Università degli Studi di Cagliari

PhD DEGREE

in Mathematics and Computer Science

Cycle XXXIV

On the computation of the minimal-norm solution of linear and nonlinear problems

Scientific and Disciplinary Sector

MAT/08

PhD Student: Federica Pes

Supervisor Prof. Giuseppe Rodriguez

Final exam. Academic Year 2020/2021

Thesis defence: February 2022 Session

Acknowledgments

During the three years of my Ph.D. I had the possibility to increase my mathematics knowledge and to understand what it means to do mathematical research, but not only. I have met several people, who have become friends, to whom I am very thankful.

Firstly, I would like to say a very big thanks to my supervisor Prof. Giuseppe Rodriguez, who has inspired me in this discipline, for the continuous support of my Ph.D. study and related research, for his patience, motivation, encouragement, and immense knowledge.

I would like to express my gratitude to Prof. Lothar Reichel, for the welcome during my period spent at Kent State University and for the ongoing collaboration.

I would like to convey my gratefulness to my co-workers Prof. Luisa Fermo and Dr. Patricia Díaz de Alba. It is always a pleasure to work in a friendly atmosphere.

Thanks to all the other people, not yet mentioned, who "live" in the Department of Mathematics and Computer Science of the University of Cagliari, in particular: Silvia Frassu, Giuseppe Viglialoro, Alessandro Buccini, Anna Concas, Caterina Fenu, and Rafael Díaz Fuentes. Without you, my Ph.D. experience would not have been the same.

I would like to thank the reviewers Prof. Elisa Francomano and Prof. Marcello Lucia, for their suggestions which have improved this work.

Last, but not least, I would like to deeply thank my parents Pina and Nazario, and my sister Claudia, who have always supported and motivated me to never give up.

Contents

Introduction

The main topic of the thesis is the study of inverse problems and, in particular, it is the study of numerical methods for the computation of the minimal-norm solution of linear inverse problems in the continuous case and nonlinear ones in the discrete case.

Inverse problems arise in many areas of science and engineering, from the need to interpret indirect and incomplete measurements. Inverse problems are the opposites of direct problems. Informally, in a direct problem, one finds an effect from a cause, and in an inverse problem, one is given the effect and wants to recover the cause.

The most usual situation giving rise to an inverse problem is the need to interpret indirect physical measurements of an unknown object of interest, for instance, if one is interested in determining the internal structure of a physical system from the system's measured behavior, or in determining the unknown input that gives rise to a measured output signal.

Inverse problems are a recent topic in mathematics. Their study is motivated by the technological development of the last decades; for example, some of the more sophisticated medical diagnostic machines solve inverse problems, such as X-ray computed tomography, in which the inverse problem is to reconstruct the inner structure of an unknown physical body from the knowledge of X-ray images taken from different directions.

An example of inverse problem, which will be treated in some numerical experiments of this thesis, concerns the study of the subsoil in a non-destructive way, through the propagation of electromagnetic waves, in order to know some properties of the subsoil. Another example concerns image processing, where the goal is to find the sharp photograph from a given blurry image.

An inverse problem takes the form $F(x) = b$, where F is a linear or nonlinear operator, x represents the unknown solution, and b is the information available, that is, the measurements dataset. The goal is to reconstruct x starting from b. Inverse problems are closely related to the concept of ill-posed problems. To understand this concept we need to resort to the definition given by Hadamard at the beginning of the last century: such problems may not have a solution, or may have more

than one, or that solution is not stable with respect to perturbation in the data. In applications, ill-posed problems are common whenever there is little available measured data compared to the number of unknowns. In this case, it is necessary to reformulate the original ill-posed problem into a well-posed problem. A typical approach is to resort to a least-squares problem, in which the mean squared error between $F(x)$ and b is required to be minimal, i.e.,

$$\min \|F(x) - b\|^2,$$

where $\|\cdot\|$ is the Euclidean norm.

In this thesis, we are concerned with problems that have no unique solution. Among the different solutions, we want to determine the minimal-norm solution.

Definition 1. *Let F be a linear or nonlinear operator defined in a Banach space $\mathcal{B}$ with norm $\|\cdot\|$. Then, $x^\dagger \in \mathcal{B}$ is a minimal-norm solution to $F(x) = b$ if*

$$\|x^\dagger\| \leq \|x^*\|,$$

for any other solution $x^* \in \mathcal{B}$.

The subjects discussed in this thesis can be divided into two themes: *nonlinear least-squares problems* and *systems of linear integral equations of the first kind.*

It is common to solve **nonlinear least-squares problems** by Newton's method or one of its variants such as the Gauss–Newton algorithm. The idea of constructing an iterative method for the computation of the minimal-norm solution of such problems was first studied by Eriksson et al. They analyzed the cases of rank-deficient and ill-conditioned problems and proposed different solution techniques based on the Gauss–Newton method and on Tikhonov regularization in standard form. In this thesis, we review the results obtained by Eriksson et al. and extend them by introducing the minimization of a seminorm. We name our algorithm the *minimal-norm Gauss–Newton method* (MNGN). We further analyze the computation of the regularized minimal-norm solution of ill-conditioned nonlinear least-squares problems by two standard procedures, namely, the truncated generalized singular value decomposition applied to the Gauss–Newton method, and Tikhonov regularization in general form.

Then, since occasionally the MNGN method does not converge, we propose an improved version of this method. The problem of non-convergence has been considered by Campbell et al. They introduce a single parameter to enhance the convergence. In this thesis, we develop an algorithm based on two parameters to control both the Gauss–Newton and the projection step. We call our new algorithm the *doubly relaxed minimal-norm Gauss–Newton method* (MNGN2).

The other topic dealt with in this thesis concerns the solution of **systems of linear integral equations of the first kind**. We focus on overdetermined systems, that is, at least two integral equations whose solution is a single unknown

function that satisfies known boundary constraints. According to our knowledge, this problem has not been addressed before in the literature, although it arises in a variety of applications.

In an experimental setting, the available data is represented by the right-hand side evaluated at a finite set of points. This leads to a system of integral equations with discrete data. We show that this problem has infinitely many solutions. Then, we reformulate it as a minimal-norm solution problem and solve the latter in suitable function spaces. Specifically, we consider a *reproducing kernel Hilbert space* where, by using *Riesz theory*, the minimal-norm solution can be written as a linear combination of the so-called Riesz representers. While this approach is rather standard in functional analysis, it has never been applied before to an overdetermined system.

The thesis is divided into 6 chapters.

Chapter 1 is devoted to some recalls of numerical linear algebra and functional analysis. We remind the concept of ill-conditioned problem and we describe some regularization methods, useful to obtain a well-conditioned problem.

Chapter 2 introduces an iterative method to solve nonlinear least-squares problems, based on the linearization of the residual function. We study the computation of the minimal-norm solution, as well as the case where the solution minimizes a suitable seminorm. When the nonlinear function is ill-conditioned, we consider some regularization techniques for the solution.

Chapter 3 deals with the computation of the minimal-norm solution of nonlinear least-squares problems in the large-scale case.

Chapter 4 illustrates a numerical method to compute the minimal-norm solution of a linear system of integral equations of the first kind in the presence of boundary constraints. The problem is solved in particular Hilbert spaces.

Chapter 5 collects several test problems both artificial and deriving from real-world applications, of which some properties are studied.

Chapter 6 presents the numerical experiments that test the performance of the methods introduced in this thesis. The methods are applied to the problems presented in the above chapter.

CHAPTER 1

A review of some concepts

In this preliminary chapter, with the purpose of making this thesis self-contained, we recall some basic concepts and results in numerical linear algebra and in functional analysis that will be used in the forthcoming chapters.

1.1 Linear algebra

Here we report some basic concepts of linear algebra that will be useful throughout the thesis. A *subspace* of $\mathbb{R}^m$ is a subset that is also a vector space. Given a collection of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$, the set of all linear combinations of these vectors is a subspace referred to as the *span* of $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$:

$$\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \left\{ \mathbf{x} \in \mathbb{R}^m : \mathbf{x} = \sum_{j=1}^{n} \alpha_j \mathbf{v}_j \,,\ \alpha_j \in \mathbb{R} \right\}.$$

There are two important subspaces associated with a matrix $A \in \mathbb{R}^{m \times n}$. The **range** of A is defined as the subspace of $\mathbb{R}^m$ spanned by the columns of A:

$$\mathcal{R}(A) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n\},$$

and the **null space** of A is defined as the subspace of $\mathbb{R}^n$ spanned by the vectors mapped to the zero vector, i.e., those vectors $\mathbf{x}$ for which $A\mathbf{x} = 0$:

$$\mathcal{N}(A) = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = 0\}.$$

Similarly, we can define the range and the null space of A^T, where the superscript ^{T} stands for transposition. For a matrix A, these spaces are sometimes referred to as the *four fundamental subspaces*.

If $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ is a column partitioning, then

$$\mathcal{R}(A) = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}.$$

The **rank** of a matrix A is defined by

$$\text{rank}(A) = \dim(\mathcal{R}(A)).$$

It can be shown that $\text{rank}(A) = \text{rank}(A^T) \leq \min(m, n)$. A matrix $A \in \mathbb{R}^{m\times n}$ is *rank-deficient* if $\text{rank}(A) < \min(m, n)$. The rank of a matrix is the number of linearly independent columns or rows. Moreover, from the *Rank-Nullity Theorem*, it follows that

$$\dim(\mathcal{N}(A)) + \text{rank}(A) = n.$$

Norms furnish a measure of distance on vector spaces. More precisely, $\mathbb{R}^n$ together with a norm on $\mathbb{R}^n$ defines a *metric space*.

A **vector norm** on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following conditions:

- $\|\mathbf{x}\| \geq 0$, $\mathbf{x} \in \mathbb{R}^n$, ($\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$),
- $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, $\alpha \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$,
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

A useful class of vector norms are the Hölder *p-norms* defined by

$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}, \quad 1 \leq p < \infty.$$

The most commonly used norms for vectors are the 1, 2, and ∞ norms:

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^{n} |x_i|, \\
\|\mathbf{x}\|_2 &= \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} = \left(\mathbf{x}^T\mathbf{x}\right)^{1/2}, \\
\|\mathbf{x}\|_\infty &= \max_{1\leq i\leq n} |x_i|.
\end{aligned}$$

The 2-norm, also called the Euclidean norm, is invariant under orthogonal transformation, i.e., if $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\mathbf{x} \in \mathbb{R}^n$, then $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$. A sequence $\{\mathbf{x}^{(k)}\}$ of vectors *converges* to a vector $\mathbf{x}$ if and only if

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0$$

for any norm.

Let $A \in \mathbb{R}^{m\times n}$ be a matrix. A **matrix norm** is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that satisfies, in addition to the analogous three vector norm properties, two further properties:

- a matrix norm is *submultiplicative* if it holds for each pair of matrices of compatible size $\|AB\| \leq \|A\|\|B\|$;
- a matrix norm is *consistent* with the vector norms $\|\cdot\|_a$ of $\mathbb{R}^n$ and $\|\cdot\|_b$ of $\mathbb{R}^m$ if $\|A\mathbf{x}\|_b \leq \|A\|\|\mathbf{x}\|_a$, with $A \in \mathbb{R}^{m\times n}$ and $\mathbf{x} \in \mathbb{R}^n$.

One of the most used matrix norms is the Frobenius norm

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} = \left(\text{Trace}(A^TA)\right)^{1/2}.$$

A matrix norm can be constructed from any vector norm by defining

$$\|A\| = \max_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1}\|A\mathbf{x}\|.$$

Such a norm is generally called *natural*. The Frobenius norm does not fall in this definition. Formulas for the p-norms are known for $p = 1, 2, \infty$:

$$\|A\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|,$$
$$\|A\|_\infty = \max_{1\leq i\leq m}\sum_{j=1}^{n}|a_{ij}|,$$
$$\|A\|_2 = \left(\lambda_{\max}(A^TA)\right)^{1/2},$$

where $\lambda_{\max}(A^TA)$ is the largest eigenvalue of the matrix A^TA. An important property of the Frobenius norm and the 2-norm is that they are invariant with respect to orthogonal transformations.

Throughout the thesis, from now on, we indicate the Euclidean norm as $\|\cdot\|$ without subscript.

Now, we introduce the concept of orthogonality. A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ in $\mathbb{R}^n$ is *orthogonal* if $\mathbf{x}_i^T\mathbf{x}_j = 0$ whenever $i \neq j$ and *orthonormal* if $\mathbf{x}_i^T\mathbf{x}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. A collection of subspaces $S_1, \ldots, S_p$ in $\mathbb{R}^n$ is *mutually orthogonal* if $\mathbf{x}^T\mathbf{y} = 0$ whenever $\mathbf{x} \in S_i$ and $\mathbf{y} \in S_j$ for $i \neq j$.

The **orthogonal complement** of a subspace $S \subseteq \mathbb{R}^n$ is defined by

$$S^\perp = \left\{\mathbf{y} \in \mathbb{R}^n : \mathbf{y}^T\mathbf{x} = 0 \text{ for all } \mathbf{x} \in S\right\}.$$

It is not hard to show that $\mathcal{R}(A)^\perp = \mathcal{N}(A^T)$ and $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$. Moreover,

$$\mathbb{R}^n = \{\mathbf{w} : \mathbf{w} = \mathbf{u} + \mathbf{v}, \mathbf{u} \in \mathcal{R}(A^T), \mathbf{v} \in \mathcal{N}(A)\},$$

that is,

$$\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A),$$

where $\oplus$ denotes the direct sum. Similarly,

$$\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T).$$

Let $S \subseteq \mathbb{R}^n$ be a subspace. $\mathcal{P} \in \mathbb{R}^{n\times n}$ is the **orthogonal projection** onto S if $\mathcal{R}(\mathcal{P}) = S$, $\mathcal{P}^2 = \mathcal{P}$, and $\mathcal{P}^T = \mathcal{P}$. From this definition it is easy to show that if $\mathbf{x} \in \mathbb{R}^n$, then $\mathcal{P}\mathbf{x} \in S$ and $(I_n - \mathcal{P})\mathbf{x} \in S^\perp$. The last expression means that $(I_n - \mathcal{P})$ is the projector for the subspace complementary to that of S.

If $\mathcal{P}_1$ and $\mathcal{P}_2$ are orthogonal projections, then for any $\mathbf{z} \in \mathbb{R}^n$ we have

$$\|(\mathcal{P}_1 - \mathcal{P}_2)\mathbf{z}\|^2 = (\mathcal{P}_1\mathbf{z})^T(I_n - \mathcal{P}_2)\mathbf{z} + (\mathcal{P}_2\mathbf{z})^T(I_n - \mathcal{P}_1)\mathbf{z}.$$

If $\mathcal{R}(\mathcal{P}_1) = \mathcal{R}(\mathcal{P}_2) = S$, then the right-hand side of this expression is zero, showing that the orthogonal projection for a subspace is unique. If the columns of $V = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ are an orthonormal basis for a subspace S, then it is easy to show that $\mathcal{P} = VV^T$ is the unique orthogonal projection onto S.

1.1.1 Matrix decompositions

In the following, we review some of the most important factorizations of a matrix A [10, 45].

We start with the **eigendecomposition**. Symmetry guarantees that all eigenvalues are real and that there is an orthonormal basis of eigenvectors. If $A \in \mathbb{R}^{n\times n}$ is symmetric, then there exists a real orthogonal $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$ such that

$$A = Q\Lambda Q^T, \tag{1.1}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ whose elements are the *eigenvalues* of A. The columns $\mathbf{q}_i$ of Q are the *eigenvectors* of A. Moreover, for $i = 1, \ldots, n$, they satisfy $A\mathbf{q}_i = \lambda_i\mathbf{q}_i$.

As the eigenvectors are the non-trivial solutions of the linear system $(A - \lambda I_n)\mathbf{q} = \mathbf{0}$, the eigenvalues have to be the roots of the characteristic polynomial of A, defined by $p_A(\lambda) = \det(A - \lambda I_n)$. The set of these roots, solutions of the characteristic equation $\det(A - \lambda I_n) = 0$, is the so-called *spectrum* of A and the *spectral radius* of A is the largest absolute value of its eigenvalues. The decomposition (1.1) is called **spectral factorization**.

Then, we recall a matrix decomposition of great importance for the treatment of least-squares problems. For a matrix $A \in \mathbb{R}^{m\times n}$ with $r = \operatorname{rank}(A)$, the **singular value decomposition** (SVD) is a matrix decomposition of the form

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{1.2}$$

where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m\times m}$ and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$ are matrices with orthonormal columns, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix. The non-zero diagonal

elements of the matrix Σ are the *singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The $\mathbf{u}_i$ and the $\mathbf{v}_i$ are the *left* and the *right singular vectors* of A, respectively, associated with σ_i, $i = 1, \ldots, r$. The SVD gives an expression to write a matrix A as the sum of r rank-1 matrices; see equation (1.2).

The SVD of a matrix A furnishes orthonormal bases for its null space and its range, as well as for their orthogonal complements

$$\begin{aligned} \mathcal{R}(A) &= \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}, & \mathcal{N}(A^T) &= \operatorname{span}\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n\}, \\ \mathcal{R}(A^T) &= \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}, & \mathcal{N}(A) &= \operatorname{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}. \end{aligned}$$

From (1.2) it follows that

$$\begin{aligned} A\mathbf{v}_i &= \sigma_i \mathbf{u}_i, \quad \|A\mathbf{v}_i\| = \sigma_i, \qquad i = 1, \ldots, r, \\ A^T\mathbf{u}_i &= \sigma_i \mathbf{v}_i, \quad \|A^T\mathbf{u}_i\| = \sigma_i, \qquad i = 1, \ldots, r. \end{aligned}$$

From these relations, we see that for each small singular value σ_i, compared to $\sigma_1 = \|A\|$, there exists a linear combination of the columns of A such that $\|A\mathbf{v}_i\| = \sigma_i$ is small. This means that A is nearly rank-deficient, and the vectors $\mathbf{v}_i$ associated with the small singular values are vectors in the numerical null space of A. The same holds for the rows of A, and the vectors $\mathbf{u}_i$ associated with the small σ_i are vectors in the numerical null space of A^T.

From the relations $A^TA = V\Sigma^T\Sigma V^T$ and $AA^T = U\Sigma\Sigma^T U^T$ we see that the SVD of A is linked to the eigenvalue decompositions of the symmetric semidefinite matrices A^TA and AA^T. Thus $\sigma_1^2, \ldots, \sigma_r^2$ are the non-zero eigenvalues of the matrices A^TA and AA^T, and $\mathbf{v}_i$ and $\mathbf{u}_i$ are the corresponding eigenvectors.

The SVD has a critical role to play because it can be used to identify an approximation of a matrix by another one of lower rank.

Theorem 1.1.1. *(Eckhart-Young's Theorem [45, Theorem 2.4.8]). If* $k < r = \operatorname{rank}(A)$ *and*

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

then

$$\begin{aligned} \min_{\operatorname{rank}(Z_k)=k} \|A - Z_k\| &= \|A - A_k\| = \sigma_{k+1}, \\ \min_{\operatorname{rank}(Z_k)=k} \|A - Z_k\|_F &= \|A - A_k\|_F = \left(\sigma_{k+1}^2 + \cdots + \sigma_n^2\right)^{1/2}. \end{aligned}$$

The SVD gives the explicit expression for projectors. Indeed, matrices U and V can be partitioned

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \qquad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix},$$

where U_1 and V_1 consist of the first r columns of U and V, respectively, U_2 is composed of the $m - r$ remaining columns of U, and V_2 of the remaining $n - r$ of

V. We can write the orthogonal projectors onto the four fundamental subspaces of A in terms of its singular vectors:

$$\begin{aligned} \mathcal{P}_{\mathcal{R}(A)} &= U_1 U_1^T, \quad \mathcal{P}_{\mathcal{N}(A^T)} = U_2 U_2^T, \\ \mathcal{P}_{\mathcal{R}(A^T)} &= V_1 V_1^T, \quad \mathcal{P}_{\mathcal{N}(A)} = V_2 V_2^T. \end{aligned}$$

Another important concept related to least-squares problems is that of **pseudoinverse** of a matrix $A \in \mathbb{R}^{m\times n}$ (or Moore-Penrose inverse), which can be defined by using the SVD. Assuming that $r = \text{rank}(A)$ and $m \geq n$, it is the matrix $A^\dagger \in \mathbb{R}^{n\times m}$ such that

$$A^\dagger = V \Sigma^\dagger U^T,$$

where

$$\Sigma^\dagger = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \in \mathbb{R}^{n\times m}.$$

The pseudoinverse $A^\dagger$ is defined to be the unique matrix $X \in \mathbb{R}^{n\times m}$ that satisfies the four Moore-Penrose conditions:

$$\begin{aligned} AXA &= A, & XAX &= X, \\ (AX)^T &= AX, & (XA)^T &= XA. \end{aligned}$$

It can be characterized as the unique minimal Frobenius norm solution to the problem

$$\min_{X\in\mathbb{R}^{n\times m}} \|AX - I_m\|_F.$$

The pseudoinverse has interesting properties, similar to those of the ordinary inverse, that are summarized in [10, Theorem 1.2.12]. On the contrary, the pseudoinverse does not share some other properties of the ordinary inverse, for instance $AA^\dagger \neq A^\dagger A$.

The pseudoinverse gives simple expressions for the orthogonal projectors onto the four fundamental subspaces of A:

$$\begin{aligned} \mathcal{P}_{\mathcal{R}(A)} &= AA^\dagger, \quad \mathcal{P}_{\mathcal{N}(A^T)} = I_m - AA^\dagger \\ \mathcal{P}_{\mathcal{R}(A^T)} &= A^\dagger A, \quad \mathcal{P}_{\mathcal{N}(A)} = I_n - A^\dagger A. \end{aligned}$$

If $m > n$ and $\text{rank}(A) = n$, then $A^\dagger = (A^T A)^{-1} A^T$, while if $m = n = \text{rank}(A)$, then $A^\dagger = A^{-1}$. If $m < n$ and $\text{rank}(A) = m$, then $A^\dagger = A^T (AA^T)^{-1}$.

We recall two other important factorizations. The **QR decomposition**: let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, then there exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times n}$ with nonnegative diagonal elements such that

$$A = QR.$$

The **Cholesky factorization**: a symmetric and positive definite matrix $A \in \mathbb{R}^{n\times n}$ can be decomposed as

$$A = R^T R,$$

where R is an upper triangular matrix with positive diagonal entries.

We conclude the paragraph on matrix decompositions with a review of a simultaneous decomposition of a pair of matrices. Let $A \in \mathbb{R}^{m\times n}$ and $L \in \mathbb{R}^{p\times n}$ be matrices with $\text{rank}(A) = r$ and $\text{rank}(L) = p$. Assume that $m + p \geq n$ and

$$\text{rank}\left(\begin{bmatrix} A \\ L \end{bmatrix}\right) = n,$$

which corresponds to requiring that $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$. The **generalized singular value decomposition** (GSVD) of the matrix pair (A, L) is a decomposition of the form

$$A = U\Sigma_A W^{-1}, \qquad L = V\Sigma_L W^{-1}, \tag{1.3}$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$ are matrices with orthonormal columns $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, and $W \in \mathbb{R}^{n\times n}$ is nonsingular. If $m \geq n \geq r$, then the matrices $\Sigma_A \in \mathbb{R}^{m\times n}$ and $\Sigma_L \in \mathbb{R}^{p\times n}$ have the form

$$\Sigma_A = \left[\begin{array}{c} O_{n-r} \\ \hline C \\ \hline I_d \\ \hline O_{(m-n)\times n} \end{array}\right], \qquad \Sigma_L = \left[\begin{array}{ccc|c} I_{p-r+d} & & & \\ & & & O_{p\times d} \\ & & S & \end{array}\right], \tag{1.4}$$

where $d = n - p$,

$$\begin{aligned} C &= \text{diag}(c_1, \ldots, c_{r-d}), \qquad & 0 < c_1 \leq c_2 \leq \cdots \leq c_{r-d} < 1, \\ S &= \text{diag}(s_1, \ldots, s_{r-d}), \qquad & 1 > s_1 \geq s_2 \geq \cdots \geq s_{r-d} > 0, \end{aligned} \tag{1.5}$$

with $c_i^2 + s_i^2 = 1$, for $i = 1, \ldots, r-d$. The identity matrix of size $k \times k$ is denoted by I_k, while O_k and $O_{k\times\ell}$ are zero matrices of size $k \times k$ and $k \times \ell$, respectively; a matrix block has to be omitted when one of its dimensions is zero. The scalars $\gamma_i = \frac{c_i}{s_i}$ are called *generalized singular values*, and they appear in non-decreasing order.

If $r \leq m < n$, then the matrices $\Sigma_A \in \mathbb{R}^{m\times n}$ and $\Sigma_L \in \mathbb{R}^{p\times n}$ take the form

$$\Sigma_A = \left[\begin{array}{c|ccc} O_{m\times(n-m)} & O_{m-r} & & \\ & & C & \\ & & & I_d \end{array}\right], \qquad \Sigma_L = \left[\begin{array}{ccc|c} I_{p-r+d} & & & \\ & & & O_{p\times d} \\ & & S & \end{array}\right], \tag{1.6}$$

where the blocks are defined as above.

When L is the identity matrix, then the matrices U and V of the GSVD are the same U and V of the SVD, and the generalized singular values of (A, I_n) coincide with the singular values of A, except for the reverse ordering.

The GSVD is connected to regularization methods in general form, in which the regularization term takes the form $\|L\mathbf{x}\|$; see equations (1.14) and (1.15) in Section 1.2 for examples of matrices L.

1.1.2 Linear systems and least-squares problems

A system of linear equations (or **linear system**) is a collection of one or more linear equations involving the same set of variables. A system of m linear equations with n unknowns can be written in matrix form as

$$A\mathbf{x} = \mathbf{b},$$

where $A \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. When there are more equations than unknowns, i.e., $m \geq n$, we say that the system is *overdetermined*. On the contrary, if $m < n$, we are dealing with the so-called *underdetermined* systems.

We say that a linear system $A\mathbf{x} = \mathbf{b}$ is compatible if the right-hand side $\mathbf{b} \in \mathcal{R}(A)$. If $m \leq n$ and A has full rank, then $\mathcal{R}(A) = \mathbb{R}^m$ and all systems are compatible.

If $m \geq n$ and A has full rank, then it has the trivial null space $\mathcal{N}(A) = \{\mathbf{0}\}$. An overdetermined system has no exact solution if $\mathbf{b}$ is not an element of $\mathcal{R}(A)$. This suggest to consider the **least-squares problem**

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|, \tag{1.7}$$

where the matrix A and the vector $\mathbf{b}$ are given. We want to find a vector $\mathbf{x}$ such that $A\mathbf{x}$ is the "best" approximation to $\mathbf{b}$. The least-squares problem originally arose from the need to fit a linear mathematical model to given observations.

We now characterize the set of all solutions to the least-squares problem (1.7).

Theorem 1.1.2. *[10, Theorem 1.1.2] Denote the set of all solutions by*

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n : \|A\mathbf{x} - \mathbf{b}\| = \min\}. \tag{1.8}$$

Then $\mathbf{x} \in \mathcal{S}$ *if and only if*

$$A^T(\mathbf{b} - A\mathbf{x}) = 0.$$

Denoted by $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ the residual vector, the theorem shows that for every least-squares solution it is true that $\mathbf{r} \in \mathcal{N}(A^T)$, and viceversa. Any least-squares solution $\mathbf{x}$ uniquely decomposes the right-hand side $\mathbf{b}$ into two orthogonal components

$$\mathbf{b} = A\mathbf{x} + \mathbf{r}, \qquad A\mathbf{x} \in \mathcal{R}(A), \qquad \mathbf{r} \in \mathcal{N}(A^T).$$

From Theorem 1.1.2 it follows that a least-squares solution satisfies the *normal equations*

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

The matrix $A^T A \in \mathbb{R}^{n\times n}$ is symmetric and nonnegative definite. The normal equations are always consistent since

$$A^T\mathbf{b} \in \mathcal{R}(A^T) = \mathcal{R}(A^T A).$$

If $m \geq n$ and $\text{rank}(A) = n$ the matrix $A^T A$ is positive definite, then the unique least-squares solution and the corresponding residual are given by

$$\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}, \qquad \mathbf{r} = \mathbf{b} - A(A^T A)^{-1} A^T \mathbf{b}.$$

Since $\mathbf{r} \in \mathcal{N}(A^T)$, then $\mathbf{r} = (I_m - \mathcal{P}_{\mathcal{R}(A)})\mathbf{b}$. In the full rank case

$$\mathcal{P}_{\mathcal{R}(A)} = A(A^T A)^{-1} A^T.$$

If $\text{rank}(A) = r < n$ then A has a non-trivial null space and the least-squares solution is not unique. If $\widehat{\mathbf{x}}$ is a particular least-squares solution then the set of all least-squares solutions is

$$\mathcal{S} = \{\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{z} : \mathbf{z} \in \mathcal{N}(A)\}.$$

If $\widehat{\mathbf{x}} \perp \mathcal{N}(A)$ then $\|\mathbf{x}\|^2 = \|\widehat{\mathbf{x}}\|^2 + \|\mathbf{z}\|^2$, and therefore $\widehat{\mathbf{x}}$ is the least-squares solution of minimal norm. It can be proved that it is unique.

However, even if $\text{rank}(A) = n$, difficulties may arise if A is nearly rank deficient. We use the SVD (1.2) to analyze least-squares problems. The solution is given by

$$\mathbf{x} = A^\dagger \mathbf{b} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i.$$

If σ_n is small, small changes in A or $\mathbf{b}$ can induce relatively large changes in $\mathbf{x}$. The presence of one or more small non-zero singular values is the problem, because $\|\mathbf{x}\|^2$ may become very large due to a small σ_i. At the end of this section, we explain how to measure this sensitivity.

The problem of computing the minimal-norm solution $\mathbf{y} \in \mathbb{R}^m$ to an underdetermined system of linear equations

$$\min \|\mathbf{y}\|, \qquad A^T \mathbf{y} = \mathbf{c},$$

where $A \in \mathbb{R}^{m \times n}$, $m > n$, occurs as a subproblem in optimization algorithms. If $\text{rank}(A) = n$, then the system $A^T \mathbf{y} = \mathbf{c}$ is consistent and the unique solution is given by the normal equations of the second kind

$$A^T A \mathbf{z} = \mathbf{c}, \qquad \mathbf{y} = A\mathbf{z},$$

that is, $\mathbf{y} = A(A^T A)^{-1} \mathbf{c}$. If $\text{rank}(A) < n$, then at least one row of A^T is a linear combination of the others. If the right-hand side $\mathbf{c}$ does not satisfy the same linear combination, then the system is incompatible.

Since the set (1.8) of all minimizers is convex, it follows that among all least-squares solutions there is a unique solution which minimizes $\|\mathbf{x}\|$. We have the following fundamental result, which applies to all cases, either overdetermined or underdetermined linear systems, full rank or rank-deficient.

Theorem 1.1.3. *[45, Theorem 5.5.1] Suppose $A = U\Sigma V^T$ is the SVD of $A \in \mathbb{R}^{m\times n}$ with $r = \text{rank}(A)$. Then*

$$\mathbf{x}_{LS} = A^\dagger \mathbf{b} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i \tag{1.9}$$

minimizes $\|A\mathbf{x} - \mathbf{b}\|$ and has the smallest norm among all minimizers. Moreover

$$\|A\mathbf{x}_{LS} - \mathbf{b}\|^2 = \sum_{i=r+1}^{n} (\mathbf{u}_i^T \mathbf{b})^2.$$

Conditioning. Here we give some results on the sensitivity of least-squares solutions to perturbations in the right-hand side $\mathbf{b}$. In [10, Chapter 1] there are the details about the results on the sensitivity to perturbation in A and $\mathbf{b}$.

It is needed to introduce the concept of **condition number** of a matrix $A \in \mathbb{R}^{m\times n}$, defined by

$$\kappa(A) = \|A\|\|A^\dagger\| = \frac{\sigma_1}{\sigma_r},$$

where $r = \text{rank}(A)$. Herein we consider the 2-norm condition number and it is equivalent to the ratio between the largest singular value σ_1 and the smallest nonzero singular value σ_r. The quantity $\kappa(A)$ measures the sensitivity of the solution to perturbations of $\mathbf{b}$. Indeed, assume that the exact $\mathbf{x}_{\text{exact}}$ and the perturbed solution $\mathbf{x}$ satisfy

$$A\mathbf{x}_{\text{exact}} = \mathbf{b}_{\text{exact}}, \qquad A\mathbf{x} = \mathbf{b} = \mathbf{b}_{\text{exact}} + \mathbf{e},$$

where $\mathbf{e}$ denotes the perturbation. If A has full rank, then the perturbed solution is given by $\mathbf{x} = A^\dagger \mathbf{b} = \mathbf{x}_{\text{exact}} + A^\dagger \mathbf{e}$, and an upper bound for the relative perturbation is given by

$$\frac{\|\mathbf{x}_{\text{exact}} - \mathbf{x}\|}{\|\mathbf{x}_{\text{exact}}\|} \leq \kappa(A) \frac{\|\mathbf{e}\|}{\|\mathbf{b}_{\text{exact}}\|}.$$

The larger the condition number, the more sensitive the system is to perturbations in the right-hand side. If $\kappa(A)$ is large, this implies that $\mathbf{x}$ can be very far from $\mathbf{x}_{\text{exact}}$. For any of the p-norms, we have $\kappa_p(A) = \|A\|_p\|A^\dagger\|_p \geq 1$.

If $\kappa(A)$ is large, then A is said to be an *ill-conditioned* matrix. Matrices with small condition number are said to be *well-conditioned*. Orthogonal matrices are perfectly conditioned in the 2-norm, indeed, if Q is orthogonal, then $\kappa(Q) = \|Q\|\|Q^T\| = 1$.

1.2 Regularization

When talking about ill-conditioned matrices, knowledge of the SVD is useful. In particular, the condition number of a matrix is defined as the ratio between the largest

and the smallest singular values. The numerical treatment of systems of equations with an ill-conditioned coefficient matrix depends on the type of ill-conditioning of A. There are two important classes of problems, and many practical problems belong to one of these two classes [57].

Rank-deficient problems are characterized by the matrix A having a cluster of small singular values, and there is a well-determined gap between large and small singular values. In this case, one or more rows and columns of A are nearly linear combinations of some or all of the remaining rows and columns. In other words, the matrix A contains almost redundant information, and the key to the numerical treatment of such problems is to consider only the linearly independent information in A, to arrive at another problem with a well-conditioned matrix. The small singular values approximate the zero singular values. Once the numerical rank has been identified, the matrix could be well-conditioned.

Discrete ill-posed problems arise from the discretization of ill-posed problems such as Fredholm integral equations of the first kind. Here all the singular values of A, as well as the SVD components of the solution, decay gradually to zero, and we say that a discrete Picard condition is satisfied. Since there is no gap in the singular value spectrum, there is no notion of numerical rank for these matrices. For discrete ill-posed problems, the goal is to find a balance between the residual norm and the size of the solution. The word "size" should be interpreted in a broad sense; e.g., size can be a norm, a seminorm, or a Sobolev norm.

Figure 1.1 shows these two different behaviors: on the left we can clearly see the gap between large and small singular values of a rank-deficient problem, while on the right picture, the singular values of a discrete ill-posed problem decay gradually to zero. The test problems are included in the "Regularization Tools" [58] package as functions `heat` and `baart`.

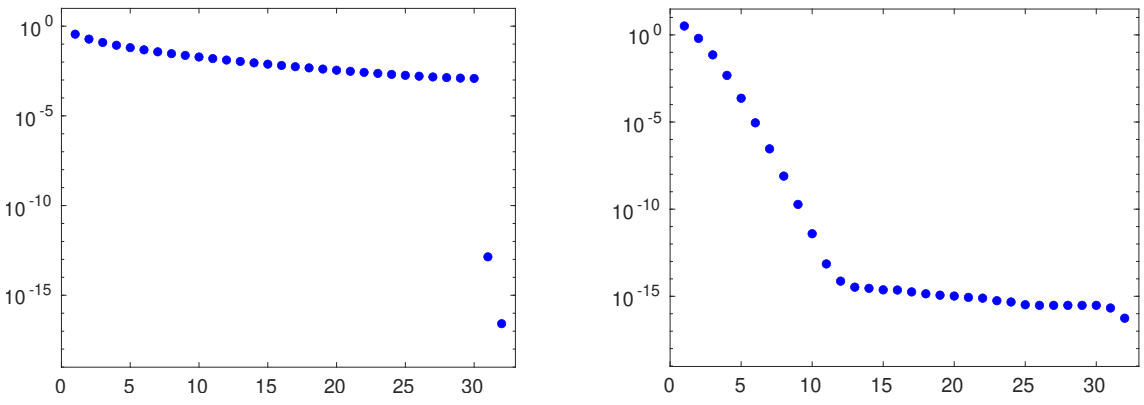

Figure 1.1: The 32 singular values σ_i of A for the `heat` test problem (left) and for the `baart` test problem (right).

The word *ill-posed* is almost automatically associated with inverse problem. To explain the concept of ill-posed problem, we need the notion of a *well-posed problem*

introduced by Hadamard [52]. Problems of this type must verify the following conditions:

- **Existence.** There should be at least one solution.
- **Uniqueness.** There should be at most one solution.
- **Stability.** The solution must depend continuously on data.

A problem is said **ill-posed** if at least one of these conditions fails. If the last condition is not verified, it means that an arbitrarily small perturbation of the data can cause an arbitrarily large perturbation of the solution.

A classical example of a linear ill-posed problem is a *Fredholm integral equation of the first kind* with a square-integrable kernel

$$\int_a^b k(s,t)f(t)\,dt = g(s), \qquad c \le s \le d, \tag{1.10}$$

where the right-hand side g and the kernel k are known functions, while f is the unknown. Let $L^2([a,b])$ be the Hilbert space of square-integrable functions in the interval $[a,b]$. The corresponding integral operator is $K : L^2([a,b]) \to L^2([c,d])$ (see Section 1.3)

$$(Kf)(s) := \int_a^b k(s,t)f(t)\,dt.$$

In many practical applications the kernel k is given exactly by the underlying mathematical model, while the right-hand side g typically consists of measured quantities, i.e., it is only known in a finite set of points $s_1, \ldots, s_m$

$$\int_a^b k(s_i,t)f(t)\,dt = g(s_i), \qquad i = 1, \ldots, m. \tag{1.11}$$

The analytical tool for the analysis of first kind Fredholm integral equations (1.10) with square-integrable kernels is the **singular value expansion** (SVE) of the kernel. First, we need to introduce a bit of notation. Given two functions ϕ and ψ defined on the interval $[a,b]$, their inner product is defined in L^2 as

$$\langle \phi, \psi \rangle = \int_a^b \phi(t)\psi(t)\,dt,$$

and the norm of the function ϕ is defined as

$$\|\phi\|_{L^2} = \langle \phi, \phi \rangle^{1/2} = \left(\int_a^b \phi(t)^2\,dt \right)^{1/2}.$$

A kernel k is square-integrable if the norm

$$\|k\|^2 = \int_a^b \int_c^d k(s,t)^2\,ds\,dt$$

is finite. For any square-integrable kernel k the SVE takes the form

$$k(s,t) = \sum_{i=1}^{\infty} \mu_i u_i(s) v_i(t). \tag{1.12}$$

The functions u_i and v_i are termed the *singular functions*. They are orthonormal with respect to the inner product, i.e.,

$$\langle u_i, u_j \rangle = \langle v_i, v_j \rangle = \delta_{ij},$$

where δ_{ij} is the Kronecker delta. The numbers μ_i are the *singular values*; they are nonnegative, they are ordered in non-increasing order such that

$$\mu_1 \geq \mu_2 \geq \cdots \geq 0,$$

and they satisfy the relation $\|k\|^2 = \sum_{i=1}^{\infty} \mu_i^2$.

The most important relation between singular values and functions is the following fundamental relation:

$$\int_a^b k(s,t) v_i(t)\, dt = \mu_i u_i(s), \quad i = 1, 2, \ldots,$$

which shows that any singular function v_i is mapped onto the corresponding u_i, and that the singular value μ_i is the amplification of this particular mapping. The integral in the above equation can be written as an operator and by considering the concept of adjoint K^* (see Section 1.3), the following relations hold

$$Kv_i = \mu_i u_i, \qquad K^* u_i = \mu_i v_i, \qquad i = 1, 2, \ldots,$$

$$Kf = \sum_{i=1}^{\infty} \mu_i \langle f, v_i \rangle u_i, \qquad K^* g = \sum_{i=1}^{\infty} \mu_i \langle g, u_i \rangle v_i, \qquad i = 1, 2, \ldots.$$

Each system $\{\mu_i, u_i, v_i\}$ with these properties is called a **singular system** of K.

The left and right singular functions u_i and v_i form bases for the function spaces $L^2([c,d])$ and $L^2([a,b])$, respectively. Hence, we can expand both f and g in terms of these functions:

$$f(t) = \sum_{i=1}^{\infty} \langle v_i, f \rangle v_i(t), \quad \text{and} \quad g(s) = \sum_{i=1}^{\infty} \langle u_i, g \rangle u_i(s).$$

If the SVE (1.12) is inserted into the integral equation (1.10), then we obtain, after some computation, the following expression for the solution to (1.10)

$$f(t) = \sum_{i=1}^{\infty} \frac{\langle u_i, g \rangle}{\mu_i} v_i(t). \tag{1.13}$$

There is also a system of m triplets $\{\mu_i, u_i, v_i\}$ associated with the operator in the Fredholm integral equation (1.11) with a discrete right-hand side. In this case, u_i are m orthonormal functions while v_i are m orthonormal vectors in $\mathbb{R}^m$.

The Picard condition. In order that there exists a square-integrable solution f to the integral equation (1.10), the right-hand side g must satisfy the Picard condition

$$\sum_{i=1}^{\infty} \left(\frac{\langle u_i, g \rangle}{\mu_i} \right)^2 < \infty.$$

This condition says that from some point in the summation in (1.13), the absolute value of the coefficients $\langle u_i, g \rangle$ must decay faster than the corresponding singular values μ_i in order that a square-integrable solution exists. The trouble with first kind Fredholm integral equations is that, even if the exact data satisfies the Picard condition, the measured and noisy data g usually violates the condition.

As we have seen, the primary difficulty with ill-posed problems is the presence of the cluster of small singular values. Hence, it is necessary to incorporate further information about the desired solution in order to stabilize the problem and to single out a useful and stable solution. This is the purpose of regularization.

In order to solve the problem numerically, we must discretize it. There are essentially two main classes of methods, namely, quadrature methods and Galerkin methods, to discretize integral equations. Both methods compute an approximation to f. In the *quadrature method* (see Section 1.3), a quadrature rule with abscissas $t_1, \ldots, t_n$ and corresponding weights $w_1, \ldots, w_n$ is used to approximate an integral as

$$\int_a^b \phi(t)\, dt \approx \sum_{j=1}^{n} w_j \phi(t_j),$$

and when this rule is applied to the integral equation (1.10) for m distinct values $s_1, \ldots, s_m$, then we obtain an $m \times n$ matrix A and a right-hand side $\mathbf{b}$ with elements given by

$$a_{ij} = w_j k(s_i, t_j), \quad b_i = g(s_i).$$

The discretization leads to a system

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|, \qquad A \in \mathbb{R}^{m \times n},$$

where the vector $\mathbf{x}$ represents the function f.

When a rank-deficient or ill-posed problem is discretized, then the difficulties carry over to the discrete problem in the sense that the coefficient matrix will also have either a cluster of small singular values or singular values that decay gradually to zero. Hence, some kind of regularization is also required to solve the discretized problem.

There exist different approaches to regularization: applying the regularization means

- minimizing the residual norm $\|A\mathbf{x} - \mathbf{b}\|$ subject to the constraint that the solution belongs to a specified subset;

- minimizing the residual norm subject to the constraint that the "size" of the solution is less than some specified upper bound;
- minimizing the "size" of the solution subject to a constraint on the residual norm;
- minimizing a linear combination of the residual norm and the "size" of the solution.

In order to select solutions exhibiting different degrees of regularity, the "size" of the solution is often of the form $\|L\mathbf{x}\|$, where the matrix L is typically either the identity matrix, a diagonal weighting matrix, or a $p \times n$ discrete approximation of a derivative operator, in which case L is a banded matrix with full row rank. For example, the matrices

$$D_1 = \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix} \quad \text{and} \quad D_2 = \begin{bmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix}, \tag{1.14}$$

of size $(n-1) \times n$ and $(n-2) \times n$, respectively, are approximations to the first and second derivative operators. Regularization operators of this form are often referred to as smoothing operators.

An effective choice of L is such that the solution $\mathbf{x}$ is (at least approximately) in the null space of L. For $p < n$ the matrix L always has a non-trivial $\mathcal{N}(L)$. If $L = D_1$, then $\mathcal{N}(L)$ contains constant vectors, while $\mathcal{N}(D_2)$ includes constant and linearly varying vectors, i.e., $\mathcal{N}(D_2)$ is spanned by the vectors $[1, 1, \ldots, 1]^T$ and $[1, 2, \ldots, n]^T$. In this case, $\|L \cdot\|$ is said to be a *seminorm*, i.e., there exist vectors $\mathbf{x} \neq \mathbf{0}$, in the null space of L, i.e., such that $\|L\mathbf{x}\| = 0$.

The constraint $p \leq n$ is not restrictive. Indeed (see, e.g., [57]), if $p > n$ it is possible to perform the compact QR factorization $L = QR$ with $Q \in \mathbb{R}^{p \times p_1}$, $R \in \mathbb{R}^{p_1 \times n}$, and $p_1 = \text{rank}(L) \leq n$. In this case the matrix L can be substituted by the triangular matrix R, as $\|L\mathbf{x}\| = \|R\mathbf{x}\|$ for any vector $\mathbf{x}$.

Other regularization matrices might be used. For instance, the regularization matrix

$$L = \begin{bmatrix} I_n \otimes D_1 \\ D_1 \otimes I_n \end{bmatrix}, \tag{1.15}$$

where I_n denotes the identity matrix of order n and $\otimes$ stands for the Kronecker product, is commonly used in image restoration [19, 70].

If an a priori estimate $\overline{\mathbf{x}}$ of the desired regularized solution is available, then this information can be taken into account by the following term

$$\|L(\mathbf{x} - \overline{\mathbf{x}})\|. \tag{1.16}$$

From general to standard form. A regularization problem with regularization term $\|L(\mathbf{x}-\overline{\mathbf{x}})\|$ is said to be in *standard form* if the matrix L is the identity matrix I_n. In many applications, regularization in standard form is not the best choice. Since only the matrix A is involved, instead of the matrices A and L, it is much simpler to treat problems in standard form, from a numerical point of view. A given regularization problem with residual $\|A\mathbf{x}-\mathbf{b}\|$ and regularization term $\|L(\mathbf{x}-\overline{\mathbf{x}})\|$ can be transformed into one in standard form, with a new residual $\|\widehat{A}\widehat{\mathbf{x}}-\widehat{\mathbf{b}}\|$ and a new regularization term $\|\widehat{\mathbf{x}}-\widehat{\overline{\mathbf{x}}}\|$, where

$$\widehat{A} = AL_A^\dagger, \quad \widehat{\mathbf{b}} = \mathbf{b} - A\mathbf{x}_0, \quad \widehat{\overline{\mathbf{x}}} = L\overline{\mathbf{x}},$$

while the transformation back to the general-form setting becomes

$$\mathbf{x} = L_A^\dagger \widehat{\mathbf{x}} + \mathbf{x}_0.$$

In the above equations, $L_A^\dagger$ is the A-weighted pseudoinverse of L, defined as

$$L_A^\dagger = \left(I_n - \left(A(I_n - L^\dagger L)\right)^\dagger A\right) L^\dagger,$$

and $\mathbf{x}_0 = \left(A(I_n - L^\dagger L)\right)^\dagger \mathbf{b}$ is the component of the regularized solution in $\mathcal{N}(L)$. Note that, if $p \geq n$ then $L_A^\dagger = L^\dagger$; if $p = n$ then $L_A^\dagger = L^{-1}$ and $\mathbf{x}_0 = \mathbf{0}$.

1.2.1 Rank-deficient problems

In this section, we discuss numerical methods that are suited for the solution of problems with a numerically rank-deficient coefficient matrix A, i.e., problems for which there is a well-determined gap between the large and small singular values of A.

The rank of a matrix A is defined as the number of linearly independent columns (or rows) of A. It is immediate to prove that the rank is equal to the number of non-zero singular values of A: $r = \text{rank}(A)$ means that $\sigma_r > 0$ and $\sigma_{r+1} = 0$. The matrix A has *full rank* only if all of its singular values are non-zero. In applications, the presence of errors of various kind (measurement errors, approximation and discretization errors, as well as rounding errors) makes this definition not appropriate. Columns of A that are theoretically strictly linearly independent, may result to be linearly dependent from a numerical point of view. To correctly describe the situation, it is necessary to introduce the concept of numerical rank: it is the number of rows or columns of A that are linearly independent with respect to some error level.

The **numerical ϵ-rank** r_ϵ of a matrix A, with respect to the tolerance ϵ, is defined by

$$r_\epsilon = \min_{\|E\| \leq \epsilon} \text{rank}(A + E).$$

It is equal to the number of columns of A that are guaranteed to be linearly independent for any perturbation E of A with norm lesser than or equal to the tolerance ϵ. In terms of the singular values of A, the numerical ϵ-rank satisfies

$$\sigma_{r_\epsilon} > \epsilon \geq \sigma_{r_\epsilon+1}.$$

The definition of r_ϵ is satisfactory only when there is a well-determined gap between σ_{r_ϵ} and $\sigma_{r_\epsilon+1}$.

Associated with the numerical rank $r_\epsilon = k$ are the *numerical null space* and the *numerical range* of A, defined as

$$\mathcal{N}_k(A) = \text{span}\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}, \tag{1.17}$$
$$\mathcal{R}_k(A) = \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}. \tag{1.18}$$

Truncated SVD. In the ideal setting, without perturbations and rounding errors, the treatment of rank-deficient least-squares problems is easy: simply ignore the SVD components associated with the zero singular values and compute the solution by means of (1.9).

In practice, A is never exactly rank-deficient, but instead numerically rank-deficient, i.e., it has one or more small non-zero singular values such that $r_\epsilon < \text{rank}(A) = \min(m, n)$. The small singular values give rise to difficulties. Indeed the norm of the solution is given by

$$\|\mathbf{x}_{\text{LS}}\|^2 = \sum_{i=1}^{n} \left(\frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \right)^2,$$

Hence, the norm is very large due to the small σ_i, unless the coefficients $\mathbf{u}_i^T \mathbf{b}$ satisfy $|\mathbf{u}_i^T \mathbf{b}| < \sigma_i$, for $i = r_\epsilon + 1, \ldots, n$. This requirement is very unlikely to be satisfied, whenever errors are present in $\mathbf{b}$. To deal with this phenomenon, it is necessary to introduce the Picard condition for the discretized problem.

Discrete Picard condition. Let τ denote the value at which the singular values σ_i level off due to rounding errors. The discrete Picard condition is satisfied if, for all singular values larger than τ, the corresponding coefficients $|\mathbf{u}_i^T \mathbf{b}|$, on average, decay faster than the σ_i.

In Figure 1.2 we plotted the singular values, the coefficients, and the ratios between them for the `shaw` test problem [58]. On the left pane, without noise, the singular values σ_i and the coefficients $|\mathbf{u}_i^T \mathbf{b}|$ both level off at the machine precision. On the contrary, if we introduce the noise on the data, that is, $\mathbf{b} = \mathbf{b} + 10^{-5}\mathbf{w}$, where $\mathbf{w}$ is a normally distributed random vector, the right part of the figure shows that the coefficients $|\mathbf{u}_i^T \mathbf{b}|$ level off at the noise level 10^{-5} and only ≈ 11 SVD components are reliable.

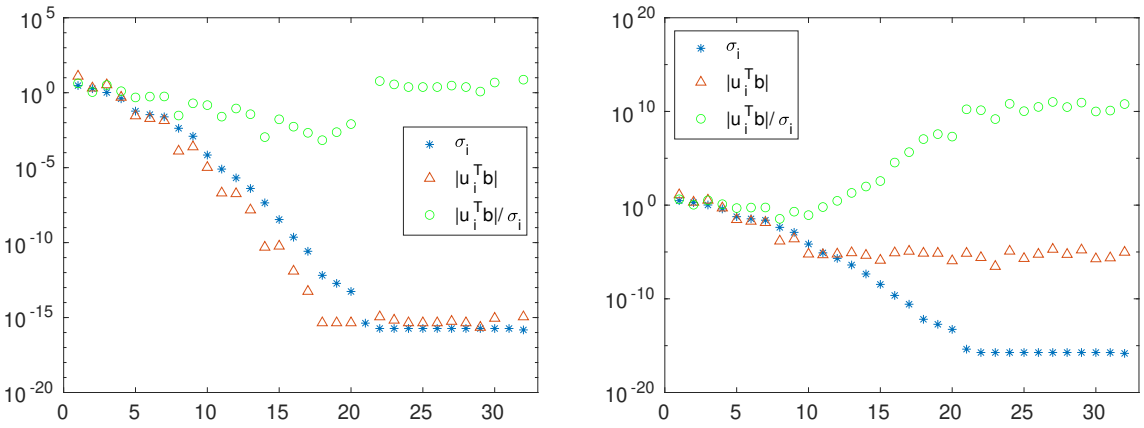

Figure 1.2: The 32 singular values σ_i of A (asterisks), the coefficients $|\mathbf{u}_i^T\mathbf{b}|$ (triangles), and the ratios $|\mathbf{u}_i^T\mathbf{b}|/\sigma_i$ (circles) for the `shaw` test problem. On the left: without noise. On the right: with noise level 10^{-5}.

The most common approach to the regularization of numerically rank-deficient problems is to replace A with a matrix that is close to A and mathematically rank-deficient. The standard choice is the matrix A_k with $\operatorname{rank}(A_k) = k$ defined as

$$A_k = \sum_{i=1}^{k} \mathbf{u}_i \sigma_i \mathbf{v}_i^T,$$

i.e., we set to zero the small non-zero singular values $\sigma_{k+1}, \ldots, \sigma_n$. Among all rank-$k$ matrices Z_k, the matrix A_k minimizes both the 2-norm and the Frobenius norm of the difference $A - Z_k$, i.e., A_k is the best rank-k approximation of A; see Theorem 1.1.1.

It is natural to choose the rank k of A_k as the numerical ϵ-rank of A, i.e., $k = r_\epsilon$, because $k < r_\epsilon$ leads to loss of information associated with large singular values, while $k > r_\epsilon$ leads to a solution with large norm. Unfortunately choosing ϵ is not easy.

When the matrix A is replaced by A_k, then we obtain a new least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A_k\mathbf{x} - \mathbf{b}\|.$$

The corresponding solution is given by

$$\mathbf{x}_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i,$$

which is the *truncated SVD* solution. The method is referred to as *truncated SVD* (TSVD) [55]. The truncation parameter k should be chosen such that all the noise-dominated SVD coefficients are discarded. A suitable value of k often can be found from an inspection of the Picard plot, as we have seen above.

Since the TSVD solution $\mathbf{x}_k$ is a regularized solution with minimal-norm, it is connected with regularization in standard form, i.e., with the regularization term $\|\mathbf{x}\|$. However, it is common in regularization problems to use the more general constraint $\|L\mathbf{x}\|$. To deal with such problems we can use a standard-form transformation to compute the matrix $\widehat{A}$ and the corresponding right-hand side $\widehat{\mathbf{b}}$, and then apply the TSVD method to $\widehat{A}$ and $\widehat{\mathbf{b}}$. After some computation (see [57, Section 3.2]), the solution is given by

$$\mathbf{x}_{L,k} = \sum_{i=p-k+1}^{p} \frac{\mathbf{u}_i^T\mathbf{b}}{c_i}\mathbf{w}_i + \sum_{i=p+1}^{n} (\mathbf{u}_i^T\mathbf{b})\mathbf{w}_i,$$

where c_i and $\mathbf{w}_i$ are elements deriving from the GSVD of the matrix pair (A, L); see (1.3) and (1.5). It is referred to as the *truncated GSVD* (TGSVD) solution.

Tikhonov regularization can also be successfully applied to rank-deficient problems, despite the fact that this method does not seem to involve the numerical rank of the matrix.

1.2.2 Problems with ill-determined rank

In this section, we see how to regularize discrete ill-posed problems. The main feature of these problems is that all the singular values of the coefficient matrix decay gradually to zero, with no gap. Obviously, in this case, the numerical rank of the matrix is not well-determined. Problems which yield matrices that lack a well-determined numerical rank are often discretizations of continuous ill-posed problems.

Also in this case, the last SVD components of the solution, corresponding to small singular values, are dominated by the errors and they should be filtered out in the regularized solution. It is useful to introduce the concept of filter factors.

If A has full rank then we can always write the regularized solution $\mathbf{x}_{\text{reg}}$ in terms of the SVD (or GSVD). Specifically, for regularization methods in standard form with $L = I_n$, the regularized solution can be written in terms of the SVD of A as

$$\mathbf{x}_{\text{reg}} = \sum_{i=1}^{n} f_i \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i,$$

where the elements f_i are called the *filter factors*. The filter factors are typically close to 1 for a large σ_i and much smaller than 1 for a small σ_i. In this way, the contributions to the regularized solution corresponding to the smaller singular values are effectively filtered out.

Similarly, for regularization methods in general form with $L \neq I_n$, we can write the regularized solution in terms of the GSVD of (A, L) as

$$\mathbf{x}_{L,\text{reg}} = \sum_{i=1}^{p} f_i \frac{\mathbf{u}_i^T\mathbf{b}}{c_i}\mathbf{w}_i + \sum_{i=p+1}^{n} (\mathbf{u}_i^T\mathbf{b})\mathbf{w}_i.$$

The difference between various regularization methods with the same L-matrix lies in the way the filter factors are defined. For example, the filter factors for the TSVD and TGSVD methods are

$$f_i = \begin{cases} 1, & i \le k, \\ 0, & i > k, \end{cases} \qquad f_i = \begin{cases} 0, & i \le n-k, \\ 1, & i > n-k, \end{cases}$$

respectively, where k is the truncation parameter. For the expression of the filter factors for Tikhonov regularization, the reader is referred to the next paragraph.

Tikhonov regularization. The idea in Tikhonov's method is to incorporate a priori assumptions about the size and smoothness of the desired solution. For discrete ill-posed problems, Tikhonov regularization in general form leads to the minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \|A\mathbf{x} - \mathbf{b}\|^2 + \lambda^2 \|L\mathbf{x}\|^2 \right\}, \tag{1.19}$$

where the regularization parameter λ controls the weight given to minimization of the regularization term, relative to the minimization of the residual norm. The Tikhonov problem (1.19) has two important alternative formulations:

$$(A^T A + \lambda^2 L^T L)\mathbf{x} = A^T \mathbf{b} \quad \text{and} \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\| \begin{bmatrix} A \\ \lambda L \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} \right\|.$$

The assumption $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$ leads to a unique Tikhonov solution

$$\mathbf{x}_{L,\lambda} = (A^T A + \lambda^2 L^T L)^{-1} A^T \mathbf{b}.$$

The ill-conditioning of A is bypassed by introducing a new problem with a new well-conditioned coefficient matrix $\left[\begin{smallmatrix} A \\ \lambda L \end{smallmatrix}\right]$ with full rank.

By inserting the SVD of A or the GSVD of (A, L), depending on whether the regularization method is in standard or in general form, into the above equation, it can be shown that the filter factors for Tikhonov regularization are

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \quad \text{if } L = I_n, \quad i = 1, \ldots, n,$$

and

$$f_i = \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2} \quad \text{if } L \ne I_n, \quad i = 1, \ldots, p,$$

where σ_i are the singular values of A (see (1.2)) and γ_i are the generalized singular values of (A, L) (see (1.3)). If the regularization term includes an a priori estimate $\overline{\mathbf{x}}$ of the desired solution, as in (1.16), the formulation takes the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\| \begin{bmatrix} A \\ \lambda L \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \lambda L \overline{\mathbf{x}} \end{bmatrix} \right\|.$$

Other regularization methods. Regularization methods in norms different from the 2-norm are also important. For example, the ℓ_p-ℓ_q minimization problem takes the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{p}\|A\mathbf{x}-\mathbf{b}\|_p^p + \frac{\lambda}{q}\|L\mathbf{x}\|_q^q \right\},$$

where $0 < p, q < \infty$. The interested reader can consult [14, 15, 16, 17, 18, 67, 73] and the references therein for a detailed discussion.

In image processing, the *total variation* (TV) functional is useful as a measure of the "size" of the regularized solution. For a one-dimensional function u defined on an interval $[a, b] \subset \mathbb{R}$, the TV functional is

$$\mathcal{J}_{\mathrm{TV}}(u) = \int_a^b \left| \frac{du}{dt} \right| \, dt,$$

and in multidimensional case is

$$\mathcal{J}_{\mathrm{TV}}(u) = \int_\Omega |\nabla u| \, d\Omega,$$

where $\Omega \subseteq \mathbb{R}^n$ is a bounded open set and ∇u is the gradient of u.

1.2.3 Choosing the regularization parameter

A regularization method, to be complete, must include a method for choosing the regularization parameter, either the continuous parameter λ or the discrete parameter k. In this section, we discuss several parameter-choice methods.

Most of the parameter-choice methods are based on residual norms and, in the case of the L-curve, also on the (semi)norm of the solution.

Parameter-choice methods can be divided into two classes depending on their assumptions about the error norm $\|\mathbf{e}\|$:

- a posteriori methods based on the knowledge, or on a good estimate, of $\|\mathbf{e}\|$, like the discrepancy principle;
- methods that do not require $\|\mathbf{e}\|$, but instead seek to extract this information from the given right-hand side, sometimes called heuristic methods, such as the L-curve criterion. For an analysis of other heuristic methods see [66, 81].

The interested reader can see [87] for a review of parameter-choice methods.

Discrepancy principle. The best known $\|\mathbf{e}\|$-based method is the *discrepancy principle*, introduced by Morozov [77]. The idea is to choose the regularization parameter such that the residual norm equals the "discrepancy" in the data, as measured by $\tau\|\mathbf{e}\|$, where $\tau > 1$ is a constant independent of $\|\mathbf{e}\|$. For a discrete

regularization parameter k, the relation can not be satisfied exactly, so instead, we choose the smallest k such that

$$\|A\mathbf{x}_k - \mathbf{b}\| \leq \tau\|\mathbf{e}\|, \qquad \tau > 1.$$

The same strategy applies to the Tikhonov method and here the equality can be obtained:

$$\|A\mathbf{x}_\lambda - \mathbf{b}\| = \tau\|\mathbf{e}\|.$$

Generalized cross-validation. (GCV) [44] is a popular $\|\mathbf{e}\|$-free method for choosing the regularization parameter. The GCV method is based on statistical considerations, namely, that a good value of the regularization parameter should predict missing data values.

The GCV method seeks to minimize the residual error $\|A\mathbf{x}_\lambda - \mathbf{b}_{\text{exact}}\|$. Since $\mathbf{b}_{\text{exact}}$ is unknown, the GCV method chooses a regularization parameter that minimizes the *GCV function*

$$\mathcal{G}(\lambda) = \frac{\|A\mathbf{x}_\lambda - \mathbf{b}\|^2}{(\text{Trace}(I_m - A(\lambda)))^2},$$

where the *influence matrix* for Tikhonov regularization is defined by

$$A(\lambda) = A(A^TA + \lambda^2 L^TL)^{-1}A^T.$$

In Figure 1.3 we represent the GCV function.

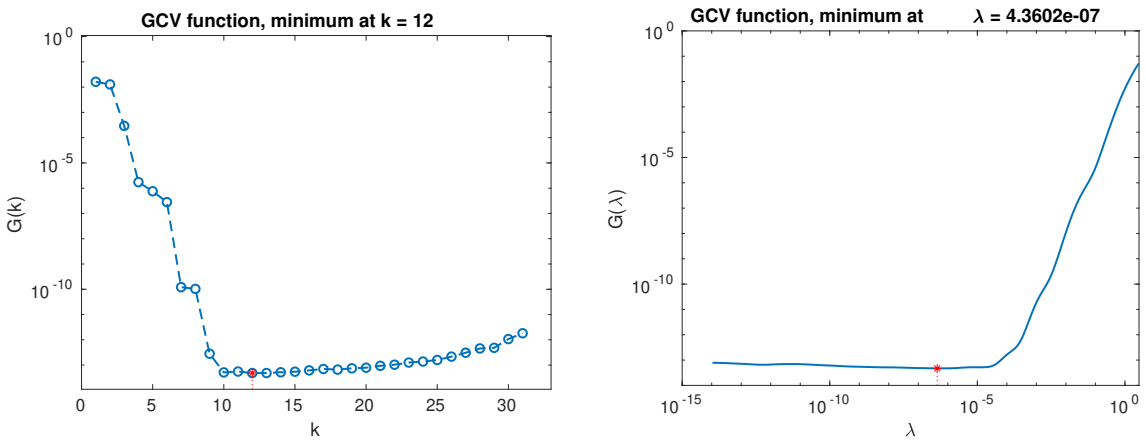

Figure 1.3: Representation of the GCV function. The minimum corresponds to the regularization parameter for TSVD (left) and for Tikhonov regularization (right).

L-curve. Here we discuss another $\|\mathbf{e}\|$-free parameter-choice method. It is based on the so-called *L-curve* [56, 59, 60], which is a plot of the logarithm of (semi)norm $\|L\mathbf{x}_{\text{reg}}\|$ of the regularized solution versus the logarithm of corresponding residual

norm $\|A\mathbf{x}_{\text{reg}} - \mathbf{b}\|$. It is also perhaps the most useful graphical tool for analysis of discrete ill-posed problems. The L-curve displays the compromise between the minimization of these two quantities, which is the heart of any regularization method. The L-curve is a continuous curve when the regularization parameter is continuous, as in Tikhonov regularization. For regularization methods with a discrete regularization parameter, such as TSVD, the L-curve consists of a discrete set of points

$$(\log \|A\mathbf{x}_{\text{reg}} - \mathbf{b}\|, \log \|L\mathbf{x}_{\text{reg}}\|) .$$

We mention that, for different regularization algorithms, it can sometimes be advantageous to plot the L-curve. In these cases, different norms, seminorms, and other measures of "size" of the regularized solution, instead of $\|L\mathbf{x}_{\text{reg}}\|$, are plotted on the y-axis of the graph in log-scale. The name of the curve derives from the characteristic L-shaped appearance it often assumes when plotted in doubly logarithmic scale; see Figure 1.4.

The error $\mathbf{x}_{\text{exact}} - \mathbf{x}_{\text{reg}}$ consists of two components, namely, the perturbation error and the regularization error. When very little regularization is introduced, most of the filter factors are approximately 1, and the error is dominated by the perturbation error. This situation corresponds to the uppermost part of the L-curve above the middle "corner". When a large amount of regularization is introduced, then most filter factors are small, $f_i \ll 1$, and the error is dominated by the regularization error. This situation corresponds to the rightmost part of the L-curve to the right of the "corner". There is an optimal regularization parameter that balances the perturbation error and the regularization error. The idea is to determine it near the "corner" of this L-curve. It also represents a compromise between a small residual and a small (semi)norm of the solution.

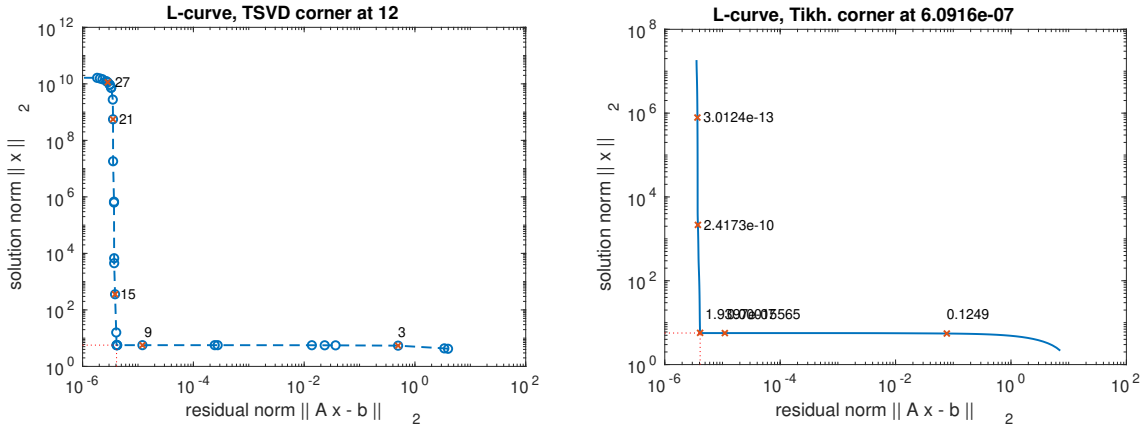

Figure 1.4: Representation of the L-curve. The corner corresponds to the regularization parameter of TSVD (left) and of Tikhonov regularization (right).

1.3 Linear operators and integral equations

In this section, we review some concepts of functional analysis related to bounded linear operators in Hilbert spaces. For a more complete discussion of linear operators and their properties, such as boundedness and compactness, the reader is referred to [13, 50, 72, 101].

A **Hilbert space** is a vector space H equipped with an inner product such that H is complete for the norm $\|u\|_H = \langle u, u\rangle^{1/2}$.

An example of Hilbert space is the set of square-integrable functions $L^2([a,b])$ with inner product

$$\langle f, g\rangle = \int_a^b f(x)g(x)\,dx.$$

We recall that two elements φ and ψ of a Hilbert space H are called *orthogonal* if $\langle \varphi, \psi\rangle = 0$. Let U be a subspace of H. The set

$$U^\perp = \{\psi \in H : \psi \perp U\}$$

is called the *orthogonal complement* of U.

Theorem 1.3.1. *[50, Theorem 3.1.4] If U is a closed subspace of a Hilbert space H, then H can be written as the direct sum of U and its orthogonal complement $U^\perp$, denoted $H = U \oplus U^\perp$, meaning that each $\varphi \in H$ can be written uniquely as $\varphi = \varphi_1 + \varphi_2$, where $\varphi_1 \in U$ and $\varphi_2 \in U^\perp$.*

Now, we recall some basic facts about linear operators on Hilbert spaces.

Definition 1.3.2. *A linear operator $A : X \to Y$ from a normed space X into a normed space Y is called* bounded *if there exists a positive number C such that*

$$\|A\varphi\|_Y \leq C\|\varphi\|_X$$

for all $\varphi \in X$.

Theorem 1.3.3. *[72, Theorem 2.6] Each linear operator $A : X \to Y$ from a finite-dimensional normed space X into a normed space Y is bounded.*

If H_1 and H_2 are Hilbert spaces, we denote the space of all bounded linear operators from H_1 into H_2 by $B(H_1, H_2)$.

An example of bounded linear operator is the *integral operator* K defined on $L^2([a,b])$ space by

$$(Kf)(s) := \int_a^b k(s,t)f(t)\,dt,$$

where the function $k(s,t) \in L^2([a,b]\times[a,b])$ is called the *kernel* of the operator K. Moreover, K is a compact operator.

The next result is known as the Riesz Representation Theorem.

Theorem 1.3.4. *[72, Theorem 4.8] Let $F : H \to \mathbb{R}$ be a bounded linear functional on a Hilbert space H. Then, there exists a unique $f \in H$ such that*

$$F(\varphi) = \langle \varphi, f \rangle$$

for all $\varphi \in H$.

Definition 1.3.5. *Let H_1 and H_2 be Hilbert spaces and $A : H_1 \to H_2$ is a linear operator. The* null space *of A is the set $\mathcal{N}(A) = \{\varphi \in H_1 : A\varphi = 0\}$. The* range *of A is the set $\mathcal{R}(A) = \{\psi \in H_2 : \psi = A\varphi$ for some $\varphi \in H_1\}$.*

$\mathcal{N}(A)$ and $\mathcal{R}(A)$ are subspaces of H_1 and H_2, respectively. If A is bounded, then $\mathcal{N}(A)$ is a closed subspace.

Definition 1.3.6. *Let H_1 and H_2 be Hilbert spaces. Given $A \in B(H_1, H_2)$, the unique linear operator $A^* \in B(H_2, H_1)$ satisfying*

$$\langle A\varphi, \psi \rangle = \langle \varphi, A^*\psi \rangle$$

for all $\varphi \in H_1$ and $\psi \in H_2$ is called the adjoint *of A.*

Here we present results which relate the concepts of range, null space and, adjoint.

Theorem 1.3.7. *[50, Theorem 3.3.2] If $A \in B(H_1, H_2)$, then*

$$\mathcal{R}(A)^\perp = \mathcal{N}(A^*), \quad \mathcal{N}(A)^\perp = \overline{\mathcal{R}(A^*)},$$
$$\mathcal{R}(A^*)^\perp = \mathcal{N}(A), \quad \mathcal{N}(A^*)^\perp = \overline{\mathcal{R}(A)}.$$

Two of the most standard forms of **linear integral equations** are

$$\int_a^b k(x, y) f(y)\, dy = g(x) \tag{1.20}$$

and

$$f(x) - \gamma \int_a^b k(x, y) f(y)\, dy = g(x), \tag{1.21}$$

where γ is a given non-zero constant, $c \le x \le d$, and $a \le y \le b$. Both, the function $k(x, y)$, called the kernel, and the function $g(x)$ are known. The function f is to be determined. Equation (1.20) is a linear *Fredholm integral equation of the first kind*, while equation (1.21) is *of the second kind*. In the first equation the unknown function only occurs under the integral whereas in the second equation it also appears outside the integral. In terms of the compact operator K, the above equations may be written as

$$Kf = g,$$

and

$$f - \gamma K f = g,$$

respectively. We have seen in Section 1.2 that integral equations of the first kind are ill-posed.

In an applicative context, the function g in (1.20) often represents experimental data, measurable only at a finite set of points x_i. Frequently these points are few, and the data usually contain errors.

An important tool for the numerical solution of integral equations is provided by the quadrature formulae. Numerical integration formulae, or **quadrature formulae**, are methods for the approximate evaluation of definite integrals. They are needed for the computation of those integrals for which either the primitive function of the integrand cannot be expressed in terms of elementary functions or for which the integrand is available only at discrete points.

Let f be a continuous function over the interval $[a, b]$. A quadrature formula is a numerical method for approximating a definite integral

$$I(f) := \int_a^b f(x)\, dx$$

by a weighted sum

$$I_n(f) := \sum_{k=0}^{n} a_k f(x_k) \tag{1.22}$$

with $n+1$ distinct *quadrature points* $x_0, \ldots, x_n \in [a, b]$ and $n+1$ *quadrature weights* $a_0, \ldots, a_n \in \mathbb{R}$.

An important group of quadrature formulae, called *polynomial interpolatory quadrature*, is obtained by integrating an interpolating polynomial instead of the integrand f, i.e., by approximating $f(x) \approx (L_n f)(x)$, where $L_n : C([a, b]) \to P_n$ denotes the polynomial interpolation operator with $n + 1$ interpolation points $x_0, \ldots, x_n$, $C([a, b])$ is the space of continuous functions on the interval $[a, b]$, and P_n is the space of polynomials $p(x) = \sum_{k=0}^{n} a_k x^k$. A polynomial interpolatory quadrature formula takes the form

$$I_n(f) = \int_a^b (L_n f)(x)\, dx. \tag{1.23}$$

Expressing the interpolating polynomial in the form of Lagrange

$$(L_n f)(x) = \sum_{k=0}^{n} f(x_k) L_k(x), \qquad \text{with } L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j},$$

we obtain $I_n(f)$ in the form (1.22) with the weights given by

$$a_k = \int_a^b L_k(x)\, dx.$$

Theorem 1.3.8. *[71, Theorem 9.2] Given $n+1$ distinct quadrature points $x_0, \dots, x_n \in [a,b]$, the interpolatory quadrature (1.23) of order n is uniquely determined by its property of integrating all polynomials $p \in P_n$ exactly, i.e.,*

$$I_n(p) = I(p).$$

The polynomial interpolatory quadrature with equidistant quadrature points $x_k = a + kh$, for $k = 0, \dots, n$, and step width $h = (b-a)/n$ is called the *Newton–Cotes quadrature formula*. Details and properties, like the expression of the weights and the errors, can be found in [71].

Gaussian quadrature formulae. We just saw that given $n+1$ arbitrary quadrature points, the quadrature weights of a polynomial interpolatory quadrature are determined such that all polynomials of degree less than or equal to n are integrated exactly. Now we see that by choosing the quadrature points in a certain way, it is possible to construct a quadrature formula with precision equal to $2n+1$.

To achieve this degree of exactness the quadrature points and the quadrature weights have to satisfy the conditions

$$\sum_{k=0}^{n} a_k x_k^i = \int_a^b x^i \, dx, \qquad i = 0, \dots, 2n+1.$$

We shall proceed with a more general treatment considering quadrature formulae for the integral

$$I(f) := \int_a^b w(x) f(x) \, dx,$$

where w denotes some *weight function*.

Definition 1.3.9. *A quadrature formula with $n+1$ distinct quadrature points is called a* Gaussian quadrature formula *if it integrates all polynomials $p \in P_{2n+1}$ exactly.*

For an in-depth discussion, the reader is referred to [24, 71]. Here we report some properties without proof.

The quadrature points of the Gauss quadrature are given by the $n+1$ distinct zeros in (a,b) of the polynomial p_{n+1} of order $n+1$ that is orthogonal to all polynomials $p \in P_n$ of degree less than $n+1$ with respect to the scalar product $\langle f, g \rangle = \int_a^b w(x) f(x) g(x) \, dx$, that is $\langle p_{n+1}, p \rangle = 0$. The weights of the Gaussian quadrature formulae are all positive. As for the error of the Gaussian formulas, the following theorem holds.

Theorem 1.3.10. *[71, Theorem 9.20] Let $f \in C^{2n+2}[a,b]$, the space of $(2n+2)$-times continuously differentiable functions. Then the error for the Gaussian quadrature*

formula of order n is given by

$$\int_a^b w(x)f(x)\,dx - \sum_{k=0}^{n} a_k f(x_k) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b w(x)[p_{n+1}(x)]^2\,dx,$$

for some $\xi \in [a,b]$.

Gauss–Legendre quadrature. Now, we consider the weight function $w(x) = 1$ in the interval $[-1,1]$. The *Legendre polynomial* L_n of degree n is defined by

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \qquad L_n \in P_n.$$

If $m < n$, the following relation is verified

$$\int_{-1}^{1} x^m \frac{d^n}{dx^n}(x^2-1)^n\,dx = 0.$$

Therefore L_n is orthogonal to all polynomials of degree less than n with respect to the scalar product on $[-1,1]$. The quadrature points of the Gauss–Legendre quadrature are given by the n zeros $-1 < x_1 < x_2 < \cdots < x_n < 1$ of L_n and the corresponding quadrature weights can be computed as

$$a_k = \frac{2(1-x_k^2)}{[nL_{n-1}(x_k)]^2}, \quad k = 1,\ldots,n.$$

From the Gauss–Legendre quadrature formula for the interval $[-1,1]$, we can obtain the quadrature formula for an arbitrary interval $[a,b]$

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} \lambda_k f(t_k),$$

by setting

$$t_k = a + \frac{b-a}{2}(x_k+1), \qquad \lambda_k = \frac{b-a}{2}a_k, \qquad k = 1,\ldots,n, \tag{1.24}$$

where $x_k \in (-1,1)$ are the zeros of the Legendre polynomial in $[-1,1]$, and a_k the corresponding weights.

Other types of orthogonal polynomials provide useful Gaussian quadrature formulas when the integrand has a particular form. Some of these are tabulated in Table 1.1.

Table 1.1: Gaussian quadrature formulae for particular weight functions.

$[a, b]$	$w(x)$	orthogonal polynomials
$[-1, 1]$	1	Legendre
$[-1, 1]$	$(1-x)^\alpha(1+x)^\beta, \quad \alpha, \beta > -1$	Jacobi
$[-1, 1]$	$\dfrac{1}{\sqrt{1-x^2}}$	Chebyshev (first kind)
$[-1, 1]$	$\sqrt{1-x^2}$	Chebyshev (second kind)
$[0, \infty)$	e^{-x}	Laguerre
$[0, \infty)$	$x^\alpha \mathrm{e}^{-x}, \quad \alpha > -1$	Generalized Laguerre
$(-\infty, +\infty)$	e^{-x^2}	Hermite

CHAPTER 2

Minimal-norm Gauss–Newton method

2.1 Introduction

Let us assume that $F(\mathbf{x}) = [F_1(\mathbf{x}), \ldots, F_m(\mathbf{x})]^T$ is a nonlinear twice continuously Fréchet-differentiable function, with values in $\mathbb{R}^m$ for any $\mathbf{x} \in \mathbb{R}^n$. For a given $\mathbf{b} \in \mathbb{R}^m$, we consider the nonlinear least-squares data fitting problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{r}(\mathbf{x})\|^2, \qquad \mathbf{r}(\mathbf{x}) = F(\mathbf{x}) - \mathbf{b}, \tag{2.1}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\mathbf{r}(\mathbf{x}) = [r_1(\mathbf{x}), \ldots, r_m(\mathbf{x})]^T$ is the residual vector function between the model expectation $F(\mathbf{x})$ and the vector $\mathbf{b}$ of measured data. The solution to the nonlinear least-squares problem gives the best model fit to the data in the sense of the minimum sum of squared errors. Classical approaches to the numerical solution of a nonlinear least-squares problem consist of applying Newton's method and its variants such as the Gauss–Newton method [10, 61, 78].

Linear least-squares problems have been widely studied; an exhaustive review can be found in [10]. There also exists a vast literature concerning regularization methods for discrete linear inverse problems; see [35, 57]. The same references discuss numerical methods for the solution of nonlinear least-squares problems, as well as suitable regularization techniques.

The Gauss–Newton method and its variants have been investigated in many papers; see, e.g., [21, 54, 69, 75, 90]. The application of the Levenberg–Marquardt method to ill-posed problems was studied in [22, 64], and in [53] it was applied to an inverse problem in groundwater hydrology. In [84], an iterative algorithm based on the minimization of the Tikhonov functional by the gradient method was developed. The application of Tikhonov regularization to nonlinear inverse problems has been further investigated in [74, 85]. The case where the regularizing term is substituted

by a penalty term which promotes the selection of a sparse solution was analyzed in [86].

At the kth step of the Gauss–Newton method, the current approximation is computed by solving, in the least-squares sense, a linearization of the original nonlinear problem. When the Jacobian of the residual function does not have full column rank, the solution is not unique, and the usual approach is to select the one having minimal-norm. This ensures that each update of the solution of the nonlinear least-squares problem has a minimal-norm, but this property does not apply to the solution itself. The same is true when a regularization technique is introduced.

The idea of constructing an iterative method for the computation of the minimal-norm solution of a nonlinear least-squares problem was first studied by Eriksson et al. In [36, 37, 38], the case where the Jacobian is rank-deficient or ill-conditioned was analyzed, and the solution techniques based on the Gauss–Newton method and on Tikhonov regularization in standard form were proposed.

In this chapter and in [82], we review the results obtained by Eriksson et al. and extend them by introducing the minimization of the seminorm $\|L\mathbf{x}\|$, where L is a regularization matrix; see equation (2.14). In case of lack of a unique solution, the employment of such a seminorm is often essential to select an effective reconstruction when suitable a priori information is available. We further analyze the computation of the regularized minimal-L-norm solution by two standard procedures for approximating the solution of ill-conditioned nonlinear least-squares problems, namely, the truncated generalized singular value decomposition (TGSVD) applied to the Gauss–Newton method, and Tikhonov regularization in general form, whose solutions are given by (2.27) and (2.37), respectively. Though the two regularized solutions are different, they both converge to the minimal-L-norm solution when the regularization level decreases. The algorithms are applied to a small-scale test problem and to the inversion of a medium-size nonlinear model typical in applied geophysics. The numerical results are compared to those produced by the classical approaches.

To ensure the computation of the minimal-norm solution [36, 37, 38, 82], at the kth iteration, the Gauss–Newton approximation is orthogonally projected onto the null space of the Jacobian matrix. Unfortunately, the algorithms developed in the above papers occasionally lack to converge. They take the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}^{(k)} - \mathcal{P}_{\mathcal{N}(J_k)} \mathbf{x}^{(k)},$$

where $\widetilde{\mathbf{s}}^{(k)}$ is the solution of (2.9), α_k is a step length, and $\mathcal{P}_{\mathcal{N}(J_k)}$ is the orthogonal projector onto the null space of $J_k = J(\mathbf{x}^{(k)})$. In [82], the damping parameter α_k is estimated by the Armijo–Goldstein principle; we refer to this method as the MNGN algorithm. One reason for the non-convergence of such methods is that the projection step may cause the residual to increase considerably at particular iterations. Moreover, the rank of the Jacobian may vary as the iteration progresses, and its incorrect estimation often leads to the presence of small singular values for the Jacobian, which amplify computational errors.

This problem of non-convergence has been considered by Campbell, Kunkel, Bobinyec in [20] using a method which will be denoted CKB in the following. The authors consider a convex combination of the Gauss–Newton approximation and its orthogonal projection, and apply a relaxation parameter γ_k to this search direction, chosen according to a given rule. After some manipulation, the method can be written as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)} - \gamma_k \mathcal{P}_{\mathcal{N}(J_k)} \mathbf{x}^{(k)}.$$

This approach makes the computation of the minimal-norm solution more robust, but it may not converge in some situation; see Section 2.9. Moreover, both the MNGN and the CKB methods suffer from serious convergence problems caused by the variation of the rank of the Jacobian along the iterations. The rank often drops to a small value in a neighborhood of the solution, while the two methods consider a fixed rank, generally assumed to be the smaller dimension of the Jacobian.

In the second part of this chapter, we aim at improving the convergence of the methods presented in [20] and [82]. We do this by first introducing in the MNGN method a technique to estimate the rank of the matrix $J(\mathbf{x}^{(k)})$ at each iteration. This procedure has the effect of improving the convergence of the method, reducing the possibility that the iteration diverges because of error amplification. Then, we introduce a second relaxation parameter for the projection term, as well as a strategy to automatically tune it, besides the usual damping parameter for the Gauss–Newton search direction. This approach produces, on the average, solutions closer to optimality, i.e., with smaller norms, than those computed by the CKB method. Furthermore, we consider a model profile $\overline{\mathbf{x}}$ for the solution, which is useful in applications where sufficient a priori information on the physical system under investigation is available.

The chapter can be divided into two parts. The content of the first part is based on our work [82], while the second part is based on our paper [83]. The first part is organized as follows: Section 2.2 recalls Newton and Gauss–Newton methods as well as some basic computational tools. In Sections 2.3 we review the results from Eriksson et al. on the computation of the minimal-norm solution to a nonlinear least-squares problem, and in Section 2.4 we extend them for the computation of the minimal-L-norm solution. Two regularization techniques for ill-conditioned problems are introduced in Section 2.5, and we discuss in Section 2.6 some details of our implementation. The reader is referred to Section 6.1 for the results of numerical experiments regarding the M(L)NGN method and the regularized variants. The second part is structured as follows: in Section 2.7, we revise the MNGN method and reformulate Theorem 2.3.1 ([82, Theorem 3.1]) by introducing a model profile for the solution. Then, we give a theoretical justification for the fact that the convergence of the method may not be ensured. Section 2.8 explains how to estimate the numerical rank of the Jacobian $J(\mathbf{x}^{(k)})$ at each iteration. In Section 2.9, we describe an algorithm which introduces a second parameter to control the size of the correction vector that provides the minimal-norm solution, and which estimates

automatically such parameter. In Section 2.10, we extend the discussion to the minimal-L-norm solution, where L is a regularization matrix. Numerical examples regarding the application of the "doubly relaxed" M(L)NGN method can be found in Section 6.2. Section 2.11 contains concluding remarks.

2.2 Mathematical preliminaries

We will rewrite the minimization problem (2.1) as

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \qquad \text{where } f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|^2 = \frac{1}{2}\sum_{i=1}^{m} r_i(\mathbf{x})^2.$$

Let the Jacobian of the residual vector function $\mathbf{r}(\mathbf{x})$ be $J(\mathbf{x}) \in \mathbb{R}^{m\times n}$, defined by

$$[J(\mathbf{x})]_{ij} = \frac{\partial r_i(\mathbf{x})}{\partial x_j}, \qquad i = 1,\dots,m,\ j = 1,\dots,n,$$

and the Hessian matrix of $r_i(\mathbf{x})$ be $\nabla^2 r_i(\mathbf{x}) \in \mathbb{R}^{n\times n}$, $i = 1,\dots,m$, with entries given by

$$[\nabla^2 r_i(\mathbf{x})]_{jk} = \frac{\partial^2 r_i(\mathbf{x})}{\partial x_j \partial x_k}, \qquad j,k = 1,\dots,n.$$

Then, the gradient and the Hessian of $f(\mathbf{x})$, written in matrix form, are given by

$$\nabla f(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = J(\mathbf{x})^T \mathbf{r}(\mathbf{x}), \tag{2.2}$$

and

$$\nabla^2 f(\mathbf{x}) = J(\mathbf{x})^T J(\mathbf{x}) + \mathcal{Q}(\mathbf{x}), \qquad \text{where } \mathcal{Q}(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x})\nabla^2 r_i(\mathbf{x}). \tag{2.3}$$

Indeed, using the chain rule,

$$[\nabla f(\mathbf{x})]_j = \frac{\partial f(\mathbf{x})}{\partial x_j} = \sum_{i=1}^{m} r_i(\mathbf{x})\frac{\partial r_i(\mathbf{x})}{\partial x_j},$$

$$[\nabla^2 f(\mathbf{x})]_{jk} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k} = \sum_{i=1}^{m} \frac{\partial r_i(\mathbf{x})}{\partial x_j}\frac{\partial r_i(\mathbf{x})}{\partial x_k} + \sum_{i=1}^{m} r_i(\mathbf{x})\frac{\partial^2 r_i(\mathbf{x})}{\partial x_j \partial x_k}.$$

If the point $\mathbf{x}^*$ is a local minimum for a twice continuously differentiable function $f(\mathbf{x})$, then $\mathbf{x}^*$ is a stationary point, i.e., the gradient (2.2) of f at $\mathbf{x}^*$ is zero. Conversely, a sufficient condition for a stationary point to be a local minimum is that the Hessian $\nabla^2 f(\mathbf{x}^*)$ is positive definite.

Newton's method for optimization [10] is based on the minimization of the second-order Taylor approximation of the function $f(\mathbf{x})$,

$$\tilde{f}(\mathbf{x}+\mathbf{s}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T\nabla^2 f(\mathbf{x})\mathbf{s}.$$

The minimizer is obtained by equating to zero the derivative with respect to $\mathbf{s}$,

$$\frac{\partial \tilde{f}(\mathbf{x}+\mathbf{s})}{\partial \mathbf{s}} = \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\mathbf{s} = 0.$$

Starting from an initial guess $\mathbf{x}^{(0)}$, assuming that the Hessian of $f(\mathbf{x})$ is invertible in $\mathbf{x}^{(k)}$, and substituting (2.2) and (2.3), the iteration of Newton's method is obtained:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(J_k^T J_k + \mathcal{Q}(\mathbf{x}^{(k)})\right)^{-1} J_k^T \mathbf{r}_k, \tag{2.4}$$

where $J_k = J(\mathbf{x}^{(k)})$ is the Jacobian of F in $\mathbf{x}^{(k)}$ and $\mathbf{r}_k = \mathbf{r}(\mathbf{x}^{(k)})$ is the residual vector. Newton's method is rarely used for nonlinear least-squares problems because computing the mn^2 derivatives appearing in $\mathcal{Q}(\mathbf{x})$ is often computationally too expensive and it is unfeasible for large-scale problems.

Initially, we assume that the problem is *overdetermined*, i.e., $m \geq n$. An alternative to Newton's method is to neglect the term $\mathcal{Q}(\mathbf{x}^{(k)})$ in (2.4), obtaining the **Gauss–Newton method**. If $m \geq n$ and J_k is full rank, then the matrix $J_k^T J_k$ is nonsingular, and we can write

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(J_k^T J_k\right)^{-1} J_k^T \mathbf{r}_k, \qquad k = 0, 1, 2, \ldots$$

In this case, the matrix $J_k^\dagger = \left(J_k^T J_k\right)^{-1} J_k^T$ is the Moore–Penrose pseudoinverse of J_k; see Section 1.1. If $m = n$, then the iteration simplifies to $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - J_k^{-1}\mathbf{r}_k$. For *underdetermined* full rank problems ($m < n$) the iteration of the Gauss–Newton method becomes

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - J_k^T \left(J_k J_k^T\right)^{-1} \mathbf{r}_k.$$

The behavior of the Gauss–Newton method can be expected to be similar to that of Newton's method when the term $\mathcal{Q}(\mathbf{x})$ is negligible, i.e., when the quantities $|r_i(\mathbf{x})|\|\nabla^2 r_i(\mathbf{x})\|$, $i = 1, \ldots, m$, are small compared to $J^T J$, where $J = J(\mathbf{x})$. This happens if the functions $r_i(\mathbf{x})$ are mildly nonlinear in a neighborhood of the solution or if the problem is consistent.

We can give a different characterization of the Gauss–Newton method. It replaces the nonlinear problem by a sequence of linear approximations of $\mathbf{r}(\mathbf{x})$, obtained through a first-order Taylor series expansion. The residual at the new iterate is approximated by

$$\mathbf{r}(\mathbf{x}^{(k+1)}) \simeq \mathbf{r}_k + J_k\mathbf{s}.$$

Chosen an initial point $\mathbf{x}^{(0)}$, if $\mathbf{x}^{(k)}$ denotes the current approximation, then the new approximation is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}, \qquad k = 0, 1, 2, \ldots, \tag{2.5}$$

where the step $\mathbf{s}^{(k)}$ is computed as a solution to the linear least-squares problem

$$\min_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2. \tag{2.6}$$

In order to ensure convergence, (2.5) is replaced by the damped Gauss–Newton method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{s}^{(k)}, \tag{2.7}$$

where the scalar α_k is a step length. We estimate it by the *Armijo–Goldstein principle* [3, 42], but it can be chosen by any strategy which guarantees a reduction in the norm of the residual. In our case, the Armijo condition [3, 30] implies

$$f(\mathbf{x}^{(k)} + \alpha_k\mathbf{s}^{(k)}) \le f(\mathbf{x}^{(k)}) + \mu\alpha_k\nabla f(\mathbf{x}^{(k)})^T\mathbf{s}^{(k)},$$

where μ is a constant in $(0, 1)$. Since $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|^2$ and $\nabla f(\mathbf{x}) = J(\mathbf{x})^T\mathbf{r}(\mathbf{x})$, it reads

$$\|\mathbf{r}(\mathbf{x}^{(k)} + \alpha_k\mathbf{s}^{(k)})\|^2 \le \|\mathbf{r}_k\|^2 + 2\mu\alpha_k\mathbf{r}_k^T J_k\mathbf{s}^{(k)}.$$

Note that, as $\mathbf{s}^{(k)}$ satisfies the normal equations associated to problem (2.6) and $J_kJ_k^\dagger$ is an orthogonal projector, and therefore symmetric, it holds

$$\mathbf{r}_k^T J_k\mathbf{s}^{(k)} = -\mathbf{r}_k^T J_kJ_k^\dagger\mathbf{r}_k = -\mathbf{r}_k^T J_kJ_k^\dagger J_kJ_k^\dagger\mathbf{r}_k = -\|J_kJ_k^\dagger\mathbf{r}_k\|^2 = -\|J_k\mathbf{s}^{(k)}\|^2.$$

The Armijo–Goldstein principle [10, 42] sets $\mu = \frac{1}{4}$ and determines the scalar α_k as the largest number in the sequence 2^{-i}, $i = 0, 1, \ldots$, for which it holds

$$\|\mathbf{r}_k\|^2 - \|\mathbf{r}(\mathbf{x}^{(k)} + \alpha_k\mathbf{s}^{(k)})\|^2 \ge \frac{1}{2}\alpha_k\|J_k\mathbf{s}^{(k)}\|^2. \tag{2.8}$$

The step length α_k may also be determined by solving the minimization problem

$$\min_{\alpha} \|\mathbf{r}(\mathbf{x}^{(k)} + \alpha\mathbf{s}^{(k)})\|^2.$$

In [90], this approach is denoted as *Gauss–Newton algorithm with line search.*

The solution to (2.6) may not be unique: this situation happens when the matrix J_k does not have full column rank, in particular, when $m < n$. To make the solution unique, the new iterate $\mathbf{x}^{(k+1)}$ is often obtained by solving the following minimal-norm linear least-squares problem

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|\mathbf{s}\|^2 \\ \mathbf{s} \in \left\{\arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2\right\}, \end{cases} \tag{2.9}$$

where the set in the lower line contains all the solutions to problem (2.6). Problem (2.9) has the solution

$$\mathbf{s}^{(k)} = -J_k^\dagger\mathbf{r}_k.$$

Such *minimal-norm solution* is orthogonal to the null space $\mathcal{N}(J_k)$ of J_k. This is generally assumed to be a good choice among the infinitely many solutions to the problem unless other constraints for the solution are available.

In order to select solutions with different degrees of regularity, the term $\|\mathbf{s}\|^2$ in (2.9) is sometimes substituted by $\|L\mathbf{s}\|^2$, where $L \in \mathbb{R}^{p\times n}$ $(p \leq n)$ is a matrix which incorporates available a priori information on the solution. Examples of regularization matrices L are reported in equations (1.14) and (1.15). When a regularization matrix is introduced, formulation (2.9) becomes

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|L\mathbf{s}\|^2 \\ \mathbf{s} \in \left\{ \arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2 \right\}. \end{cases} \tag{2.10}$$

It is important to remark that both (2.9) and (2.10) impose some kind of regularity on the update vector $\mathbf{s}$ for the solution $\mathbf{x}^{(k)}$, and not on the solution itself. We will explore in this chapter what the consequence is of imposing a regularity constraint directly on the solution $\mathbf{x}$ of problem (2.1). Approaches of this kind were studied by Eriksson and Wedin [36, 37, 38]: they proposed a minimal-norm Gauss–Newton method and a Tikhonov regularization method in standard form. We extend, in Theorem 2.4.2, the minimal-norm Gauss–Newton method by introducing a regularization matrix L. Moreover, in Section 2.5 we investigate Tikhonov regularization in general form and the use of truncated SVD/GSVD in the minimal-norm Gauss–Newton method.

In Section 2.5.2 we will see that, in the limit, the minimal-norm Gauss–Newton iteration and the iteration obtained through Tikhonov regularization in standard form are closely related; the same happens for the minimal-L-norm Gauss–Newton iteration and Tikhonov regularization in general form.

2.3 Nonlinear minimal-norm solution

Let us discuss the computation of the minimal-norm solution to the nonlinear problem (2.1),

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|^2 \\ \mathbf{x} \in \left\{ \arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} \|F(\mathbf{x}) - \mathbf{b}\|^2 \right\}. \end{cases} \tag{2.11}$$

The problem of imposing a regularity constraint directly on the solution $\mathbf{x}$ of problem (2.1) is studied in [36, 37, 38, 82]. These papers are based on the application of the damped Gauss–Newton method to the solution of (2.11).

We consider the following iterative method of type (2.5), based on a first-order linearization of the problem:

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|\mathbf{x}^{(k)} + \mathbf{s}\|^2 \\ \mathbf{s} \in \left\{ \arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2 \right\}. \end{cases} \tag{2.12}$$

We will denote this as the *minimal-norm Gauss–Newton* (MNGN) method.

A theorem similar to the following one is presented, in a slightly general form, in [36, 37, 38]. We provide here a statement and a proof in terms of the SVD, which is useful from a computational point of view.

Theorem 2.3.1. *Let* $\mathbf{x}^{(k)} \in \mathbb{R}^n$ *and let* $\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)}$ *be the Gauss–Newton iteration for* (2.1)*, where the step* $\widetilde{\mathbf{s}}^{(k)}$ *is determined by solving* (2.9)*. Then, the iteration* $\mathbf{x}^{(k+1)}$ *for* (2.12)*, starting from the same point* $\mathbf{x}^{(k)}$*, is given by*

$$\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - V_2 V_2^T \mathbf{x}^{(k)},$$

where $\operatorname{rank}(J_k) = r_k$ *and the columns of the matrix* $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$ *are orthonormal vectors in* $\mathbb{R}^n$ *spanning the null space of* J_k*.*

Proof. Let $U\Sigma V^T$ be the singular value decomposition of the matrix J_k. The norm of the solution $\mathbf{x}^{(k+1)}$ of (2.12) may be expressed as

$$\|\mathbf{x}^{(k+1)}\|^2 = \|V^T(\mathbf{x}^{(k)} + \mathbf{s})\|^2 = \|\mathbf{y} + \mathbf{z}^{(k)}\|^2,$$

with $\mathbf{y} = V^T\mathbf{s}$ and $\mathbf{z}^{(k)} = V^T\mathbf{x}^{(k)}$. Replacing J_k by its SVD and setting $\mathbf{g}^{(k)} = U^T\mathbf{r}_k$, we can rewrite (2.12) as the following diagonally constrained least-squares problem

$$\begin{cases} \min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{y} + \mathbf{z}^{(k)}\|^2 \\ \mathbf{y} \in \left\{ \arg\min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma\mathbf{y} + \mathbf{g}^{(k)}\|^2 \right\}. \end{cases}$$

Solving the second minimization problem uniquely determines the components $y_i = -\sigma_i^{-1} g_i^{(k)}$, $i = 1, \ldots, r_k$, while the entries y_i, $i = r_k + 1, \ldots, n$, are undetermined. In order to minimize the norm of the solution

$$\|\mathbf{y} + \mathbf{z}^{(k)}\|^2 = \sum_{i=1}^{r_k} \left(-\frac{g_i^{(k)}}{\sigma_i} + z_i^{(k)} \right)^2 + \sum_{i=r_k+1}^{n} \left(y_i + z_i^{(k)} \right)^2,$$

we set $y_i = -z_i^{(k)} = -\mathbf{v}_i^T\mathbf{x}^{(k)}$, $i = r_k + 1, \ldots, n$. The solution to (2.12), that is, the next approximation to the solution of (2.11), is then

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{g_i^{(k)}}{\sigma_i} \mathbf{v}_i - \sum_{i=r_k+1}^{n} (\mathbf{v}_i^T \mathbf{x}^{(k)}) \mathbf{v}_i, \tag{2.13}$$

where the last summation can be written in matrix form as $V_2 V_2^T \mathbf{x}^{(k)}$.

Similarly, we rewrite (2.9) as the following diagonal least-squares problem

$$\begin{cases} \min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{y}\|^2 \\ \mathbf{y} \in \left\{ \arg\min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma\mathbf{y} + \mathbf{g}^{(k)}\|^2 \right\}, \end{cases}$$

with $\mathbf{y} = V^T\mathbf{s}$, obtaining

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)} = \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{g_i^{(k)}}{\sigma_i}\mathbf{v}_i.$$

Then,

$$\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - V_2V_2^T\mathbf{x}^{(k)},$$

where the columns of $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$ are a basis for $\mathcal{N}(J_k)$. This completes the proof. □

Remark 2.3.2. Let $\mathcal{P}_{\mathcal{N}(J_k)}$ represent the orthogonal projector onto $\mathcal{N}(J_k)$. Since $\mathcal{P}_{\mathcal{N}(J_k)} = V_2V_2^T$ (see Section 1.1), the above theorem shows that, unlike the Gauss–Newton method, the $(k+1)$th iterate of the MNGN method is orthogonal to the null space of J_k. Then, equation (2.13) may be expressed in the more general form [36, 37, 38]

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \begin{bmatrix} J_k^\dagger & \mathcal{P}_{\mathcal{N}(J_k)} \end{bmatrix} \begin{bmatrix} \mathbf{r}_k \\ \mathbf{x}^{(k)} \end{bmatrix}.$$

Corollary 2.3.3. *If $\mathbf{x}^{(k)}$ is orthogonal to the null space of J_k, then the solution $\mathbf{x}^{(k+1)}$ of (2.12) is the same as that of (2.9).*

Remark 2.3.4. It is useful to remember that $V_2V_2^T = I_n - V_1V_1^T$ with the matrix $V_1 = [\mathbf{v}_1, \ldots, \mathbf{v}_{r_k}]$. So, the updated solution can be obtained without necessarily computing the singular vectors $\mathbf{v}_i$, $i = r_k + 1, \ldots, n$, i.e., when a compact SVD is available

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)} - V_2V_2^T\mathbf{x}^{(k)} = \widetilde{\mathbf{s}}^{(k)} + V_1V_1^T\mathbf{x}^{(k)}.$$

Remark 2.3.5. In the first numerical example of Section 6.1, the approach of projecting the iterates orthogonally to the null space of J_k will also be applied to Newton's method. This approach is only heuristic in this case. It will be shown that the solution at convergence coincides with the one produced by the MNGN method but that the speed of convergence of Newton's method degrades.

2.4 Nonlinear minimal-L-norm solution

The introduction of a regularization matrix $L \in \mathbb{R}^{p\times n}$, $p \leq n$, in least-squares problems was originally connected to the numerical treatment of linear discrete ill-posed problems, and in particular to Tikhonov regularization. The use of a regularization matrix is also justified in underdetermined least-squares problems to select a solution with particular features, such as smoothness or sparsity, among the infinitely many possible solutions.

Here, we seek to compute the minimal-L-norm solution to the nonlinear problem (2.1), that is the vector $\mathbf{x}$ which solves the constrained problem

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} \|L\mathbf{x}\|^2 \\ \mathbf{x} \in \left\{\arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} \|F(\mathbf{x}) - \mathbf{b}\|^2\right\}. \end{cases} \tag{2.14}$$

Similarly to the previous section, we consider an iterative method of type (2.5), where the step $\mathbf{s}^{(k)}$ is the solution of the linearized problem

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|L(\mathbf{x}^{(k)} + \mathbf{s})\|^2 \\ \mathbf{s} \in \left\{\arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2\right\}. \end{cases} \tag{2.15}$$

We will denote this as the *minimal-L-norm Gauss–Newton* (MLNGN) method.

Let $J_k = U\Sigma_J W^{-1}$, $L = V\Sigma_L W^{-1}$ be the generalized singular value decomposition of the matrix pair (J_k,L). We indicate by $\mathbf{w}_i$ the column vectors of the matrix W, and by $\widehat{\mathbf{w}}^j$ the rows of W^{-1}, that is,

$$W = [\mathbf{w}_1, \ldots, \mathbf{w}_n], \qquad W^{-1} = \begin{bmatrix} \widehat{\mathbf{w}}^1 \\ \vdots \\ \widehat{\mathbf{w}}^n \end{bmatrix}.$$

The columns of W and the rows of W^{-1} form a pair of biorthogonal bases, i.e., $\widehat{\mathbf{w}}^i\mathbf{w}_j = \delta_{ij}$.

Lemma 2.4.1. *If* $r_k = \operatorname{rank}(J_k)$, *then*

$$\mathcal{N}(J_k) = \operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_{n-r_k}).$$

Proof. Any vector $\mathbf{x}$ in $\mathbb{R}^n$ can be represented in the basis $\{\mathbf{w}_i\}_{i=1}^n$ by writing

$$\mathbf{x} = W(W^{-1}\mathbf{x}) = \sum_{j=1}^{n} (\widehat{\mathbf{w}}^j\mathbf{x})\mathbf{w}_j. \tag{2.16}$$

From the GSVD of (J_k, L), we obtain

$$J_k\mathbf{x} = \sum_{i=1}^{m} \delta_i \mathbf{u}_i,$$

where $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_m]^T = \Sigma_J W^{-1}\mathbf{x}$. When $m \geq n$, Σ_J is of the form (1.4) and it leads to

$$\delta_i = \begin{cases} 0, & i = 1, \ldots, n - r_k, \\ c_{i-n+r_k}(\widehat{\mathbf{w}}^i\mathbf{x}), & i = n - r_k + 1, \ldots, p, \\ \widehat{\mathbf{w}}^i\mathbf{x}, & i = p + 1, \ldots, n, \\ 0, & i = n + 1, \ldots, m, \end{cases}$$

so that $J_k\mathbf{x} = 0$ if and only if $\widehat{\mathbf{w}}^i\mathbf{x} = 0$, for $i = n - r_k + 1, \ldots, n$. By (2.16), this means that

$$\mathbf{x} \in \operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_{n-r_k}).$$

When $m < n$, Σ_J is given by (1.6), then we obtain

$$\delta_i = \begin{cases} 0, & i = 1, \ldots, m - r_k, \\ c_{i-m+r_k}(\widehat{\mathbf{w}}^{i-m+n}\mathbf{x}), & i = m - r_k + 1, \ldots, m + p - n, \\ \widehat{\mathbf{w}}^{i-m+n}\mathbf{x}, & i = m + p - n + 1, \ldots, m, \end{cases}$$

so the same conclusion holds true: $\mathbf{x} \in \mathcal{N}(J_k)$ if and only if $\mathbf{x} \in \operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_{n-r_k})$. □

Theorem 2.4.2. *Let $\mathbf{x}^{(k)} \in \mathbb{R}^n$ and let $\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)}$ be the Gauss–Newton iteration for (2.1), where the step $\widetilde{\mathbf{s}}^{(k)}$ has been determined by solving (2.10). Then, the iteration $\mathbf{x}^{(k+1)}$ for (2.15), starting from the same point $\mathbf{x}^{(k)}$, is given by*

$$\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - W_1\widehat{W}_1\mathbf{x}^{(k)}, \tag{2.17}$$

where $\widehat{W}_1 \in \mathbb{R}^{(n-r_k)\times n}$ contains the first $n - r_k$ rows of W^{-1} and $W_1 \in \mathbb{R}^{n\times(n-r_k)}$ is composed of the first $n - r_k$ columns of W.

Proof. Replacing J_k and L with their GSVD and setting $\mathbf{y} = W^{-1}\mathbf{s}$, $\mathbf{z}^{(k)} = W^{-1}\mathbf{x}^{(k)}$, and $\mathbf{g}^{(k)} = U^T\mathbf{r}_k$, (2.15) can be rewritten as the following diagonal least-squares problem

$$\begin{cases} \min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma_L(\mathbf{y} + \mathbf{z}^{(k)})\|^2 \\ \mathbf{y} \in \left\{\arg\min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma_J\mathbf{y} + \mathbf{g}^{(k)}\|^2\right\}. \end{cases} \tag{2.18}$$

When $m \geq n$, the diagonal linear system in the constraint is solved by a vector $\mathbf{y}$ with entries

$$y_i = \begin{cases} -\dfrac{g_i^{(k)}}{c_{i-n+r_k}}, & i = n - r_k + 1, \ldots, p, \\ -g_i^{(k)}, & i = p + 1, \ldots, n, \end{cases}$$

while the components y_i, for $i = 1, \ldots, n - r_k$, can be determined by minimizing the norm

$$\|\Sigma_L(\mathbf{y} + \mathbf{z}^{(k)})\|^2 = \sum_{i=1}^{n-r_k}\left(y_i + z_i^{(k)}\right)^2 + \sum_{i=n-r_k+1}^{p}\left(-\frac{g_i^{(k)}}{\gamma_{i-n+r_k}} + s_{i-n+r_k}z_i^{(k)}\right)^2, \tag{2.19}$$

where $\gamma_i = \frac{c_i}{s_i}$ are the generalized singular values of the matrix pair (J_k, L). The minimum of (2.19) is reached for

$$y_i = -z_i^{(k)} = -\widehat{\mathbf{w}}^i\mathbf{x}^{(k)}, \quad i = 1, \ldots, n - r_k,$$

and the solution to (2.15), that is, the next approximation for the solution of (2.14), is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{n-r_k} z_i^{(k)} \mathbf{w}_i - \sum_{i=n-r_k+1}^{p} \frac{g_i^{(k)}}{c_{i-n+r_k}} \mathbf{w}_i - \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i, \tag{2.20}$$

where the first summation at the right-hand side can be rewritten in the form $W_1 \widehat{W}_1 \mathbf{x}^{(k)}$. Applying the same procedure to (2.10), we obtain

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \frac{g_i^{(k)}}{c_{i-n+r_k}} \mathbf{w}_i - \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i, \tag{2.21}$$

from which (2.17) follows. We note that the last summation in (2.20) and (2.21) is the component of the update vector $\mathbf{s}$ in the null space of L.

When $m < n$, (1.6) yields the following solution for the diagonal system in (2.18),

$$y_i = \begin{cases} -\dfrac{g_{i-n+m}^{(k)}}{c_{i-n+r_k}}, & i = n - r_k + 1, \ldots, p, \\ -g_{i-n+m}^{(k)}, & i = p+1, \ldots, n, \end{cases}$$

from which, after minimizing the weighted norm like in (2.19), we obtain

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{n-r_k} z_i^{(k)} \mathbf{w}_i - \sum_{i=n-r_k+1}^{p} \frac{g_{i-n+m}^{(k)}}{c_{i-n+r_k}} \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-n+m}^{(k)} \mathbf{w}_i. \tag{2.22}$$

Since solving (2.10) when $m < n$ leads to a formula similar to (2.21) with $g_{i-n+m}^{(k)}$ in place of $g_i^{(k)}$, the validity of (2.17) is confirmed. □

2.5 Regularization

The nonlinear function $F(\mathbf{x})$ is considered ill-conditioned in a domain $\mathcal{D} \subset \mathbb{R}^n$ when the condition number $\kappa(J)$ of the Jacobian $J = J(\mathbf{x})$ is very large for any $\mathbf{x} \in \mathcal{D}$. In this situation, it is common to apply a regularization procedure to each step of the Gauss–Newton method. Section 1.2 contains a reminder of the reasons that lead to the need to regularize a problem, as well as an introduction to regularization methods.

The truncated singular value decomposition (TSVD) solves (2.9) after substituting J_k by its best rank-ℓ approximation (see Theorem 1.1.1), that is,

$$J_k^{(\ell)} = \arg \min_{\operatorname{rank}(M)=\ell} \|J_k - M\| = \sum_{i=1}^{\ell} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{2.23}$$

where $(\sigma_i, \mathbf{u}_i, \mathbf{v}_i)$ is the ith singular triplet for J_k; see (1.2). Here, ℓ plays the role of a regularization parameter, which has to be carefully chosen. Its role is to approximate the initial least-squares problem by a better-conditioned problem. Choosing its value amounts to finding a compromise between fidelity to the original model and numerical stability.

Another classical approach is Tikhonov regularization, in which the minimization problem (2.6) is replaced by

$$\min_{\mathbf{s}\in\mathbb{R}^n} \left\{ \|J_k\mathbf{s} + \mathbf{r}_k\|^2 + \lambda^2\|\mathbf{s}\|^2 \right\}, \tag{2.24}$$

for a fixed value of the parameter $\lambda > 0$. The regularization parameter λ controls the balance between the two terms of the functional, i.e., the weights attributed to the residual term and to the regularization term.

If a regularization matrix $L \in \mathbb{R}^{p\times n}$ is introduced, (2.9) becomes (2.10), and the regularized solution is computed by the truncated generalized singular value decomposition (TGSVD). If the Tikhonov approach is followed, then the standard form functional (2.24) is expressed in general form

$$\min_{\mathbf{s}\in\mathbb{R}^n} \left\{ \|J_k\mathbf{s} + \mathbf{r}_k\|^2 + \lambda^2\|L\mathbf{s}\|^2 \right\}. \tag{2.25}$$

We stress that both (2.24) and TSVD applied to (2.9) impose a regularity constraint on the update vector $\mathbf{s}$ for the solution $\mathbf{x}^{(k)}$ and not on the solution itself in the same matter as (2.25) and TGSVD applied to (2.10) do.

2.5.1 Truncated minimal-norm solution

When the function F is ill-conditioned, we propose a *truncated minimal-norm Gauss–Newton* (TMNGN) method to solve (2.12). We choose a value for the truncation parameter $1 \le \ell \le r_k$, an initial solution $\mathbf{x}_\ell^{(0)}$, and compute

$$\mathbf{x}_\ell^{(k+1)} = \mathbf{x}_\ell^{(k)} - \sum_{i=1}^{\ell} \frac{g_i^{(\ell,k)}}{\sigma_i}\mathbf{v}_i - V_{2,\ell}V_{2,\ell}^T\mathbf{x}_\ell^{(k)}, \qquad k = 0, 1, 2, \ldots, \tag{2.26}$$

where $V_{2,\ell} = [\mathbf{v}_{\ell+1}, \ldots, \mathbf{v}_n]$, until convergence. In the above formula, $\mathbf{g}^{(\ell,k)} = U^T\mathbf{r}(\mathbf{x}_\ell^{(k)})$ as in the proof of Theorem 2.3.1. Notice that the columns of $V_{2,\ell}$ form a basis for the null space of the rank-ℓ approximation (2.23) of the Jacobian.

In case a partial SVD is computed, say, up to the truncation index ℓ, the last term may be expressed as $(I_n - V_{1,\ell}V_{1,\ell}^T)\mathbf{x}_\ell^{(k)}$, where $V_{1,\ell} = [\mathbf{v}_1, \ldots, \mathbf{v}_\ell]$. There are several methods for computing a partial SVD for large-scale problems [7, 8, 9, 65, 92].

To solve (2.15), we employ a *truncated minimal-L-norm Gauss–Newton* (TML-NGN) method. This consists of choosing an integer $0 \le \ell \le p - n + r_k = r_k - d$

(see (2.20) and (2.22)), and computing, for $k = 0, 1, 2, \ldots$, until convergence, the iterates

$$\mathbf{x}_\ell^{(k+1)} = \mathbf{x}_\ell^{(k)} - \sum_{i=p-\ell+1}^{p} \frac{g_{i-N}^{(\ell,k)}}{c_{i-n+r_k}} \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-N}^{(\ell,k)} \mathbf{w}_i - W_{1,\ell} \widehat{W}_{1,\ell} \mathbf{x}_\ell^{(k)}, \tag{2.27}$$

where $N = \max(n - m, 0)$. The matrix $W_{1,\ell} \in \mathbb{R}^{n \times (p-\ell)}$ contains the first $p - \ell$ columns of W, and $\widehat{W}_{1,\ell} \in \mathbb{R}^{(p-\ell) \times n}$ the first $p - \ell$ rows of W^{-1}. Again, the columns of $W_{1,\ell}$ span the null space of $J_k^{(\ell)}$.

In formulas (2.26) and (2.27), the solution at convergence will be denoted by $\mathbf{x}_\ell$. Under the assumption that the exact data vector $\mathbf{b}_{\text{exact}}$ is perturbed by noise

$$\mathbf{b} = \mathbf{b}_{\text{exact}} + \mathbf{e}, \tag{2.28}$$

and that the noise level $\|\mathbf{e}\|$ is known, the classical discrepancy principle introduced by Morozov [77] can be used to estimate the optimal value of the regularization parameter, namely, selecting the smallest truncation parameter ℓ such that

$$\|F(\mathbf{x}_\ell) - \mathbf{b}\| \leq \tau \|\mathbf{e}\|, \tag{2.29}$$

where $\tau > 1$ is a constant independent of the noise level $\|\mathbf{e}\|$.

When the noise level is unknown, heuristic methods are commonly used. We use the L-curve criterion [56, 60] which selects the regularization parameter at the "corner" of the curve obtained by joining the points

$$\left(\log \|F(\mathbf{x}_\ell) - \mathbf{b}\|,\ \log \|\mathbf{x}_\ell\|\right). \tag{2.30}$$

When solving discrete ill-posed problems, this curve often exhibits a typical L-shape. We determine its corner by the method described in [59] and implemented in [58].

For a summary of the criteria for choosing the regularization parameter, the reader is referred to Subsection 1.2.3.

2.5.2 Minimal-norm Tikhonov solution

We assume here that a regularizing term is added to the least-squares problem (2.1), transforming it into the minimization of the nonlinear Tikhonov functional

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \|F(\mathbf{x}) - \mathbf{b}\|^2 + \lambda^2 \|L\mathbf{x}\|^2 \right\}, \tag{2.31}$$

where $\lambda > 0$ is a continuous regularization parameter and $L \in \mathbb{R}^{p \times n}$ is a regularization matrix. We will apply the Gauss–Newton method to the solution of (2.31) and compare the iterates to those derived from the application of the same method to (2.1) followed by the Tikhonov regularization of each step as in (2.25).

Linearizing (2.31), we obtain

$$\min_{\mathbf{s}\in\mathbb{R}^n} \left\{ \|J_k\mathbf{s} + \mathbf{r}_k\|^2 + \lambda^2 \|L(\mathbf{x}^{(k)} + \mathbf{s})\|^2 \right\}. \tag{2.32}$$

We first analyze the case $L = I_n$.

Theorem 2.5.1. *Let* $\operatorname{rank}(J_k) = r_k$. *The iteration for* (2.32) *is given by*

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{\sigma_i g_i^{(k)} + \lambda^2 z_i^{(k)}}{\sigma_i^2 + \lambda^2} \mathbf{v}_i - V_2 V_2^T \mathbf{x}^{(k)}, \tag{2.33}$$

where $\mathbf{g}^{(k)} = U^T\mathbf{r}_k$, $\mathbf{z}^{(k)} = V^T\mathbf{x}^{(k)}$, *and* $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$ *is defined as in Theorem 2.3.1.*

Proof. Computing the gradient of the function (2.32) with $L = I_n$ yields the normal equations associated to the penalized least-squares problem

$$(J_k^T J_k + \lambda^2 I_n)\mathbf{s} = -J_k^T \mathbf{r}_k - \lambda^2 \mathbf{x}^{(k)}. \tag{2.34}$$

By employing the singular value decomposition $J_k = U\Sigma V^T$, the normal equations (2.34) become

$$(\Sigma^T\Sigma + \lambda^2 I_n)\mathbf{y} = -\Sigma^T \mathbf{g}^{(k)} - \lambda^2 \mathbf{z}^{(k)}, \tag{2.35}$$

with $\mathbf{y} = V^T\mathbf{s}$, $\mathbf{g}^{(k)} = U^T\mathbf{r}_k$, and $\mathbf{z}^{(k)} = V^T\mathbf{x}^{(k)}$. The solution to the diagonal normal equations (2.35),

$$y_i = \begin{cases} -\dfrac{\sigma_i g_i^{(k)} + \lambda^2 z_i^{(k)}}{\sigma_i^2 + \lambda^2}, & i = 1, \ldots, r_k, \\ -z_i^{(k)}, & i = r_k + 1, \ldots, n, \end{cases}$$

leads to the *Tikhonov–Gauss–Newton* (TikGN) iterate, which solves (2.32):

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{\sigma_i g_i^{(k)} + \lambda^2 z_i^{(k)}}{\sigma_i^2 + \lambda^2} \mathbf{v}_i - \sum_{i=r_k+1}^{n} z_i^{(k)} \mathbf{v}_i.$$

The last summation can be rewritten in matrix form as $V_2 V_2^T \mathbf{x}^{(k)}$, where $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$. This completes the proof. □

The normal equations associated to (2.24) are

$$(J_k^T J_k + \lambda^2 I_n)\mathbf{s} = -J_k^T \mathbf{r}_k,$$

which become after substituting the SVD of J_k,

$$(\Sigma^T\Sigma + \lambda^2 I_n)\mathbf{y} = -\Sigma^T \mathbf{g}^{(k)}.$$

The solution to this diagonal system

$$y_i = \begin{cases} -\dfrac{\sigma_i g_i^{(k)}}{\sigma_i^2 + \lambda^2}, & i = 1, \dots, r_k, \\ 0, & i = r_k + 1, \dots, n, \end{cases}$$

produces the iterate

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{\sigma_i g_i^{(k)}}{\sigma_i^2 + \lambda^2} \mathbf{v}_i. \tag{2.36}$$

Comparing equation (2.36), where the approximate solution is obtained by imposing the regularity constraint on the update vector $\mathbf{s}$, to the iteration (2.33), where the regularity constraint is imposed on the approximate solution $\mathbf{x}^{(k+1)}$, we see that the TikGN method implements a different filtering technique with respect to the standard application of Tikhonov regularization to the Gauss–Newton iteration and produces approximate solutions which are orthogonal to the null space of the Jacobian matrix J_k.

Remark 2.5.2. Since $V_2 V_2^T = I_n - V_1 V_1^T$, the updated solution (2.33) can be expressed without the explicit use of the singular vectors $\mathbf{v}_i$, $i = r_k + 1, \dots, n$, in the form

$$\mathbf{x}^{(k+1)} = V_1 V_1^T \mathbf{x}^{(k)} - \sum_{i=1}^{r_k} \frac{\sigma_i g_i^{(k)} + \lambda^2 z_i^{(k)}}{\sigma_i^2 + \lambda^2} \mathbf{v}_i.$$

This is useful when a compact SVD is available.

Formula (2.33) immediately yields the following result.

Corollary 2.5.3. *When the regularization parameter λ approaches zero, the TikGN iterate computed by (2.33) converges to the MNGN solution (2.13), that is*

$$\mathbf{x}_{MNGN}^{(k+1)} = \lim_{\lambda \to 0^+} \mathbf{x}_{TikGN}^{(k+1)}.$$

We now turn to the case $L \neq I_n$. We will denote the resulting method by TikLGN.

Theorem 2.5.4. *Let* $\operatorname{rank}(J_k) = r_k$. *The iteration for the TikLGN approach (2.32) is*

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \xi_i \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-N}^{(k)} \mathbf{w}_i - W_1 \widehat{W}_1 \mathbf{x}^{(k)}, \tag{2.37}$$

with

$$\xi_i = \frac{c_{i-n+r_k} g_{i-N}^{(k)} + \lambda^2 s_{i-n+r_k}^2 z_i^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2}, \qquad i = n - r_k + 1, \dots, p,$$

where $N = \max(n - m, 0)$ and W_1 and $\widehat{W}_1$ are defined as in Theorem 2.4.2.

Proof. Let us consider the generalized singular value decomposition (1.3) of the matrix pair (J_k, L). We initially assume that $m \geq n \geq r_k = \operatorname{rank}(J_k)$ and that L has full rank, i.e., $\operatorname{rank}(L) = p$. We have $J_k = U\Sigma_J W^{-1}$ and $L = V\Sigma_L W^{-1}$, with Σ_J and Σ_L given by (1.4).

Substituting the GSVD in the normal equations associated to (2.32),

$$(J_k^T J_k + \lambda^2 L^T L)\mathbf{s} = -J_k^T \mathbf{r}_k - \lambda^2 L^T L \mathbf{x}^{(k)},$$

leads to

$$(D + \lambda^2 H)\mathbf{y} = -\Sigma_J^T \mathbf{g}^{(k)} - \lambda^2 H \mathbf{z}^{(k)}, \tag{2.38}$$

where

$$D = \begin{bmatrix} O_{n-r} & & \\ & C^2 & \\ & & I_{n-p} \end{bmatrix}, \qquad H = \begin{bmatrix} I_{n-r} & & \\ & S^2 & \\ & & O_{n-p} \end{bmatrix},$$

$\mathbf{y} = W^{-1}\mathbf{s}$, $\mathbf{g}^{(k)} = U^T \mathbf{r}_k$, and $\mathbf{z}^{(k)} = W^{-1}\mathbf{x}^{(k)}$. The diagonal system (2.38) yields the iterate

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \xi_i \mathbf{w}_i - \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i - W_1 \widehat{W}_1 \mathbf{x}^{(k)}, \tag{2.39}$$

where

$$\xi_i = \frac{c_{i-n+r_k} g_i^{(k)} + \lambda^2 s_{i-n+r_k}^2 z_i^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2}, \qquad i = n - r_k + 1, \ldots, p.$$

Similarly, when $r_k \leq m < n$, the TikLGN approach leads to the iterate

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \xi_i' \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-n+m}^{(k)} \mathbf{w}_i - W_1 \widehat{W}_1 \mathbf{x}^{(k)}, \tag{2.40}$$

with

$$\xi_i' = \frac{c_{i-n+r_k} g_{i-n+m}^{(k)} + \lambda^2 s_{i-n+r_k}^2 z_i^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2}, \qquad i = n - r_k + 1, \ldots, p.$$

Introducing $N = n - m$ if $m < n$ and zero otherwise, the overdetermined (2.39) and the underdetermined (2.40) cases may be condensed into the single expression (2.37), and this completes the proof. □

The normal equations associated to (2.25), if $m \geq n \geq r_k$, are

$$(J_k^T J_k + \lambda^2 L^T L)\mathbf{s} = -J_k^T \mathbf{r}_k,$$

that is,

$$(D + \lambda^2 H)\mathbf{y} = -\Sigma_J^T \mathbf{g}^{(k)},$$

where D, H, and $\mathbf{g}^{(k)}$ are defined as above. This diagonal system yields

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \frac{c_{i-n+r_k} g_i^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2} \mathbf{w}_i - \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i. \tag{2.41}$$

When $r_k \leq m < n$, the iteration induced by (2.25) is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \frac{c_{i-n+r_k} g_{i-n+m}^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2} \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-n+m}^{(k)} \mathbf{w}_i. \tag{2.42}$$

We can condense the overdetermined (2.41) and the underdetermined (2.42) cases into a single expression, introducing N defined as in the Theorem 2.5.4:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \sum_{i=n-r_k+1}^{p} \frac{c_{i-n+r_k} g_{i-N}^{(k)}}{c_{i-n+r_k}^2 + \lambda^2 s_{i-n+r_k}^2} \mathbf{w}_i - \sum_{i=p+1}^{n} g_{i-N}^{(k)} \mathbf{w}_i.$$

Comparing (2.37) to this formula shows, as in the case $L = I_n$, that the minimal-L-norm approach and the traditional Tikhonov method produce different reconstructions. Also, when the regularization parameter λ approaches zero, the TikLGN solution converges to the MLNGN solution.

Corollary 2.5.5. *For the iterations computed by the MLNGN method* (2.20) *and by the TikLGN method* (2.39)*, it holds that*

$$\mathbf{x}_{MLNGN}^{(k+1)} = \lim_{\lambda \to 0^+} \mathbf{x}_{TikLGN}^{(k+1)}.$$

In formulas (2.33) and (2.37), the solution at convergence will be denoted by $\mathbf{x}_\lambda$, and also in this case we will consider the right-hand side $\mathbf{b}$ to be affected by noise as in (2.28). In this case, the regularization parameter λ can be estimated by the discrepancy principle, substituting $F(\mathbf{x}_\ell)$ by $F(\mathbf{x}_\lambda)$ in (2.29). The L-curve criterion can also be adapted by substituting $F(\mathbf{x}_\ell)$ and $\mathbf{x}_\ell$ by $F(\mathbf{x}_\lambda)$ and $\mathbf{x}_\lambda$, respectively, in (2.30).

We adopt the following stopping rule for all the iterative methods (M(L)NGN, TM(L)NGN, Tik(L)GN). We iterate until either the difference between two successive approximations is small enough

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \delta \|\mathbf{x}^{(k+1)}\|,$$

or until a chosen maximum number of iterations $N_{\max}$ is reached.

In the case of ill-conditioned problems, it is useful to consider an additional stopping criterion in order to detect the unboundedness of the solution for a particular value of the regularization parameter. The iteration is interrupted when one of the preceding conditions is reached or when the ratio between the norms of the kth approximate solution and the initial point is larger than a certain fixed threshold.

2.6 Implementation details

In some situations, the `gsvd` routine provided by Matlab produces unexpected results. We observed that when the norm of the Jacobian matrix J_k is very small,

the GSVD of (J_k, L) may produce an inaccurate factor W, which prevents the Gauss–Newton method (2.15) to converge. To overcome such numerical issues, when $\|J_k\|_\infty < \rho$ (in the experiments we set $\rho = 10^{-6}$), we rescale the least-squares problem (2.15) to obtain

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|L(\mathbf{x}^{(k)} + \mathbf{s})\|^2 \\ \mathbf{s} \in \left\{ \arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|\widetilde{J}_k\mathbf{s} + \widetilde{\mathbf{r}}_k\|^2 \right\}, \end{cases}$$

with $\widetilde{J}_k = \rho^{-1} J_k$ and $\widetilde{\mathbf{r}}_k = \rho^{-1}\mathbf{r}_k$, before applying the algorithms described in the preceding sections. The Armijo–Goldstein principle (2.8) is modified accordingly. The Tikhonov approach (2.32) is rescaled similarly.

2.7 Doubly relaxed nonlinear minimal-norm solution

Let us now briefly review the computation of the minimal-norm solution to the nonlinear problem (2.1) by the *minimal-norm Gauss–Newton* (MNGN) method, presented in [82] and reported in Section 2.3. Our aim is showing the reason for the possible lack of convergence of such method. Here, we extend the discussion by introducing a model profile $\overline{\mathbf{x}} \in \mathbb{R}^n$, which represents an a priori estimate of the desired solution, and formulate the problem in the form

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x} - \overline{\mathbf{x}}\|^2 \\ \mathbf{x} \in \left\{ \arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} \|F(\mathbf{x}) - \mathbf{b}\|^2 \right\}. \end{cases} \tag{2.43}$$

When such an approximation is unavailable, we just set $\overline{\mathbf{x}} = \mathbf{0}$. Similarly to Section 2.3, we consider an iterative method of the type (2.7) based on the following first-order linearization of the problem

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|\mathbf{x}^{(k)} - \overline{\mathbf{x}} + \alpha_k\mathbf{s}\|^2 \\ \mathbf{s} \in \left\{ \arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2 \right\}, \end{cases} \tag{2.44}$$

where J_k is the Jacobian of F in $\mathbf{x}^{(k)}$ and $\mathbf{r}_k$ is the residual vector. The damping parameter α_k is indispensable to ensure the convergence of the Gauss–Newton method. We determine it by the Armijo–Goldstein principle.

The iteration resulting from the solution of (2.44) is defined by the following theorem.

Theorem 2.7.1. *Let $\mathbf{x}^{(k)} \in \mathbb{R}^n$ and let $\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\widetilde{\mathbf{s}}^{(k)}$ be the Gauss–Newton iteration for (2.1), where the step $\widetilde{\mathbf{s}}^{(k)}$ is determined by solving (2.9) and the step*

length α_k *by the Armijo–Goldstein principle. Then, the iteration* $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{s}^{(k)}$ *defined by* (2.44) *is given by*

$$\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - V_2 V_2^T \big(\mathbf{x}^{(k)} - \overline{\mathbf{x}}\big), \tag{2.45}$$

where $\operatorname{rank}(J_k) = r_k$ *and the columns of the matrix* $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$ *are orthonormal vectors in* $\mathbb{R}^n$ *spanning the null space of* J_k.

Proof. The proof follows the pattern of that of Theorem 2.3.1. Let $U\Sigma V^T$ be the singular value decomposition of the matrix J_k. The upper-level problem in (2.44) can be expressed as

$$\|\mathbf{x}^{(k)} - \overline{\mathbf{x}} + \alpha_k \mathbf{s}\|^2 = \|V^T(\mathbf{x}^{(k)} - \overline{\mathbf{x}} + \alpha_k \mathbf{s})\|^2 = \|\alpha_k \mathbf{y} + \mathbf{z}^{(k)}\|^2,$$

with $\mathbf{y} = V^T \mathbf{s}$ and $\mathbf{z}^{(k)} = V^T \big(\mathbf{x}^{(k)} - \overline{\mathbf{x}}\big)$. Replacing J_k by its SVD and setting $\mathbf{g}^{(k)} = U^T \mathbf{r}_k$, we can rewrite (2.44) as the following diagonal linear least-squares problem

$$\begin{cases} \min\limits_{\mathbf{y} \in \mathbb{R}^n} \|\alpha_k \mathbf{y} + \mathbf{z}^{(k)}\|^2 \\ \mathbf{y} \in \big\{ \arg\min\limits_{\mathbf{y} \in \mathbb{R}^n} \|\Sigma \mathbf{y} + \mathbf{g}^{(k)}\|^2 \big\}. \end{cases}$$

Solving the lower-level minimization problem uniquely determines the components $y_i = -\sigma_i^{-1} g_i^{(k)}$, $i = 1, \ldots, r_k$, while the entries y_i, $i = r_k + 1, \ldots, n$, are left undetermined. Their values can be found by solving the upper-level problem. From

$$\|\alpha_k \mathbf{y} + \mathbf{z}^{(k)}\|^2 = \sum_{i=1}^{r_k} \left(-\alpha_k \frac{g_i^{(k)}}{\sigma_i} + z_i^{(k)} \right)^2 + \sum_{i=r_k+1}^{n} \left(\alpha_k y_i + z_i^{(k)} \right)^2,$$

we obtain $y_i = -\frac{z_i^{(k)}}{\alpha_k} = -\frac{1}{\alpha_k} \mathbf{v}_i^T (\mathbf{x}^{(k)} - \overline{\mathbf{x}})$, $i = r_k + 1, \ldots, n$. Then, the solution to (2.44), that is, the next approximation to the solution of (2.43), is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k V \mathbf{y} = \mathbf{x}^{(k)} - \alpha_k \sum_{i=1}^{r_k} \frac{g_i^{(k)}}{\sigma_i} \mathbf{v}_i - \sum_{i=r_k+1}^{n} (\mathbf{v}_i^T (\mathbf{x}^{(k)} - \overline{\mathbf{x}})) \mathbf{v}_i,$$

where the last summation can be written in matrix form as $V_2 V_2^T \big(\mathbf{x}^{(k)} - \overline{\mathbf{x}}\big)$, and the columns of $V_2 = [\mathbf{v}_{r_k+1}, \ldots, \mathbf{v}_n]$ are a basis for $\mathcal{N}(J_k)$.

It is immediate (see Theorem 2.3.1) to prove that

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}^{(k)} = \mathbf{x}^{(k)} - \alpha_k \sum_{i=1}^{r_k} \frac{g_i^{(k)}}{\sigma_i} \mathbf{v}_i,$$

from which (2.45) follows. □

Summarizing, the MNGN method consists of the iteration

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{s}^{(k)},$$

where the step is

$$\mathbf{s}^{(k)} = \widetilde{\mathbf{s}}^{(k)} - \frac{1}{\alpha_k}\mathbf{t}^{(k)},$$

with

$$\widetilde{\mathbf{s}}^{(k)} = -\sum_{i=1}^{r_k} \frac{g_i^{(k)}}{\sigma_i}\mathbf{v}_i, \qquad \mathbf{t}^{(k)} = V_2 V_2^T\big(\mathbf{x}^{(k)} - \overline{\mathbf{x}}\big). \tag{2.46}$$

Theorem 2.7.1 shows that the correction vector $\mathbf{t}^{(k)}$ defined in (2.46), which allows to compute the minimal-norm solution at each step, is not damped by the parameter α_k. As a result, in some numerical examples, the method fails to converge because projecting the solution orthogonally to the null space of J_k causes the residual to increase. To understand how this can happen, a second-order analysis of the objective function is required.

The second-order Taylor approximation to the function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|^2$ at $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\mathbf{s}$ is

$$f(\mathbf{x}^{(k+1)}) \simeq f(\mathbf{x}^{(k)}) + \alpha\nabla f(\mathbf{x}^{(k)})^T\mathbf{s} + \frac{1}{2}\alpha^2\mathbf{s}^T\nabla^2 f(\mathbf{x}^{(k)})\mathbf{s}. \tag{2.47}$$

The gradient and the Hessian of $f(\mathbf{x})$, written in matrix form, are given by (2.2) and (2.3), respectively; see Section 2.2. By replacing the expression of f and $\alpha\mathbf{s} = \alpha\widetilde{\mathbf{s}} - \mathbf{t}$ in (2.47), where $\widetilde{\mathbf{s}}$ is the Gauss–Newton step and $\mathbf{t}$ is in the null space of J_k, and letting $\mathcal{Q}_k = \mathcal{Q}(\mathbf{x}^{(k)})$, the following approximation is obtained

$$\begin{aligned}\frac{1}{2}\|\mathbf{r}_{k+1}\|^2 &\simeq \frac{1}{2}\|\mathbf{r}_k\|^2 + \alpha\mathbf{r}_k^T J_k\mathbf{s} + \frac{1}{2}\alpha^2\mathbf{s}^T\left(J_k^T J_k + \mathcal{Q}_k\right)\mathbf{s}\\ &= \frac{1}{2}\|\mathbf{r}_k\|^2 + \alpha\mathbf{r}_k^T J_k\widetilde{\mathbf{s}} + \frac{1}{2}\alpha^2\widetilde{\mathbf{s}}^T\left(J_k^T J_k + \mathcal{Q}_k\right)\widetilde{\mathbf{s}} - \alpha\mathbf{t}^T\mathcal{Q}_k\widetilde{\mathbf{s}} + \frac{1}{2}\mathbf{t}^T\mathcal{Q}_k\mathbf{t}.\end{aligned}$$

The first two terms containing second derivatives (the matrix $\mathcal{Q}_k$) are damped by the α parameter. If the function F is mildly nonlinear, the third term $\frac{1}{2}\mathbf{t}^T\mathcal{Q}_k\mathbf{t}$ is negligible. In the presence of a strong nonlinearity, its contribution to the residual is significant and may lead to its growth. This shows that a damping parameter is required to control the step length for both the Gauss–Newton step $\widetilde{\mathbf{s}}$ and the correction vector $\mathbf{t}$. If a relaxation parameter is introduced for $\mathbf{t}$, Theorem 2.7.1 implies that the minimal-norm solution of (2.44) can only be approximated.

Remark 2.7.2. We report a simple low dimensional example for which the MNGN method may not converge. Let us consider the function $F : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F(\mathbf{x}) = \delta^2\left[(x_1 - \gamma)^2 + (x_2 - \gamma)^2\right] - 1,$$

depending on the parameters $\delta, \gamma \in \mathbb{R}$. Since the Hessian matrix of the residual is given by

$$\nabla^2 r(\mathbf{x}) = \begin{bmatrix} 2\delta^2 & 0 \\ 0 & 2\delta^2 \end{bmatrix},$$

the second-order term $\frac{1}{2}\mathbf{t}^T \mathcal{Q}_k \mathbf{t}$ is not negligible, in general, when δ is relatively large. For example, setting $\delta = 0.7$, $\gamma = 2$, and choosing an initial vector $\mathbf{x}^{(0)}$ with random components in $(-5, 5)$, the MNGN method converges with a large number of iterations (350 on average). Setting $\delta = 0.75$, the same method does not converge within 500 iterations.

2.8 Estimating the rank of the Jacobian

In order to apply Theorem 2.7.1 to computing the minimal-norm solution by (2.45), the rank of the Jacobian matrix J_k should be known in advance. As the rank may vary along the iterations, we set $r_k = \operatorname{rank}(J_k)$. The knowledge of r_k for each $k = 0, 1, \ldots$, is not generally available, making it necessary to estimate its value at each iteration step, to avoid non-convergence or a breakdown of the algorithm.

In such situations, it is common to consider the numerical rank $r_{\epsilon,k}$ of J_k, where ϵ represents a chosen tolerance; see Subsection 1.2.1. The numerical rank is defined in terms of the singular values $\sigma_i^{(k)}$ of J_k, as the integer $r_{\epsilon,k}$ such that

$$\sigma_{r_{\epsilon,k}}^{(k)} > \epsilon \geq \sigma_{r_{\epsilon,k}+1}^{(k)}.$$

Theorem 2.7.1 can be adapted to this setting, by simply replacing at each iteration the rank r_k with the numerical rank $r_{\epsilon,k}$.

Determining the numerical rank is a difficult task for discrete ill-posed problems, in which the singular values decay monotonically to zero. In such a case, the numerical rank plays the role of a regularization parameter.

When the problem is locally rank-deficient, meaning that the rank of $J(\mathbf{x})$ depends on the evaluation vector $\mathbf{x}$, the numerical rank $r_{\epsilon,k}$ can be determined, in principle, by choosing a suitable value of ϵ. Numerical experiments show that a fixed value of ϵ does not always lead to a correct estimation of $r_{\epsilon,k}$, and that it is preferable to determine the ϵ-rank by searching for a sensible gap between $\sigma_{r_{\epsilon,k}}^{(k)}$ and $\sigma_{r_{\epsilon,k}+1}^{(k)}$.

To locate such a gap, we adopt a heuristic approach already applied in [23] for the same purpose, in a different setting. At each step, we compute the ratios

$$\rho_i^{(k)} = \frac{\sigma_i^{(k)}}{\sigma_{i+1}^{(k)}}, \qquad i = 1, 2, \ldots, q-1,$$

where $q = \min(m, n)$. Then, we consider the index set

$$\mathcal{I}_k = \left\{ i \in \{1, 2, \ldots, q-1\} : \rho_i^{(k)} > R \text{ and } \sigma_i^{(k)} > \tau \right\}.$$

An index i belongs to $\mathcal{I}_k$ if there is a significant "jump" between $\sigma_i^{(k)}$ and $\sigma_{i+1}^{(k)}$, and $\sigma_i^{(k)}$ is numerically non-zero. If the set $\mathcal{I}_k$ is empty, we set $r_{\epsilon,k} = q$. Otherwise, we consider

$$\rho_j^{(k)} = \max_{i \in \mathcal{I}_k} \rho_i^{(k)}, \tag{2.48}$$

and we define $r_{\epsilon,k} = j$. This amounts to selecting the largest gap between "large" and "small" singular values. In our numerical simulations (Section 6.2), we set $R = 10^2$ and $\tau = 10^{-8}$. We observed that the value of these parameters is not critical for problems characterized by a rank-deficient Jacobian. Estimating the rank becomes increasingly difficult as the gap between "large" and "small" singular values gets smaller. This condition usually corresponds to ill-conditioned problems, which require specific regularization methods.

2.9 Choosing the projection step length

The occasional non-convergence in the computation of the minimal-norm solution to a nonlinear least-squares problem was discussed in [20], where the authors propose an iterative method based on a convex combination of the Gauss–Newton and the minimal-norm Gauss–Newton iterates, which we denote by CKB. Following our notation, it can be expressed in the form

$$\mathbf{x}^{(k+1)} = (1 - \gamma_k) \left[\mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)}\right] + \gamma_k \left[\mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)} - V_2 V_2^T \mathbf{x}^{(k)}\right], \tag{2.49}$$

where the parameters $\gamma_k \in [0, 1]$, for $k = 0, 1, \ldots$, form a sequence converging to zero. The standard Gauss–Newton method is obtained by setting $\gamma_k = 0$, while $\gamma_k = 1$ leads to the minimal-norm Gauss–Newton method. In their numerical examples, the authors adopt the sequences $\gamma_k = (0.5)^{k+1}$ and $\gamma_k = (0.5)^{2^k}$.

It is immediate to rewrite (2.49) in the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \widetilde{\mathbf{s}}^{(k)} - \gamma_k V_2 V_2^T \mathbf{x}^{(k)}, \tag{2.50}$$

showing that the method proposed in [20] is equivalent to the application of the undamped Gauss–Newton method, whose convergence is not theoretically guaranteed [10], with a damped correction to favor the decrease of the norm of the solution. The numerical experiments reported in the paper show that the minimization of the residual is sped up if γ_k quickly converges to zero, while the norm of the solution decreases faster if γ_k has a slower decay. The choice of the sequence of parameters appears to be critical to tune the performance of the algorithm, and no adaptive choice for γ_k is proposed.

In this section, following [83], we propose to introduce a second relaxation parameter, β_k, to control the step length of the minimal-norm correction $\mathbf{t}^{(k)}$ defined in (2.46). The new iterative method is denoted by MNGN2 and it takes the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}^{(k)} - \beta_k \mathbf{t}^{(k)}, \tag{2.51}$$

where $\widetilde{\mathbf{s}}^{(k)}$ is the step vector produced by the Gauss–Newton method and $\mathbf{t}^{(k)}$ is the projection vector which makes the norm of $\mathbf{x}^{(k+1)}$ minimal, without changing the value of the linearized residual.

The second-order analysis reported at the end of Section 2.7 may be adapted for the CKB method (2.50)

$$\frac{1}{2}\|\mathbf{r}_{k+1}\|^2 = \frac{1}{2}\|\mathbf{r}_k\|^2 + \mathbf{r}_k^T J_k \widetilde{\mathbf{s}} + \frac{1}{2}\widetilde{\mathbf{s}}^T \left(J_k^T J_k + \mathcal{Q}_k\right) \widetilde{\mathbf{s}} - \gamma \mathbf{t}^T \mathcal{Q}_k \widetilde{\mathbf{s}} + \frac{1}{2}\gamma^2 \mathbf{t}^T \mathcal{Q}_k \mathbf{t}.$$

It shows that neither the CKB nor the MNGN method are guaranteed to converge, as both the Gauss–Newton search direction and the projection step should be damped to ensure that the residual decreases. The MNGN2 method locally converges if α_k and β_k are suitably chosen, but it will recover the minimal-norm solution only if $\beta_k \simeq 1$ for k close to convergence.

Our numerical tests showed that it is important to choose both α_k and β_k adaptively along the iterations. A simple solution is to let $\beta_k = \alpha_k$ and estimate α_k by the Armijo–Goldstein principle (2.8), with $\mathbf{s}^{(k)} = \widetilde{\mathbf{s}}^{(k)} - \mathbf{t}^{(k)}$ in place of $\mathbf{s}^{(k)}$. This approach proves to be effective in the computation of the minimal-norm solution, but its convergence is often rather slow. To speed up iteration we propose a procedure to adaptively choose the value of β_k.

This procedure is outlined in Algorithm 1. Initially, we set $\beta = 1$. At each iteration, we compute the residual at the Gauss–Newton iteration $\widetilde{\mathbf{x}}^{(k+1)}$ and at the tentative iteration $\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - \beta \mathbf{t}^{(k)}$. Subtracting the vector $\beta \mathbf{t}^{(k)}$ may cause the residual to increase. We accept such an increase if

$$\|\mathbf{r}(\mathbf{x}^{(k+1)})\| \le \|\mathbf{r}(\widetilde{\mathbf{x}}^{(k+1)})\| + \delta\big(\|\mathbf{r}(\widetilde{\mathbf{x}}^{(k+1)})\|, \eta\big), \tag{2.52}$$

where $\delta(\rho, \eta)$ is a function determining the maximal increase allowed in the residual $\rho = \|\mathbf{r}(\widetilde{\mathbf{x}}^{(k+1)})\|$, and $\eta > 0$ is a chosen tolerance. On the contrary, β is halved and the residual is recomputed until (2.52) is verified or β becomes excessively small. To allow β to increase, we tentatively double it at each iteration (see line 9 in the algorithm) before applying the above procedure. At line 12 of the algorithm we add the machine epsilon ε_M to the actual residual $\widetilde{\rho}_{k+1}$ to avoid that $\delta(\widetilde{\rho}_{k+1}, \eta)$ becomes zero.

A possible choice for the value of the residual increase is $\delta(\rho, \eta) = \eta\rho$, with η suitably chosen. Our experiments showed that it is possible to find, by chance, a value of η which produces good results, but its choice is strongly dependent on the particular example. We also noticed that, in cases where the residual stagnates, accepting a large increase in the residual may lead to non-convergence. In such situations, a fixed multiple of the residual is not well suited to model its increase. Indeed, if the residual is large, one is prone to accept only a small increase, while if the residual is very small, a relatively large growth may be acceptable.

To overcome these difficulties, we consider $\delta(\rho, \eta) = \rho^\eta$, and choose η at each step by the adaptive procedure described in Algorithm 2. When at least k_{res} iterations

Algorithm 1 Outline of the MNGN2 method.

Require: nonlinear function F, data vector $\mathbf{b}$,
Require: initial solution $\mathbf{x}^{(0)}$, model profile $\overline{\mathbf{x}}$, tolerance η for residual increase
Ensure: approximation $\mathbf{x}^{(k+1)}$ of minimal-norm least-squares solution

$k = 0$, $\beta = 1$
repeat
  $k = k + 1$
  estimate $r_k = \operatorname{rank}(J(\mathbf{x}^{(k)}))$ by (2.48)
  compute $\widetilde{\mathbf{s}}^{(k)}$ by the Gauss–Newton method (2.6)
  compute α_k by the Armijo–Goldstein principle (2.8)
  compute $\mathbf{t}^{(k)}$ by (2.46)
  if $\beta < 1$ **then**
    $\beta = 2\beta$
  end if
  $\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}^{(k)}$
  $\widetilde{\rho}_{k+1} = \|F(\widetilde{\mathbf{x}}^{(k+1)}) - \mathbf{b}\| + \varepsilon_M$
  $\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - \beta \mathbf{t}^{(k)}$
  $\rho_{k+1} = \|F(\mathbf{x}^{(k+1)}) - \mathbf{b}\|$
  while $(\rho_{k+1} > \widetilde{\rho}_{k+1} + \delta(\widetilde{\rho}_{k+1}, \eta))$ **and** $(\beta > 10^{-8})$ **do**
    $\beta = \beta/2$
    $\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - \beta \mathbf{t}^{(k)}$
    $\rho_{k+1} = \|F(\mathbf{x}^{(k+1)}) - \mathbf{b}\|$
  end while
  $\beta_k = \beta$
until convergence

have been performed, we compute the linear polynomial which fits the logarithm of the last k_{res} residuals in the least-squares sense. To detect if the residual stagnates or increases, we check if the slope M of the regression line exceeds -10^{-2}. If this happens, the value of η is doubled. The effect on the algorithm is to enhance the importance of the decrease of the residual and reduce that of the norm. To recover a sensible decrease in the norm, if at a subsequent step the residual reduction accelerates (e.g., $M < -\frac{1}{2}$), the value of η is halved. In our experiments, we initialize η to $\frac{1}{8}$ and set $k_{\text{res}} = 5$.

Remark 2.9.1. The adaptive estimation of $\delta(\rho, \eta)$ does not significantly increase the complexity of Algorithm 1, as line 3 of Algorithm 2 implies the solution of a 2×2 linear system whose matrix is fixed and can be computed in advance, while forming the right-hand side requires $4k_{\text{res}}$ floating point operations.

Algorithm 2 Adaptive determination of the residual increase $\delta(\rho, \eta)$.

Require: actual residual $\rho = \|\mathbf{r}(\widetilde{\mathbf{x}}^{(k+1)})\|$, starting tolerance η
Require: iteration index k, residuals $\theta_j = \|\mathbf{r}(\widetilde{\mathbf{x}}^{(k-k_{\text{res}}+j)})\|$, $j = 1, \ldots, k_{\text{res}}$
Ensure: residual increase $\delta(\rho, \eta)$
1: $M_{\min} = -10^{-2}$, $M_{\max} = -\frac{1}{2}$
2: **if** $k \geq k_{\text{res}}$ **then**
3: compute regression line $p_1(t) = Mt + N$ of $(j, \log(\theta_j))$, $j = 1, \ldots, k_{\text{res}}$
4: **if** $M > M_{\min}$ **then**
5: $\eta = 2\eta$
6: **else if** $M < M_{\max}$ **then**
7: $\eta = \eta/2$
8: **end if**
9: **end if**
10: $\delta(\rho, \eta) = \rho^{\eta}$

To detect convergence, we interrupt the iteration as soon as

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \tau \|\mathbf{x}^{(k+1)}\| \qquad \text{or} \qquad \|\alpha_k \widetilde{\mathbf{s}}^{(k)}\| < \tau, \tag{2.53}$$

or when a fixed number of iterations $N_{\max}$ is exceeded. The second stop condition in (2.53) detects the slow progress of the relaxed Gauss–Newton iteration algorithm. This often happens close to the solution. The stop tolerance is set to $\tau = 10^{-8}$.

2.10 Doubly relaxed nonlinear minimal-L-norm solution

In this section we extend the discussion made in Section 2.4, by introducing in the analysis the step length and an a priori estimate of the solution. Let us introduce

a regularization matrix $L \in \mathbb{R}^{p\times n}$, $p \le n$. While in (2.10) the seminorm $\|L\mathbf{s}\|$ is minimized over all the updating vectors $\mathbf{s}$ which minimize the linearized residual, here we seek to compute the minimal-L-norm solution to the nonlinear problem (2.1), that is the vector $\mathbf{x}$ which solves the constrained problem

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} \|L(\mathbf{x}-\overline{\mathbf{x}})\|^2 \\ \mathbf{x} \in \left\{\arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} \|F(\mathbf{x})-\mathbf{b}\|^2\right\}. \end{cases} \tag{2.54}$$

Similarly to Section 2.7, we consider an iterative method of the type (2.5), where the step $\mathbf{s}^{(k)}$ is the solution of the linearized problem

$$\begin{cases} \min\limits_{\mathbf{s}\in\mathbb{R}^n} \|L(\mathbf{x}^{(k)}-\overline{\mathbf{x}}+\alpha\mathbf{s})\|^2 \\ \mathbf{s} \in \left\{\arg\min\limits_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s}+\mathbf{r}_k\|^2\right\}. \end{cases} \tag{2.55}$$

We will denote the iteration resulting from the solution of (2.55) as the *minimal-L-norm Gauss–Newton* (MLNGN) method.

Let $J_k = U\Sigma_J W^{-1}$, $L = V\Sigma_L W^{-1}$ be the GSVD of the matrix pair (J_k,L). We indicate by $\mathbf{w}_i$ the column vectors of the matrix W, and by $\widehat{\mathbf{w}}^j$ the rows of W^{-1}, that is

$$W = [\mathbf{w}_1,\dots,\mathbf{w}_n], \qquad W^{-1} = \begin{bmatrix} \widehat{\mathbf{w}}^1 \\ \vdots \\ \widehat{\mathbf{w}}^n \end{bmatrix}.$$

Theorem 2.10.1. *Let $\mathbf{x}^{(k)} \in \mathbb{R}^n$ and let $\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\widetilde{\mathbf{s}}^{(k)}$ be the Gauss–Newton iteration for (2.1), where the step $\widetilde{\mathbf{s}}^{(k)}$ is determined by solving (2.10) and the step length α_k by the Armijo–Goldstein principle. Then, the iteration $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{s}^{(k)}$ for (2.55), is given by*

$$\mathbf{x}^{(k+1)} = \widetilde{\mathbf{x}}^{(k+1)} - W_1\widehat{W}_1\left(\mathbf{x}^{(k)}-\overline{\mathbf{x}}\right), \tag{2.56}$$

where $\widehat{W}_1 \in \mathbb{R}^{(n-r_k)\times n}$ contains the first $n-r_k$ rows of W^{-1}, and $W_1 \in \mathbb{R}^{n\times(n-r_k)}$ is composed of the first $n-r_k$ columns of W.

Proof. The proof proceeds analogously to that of Theorem 2.4.2. Replacing J_k and L with their GSVD and setting $\mathbf{y} = W^{-1}\mathbf{s}$, $\mathbf{z}^{(k)} = W^{-1}\left(\mathbf{x}^{(k)}-\overline{\mathbf{x}}\right)$, and $\mathbf{g}^{(k)} = U^T\mathbf{r}_k$, (2.55) can be rewritten as the following diagonal least-squares problem

$$\begin{cases} \min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma_L(\alpha_k\mathbf{y}+\mathbf{z}^{(k)})\|^2 \\ \mathbf{y} \in \left\{\arg\min\limits_{\mathbf{y}\in\mathbb{R}^n} \|\Sigma_J\mathbf{y}+\mathbf{g}^{(k)}\|^2\right\}. \end{cases}$$

When $m \geq n$, the diagonal linear system in the constraint is solved by a vector $\mathbf{y}$ with entries

$$y_i = \begin{cases} -\dfrac{g_i^{(k)}}{c_{i-n+r_k}}, & i = n - r_k + 1, \ldots, p, \\ -g_i^{(k)}, & i = p+1, \ldots, n. \end{cases}$$

The components y_i, for $i = 1, \ldots, n - r_k$, can be determined by minimizing the norm

$$\begin{aligned} \|\Sigma_L(\alpha_k \mathbf{y} + \mathbf{z}^{(k)})\|^2 = &\sum_{i=1}^{n-r_k} \left(\alpha_k y_i + z_i^{(k)}\right)^2 \\ &+ \sum_{i=n-r_k+1}^{p} \left(-\alpha_k \frac{g_i^{(k)}}{\gamma_{i-n+r_k}} + s_{i-n+r_k} z_i^{(k)}\right)^2, \end{aligned} \tag{2.57}$$

where $\gamma_i = \frac{c_i}{s_i}$ are the generalized singular values of the matrix pair (J_k, L). The minimum of (2.57) is reached for $y_i = -\frac{1}{\alpha_k} z_i^{(k)} = -\frac{1}{\alpha_k} \widehat{\mathbf{w}}^i (\mathbf{x}^{(k)} - \overline{\mathbf{x}})$, $i = 1, \ldots, n - r_k$, and the solution to (2.55), that is, the next approximation to the solution of (2.54), is

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k W \mathbf{y} \\ &= \mathbf{x}^{(k)} - \sum_{i=1}^{n-r_k} z_i^{(k)} \mathbf{w}_i - \alpha_k \sum_{i=n-r_k+1}^{p} \frac{g_i^{(k)}}{c_{i-n+r_k}} \mathbf{w}_i - \alpha_k \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i, \end{aligned} \tag{2.58}$$

where the first summation in the right-hand side can be rewritten as $W_1 \widehat{W}_1 (\mathbf{x}^{(k)} - \overline{\mathbf{x}})$. Applying the same procedure to (2.10), we obtain

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \sum_{i=n-r_k+1}^{p} \frac{g_i^{(k)}}{c_{i-n+r_k}} \mathbf{w}_i - \alpha_k \sum_{i=p+1}^{n} g_i^{(k)} \mathbf{w}_i,$$

from which (2.56) follows. Since solving (2.55) for $m < n$ leads to a formula similar to (2.58), with $g_{i-n+m}^{(k)}$ in place of $g_i^{(k)}$, the validity of (2.56) is confirmed. □

As in the computation of the minimal-norm solution, the iteration based on (2.56) fails to converge without a suitable relaxation parameter β_k for the projection vector $\mathbf{t}^{(k)} = W_1 \widehat{W}_1 (\mathbf{x}^{(k)} - \overline{\mathbf{x}})$. We adopted an iteration similar to (2.51), choosing β_k by adapting Algorithms 1 and 2 to this setting. It is important to note that $\widetilde{\mathcal{P}}_{\mathcal{N}(J_k)} = W_1 \widehat{W}_1$ is an oblique projector onto $\mathcal{N}(J_k)$.

At the same time, the rank of the Jacobian is estimated at each step by applying the procedure described in Section 2.8 to the diagonal elements $c_j^{(k)}$, $j = 1, \ldots, q - d$, of the GSVD factor Σ_J of J_k; see equations (1.4) and (1.6). In this case, at each step, we compute the ratios

$$\rho_i^{(k)} = \frac{c_{i+1}^{(k)}}{c_i^{(k)}}, \qquad i = 1, 2, \ldots, q - d - 1,$$

where $q = \min(m, n)$.

Actually, the GSVD routine computes the matrix W^{-1}, but the matrix W is needed for the computation of both the vectors $\widetilde{\mathbf{s}}^{(k)}$ and $\mathbf{t}^{(k)}$. To reduce the computational load, we compute at each iteration the LU factorization $PW^{-1} = LU$, and we use it to solve the linear system with two right-hand sides

$$W^{-1} \begin{bmatrix} \mathbf{t}^{(k)} & \widetilde{\mathbf{s}}^{(k)} \end{bmatrix} = \begin{bmatrix} \widehat{W}_1(\mathbf{x}^{(k)} - \overline{\mathbf{x}}) & \mathbf{0}_{n-r} \\ \mathbf{0}_r & \widetilde{\mathbf{y}} \end{bmatrix},$$

where $\widetilde{\mathbf{y}} \in \mathbb{R}^r$ contains the last r components of the vector $\mathbf{y}$ appearing in (2.58), and $\mathbf{0}_k$ denotes the zero vector of size k.

2.11 Conclusions

This chapter explores the solution of a nonlinear least-squares problem in the case its solution lacks uniqueness. The usual approach is to compute the minimal-norm solution to a linearization of the problem, generating an iterative method which does not guarantee that the converged solution itself has a minimal-norm, or minimizes a suitable seminorm. Here, we develop various techniques to impose such constraint on the solution.

In the case of ill-conditioned problems, we also propose two regularization algorithms, namely the truncated minimal-L-norm Gauss–Newton method and the minimal-L-norm Tikhonov–Gauss–Newton method. In the numerical experiments (see Section 6.1), we compare the newly proposed methods to the classical approaches. The results show that the two classes of methods produce, in general, different results. The new methods are in some cases less sensitive to the initial guess without a significant increase in the computational load.

The second part of the chapter explores the reasons for the occasional lack of convergence of the minimal-norm Gauss–Newton method. We propose an automatic procedure to estimate the rank of the Jacobian along the iteration, and the introduction of two different relaxation parameters that improve the efficiency of the iterative method. The first parameter is determined by applying the Armijo–Goldstein principle, while three techniques are investigated to estimate the second one. In numerical experiments (see Section 6.2) performed on various test problems, the new methods prove to be very effective, compared to other approaches based on a single damping parameter. In particular, the variant which automatically estimates the projection parameter gives satisfactory results in all the examples.

CHAPTER 3

Large-scale minimal-norm solution

In this chapter, we see how to handle the problem of computing the minimal-norm solution of a nonlinear least-squares problem in a large-scale setting.

3.1 Golub–Kahan bidiagonalization

In [82, 83], formula

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}^{(k)} - \beta_k \mathcal{P}_{\mathcal{N}(J_k)}\big(\mathbf{x}^{(k)} - \overline{\mathbf{x}}\big),$$

is implemented by computing the singular value decomposition (SVD) of the matrix J_k at each step of the iterative methods. Although the SVD is a powerful tool for the analysis of inverse problems, it is feasible to compute it only for small-scale problems. For large-scale problems, one must turn to iterative methods that realize a partial factorization of the matrix J_k. In such a way, large-scale least-squares problems

$$\min_{\mathbf{s}\in\mathbb{R}^n} \|J_k\mathbf{s} + \mathbf{r}_k\|^2 \tag{3.1}$$

are reduced to small size by carrying out a few steps of the Golub–Kahan bidiagonalization process [43]. This is the basis for the LSQR algorithm by Paige and Saunders [79, 80], where the matrix J_k is only used to compute products of the form $J_k\mathbf{v}$ and $J_k^T\mathbf{u}$ for various vectors $\mathbf{v}$ and $\mathbf{u}$.

Application of ℓ Golub–Kahan bidiagonalization steps to J_k with initial vector $\mathbf{r}_k$ yields the decompositions

$$\begin{aligned} J_k V_\ell &= U_{\ell+1} C_{\ell+1,\ell} \\ J_k^T U_\ell &= V_\ell C_{\ell,\ell}^T, \end{aligned} \tag{3.2}$$

where the matrices $U_{\ell+1} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{\ell+1}] \in \mathbb{R}^{m\times(\ell+1)}$ and $V_\ell = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\ell] \in \mathbb{R}^{n\times\ell}$ have orthonormal columns, with $\mathbf{u}_1 = \mathbf{r}_k/\|\mathbf{r}_k\|$, and $U_\ell \in \mathbb{R}^{m\times\ell}$ consists of the

first ℓ columns of $U_{\ell+1}$. Finally,

$$C_{\ell+1,\ell} = \begin{bmatrix} \rho_1 & & & \\ \sigma_2 & \rho_2 & & \\ & \ddots & \ddots & \\ & & \sigma_\ell & \rho_\ell \\ & & & \sigma_{\ell+1} \end{bmatrix} \in \mathbb{R}^{(\ell+1)\times\ell} \tag{3.3}$$

is lower bidiagonal and $C_{\ell,\ell}$ is its leading $\ell \times \ell$ submatrix. In the decomposition (3.2), to ease the notation, we drop the dependence on the iteration index k.

Starting with a vector $\mathbf{u}_1 \in \mathbb{R}^m$, with $\|\mathbf{u}_1\| = 1$, and setting $\mathbf{v}_0 = \mathbf{0}$, the algorithm recursively generates the vectors $\mathbf{v}_i$, $\mathbf{u}_{i+1}$, $i = 1, \ldots, \ell$, and the non-zero elements in $C_{\ell+1,\ell}$ by the recursion

$$\begin{aligned} \widetilde{\mathbf{v}}_i &= J_k^T \mathbf{u}_i - \sigma_i \mathbf{v}_{i-1}, & \rho_i &= \|\widetilde{\mathbf{v}}_i\|, & \mathbf{v}_i &= (\rho_i)^{-1} \widetilde{\mathbf{v}}_i, \\ \widetilde{\mathbf{u}}_{i+1} &= J_k \mathbf{v}_i - \rho_i \mathbf{u}_i, & \sigma_{i+1} &= \|\widetilde{\mathbf{u}}_{i+1}\|, & \mathbf{u}_{i+1} &= (\sigma_{i+1})^{-1} \widetilde{\mathbf{u}}_{i+1}, \end{aligned} \tag{3.4}$$

for $i = 1, 2, \ldots, \ell$. The computation requires ℓ matrix-vector product evaluations with J_k, and ℓ matrix-vector product evaluations with J_k^T.

If exact arithmetic were used, then one would have $U_{\ell+1}^T U_{\ell+1} = I_{\ell+1}$ and $V_\ell^T V_\ell = I_\ell$. In finite precision arithmetic, to avoid loss of orthogonality in the columns of $U_{\ell+1}$ and V_ℓ, reorthogonalization is usually employed.

For the moment, we assume that the non-trivial entries of the bidiagonal matrix (3.3) are positive for all ℓ. When either ρ_ℓ or $\sigma_{\ell+1}$ are zero at step ℓ a *breakdown* occurs. The Golub–Kahan process will lead to a breakdown for $i > r_k = \operatorname{rank}(J_k)$. Anyway, a breakdown may happen even for $i \leq r_k$, we will comment later on its consequences.

The columns of $U_{\ell+1}$ and V_ℓ form an orthonormal basis for the following Krylov subspaces

$$\begin{aligned} \mathcal{K}_{\ell+1}^{U,k} &:= \mathcal{K}_{\ell+1}(J_k J_k^T, \mathbf{r}_k) = \mathcal{R}(U_{\ell+1}), \\ \mathcal{K}_{\ell}^{V,k} &:= \mathcal{K}_{\ell}(J_k^T J_k, J_k^T \mathbf{r}_k) = \mathcal{R}(V_\ell), \end{aligned} \tag{3.5}$$

where

$$\mathcal{K}_\ell(A, \mathbf{b}) = \operatorname{span}\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \ldots, A^{\ell-1}\mathbf{b}\}.$$

It follows from (3.2) and $U_{\ell+1}^T U_{\ell+1} = I_{\ell+1}$ that

$$U_{\ell+1}^T J_k V_\ell = C_{\ell+1,\ell}.$$

LSQR is an iterative method for solving (3.1) based on the decomposition (3.2), which is equivalent to applying the conjugate gradient iteration to the normal equations associated with the problem. Let the initial iterate be $\mathbf{s}_0 = \mathbf{0}$. Then, the ℓth iterate $\mathbf{s}_\ell$ determined by the LSQR method satisfies

$$\|J_k \mathbf{s}_\ell + \mathbf{r}_k\| = \min_{\mathbf{s} \in \mathcal{K}_\ell^{V,k}} \|J_k \mathbf{s} + \mathbf{r}_k\|. \tag{3.6}$$

This shows that LSQR is a so-called minimal residual method: the iterate $\mathbf{s}_\ell$ minimizes the residual error over the Krylov subspace $\mathcal{K}_\ell^{V,k}$ defined in (3.5). Substituting $\mathbf{s} = V_\ell \mathbf{y}$ into the right-hand side of (3.6) and using the decompositions (3.2), yields

$$\begin{aligned} \min_{\mathbf{s}\in\mathcal{K}_\ell^{V,k}} \|J_k\mathbf{s} + \mathbf{r}_k\| &= \min_{\mathbf{y}\in\mathbb{R}^\ell} \|J_k V_\ell \mathbf{y} + \mathbf{r}_k\| = \min_{\mathbf{y}\in\mathbb{R}^\ell} \|U_{\ell+1} C_{\ell+1,\ell}\mathbf{y} + \mathbf{r}_k\| \\ &= \min_{\mathbf{y}\in\mathbb{R}^\ell} \left\| C_{\ell+1,\ell}\mathbf{y} + \|\mathbf{r}_k\|\mathbf{e}_1 \right\|, \end{aligned} \tag{3.7}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{\ell+1}$.

By employing the QR factorization

$$C_{\ell+1,\ell} = Q_{\ell+1} R_{\ell+1,\ell},$$

in the reduced least-squares problem on the right-hand side of (3.7), the ℓth iterate can be expressed as $\mathbf{s}_\ell = V_\ell \mathbf{y}_\ell$, where $\mathbf{y}_\ell$ solves the linear system

$$R_{\ell,\ell}\mathbf{y}_\ell = -\|\mathbf{r}_k\|\mathbf{q}_1,$$

$R_{\ell,\ell}$ is the bidiagonal $\ell \times \ell$ upper block of $R_{\ell+1,\ell}$, and $\mathbf{q}^T = \left[\mathbf{q}_1^T, q_{\ell+1}\right]$ is the first row of the matrix $Q_{\ell+1}$. Moreover,

$$\min_{\mathbf{s}\in\mathcal{K}_\ell^{V,k}} \|J_k\mathbf{s}_\ell + \mathbf{r}_k\|^2 = (\|\mathbf{r}_k\| q_{\ell+1})^2.$$

Since $\mathcal{K}_{\ell-1}^{V,k} \subseteq \mathcal{K}_\ell^{V,k}$ the method generates a sequence of approximations $\mathbf{s}_\ell$ such that the residual error $\|J_k\mathbf{s}_\ell + \mathbf{r}_k\|$ decreases monotonically when ℓ increases.

The LSQR is outlined in Algorithm 3. For further background on Golub–Kahan bidiagonalization, Lanczos methods, and Krylov subspaces, see [10, 91].

A proof of the fact that the LSQR algorithm converges to the minimum norm solution can be found in [40, Theorem 4.2]. Since it is essential to solve problem (2.12), we review the following results.

Theorem 3.1.1. *The Krylov spaces $\mathcal{K}_\ell^{V,k}$ generated by the Golub–Kahan process for $\ell = 1, \ldots, r_k = \operatorname{rank}(J_k)$, are orthogonal to the null space of J_k.*

Proof. From (3.4), we have that $\mathbf{u}_i \in \mathcal{R}(J_k)$ if and only if $\mathbf{r}_k \in \mathcal{R}(J_k)$ and $\mathbf{v}_i \in \mathcal{R}(J_k^T) = \mathcal{N}(J_k)^\perp$, independently of $\mathbf{r}_k$, for $i = 1, \ldots, r_k$. Since $\mathcal{K}_\ell^{V,k} = \mathcal{R}(V_\ell)$, the result follows. □

Theorem 3.1.2. *The LSQR method converges to the minimal-norm solution of (2.6). If a breakdown occurs, then LSQR finds the exact solution.*

Proof. The fact that LSQR converges to a minimizer of the residual $\|J_k\mathbf{s} + \mathbf{r}_k\|$ has been proved in [80]. In [68, Theorem 2], it is shown that, for each ℓ,

$$\|J_k^T(J_k\mathbf{s}_\ell + \mathbf{r}_k)\| = \rho_{\ell+1}\sigma_{\ell+1}|\mathbf{e}_\ell^T\mathbf{y}_\ell|, \tag{3.8}$$

Algorithm 3 Outline of the LSQR algorithm.

Require: Jacobian matrix J_k, residual vector $\mathbf{r}_k$, threshold *tol* for breakdown
Ensure: U_ℓ, V_ℓ, $C_{\ell+1,\ell}$, and $\mathbf{s}_\ell^{(k)}$

$\mathbf{v}_0 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{r}_k / \|\mathbf{r}_k\|$, and $i = 1$
repeat
  $\widetilde{\mathbf{v}}_i = J_k^T \mathbf{u}_i - \sigma_i \mathbf{v}_{i-1}$
  reorthogonalization $\widetilde{\mathbf{v}}_i = \widetilde{\mathbf{v}}_i - \sum_{j=1}^{i-1} \langle \mathbf{v}_j, \widetilde{\mathbf{v}}_i \rangle \mathbf{v}_j$
  $\rho_i = \|\widetilde{\mathbf{v}}_i\|$
  if $\rho_i < tol$ **then**
    breakdown
  end if
  $\mathbf{v}_i = (\rho_i)^{-1} \widetilde{\mathbf{v}}_i$
  $\widetilde{\mathbf{u}}_{i+1} = J_k \mathbf{v}_i - \rho_i \mathbf{u}_i$
  reorthogonalization $\widetilde{\mathbf{u}}_{i+1} = \widetilde{\mathbf{u}}_{i+1} - \sum_{j=1}^{i} \langle \mathbf{u}_j, \widetilde{\mathbf{u}}_i \rangle \mathbf{u}_j$
  $\sigma_{i+1} = \|\widetilde{\mathbf{u}}_{i+1}\|$
  if $\sigma_{i+1} < tol$ **then**
    breakdown
  end if
  $\mathbf{u}_{i+1} = (\sigma_{i+1})^{-1} \widetilde{\mathbf{u}}_{i+1}$
  solve $\min_{\mathbf{y} \in \mathbb{R}^i} \left\| C_{i+1,i} \mathbf{y} + \|\mathbf{r}_k\| \mathbf{e}_1 \right\|$
  $i = i + 1$
until convergence
$\ell = i, \quad \mathbf{s}_\ell^{(k)} = V_\ell \mathbf{y}_\ell$

so that if the algorithm breaks down at step $\ell < r_k \leq \min(m,n)$, then the ℓth iterate of LSQR $\mathbf{s}_\ell = V_\ell \mathbf{y}_\ell \in \mathcal{K}_\ell^{V,k}$ minimizes the residual, as (3.8) is equivalent to normal equations of (3.1). If this is the case, $\mathbf{s}_\ell$ belongs to $\mathcal{N}(J_k)^\perp$ by Theorem 3.1.1, so it is the minimal-norm solution. The same happens, with $\ell = r_k$, if no breakdown occurs and $r_k = m < n$. If $m \geq n$, the least-squares solution is unique. □

3.2 Breakdowns

Some properties related to the breakdown will now be analyzed. Combining the decompositions (3.2) we obtain

$$\begin{aligned} J_k J_k^T U_\ell &= U_{\ell+1} T_{\ell+1,\ell}, \\ J_k^T J_k V_\ell &= V_{\ell+1} \widehat{T}_{\ell+1,\ell}, \end{aligned} \tag{3.9}$$

where

$$T_{\ell+1,\ell} = C_{\ell+1,\ell} C_{\ell,\ell}^T, \qquad \widehat{T}_{\ell+1,\ell} = C_{\ell+1,\ell+1}^T C_{\ell+1,\ell},$$

are tridiagonal matrices of the form

$$\begin{bmatrix} \gamma_1 & \delta_2 & & & \\ \delta_2 & \gamma_2 & \delta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{\ell-1} & \gamma_{\ell-1} & \delta_\ell \\ & & & \delta_\ell & \gamma_\ell \\ & & & & \delta_{\ell+1} \end{bmatrix} \in \mathbb{R}^{(\ell+1)\times \ell},$$

and $T_{\ell,\ell}$ and $\widehat{T}_{\ell,\ell}$ are their leading $\ell \times \ell$ symmetric submatrices. Equations (3.9) are Lanczos decompositions of the symmetric positive semidefinite matrices $J_k J_k^T$ and $J_k^T J_k$, respectively. They allow one to apply the theorem stated for the Lanczos decomposition in [95, Theorem 36.1] to the Golub–Kahan decomposition.

Theorem 3.2.1. *As long as the Golub–Kahan iteration does not break down, the characteristic polynomial of $T_{j,j}$ is the unique polynomial p_j of degree j such that $\|p_j(J_k J_k^T)\mathbf{r}_k\|$ is minimum, and the characteristic polynomial of $\widehat{T}_{j,j}$ is the unique polynomial $\widehat{p}_j$ of degree j such that $\|\widehat{p}_j(J_k^T J_k) J_k^T \mathbf{r}_k\|$ is minimum.*

If a breakdown happens for $\sigma_{\ell+1}$ at step ℓ, then $p_\ell(J_k J_k^T)\mathbf{r}_k = \mathbf{0}$.

If a breakdown happens for ρ_ℓ at step ℓ, then $\widehat{p}_{\ell-1}(J_k^T J_k) J_k^T \mathbf{r}_k = \mathbf{0}$.

If the breakdown occurs at the first steps in the Golub–Kahan process, it is possible to obtain an explicit expression for the above polynomial equations.

Corollary 3.2.2. *If at the first step of the Golub–Kahan bidiagonalization $\sigma_2 = 0$, then*

$$J_k J_k^T \mathbf{r}_k - \rho_1^2 \mathbf{r}_k = \mathbf{0}, \tag{3.10}$$

that is, $\mathbf{r}_k$ *is an eigenvector of* $J_k J_k^T$ *and* ρ_1 *is a singular value of* J_k*. If a breakdown occurs at the second step, then if* $\rho_2 = 0$

$$(J_k^T J_k) J_k^T \mathbf{r}_k - (\rho_1^2 + \sigma_2^2) J_k^T \mathbf{r}_k = \mathbf{0},$$

and if $\sigma_3 = 0$

$$\left[(J_k J_k^T)^2 - (\rho_1^2 + \rho_2^2 + \sigma_2^2) J_k J_k^T + \rho_1^2 \rho_2^2 I_m \right] \mathbf{r}_k = \mathbf{0}. \tag{3.11}$$

Proof. The results follow easily from (3.2). For example, if $\sigma_3 = 0$ we have

$$J_k \mathbf{v}_2 = \rho_2 \mathbf{u}_2. \tag{3.12}$$

Substituting $\mathbf{v}_2$, $\mathbf{u}_2$, and $\mathbf{v}_1$, retracing the algorithm (3.4) in reverse, we get

$$J_k \mathbf{v}_2 = \frac{1}{\rho_1 \rho_2 \sigma_2} \left[(J_k J_k^T)^2 - (\rho_1^2 + \sigma_2^2) J_k J_k^T \right] \mathbf{u}_1,$$

and

$$\rho_2 \mathbf{u}_2 = \frac{\rho_2}{\rho_1 \sigma_2} \left[J_k J_k^T \mathbf{u}_1 - \rho_1^2 \mathbf{u}_1 \right].$$

Replacing these two expressions in (3.12), since $\mathbf{u}_1 = \mathbf{r}_k / \|\mathbf{r}_k\|$, yields (3.11). □

In the next example, we illustrate a particular function which presents a breakdown at the first step ($\sigma_2 = 0$).

Example 3.2.3. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function (5.5) defined in Test Function 4 (see Section 5.1). The function (5.5) can be written in vectorial form as

$$F(\mathbf{x}) = S(\mathbf{x}) I_{m \times n} (\mathbf{x} - \mathbf{c}),$$

and its Jacobian matrix is

$$J(\mathbf{x}) = I_{m \times n} \left[S(\mathbf{x}) I_n + 2(\mathbf{x} - \mathbf{c})(\mathbf{x} - \mathbf{c})^T D \right],$$

where

$$S(\mathbf{x}) = \sum_{j=1}^{n} \left(\frac{x_j - c_j}{a_j} \right)^2 - 1$$

is the n-ellipsoid with center $\mathbf{c} = [c_1, \ldots, c_n]^T$ and whose semiaxes are the components of the vector $\mathbf{a} = [a_1, \ldots, a_n]^T$, and

$$D = \operatorname{diag} \left(\frac{1}{a_1^2}, \frac{1}{a_2^2}, \ldots, \frac{1}{a_n^2} \right).$$

If $\mathbf{a} = \mathbf{e} = [1, 1, \ldots, 1]^T$, i.e., $D = I_n$, that is, $S(\mathbf{x})$ is a sphere, a breakdown for σ_ℓ occurs at the first step of the Golub–Kahan bidiagonalization, i.e., $\sigma_2 = 0$; see Corollary 3.2.2. Since $\mathbf{b} = \mathbf{0}$, then $\mathbf{r}(\mathbf{x}) = F(\mathbf{x})$. From (3.4) it follows $\rho_1^2 = \|J^T \mathbf{r}\|^2 / \|\mathbf{r}\|^2$. It is simple to verify that $J J^T \mathbf{r}$ and $\rho_1^2 \mathbf{r}$ are both equal to

$$S(\mathbf{x}) \left[S(\mathbf{x})^2 + 4(2S(\mathbf{x}) + 1) \|I_{m \times n} (\mathbf{x} - \mathbf{c})\|^2 \right] I_{m \times n} (\mathbf{x} - \mathbf{c}),$$

that is, equation (3.10) is verified.

3.3 Minimal-norm solution

Let $\mathcal{P}_{\mathcal{N}(J_k)}$ represent the orthogonal projector onto $\mathcal{N}(J_k)$. Following [83], the approximate solution of (2.12) is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \widetilde{\mathbf{s}}_{\ell_k}^{(k)} - \beta_k \mathcal{P}_{\mathcal{N}(J_k)}(\mathbf{x}^{(k)} - \overline{\mathbf{x}}),$$

where k is the iteration index of the MNGN2 method. It follows that the $(k+1)$th iterate of the MNGN2 method, when $\beta_k = 1$, is orthogonal to the null space of J_k. The parameter β_k is chosen by applying the method described in Section 2.9.

We remark that the orthogonal projector onto the null space of J_k is approximated by the orthogonal projector onto the null space of $J_k^{(\ell)}$, where the matrix $J_k^{(\ell)} = U_{\ell+1} C_{\ell+1,\ell} V_\ell^T$ is a rank-ℓ approximation of the matrix J_k obtained after ℓ steps of the Golub–Kahan bidiagonalization.

Lemma 3.3.1. *The null space of $J_k^{(\ell)}$ is orthogonal to the Krylov subspace $\mathcal{K}_\ell(J_k^T J_k, J_k^T \mathbf{r}_k)$.*

Proof. Since the null space of $J_k^{(\ell)}$ coincides with the null space of V_ℓ^T, indeed $J_k^{(\ell)}\mathbf{x} = \mathbf{0}$ if and only if $V_\ell^T \mathbf{x} = \mathbf{0}$, and $\mathcal{N}(V_\ell^T)$ is orthogonal to the range $\mathcal{R}(V_\ell)$, equation (3.5) completes the proof. □

From this result, it follows that the orthogonal projector onto the null space of $J_k^{(\ell)}$ is given by

$$\mathcal{P}_{\mathcal{N}(J_k)} \approx \mathcal{P}_{\mathcal{N}(J_k^{(\ell)})} = I_n - V_\ell V_\ell^T.$$

A natural extension is to develop an analogous method for solving the minimization problem (2.15), where the seminorm of the solution is minimized.

3.4 Tikhonov regularization

Regularization is achieved by terminating the iterations sufficiently early. Another approach is to apply the Tikhonov regularization.

The minimization problem (2.6) is replaced by a nearby problem

$$\min_{\mathbf{s}\in\mathbb{R}^n} \left\{ \|J_k \mathbf{s} + \mathbf{r}_k\|^2 + \lambda^2 \|\mathbf{s}\|^2 \right\},$$

where $\lambda > 0$ is a regularization parameter. The unique solution is given by

$$\mathbf{s}_\lambda = -\left(J_k^T J_k + \lambda^2 I_n\right)^{-1} J_k^T \mathbf{r}_k.$$

To choose λ, the quantity $\|J_k \mathbf{s}_\ell + \mathbf{r}_k\|$ has to be evaluated for several λ-values. This can be expensive when the matrix J_k is large. We first reduce J_k to a small

bidiagonal matrix with the aid of Golub–Kahan bidiagonalization. After ℓ steps of Golub–Kahan bidiagonalization, the functional becomes

$$\min_{\mathbf{y}\in\mathbb{R}^{\ell}} \left\{ \left\| C_{\ell+1,\ell}\mathbf{y} + \|\mathbf{r}_k\|\mathbf{e}_1 \right\|^2 + \lambda^2 \|\mathbf{y}\|^2 \right\}.$$

The normal equations related to this functional are

$$\left(C_{\ell+1,\ell}^T C_{\ell+1,\ell} + \lambda^2 I_{\ell}\right)\mathbf{y} = -\|\mathbf{r}_k\| C_{\ell+1,\ell}^T \mathbf{e}_1$$

If the Tikhonov functional is in general form, i.e., there is a regularization matrix $L \in \mathbb{R}^{p\times n}$,

$$\min_{\mathbf{s}\in\mathbb{R}^{n}} \left\{ \|J_k\mathbf{s} + \mathbf{r}_k\|^2 + \lambda^2 \|L\mathbf{s}\|^2 \right\},$$

it is prohibitively expensive to compute the GSVD of a pair of large matrices. Tikhonov regularization problems with large matrices J_k and L have to be reduced to problems of small size. We first reduce J_k by applying ℓ steps of Golub–Kahan bidiagonalization (3.2). Then, we introduce the QR factorization of the "slim" matrix

$$LV_{\ell} = Q_{\ell}R_{\ell},$$

where $Q_{\ell} \in \mathbb{R}^{p\times\ell}$ has orthonormal columns and $R_{\ell} \in \mathbb{R}^{\ell\times\ell}$ is upper triangular; see [81]. This approach is computationally convenient because $\ell \ll p$ and the QR factorization can be realized by only ℓ Householder transformations. We seek a solution in the subspace $\mathcal{K}_{\ell}^{V,k}$. Thus, we solve the Tikhonov minimization problem

$$\min_{\mathbf{y}\in\mathbb{R}^{\ell}} \left\{ \left\| C_{\ell+1,\ell}\mathbf{y} + \|\mathbf{r}_k\|\mathbf{e}_1 \right\|^2 + \lambda^2 \|R_{\ell}\mathbf{y}\|^2 \right\},$$

with $\mathbf{s} = V_{\ell}\mathbf{y}$.

CHAPTER 4

Minimal-norm solution of first kind integral equations

4.1 Introduction

Fredholm integral equations of the first kind model several physical problems arising in different contexts such as medical imaging, image processing, signal processing and geophysics. Their standard form is

$$\int_a^b k(x,t)f(t)\,dt = g(x), \qquad x \in [c,d], \tag{4.1}$$

where the right-hand side g, usually given at a finite set of points $x = x_i$, $i = 1,\dots,n$, represents the experimental data, the kernel k, often analytically known, stands for the impulse response of the experimental equipment, and the function f is the signal to recover.

From a theoretical point of view, they are treated in a Hilbert space setting which typically coincides with the space of square-integrable functions. The corresponding integral operator

$$(Kf)(x) = \int_a^b k(x,t)f(t)\,dt,$$

is a bounded linear operator from a Hilbert space H_1 into a Hilbert space H_2, and a solution f of (4.1) exists only if the right-hand side g belongs to the range of K, $\mathcal{R}(K) \subset H_2$. Consequently, the existence of the solution of (4.1) cannot be guaranteed for any right-hand side, but only for a restricted class of functions g [51]. The uniqueness of the solution depends upon the structure of the null space of the operator K, but even when it is ensured the problem is still ill-posed since the stability is missing; see [50, pag. 155].

In an experimental setting, g is certainly an element of $\mathcal{R}(K)$, since it represents the data $g(x_i)$ produced by an operator K which reproduces a real situation. This

leads to the integral equation with discrete data

$$\int_a^b k(x_i,t)f(t)\,dt = g(x_i), \quad i = 1,\dots,n. \tag{4.2}$$

However, even when $g \in \mathcal{R}(K)$, the data values in (4.2) are affected by perturbations due to measuring and rounding errors, so one cannot be sure that the perturbed right-hand side lies exactly in the range of K. Moreover, the solution of (4.2) is not unique and it does not depend continuously on the data. In other words, a discretization (4.2) of equation (4.1) is an ill-posed problem [52, 100]. This fact makes its numerical treatment rather delicate, especially if compared to the discretization of integral equations of the second kind, a typical example of well-posed problem [5].

The non-uniqueness of the solution of (4.2) can be stated as follows. Let us consider the functions $k_i(t) = k(x_i,t)$, $i = 1,\dots,n$. By the Gram–Schmidt process it is possible to construct a set of orthonormal functions $\phi_j(t)$, $j = 1,\dots,\bar{n} \le n$, such that

$$\mathcal{S} = \operatorname{span}\{\phi_1,\dots,\phi_{\bar{n}}\} = \operatorname{span}\{k_1,\dots,k_n\}.$$

Chosen any function $\psi(t)$ linearly independent of $k_i(t)$, $i = 1,\dots,n$, the function

$$\phi_{\bar{n}+1}(t) = \psi(t) - \sum_{j=1}^{\bar{n}} \langle \psi, \phi_j \rangle \phi_j$$

is orthogonal to $\mathcal{S}$, so that whenever $f(t)$ is a solution of (4.2) also $f(t) + \alpha\phi_{\bar{n}+1}(t)$ is, for any $\alpha \in \mathbb{R}$.

Of course, if we move to a system of integral equations of the first kind the situation is not different, at least if the number of equations and unknown functions is the same.

Overdetermined systems of such equations, that is, at least two equations whose solution is a single unknown function, arise in a variety of applications. Indeed, specific physical systems can be observed by different devices, or by the same device with different configurations, and this fact results in writing distinct equations with the same unknown. An example is given by the geophysical model presented in [76]; see also Test Function 10 in Section 5.3. The model reproduces the readings of a ground conductivity meter, a device composed of two coils, a transmitter and a receiver, placed at a fixed distance from each other. It reads as two integral equations of the first kind involving the same unknown function, representing the electrical conductivity of the soil at a certain depth; see equations (5.20). The first equation describes the situation in which both coil axes are aligned vertically with respect to the ground level, while the second one corresponds to the horizontal orientation of the coils. This system has been studied in [33], under the assumption that the values of the unknown function at the boundaries are known, either on the basis of additional measurements or of known geophysical properties of the subsoil.

Further applications are the model considered in [62], and the Radon transform [97, 98]. In all these situations, the model is written in terms of an overdetermined system and boundary a priori information on the signal to recover may be known.

In this chapter and in [32], motivated by these applications and with the purpose of developing a method that can be applied to different physical models, we focus on the following system of m integral equations of the first kind

$$\begin{cases} \displaystyle\int_a^b k_\ell(x,t)\, f(t)\, dt = g_\ell(x), \quad \ell = 1,\dots,m, \quad x \in [c,d], \\ f(a) = f_0,\ \ f(b) = f_1, \end{cases} \tag{4.3}$$

where k_ℓ and g_ℓ are the given kernel and right-hand side of the ℓth equation, respectively, and f is the function to be determined satisfying known constraints at the boundary. Specifically, given the data at a finite (and often small in applications) set of points $x_{\ell,i} \in [c_\ell, d_\ell]$, $i = 1,\dots,n_\ell$, we aim at solving the problem with discrete data

$$\begin{cases} \displaystyle\int_a^b k_\ell(x_{\ell,i},t)\, f(t)\, dt = g_\ell(x_{\ell,i}), \qquad \ell = 1,\dots,m, \quad i = 1,\dots,n_\ell, \\ f(a) = f_0,\ \ f(b) = f_1. \end{cases} \tag{4.4}$$

As it is well-known, such a problem has infinitely many solutions. Our aim is to construct a method that selects, among all the possible solutions, the one having a certain degree of regularity. We reformulate the problem as a minimal-norm least-squares problem, and solve the latter in suitable function spaces. While this approach is rather standard in functional analysis, it has never been applied to an overdetermined system. Moreover, as we will show, the corresponding algorithm proves to be very accurate in the absence of experimental errors and it naturally leads to an effective regularization technique, when the data is affected by noise.

Specifically, we consider a reproducing kernel Hilbert space where, by using the Riesz theory, the minimal-norm solution can be written as a linear combination of the so-called Riesz representers. Then, the main issue is to determine the Riesz functions as well as the coefficients of such a linear combination. The first ones, which are determined by the reproducing kernel, are expressed in terms of integrals which need suitable quadrature schemes, whenever they cannot be evaluated analytically. The latter ones are obtained by solving a square ill-conditioned linear system. If the data is only affected by rounding errors, this representation proves to be accurate. If the noise level is realistic, as one would expect, the error propagation completely cancels the solution and a regularized approach is required.

To this end, we construct a regularization method to solve problem (4.4), based on a truncated expansion in terms of the singular functions of the corresponding integral operator. To improve steadiness, the singular system is not explicitly used in the construction of the regularized solution, which is still represented as a linear combination of the Riesz representers instead. We prove that the coefficients of such

regularized expansion are obtained by applying the truncated eigenvalue decomposition to the initial ill-conditioned linear system. The truncation index is, in fact, a regularization parameter, which we determine by different estimation approaches. The effectiveness of the resulting solution method is confirmed by numerical experiments, which involve both artificial examples and an integral model reproducing the propagation of an electromagnetic field in the earth soil.

We remark that a preliminary version of the above procedure, still not completely motivated from a theoretical point of view, has been applied to the solution of a single equation in a specific applicative context in [31].

The structure of the chapter is as follows. In Section 4.2, we reformulate (4.4) as a minimal-norm solution problem in suitable Hilbert spaces. Then, in Section 4.3, we develop a solution method which leads to a linear ill-conditioned system, whose regularized solution is characterized in Section 4.4. The reader is referred to Section 6.3 to analyze the performance of our method applied to some numerical examples, including the application of the proposed numerical approach to a geophysical model.

4.2 Statement of the problem

Let us consider problem (4.4) and, from now on, let us assume that $f_0 = f_1 = 0$. This assumption does not affect the generality. Indeed, if it is not fulfilled, by introducing the linear function

$$\gamma(t) = \frac{b-t}{b-a} f_0 + \frac{t-a}{b-a} f_1, \tag{4.5}$$

we can rewrite problem (4.3) into an equivalent one with vanishing boundary conditions

$$\begin{cases} \displaystyle\int_a^b k_\ell(x,t)\,\xi(t)\,dt = \varphi_\ell(x), \qquad \ell = 1,\dots,m, \\ \xi(a) = 0,\ \ \xi(b) = 0, \end{cases} \tag{4.6}$$

where

$$\xi(t) = f(t) - \gamma(t), \qquad \varphi_\ell(x) = g_\ell(x) - \int_a^b k_\ell(x,t)\,\gamma(t)\,dt, \tag{4.7}$$

are the new unknown function and right-hand side, respectively.

Let us now introduce the integral operators

$$(K_\ell f)(x) := \int_a^b k_\ell(x,t)\,f(t)\,dt, \qquad \ell = 1,\dots,m, \tag{4.8}$$

so that problem (4.4) can be written as

$$\begin{cases} (K_\ell f)(x_{\ell,i}) = g_\ell(x_{\ell,i}), \qquad \ell = 1,\dots,m, \quad i = 1,\dots,n_\ell, \\ f(a) = 0,\ \ f(b) = 0, \end{cases} \tag{4.9}$$

or, equivalently,

$$\begin{cases} \mathbf{K}f = \mathbf{g}, \\ f(a) = 0, \;\; f(b) = 0, \end{cases} \tag{4.10}$$

where

$$\mathbf{K}f = \begin{bmatrix} \mathbf{K}_1 f \\ \vdots \\ \mathbf{K}_m f \end{bmatrix}, \qquad \mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_m \end{bmatrix}, \tag{4.11}$$

and

$$\mathbf{K}_\ell f = [(K_\ell f)(x_{\ell,1}), \ldots, (K_\ell f)(x_{\ell,n_\ell})]^T,$$
$$\mathbf{g}_\ell = [g_\ell(x_{\ell,1}), \ldots, g_\ell(x_{\ell,n_\ell})]^T,$$

are vectors in $\mathbb{R}^{n_\ell}$ for $\ell = 1, \ldots, m$.

As already remarked, the above problem is ill-posed. If the right-hand side does not belong to the range of the operator the solution does not exist; this happens, in particular, when the data are affected by errors. Moreover, the solution is not unique. Indeed, if $f(t)$ is a solution of (4.10) and $h(t)$ is orthogonal to the functions $k_\ell(x_{\ell,i}, t)$, for $\ell = 1, \ldots, m$ and $i = 1, \ldots, n_\ell$, then $f(t) + h(t)$ is a solution as well.

Because of this, we reformulate (4.10) in terms of the following least-squares problem

$$\min_f \|\mathbf{K}f - \mathbf{g}\|^2, \tag{4.12}$$

where $\|\cdot\|$ is the Euclidean norm. Problem (4.12) has infinitely many solutions and among them we look for a function $f(t)$ which verifies

$$\min \int_a^b (f''(t))^2 \, dt, \tag{4.13}$$

that is, we look for the solution having second derivative with minimal norm over the interval $[a, b]$. The result of this choice is to minimize the curvature of $f(t)$ and promote the selection of a smooth solution. In the space of square-integrable functions, this solution may not be unique. It is necessary to introduce a suitable function space in which (4.13) represents a strictly convex norm. In this way, the uniqueness of the solution is ensured.

Remark 4.2.1. Let us observe that in case f does not satisfy homogeneous boundary conditions, so that we have to reformulate the original problem as (4.6), from (4.5) and (4.7), we obtain

$$\min \int_a^b (f''(t))^2 \, dt = \min \int_a^b (\xi''(t))^2 \, dt.$$

This means that, after collocation, selecting the solution f of (4.4) satisfying (4.13) corresponds to computing the minimal-norm solution of (4.9) in a suitable Hilbert space.

Let us now introduce a function space for the solution of such a problem. Let $L^2([a,b])$ be the Hilbert space of square-integrable functions $f : [a,b] \to \mathbb{R}$, equipped with the inner product

$$\langle f, g \rangle_{L^2} = \int_a^b f(x)g(x)\,dx,$$

and the induced norm

$$\|f\|_{L^2} = \sqrt{\langle f, f \rangle_{L^2}}.$$

Let us also define the Hilbert space

$$W = \{f \in L^2 : f, f' \in AC([a,b]), f'' \in L^2, f(a) = f(b) = 0\},$$

where $AC([a,b])$ denotes the set of all functions f which are absolutely continuous on $[a,b]$, with inner product

$$\langle f, g \rangle_W = \langle f'', g'' \rangle_{L^2}, \tag{4.14}$$

and induced norm

$$\|f\|_W = \|f''\|_{L^2}.$$

The space W is a *reproducing kernel Hilbert space* (RKHS), i.e., there exists a bivariate function $G : [a,b] \times [a,b] \to \mathbb{R}$, called the *reproducing kernel*, satisfying the following properties:

(i) for any $y \in [a,b]$, we have $G_y(x) = G(x,y) \in W$;

(ii) for any $y \in [a,b]$, each function f belonging to W can be written as

$$f(y) = \langle G_y, f \rangle_W. \tag{4.15}$$

The expression of G, for any $x, y \in [a,b]$, is given by

$$G(x,y) = G_y(x) = \int_a^b G_x''(z) G_y''(z)\,dz,$$

where

$$G_y''(z) = \frac{\partial^2 G_y(z)}{\partial z^2} = \begin{cases} \dfrac{(z-a)(y-b)}{b-a}, & a \le z < y, \\ \dfrac{(y-a)(z-b)}{b-a}, & y \le z \le b. \end{cases}$$

It is easy to check that from (4.14) and (4.15) it follows

$$f(y) = \int_a^b G_y''(z) f''(z)\,dz.$$

For further examples and properties concerning reproducing kernels, the interested reader can consult [4, 63, 88, 89, 101].

Let us now consider problem (4.10) in W. This means that the bounded linear functional $\mathbf{K}$ is such that

$$\begin{aligned} \mathbf{K} : W &\longrightarrow \mathbb{R}^{N_m} \\ f &\longmapsto \mathbf{K}f, \end{aligned}$$

with

$$(\mathbf{K}f)_j = (K_\ell f)(x_{\ell,i}), \quad j = i + N_{\ell-1}, \qquad N_r = \sum_{k=1}^{r} n_k, \tag{4.16}$$

$\ell = 1, \dots, m$, $i = 1, \dots, n_\ell$, and $N_0 = 0$. We want to solve the following problem

$$\begin{cases} \min\limits_f \|f\|_W^2 \\ f \in \left\{ \arg\min\limits_f \|\mathbf{K}f - \mathbf{g}\|^2 \right\}. \end{cases} \tag{4.17}$$

For a brief summary of linear operators, the reader is referred to Section 1.3.

By the Riesz representation theorem (see Theorem 1.3.4), there exist N_m functions $\{\eta_j\}_{j=1}^{N_m} \in W$, named *Riesz representers*, such that the jth component of the array $\mathbf{K}f$ is given by

$$(\mathbf{K}f)_j = \langle \eta_j, f \rangle_W, \quad j = 1, \dots, N_m. \tag{4.18}$$

Moreover, let us denote by $\mathbf{K}^* : \mathbb{R}^{N_m} \to W$ the adjoint operator of $\mathbf{K}$, defined by

$$\langle \mathbf{K}f, \mathbf{g} \rangle_2 = \langle f, \mathbf{K}^*\mathbf{g} \rangle_W, \qquad \text{for any } \mathbf{g} \in \mathbb{R}^{N_m}, \tag{4.19}$$

where $\langle \cdot, \cdot \rangle_2$ is the usual Euclidean inner product in $\mathbb{R}^{N_m}$. Let us also introduce the null space of $\mathbf{K}$

$$\mathcal{N}(\mathbf{K}) = \{ f \in W : \mathbf{K}f = \mathbf{0} \},$$

and its orthogonal complement

$$\mathcal{N}(\mathbf{K})^\perp = \{ f \in W : \langle f, g \rangle_W = 0, \forall g \in \mathcal{N}(\mathbf{K}) \}.$$

The latter space is spanned by the Riesz representers, as the following lemma states.

Lemma 4.2.2. *Let $\mathbf{K}$ be a bounded linear operator from a Hilbert space to a finite-dimensional Hilbert space, then $\mathcal{N}(\mathbf{K})^\perp$ coincides with the range of the adjoint operator $\mathcal{R}(\mathbf{K}^*)$*

$$\mathcal{N}(\mathbf{K})^\perp = \mathcal{R}(\mathbf{K}^*) = \{ f \in W : f = \mathbf{K}^*\mathbf{g}, \ \textit{for } \mathbf{g} \in \mathbb{R}^{N_m} \},$$

and, in addition,

$$\mathcal{N}(\mathbf{K})^\perp = \operatorname{span}\{\eta_1, \dots, \eta_{N_m}\}.$$

Proof. We recall from Theorem 1.3.7 that in [50, Theorem 3.3.2] it is proved that $\mathcal{N}(\mathbf{K})^\perp = \overline{\mathcal{R}(\mathbf{K}^*)}$. In our case, $\mathcal{R}(\mathbf{K}^*)$ is finite-dimensional, so the closure is not needed. For any $f \in W$ and $\mathbf{g} \in \mathbb{R}^{N_m}$, we have

$$\langle \mathbf{K}f, \mathbf{g}\rangle_2 = \sum_{\ell=1}^{m} \langle \mathbf{K}_\ell f, \mathbf{g}_\ell\rangle_2 = \sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell} (K_\ell f)(x_{\ell,i})\, g_\ell(x_{\ell,i}).$$

Then, by combining (4.16) and (4.18), we can assert

$$\begin{aligned}\langle \mathbf{K}f, \mathbf{g}\rangle_2 &= \sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell} \langle \eta_{i+N_{\ell-1}}, f\rangle_W\, g_\ell(x_{\ell,i}) = \left\langle f, \sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell} g_\ell(x_{\ell,i})\, \eta_{i+N_{\ell-1}} \right\rangle_W \\ &= \langle f, \mathbf{K}^*\mathbf{g}\rangle_W\,,\end{aligned}$$

where the last equality follows by virtue of (4.19). This shows that any function in the range of $\mathbf{K}^*$ can be expressed as a linear combination of the Riesz representers η_j, $j = 1, \ldots, N_m$. □

4.3 Computing the minimal-norm solution

In this section, we develop a projection method to compute the minimal-norm solution of (4.10).

As a consequence of Lemma 4.2.2, such a solution can be expressed as a linear combination of the Riesz representers, as the following theorem shows.

Theorem 4.3.1. *The minimal-norm solution $f^\dagger$ of (4.10) is given by*

$$f^\dagger = \sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell} c_{i+N_{\ell-1}}\eta_{\ell,i}, \tag{4.20}$$

with $\eta_{\ell,i} := \eta_{i+N_{\ell-1}}$.

Proof. Since the minimal-norm solution $f^\dagger$ belongs to $\mathcal{N}(\mathbf{K})^\perp$ [101], from Lemma 4.2.2 we can write

$$f^\dagger = \sum_{j=1}^{N_m} c_j\eta_j = \sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell} c_{i+N_{\ell-1}}\eta_{\ell,i}, \qquad \text{with } \eta_{\ell,i} := \eta_{i+N_{\ell-1}}.$$

□

Since the Riesz representers are functions in the space W, we have

$$\eta_{\ell,i}(t) = \langle G_t, \eta_{i+N_{\ell-1}}\rangle_W \quad \text{and} \quad \eta_{\ell,i}(a) = \eta_{\ell,i}(b) = 0. \tag{4.21}$$

Given the definition (4.14) of the inner product, to obtain the Riesz representers $\eta_{\ell,i}(t)$ the expressions of functions $\eta''_{\ell,i}$ are needed, for $\ell = 1, \dots, m$ and $i = 1, \dots, n_\ell$. To this end, we consider (4.8) and write the unknown function by (4.15)

$$\begin{aligned}(K_\ell f)(x_{\ell,i}) &= \int_a^b k_\ell(x_{\ell,i}, t) \int_a^b G''_t(z)\, f''(z)\, dz\, dt \\ &= \int_a^b f''(z) \int_a^b G''_t(z)\, k_\ell(x_{\ell,i}, t)\, dt\, dz,\end{aligned}$$

from which, by (4.18), we deduce

$$\eta''_{\ell,i}(z) = \int_a^b G''_t(z)\, k_\ell(x_{\ell,i}, t)\, dt, \tag{4.22}$$

for $\ell = 1, \dots, m$ and $i = 1, \dots, n_\ell$.

Let us mention that, depending on the expression of the kernels k_ℓ, the above integrals may be analytically computed. Whenever this is not possible, we employ a Gaussian quadrature formula of suitable order to approximate (4.22). The two examples Test Functions 8 and 9, described in Chapter 5, illustrate both situations. In the same chapter, the expressions of the Riesz representers and of their second derivatives are reported.

Let us now compute the coefficient of the expansion (4.20) of the minimal-norm solution. By substituting expression (4.20) in place of f in (4.9), we obtain

$$(K_\ell f^\dagger)(x_{\ell,i}) = g_\ell(x_{\ell,i}), \qquad \ell = 1, \dots, m, \quad i = 1, \dots, n_\ell,$$

namely,

$$\sum_{\ell=1}^{m} \sum_{k=1}^{n_\ell} (K_\ell \eta_{\ell,k})(x_{\ell,i}) c_{k+N_{\ell-1}} = g_\ell(x_{\ell,i}),$$

where $\eta_{\ell,k} := \eta_{k+N_{\ell-1}}$ and the integers N_ℓ are defined in (4.16). By renumbering the Riesz representers, we obtain the square linear system

$$\sum_{j=1}^{N_m} (K_\ell \eta_j)(x_{\ell,i}) c_j = g_\ell(x_{\ell,i}), \qquad \ell = 1, \dots, m, \quad i = 1, \dots, n_\ell.$$

Taking into account (4.18), the above linear system can be written in matrix form as

$$\mathcal{G}\mathbf{c} = \mathbf{g}, \tag{4.23}$$

where $\mathbf{g}$ is defined in (4.11) and $\mathbf{c} = [c_j]_{j=1}^{N_m}$ is the vector of the unknowns. The Gram matrix $\mathcal{G} \in \mathbb{R}^{N_m \times N_m}$ is defined as

$$\mathcal{G} = \begin{bmatrix} \mathcal{G}^1 & \Gamma^{1,2} & \dots & \Gamma^{1,m} \\ (\Gamma^{1,2})^T & \mathcal{G}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ (\Gamma^{1,m})^T & \dots & \dots & \mathcal{G}^m \end{bmatrix}, \tag{4.24}$$

where the entries of the m diagonal blocks $\mathcal{G}^\ell$, $\ell = 1, \ldots, m$, are given by

$$\mathcal{G}^\ell_{ij} = \langle \eta_{\ell,i}, \eta_{\ell,j} \rangle_W, \tag{4.25}$$

and the off-diagonal blocks $\Gamma^{\ell,k}$, with $\ell, k = 1, \ldots, m$, $k > \ell$, have entries

$$\Gamma^{\ell,k}_{ij} = \langle \eta_{\ell,i}, \eta_{k,j} \rangle_W, \tag{4.26}$$

for $i = 1, \ldots, n_\ell$ and $j = 1, \ldots, n_k$.

The inner products in (4.25) and (4.26) involve the second derivatives $\eta''_{\ell,i}$. Whenever they can be computed analytically, the elements of the Gram matrix $\mathcal{G}$ can be obtained by symbolic computation; we used the `integral` function of Matlab. If this is not possible, a Gaussian quadrature formula is adopted.

As it is well-known, the Gram matrix $\mathcal{G}$ defined in (4.24) is symmetric positive definite. Then, a natural approach for solving system (4.23) would be to apply Cholesky factorization. However, as this linear system results from the discretization of an ill-posed problem, the matrix $\mathcal{G}$ is severely ill-conditioned. Since experimental data is typically contaminated by noise, the numerical solution of (4.23) is subject to strong error propagation and can deviate substantially from the exact solution. Moreover, because of ill-conditioning, the numerical computation of the Cholesky factorization may lead to computing the square root of small negative quantities, making it impossible to construct the Cholesky factor.

We adopted a different approach, consisting of writing the Gram matrix in terms of its spectral factorization

$$\mathcal{G} = U \Lambda U^T, \tag{4.27}$$

where the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{N_m})$ contains the eigenvalues of $\mathcal{G}$ sorted by decreasing value, and $U = [\mathbf{u}_1, \ldots, \mathbf{u}_{N_m}]$ is the eigenvector matrix with orthonormal columns; see Section 1.1 and [94].

Then, by employing factorization (4.27) in system (4.23), we obtain the following representation for the coefficients

$$\mathbf{c} = [c_1, \ldots, c_{N_m}]^T = \sum_{\ell=1}^{N_m} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\lambda_\ell} \mathbf{u}_\ell, \tag{4.28}$$

of the minimal-norm solution

$$f^\dagger = \sum_{j=1}^{N_m} c_j \eta_j, \tag{4.29}$$

resulting from Theorem 4.3.1.

4.4 Regularized minimal-norm solution

The severe ill-conditioning of the matrix $\mathcal{G}$ produces a strong error propagation in (4.28) and, consequently, in the solution (4.29). A regularized solution is needed, instead.

In what follows, it is convenient to write $f^\dagger$ as a linear combination of orthonormal functions. The orthonormalization of family of functions is a classical topic in functional analysis. The properties arising from the orthogonalization of the translates of a given function, and the connections of such process to the factorization of the associated Gram matrix have been investigated in [46, 48], and later generalized to multivariate functions in [49]. A review of the available algorithms for the spectral factorization of infinite Gram matrices is contained in [47].

The following Theorem shows how an orthonormal expansion for the minimal-norm solution can be constructed by (4.27), and gives the expression of such orthonormal functions, which are, in fact, the singular functions [35, 72] of the integral operator $\mathbf{K}$. For a summary on singular systems we refer the reader to Section 1.2.

Theorem 4.4.1. *The minimal-norm solution $f^\dagger$ of (4.10) can be written as a linear combination of orthonormal functions $\widehat{\eta}_\ell$*

$$f^\dagger = \sum_{\ell=1}^{N_m} \widehat{c}_\ell \widehat{\eta}_\ell, \tag{4.30}$$

where

$$\widehat{c}_\ell = \frac{\mathbf{u}_\ell^T \mathbf{g}}{\sqrt{\lambda_\ell}}, \qquad \widehat{\eta}_\ell = \sum_{j=1}^{N_m} \frac{u_{j\ell}}{\sqrt{\lambda_\ell}} \eta_j, \qquad \ell = 1, \dots, N_m, \tag{4.31}$$

and $u_{j\ell}$ denotes the jth component of the eigenvector $\mathbf{u}_\ell$ with eigenvalue λ_ℓ in the spectral factorization (4.27). Moreover, the set of the triplets $\left\{\sqrt{\lambda_\ell}, \widehat{\eta}_\ell, \mathbf{u}_\ell\right\}$, $\ell = 1, \dots, N_m$, is the singular system of the operator $\mathbf{K}$ (4.10).

Proof. Starting from (4.29) and (4.28), changing the order of summation, we obtain

$$f^\dagger = \sum_{j=1}^{N_m} c_j \eta_j = \sum_{j=1}^{N_m} \sum_{\ell=1}^{N_m} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\lambda_\ell} u_{j\ell} \eta_j = \sum_{\ell=1}^{N_m} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\sqrt{\lambda_\ell}} \sum_{j=1}^{N_m} \frac{u_{j\ell}}{\sqrt{\lambda_\ell}} \eta_j.$$

Equation (4.30) follows by defining $\widehat{c}_\ell$ and $\widehat{\eta}_\ell$ as in (4.31).

Let us now prove the final statement of the theorem. It is immediate to verify that the functions $\widehat{\eta}_\ell$, $\ell = 1, \dots, N_m$, form an orthonormal basis for $\mathcal{N}(\mathbf{K})^\perp$. Indeed, letting $\mathcal{G}_{ij} = \langle \eta_i, \eta_j \rangle_W$ be the elements of $\mathcal{G}$, we have

$$\begin{aligned} \langle \widehat{\eta}_k, \widehat{\eta}_h \rangle_W &= \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \frac{u_{ik}}{\sqrt{\lambda_k}} \frac{u_{jh}}{\sqrt{\lambda_h}} \langle \eta_i, \eta_j \rangle_W = \frac{1}{\sqrt{\lambda_k \lambda_h}} \sum_{i=1}^{N_m} u_{ik} \sum_{j=1}^{N_m} \mathcal{G}_{ij} u_{jh} \\ &= \frac{1}{\sqrt{\lambda_k \lambda_h}} (U^T \mathcal{G} U)_{kh} = \frac{1}{\sqrt{\lambda_k \lambda_h}} \Lambda_{kh} = \delta_{kh}, \end{aligned}$$

where δ_{kh} is the Kronecker delta and, in the last equality, the matrix $\mathcal{G}$ is replaced by its spectral decomposition (4.27). The orthonormality of the vectors $\mathbf{u}_\ell$, $\ell = 1, \dots, N_m$, immediately follows from factorization (4.27).

From the definition (4.18) of the Riesz representers, we can write

$$(\mathbf{K}\widehat{\eta}_\ell)_j = \langle \eta_j, \widehat{\eta}_\ell \rangle_W = \sum_{s=1}^{N_m} \frac{u_{s\ell}}{\sqrt{\lambda_\ell}} \langle \eta_j, \eta_s \rangle_W = \sum_{s=1}^{N_m} \frac{u_{s\ell}}{\sqrt{\lambda_\ell}} \mathcal{G}_{js} = \frac{1}{\sqrt{\lambda_\ell}} (\mathcal{G}U)_{j\ell}$$
$$= \frac{1}{\sqrt{\lambda_\ell}} (U\Lambda)_{j\ell} = \sqrt{\lambda_\ell}\, u_{j\ell}, \qquad j = 1, \ldots, N_m,$$

where the spectral factorization (4.27) of $\mathcal{G}$ is employed again. Then, $\mathbf{K}\widehat{\eta}_\ell = \sqrt{\lambda_\ell}\mathbf{u}_\ell$.

Now, let $f \in W$. Then, $f = f_0 + f_1$, with $f_0 \in \mathcal{N}(\mathbf{K})$, $f_1 \in \mathcal{N}(\mathbf{K})^\perp$, and

$$f_1 = \sum_{j=1}^{N_m} \alpha_j \widehat{\eta}_j, \qquad \text{with } \alpha_j = \langle f_1, \widehat{\eta}_j \rangle_W.$$

By the definition (4.19) of the adjoint operator, we obtain

$$\langle f, \mathbf{K}^*\mathbf{u}_\ell \rangle_W = \langle \mathbf{K}f, \mathbf{u}_\ell \rangle_2 = \langle \mathbf{K}f_1, \mathbf{u}_\ell \rangle_2 = \sum_{j=1}^{N_m} \alpha_j \langle \mathbf{K}\widehat{\eta}_j, \mathbf{u}_\ell \rangle_2$$
$$= \sum_{j=1}^{N_m} \alpha_j \langle \sqrt{\lambda_j}\mathbf{u}_j, \mathbf{u}_\ell \rangle_2 = \alpha_\ell \sqrt{\lambda_\ell} = \langle f, \sqrt{\lambda_\ell}\widehat{\eta}_\ell \rangle_W,$$

since $\alpha_\ell = \langle f_1, \widehat{\eta}_\ell \rangle_W = \langle f, \widehat{\eta}_\ell \rangle_W$. Then $\mathbf{K}^*\mathbf{u}_\ell = \sqrt{\lambda_\ell}\widehat{\eta}_\ell$. It follows that

$$\mathbf{K}^*\mathbf{K}\widehat{\eta}_\ell = \lambda_\ell \widehat{\eta}_\ell, \qquad \mathbf{K}\mathbf{K}^*\mathbf{u}_\ell = \lambda_\ell \mathbf{u}_\ell, \qquad \ell = 1, \ldots, N_m.$$

This completes the proof. □

We remark that Theorem 4.4.1 is applicable under the assumption that the Gram matrix $\mathcal{G}$ is positive definite. In practice, because of error propagation, the smallest numerical eigenvalues of $\mathcal{G}$ may become zero, or even negative. In this case, that is, if $\lambda_{N_m} \leq 0$, we substitute to N_m, in all summations, an integer $N < N_m$ such that $\lambda_N > 0 \geq \lambda_{N+1}$.

From (4.30) and from the definition of $\widehat{\mathbf{c}}$ in (4.31), it follows that

$$\|f^\dagger\|_W = \|\widehat{\mathbf{c}}\| = \|L\mathbf{c}\|, \qquad \text{with } L = \Lambda^{1/2}U^T, \tag{4.32}$$

where the relation between $\mathbf{c}$ and $\widehat{\mathbf{c}}$ is obtained by (4.31), writing $\widehat{\mathbf{c}}$ in matrix form

$$\widehat{\mathbf{c}} = \Lambda^{-1/2}U^T\mathbf{g} = \Lambda^{-1/2}U^T\mathcal{G}\mathbf{c} = \Lambda^{-1/2}U^TU\Lambda U^T\mathbf{c} = \Lambda^{1/2}U^T\mathbf{c}.$$

As it is customary, in order to face ill-conditioning, we replace the original problem by a nearby one, whose solution is less sensitive to the error present in the data. The representation (4.30) is particularly suitable to construct a regularized solution. Indeed, according to the Picard condition [93], the sum of the coefficients

$\widehat{c}_\ell$ should theoretically be bounded. Anyway, the presence of noise in the right-hand side $\mathbf{g}$ will prevent the projections $\mathbf{u}_\ell^T\mathbf{g}$ from decaying when ℓ increases, leading to a severe growth in the values of the coefficients. Truncating the summation in (4.30) removes the noisy components of the solution that are enhanced by ill-conditioning. Moreover, it damps the high frequency components represented by the $\widehat{\eta}_\ell$ functions with a large index ℓ.

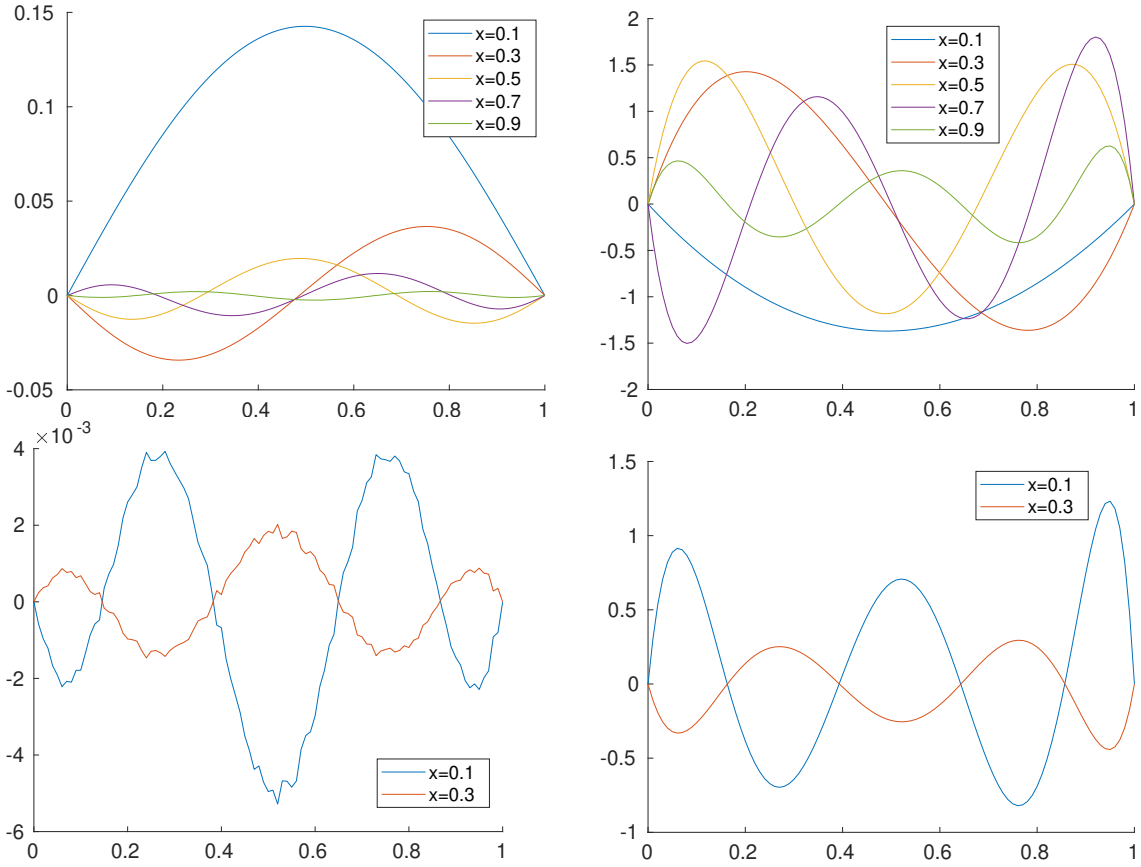

Figure 4.1: Orthonormalized Riesz functions for the system (5.10): $\widehat{\eta}_{1,i}$ (top-left) and $\widehat{\eta}'_{1,i}$ (top-right) for $x_i = 0.1 + 0.2(i-1)$, $i = 1, \ldots, 5$; $\widehat{\eta}_{2,i}$ (bottom-left) and $\widehat{\eta}'_{2,i}$ (bottom-right) are displayed only for x_1 and x_2.

It is well-known that singular functions associated with first kind integral equations with a smooth kernel oscillate at an increasing frequency as the singular values decrease. For example, Figure 4.1 displays the functions $\widehat{\eta}_\ell$ obtained by applying formula (4.31) to the Riesz functions constructed in Test Function 8 of Chapter 5. In the summation (4.31), the upper bound for the index is fixed at $N = 7$, to preserve the positivity of the eigenvalues. The graphs in the left column depict the orthonormal functions, and the ones in the right column their second derivatives. It is immediate to observe the increasing frequency of the orthonormal basis. It is also clear that there is a strong error propagation in the numerical construction of

such functions; see, in particular, the graphs in the bottom-left panel. This deters from employing the orthonormal basis in the real computation, unless a more stable orthonormalization process is implemented. Anyway, as we will show, the functions $\widehat{\eta}_\ell$ are only implicitly used in the construction of the regularized solution.

Indeed, the regularized solution is obtained by choosing an index κ to truncate the summation in (4.30), i.e., $1 \leq \kappa \leq N$, leading to the expression

$$f^{(\kappa)} = \sum_{\ell=1}^{\kappa} \widehat{c}_\ell \widehat{\eta}_\ell = \sum_{\ell=1}^{\kappa} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\sqrt{\lambda_\ell}} \sum_{j=1}^{N} \frac{u_{j\ell}}{\sqrt{\lambda_\ell}} \eta_j = \sum_{j=1}^{N} \sum_{\ell=1}^{\kappa} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\lambda_\ell} u_{j\ell} \eta_j = \sum_{j=1}^{N} c_j^{(\kappa)} \eta_j. \tag{4.33}$$

This shows that $f^{(\kappa)}$ can be expressed as a linear combination of the Riesz representers η_j and there is no need to explicitly construct the singular functions $\widehat{\eta}_\ell$.

The coefficients in the last summation correspond to the *truncated eigendecomposition* (TEIG) solution of system (4.23) (see [1, 41] for more details) with parameter $\kappa \leq N$, defined to be the components of the vector

$$\mathbf{c}^{(\kappa)} = U \Lambda_\kappa^\dagger U^T \mathbf{g} = \sum_{\ell=1}^{\kappa} \frac{\mathbf{u}_\ell^T \mathbf{g}}{\lambda_\ell} \mathbf{u}_\ell, \tag{4.34}$$

where $\Lambda_\kappa^\dagger$ denotes the Moore-Penrose pseudoinverse [10] of $\Lambda_\kappa = \operatorname{diag}(\lambda_1, \ldots, \lambda_\kappa, 0, \ldots, 0)$. We observe herein that, because of the orthonormality of the functions $\widehat{\eta}_\ell$, $\|f^{(\kappa)}\|_W \leq \|f^{(\kappa+1)}\|_W \leq \|f^\dagger\|_W$.

It is possible to show that the above vector $\mathbf{c}^{(\kappa)}$ solves the optimization problem

$$\begin{cases} \min\limits_{\mathbf{c}} \|L\mathbf{c}\| \\ \mathbf{c} \in \left\{ \arg\min\limits_{\mathbf{c}} \|\mathcal{G}_\kappa \mathbf{c} - \mathbf{g}\| \right\}, \end{cases}$$

where $\mathcal{G}_\kappa = U \Lambda_\kappa U^T$ is the TEIG of $\mathcal{G}$. Therefore, from the algebraic point of view, the computation of $f^{(\kappa)}$ corresponds to selecting the minimal-L-norm vector among the solutions of the best rank-κ approximation of system (4.23). Equation (4.32) shows that the regularized solution $f^{(\kappa)}$ has minimal-norm in W.

Remark 4.4.2. If the problem does not verify homogeneous boundary conditions and it has to be transformed in an equivalent one, then the solution of the original problem is given by

$$f^{(\kappa)} = \sum_{j=1}^{N} c_j^{(\kappa)} \eta_j + \gamma,$$

where the function γ is defined in (4.5).

A crucial point in the regularization process, in order to get an accurate solution, is the estimation of the truncation parameter κ in (4.33) and (4.34). There exist many methods, either a posteriori and heuristic, aiming at this. In this chapter,

we focus our attention on the discrepancy principle and the L-curve method. For a summary of the criteria for choosing the regularization parameter, the reader is referred to Subsection 1.2.3 and [35, 57, 87].

We assume that the exact right-hand side vector $\mathbf{g}_{\text{exact}}$ is contaminated by an unknown normally distributed noise vector $\mathbf{e}$, i.e.,

$$\mathbf{g} = \mathbf{g}_{\text{exact}} + \mathbf{e}. \tag{4.35}$$

If $\|\mathbf{e}\|$ is known, a widely used method to estimate κ is the classical discrepancy principle, introduced by Morozov [77], which selects the smallest truncation parameter κ_{d} such that

$$\|\mathcal{G}\mathbf{c}^{(\kappa_{\text{d}})} - \mathbf{g}\| \leq \tau\|\mathbf{e}\|, \tag{4.36}$$

where $\tau > 1$ is a constant independent of the noise level $\|\mathbf{e}\|$, and $\|\cdot\|$ denotes the Euclidean norm. Note that from (4.27) and (4.34), we can write the residual norm as

$$\|\mathcal{G}\mathbf{c}^{(\kappa)} - \mathbf{g}\|^2 = \|U(\Lambda\Lambda_\kappa^\dagger - I)U^T\mathbf{g}\|^2 = \sum_{j=\kappa+1}^{N_m} (\mathbf{u}_j^T\mathbf{g})^2. \tag{4.37}$$

Besides reducing the computational load, this relation shows that the residual is non-decreasing when κ decreases.

When the noise level is unknown, heuristic methods are commonly used. We use the L-curve criterion [56, 60] which selects the regularization parameter κ_{lc} at the "corner" of the curve obtained by joining the points

$$\left(\log\|\mathcal{G}\mathbf{c}^{(\kappa)} - \mathbf{g}\|,\ \log\|f^{(\kappa)}\|_W\right), \qquad \kappa = 1, \ldots, N, \tag{4.38}$$

where $f^{(\kappa)}$ is the function defined in (4.33) and its W-norm can be expressed in the form

$$\|f^{(\kappa)}\|_W = \|L\mathbf{c}^{(\kappa)}\| = \sqrt{(\mathbf{c}^{(\kappa)})^T\mathcal{G}\mathbf{c}^{(\kappa)}}.$$

When solving discrete ill-posed problems, this curve often exhibits a typical L-shape. We determine its corner by the method described in [59] and implemented in [58].

When the exact solution f is available, to ascertain the best possible performance of the algorithms independently of the strategy adopted for the estimation of the regularization parameter, in the numerical experiments we also consider the parameter κ_{best} which minimizes the norm of the error, that is

$$\kappa_{\text{best}} = \arg\min_\kappa \|f - f^{(\kappa)}\|_W = \arg\min_\kappa \|L(\mathbf{c} - \mathbf{c}^{(\kappa)})\|. \tag{4.39}$$

Remark 4.4.3. We observe that the operator $\mathcal{F}_\kappa$ which assigns to a noisy right-hand side $\mathbf{g}$ (see (4.35) and (6.6)) the regularized solution $f^{(\kappa_{\text{d}})}$ (4.33), corresponding to the regularization parameter $\kappa_{\text{d}} = \kappa_{\text{d}}(\delta, \mathbf{g})$ estimated by the discrepancy principle, is trivially a regularization method in the sense of [35, Definition 3.1]. Indeed, from (4.36) and (4.37), $\kappa_{\text{d}} = N_m$ when $\delta \to 0$, and $f^{(N_m)}$ coincides with the minimal-norm solution $f^\dagger$.

CHAPTER 5

Test problems

In this chapter, we collect the test problems which were adopted, in [31, 32, 82, 83], as well as in the following chapter, to illustrate the performance of the algorithms proposed in the same papers and in this thesis. Test Functions 1, 2, 3, 4, 5, 8, and 9 are nonlinear functions introduced by the authors of the above papers. Test Function 6 is a nonlinear function already described in [20]. Test Function 7 is a nonlinear model describing an engineering application. Finally, Test Functions 10 and 11 are a linear and a nonlinear model, respectively, involved in soil surveying.

For each test function, the "ingredients" needed to apply the algorithms are reported. For instance, we compute the Jacobian matrix when the function is tested in "MNGN" algorithms and we report the computation of the Riesz representers when we want to solve linear integral equations in a RKHS.

5.1 Nonlinear least-squares

We start with nonlinear functions of very small size.

Test Function 1. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be the nonlinear function defined by

$$F(\mathbf{x}) = \left[\alpha(x_1 - 1)^2 + \beta(x_2 - 1)^2 - 1\right]^2, \qquad b = -1, \tag{5.1}$$

depending upon the parameters $\alpha, \beta \in \mathbb{R}$, $\mathbf{x} = [x_1, x_2]^T$, and $r(\mathbf{x}) = F(\mathbf{x}) - b$ is the residual function. The function is constructed in such a way as to assign different values to α and β and thus obtain different functions. The objective function we want to minimize is

$$f(\mathbf{x}) = \frac{1}{2} r(\mathbf{x})^2 = \frac{1}{2} \left\{ \left[\alpha(x_1 - 1)^2 + \beta(x_2 - 1)^2 - 1\right]^2 + 1 \right\}^2.$$

The Jacobian matrix of $F(\mathbf{x})$ is

$$J(\mathbf{x}) = 4 \left(\alpha(x_1 - 1)^2 + \beta(x_2 - 1)^2 - 1\right) \begin{bmatrix} \alpha(x_1 - 1) & \beta(x_2 - 1) \end{bmatrix},$$

while the Hessian matrix is

$$H(\mathbf{x}) = \begin{bmatrix} 12\alpha^2(x_1-1)^2 + 4\alpha\beta(x_2-1)^2 - 4\alpha & 8\alpha\beta(x_1-1)(x_2-1) \\ 8\alpha\beta(x_1-1)(x_2-1) & 4\alpha\beta(x_1-1)^2 + 12\beta^2(x_2-1)^2 - 4\beta \end{bmatrix}.$$

If $\alpha, \beta > 0$ or $\alpha > 0, \beta < 0$, then any point on the conic section

$$\alpha(x_1-1)^2 + \beta(x_2-1)^2 = 1$$

is a minimum of $f(\mathbf{x})$. Whereas if $\alpha, \beta < 0$, then the minimum is unique and given by $\mathbf{x}^\dagger = [1,1]^T$.

Test Function 2. A second function concern an underdetermined least-squares problem with solution in $\mathbb{R}^3$ and values in $\mathbb{R}^2$. The nonlinear function has the expression

$$F(\mathbf{x}) = \begin{bmatrix} (x_1-1)^2 + x_2^2 + x_3^2 \\ x_3 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{5.2}$$

with $\mathbf{x} = [x_1, x_2, x_3]^T$ and $\mathbf{r}(\mathbf{x}) = F(\mathbf{x}) - \mathbf{b}$ is the residual function. The Jacobian matrix of $F(\mathbf{x})$ is

$$J(\mathbf{x}) = \begin{bmatrix} 2(x_1-1) & 2x_2 & 2x_3 \\ 0 & 0 & 1 \end{bmatrix}.$$

The minimal-norm solution is $\mathbf{x}^\dagger = [0,0,0]^T$.

Test Function 3. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function

$$F(\mathbf{x}) = [F_1(\mathbf{x}), F_2(\mathbf{x}), \ldots, F_m(\mathbf{x})]^T, \qquad m \leq n, \tag{5.3}$$

defined by

$$F_i(\mathbf{x}) = \frac{1}{2} S(\mathbf{x}) \left(x_i^2 + 1\right), \qquad i = 1, \ldots, m, \tag{5.4}$$

where

$$S(\mathbf{x}) = \sum_{j=1}^{n} \left(\frac{x_j - c_j}{a_j} \right)^2 - 1$$

is the n-ellipsoid with center $\mathbf{c} = [c_1, \ldots, c_n]^T$ and whose semiaxes are the components of the vector $\mathbf{a} = [a_1, \ldots, a_n]^T$. The locus of the solutions is the n-ellipsoid.

Setting $y_i = x_i^2 + 1$, for $i = 1, \ldots, m$, and $z_j = \frac{x_j - c_j}{a_j^2}$, for $j = 1, \ldots, n$, the Jacobian matrix can be expressed as

$$J(\mathbf{x}) = S(\mathbf{x}) D_{m,n}(\mathbf{x}) + \mathbf{y}\mathbf{z}^T,$$

where $D_{m,n}(\mathbf{x})$ is an $m \times n$ diagonal matrix whose main diagonal consists of the vector $\mathbf{x}$. Indeed,

$$\frac{\partial F_i}{\partial x_k} = \begin{cases} x_i S(\mathbf{x}) + \dfrac{x_i - c_i}{a_i^2}\left(x_i^2 + 1\right), & k = i, \\ \dfrac{x_k - c_k}{a_k^2}\left(x_i^2 + 1\right), & k \neq i. \end{cases}$$

When $S(\mathbf{x}) = 0$, $\operatorname{rank}(J(\mathbf{x})) = 1$, so we expect the Jacobian to be rank-deficient in a neighborhood of the solution.

If $\mathbf{a} = \mathbf{e} = [1, \ldots, 1]^T$, the locus of the solutions is the n-sphere centered in $\mathbf{c}$ with unitary radius. If $\mathbf{c} = 2\mathbf{e}$, the minimal-norm solution is

$$\mathbf{x}^\dagger = \left(2 - \frac{\sqrt{n}}{n}\right)\mathbf{e},$$

while if $\mathbf{c} = [2, 0, \ldots, 0]^T$ it is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$.

Test Function 4. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a nonlinear function such as (5.3), with

$$F_i(\mathbf{x}) = S(\mathbf{x})\left(x_i - c_i\right), \qquad i = 1, \ldots, m, \tag{5.5}$$

and $S(\mathbf{x})$ defined as in the previous example. The first-order derivatives of $F_i(\mathbf{x})$ are

$$\frac{\partial F_i}{\partial x_k} = \begin{cases} \dfrac{2}{a_i^2}(x_i - c_i)^2 + S(\mathbf{x}), & k = i, \\ \dfrac{2}{a_k^2}(x_k - c_k)(x_i - c_i), & k \neq i. \end{cases}$$

Setting $y_i = x_i - c_i$, for $i = 1, \ldots, m$, and $z_j = \frac{x_j - c_j}{a_j^2}$, for $j = 1, \ldots, n$, the Jacobian matrix can be represented as

$$J(\mathbf{x}) = S(\mathbf{x}) I_{m \times n} + 2\mathbf{y}\mathbf{z}^T,$$

where $I_{m \times n}$ includes the first m rows of an identity matrix of size n. The Jacobian turns out to be a diagonal plus rank-1 matrix. This structure may be useful to reduce complexity when solving large-scale problems.

When $S(\mathbf{x}) = 0$, the matrix $J(\mathbf{x})$ has rank 1. Indeed, in this case, the compact SVD of the Jacobian is

$$J(\mathbf{x}) = \frac{\mathbf{y}}{\|\mathbf{y}\|}(2\|\mathbf{y}\|\|\mathbf{z}\|)\frac{\mathbf{z}^T}{\|\mathbf{z}\|},$$

so that the only non-zero singular value is $2\|\mathbf{y}\|\|\mathbf{z}\|$. Therefore, the pseudoinverse is

$$J(\mathbf{x})^\dagger = \frac{\mathbf{z}}{\|\mathbf{z}\|}\left(\frac{1}{2\|\mathbf{y}\|\|\mathbf{z}\|}\right)\frac{\mathbf{y}^T}{\|\mathbf{y}\|}.$$

As in the preceding example, we may assume that the Jacobian is rank-deficient in the surroundings of a solution.

The locus of the solutions is the union of the n-ellipsoid and the intersection between the planes $x_i = c_i$, $i = 1, \ldots, m$.

If $\mathbf{a} = \mathbf{e}$ and $\mathbf{c} = 2\mathbf{e}$, the minimal-norm solution $\mathbf{x}^\dagger$ depends on the dimensions m and n: if $m < n - \sqrt{n} + \frac{1}{4}$, then it is

$$\mathbf{x}^\dagger = [\underbrace{2, 2, \ldots, 2}_{m}, \underbrace{0, \ldots, 0}_{n-m}]^T, \qquad \text{with } \|\mathbf{x}^\dagger\| = 4m,$$

otherwise, it is

$$\mathbf{x}^\dagger = \left(2 - \frac{\sqrt{n}}{n}\right)\mathbf{e}, \qquad \text{with } \|\mathbf{x}^\dagger\| = 4n - 4\sqrt{n} + 1.$$

From the comparison of the two norms reported in the above equations, the inequality on the dimensions is deduced. If $\mathbf{c} = [2, 0, \ldots, 0]^T$, it is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$.

In the case of a square problem, i.e., $m = n$, the locus of the solutions is the union of the n-ellipsoid and the point $\mathbf{x} = \mathbf{c}$. The spectrum of $J(\mathbf{x})$ is

$$\sigma(J(\mathbf{x})) = \left\{S(\mathbf{x}) + 2\mathbf{y}^T\mathbf{z}, S(\mathbf{x}), \ldots, S(\mathbf{x})\right\},$$

where the eigenvalue $S(\mathbf{x})$ has algebraic multiplicity $n - 1$. The Jacobian matrix is invertible if and only if $S(\mathbf{x}) \neq 0$. If this condition is met, the inverse is obtained by the Sherman–Morrison formula

$$J(\mathbf{x})^{-1} = \frac{1}{S(\mathbf{x})} I_n - \frac{2}{S(\mathbf{x})(S(\mathbf{x}) + 2\mathbf{z}^T\mathbf{y})}\mathbf{y}\mathbf{z}^T.$$

Test Function 5. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function (5.3) with components

$$F_i(\mathbf{x}) = \begin{cases} S(\mathbf{x}), & i = 1, \\ x_{i-1}(x_i - c_i), & i = 2, \ldots, m, \end{cases} \tag{5.6}$$

and $S(\mathbf{x})$ defined as above. The first-order partial derivatives of $F_i(\mathbf{x})$ are

$$\frac{\partial F_i}{\partial x_k} = \begin{cases} \dfrac{2}{a_k^2}(x_k - c_k), & i = 1,\ k = 1, \ldots, n, \\ x_i - c_i, & i = 2, \ldots, m,\ k = i - 1, \\ x_{i-1}, & i = k = 2, \ldots, m, \\ 0, & \text{otherwise.} \end{cases}$$

Setting $z_j = 2\frac{x_j - c_j}{a_j^2}$ and $y_j = x_j - c_j$, for $j = 1, \ldots, n$, the Jacobian matrix of F is

$$J(\mathbf{x}) = \begin{bmatrix} z_1 & z_2 & z_3 & \cdots & z_{m-1} & z_m & \cdots & z_n \\ y_2 & x_1 & & & & & & \\ & y_3 & x_2 & & & & & \\ & & \ddots & \ddots & & & & \\ & & & \ddots & \ddots & & & \\ & & & & y_m & x_{m-1} & & \end{bmatrix}. \tag{5.7}$$

The locus of the solutions is the intersection between the hypersurfaces defined by $S(\mathbf{x}) = 0$ and by the pairs of planes $x_{i-1} = 0$, $x_i - c_i = 0$, $i = 2, \ldots, m$.

If $\mathbf{a} = \mathbf{e} = [1, \ldots, 1]^T$ and $\mathbf{c} = 2\mathbf{e}$, the minimal-norm solution is

$$\mathbf{x}^\dagger = [\xi_{n,m}, \underbrace{2, \ldots, 2}_{m-1}, \underbrace{\xi_{n,m}, \ldots, \xi_{n,m}}_{n-m}]^T,$$

with $\xi_{n,m} = 2 - (n - m + 1)^{-1/2}$, while if $\mathbf{c} = [2, 0, \ldots, 0]^T$ it is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$. It is immediate to observe that in the last situation the Jacobian (5.7) is rank-deficient at $\mathbf{x}^\dagger$.

In particular, if $m = n$, $\mathbf{a} = \mathbf{e} = [1, \ldots, 1]^T$ and $\mathbf{c} = 2\mathbf{e}$, the locus of the solutions consists of two points of the n-sphere. In this case, the minimal-norm solution is $\mathbf{x}^\dagger = [1, 2, \ldots, 2]^T$.

Test Function 6. Here we consider a test problem introduced in [20]. Let $F : \mathbb{R}^3 \to \mathbb{R}$ be the nonlinear function defined by

$$F(\mathbf{x}) = x_3 - (x_1 - 1)^2 - 2(x_2 - 2)^2 - 3. \tag{5.8}$$

The equation $F(\mathbf{x}) = 0$ represents an elliptic paraboloid in $\mathbb{R}^3$ with vertex $\mathbf{V} = [1, 2, 3]^T$. The Jacobian matrix of $F(\mathbf{x})$ is

$$J(\mathbf{x}) = \begin{bmatrix} -2(x_1 - 1) & -4(x_2 - 2) & 1 \end{bmatrix}.$$

We remark that the minimal-norm solution is the point

$$\mathbf{x}^\dagger \approx [0.859754, 1.849178, 3.065164]^T,$$

and not the vector $\widehat{\mathbf{x}}$ reported in [20, Sec. 4.2]. Indeed, $\|\mathbf{x}^\dagger\| \approx 3.681558$, whereas $\|\widehat{\mathbf{x}}\| \approx 3.706359$.

Test Function 7. Here we consider a nonlinear model that describes the behavior of a redundant parallel robot. It is a problem that concerns the inverse kinematics of position [2], and is defined by the following function $F : \mathbb{R}^4 \to \mathbb{R}^2$

$$F(\mathbf{x}) = \begin{bmatrix} (X - A\cos(x_1))^2 + (Y - A\sin(x_1))^2 - x_2^2 \\ (X - A\cos(x_3) - H)^2 + (Y - A\sin(x_3))^2 - x_4^2 \end{bmatrix}. \tag{5.9}$$

The model describes the kinematic of a robotic arm moved by four motors, whose position is identified by the unknowns $\{x_i\}_{i=1}^4$, which must reach a point with given coordinates (X, Y); A and H are parameters describing the system.

The Jacobian matrix of F is

$$J(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1} & \dfrac{\partial F_1}{\partial x_2} & 0 & 0 \\ 0 & 0 & \dfrac{\partial F_2}{\partial x_3} & \dfrac{\partial F_2}{\partial x_4} \end{bmatrix},$$

with

$$\begin{aligned} \frac{\partial F_1}{\partial x_1} &= 2A(X - A\cos(x_1))\sin(x_1) - 2A(Y - A\sin(x_1))\cos(x_1), \\ \frac{\partial F_2}{\partial x_3} &= 2A(X - A\cos(x_3) - H)\sin(x_3) - 2A(Y - A\sin(x_3))\cos(x_3), \\ \frac{\partial F_1}{\partial x_2} &= -2x_2, \quad \frac{\partial F_2}{\partial x_4} = -2x_4. \end{aligned}$$

5.2 Systems of integral equations

Now, we introduce some examples of systems of integral equations. In Test Functions 8 and 9, by using the notation of Chapter 4, we assume $m = 2$, $n_1 = n_2 = n$, so that $N_m = 2n$, and $x_{1,i} = x_{2,i} = x_i$, for $i = 1, \ldots, n$.

Test Function 8. Let us consider the system of integral equations

$$\begin{cases} \displaystyle\int_0^1 \frac{x}{t+1} f(t)\, dt = x\left(\log 4 - \frac{1}{2}\right), \\ \displaystyle\int_0^1 \cos(xt) f(t)\, dt = \frac{2}{x^3}\left(x\cos x + (x^2 - 1)\sin x\right), \end{cases} \tag{5.10}$$

with $x \in (0, 1]$, whose exact solution is $f(t) = t^2 + 1$. We note that the function $1/x^3$ has a discontinuity of the second kind at $x = 0$. Anyway, the right-hand side of the second equation is a square-integrable function. We introduce the function (4.5)

$$\gamma(t) = t + 1,$$

to reformulate the original problem as the following one

$$\begin{cases} \displaystyle\int_0^1 \frac{x}{t+1} \xi(t)\, dt = x\left(\log 4 - \frac{3}{2}\right), \\ \displaystyle\int_0^1 \cos(xt) \xi(t)\, dt = \frac{1}{x^2}\left(\cos x + 1 - \frac{2\sin x}{x}\right), \end{cases}$$

where $\xi(t) = f(t) - \gamma(t)$ satisfies homogeneous boundary conditions.

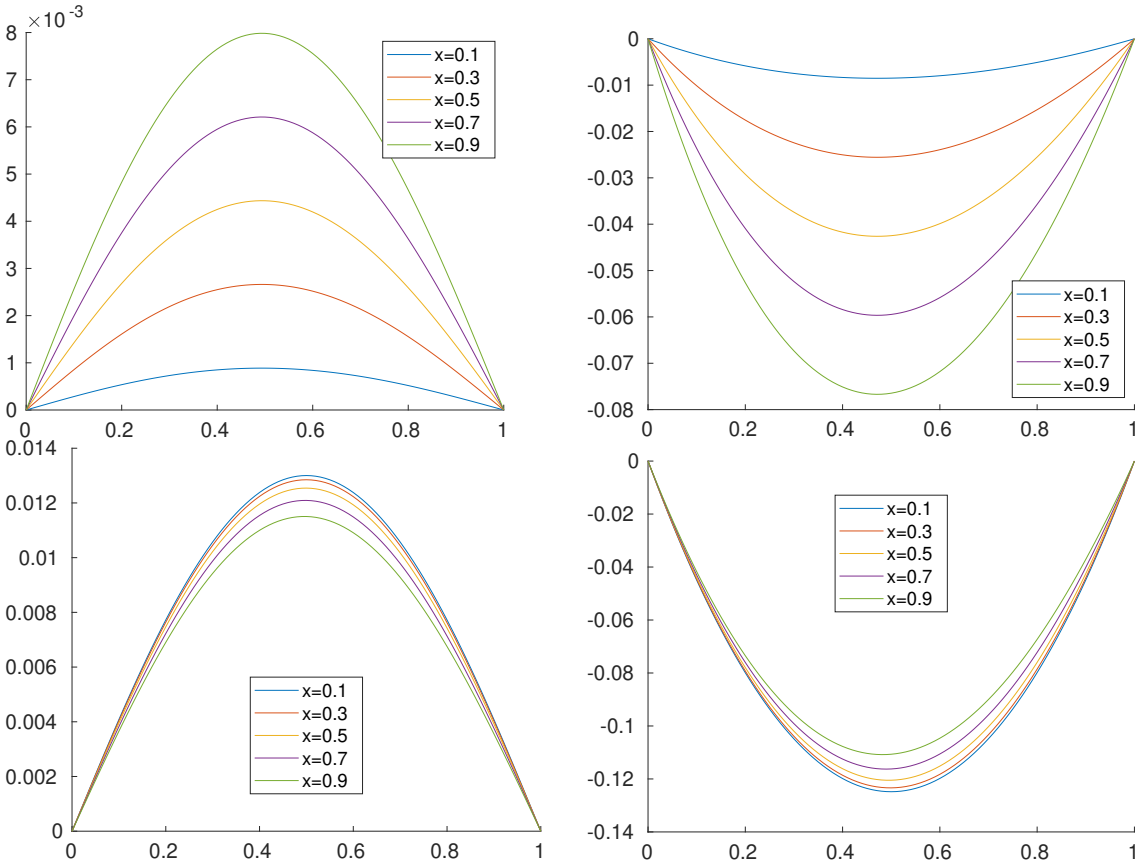

Figure 5.1: Riesz functions for the system (5.10): $\eta_{1,i}$ (top-left), $\eta''_{1,i}$ (top-right), $\eta_{2,i}$ (bottom-left), and $\eta''_{2,i}$ (bottom-right), with $x_i = 0.1 + 0.2(i-1)$ for $i = 1, \dots, 5$.

From (4.22), after some computation, we obtain, for $i = 1, \dots, n$,

$$\eta''_{1,i}(z) = x_i \left[(1-z)\log(1+z) - z \log\left(\frac{4}{(1+z)^2} \right) \right], \tag{5.11}$$

$$\eta''_{2,i}(z) = \frac{1}{x_i^2} \left(z \cos x_i - \cos(x_i z) - z + 1 \right). \tag{5.12}$$

Then, from (4.21),

$$\eta_{1,i}(y) = \frac{x_i}{36} \Big\{ 6(1+y)^3 \log(1+y) - y \left[y^2(5 + 12\log 2) + 15y + 4(9\log 2 - 5) \right] \Big\}, \tag{5.13}$$

$$\eta_{2,i}(y) = \frac{y(y-1)}{6x_i^2} \left[(y+1)\cos(x_i) - y + 2 \right] + \frac{1}{x_i^4} \left[y(1 - \cos(x_i)) - 1 + \cos(x_i y) \right]. \tag{5.14}$$

Figure 5.1 displays, in the top row, the functions $\eta_{1,i}$ (on the left) and $\eta''_{1,i}$ (on the right), while the bottom row depicts the functions $\eta_{2,i}$ (on the left) and $\eta''_{2,i}$ (on the

right) for different collocation points $x_{\ell,i}$. We see from Figure 5.1 that the Riesz functions verify the boundary conditions, i.e., $\eta_{\ell,i}(0) = \eta_{\ell,i}(1) = 0$, for $\ell = 1, 2$ and $i = 1, \ldots, 5$. From the same figure we can observe that, in this case, it also holds $\eta''_{\ell,i}(0) = \eta''_{\ell,i}(1) = 0$.

Test Function 9. Let us consider the system

$$\begin{cases} \displaystyle\int_0^\pi \mathrm{e}^{x\cos t} f(t)\,dt = 2\frac{\sinh x}{x}, \\ \displaystyle\int_0^\pi (xt + \mathrm{e}^{xt}) f(t)\,dt = \pi x + \frac{1+\mathrm{e}^{\pi x}}{1+x^2}, \end{cases} \tag{5.15}$$

with $x \in (0, \pi/2]$, whose exact solution is $f(t) = \sin t$. This system has been obtained by coupling the well-known Baart test problem [6, 58] with another equation having the same solution. The Green function, for the interval $[0, \pi]$, is

$$G''_y(z) = \begin{cases} \dfrac{z(y-\pi)}{\pi}, & 0 \le z < y, \\ \dfrac{y(z-\pi)}{\pi}, & y \le z \le \pi. \end{cases}$$

From (4.22) we have, for $i = 1, \ldots, n$,

$$\begin{aligned} \eta''_{2,i}(z) &= \int_0^z \frac{t(z-\pi)}{\pi}(x_i t + \mathrm{e}^{x_i t})\,dt + \int_z^\pi \frac{z(t-\pi)}{\pi}(x_i t + \mathrm{e}^{x_i t})\,dt \\ &= \frac{z(1-\mathrm{e}^{\pi x_i})}{\pi x_i^2} + \frac{x_i z(z^2-\pi^2)}{6} + \frac{\mathrm{e}^{x_i z} - 1}{x_i^2}, \end{aligned} \tag{5.16}$$

and from (4.21)

$$\begin{aligned} \eta_{2,i}(y) &= \int_0^y \frac{z(y-\pi)}{\pi}\eta''_{2,i}(z)\,dz + \int_y^\pi \frac{y(z-\pi)}{\pi}\eta''_{2,i}(z)\,dz \\ &= \frac{\pi^2 x_i y}{36}\left(\frac{7}{10}\pi^2 - y^2\right) + \frac{y}{6\pi x_i^4}(1-\mathrm{e}^{\pi x_i})(x_i^2 y^2 + 6) \\ &\quad + \frac{\pi y}{6x_i^2}(\mathrm{e}^{\pi x_i} + 2) + \frac{y^2}{2}\left(\frac{x_i y^3}{60} - \frac{1}{x_i^2}\right) + \frac{\mathrm{e}^{x_i y} - 1}{x_i^4}. \end{aligned} \tag{5.17}$$

The functions $\eta''_{1,i}(z)$ and $\eta_{1,i}(y)$ do not have an analytic representation, so they should be approximated by a quadrature formula. Here, we approximate the integrals by a Gauss–Legendre quadrature formula; see Section 1.3 for a summary on quadrature methods. The functions $\eta''_{1,i}(z)$ and $\eta_{1,i}(y)$ are approximated in the following way:

$$\begin{aligned} \eta''_{1,i}(z) &= \int_0^z \frac{t(z-\pi)}{\pi}\mathrm{e}^{x_i\cos t}\,dt + \int_z^\pi \frac{z(t-\pi)}{\pi}\mathrm{e}^{x_i\cos t}\,dt \\ &\approx \frac{z-\pi}{\pi}\sum_{j=1}^m \widehat{\lambda}_j \widehat{t}_j \mathrm{e}^{x_i\cos\widehat{t}_j} + \frac{z}{\pi}\sum_{j=1}^m \widetilde{\lambda}_j(\widetilde{t}_j - \pi)\mathrm{e}^{x_i\cos\widetilde{t}_j}, \end{aligned} \tag{5.18}$$

and

$$\begin{aligned}\eta_{1,i}(y) &= \int_0^y \frac{z(y-\pi)}{\pi}\eta''_{1,i}(z)\,dz + \int_y^\pi \frac{y(z-\pi)}{\pi}\eta''_{1,i}(z)\,dz \\ &\approx \frac{y-\pi}{\pi}\sum_{j=1}^{m}\widehat{\lambda}_j\widehat{t}_j\eta''_{1,i}(\widehat{t}_j) + \frac{y}{\pi}\sum_{j=1}^{m}\widetilde{\lambda}_j(\widetilde{t}_j-\pi)\eta''_{1,i}(\widetilde{t}_j),\end{aligned} \tag{5.19}$$

where, for both equations, $\widehat{t}_j$ are the quadrature points in the integration interval of the first integral and $\widehat{\lambda}_j$ are the corresponding quadrature weights, and $\widetilde{t}_j$ are the quadrature points in the integration interval of the second integral and $\widetilde{\lambda}_j$ are the corresponding weights. All these elements are obtained by equation (1.24) starting from Legendre quadrature points and weights defined in $[-1, 1]$.

5.3 FDEM data inversion

In the following two paragraphs, Test Functions 10 and 11, we report a linear and a nonlinear model involved in applied geophysics. Electromagnetic induction (EMI) techniques are used to investigate soil properties in a non-destructive way.

In 1980 McNeill [76] developed a linear model to reproduce the readings of one of the first available ground conductivity meters (GCM), the Geonics EM-38. During the last decades, much effort has been made to retrieve the electrical conductivity distribution $\sigma(z)$ by the above described linear model. In [12], a Tikhonov regularization technique was implemented to reconstruct the conductivity profile from measurements obtained by positioning a GCM at various heights above the ground, while in [25] the Tikhonov approach was optimized by a projected conjugate gradient algorithm. A nonlinear forward model for predicting the EM response of the subsoil was described in 1982 by Wait in [96]. A regularized inversion algorithm was studied in [27, 29] and recently extended to process complex-valued data sets [28]. The algorithm, as well as the forward model, were coded in Matlab and included in a publicly available software package [26], which has already been employed in real-world applications [11, 28, 34]. In [62], the technique adopted in [12] was extended and applied to a nonlinear model for the same physical system, previously described in [99]. For other integral models in applied geophysics, see [39].

In the models, the soil is assumed to have a layered structure with n layers below ground level ($z_1 = 0$). Each subsoil layer, of thickness d_k (meters), ranges from depth z_k to z_{k+1}, $k = 1, \ldots, n-1$, and it is characterized by an electrical conductivity σ_k (Siemens/meter) and a magnetic permeability μ_k (Henry/meter), for $k = 1, \ldots, n$. The thickness of the deepest layer d_n, starting at z_n, is considered infinite.

The GCM is an electromagnetic (EM) device composed of two coils, a transmitter and a receiver, placed at a fixed distance ρ from each other. The two coils, operating at frequency f in Hertz, are at height h above the ground with their axes oriented either vertically or horizontally with respect to the ground surface. Both the depth z

and the height h are measured in meters. The measuring device generates a primary EM field that induces eddy currents in the conductive parts of the subsurface and measures the ratio between the secondary EM field produced by such currents and the primary field.

Test Function 10. Linear model. In this simplified linear model, the readings of the device only depend upon the electrical conductivity σ of the soil as a function of depth z. Mathematically, the model consists of two integral equations of the first kind:

$$\begin{cases} \displaystyle\int_0^\infty k^V(z+h)\sigma(z)\,dz = g^V(h), & h \in [0,\infty), \\ \displaystyle\int_0^\infty k^H(z+h)\sigma(z)\,dz = g^H(h), & h \in [0,\infty). \end{cases} \tag{5.20}$$

The first one describes the situation in which both coil axes are aligned vertically with respect to the ground level, whereas the second one corresponds to the horizontal orientation of the coils. In the above equations,

$$k^V(z) = \frac{4z}{(4z^2+1)^{3/2}}, \qquad k^H(z) = 2 - \frac{4z}{(4z^2+1)^{1/2}} \tag{5.21}$$

are the kernel functions corresponding to the vertical and horizontal orientation of the coils, respectively, $\sigma(z) \geq 0$ is the unknown function that represents the electrical conductivity of the subsoil at depth z below the ground surface, and $g^V(h)$, $g^H(h)$ are given right-hand sides that represent the apparent conductivity of the soil sensed by the device when it is held at height $h > 0$ over the ground, in correspondence to the two possible orientations of the coils.

Let us rewrite (5.20) in operatorial form

$$\begin{cases} K^V\sigma = g^V, \\ K^H\sigma = g^H, \end{cases}$$

where

$$(K^J\sigma)(h) = \int_0^\infty k^J(z+h)\sigma(z)\,dz, \qquad J \in \{V,H\},$$

are self adjoint operators from $L^2([0,\infty))$ into itself. Both operators K^V and K^H, seen as functions from $L^2([0,\infty))$ into itself, are compact and hence bounded; see [33]. Indeed

$$\begin{aligned} \int_0^\infty \int_0^\infty |k^V(z+h)|^2\,dz\,dh &= \int_0^\infty t\,|k^V(t)|^2\,dt = \frac{1}{4}, \\ \int_0^\infty \int_0^\infty |k^H(z+h)|^2\,dz\,dh &= \int_0^\infty t\,|k^H(t)|^2\,dt = \log 2 - \frac{1}{2}, \end{aligned}$$

and, by virtue of the Hilbert-Schmidt integral operator theorem, they define compact and hence bounded maps of $L^2([0,\infty))$ into itself.

From the numerical point of view, the numerical treatment of equations (5.20) is quite delicate, due to the ill-posed nature of first kind integral equations [52].

In [33], the authors proved the well-posedness of the model in suitable function spaces; three different collocation methods were also proposed and compared. To face the ill-conditioning of the resulting linear systems, especially when the data are affected by experimental errors, the authors resorted to the truncated (generalized) singular value decomposition as a regularization method.

Now, we explain how to modify the model, in order to apply the numerical method proposed in Chapter 4 and in [32]. Following [33], assuming the a priori information $\sigma(z) \leq \beta$, for $z > z_0$, we split each integral appearing in (5.20) into the sum

$$\int_0^\infty k_\ell(h,z)\sigma(z)\,dz = \int_0^{z_0} k_\ell(h,z)\sigma(z)\,dz + \int_{z_0}^\infty k_\ell(h,z)\sigma(z)\,dz, \qquad \ell = 1,2,$$

where $k_1(h,z) = k^V(h+z)$ and $k_2(h,z) = k^H(h+z)$. Moreover, we may assume that at a sufficient depth z_0 the electrical conductivity stabilizes, converging to a value β. There are significant cases where this value can be estimated a priori on the basis of the geophysical properties of the subsoil. Let us also mention that the value of α can be obtained by direct conductivity measures at $z = 0$, that is in the soil portion which corresponds to the first layer of the subsurface discretization. Given the expression (5.21) of the kernels, setting $\sigma(z) \simeq \beta$, for $z > z_0$ and z_0 sufficiently large, the last integral can be analytically computed. Then, the system becomes

$$\int_0^{z_0} k_\ell(h,z)\sigma(z)\,dz = g_\ell(h) - \beta \int_{z_0}^\infty k_\ell(h,z)\,dz, \qquad \ell = 1,2,$$

with $g_1(h) = g^V(h)$ and $g_2(h) = g^H(h)$. In this way, system (5.20) is replaced by

$$\begin{cases} (K_1\sigma)(h) := \displaystyle\int_0^{z_0} k_1(h,z)\sigma(z)\,dz = g_1(h) - \frac{\beta}{\theta(z_0,h)}, \\ (K_2\sigma)(h) := \displaystyle\int_0^{z_0} k_2(h,z)\sigma(z)\,dz = g_2(h) - \beta\left(\theta(z_0,h) - 2(h+z_0)\right), \end{cases} \tag{5.22}$$

where the right-hand sides are corrected by the available a priori information on the unknown function and (see [33])

$$\theta(z,h) = \sqrt{4(z+h)^2 + 1}. \tag{5.23}$$

We remark that $(-\theta(z,h))^{-1}$ is the primitive function of $k_1(h,z)$, and $2z - \theta(z,h)$ is that of $k_2(h,z)$.

To determine a solution by applying the theory developed in Sections 4.2 and 4.3, it is necessary to introduce the linear function (4.5)

$$\gamma(z) = \left(1 - \frac{z}{z_0}\right)\alpha + \frac{z}{z_0}\beta,$$

and assume that the values of the electrical conductivity at the endpoints of the integration interval are known, e.g., $\sigma(0) = \alpha$ and $\sigma(z_0) = \beta$. The boundary values can usually be approximated in applications; see [33].

By collocating equations (5.22) at the points h_i, assuming $n_1 = n_2 = n$ and $h_{1,i} = h_{2,i} = h_i$, for $i = 1, \dots, n$, we obtain the following system of integral equations with discrete data

$$\begin{cases} \displaystyle\int_0^{z_0} k_1(h_i, z)\phi(z)\,dz = \psi_1(h_i), \quad i = 1, \dots, n, \\ \displaystyle\int_0^{z_0} k_2(h_i, z)\phi(z)\,dz = \psi_2(h_i), \quad i = 1, \dots, n, \end{cases}$$

where

$$\phi(z) = \sigma(z) - \gamma(z)$$

is the new unknown function, and considering the fact that

$$\int k_1(h_i, z)z\,dz = \frac{1}{2}\operatorname{arcsinh}(2(z + h_i)) - \frac{z}{\theta(z, h_i)},$$

$$\int k_2(h_i, z)z\,dz = \frac{\theta(z, h_i)}{2}(h_i - z) + z^2 - h_i^2 + \frac{1}{4}\operatorname{arcsinh}(2(z + h_i)),$$

then

$$\begin{aligned} \psi_1(h_i) &= g_1(h_i) - \frac{\beta}{\theta(z_0, h_i)} - \int_0^{z_0} k_1(h_i, z)\gamma(z)\,dz \\ &= g_1(h_i) - \frac{\alpha}{\theta(0, h_i)} - \frac{\alpha - \beta}{2z_0}\left[\operatorname{arcsinh}(2h_i) - \operatorname{arcsinh}(2(z_0 + h_i))\right], \\ \psi_2(h_i) &= g_2(h_i) - \beta\left(\theta(z_0, h_i) - 2(h_i + z_0)\right) - \int_0^{z_0} k_2(h_i, z)\gamma(z)\,dz \\ &= g_2(h_i) - \left[\frac{(\alpha - \beta)h_i}{2z_0} + \alpha\right]\theta(0, h_i) + \frac{\alpha - \beta}{2}\left[\frac{h_i}{z_0} + 1\right]\theta(z_0, h_i) \\ &\quad + 2\beta h_i - z_0(\alpha - \beta) - \frac{a - b}{4z_0}\left[\operatorname{arcsinh}(2h_i) - \operatorname{arcsinh}(2(z_0 + h_i))\right] \end{aligned}$$

are the new right-hand sides.

The Green function, for the interval $[0, z_0]$, is

$$G_y''(z) = \begin{cases} \dfrac{z(y - z_0)}{z_0}, & 0 \le z < y, \\ \dfrac{y(z - z_0)}{z_0}, & y \le z \le z_0. \end{cases}$$

The second derivative of the Riesz representers can be computed analytically. Indeed, from (4.22), it follows that

$$\begin{aligned}\eta''_{1,i}(x) &= \int_0^{z_0} G'''_z(x) k_1(h_i, z)\, dz \\ &= \frac{1}{2}\left[\left(1 - \frac{x}{z_0}\right)\operatorname{arcsinh}(2h_i) - \operatorname{arcsinh}(2(x+h_i))\right. \\ &\qquad \left. + \frac{x}{z_0}\operatorname{arcsinh}(2(z_0+h_i))\right]\end{aligned}$$

and

$$\begin{aligned}\eta''_{2,i}(x) &= \int_0^{z_0} G'''_z(x) k_2(h_i, z)\, dz \\ &= \frac{1}{2}\left[2x(x-z_0) + x\left(1 + \frac{h_i}{z_0}\right)\theta(z_0, h_i) - (x+h_i)\theta(x, h_i)\right. \\ &\qquad \left. + h_i\left(1 - \frac{x}{z_0}\right)\theta(0, h_i) + \eta''_{1,i}(x)\right],\end{aligned}$$

where $\theta(x, h)$ is the function defined in (5.23). From the above second derivatives, we can compute the Riesz functions

$$\begin{aligned}\eta_{1,i}(y) &= \int_0^{z_0} G''_y(x)\eta''_{1,i}(x)\, dx \\ &= \frac{3}{16}\left[(y+h_i)\theta(y, h_i) - y\left(1 + \frac{h_i}{z_0}\right)\theta(z_0, h_i) + h_i\left(\frac{y}{z_0} - 1\right)\theta(0, h_i)\right] \\ &\quad + \frac{1}{2}\left\{\left[\frac{1}{2}\left(\frac{y}{z_0} - 1\right)\left(\frac{1}{8} - h_i^2 - \frac{y^2}{3}\right) + \frac{y}{3}(y - z_0)\right]\operatorname{arcsinh}(2h_i)\right. \\ &\quad + \left[-\frac{y}{2z_0}\left(\frac{1}{8} - h_i^2 - \frac{y^2}{3}\right) + y\left(h_i + \frac{z_0}{3}\right)\right]\operatorname{arcsinh}(2(z_0+h_i)) \\ &\quad \left. + \frac{1}{2}\left[\frac{1}{8} - (y+h_i)^2\right]\operatorname{arcsinh}(2(y+h_i))\right\}\end{aligned}$$

and

$$
\begin{aligned}
\eta_{2,i}(y) &= \int_0^{z_0} G''_y(x)\eta''_{2,i}(x)\,dx \\
&= \frac{1}{192z_0}\Big\{ z_0\left[h_i\left(13-8(3h_iy+h_i^2+3y^2)\right)+y(13-8y^2)\right]\theta(y,h_i) \\
&\quad +y\left[h_i\left(8(3h_iz_0+h_i^2+2y^2+z_0^2)-13\right)+z_0\left(8(2y^2-z_0^2)-13\right)\right]\theta(z_0,h_i) \\
&\quad +h_i\left[z_0\left(8(h_i^2+6y^2-4yz_0)-13\right)+y\left(13-8(h_i^2+2y^2)\right)\right]\theta(0,h_i) \\
&\quad +16yz_0\left(y^3-2y^2z_0+z_0^3\right)\Big\} \\
&\quad +\frac{1}{128z_0}\left\{(y-z_0)\left[1-16\left(h_i^2+\frac{y^2}{3}-\frac{2yz_0}{3}\right)\right]\operatorname{arcsinh}(2h_i)\right. \\
&\quad +z_0\left[1-16\,(y+h_i)^2\right]\operatorname{arcsinh}(2(y+h_i) \\
&\quad \left.-y\left[1-16\left(h_i^2+\frac{y^2}{3}+2h_iz_0+\frac{2z_0^2}{3}\right)\right]\operatorname{arcsinh}(2(z_0+h_i))\right\}.
\end{aligned}
$$

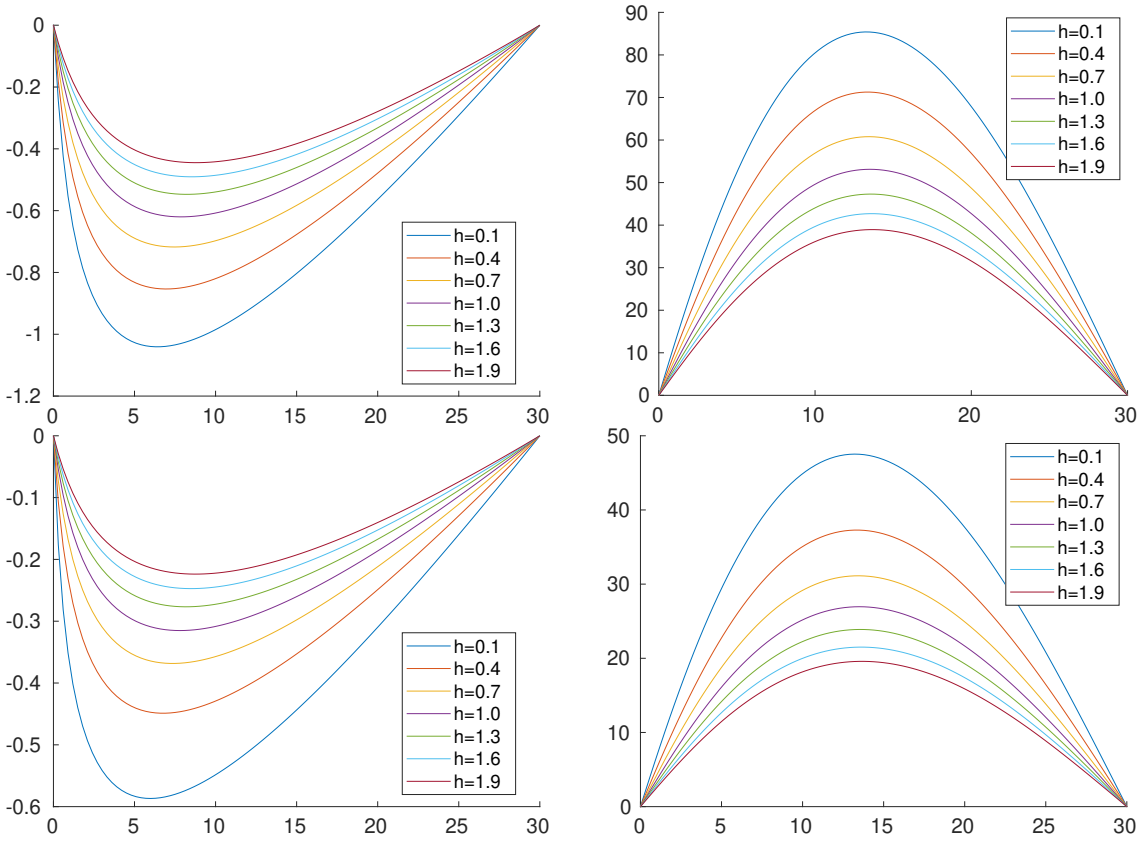

Figure 5.2: The functions $\eta''_{1,i}$ (top-left), $\eta_{1,i}$ (top-right), $\eta''_{2,i}$ (bottom-left), and $\eta_{2,i}$ (bottom-right), with $h_i = 0.1 + (i-1)\frac{3}{10}$ and $i = 1, \ldots, 7$.

Figure 5.2 shows the behavior of $\eta''_{1,i}$ and $\eta_{1,i}$ for different values of h_i, in the case $z_0 = 30$. We also report the graphs of the Riesz representers for the horizontal orientation $\eta''_{2,i}$ and $\eta_{2,i}$. It is straightforward to verify that $\eta_{1,i}(0) = \eta_{1,i}(z_0) = \eta_{2,i}(0) = \eta_{2,i}(z_0) = 0$. Figure 5.3 displays the orthonormal functions $\widehat{\eta}_{1,i}$ and $\widehat{\eta}_{2,i}$ defined in (4.31), together with their second derivatives $\widehat{\eta}''_{1,i}$ and $\widehat{\eta}''_{2,i}$. In the summation (4.31), the upper bound is set to $N = 12$ for preserving the positivity of the eigenvalues.

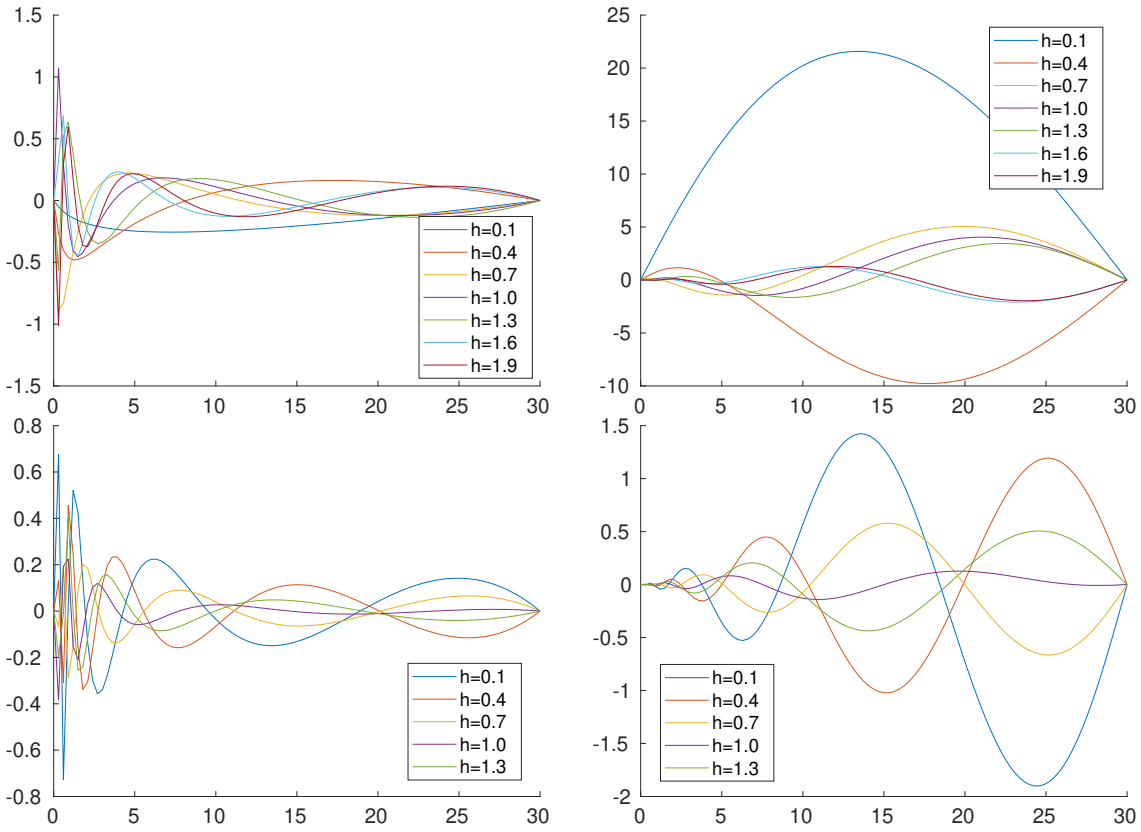

Figure 5.3: The orthonormal functions $\widehat{\eta}''_{1,i}$ (top-left), $\widehat{\eta}_{1,i}$ (top-right), $\widehat{\eta}''_{2,i}$ (bottom-left), and $\widehat{\eta}_{2,i}$ (bottom-right), with $h_i = 0.1 + (i-1)\frac{3}{10}$ and $i = 1, \ldots, 7$.

Test Function 11. Nonlinear model. We report here a nonlinear model for the same applicative setting. It is derived from Maxwell's equations.

Let $u_k(\lambda) = \sqrt{\lambda^2 + \mathrm{i}\sigma_k\mu_k\omega}$ be the propagation constant, where $\omega = 2\pi f$ is the angular frequency of the instrument. The variable λ ranges from zero to infinity and it measures the ratio between the depth below the ground surface and the inter-coil distance ρ. If we denote the characteristic admittance in the kth layer by

$$N_k(\lambda) = \frac{u_k(\lambda)}{\mathrm{i}\mu_k\omega}, \quad k = 1, \ldots, n,$$

then it is shown in [96] that the surface admittance $Y_k(\lambda)$ at the top of the same layer verifies the recursion

$$\begin{cases} Y_n(\lambda) = N_n(\lambda), \\ Y_k(\lambda) = N_k(\lambda)\dfrac{Y_{k+1}(\lambda) + N_k(\lambda)\tanh(d_k u_k(\lambda))}{N_k(\lambda) + Y_{k+1}(\lambda)\tanh(d_k u_k(\lambda))}, \quad k = n-1, \dots, 1. \end{cases} \tag{5.24}$$

Let us define the reflection factor as

$$R_\omega(\lambda) = \frac{N_0(\lambda) - Y_1(\lambda)}{N_0(\lambda) + Y_1(\lambda)},$$

where $Y_1(\lambda)$ is computed by the recursion (5.24) and $N_0(\lambda) = \lambda/(\mathrm{i}\mu_0\omega)$, with $\mu_0 = 4\pi 10^{-7}\mathrm{H/m}$, that is, the value of the magnetic permeability in the empty space. The ratio of the secondary to the primary field for the vertical and horizontal orientation of the coils, respectively, is given by

$$\begin{cases} M_0(\boldsymbol{\sigma}, \boldsymbol{\mu}; h, \omega, \rho) = -\rho^3 \displaystyle\int_0^\infty \lambda^2 e^{-2h\lambda} R_\omega(\lambda) J_0(\rho\lambda)\, d\lambda, \\ M_1(\boldsymbol{\sigma}, \boldsymbol{\mu}; h, \omega, \rho) = -\rho^2 \displaystyle\int_0^\infty \lambda e^{-2h\lambda} R_\omega(\lambda) J_1(\rho\lambda)\, d\lambda, \end{cases}$$

where $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_n]^T$, $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n]^T$, and J_0, J_1 are Bessel functions of the first kind of order 0 and 1, respectively.

The functions M_0 and M_1 can be expressed in a more compact form in terms of the Hankel transform

$$\mathcal{H}_\nu[f](\rho) = \int_0^\infty f(\lambda) J_\nu(\rho\lambda)\lambda\, d\lambda,$$

as follows

$$M_\nu(\boldsymbol{\sigma}, \boldsymbol{\mu}; h, \omega, \rho) = -\rho^{3-\nu}\mathcal{H}_\nu[\lambda^{1-\nu}\, e^{-2h\lambda}\, R_\omega(\lambda)](\rho), \qquad \nu = 0, 1.$$

As it is usual in many applications, we let the magnetic permeability take the constant value μ_0, that is, the value in the empty space. According to the instrument configuration (orientation of the coils, height above the ground, inter-coil distance, alternating current frequency) multiple measurements are available. We will denote them by b_i, $i = 1, \dots, m$, and the model prediction by $F(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_n]^T$. Then, the problem of data inversion consists of computing the conductivity vector $\boldsymbol{\sigma}$ which determines the best fit to the data vector $\mathbf{b}$, that is, the one which solves the problem

$$\min_{\boldsymbol{\sigma} \in \mathbb{R}^n} \|\mathbf{r}(\boldsymbol{\sigma})\|^2, \qquad \text{with } \mathbf{r}(\boldsymbol{\sigma}) = F(\boldsymbol{\sigma}) - \mathbf{b}.$$

CHAPTER 6

Numerical experiments

This chapter is devoted to testing the algorithms proposed in this thesis. In Section 6.1 we apply the MNGN method and its regularized variants, described in the first part of Chapter 2, to nonlinear least-squares problems, while in Section 6.2, the "doubly relaxed" approaches of the MNGN method, analyzed in the second part of Chapter 2, are tested. In Section 6.3 we pass to solving some systems of integral equations by applying the Riesz theory, developed in Chapter 4. All computational codes are implemented in Matlab. The numerical experiments were performed on an Intel Core i5 system with 16Gb RAM, running the Debian GNU/Linux operating system.

6.1 The MNGN method in action

In this section, we present two classes of numerical examples. In the first one, we apply the minimal-norm Gauss–Newton (MNGN) method of Section 2.3 to two well-conditioned problems of small dimension in order to visualize its convergence and compare it to the standard Gauss–Newton iteration. In the second class, we apply the regularization techniques of Sections 2.5.1 and 2.5.2, that is, the truncated minimal-L-norm Gauss–Newton (TMLNGN) method and the minimal-L-norm Tikhonov–Gauss–Newton (TikLGN) method, to an ill-conditioned nonlinear problem of larger size. In each experiment, we solve problem (2.1) for a particular function $F(\mathbf{x})$ with values in $\mathbb{R}^m$ for $\mathbf{x} \in \mathbb{R}^n$.

Regarding the stopping rule, the following ones are adopted for all iterative methods. We iterate until either the difference between two successive approximations is small enough

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \delta \|\mathbf{x}^{(k+1)}\|, \qquad \delta = 10^{-8},$$

or until the maximum number of iterations $N_{\max} = 60$ is reached.

For ill-conditioned problems, it is useful to consider an additional stopping criterion in order to detect the unboundedness of the solution for a particular value of

the regularization parameter. The iteration is interrupted when one of the preceding conditions is reached or when the ratio between the norms of the kth approximate solution and the initial point is larger than 10^8.

Example 6.1.1. In the first example, we consider the nonlinear function (5.1) defined in Test Function 1. This is a well-conditioned example. We recall that the residual $r : \mathbb{R}^2 \to \mathbb{R}$ is the nonlinear function

$$r(\mathbf{x}) = F(\mathbf{x}) - b = \left[\alpha(x_1 - 1)^2 + \beta(x_2 - 1)^2 - 1\right]^2 + 1,$$

depending upon the parameters $\alpha, \beta \in \mathbb{R}$. We minimize the objective function

$$f(\mathbf{x}) = \frac{1}{2} r(\mathbf{x})^2 = \frac{1}{2} \left\{ \left[\alpha(x_1 - 1)^2 + \beta(x_2 - 1)^2 - 1\right]^2 + 1 \right\}^2, \tag{6.1}$$

which can be graphically represented by a surface; see Figure 6.1.

In this case, the least-squares problem (2.1) is underdetermined, so it has infinitely many solutions. We solve it by Newton's method (2.4), the Gauss–Newton method (2.7), the minimal-norm Gauss–Newton method (2.13), and the "projected" Newton method discussed in Remark 2.3.5.

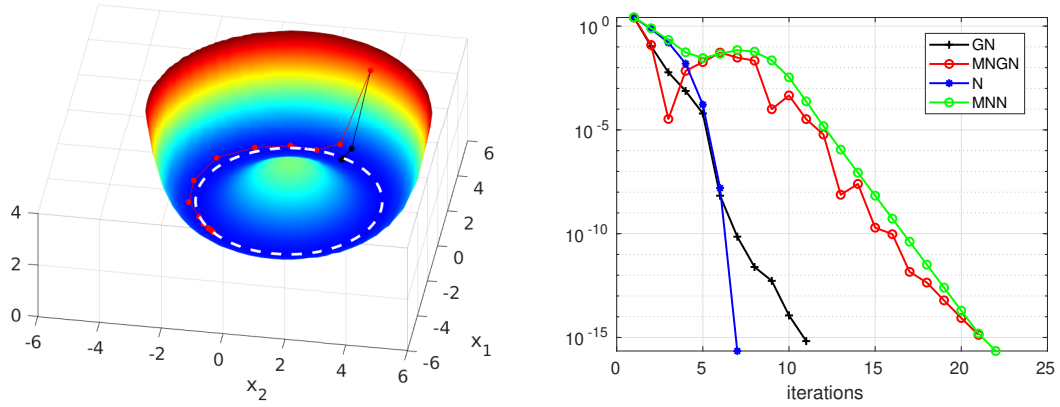

Figure 6.1: Convergence of problem (6.1) with $\alpha = \beta = \frac{1}{9}$ and $\mathbf{x}^{(0)} = [5, 3]^T$. In the 3D graph on the left, the white dashed line represents the locus of the solutions, the red dots are the iterations of the MNGN method, and the black ones correspond to the GN method. The graph on the right reports the residuals for each method.

First we consider $\alpha = \beta = \frac{1}{9}$. In this case, the minimal-norm solution is $\mathbf{x}^\dagger \approx [-1.12, -1.12]^T$, with $\|\mathbf{x}^\dagger\| \approx 1.58$. Figure 6.1 illustrates the progress of the iterations: the graph on the left displays the iterates produced by the MNGN and the GN methods in a 3D representation of $f(\mathbf{x})$. The one on the right reports the residuals corresponding to the above methods, to Newton's method (N), and to the "projected" Newton method (MNN). The last two methods converge to the same

solutions as the GN and MNGN methods, respectively, so they are not represented in the 3D plot.

All the methods reach convergence as the residuals converge to zero. We see that MNGN takes longer to converge as it must "travel" across the solutions locus to reach the minimal-norm solution, which is the one nearby the origin. On the contrary, GN converges to the solution closer to the initial point. This fact is even clearer in the contour plot on the left of Figure 6.2.

Observing the residuals, we see that Newton's method has the highest convergence rate. Anyway, if we trivially project its iterates orthogonally to the null space of the Jacobian (see Remark 2.3.5), then it converges to the minimal-norm solution, but its speed of convergence degrades and equals the MNGN method. So, no computational gain derives from its higher complexity.

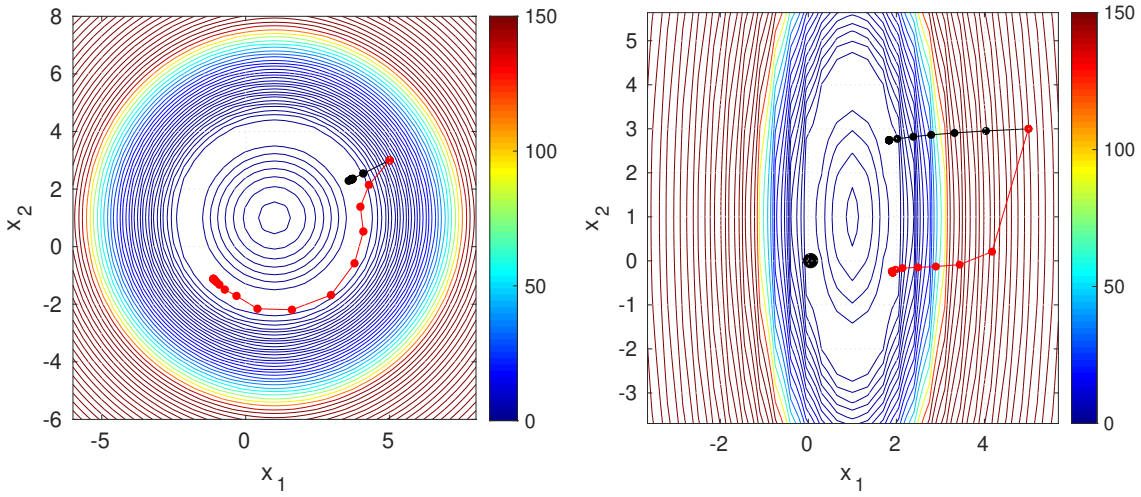

Figure 6.2: Contour plots for problem (6.1): on the left $\alpha = \beta = \frac{1}{9}$, on the right $\alpha = 1$ and $\beta = \frac{1}{10}$. The red dots are the iterates of the MNGN method and the black ones the approximations produced by the GN method. The thick black dot in the graph on the right is the minimal-norm solution.

It is also interesting to observe that the residuals of the MNGN method are not monotonically decreasing. The method, in some measure, is able to step away from the local attraction basin in order to chase the minimal-norm solution. Anyway, the dependence upon the initial point $\mathbf{x}^{(0)}$ is obviously maintained. This is shown in the contour plot on the right of Figure 6.2, which illustrates the convergence of the GN and MNGN methods when $\alpha = 1$ and $\beta = \frac{1}{10}$, starting from the same initial point. The MNGN method converges, in this case, to a solution with a smaller norm (i.e., with a smaller distance from the origin) than the one computed by the GN method but not to the minimal-norm solution, identified by a thick black dot in the graph.

Example 6.1.2. As a second well-conditioned example we consider the nonlinear function (5.2) defined in Test Function 2. The residual function $\mathbf{r} : \mathbb{R}^3 \to \mathbb{R}^2$ has

the expression

$$\mathbf{r}(\mathbf{x}) = \begin{bmatrix} (x_1 - 1)^2 + x_2^2 + x_3^2 - 1 \\ x_3 \end{bmatrix},$$

with $\mathbf{x} = [x_1, x_2, x_3]^T$.

The objective function f is represented in Figure 6.3 by a contour-slice volume plot, from two different points of view, together with the iterates of the GN and MNGN methods. Contour-slices are contour plots drawn at specific planes within the volume and they show where data values are equal on these planes.

As it is expected, the first method converges to the solution closer to the initial point, while MNGN converges to the minimal-norm solution.

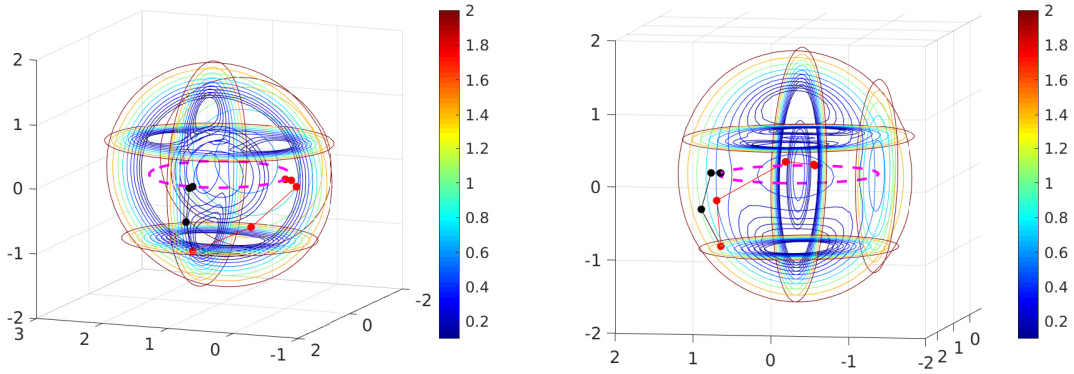

Figure 6.3: Contour-slice volume plot for problem (5.2): iterates of the MNGN method (red dots) and of the GN method (black dots), starting from the initial point $\mathbf{x}^{(0)} = [1.01, 1, -1]^T$. The dashed circle represents the locus of solutions.

Example 6.1.3. Here we consider a nonlinear model applied to investigate soil properties, which involves the use of electromagnetic induction techniques. A brief description of the model can be found in Test Function 11. We denote the measurements by b_i, $i = 1, \ldots, m$, and the model prediction by $F(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_n]^T$ represents the electrical conductivity in the discretization layers. Then, the problem of data inversion consists of computing the conductivity vector $\boldsymbol{\sigma}$ such that

$$\min_{\boldsymbol{\sigma} \in \mathbb{R}^n} \|\mathbf{r}(\boldsymbol{\sigma})\|^2, \qquad \text{with } \mathbf{r}(\boldsymbol{\sigma}) = F(\boldsymbol{\sigma}) - \mathbf{b}. \tag{6.2}$$

This is an ill-conditioned example.

In our numerical simulation, we fix the following test model for the electrical conductivity as a function of depth,

$$\sigma(z) = \mathrm{e}^{-(z-1.2)^2}. \tag{6.3}$$

We discretize the soil by $n = 20$ uniformly spaced layers up to the depth of 3.5m and we assign to each layer the conductivity $\sigma_i = \sigma(z_i)$, $i = 1, \ldots, n$, with $z_1 = 0$m and $z_n = 3.5$m. We choose the configuration of an existing device (the Geophex GEM-2), using a single pair of coils at 1.66m distance and 5 different current frequencies. This means that it can acquire 5 measurements for each sampling. The forward model generates a noise-free data vector $\mathbf{b}_{\text{exact}}$ of m synthetic measurements, corresponding to placing the instrument at two different heights above the ground (0.75m and 1.5m) with the coils either in vertical orientation ($m = 10$) or in vertical and horizontal orientations ($m = 20$). To simulate experimental errors, the noise-free data vector $\mathbf{b}_{\text{exact}}$ is perturbed by

$$\mathbf{b} = \mathbf{b}_{\text{exact}} + \frac{\varepsilon \|\mathbf{b}_{\text{exact}}\|}{\sqrt{m}} \mathbf{w},$$

where $\mathbf{w}$ is a normally distributed random vector with zero mean and unitary variance and ε represents the noise level.

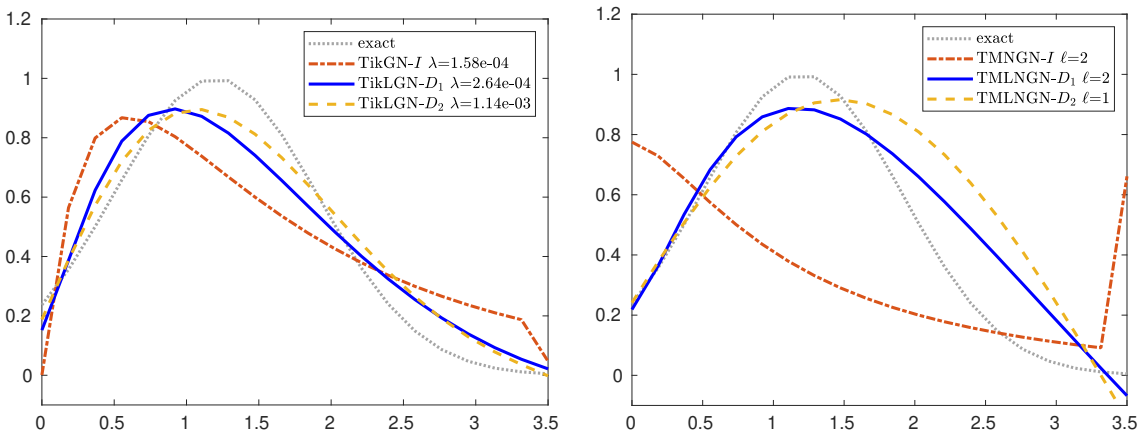

Figure 6.4: EM data inversion: $m = n = 20$, noise level $\varepsilon = 10^{-2}$, comparison of the solution corresponding to the regularization matrices $L = I_n$, D_1, and D_2. The initial point $\boldsymbol{\sigma}^{(0)}$ has random components uniformly distributed in the interval $(49.5, 50.5)$. The exact solution is compared to the solutions computed by TikGN/TikLGN on the left and by TMNGN/TMLNGN on the right. The parameters λ and ℓ are the best possible.

We solve problem (6.2) by the damped Gauss–Newton method with the damping parameter determined by the Armijo–Goldstein principle. Each step of the iterative method is regularized by one of the methods described in this chapter. In the standard case, when $L = I_n$, we display the solutions computed by the TMNGN (2.26) and TikGN (2.33) methods; when a regularization matrix is present, that is, when $L = D_1$ or D_2 (see (1.14)), we apply the TMLNGN (2.27) and TikLGN (2.37) methods. The regularization parameters λ and ℓ are chosen by different criteria: by minimizing the 2-norm error with respect to the exact solution in order to ascertain

the best possible performance of the methods, by the discrepancy principle and by the L-curve criterion to test the algorithms in a realistic situation (see Section 2.5).

We start by discussing the importance of the regularization matrix L for the accuracy of the solution. The data set is composed by $m = 20$ measurements, the noise level is $\varepsilon = 10^{-2}$, a value consistent with experimental data sets, and the initial vector is $\boldsymbol{\sigma}^{(0)}$, whose components are uniformly distributed random numbers in the interval $(49.5, 50.5)$.

The model function (6.3) is smooth, favoring a regularizing term based on the approximation of the first or second derivatives. The graphs in Figure 6.4 compare the solutions obtained by the regularization matrices $L = I_n$, D_1, and D_2. The computation is performed by using a Tikhonov approach (graph on the left) and by truncating the SVD/GSVD in the minimal-norm Gauss–Newton iteration (on the right). In the first case the solution corresponding to $L = I_n$ is evidently less accurate than the others. In the second one, the minimal-norm method TMNGN converges to a solution which is totally different from the model function, while the other two reconstructions are close to it. We also observe that in this case, as it happened in other experiments, Tikhonov regularization can reach a higher accuracy than the truncated SVD/TSVD approach. This is due to the fact that the regularization parameter λ can be varied continuously, while the parameter ℓ can only take integer values. In this example and in the following one, both parameters are chosen by minimizing the 2-norm error.

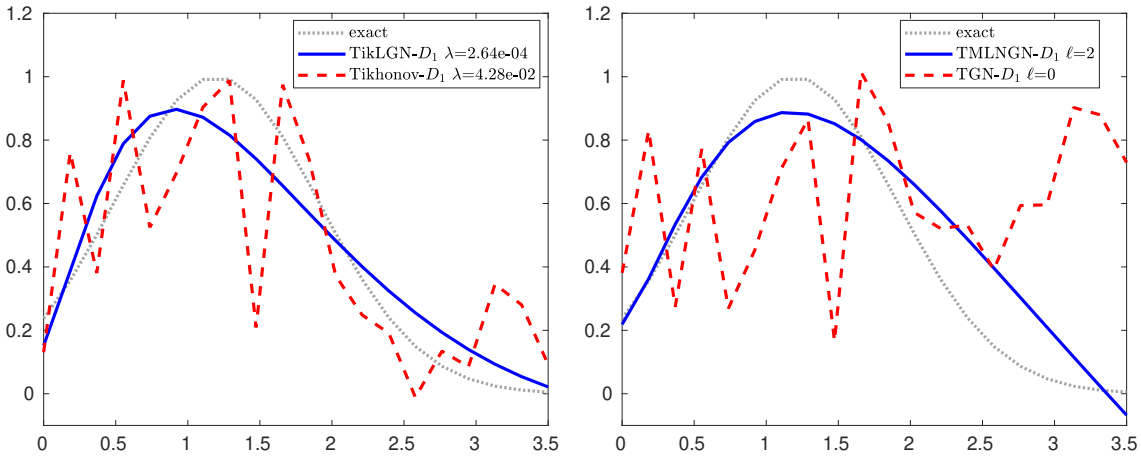

Figure 6.5: EM data inversion: $m = n = 20$, noise level $\varepsilon = 10^{-2}$, regularization matrix $L = D_1$, initial point $\boldsymbol{\sigma}^{(0)}$ with random components uniformly distributed in the interval $(49.5, 50.5)$. The exact solution is compared to the solutions computed by TikLGN and the standard Tikhonov method (on the left) and by TMLNGN and by the Gauss–Newton method regularized by TGSVD, labelled as TGN, (on the right). The parameters λ and ℓ are the best possible.

In many cases, especially when the initial vector used to initialize the iteration is close enough to the solution of the problem, the minimal-norm and the standard ap-

proaches produce similar approximations. Anyway, when the initial vector is rather far away from the solution, there are cases in which the minimal-norm methods are significantly more accurate and less sensible to the presence of local minima than the traditional approaches.

Figure 6.5 shows one of these cases. Here the minimal-norm algorithms are compared to the traditional approaches, namely, Tikhonov regularization (2.25) and the Gauss–Newton method regularized by TGSVD, labeled as TGN. The regularization matrix is the discretization of the first derivative operator; the other parameters are the same as in the previous example. We observe from both graphs of Figure 6.5 that the minimal-L-norm approaches, i.e., the TikLGN and the TMLNGN methods, reproduce acceptable solutions, while the approximations from the Tikhonov approach (2.25) and the TGN method are completely wrong.

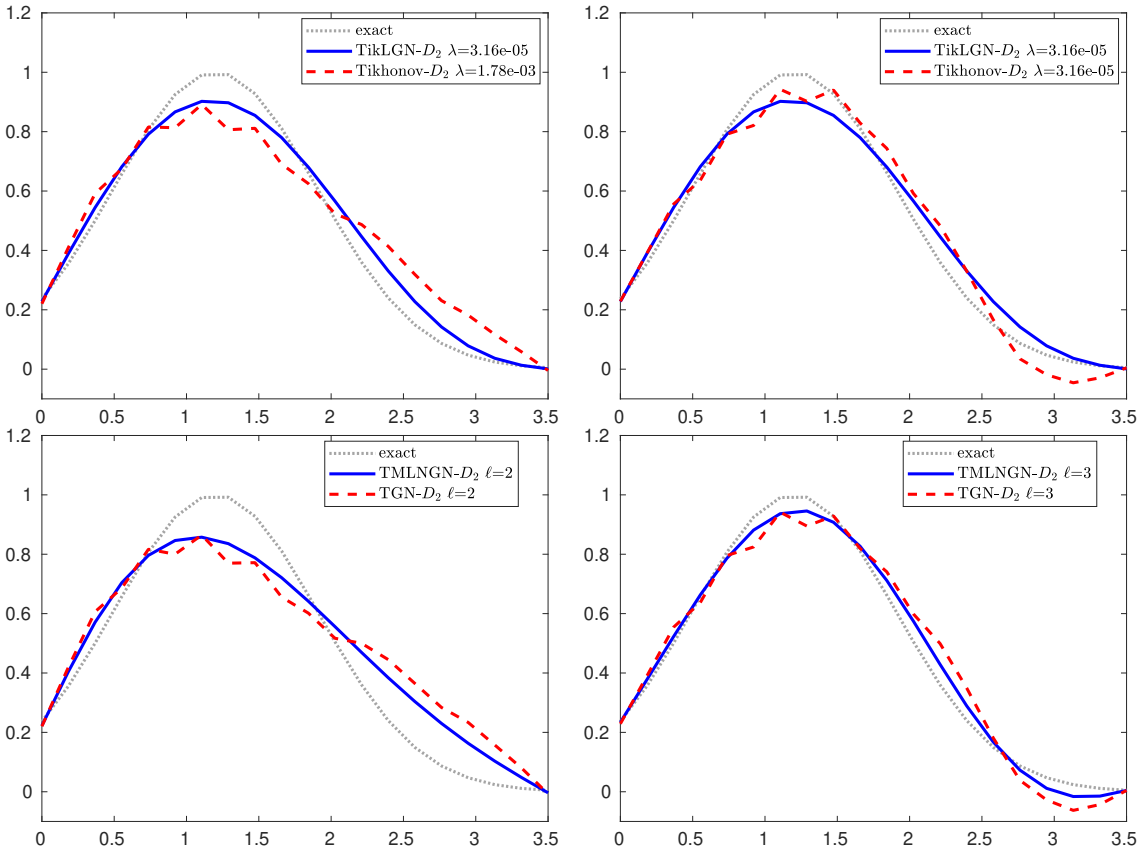

Figure 6.6: EM data inversion: $m = 10$, $n = 20$, noise level $\varepsilon = 10^{-4}$, regularization matrix $L = D_2$. The initial point has $\boldsymbol{\sigma}^{(0)}$ with random components uniformly distributed in the interval $(0.45, 0.55)$. The exact solution is compared to the solutions computed by TikLGN and the standard Tikhonov method (top row) and by TMLNGN/TGN (bottom row). The parameters λ and ℓ have been chosen by the discrepancy principle (left column) and by the L-curve (right column).

In Figure 6.6 we illustrate the performance of the discrepancy principle, with $\tau = 1.6$, and of the L-curve criterion in estimating the regularization parameters λ and ℓ. In this example, we consider $m = 10$ data values, the initial solution is $\boldsymbol{\sigma}^{(0)}$, whose components are uniformly distributed random numbers in the interval $(0.45, 0.55)$, and the noise level $\varepsilon = 10^{-4}$. The regularization matrix is the discretization of the second derivative D_2. The graphs in the top row concern the reconstructions obtained by the TikLGN and Tikhonov methods. In the bottom row we report the results obtained by the TMLNGN and the TGN methods. The regularization parameters for the graphs in the left column are determined by the discrepancy principle, while the right column shows the reconstructions corresponding to the regularization parameters estimated by the L-curve. From all four plots, we can see that the minimal-L-norm solutions are more accurate.

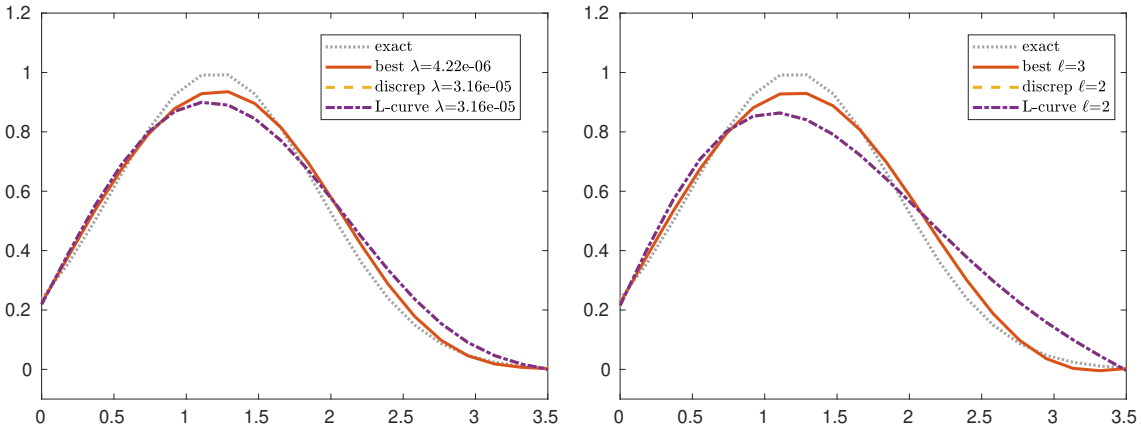

Figure 6.7: EM data inversion: $m = 10$ and $n = 20$, noise level $\varepsilon = 10^{-4}$, regularization matrix $L = D_2$, initial point $\boldsymbol{\sigma}^{(0)}$ with random components uniformly distributed in the interval $(0.4, 0.6)$. The exact solution is compared to the solutions computed by TikLGN (on the left) and by TMLNGN (on the right).

In Figure 6.7, we compare the "best" solution for the noise level $\varepsilon = 10^{-4}$ to the ones obtained by estimating the regularization parameter by the discrepancy principle and by the L-curve criterion. Here, we consider $m = 10$ data values, the initial solution is $\boldsymbol{\sigma}^{(0)}$, whose components are uniformly distributed random numbers in the interval $(0.4, 0.6)$. The regularization matrix is the discretization of the second derivative operator D_2. On the left we report the approximate solutions obtained by the TikLGN method, while the right pane shows the reconstructions obtained by the TMLNGN method. Approximate solutions from all parameter selection criteria are acceptable.

6.2 Performance of the doubly relaxed MNGN method

This section is devoted to analyzing the behavior of the "doubly relaxed MNGN" approach for several test problems.

The developed Matlab functions implement all the variants of the MNGN2 algorithm, defined by (2.51), as well as the MNGN and CKB methods developed in [82] and [20], respectively, and reported in this thesis.

In the following, the MNGN2 algorithm will be denoted by different names, according to the particular implementation. In the method denoted by MNGN2$_\alpha$, we let $\beta_k = \alpha_k$ in (2.51), and determine α_k by the Armijo–Goldstein principle. Algorithm 1 is denoted by MNGN2$_{\alpha\beta}$, when $\delta(\rho, \eta) = \eta\rho$, with a fixed value of η. The same algorithm with $\delta(\rho, \eta) = \rho^\eta$, and η estimated by Algorithm 2, is labeled as MNGN2$_{\alpha\beta\delta}$. The algorithm (2.49) developed in [20] is denoted by CKB$_1$ when $\gamma_k = (0.5)^{k+1}$, and by CKB$_2$ when $\gamma_k = (0.5)^{2^k}$. The same algorithms are denoted by rCKB$_1$ and rCKB$_2$ when they are applied with the automatic estimation of the rank of the Jacobian, discussed in Section 2.8.

We note that the computational cost of each iteration is roughly the same for all the methods considered. Indeed, the additional complexity required by the MNGN2 algorithms consists of the estimation of the numerical rank $r_{\epsilon,k}$, of the residual increase $\delta(\rho, \eta)$, and of the projection parameter β_k. All these computations involve a small number of floating point operations; see also Remark 2.9.1.

To compare the methods and investigate their performance, we performed numerical experiments on various test problems that highlight particular difficulties in the computation of the minimal-norm solution. Example 6.2.1 illustrates a situation where the MNGN method either fails or produces unacceptable results, while the other methods perform well; in Example 6.2.2, we investigate the dependence of the MNGN2$_{\alpha\beta}$ method on the choice of the parameter η; Example 6.2.3 is the first medium-size test problem we consider, it shows the importance of the Jacobian rank estimation for the effectiveness of the algorithms; in Example 6.2.4, the methods are compared in the solution of minimal-L-norm problems with different regularization matrices; finally, in Example 6.2.5, we let the dimension of the problem vary and we explore the dependence of the computed solution on the availability of a priori information in the form of a model profile.

In this section, we apply the MNGN2 methods also in the large-scale case in Example 6.2.4 and Example 6.2.5. The different approaches are denoted as GK-MNGN2$_\alpha$, GK-MNGN2$_{\alpha\beta}$, and GK-MNGN2$_{\alpha\beta\delta}$, where GK stands for "Golub–Kahan". We consider also the CKB methods and we denote them as GK-CKB$_1$ and GK-CKB$_2$. To detect a breakdown we set $tol = 10^{-8}$ in Algorithm 3.

For each experiment, we repeated the computation 100 times, varying the starting point $\mathbf{x}^{(0)}$ by letting its components be uniformly distributed random numbers in $(-5, 5)$. The model profile $\overline{\mathbf{x}}$ was set to the zero vector except in Example 6.2.5.

We consider a numerical test a "success" if the algorithm converges according to condition (2.53), with stop tolerance $\tau = 10^{-8}$ and maximum number of iterations $N_{\max} = 500$. A failure is not a serious problem, in general, because non-convergence simply suggests to try a different starting vector. Anyway, if this happens too often, it increases the computational load. At the same time, a success of a method does not imply that it recovers the minimal-norm solution, as the convergence is only local. So, to give an idea of the performance of the methods, we measure over all the tests the average of both the number of iterations required and the norm of the converged solution $\|\widetilde{\mathbf{x}}\|$. We also report the number of successes.

Example 6.2.1. In this first example, we consider the nonlinear model (5.9) described in Test Function 7, that concerns the behavior of a redundant parallel robot. In our simulation we assume $(X, Y) = (3, 3)$, $A = 2$, $H = 10$.

Table 6.1: Results for Example 6.2.1.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
MNGN2$_\alpha$	239	8.7246	92
MNGN2$_{\alpha\beta\delta}$	38	9.0621	96
CKB$_1$	26	8.5515	100
CKB$_2$	10	9.7344	100
MNGN	182	17.6329	30

The results obtained are reported in Table 6.1. We see that the MNGN2$_\alpha$ and CKB$_1$ methods recover solutions with smaller norms, in the average, but the first one requires a large number of iterations. The MNGN2$_{\alpha\beta\delta}$ implementation, with automatic estimation of the projection step β_k, quickly converges but produces solutions with slightly larger norms. The CKB$_2$ method leads to solutions with a worse norm, testifying that the performance of the method in (2.49) is very sensitive to the choice of the sequence γ_k. The MNGN method described in Section 2.3 leads to solutions far from optimality and fails in 70% of the tests. This happens in most of the examples considered in this section, so we will involve it only in another experiment.

Example 6.2.2. Now, we consider a test problem introduced in [20] and described in Test Function 6. It consists of the nonlinear function (5.8). We recall that the minimal-norm solution is

$$\mathbf{x}^\dagger \approx [0.859754, 1.849178, 3.065164]^T,$$

with $\|\mathbf{x}^\dagger\| \approx 3.681558$.

The results obtained are reported in Table 6.2. The MNGN2$_{\alpha\beta}$ method is tested with two values of the parameter η appearing in the residual increase $\delta(\rho, \eta) = \eta\rho$;

Table 6.2: Results for Example 6.2.2.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
$\text{MNGN2}_{\alpha\beta}$ $(\eta = 8)$	174	3.6903	15
$\text{MNGN2}_{\alpha\beta}$ $(\eta = 2)$	62	3.7120	100
MNGN2_{α}	330	3.6816	100
$\text{MNGN2}_{\alpha\beta\delta}$	37	3.6832	100
CKB_1	26	3.7343	100
CKB_2	10	3.7561	100

see Algorithm 1. It is clear that it can lead to accurate solutions only if the parameter is suitably chosen ($\eta = 2$). On the contrary ($\eta = 8$), it shows a great number of failures.

As in the previous example, the best results are produced by MNGN2_{α}, and $\text{MNGN2}_{\alpha\beta\delta}$ reaches very similar solutions but is about 10 times faster. The CKB methods take a smaller number of iterations, but produce less accurate solutions.

Example 6.2.3. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function introduced in Test Function 3, whose components are defined by (5.4). The locus of the solutions is the n-ellipsoid $S(\mathbf{x})$ with center $\mathbf{c} = [c_1, \ldots, c_n]^T$ and whose semiaxes are the components of the vector $\mathbf{a} = [a_1, \ldots, a_n]^T$.

We recall here that if $\mathbf{a} = \mathbf{e} = [1, \ldots, 1]^T$, the locus of the solutions is the n-sphere centered in $\mathbf{c}$ with unitary radius. Moreover, if $\mathbf{c} = [2, 0, \ldots, 0]^T$, the minimal-norm solution is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$.

Table 6.3: Results for Example 6.2.3 with $m = 8$, $n = 10$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = [2, 0, \ldots, 0]^T$. In MNGN, CKB_1, and CKB_2, the rank is not estimated.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
MNGN2_{α}	209	1.0263	83
$\text{MNGN2}_{\alpha\beta}$ $(\eta = 8)$	208	1.0449	99
$\text{MNGN2}_{\alpha\beta\delta}$	206	1.0367	97
MNGN	70	2.1083	2
CKB_1	216	2.2002	32
CKB_2	20	2.1305	2
rCKB_1	160	2.1088	32
rCKB_2	197	1.0454	97

Table 6.3 displays the results for this case, when $m = 8$ and $n = 10$. These results aim at underlining the importance of estimating the rank of the Jacobian J_k. The implementations of the MNGN2 algorithm are more or less equivalent, recovering solutions with almost optimal norm; MNGN2_{α} fails in 17% of the tests. The value of η for $\text{MNGN2}_{\alpha\beta}$ is tailored to maximize the performance, which is

not possible in practice, while it is automatically estimated for MNGN2$_{\alpha\beta\delta}$. The MNGN and CKB methods do not perform well, because of the rank deficiency of the Jacobian. We also implemented the rank estimation in the algorithms from [20]; the corresponding methods are denoted by rCKB. It happens that rCKB$_2$ produces results comparable to the MNGN2 methods, confirming that a correct estimation of the rank is essential for the convergence, while rCKB$_1$ converges only in 32% of the tests and produces solutions with large norms. Again, this shows that the sequence adopted for the step length in (r)CKB methods is critical for the effectiveness of the computation.

The norms of the solutions, whose average is displayed in Table 6.3, are reported in the boxplot in the left pane of Figure 6.8. In each box, the red mark is the median, the edges of the blue box are the 25th and 75th percentiles, and the black whiskers extend to the most extreme data points non considered to be outliers, which are plotted as red crosses.

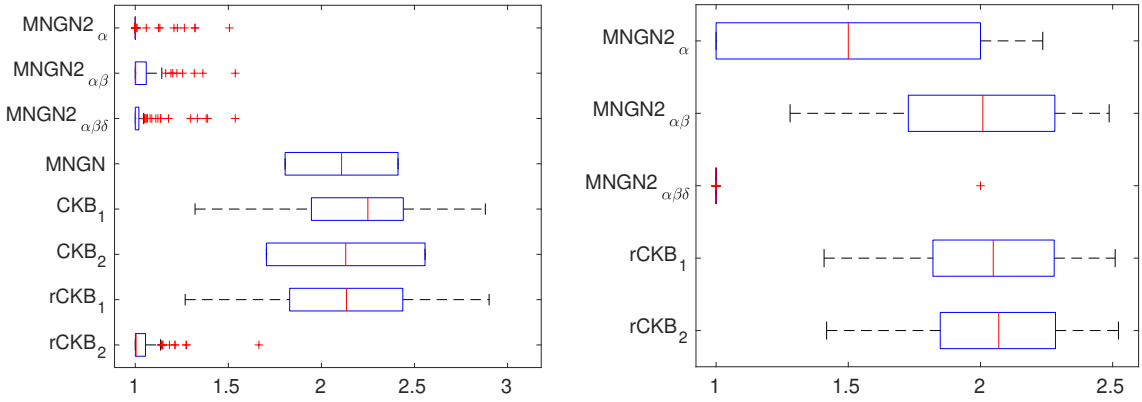

Figure 6.8: Boxplot of the norms of the solutions for Examples 6.2.3 (left) and 6.2.4 (right). The series, labeled by the methods name, are displayed in the same order of Table 6.3 and Table 6.4, respectively.

Example 6.2.4. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function introduced in Test Function 4, with components defined by (5.5). The locus of the solutions is the union of the n-ellipsoid and the intersection between the planes $x_i = c_i$, $i = 1, \ldots, m$.

As already said in Section 5, if $\mathbf{a} = \mathbf{e}$ and $\mathbf{c} = [2, 0, \ldots, 0]^T$, the minimal-norm solution is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$. The case $m = 2$, $n = 3$, is displayed in Figure 6.9, together with the iterations of the algorithms MNGN2$_{\alpha\beta\delta}$ and rCKB$_1$. In this test, the latter algorithm converges to a solution of non-minimal norm.

Table 6.4 illustrates the situation where $\mathbf{a} = \mathbf{e}$, $\mathbf{c} = [2, 0, \ldots, 0]^T$, $m = 8$ and $n = 10$. The corresponding boxplot of the norms of the solutions is displayed in the right pane of Figure 6.8. The MNGN2$_{\alpha\beta\delta}$ method is the only one which recovers

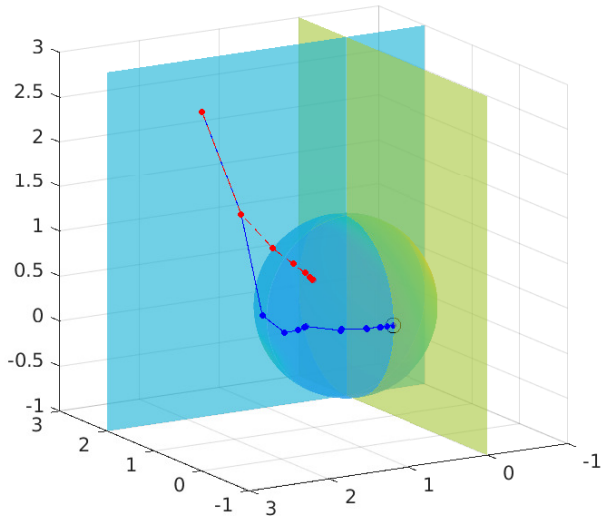

Figure 6.9: Solution of problem (5.5) (Example 6.2.4) for $m = 2$ and $n = 3$, with $\mathbf{a} = [1, 1, 1]^T$, $\mathbf{c} = [2, 0, 0]^T$, and $\mathbf{x}^{(0)} = [0, 3, 3]^T$. The locus of the solutions is the sphere and the line intersection of the two planes. The blue dots are the iterations of the MNGN2$_{\alpha\beta\delta}$ method, and the red ones correspond to the rCKB$_1$ method. The black circle encompasses the minimal-norm solution.

the correct solution; MNGN2$_\alpha$ gets close to it, but with a very small number of successes.

If $\mathbf{a} = \mathbf{e}$ and $\mathbf{c} = 2\mathbf{e}$, we recall that the minimal-norm solution $\mathbf{x}^\dagger$ depends on the dimensions m and n: if $m < n - \sqrt{n} + \frac{1}{4}$, then it is

$$\mathbf{x}^\dagger = [\underbrace{2, 2, \ldots, 2}_{m}, \underbrace{0, \ldots, 0}_{n-m}]^T,$$

otherwise, it is

$$\mathbf{x}^\dagger = \left(2 - \frac{\sqrt{n}}{n}\right) \mathbf{e}. \tag{6.4}$$

Table 6.5 reports the results obtained for $\mathbf{a} = \mathbf{e}$, $\mathbf{c} = 2\mathbf{e}$, $m = 8$, and $n = 10$. In this case, the solution is (6.4). We applied the algorithms to both the solution of the minimal-norm problem, and the computation of the minimal-L-norm solution with $L = D_2$, i.e., the discrete approximations of the second derivative (1.14). Since the solution is exactly in the null space of L, we expect the minimal-L-norm solution to perform well. No algorithm is accurate when $L = I_n$, as the minimal-norm is $2\sqrt{n} - 1 = 5.3246$. When $L = D_2$, the two MNGN2 implementations are superior to the rCKB methods, as $\|L\mathbf{x}^\dagger\| = 0$. As in the previous example, MNGN2$_\alpha$ exhibits a large number of failures.

Table 6.4: Results for Example 6.2.4 with $m = 8$, $n = 10$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = [2, 0, \ldots, 0]^T$.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
MNGN2$_{\alpha}$	215	1.5196	12
MNGN2$_{\alpha\beta}$ ($\eta = 8$)	11	1.9911	100
MNGN2$_{\alpha\beta\delta}$	47	1.0100	100
rCKB$_1$	27	2.0346	100
rCKB$_2$	11	2.0531	100

Table 6.5: Results for Example 6.2.4 with $m = 8$, $n = 10$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = 2\mathbf{e}$.

	method	iterations	$\|L\widetilde{\mathbf{x}}\|$	#success
$L = I_n$	MNGN2$_{\alpha}$	12	5.6569	23
	MNGN2$_{\alpha\beta\delta}$	45	5.4529	100
	rCKB$_1$	26	5.7274	100
	rCKB$_2$	11	5.7520	100
$L = D_2$	MNGN2$_{\alpha}$	20	0.0500	26
	MNGN2$_{\alpha\beta\delta}$	17	0.0765	100
	rCKB$_1$	27	2.1694	100
	rCKB$_2$	17	2.2761	100

Now, we consider the same function introduced in Test Function 4 in the large-scale case. We consider $\mathbf{a} = \mathbf{e}$, $\mathbf{c} = [2, 0, \ldots, 0]^T$, $m = 150$, and $n = 200$. We remember that for this function, the Golub–Kahan decomposition presents a breakdown after the first step; see Example 3.2.3. In Table 6.6 we can see that the GK-MNGN2$_{\alpha}$, GK-MNGN2$_{\alpha\beta}$, and GK-MNGN2$_{\alpha\beta\delta}$ approaches converge to the minimal-norm solution $\mathbf{x}^{\dagger} = [1, 0, \ldots, 0]^T$.

Table 6.6: Results for Example 6.2.4 with $m = 150$, $n = 200$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = [2, 0, \ldots, 0]^T$.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
GK-MNGN2$_{\alpha}$	17	1.0000	90
GK-MNGN2$_{\alpha\beta}$ ($\eta = 8$)	15	1.0000	100
GK-MNGN2$_{\alpha\beta\delta}$	64	1.0000	100
GK-CKB$_1$	27	2.4320	100
GK-CKB$_2$	14	2.7101	100

Example 6.2.5. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be the nonlinear function described in Test Function 5, with components defined by (5.6).

The locus of the solutions is the intersection between the hypersurface defined by $S(\mathbf{x}) = 0$ and by the pairs of planes $x_{i-1} = 0$, $x_i - c_i = 0$, $i = 2, \ldots, m$.

Table 6.7: Results for Example 6.2.5 with different size (m, n), $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = [2, 0, \ldots, 0]^T$.

(m, n)	method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
$(8, 10)$	MNGN2_{α}	167	1.0000	48
	$\text{MNGN2}_{\alpha\beta}$ $(\eta = 8)$	24	1.0508	100
	$\text{MNGN2}_{\alpha\beta\delta}$	37	1.0659	100
	rCKB_1	44	1.4867	100
	rCKB_2	22	1.4776	100
$(16, 20)$	MNGN2_{α}	144	1.0000	36
	$\text{MNGN2}_{\alpha\beta}$ $(\eta = 8)$	29	1.0170	99
	$\text{MNGN2}_{\alpha\beta\delta}$	34	1.0518	99
	rCKB_1	54	1.4343	100
	rCKB_2	53	1.5269	90
$(24, 30)$	MNGN2_{α}	133	1.0000	34
	$\text{MNGN2}_{\alpha\beta}$ $(\eta = 8)$	34	1.0154	99
	$\text{MNGN2}_{\alpha\beta\delta}$	32	1.0191	96
	rCKB_1	43	1.4446	100
	rCKB_2	52	1.4529	70

If $\mathbf{a} = \mathbf{e} = [1, \ldots, 1]^T$ and $\mathbf{c} = 2\mathbf{e}$, the minimal-norm solution is

$$\mathbf{x}^\dagger = [\xi_{n,m}, \underbrace{2, \ldots, 2}_{m-1}, \underbrace{\xi_{n,m}, \ldots, \xi_{n,m}}_{n-m}]^T, \tag{6.5}$$

with $\xi_{n,m} = 2 - (n - m + 1)^{-1/2}$, while if $\mathbf{c} = [2, 0, \ldots, 0]^T$ it is $\mathbf{x}^\dagger = [1, 0, \ldots, 0]^T$. This case is illustrated in Figure 6.10, where the iterations of the $\text{MNGN2}_{\alpha\beta\delta}$ and the rCKB_1 methods are reported too. The iterations performed are 20 and 24, respectively; the computed solutions are substantially coincident.

Table 6.7 displays the results obtained for the same parameter vectors of Figure 6.10, when the size of the problem varies, i.e., for $(m, n) = (8k, 10k)$, $k = 1, 2, 3$. The MNGN2 algorithms behave almost optimally, while the rCKB methods lead to solutions with larger norm. The table shows that the performance is not significantly affected by the size of the problem. This example suggests that large-scale problems could be faced by the methods discussed, but a suitable algorithm for the solution of the linearized problem should be adopted, to reduce the computational complexity of each step.

Table 6.8 investigates the effectiveness of choosing an appropriate model profile $\overline{\mathbf{x}}$ when applying the MNGN2 algorithms. We consider the case $\mathbf{a} = \mathbf{e}$, $\mathbf{c} = 2\mathbf{e}$, $m = 8$, and $n = 10$. The minimal-norm solution $\mathbf{x}^\dagger$ is (6.5), with $\xi_{8,10} \simeq 1.4226$ and $\|\mathbf{x}^\dagger\| \simeq 5.8371$.

When $\overline{\mathbf{x}} = \mathbf{0}$, the solutions produced by the considered variants of the method are almost optimal, but the number of iterations is quite large, as well as the number

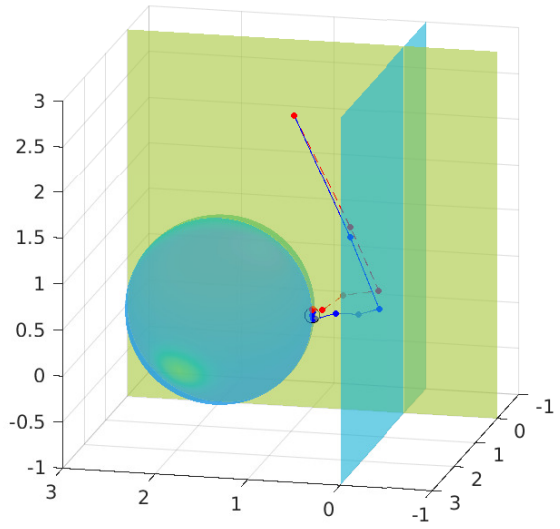

Figure 6.10: Solution of problem (5.6) (Example 6.2.5) for $m = 2$ and $n = 3$, with $\mathbf{a} = [1, 1, 1]^T$, $\mathbf{c} = [2, 0, 0]^T$, and $\mathbf{x}^{(0)} = [\frac{1}{2}, 3, 3]^T$. The solutions are in the intersection between the sphere and the union of the two planes. The blue dots are the iterations of the MNGN2$_{\alpha\beta\delta}$ method, and the red ones correspond to the rCKB$_1$ method. The black circle encompasses the minimal-norm solution.

of failures for MNGN2$_{\alpha\beta}$ (with a suitably chosen η) and MNGN2$_{\alpha\beta\delta}$. The model profile $\overline{\mathbf{x}} = 2\mathbf{e}$ reduces the number of iterations and leads to almost 100% of successes, but the average norm of the solutions is slightly larger than the optimal one. Choosing $\overline{\mathbf{x}} = 1.7\mathbf{e}$, a value which is roughly halfway between 2 and $\xi_{8,10}$, the extreme values of $\mathbf{x}^\dagger$, restores the optimality of the results. This confirms that, when a priori information is available, an accurate choice of the model profile enhances the performance of the algorithms.

Now, we apply the MNGN2 and CKB methods in the large-scale setting. We consider the same function described in Test Function 5 in the case $\mathbf{a} = \mathbf{e}$, $\mathbf{c} = [2, 0, \ldots, 0]^T$, $m = 50$, and $n = 70$. The results are reported in Table 6.9. The GK-MNGN2 algorithms behave almost optimally, while the GK-CKB methods lead to solutions with larger norm.

Table 6.8: Results for Example 6.2.5 with $m = 8$, $n = 10$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = 2\mathbf{e}$.

	method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
$\overline{\mathbf{x}} = \mathbf{0}$	MNGN2$_{\alpha}$	138	5.8371	100
	MNGN2$_{\alpha\beta}$ ($\eta = 8$)	175	5.8374	38
	MNGN2$_{\alpha\beta\delta}$	94	5.8988	67
$\overline{\mathbf{x}} = 2\mathbf{e}$	MNGN2$_{\alpha}$	37	6.1141	99
	MNGN2$_{\alpha\beta}$ ($\eta = 8$)	34	6.1144	98
	MNGN2$_{\alpha\beta\delta}$	34	6.1144	98
$\overline{\mathbf{x}} = 1.7\mathbf{e}$	MNGN2$_{\alpha}$	54	5.8371	100
	MNGN2$_{\alpha\beta}$ ($\eta = 8$)	34	5.8394	99
	MNGN2$_{\alpha\beta\delta}$	40	5.8789	99

Table 6.9: Results for Example 6.2.5 with $m = 50$, $n = 70$, $\mathbf{a} = \mathbf{e}$, and $\mathbf{c} = [2, 0, \ldots, 0]^T$.

method	iterations	$\|\widetilde{\mathbf{x}}\|$	#success
GK-MNGN2$_{\alpha}$	212	1.0049	17
GK-MNGN2$_{\alpha\beta}$ ($\eta = 8$)	93	1.0382	93
GK-MNGN2$_{\alpha\beta\delta}$	96	1.0838	87
GK-CKB$_1$	69	1.5190	100
GK-CKB$_2$	95	1.5166	37

6.3 Reproducing Kernel and Riesz representers at work

In this section, we report some numerical results obtained by applying our algorithm explained in Chapter 4 and in [32].

The following two numerical experiments are based on systems of two linear integral equations. We consider the exact right-hand side $\mathbf{g}_{\text{exact}}$ of the linear system (4.23), corresponding to the collocation nodes $x_{\ell,i}$, for $\ell = 1, \ldots, m$ and $i = 1, \ldots, n_\ell$. We add Gaussian noise as in (4.35), where the noise vector $\mathbf{e}$ is defined by

$$\mathbf{e} = \frac{\delta}{\sqrt{N_m}} \|\mathbf{g}_{\text{exact}}\| \mathbf{w}, \tag{6.6}$$

with N_m as in (4.16). The components of the vector $\mathbf{w}$ are normally distributed with zero average and unit variance, and δ represents the noise level. For the sake of simplicity, for each system we consider the same collocation nodes $x_{\ell,i}$ in both equations, so that $m = 2$, $n_1 = n_2 = n$, $N_m = 2n$, and $x_{1,i} = x_{2,i}$, for $i = 1, \ldots, n$.

Example 6.3.1. We consider the system (5.10) described in Test Function 8. It consists of two Fredholm integral equations of the first kind, with $x \in (0, 1]$ and

exact solution $f(t) = t^2 + 1$. In this example we set $x_{\ell,i} = 0.1 + 0.9\,(i-1)/(n-1)$, for $\ell = 1, 2$ and $i = 1, \ldots, n$.

The corresponding Riesz representers have been computed analytically in (5.13) and (5.14). Note that the analytic expression of $\eta''_{\ell,i}$ defined in (5.11) and (5.12) allows for an accurate computation of the elements of the Gram matrix (4.24) and for obtaining an explicit representation of the functions $\eta_{\ell,i}$, providing a fast and accurate algorithm.

We remind the reader that, by (4.7), the solution of this problem is expressed as

$$f(t) = \xi(t) + \gamma(t),$$

where $\gamma(t) = t + 1$ is the function (4.5) and $\xi(t)$ is the solution of the system (4.6).

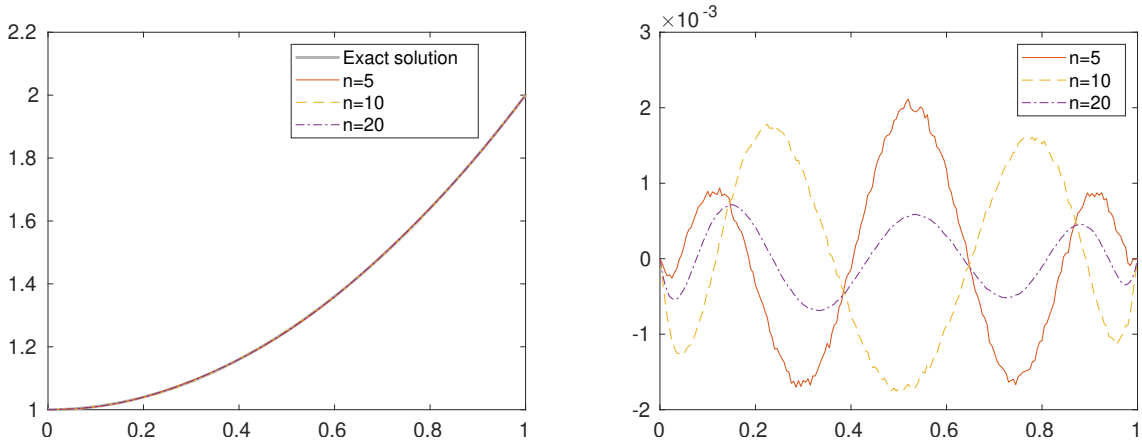

Figure 6.11: Example 6.3.1: non-regularized reconstructions of the solution of Test Function 8 (left) and corresponding errors (right), for $n = 5, 10, 20$, and without noise.

To start with, we depict in Figure 6.11 the non-regularized reconstructions of the solution, obtained for $n = 5, 10, 20$, without noise in the data, and the corresponding error curves with respect to the exact solution. By "non-regularized", we mean that we set $\kappa = N$ in (4.33) and (4.34). The fact that the errors are so small that the solutions graphically coincide is remarkable. Indeed, setting $\delta = 0$ in (6.6) only guarantees that the right-hand side is accurate up to machine precision, that is, roughly 10^{-16}. Since the estimation of the condition number of the Gram matrix $\mathcal{G}$ provided by the `cond` function of Matlab for the three problem sizes considered is $2.2 \cdot 10^{18}$, $6.9 \cdot 10^{32}$, and $1.1 \cdot 10^{19}$, respectively, the results highlight the extreme stability in the computation, as well as the effectiveness of the function space setting.

Figure 6.12 shows, in the left pane, the reconstructions obtained without regularization for $n = 5, 10, 20$, together with the exact solution, when the data vector is affected by noise with level $\delta = 10^{-4}$. Due to the large condition number, the computed solutions are polluted by noise propagation at such a point that they swing at high frequency away from the exact solution. The graph on the right of the

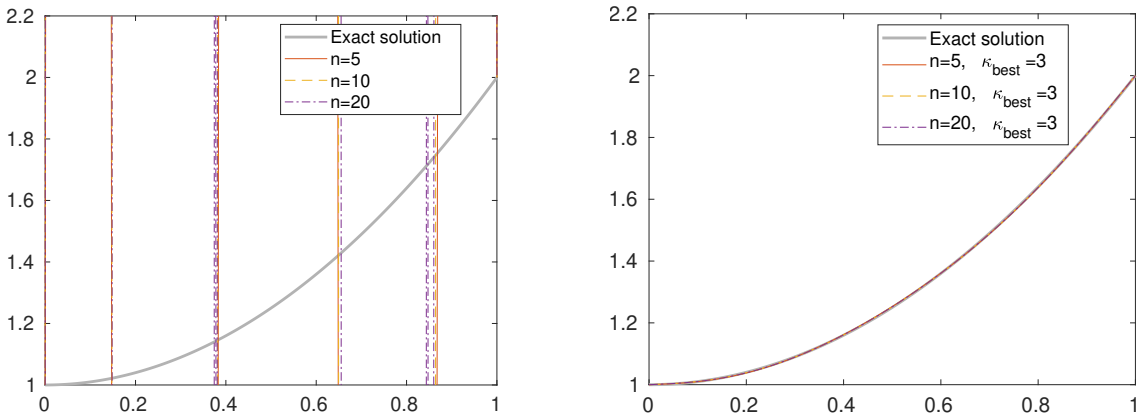

Figure 6.12: Example 6.3.1. On the left: non-regularized solutions of Test Function 8, for $n = 5, 10, 20$, and noise level $\delta = 10^{-4}$. On the right: regularized solution $f^{(\kappa_{\text{best}})}(t)$, for $n = 5, 10, 20$, and $\delta = 10^{-4}$; the optimal value κ_{best} of the regularization parameter is displayed in the legend.

same figure displays the results obtained by computing the regularized solution $f^{(\kappa)}$ defined in (4.33). Here, the truncation parameter κ coincides with the value κ_{best}, defined in (4.39), corresponding to the best possible performance of the algorithm. The quality of the results is excellent.

The graph on the left of Figure 6.13 investigates the sensitivity of the solution on the noise level. It shows the errors obtained for $n = 10$ and $\delta = 10^{-8}, 10^{-4}, 10^{-2}$. The graph confirms the accuracy and stability of the proposed regularization method. In the graph on the right, we compare the "best" solution for the noise level $\delta = 10^{-4}$ to the ones obtained by estimating the regularization parameter by the discrepancy principle (4.36) (κ_{d}), with $\tau = 1.1$, and by the L-curve criterion (4.38) (κ_{lc}). Both estimation techniques are successful.

Example 6.3.2. Let us now consider the system (5.15) introduced in Test Function 9, with $x \in (0, \pi/2]$. It pairs the well-known Baart test problem [6, 58] to an equation having the same solution $f(t) = \sin t$. The collocation points are $x_{\ell,i} = 0.1 + (\pi/2 - 0.1)\,(i-1)/(n-1)$, for $\ell = 1, 2$ and $i = 1, \ldots, n$.

In this example, we were only able to analytically compute the Riesz representers for the second equation; see (5.16) and (5.17). An approximation of the Riesz representers for the first equation was computed by a Gauss–Legendre quadrature formula; see (5.18) and (5.19).

Figure 6.14 shows that, when the data vector is only affected by rounding errors, the non-regularized solution is very accurate. On the contrary, as in the previous example, the non-regularized solution is strongly unstable when a sensible amount of noise is added to the data; we do not display the results for the sake of brevity.

The graph on the left of Figure 6.15 depicts the behavior of the best regularized solutions corresponding to the three noise levels $\delta = 10^{-8}, 10^{-4}, 10^{-2}$. All the

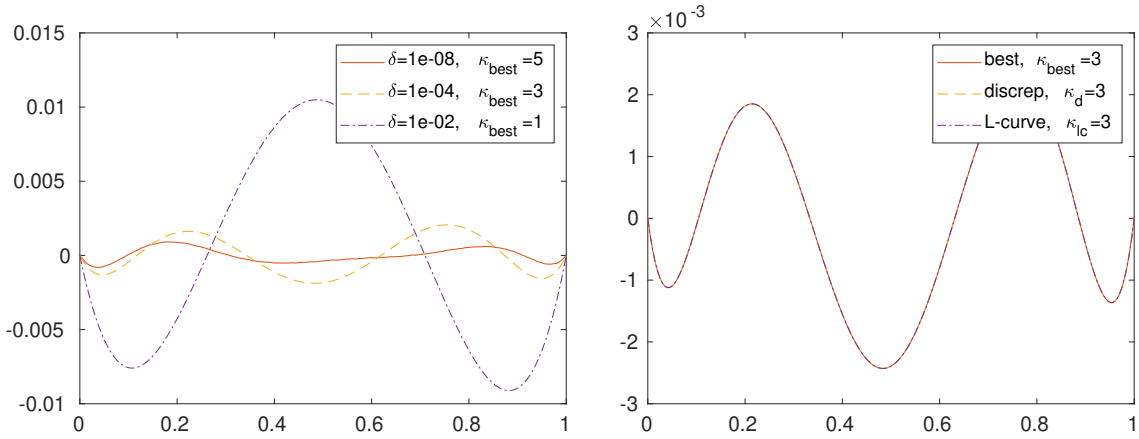

Figure 6.13: Example 6.3.1. On the left: errors corresponding to the regularized solutions $f^{(\kappa_{\text{best}})}(t)$ of Test Function 8, for $n = 10$ and $\delta = 10^{-8}, 10^{-4}, 10^{-2}$. On the right: errors for the solutions $f^{(\kappa)}(t)$, for $n = 20$, $\delta = 10^{-4}$, and different estimation methods for κ. The values of the regularization parameters κ_{best}, κ_{d}, and κ_{lc} are displayed in the legend.

reconstructions are accurate. In the second graph, we compare the reconstruction corresponding by the optimal regularization parameter to the ones produced by the discrepancy principle and the L-curve. Even if the estimated values of the parameter are slightly different, the results are satisfactory. We verified that the results are not sensibly influenced by the size of the problem.

In the next example, we apply the same algorithm to a linear model involved in applied geophysics.

Example 6.3.3. Here we consider a linear model applied to investigate soil properties, described in Test Function 10 of Chapter 5. The aim is to determine the electrical conductivity σ of the subsoil.

In order to ascertain the accuracy of our method, we consider three different profiles for the electrical conductivity $\sigma(z)$. Then, for each test function, we compute the data vector $\boldsymbol{\psi}_{\text{exact}}$, setting $z_0 = 4$ and $h_i = 0.1 + 0.9\,(i-1)/(n-1)$, $i = 1, \ldots, n$, for a chosen dimension n. The computation of the exact data vector is performed by the `quadgk` function of Matlab, which implements an adaptive Gauss-Kronrod quadrature formula.

In the case of experimental data, the available data is typically contaminated by errors. To simulate this situation, the perturbed data vector $\boldsymbol{\psi}$ is determined by adding to $\boldsymbol{\psi}_{\text{exact}}$ a noise-vector $\mathbf{e}$, obtained by substituting in (6.6) $\boldsymbol{\psi}_{\text{exact}}$ to $\mathbf{g}_{\text{exact}}$ and setting $N_m = 2n$. The noise level is determined by the parameter δ.

In the first example, we assume a smooth profile for the exact solution of (5.20)

$$\sigma_1(z) = \mathrm{e}^{-(z-1)^2} + 1.$$

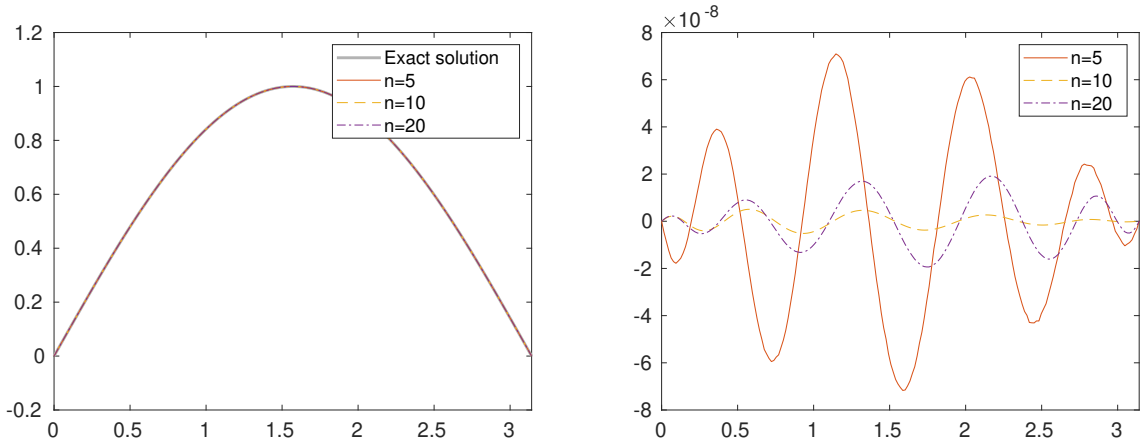

Figure 6.14: Example 6.3.2: non-regularized reconstructions of the solution of Test Function 9 (left) and corresponding errors (right), for $n = 5, 10, 20$ and without noise.

We set $\alpha = \sigma_1(0) = \mathrm{e}^{-1} + 1$ and $\beta = \sigma_1(z_0) = \mathrm{e}^{-9} + 1$. We remark that this test function is extremely smooth, so the function $\phi_1(z) = \sigma_1(z) - \gamma_1(z)$ can be assumed to approximately belong to $\mathcal{N}(\mathbf{K})^\perp = \mathrm{span}\{\eta_1, \ldots, \eta_{N_m}\}$, the space which contains the minimal-norm solution.

Figure 6.16 displays the results obtained by applying the method described in this chapter to the electromagnetic integral model (5.20) with the optimal regularization parameter. On the left-hand side, we report the approximation of the solution for different noise levels $\delta = 10^{-8}, 10^{-4}, 10^{-2}$, and $n = 10$; on the right-hand side, the results for $n = 5, 10, 20$ and $\delta = 10^{-2}$ are depicted. All the reconstructions are accurate and identify with sufficient accuracy the maximum value of the conductivity and its depth localization. The graph on the left shows that, even for an increasing noise level, the method is still able to produce reliable results. On the other hand, from the graph on the right we deduce that both the reconstructions and the optimal value of the regularization parameter are not very sensitive on the size of the data vector.

In order to test the method in realistic conditions, in Figure 6.17 we compare the optimal solution to the approximate solutions corresponding to the parameter estimated by the discrepancy principle κ_{d}, with $\tau = 1.3$, and by the L-curve criterion κ_{lc}. In this case, we have fixed $n = 10$ and a noise level $\delta = 10^{-4}$. Both estimation techniques appear to be effective.

In the second experiment, we select the following model function

$$\sigma_2(z) = \begin{cases} 0.8z + 0.2, & z \in [0, 1], \\ 0.8\mathrm{e}^{-(z-1)} + 0.2, & z \in (1, \infty), \end{cases}$$

and set $\alpha = 0.2$ and $\beta = 0.2 + 0.8\mathrm{e}^{-3}$.

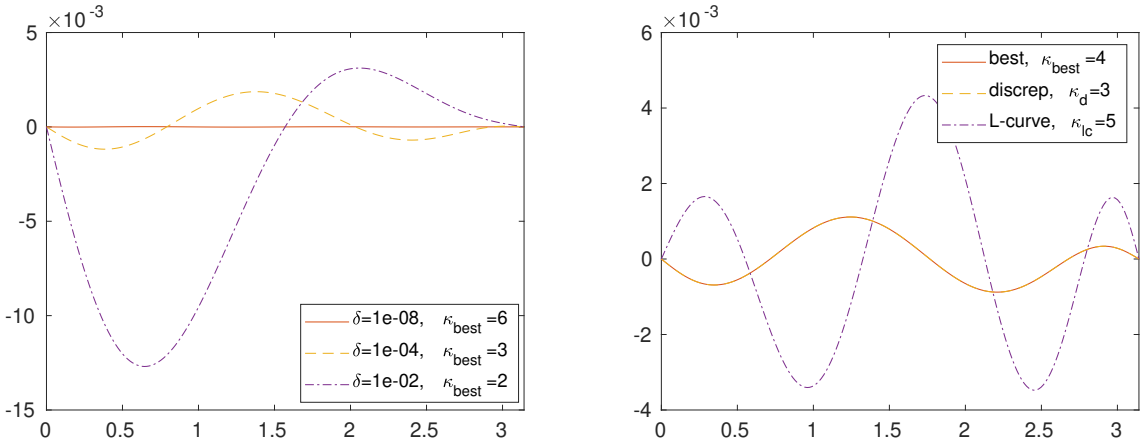

Figure 6.15: Example 6.3.2. On the left: errors corresponding to the regularized solutions $f^{(\kappa_{\text{best}})}(t)$ of Test Function 9, for $n = 10$ and $\delta = 10^{-8}, 10^{-4}, 10^{-2}$. On the right: errors for the solutions $f^{(\kappa)}(t)$, for $n = 20$, $\delta = 10^{-4}$, and different estimation methods for κ. The values of the regularization parameters κ_{best}, κ_{d}, and κ_{lc} are displayed in the legend.

The graph in the left pane of Figure 6.18 reports the optimal regularized solutions corresponding to the noise levels $\delta = 10^{-8}$, 10^{-4}, 10^{-2}, and $n = 10$. The optimal parameter is displayed in the legend. The reconstruction is not accurate as in the previous test, because the solution is non-differentiable and, consequently, it does not belong to $\mathcal{N}(\mathbf{K})^{\perp}$. Anyway, the algorithm correctly identifies the position of the maximum of the electrical conductivity at 1m depth.

The third model function is the step function

$$\sigma_3(z) = \begin{cases} 0.2, & z \in (0, 0.5), \\ 2, & z \in [0.5, 1.5], \\ 0.2, & z \in (1.5, \infty), \end{cases}$$

with $\alpha = \beta = 0.2$.

The graph on the right-hand side of Figure 6.18 reports the optimal regularized solutions for $\delta = 10^{-8}$, 10^{-4}, 10^{-2}, and $n = 10$. Since the function is discontinuous, we do not expect an accurate reconstruction and comments similar to the previous example are valid.

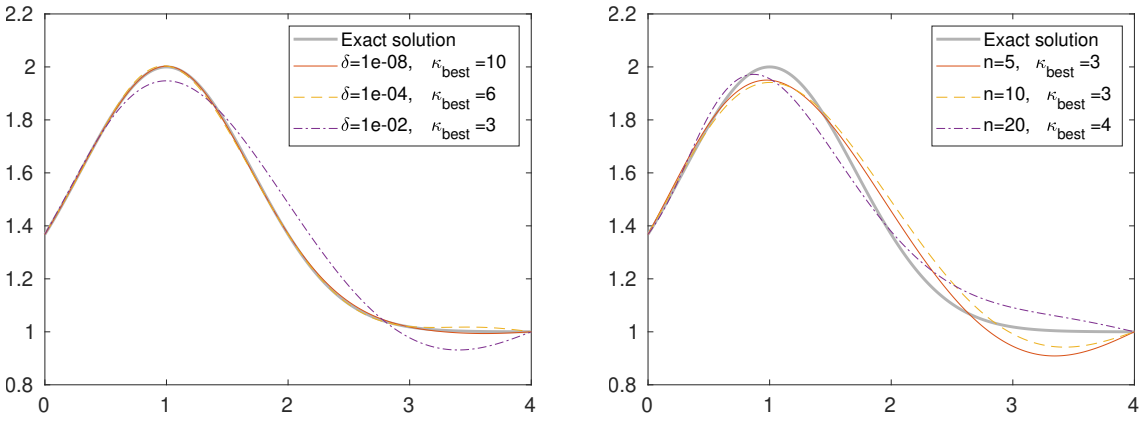

Figure 6.16: Example 6.3.3. On the left: regularized solution $\sigma_1^{(\kappa_{\text{best}})}(z)$ for noise levels $\delta = 10^{-8}, 10^{-4}, 10^{-2}$, and $n = 10$. On the right: regularized solution $\sigma_1^{(\kappa_{\text{best}})}(z)$ for $n = 5, 10, 20$, and noise level $\delta = 10^{-2}$; the optimal regularization parameter κ_{best} is displayed in the legend.

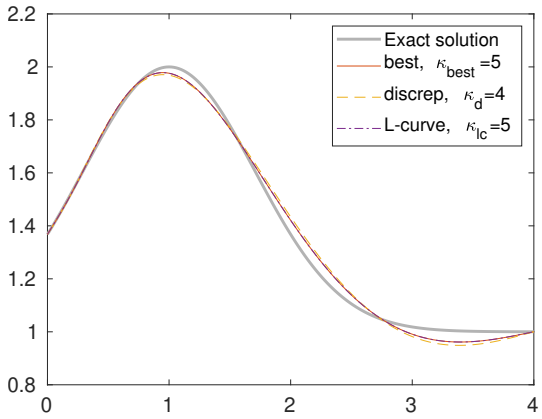

Figure 6.17: Example 6.3.3. Regularized solution $\sigma_1^{(\kappa)}(z)$ with $n = 10$ and $\delta = 10^{-4}$; the optimal regularization parameter κ_{best} is compared to those determined by the discrepancy principle κ_{d} and by the L-curve κ_{lc}.

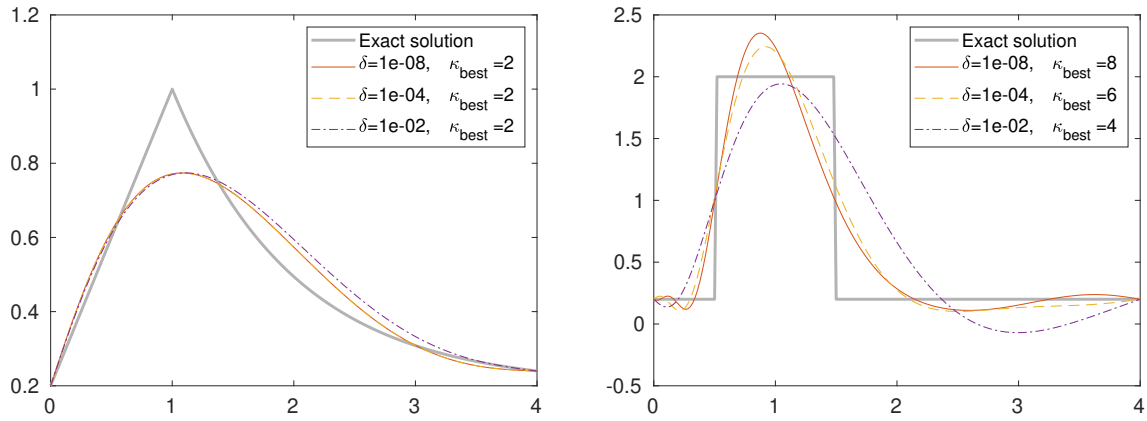

Figure 6.18: Example 6.3.3. Regularized solution $\sigma_2^{(\kappa_{\text{best}})}(z)$ (left) and $\sigma_3^{(\kappa_{\text{best}})}(z)$ (right), for $n = 10$ and different noise levels $\delta = 10^{-8}, 10^{-4}, 10^{-2}$; the optimal regularization parameter κ_{best} is displayed in the legend.

Conclusions and future work

In this thesis, we proposed different methods to compute the minimal-norm solution of nonlinear least-squares problems (Chapter 2) and of systems of linear integral equations (Chapter 4). We dedicated Chapter 5 to the description of test problems, both artificial and deriving from engineering applications. In Chapter 6 we applied various methods to verify their effectiveness and performance.

Regarding the computation of the minimal-norm solution of nonlinear least-squares problems, the large-scale case (Chapter 3) involves a work in progress. Moreover, we are planning to apply the MNGN method and its variants (MNGN2 and regularized approaches) to imaging science. In this case, the approximate solution will be a 2D reconstruction.

About solving systems of integral equations of the first kind using the tools of functional analysis such as the Riesz theory, future work will concern the computation of the solution of nonlinear integral equations.

Bibliography

[1] A. Alqahtani, S. Gazzola, L. Reichel, and G. Rodriguez, *On the block Lanczos and block Golub-Kahan reduction methods applied to discrete ill-posed problems*, Numer. Linear Algebra Appl., 2021 (2021), p. e2376.

[2] J. Angeles, *Fundamentals of Robotic Mechanical Systems: Theory, Methods, and Algorithms*, Springer Science & Business Media, fourth ed., 2014.

[3] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pac. J. Math., 16 (1966), pp. 1–3.

[4] N. Aronszajn, *Theory of reproducing kernels*, Trans. Am. Math. Soc., 68 (1950), pp. 337–404.

[5] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, vol. 552, Cambridge University Press, Cambridge, 1997.

[6] M. L. Baart, *The use of auto-correlation for pseudo-rank determination in noisy ill-conditioned linear least-squares problems*, IMA J. Numer. Anal., 2 (1982), pp. 241–247.

[7] J. Baglama and L. Reichel, *Augmented implicitly restarted Lanczos bidiagonalization methods*, SIAM J. Sci. Comput., 27 (2005), pp. 19–42.

[8] ——, *Restarted block Lanczos bidiagonalization methods*, Numer. Algorithms, 43 (2006), pp. 251–272.

[9] ——, *An implicitly restarted block Lanczos bidiagonalization method using Leja shifts*, BIT, 53 (2013), pp. 285–310.

[10] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[11] J. Boaga, M. Ghinassi, A. D'Alpaos, G. P. Deidda, G. Rodriguez, and G. Cassiani, *Geophysical investigations unravel the vestiges of ancient*

meandering channels and their dynamics in tidal landscapes, Sci. Rep., 8 (2018), p. 1708 (8 pages).

[12] B. Borchers, T. Uram, and J. M. H. Hendrickx, *Tikhonov regularization of electrical conductivity depth profiles in field soils*, Soil Sci. Soc. Am. J., 61 (1997), pp. 1004–1009.

[13] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer Science & Business Media, New York, 2011.

[14] A. Buccini, G. Huang, L. Reichel, and F. Yin, *On the choice of regularization matrix for an ℓ_2-ℓ_q minimization method for image restoration*, Appl. Numer. Math., 164 (2021), pp. 211–221.

[15] A. Buccini, M. Pasha, and L. Reichel, *Modulus-based iterative methods for constrained ℓ_p-ℓ_q minimization*, Inverse Probl., 36 (2020), p. 084001.

[16] A. Buccini and L. Reichel, *An ℓ_2-ℓ_q regularization method for large discrete ill-posed problems*, J. Sci. Comput., 78 (2019), pp. 1526–1549.

[17] ——, *An ℓ_p-ℓ_q minimization method with cross-validation for the restoration of impulse noise contaminated images*, J. Comput. Appl. Math., 375 (2020), p. 112824.

[18] ——, *Generalized cross validation for ℓ_p-ℓ_q minimization*, Numer. Algorithms, (2021), pp. 1–22.

[19] D. Calvetti, B. Lewis, and L. Reichel, *A hybrid GMRES and TV-norm-based method for image restoration*, in Advanced Signal Processing Algorithms, Architectures, and Implementations XII, vol. 4791, International Society for Optics and Photonics, 2002, pp. 192–200.

[20] S. L. Campbell, P. Kunkel, and K. Bobinyec, *A minimal norm corrected underdetermined Gauß–Newton procedure*, Appl. Numer. Math., 62 (2012), pp. 592–605.

[21] J. Chung and J. G. Nagy, *An efficient iterative approach for large-scale separable nonlinear inverse problems*, SIAM J. Sci. Comput., 31 (2010), pp. 4654–4674.

[22] C. Clason and V. H. Nhu, *Bouligand-Levenberg-Marquardt iteration for a non-smooth ill-posed inverse problem*, Electron. Trans. Numer. Anal., 51 (2019), pp. 274–314.

[23] A. Concas, S. Noschese, L. Reichel, and G. Rodriguez, *A spectral method for bipartizing a network and detecting a large anti-community*, J. Comput. Appl. Math., 373 (2020), p. 112306.

[24] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, San Diego, second ed., 1984.

[25] G. P. Deidda, E. Bonomi, and C. Manzi, *Inversion of electrical conductivity data with Tikhonov regularization approach: some considerations*, Ann. Geophys., 46 (2003), pp. 549–558.

[26] G. P. Deidda, P. Díaz de Alba, C. Fenu, G. Lovicu, and G. Rodriguez, *FDEMtools: a MATLAB package for FDEM data inversion*, Numer. Algorithms, 84 (2020), pp. 1313–1327.

[27] G. P. Deidda, P. Díaz de Alba, and G. Rodriguez, *Identifying the magnetic permeability in multi-frequency EM data inversion*, Electron. Trans. Numer. Anal., 47 (2017), pp. 1–17.

[28] G. P. Deidda, P. Díaz de Alba, G. Rodriguez, and G. Vignoli, *Inversion of multiconfiguration complex EMI data with minimum gradient support regularization: a case study*, Math. Geosci., 52 (2020), pp. 945–970.

[29] G. P. Deidda, C. Fenu, and G. Rodriguez, *Regularized solution of a nonlinear problem in electromagnetic sounding*, Inverse Probl., 30 (2014), p. 125014 (27 pages).

[30] J. E. Dennis Jr. and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, SIAM, 1996.

[31] P. Díaz de Alba, L. Fermo, F. Pes, and G. Rodriguez, *Minimal-norm RKHS solution of an integral model in geo-electromagnetism*, in Proceedings of the International Conference on Computational Science and its Applications (ICCSA), Cagliari, Italy, Sept. 2021. To appear.

[32] P. Díaz de Alba, L. Fermo, F. Pes, and G. Rodriguez, *Regularized minimal-norm solution of an overdetermined system of first kind integral equations*. Submitted, 2021.

[33] P. Díaz de Alba, L. Fermo, C. van der Mee, and G. Rodriguez, *Recovering the electrical conductivity of the soil via a linear integral model*, J. Comput. Appl. Math., 352 (2019), pp. 132–145.

[34] G. Dragonetti, A. Comegna, A. Ajeel, G. P. Deidda, N. Lamaddalena, G. Rodriguez, G. Vignoli, and A. Coppola, *Calibrating electromagnetic induction conductivities with time-domain reflectometry measurements*, Hydrol. Earth. Syst. Sc., 22 (2018), pp. 1509–1523.

[35] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[36] J. Eriksson, *Optimization and regularization of nonlinear least squares problems*. Ph.D. Thesis, Umeå University, Sweden, 1996.

[37] J. Eriksson and P. Wedin, *Regularization methods for nonlinear least squares problems. part i: Exactly rank-deficient problems*, tech. rep., Umeå University, Sweden, 1996.

[38] J. Eriksson, P. Wedin, M. Gulliksson, and I. Söderkvist, *Regularization methods for uniformly rank-deficient nonlinear least-squares problems*, J. Optim. Theory Appl., 127 (2005), pp. 1–26.

[39] L. Eskola, *Geophysical Interpretation Using Integral Equations*, Springer Science & Business Media, 2012.

[40] D. C.-L. Fong and M. A. Saunders, *LSMR: An iterative algorithm for sparse least-squares problems*, SIAM J. Sci. Comput., 33 (2011), pp. 2950–2971.

[41] S. Gazzola, E. Onunwor, L. Reichel, and G. Rodriguez, *On the Lanczos and Golub–Kahan reduction methods applied to discrete ill-posed problems*, Numer. Linear Algebra Appl., 23 (2016), pp. 187–204.

[42] A. A. Goldstein, *Constructive Real Analysis*, Harper and Row, 1967.

[43] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2 (1965), pp. 205–224.

[44] G. H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[45] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, fourth ed., 2013.

[46] T. Goodman, C. Micchelli, G. Rodriguez, and S. Seatzu, *On the Cholesky factorization of the Gram matrix of locally supported functions*, BIT, 35 (1995), pp. 233–257.

[47] ——, *Spectral factorization of Laurent polynomials*, Adv. Comput. Math., 7 (1997), pp. 429–454.

[48] ——, *On the limiting profile arising from orthonormalizing shifts of exponentially decaying functions*, IMA J. Numer. Anal., 18 (1998), pp. 331–354.

[49] ——, *On the Cholesky factorisation of the Gram matrix of multivariate functions*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 501–526.

[50] C. W. Groetsch, *Elements of Applicable Functional Analysis*, vol. 55 of Monographs and Textbooks in Pure and Applied Mathematics, Dekker, New York and Basel, 1980.

[51] ——, *Integral equations of the first kind, inverse problems and regularization: A crash course*, in Journal of Physics: Conference Series, vol. 73, 2007, p. 012001.

[52] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.

[53] M. Hanke, *A regularizing Levenberg–Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Probl., 13 (1997), p. 79.

[54] M. Hanke, J. G. Nagy, and C. Vogel, *Quasi–Newton approach to nonnegative image restorations*, Linear Alg. Appl., 316 (2000), pp. 223–236.

[55] P. C. Hansen, *The truncated SVD as a method for regularization*, BIT, 27 (1987), pp. 543–553.

[56] ——, *Analysis of the discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.

[57] ——, *Rank–Deficient and Discrete Ill–Posed Problems*, SIAM, Philadelphia, 1998.

[58] ——, *Regularization Tools: version 4.0 for Matlab 7.3*, Numer. Algorithms, 46 (2007), pp. 189–194.

[59] P. C. Hansen, T. K. Jensen, and G. Rodriguez, *An adaptive pruning algorithm for the discrete L-curve criterion*, J. Comput. Appl. Math., 198 (2007), pp. 483–492.

[60] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill–posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.

[61] P. C. Hansen, V. Pereyra, and G. Scherer, *Least Squares Data Fitting with Applications*, Johns Hopkins University Press, 2012.

[62] J. M. H. Hendrickx, B. Borchers, D. L. Corwin, S. M. Lesch, A. C. Hilgendorf, and J. Schlue, *Inversion of soil conductivity profiles from electromagnetic induction measurements*, Soil Sci. Soc. Am. J., 66 (2002), pp. 673–685.

[63] E. Hille, *Introduction to general theory of reproducing kernels*, Rocky Mt. J. Math., 2 (1972), pp. 321–368.

[64] M. Hochbruck and M. Hönig, *On the convergence of a regularizing Levenberg–Marquardt scheme for nonlinear ill-posed problems*, Numer. Math., 115 (2010), pp. 71–79.

[65] M. E. Hochstenbach, *Harmonic and refined extraction methods for the singular value problem, with applications in least squares problems*, BIT, 44 (2004), pp. 721–754.

[66] M. E. Hochstenbach, L. Reichel, and G. Rodriguez, *Regularization parameter determination for discrete ill-posed problems*, J. Comput. Appl. Math., 273 (2015), pp. 132–149.

[67] G. Huang, A. Lanza, S. Morigi, L. Reichel, and F. Sgallari, *Majorization–minimization generalized krylov subspace methods for ℓ_p-ℓ_q optimization applied to image restoration*, BIT, 57 (2017), pp. 351–378.

[68] Z. Jia, *Some properties of LSQR for large sparse linear least squares problems*, J. Syst. Sci. Complex., 23 (2010), pp. 815–821.

[69] Q. Jin, *On a class of frozen regularized Gauss–Newton methods for nonlinear inverse problems*, Math. Comp., 79 (2010), pp. 2191–2211.

[70] M. E. Kilmer, P. C. Hansen, and M. I. Espanol, *A projection-based approach to general-form Tikhonov regularization*, SIAM J. Sci. Comput., 29 (2007), pp. 315–330.

[71] R. Kress, *Numerical Analysis*, Springer, Berlin, 1998.

[72] ——, *Linear Integral Equation*, Springer, Berlin, 1999.

[73] A. Lanza, S. Morigi, L. Reichel, and F. Sgallari, *A generalized krylov subspace method for ℓ_p-ℓ_q minimization*, SIAM J. Sci. Comput., 37 (2015), pp. S30–S50.

[74] S. Lu, S. V. Pereverzev, and R. Ramlau, *An analysis of Tikhonov regularization for nonlinear ill-posed problems under a general smoothness assumption*, Inverse Probl., 23 (2007), pp. 217–230.

[75] P. Mahale and M. Nair, *A simplified generalized Gauss–Newton method for nonlinear ill-posed problems*, Math. Comp., 78 (2009), pp. 171–184.

[76] J. D. McNeill, *Electromagnetic terrain conductivity measurement at low induction numbers*, Tech. Rep. TN-6, Geonics Limited, Mississauga, Ontario, Canada, 1980.

[77] V. A. Morozov, *The choice of parameter when solving functional equations by regularization*, Dokl. Akad. Nauk. SSSR, 175 (1962), pp. 1225–1228.

[78] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[79] C. C. Paige and M. A. Saunders, *Algorithm 583: LSQR: Sparse linear equations and least squares problems*, ACM Trans. Math. Softw., 8 (1982), pp. 195–209.

[80] ——, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Softw., 8 (1982), pp. 43–71.

[81] Y. Park, L. Reichel, G. Rodriguez, and X. Yu, *Parameter determination for Tikhonov regularization problems in general form*, J. Comput. Appl. Math., 343 (2018), pp. 12–25.

[82] F. Pes and G. Rodriguez, *The minimal-norm Gauss-Newton method and some of its regularized variants*, Electron. Trans. Numer. Anal., 53 (2020), pp. 459–480.

[83] ——, *A doubly relaxed minimal-norm Gauss–Newton method for underdetermined nonlinear least-squares problems*, Appl. Numer. Math., 171 (2022), pp. 233–248.

[84] R. Ramlau, *TIGRA-an iterative algorithm for regularizing nonlinear ill-posed problems*, Inverse Probl., 19 (2003), pp. 433–465.

[85] R. Ramlau and G. Teschke, *Tikhonov replacement functionals for iteratively solving nonlinear operator equations*, Inverse Probl., 21 (2005), pp. 1571–1592.

[86] ——, *A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints*, Numer. Math., 104 (2006), pp. 177–203.

[87] L. Reichel and G. Rodriguez, *Old and new parameter choice rules for discrete ill-posed problems*, Numer. Algorithms, 63 (2013), pp. 65–87.

[88] G. Rodriguez and S. Seatzu, *Numerical solution of the finite moment problem in a reproducing kernel Hilbert space*, J. Comput. Appl. Math., 33 (1990), pp. 233–244.

[89] ——, *On the numerical inversion of the Laplace transform in reproducing kernel Hilbert spaces*, IMA J. Numer. Anal., 13 (1993), pp. 463–475.

[90] A. Ruhe, *Accelerated Gauss-Newton algorithms for nonlinear least squares problems*, BIT, 19 (1979), pp. 356–367.

[91] Y. Saad, *Iterative Methods for Sparse Linear Systems, Second Edition*, SIAM, Philadelphia, PA, 2003.

[92] H. D. Simon and H. Zha, *Low-rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput., 21 (2000), pp. 2257–2274.

[93] F. Smithies, *Integral Equations*, Cambridge University Press, Cambridge, 1958.

[94] G. Stewart, *Matrix Algorithms: Volume 1: Basic Decompositions*, SIAM, Philadelphia, PA, 1998.

[95] L. N. Trefethen and D. Bau, *Numerical linear algebra*, vol. 50, SIAM, 1997.

[96] J. R. Wait, *Geo-Electromagnetism*, Academic Press, New York, 1982.

[97] R. Wang and Y. Xu, *Functional reproducing kernel Hilbert spaces for non-point-evaluation functional data*, Appl. Comput. Harmon. Anal., 46 (2019), pp. 569–623.

[98] ——, *Regularization in a functional reproducing kernel Hilbert space*, J. Complex., (2021), p. 101567.

[99] S. H. Ward and G. W. Hohmann, *Electromagnetic theory for geophysical applications. (In Electromagnetic Methods in Applied Geophysics)*, Society of Exploration Geophysicists, Tulsa, OK, 1987.

[100] G. M. Wing, *A primer on integral equations of the first kind: the problem of deconvolution and unfolding*, SIAM, Philadelphia, PA, 1991.

[101] K. Yosida, *Functional Analysis*, Classics in Mathematics, Springer, Berlin, 1995.