frac{2z}{H}$$

and making the change of variable

$$\frac{1}{2ck_B \gamma_{L0}^2} I(\nu) \rightarrow I(\nu),$$

the integral equation (2.1) becomes

$$I(\nu) = \int_0^{\zeta_{\max}} K(\nu, \zeta) T(\zeta) \, d\zeta, \quad \nu \in [\nu_{\min}, \nu_{\max}],$$

with

$$K(\nu, \zeta) = \frac{\nu^2 (\nu - \hat{\nu})^2 \exp(-\zeta)}{\left[ (\nu - \hat{\nu})^2 + \gamma_{L0}^2 \exp(-\zeta) \right]^2}.$$
Assuming a discretization scheme with piecewise constant functions

\[ T_n(\zeta) = \sum_{j=1}^{n} T(\zeta_j) P_{nj}(\zeta) \]

we obtain the discrete equation

\[ [I_m]_i = \sum_{j=1}^{n} [K_{mn}]_{ij} [T_n]_j, \quad i = 1, \ldots, m, \quad (2.24) \]

with the forward model matrix \((N\) is the number of quadrature points\)

\[ [K_{mn}]_{ij} = \int_{0}^{\zeta_{\text{max}}} K(\nu_i, \zeta) P_{nj}(\zeta) \, d\zeta = \sum_{k=1}^{N} K(\nu_i, \zeta_k) P_{nj}(\zeta_k) \Delta \zeta_N, \quad (2.25) \]

the data vector

\[ [I_m]_i = I(\nu_i), \]

and the state vector

\[ [T_n]_j = T(\zeta_j). \]

The layer centerpoints in the data and state spaces are

\[ \nu_i = \left( i - \frac{1}{2} \right) \Delta \nu_m, \quad i = 1, \ldots, m, \]

and

\[ \zeta_j = \left( j - \frac{1}{2} \right) \Delta \zeta_n, \quad j = 1, \ldots, n, \]

respectively, while the discretization steps are \(\Delta \nu_m = (\nu_{\text{max}} - \nu_{\text{min}}) / m\) and \(\Delta \zeta_n = \zeta_{\text{max}} / n\). The integration scheme for computing the integral in (2.25) does not depend on the discretization grid in the state space and we have

\[ \zeta_k = \left( k - \frac{1}{2} \right) \Delta \zeta_N, \quad k = 1, \ldots, N, \]

with \(\Delta \zeta_N = \zeta_{\text{max}} / N\). Further, we set \(\nu_{\text{min}} = 1.98 \, \text{cm}^{-1}, \nu_{\text{max}} = 2.02 \, \text{cm}^{-1}, \zeta_{\text{max}} = 15\) and choose the exact temperature profile as

\[ T^\dagger(\zeta) = \begin{cases} 
220 + 28 \left( \frac{5}{2} - \zeta \right), & 0 \leq \zeta \leq 2.5, \\
220, & 2.5 < \zeta \leq 5, \\
220 + \frac{28}{3} (\zeta - 5), & 5 < \zeta \leq 11, \\
270, & 11 < \zeta \leq 14, \\
250 + 20 (15 - \zeta), & 14 < \zeta \leq 15. 
\end{cases} \quad (2.26) \]
Our analysis is organized as follows:

1. We fix the number of discrete data and quadrature points, \( m = 200 \) and \( N = 1000 \), respectively, and compute the exact data vector by using the relation

\[
[I_m]_i = \int_0^{\zeta_{\text{max}}} K(\nu_i, \zeta) T^\dagger(\zeta) d\zeta = \sum_{k=1}^{N} K(\nu_i, \zeta_k) T^\dagger(\zeta_k) \Delta \zeta_N;
\]

2. We generate the noisy data vector \( I_m^\delta \) as

\[
I_m^\delta = I_m + \delta I,
\]

where \( \delta I \) is a random Gaussian vector with zero mean and covariance \( C_\delta = \sigma^2 I_m \); the noise standard deviation is defined in terms of the signal-to-noise ratio SNR by

\[
\sigma = \frac{\|I_m\|}{\sqrt{m \text{SNR}}};
\]

3. For different values of the discretization index \( n \), we compute the least squares solution

\[
T^\delta_n = K^\dagger_{mn} I_m^\delta,
\]

and determine the solution error

\[
\varepsilon^2_n = \|T^\dagger - T^\delta_n\|^2 = \int_0^{\zeta_{\text{max}}} [T^\dagger(\zeta) - T^\delta_n(\zeta)]^2 d\zeta,
\]

where

\[
T^\delta_n(\zeta) = \sum_{j=1}^{n} [T^\delta_n]_j \delta j(\zeta).
\]

In the left panel of Figure 2.3 we plot the condition number of the matrix \( K_{mn} \) for different values of the discretization index \( n \). As expected, increasing the degree of discretization causes a huge condition number of the matrix and a dramatic amplification of solution errors is expected. The behavior of the Picard coefficients

\[
P^\delta_i = \frac{|u_i^T x^\delta_m|}{\sigma_i},
\]

which reflects the ill-posedness of the discrete equation, is shown in the right panel of Figure 2.3. As we will see in Chapter 3, if the sequence of Picard coefficients does not decay on average, then the reconstruction error is expected to be extremely large. For \( n = 12 \), the discrete Picard condition is satisfied on average and we anticipate reasonable small errors, while for \( n = 18 \), the Picard coefficients do not decay on average and extremely large reconstruction errors are expected.

Figure 2.4 shows the singular vectors \( v_1, v_6 \) and \( v_{14} \) corresponding to the singular values \( \sigma_1 = 1.8 \cdot 10^3 \), \( \sigma_6 = 1.2 \cdot 10^2 \) and \( \sigma_{14} = 1.6 \cdot 10^{-1} \), respectively. The results illustrate that the number of oscillations of the singular vector components increases when the corresponding singular values decrease. Note that fine structures in the profile are reproduced by singular vectors corresponding to smaller singular values, while singular vectors corresponding to larger singular values are responsible for the rough structures (see, e.g., Rodgers, 2000).
In the left panel of Figure 2.5 we plot the solution errors versus the discretization index $n$ for different values of the SNR. The results show that the solution error has a minimum for an optimal value $n^*$ of the discretization index. These values are $n^* = 12$ for SNR = 100, $n^* = 13$ for SNR = 500, and $n^* = 14$ for SNR = 1000; thus $n^*$ increases
when the SNR increases. The solution errors corresponding to the optimal values of the discretization index are reasonable small. This is apparent from the right panel of Figure 2.5, which illustrates the retrieved profile for \( n = 12 \) and SNR = 100. In contrast, up to the discretization index \( n = 14 \), the least squares solution oscillates and has a large norm.

The behavior of the solution error illustrates that projection methods have an inherent regularizing property (Natterer, 1977). If the discretization is too coarse, the finite-dimensional equation will be fairly well conditioned, but the solution will be affected by a large discretization error

\[
\varepsilon_{dn}^2 = \| T^\dagger - T_n^\dagger \|^2,
\]

where

\[
T_n^\dagger (\zeta) = \sum_{j=1}^{n} \left[ T_n^\dagger \right]_j P_{nj} (\zeta)
\]

and \( T_n^\dagger = K_{mn}^\dagger I_m \) is the least squares solution in the noise-free case. On the other hand, if the discretization is too fine, then the influence of the small singular values is significant, and the so-called noise error

\[
\varepsilon_{nn}^2 = \| T_n^\dagger - T_n^\delta \|^2
\]

explodes. Both the discretization and the noise errors contribute to the solution error \( \varepsilon_n \), and the optimal discretization index \( n^* \) balances the two error components.

The main drawback of regularization by projection is that the optimal discretization index \( n^* \) is too small and the corresponding vertical resolution is too low. Regularization methods yielding satisfactory reconstruction errors with high vertical resolutions are the topic of the next chapters.

*Fig. 2.5.* Left: relative errors versus the discretization index. Right: retrieved profiles for \( n = 12 \) and \( n = 14 \), in the case SNR = 100.
Numerical Regularization for Atmospheric Inverse Problems
Doicu, A.; Trautmann, Th.; Schreier, F.
2010, XIII, 426 p., Hardcover
ISBN: 978-3-642-05438-9