

## Classical Regularization Methods

In this section we review some of the most commonly used methods used when ill-posed inverse problems are treated. These methods are called regularization methods. Although the emphasis in this book is not on regularization techniques, it is important to understand the philosophy behind them and how the methods work. Later we analyze these methods also from the point of view of statistics which is one of the main themes in this book.

### 2.1 Introduction: Fredholm Equation

To explain the basic ideas of regularization, we consider a simple linear inverse problem. Following the traditions, the discussion in this chapter is formulated in terms of Hilbert spaces. A brief review of some of the functional analytic results can be found in Appendix A of the book.

Let  $H_1$  and  $H_2$  be separable Hilbert spaces of finite or infinite dimensions and  $A : H_1 \rightarrow H_2$  a compact operator. Consider first the problem of finding  $x \in H_1$  satisfying the equation

$$Ax = y, \tag{2.1}$$

where  $y \in H_2$  is given. This equation is said to be a *Fredholm equation of the first kind*. Since, clearly

1. the solution *exists* if and only if  $y \in \text{Ran}(A)$ , and
2. the solution is *unique* if and only if  $\text{Ker}(A) = \{0\}$ ,

both conditions must be satisfied to ensure that the problem has a unique solution. From the practical point of view, there is a third obstacle for finding a useful solution. The vector  $y$  typically represents measured data which is therefore contaminated by errors, i.e., instead of the exact equation (2.1), we have an approximate equation

$$Ax \approx y.$$

It is well known that even when the inverse of  $A$  exists, it cannot be continuous unless the spaces  $H_j$  are finite-dimensional. Thus, small errors in  $y$  may cause errors of arbitrary size in  $x$ .

**Example 1:** A classical ill-posed inverse problem is the deconvolution problem. Let  $H_1 = H_2 = L^2(\mathbb{R})$  and define

$$A : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R}), \quad (Af)(x) = \phi * f(x) = \int_{-\infty}^{\infty} \phi(x-y)f(y)dy,$$

where  $\phi$  is a Gaussian convolution kernel,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The operator  $A$  is injective, which is seen by applying the Fourier transform on  $Af$ , yielding

$$\mathcal{F}(Af)(\xi) = \int_{-\infty}^{\infty} e^{-i\xi x} Af(x)dx = \hat{\phi}(\xi)\hat{f}(\xi)$$

with

$$\hat{\phi}(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} > 0.$$

Therefore, if  $Af = 0$ , we have  $\hat{f} = 0$ , hence  $f = 0$ . Formally, the solution to the equation  $Af = g$  is

$$f(x) = \mathcal{F}^{-1}(\hat{\phi}^{-1}\hat{g})(x).$$

However, the above formula is not well defined for general  $g \in L^2(\mathbb{R})$  (or even in the space of tempered distributions) since the inverse of  $\hat{\phi}$  grows exponentially. Measurement errors of arbitrarily small  $L^2$ -norm in  $g$  can cause  $g$  to be not in  $\text{Ran}(A)$  and the integral not to converge, thus making the inversion formula practically useless.  $\diamond$

The following example shows that even when the Hilbert spaces are finite-dimensional, serious practical problems may occur.

**Example 2:** Let  $f$  be a real function defined over the interval  $[0, \infty)$ . The Laplace transform  $\mathcal{L}f$  of  $f$  is defined as the integral

$$\mathcal{L}f(s) = \int_0^{\infty} e^{-st} f(t)dt,$$

provided that the integral is convergent. We consider the following problem: Given the values of the Laplace transform at points  $s_j$ ,  $0 < s_1 < \dots < s_n < \infty$ , we want to estimate the function  $f$ . To this end, we approximate first the integral defining the Laplace transform by a finite sum,

$$\int_0^{\infty} e^{-s_j t} f(t)dt \approx \sum_{k=1}^n w_k e^{-s_j t_k} f(t_k),$$

where,  $w_k$ 's are the weights and  $t_k$ 's are the nodes of the quadrature rule, e.g., Gauss quadrature, Simpson's rule or the trapezoid rule. Let  $x_k = f(t_k)$ ,  $y_j = \mathcal{L}f(s_j)$  and  $a_{jk} = w_k e^{-s_j t_k}$ , and write the numerical approximation of the Laplace transform in the form (2.1), where  $A$  is an  $n \times n$  square matrix. Here,  $H_1 = H_2 = \mathbb{R}^n$ . In this example, we choose the data points logarithmically distributed, e.g.,

$$\log(s_j) = \left(-1 + \frac{j-1}{20}\right) \log 10, \quad 1 \leq j \leq 40,$$

to guarantee denser sampling near the origin. The quadrature rule is the 40-point Gauss-Legendre rule and the truncated interval of integration  $(0, 5)$ . Hence,  $A \in \mathbb{R}^{40 \times 40}$ .

Let the function  $f$  be

$$f(t) = \begin{cases} t, & \text{if } 0 \leq t < 1, \\ \frac{3}{2} - \frac{1}{2}t, & \text{if } 1 \leq t < 3, \\ 0, & \text{if } t \geq 3, \end{cases}$$

The Laplace transform can then be calculated analytically. We have

$$\mathcal{L}f(s) = \frac{1}{2s^2}(2 - 3e^{-s} + e^{-3s}).$$

The function  $f$  and its Laplace transform are depicted in Figure 2.1.

An attempt to estimate the values  $x_j = f(t_j)$  by direct solution of the system (2.1) even without adding any error leads to the catastrophic results shown also in Figure 2.1. The reason for the bad behaviour of this solution is that in this example, the condition number of the matrix  $A$ , defined as

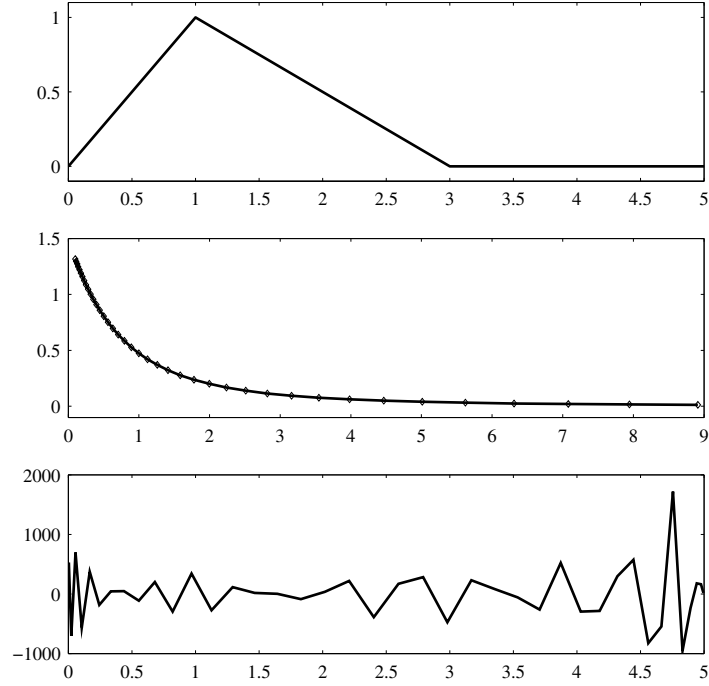
$$\kappa(A) = \|A\| \|A^{-1}\|$$

is very large, i.e.,  $\kappa(A) \approx 8.5 \times 10^{20}$ . Hence, even roundoff errors that in double precision are numerical zeroes are negatively affecting the solution.  $\diamond$

The above example demonstrates that the conditions 1 and 2 that guarantee the unique existence of a solution of equation (2.1) are not sufficient in practical applications. Even in the finite-dimensional problems, we must require further that the condition number is not excessively large. This can be formulated more precisely using the singular value decomposition of operators discussed in the following section.

Classical regularization methods are designed to overcome the obstacles illustrated in the examples above. To summarize, the basic idea of regularization methods is that, instead of trying to solve equation (2.1) exactly, one seeks to find a nearby problem that is uniquely solvable and that is robust in the sense that small errors in the data do not corrupt excessively this approximate solution.

In this chapter, we review three families of classical methods. These methods are (1) regularization by singular value truncation, (2) the Tikhonov regularization and (3) regularization by truncated iterative methods.



**Figure 2.1.** The original function (top), its Laplace transform (center) and the estimator obtained by solving the linear system (bottom).

## 2.2 Truncated Singular Value Decomposition

In this section,  $H_1$  and  $H_2$  are Hilbert spaces of finite or infinite dimension, equipped with the inner products  $\langle x, y \rangle_j$ ,  $x, y \in H_j$ ,  $j = 1, 2$ , and  $A : H_1 \rightarrow H_2$  is a compact operator. When there is no risk of confusion, the subindices in the inner products are suppressed. For the sake of keeping the notation fairly straightforward, we assume that both  $H_1$  and  $H_2$  are infinite-dimensional.

The starting point in this section is the following proposition.

**Proposition 2.1.** *Let  $H_1$ ,  $H_2$  and  $A$  be as above, and let  $A^*$  be the adjoint operator of  $A$ . Then*

1. *The spaces  $H_j$ ,  $j = 1, 2$ , allow orthogonal decompositions*

$$H_1 = \text{Ker}(A) \oplus (\text{Ker}(A))^\perp = \text{Ker}(A) \oplus \overline{\text{Ran}(A^*)},$$

$$H_2 = \overline{\text{Ran}(A)} \oplus (\text{Ran}(A))^\perp = \overline{\text{Ran}(A)} \oplus \text{Ker}(A^*).$$

2. *There exists orthonormal sets of vectors  $(v_n) \in H_1$ ,  $(u_n) \in H_2$  and a sequence  $(\lambda_j)$  of positive numbers,  $\lambda \searrow 0+$  such that*

$$\overline{\text{Ran}(A)} = \overline{\text{span}}\{u_n \mid n \in \mathbb{N}\}, \quad (\text{Ker}(A))^\perp = \overline{\text{span}}\{v_n \mid n \in \mathbb{N}\},$$

and the operator  $A$  can be represented as

$$Ax = \sum_n \lambda_j \langle x, v_n \rangle u_n.$$

The system  $(v_n, u_n, \lambda_n)$  is called the singular system of the operator  $A$ .

3. The equation  $Ax = y$  has a solution if and only if

$$y = \sum_n \langle y, u_n \rangle u_n, \quad \sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 < \infty.$$

In this case a solution is of the form

$$x = x_0 + \sum_n \frac{1}{\lambda_j} \langle y, u_n \rangle v_n,$$

where  $x_0 \in \text{Ker}(A)$  can be chosen arbitrarily.

The proofs of these results, with proper references, are briefly outlined in Appendix A.

The representation of the operator  $A$  in terms of its singular system is called the *singular value decomposition* of  $A$ , abbreviated as SVD of  $A$ . The above proposition gives a good picture of the possible difficulties in solving the equation  $Ax = y$ . First of all, let  $P$  denote the orthogonal projection on the closure of the range of  $A$ . By the above proposition, we see that  $P$  is given as

$$P : H_2 \rightarrow \overline{\text{Ran}(A)}, \quad y \mapsto \sum_n \langle y, u_n \rangle u_n. \quad (2.2)$$

It follows that for any  $x \in H_1$ , we have

$$\|Ax - y\|^2 = \|Ax - Py\|^2 + \|(1 - P)y\|^2 \geq \|(1 - P)y\|^2.$$

Hence, if  $y$  has a nonzero component in the subspace orthogonal to the range of  $A$ , the equation  $Ax = y$  cannot be satisfied exactly. Thus, the best we can do is to solve the projected equation,

$$Ax = PAx = Py. \quad (2.3)$$

This projection removes the most obvious obstruction of the solvability of the equation by replacing it with another substitute equation. However, given a noisy data vector  $y$ , there is in general no guarantee that the components  $\langle y, u_n \rangle$  tend to zero rapidly enough to guarantee convergence of the quadratic sum in the solvability condition 3 of Proposition 2.1.

Let  $P_k$  denote the finite-dimensional orthogonal projection

$$P_k : H_2 \rightarrow \text{span}\{u_1, \dots, u_k\}, \quad y \mapsto \sum_{n=1}^k \langle y, u_n \rangle u_n. \quad (2.4)$$

Since  $P_k$  is finite dimensional, we have  $P_k y \in \text{Ran}(A)$  for all  $k \in \mathbb{N}$ , and more importantly,  $P_k y \rightarrow Py$  in  $H_2$  as  $k \rightarrow \infty$ . Thus, instead of equation (2.3), we consider the projected equation

$$Ax = P_k y, \quad k \in \mathbb{N}. \quad (2.5)$$

This equation is always solvable. Taking on both sides the inner product with  $u_n$ , we find that

$$\lambda_n \langle x, v_n \rangle = \begin{cases} \langle y, u_n \rangle, & 1 \leq n \leq k, \\ 0, & n > k. \end{cases}$$

Hence, the solution to equation (2.5) is

$$x_k = x_0 + \sum_{n=1}^k \frac{1}{\lambda_j} \langle y, u_n \rangle,$$

for some  $x_0 \in \text{Ker}(A)$ . Observe that since for increasing  $k$ ,

$$\|Ax_k - Py\|^2 = \|(P - P_k)y\|^2 \rightarrow 0,$$

the residual of the projected equation can be made arbitrarily small.

Finally, to remove the ambiguity of the sought solution due to the possible noninjectivity of  $A$ , we select  $x_0 = 0$ . This choice minimizes the norm of  $x_k$ , since by orthogonality,

$$\|x_k\|^2 = \|x_0\|^2 + \sum_{j=1}^k \frac{1}{\lambda_j^2} |\langle y, u_j \rangle|^2.$$

These considerations lead us to the following definition.

**Definition 2.2.** *let  $A : H_1 \rightarrow H_2$  be a compact operator with the singular system  $(\lambda_n, v_n, u_n)$ . By the truncated SVD approximation (TSVD) of the problem  $Ax = y$  we mean the problem of finding  $x \in H_1$  such that*

$$Ax = P_k y, \quad x \perp \text{Ker}(A)$$

for some  $k \geq 1$ .

We are now ready to state the following result.

**Theorem 2.3.** *The problem given in Definition 2.2 has a unique solution  $x_k$ , called the truncated SVD (or TSVD) solution, which is*

$$x_k = \sum_{n=1}^k \frac{1}{\lambda_j} \langle y, u_n \rangle v_n.$$

Furthermore, the TSVD solution satisfies

$$\|Ax_k - y\|^2 = \|(1 - P)y\|^2 + \|(P - P_k)y\|^2 \rightarrow \|(1 - P)y\|^2$$

as  $k \rightarrow \infty$ , where the projections  $P$  and  $P_k$  are given by formulas (2.2) and (2.4), respectively.

Before presenting numerical examples, we briefly discuss the above regularization scheme in the finite-dimensional case. Therefore, let  $A \in \mathbb{R}^{m \times n}$ ,  $A \neq 0$ , be a matrix defining a linear mapping  $\mathbb{R}^n \rightarrow \mathbb{R}^m$ , and consider the matrix equation

$$Ax = y.$$

In Appendix A, it is shown that the matrix  $A$  has a singular value decomposition

$$A = U\Lambda V^T,$$

where  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  are orthogonal matrices, i.e.,

$$U^T = U^{-1}, \quad V^T = V^{-1},$$

and  $\Lambda \in \mathbb{R}^{m \times n}$  is a diagonal matrix with diagonal elements

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{\min(m,n)} \geq 0.$$

Let us denote by  $p$ ,  $1 \leq p \leq \min(m, n)$ , the largest index for which  $\lambda_p > 0$ , and let us think of  $U = [u_1, u_2, \dots, u_m]$  and  $V = [v_1, v_2, \dots, v_n]$  as arrays of column vectors. The orthogonality of the matrices  $U$  and  $V$  is equivalent to saying that the vectors  $v_j$  and  $u_j$  form orthonormal base for  $\mathbb{R}^n$  and  $\mathbb{R}^m$ , respectively. Hence, the singular system of the mapping  $A$  is  $(v_j, u_j, \lambda_j)_{1 \leq j \leq p}$ .

We observe that If  $p = n$ ,

$$\mathbb{R}^n = \text{span}\{v_1, \dots, v_n\} = \text{Ran}(A^T),$$

and consequently,  $\text{Ker}(A) = \{0\}$ . If  $p < n$ , then we have

$$\text{Ker}(A) = \text{span}\{v_{p+1}, \dots, v_n\}.$$

Hence, any vector  $x_0$  in the kernel of  $A$  is of the form

$$x_0 = V_0 c, \quad V_0 = [v_{p+1}, \dots, v_n] \in \mathbb{R}^{n \times (n-p)}$$

for some  $c \in \mathbb{R}^{n-p}$ .

In the finite-dimensional case, we need not to worry about the convergence condition 3 of Proposition 2.1; hence the projected equation (2.3) always has a solution,

$$x = x_0 + A^\dagger y,$$

where  $x_0$  is an arbitrary vector in the kernel of  $A$ . The matrix  $A^\dagger$  is called the *pseudoinverse* or *Moore–Penrose inverse* of  $A$ , and it is defined as

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

where

$$A^\dagger = \begin{bmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & & \\ & & \ddots & \\ \vdots & & & 1/\lambda_p & \vdots \\ & & & 0 & \\ 0 & & \cdots & & 0 \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Properties of the pseudoinverse are listed in the “Notes and Comments” at the end of this chapter.

When  $x_0 = 0$ , the solution  $x = A^\dagger y$  is called simply the *minimum norm solution* of the problem  $Ax = y$ , since

$$\|A^\dagger y\| = \min\{\|x\| \mid \|Ax - y\| = \|(1 - P)y\|\},$$

where  $P$  is the projection onto the range of  $A$ . Thus, the minimum norm solution is the solution that minimizes the residual error and that has the minimum norm. Observe that in this definition, there is no truncation since we keep all the nonzero singular values.

In the case of inverse problems, the minimum norm solution is often useless due to the ill-conditioning of the matrix  $A$ . The smallest positive singular values are very close to zero and the minimum norm solution is sensitive to errors in the vector  $y$ . Therefore, in practice we need to choose the truncation index  $k < p$  in Definition 2.2. The question that arises is: what is a judicious choice for the value of the for the truncation level  $k$ ? There is a rule of thumb that is often referred to as the *discrepancy principle*. Assume that the data vector  $y$  is a noisy approximation of a noiseless vector  $y_0$ . While  $y_0$  is unknown to us, we may have an estimate of the noise level, e.g., we may have

$$\|y - y_0\| \simeq \varepsilon \quad (2.6)$$

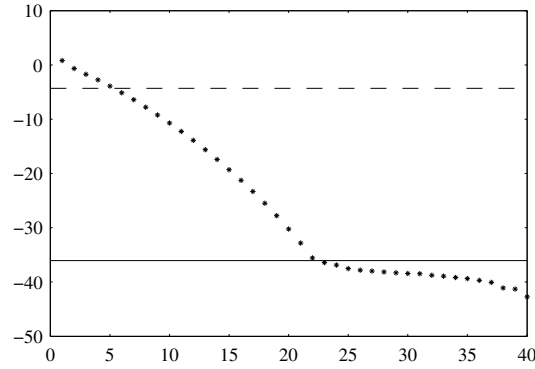
for some  $\varepsilon > 0$ . The discrepancy principle states that we cannot expect the approximate solution to yield a smaller residual error than the measurement error, since otherwise we would be fitting the solution to the noise. This principle leads to the following selection criterion for the truncation parameter  $k$ : choose  $k$ ,  $1 \leq k \leq m$  the largest index that satisfies

$$\|y - Ax_k\| = \|y - P_k y\| \leq \varepsilon.$$

In the following example, the use of the minimum norm solution and the TSVD solution are demonstrated.

**Example 3:** We return to the Laplace inversion problem of Example 2. Let  $A$  be the same matrix as before. A plot of the logarithms of its singular values is shown in Figure 2.2.





**Figure 2.2.** The singular values of the discretized Laplace transform on a logarithmic scale. The solid line indicates the level of the machine epsilon.

Let  $\varepsilon_0$  denote the *machine epsilon*, i.e., the smallest floating point number that the machine recognizes to be nonzero. In IEEE double precision arithmetic, this number is of the order  $10^{-16}$ . In Figure 2.2, we have marked this level by a solid horizontal line. The plot clearly demonstrates that the matrix is numerically singular: Singular values smaller than  $\varepsilon_0$  represent roundoff errors and should be treated as zeros.

First, we consider the case where only the roundoff error is present and the data is precise within the arithmetic. We denote in this case  $y = y_0$ . Here, the minimum norm solution  $x = A^\dagger y_0$  should give a reasonable estimate for the discrete values of  $f$ . It is also clear that although 22 of the singular values are larger than  $\varepsilon_0$ , the smallest ones above this level are quite close to  $\varepsilon_0$ .

In Figure 2.3 we have plotted the reconstruction of  $f$  with  $x = A^\dagger y_0$  computed with  $p = 20, 21$  and 22 singular values retained.

For comparison, let us add artificial noise, i.e., the data vector is

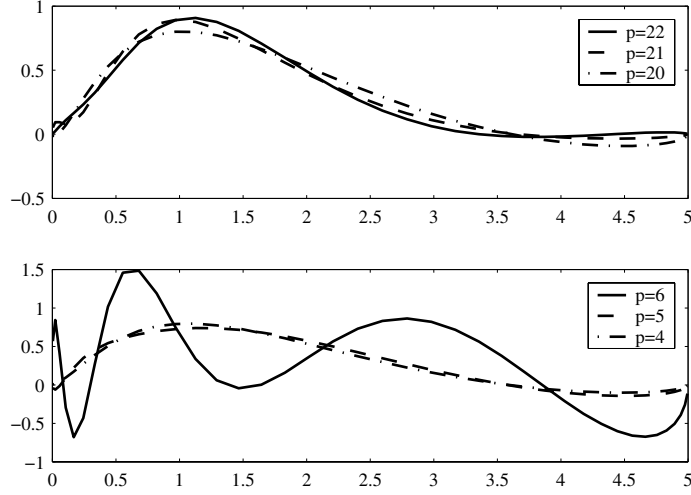
$$y = y_0 + e,$$

where the noise vector  $e$  is normally distributed zero mean noise with the standard deviation (STD)  $\sigma$  being 1% of the maximal data component, i.e.,  $\sigma = 0.01 \|y_0\|_\infty$ . The logarithm of this level is marked in Figure 2.2 by a dashed horizontal line. In this case only five singular values remain above  $\sigma$ .

When the standard deviation of the noise is given, it is not clear without further analysis how one should select the parameter  $\varepsilon$  in the discrepancy principle. In this example, expect somewhat arbitrarily the norm of the noise to be of the order of  $\sigma$ . Figure 2.3 depicts the reconstructions of  $f$  obtained from the TSVD solutions  $x_k$  with  $k = 4, 5$  and 6. We observe that for  $k = 6$ , the solution is oscillatory.

Let us remark here that the noise level criterion in the discrepancy principle does not take into account the stochastic properties of the noise. Later in this chapter, we discuss in more detail how to choose the cutoff level.

Let us further remark that single reconstructions such as those displayed in Figure 2.3 are far from giving a complete picture of the stability of the reconstruction. Instead, one should analyze the variance of the solutions by performing several runs from independently generated data. This issue will be discussed in Chapter 5, where the classical methods are revisited and analyzed from the statistical point of view.  $\diamond$



**Figure 2.3.** The inverse Laplace transform by using the singular value truncation. The top figure corresponds to no artificial noise in the data, the bottom one with 1% additive artificial noise.

### 2.3 Tikhonov Regularization

The discussion in Section 2.2 demonstrates that when solving the equation  $Ax = y$ , problems occur when the singular values of the operator  $A$  tend to zero rapidly, causing the norm of the approximate solution  $x_k$  to go to infinity when  $k \rightarrow \infty$ . The idea in the basic regularization scheme discussed in this section is to control simultaneously the norm of the residual  $r = Ax - y$  and the norm of the approximate solution  $x$ . We start with the following definition.

**Definition 2.4.** Let  $\delta > 0$  be a given constant. The Tikhonov regularized solution  $x_\delta \in H_1$  is the minimizer of the functional

$$F_\delta(x) = \|Ax - y\|^2 + \delta \|x\|^2,$$

provided that a minimizer exists. The parameter  $\delta > 0$  is called the regularization parameter.

Observe that the regularization parameter plays essentially the role of a Lagrange multiplier, i.e., we may think that we are solving a minimization problem with the constraint  $\|x\| = R$ , for some  $R > 0$ .

The following theorem shows that Definition 2.4 is reasonable.

**Theorem 2.5.** *Let  $A : H_1 \rightarrow H_2$  be a compact operator with the singular system  $(\lambda_n, v_n, u_n)$ . Then the Tikhonov regularized solution exists, is unique, and is given by the formula*

$$x_\delta = (A^*A + \delta I)^{-1}A^*y = \sum_n \frac{\lambda_n}{\lambda_n^2 + \delta} \langle y, u_n \rangle v_n. \quad (2.7)$$

*Proof:* We have

$$\langle x, (A^*A + \delta I)x \rangle \geq \delta \|x\|^2,$$

i.e., the operator  $(A^*A + \delta I)$  is bounded from below. It follows from the Riesz representation theorem (see Appendix A) that the inverse of this operator exists and

$$\|(A^*A + \delta I)^{-1}\| \leq \frac{1}{\delta}. \quad (2.8)$$

Hence,  $x_\delta$  in (2.7) is well defined. Furthermore, expressing the equation

$$(A^*A + \delta I)x = A^*y$$

in terms of the singular system of  $A$ , we have

$$\sum_n (\lambda_n^2 + \delta) \langle x, v_n \rangle v_n + Px = \sum_n \lambda_n \langle y, u_n \rangle v_n,$$

where  $P : H_1 \rightarrow \text{Ker}(A)$  is the orthogonal projector. By projecting onto the eigenspaces  $\text{sp}\{v_n\}$ , we find that  $Px = 0$  and  $(\lambda_n^2 + \delta) \langle x, v_n \rangle = \lambda_n \langle y, u_n \rangle$ .

To show that  $x_\delta$  minimizes the quadratic functional  $F_\delta$ , let  $x$  be any vector in  $H_1$ . By decomposing  $x$  as

$$x = x_\delta + z, \quad z = x - x_\delta,$$

and arranging the terms in  $F_\delta(x)$  according to the degree with respect to  $z$ , we obtain

$$\begin{aligned} F_\delta(x_\delta + z) &= F_\delta(x_\delta) + \langle z, (A^*A + \delta I)x_\delta - A^*y \rangle + \langle z, (A^*A + \delta I)z \rangle \\ &= F_\delta(x_\delta) + \langle z, (A^*A + \delta I)z \rangle \end{aligned}$$

by definition of  $x_\delta$ . The last term is nonnegative and vanishes only if  $z = 0$ . This proves the claim.  $\square$

**Remark:** When the spaces  $H_j$  are finite-dimensional and  $A$  is a matrix, we may write

$$F_\delta(x) = \left\| \begin{bmatrix} A \\ \sqrt{\delta}I \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2.$$

From the inequality (2.8) it follows that the singular values of the matrix

$$K_\delta = \begin{bmatrix} A \\ \sqrt{\delta}I \end{bmatrix}$$

are bounded from below by  $\sqrt{\delta}$ , so the minimizer of the functional  $F_\delta$  is simply

$$x_\delta = K_\delta^\dagger \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

This formula is particularly handy in numerical implementation of the Tikhonov regularization method.

The choice of the value of the regularization parameter  $\delta$  based on the noise level of the measurement  $y$  is a central issue in the literature discussing Tikhonov regularization. Several methods for choosing  $\delta$  have been proposed. Here, we discuss briefly only one of them, known as the *Morozov discrepancy principle*. This principle is essentially the same as the discrepancy principle discussed in connection with the singular value truncation method.

Let us assume that we have an estimate  $\varepsilon > 0$  of the norm of the error in the data vector as in (2.6). Then any  $x \in H_1$  such that

$$\|Ax - y\| \leq \varepsilon$$

should be considered an acceptable approximate solution. Let  $x_\delta$  be defined by (2.7), and

$$f : \mathbb{R}_+ \rightarrow \mathbb{R}_+, \quad f(\delta) = \|Ax_\delta - y\| \quad (2.9)$$

the discrepancy related to the parameter  $\delta$ . The Morozov discrepancy principle says that the regularization parameter  $\delta$  should be chosen from the condition

$$f(\delta) = \|Ax_\delta - y\| = \varepsilon, \quad (2.10)$$

if possible, i.e., the regularized solution should not try to satisfy the data more accurately than up to the noise level.

The following theorem gives a condition when the discrepancy principle can be used.

**Theorem 2.6.** *The discrepancy function (2.9) is strictly increasing and*

$$\|Py\| \leq f(\delta) \leq \|y\|, \quad (2.11)$$

where  $P : H_2 \rightarrow \text{Ker}(A^*) = \text{Ran}(A)^\perp$  is the orthogonal projector. Hence, the equation (2.10) has a unique solution  $\delta = \delta(\varepsilon)$  if and only if  $\|Py\| \leq \varepsilon \leq \|y\|$ .

*Proof:* By using the singular system representation of the vector  $x_\delta$ , we have

$$\begin{aligned} \|Ax_\delta - y\|^2 &= \sum \left( \frac{\lambda_n^2}{\lambda_n^2 + \delta} - 1 \right)^2 \langle y, u_n \rangle^2 + \|Py\|^2 \\ &= \sum \left( \frac{\delta}{\lambda_n^2 + \delta} \right)^2 \langle y, u_n \rangle^2 + \|Py\|^2. \end{aligned}$$

Since, for each term of the sum,

$$\frac{d}{d\delta} \left( \frac{\delta}{\lambda_n^2 + \delta} \right)^2 = \frac{2\delta\lambda_n^2}{(\lambda_n^2 + \delta)^3} > 0, \quad (2.12)$$

the mapping  $\delta \mapsto \|Ax_\delta - y\|^2$  is strictly increasing, and

$$\|Py\|^2 = \lim_{\delta \rightarrow 0+} \|Ax_\delta - y\|^2 \leq \|Ax_\delta - y\|^2 \leq \lim_{\delta \rightarrow \infty} \|Ax_\delta - y\|^2 = \|y\|^2,$$

as claimed.  $\square$

**Remark** The condition  $\|Py\| \leq \varepsilon$  is natural in the sense that any component in the data  $y$  that is orthogonal to the range of  $A$  must be due to noise. On the other hand, the condition  $\varepsilon < \|y\|$  can be understood in the sense that the error level should not exceed the signal level. Indeed, if  $\|y\| < \varepsilon$ , we might argue that, from the viewpoint of the discrepancy principle,  $x = 0$  is an acceptable solution.

The Morozov discrepancy principle is rather straightforward to implement numerically, apart of problems that arise from the size of the matrices. Indeed, if  $A$  is a matrix with nonzero singular values  $\lambda_1 \geq \dots \geq \lambda_r$ , one can employ e.g., Newton's method to find the unique zero of the function

$$f(\delta) = \sum_{j=1}^r \left( \frac{\delta}{\lambda_j^2 + \delta} \right)^2 \langle y, u_j \rangle^2 + \|Py\|^2 - \varepsilon^2.$$

The derivative of this function with respect to the parameter  $\delta$  can be expressed without a reference to the singular value decomposition. Indeed, from formula (2.12), we find that

$$f'(\delta) = \sum \frac{2\delta\lambda_n^2}{(\lambda_n^2 + \delta)^3} \langle u_n, y \rangle^2 = \langle x_\delta, \delta(A^*A + \delta I)^{-1}x_\delta \rangle.$$

This formula is valuable in particular when  $A$  is a large sparse matrix and the linear system with the matrix  $A^*A + \delta I$  is easier to calculate than the singular value decomposition.

**Example 4:** Anticipating the statistical analysis of the inverse problems, we consider the problem of how to set the noise level  $\varepsilon$  appearing in the discrepancy principle. Assume that we have a linear inverse problem with additive noise model, i.e.,  $A \in \mathbb{R}^{k \times m}$  is a known matrix and the model is

$$y = Ax + e = y_0 + e.$$

Furthermore, assume that we have information about the statistics of the noise vector  $e \in \mathbb{R}^k$ . The problem is, how does one determine a reasonable noise level based on the probability distribution of the noise. In principle, there are several possible candidates. Remembering that  $e$  is a random variable, we might in fact define

$$\varepsilon = \mathbb{E}\{\|e\|\}, \quad (2.13)$$

where  $\mathbb{E}$  is the expectation (see Appendix B). Equally well, one could argue that another judicious choice is to set

$$\varepsilon^2 = \mathbb{E}\{\|e\|^2\}, \quad (2.14)$$

leading to a slightly different value of  $\varepsilon$ . In general, these levels can be computed either numerically by generating randomly a sample of noise vectors and averaging, or analytically, if the explicit integrals of the probability densities are available.

For a simple illustration of how (2.13) and (2.14) differ from each other, assume that  $k = 1$ , i.e., the data  $y$  is a real number and  $e \sim \mathcal{U}(0, 1)$ , i.e.,  $e$  has a uniform probability distribution on the interval  $[0, 1]$ . The criterion (2.13) would give

$$\varepsilon = \int_0^1 t dt = \frac{1}{2},$$

while the second criterion leads to

$$\varepsilon = \left( \int_0^1 t^2 dt \right)^{1/2} = \frac{1}{\sqrt{3}}.$$

for another, more frequently encountered example, consider  $k$ -variate zero mean Gaussian noise with independent components, i.e.,  $e \sim \mathcal{N}(0, \sigma^2 I)$ , where  $\sigma^2$  is the variance and  $I$  is the unit matrix of dimension  $k$ . In this case, the criterion (2.14) immediately yields

$$\varepsilon^2 = \mathbb{E}\{\|e\|^2\} = k\sigma^2.$$

The first criterion requires more work. We have

$$\varepsilon = \frac{1}{(2\pi\sigma^2)^{k/2}} \int_{\mathbb{R}^k} \|t\| \exp\left(-\frac{1}{2\sigma^2}\|t\|^2\right) dt,$$

which, after passing to polar coordinates and properly scaling the variables, yields

$$\varepsilon = \frac{|\mathbb{S}^{k-1}|}{(2\pi)^{k/2}} \sigma \int_0^\infty t^k \exp\left(-\frac{1}{2}\|t\|^2\right) dt = \gamma_k \sigma.$$

Here,  $|\mathbb{S}^{k-1}|$  denotes the surface area of the unit ball. It is left as an exercise to evaluate the scaling factor  $\gamma_k$ . The important thing to notice is that both results scale linearly with  $\sigma$ .

The important thing to notice here that  $\|e\|$  is a random variable. For example, taking above  $k = 100$  and  $\sigma = 1$ , the probability of  $9 < \|e\| \leq 11$  is approximately 0.84.

Often, in classical regularization literature, the noise level used in the Morozov discrepancy principle is adjusted by an extra parameter  $\tau > 1$  to

avoid underregularization. Using the  $k$ -variate Gaussian white noise model, the discrepancy condition would give

$$\|Ax_\alpha - y\| = \tau\sqrt{k}\sigma.$$

A common choice is  $\tau = 1.1$ .  $\diamond$

**Example 5:** As an example of the use of the Tikhonov regularization method, consider the image deblurring problem, i.e., a deconvolution problem in the plane, introduced in Example 1. It is instructive to express the Tikhonov regularized solution using Fourier analysis. Therefore, let  $H_1 = H_2 = L^2(\mathbb{R}^2)$ . To guarantee the integrability of the image, we assume that  $f$  is compactly supported. With respect to the inner product of  $L^2(\mathbb{R}^2)$ , the adjoint of the convolution operator with a real valued kernel is

$$A^*f(x) = \int_{\mathbb{R}^2} \phi(y-x)f(y)dy.$$

Moreover, if the kernel is even, i.e.,  $\phi(-x) = \phi(x)$ ,  $A$  is self-adjoint. Since  $\widehat{Af} = \widehat{\phi}\widehat{f}$ , the operator  $A$  allows a Fourier representation

$$Af(x) = \mathcal{F}^{-1}(\widehat{\phi}\widehat{f})(\xi) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{i\langle \xi, x \rangle} \widehat{\phi}(\xi) \widehat{f}(\xi) d\xi.$$

Similarly, the adjoint operator can be written as

$$A^*f(x) = \mathcal{F}^{-1}(\overline{\widehat{\phi}}\widehat{f})(\xi).$$

Based on this representation, we have

$$(A^*A + \delta I)f(x) = \mathcal{F}^{-1}((|\widehat{\phi}|^2 + \delta)\widehat{f})(x)$$

and further, the operator defining the Tikhonov regularized solution is simply

$$(A^*A + \delta I)^{-1}A^*g(x) = \mathcal{F}^{-1}\left(\frac{\overline{\widehat{\phi}}}{|\widehat{\phi}|^2 + \delta}\widehat{g}\right)(x).$$

We have

$$\left|\frac{\overline{\widehat{\phi}(\xi)}}{|\widehat{\phi}(\xi)|^2 + \delta}\right| \leq \frac{1}{\delta},$$

so the operator is well defined in  $L^2(\mathbb{R}^2)$  as the theory predicts.

Although it is possible to determine a numerical solution based on the above formula, we represent the numerical solution here by using direct matrix discretization.

Let the image area be the unit rectangle  $[0, 1] \times [0, 1]$ , the true image  $f$  being identically zero outside this area. Assume that the image area is divided into pixels of equal size,  $P_{jk} = [(j-1)\Delta, j\Delta] \times [(k-1)\Delta, k\Delta]$ ,  $1 \leq j, k \leq N$ , where  $\Delta = 1/N$ . If  $p_{jk}$  denotes the centerpoint of  $P_{jk}$ , we write the approximation

$$g(p_{jk}) \approx \sum_{\ell,m=1}^N \Delta^2 \phi(p_{jk} - p_{\ell m}) f(p_{\ell m}),$$

or, in the matrix form,

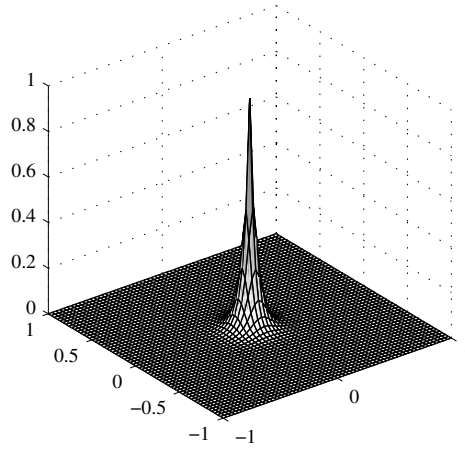
$$y = Ax,$$

where  $x, y \in \mathbb{R}^{N^2}$  are vectors with the pixel values stacked in a long vector and  $A \in \mathbb{R}^{N^2 \times N^2}$  a convolution matrix arranged accordingly.

In our numerical example, we consider the convolution kernel

$$\phi(x) = e^{-\alpha|x|}$$

with  $\alpha = 20$ . The convolution kernel is plotted in Figure 2.4.



**Figure 2.4.** The convolution kernel used for image blurring.

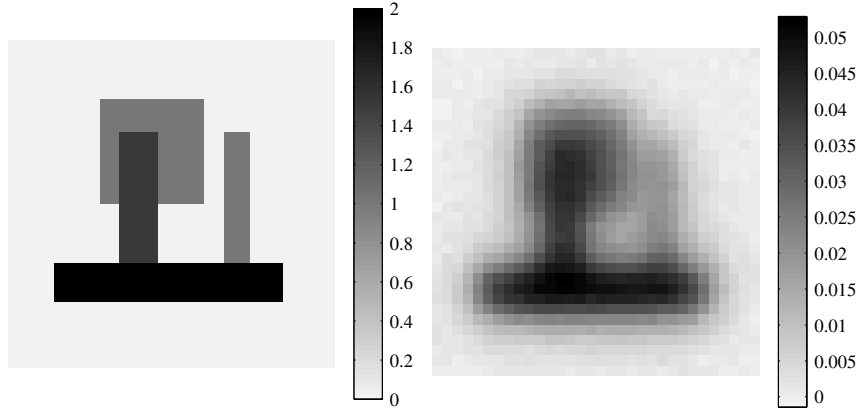
To avoid the infamous inverse crime - or at least the most evident version of it - we have computed the blurred data using a finer mesh (size  $50 \times 50$  pixels) than the one in which the blurred image is given (size  $32 \times 32$ ). The true and the blurred images are shown in Figure 2.5.

The noise model we use here is Gaussian additive noise,

$$y = Ax + e,$$

where  $e$  is a random vector with independent components, each component being zero mean normally distributed. The standard deviation of each component of  $e$  is  $\sigma = 0.01 \max(Ax)$ , i.e., 1% of the maximum pixel value in the blurred noiseless image. To fix the noise level used in the Morozov discrepancy principle, we use the criterion (2.14) of the previous example, i.e., in this example we set

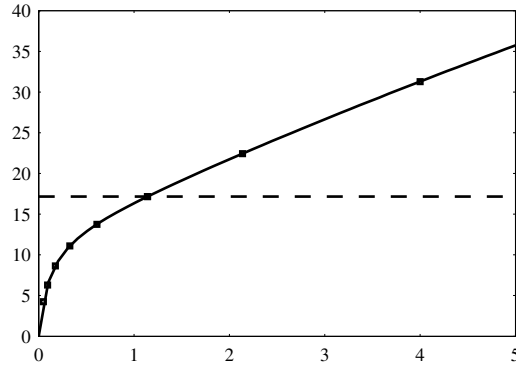




**Figure 2.5.** Original image and the noisy blurred image.

$$\varepsilon = N\sigma.$$

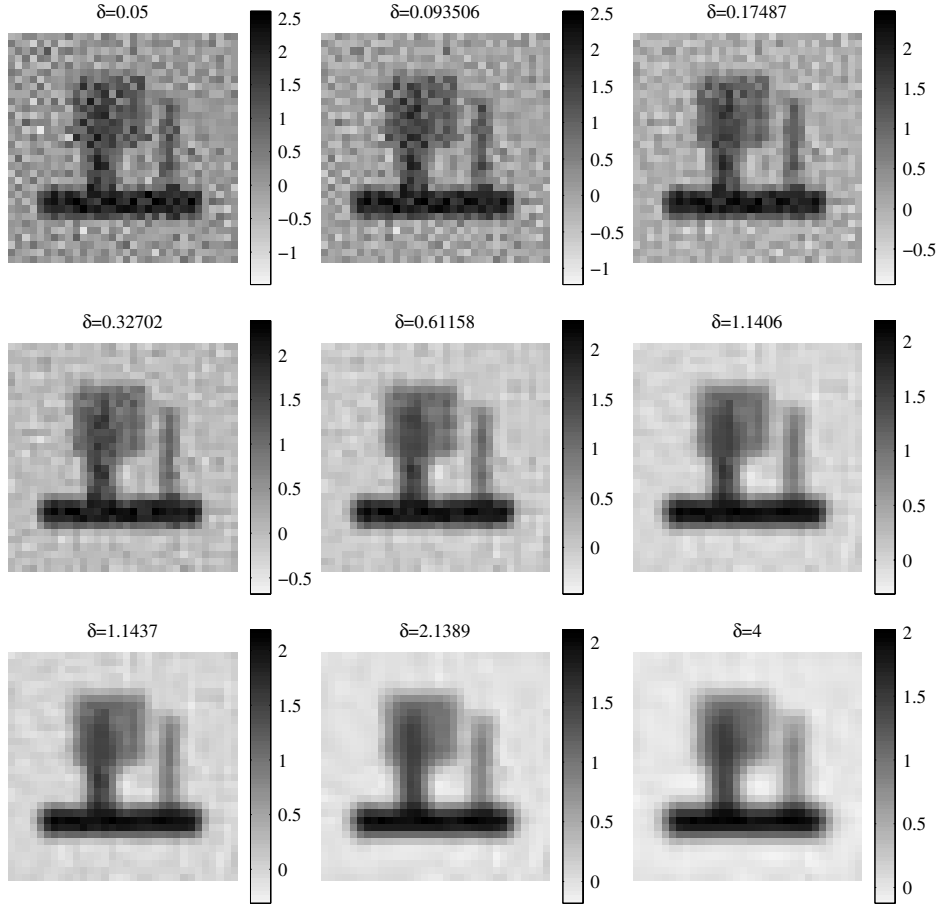
In Figure 2.6, we have plotted a piece of the curve  $\delta \mapsto \|Ax_\delta - y\|$ . The noise level is marked with a dashed line. Evidently, in this case, the condition (2.11) of the Theorem 2.6 is satisfied, so the Morozov discrepancy principle is applicable.



**Figure 2.6.** The discrepancy versus the regularization parameter  $\delta$ . The estimated noise level is marked by a dashed line. The asterisks mark the values of the regularization parameters corresponding to the regularised solutions of the Figure 2.7

The value  $\delta = \delta(\varepsilon)$  in this example is calculated using a bisection method. To illustrate the effect on the solution of the regularization parameter, we calculate Tikhonov regularized solutions with nine different values of the regularization parameter. These values of  $\delta$  are marked in Figure 2.6 by an asterisk. The outcomes are shown in Figure 2.7. When the regularization pa-

parameter is significantly below  $\delta(\varepsilon)$ , the outcome is noisy, i.e., the solution is *underregularized*, while for large values, the results get again blurred. These solutions are often said to be *overregularized*.  $\diamond$



**Figure 2.7.** Nine reconstructions from the same noisy data with various values of the regularization parameter  $\delta$ . The reconstruction that corresponds to the Morozov discrepancy principle is in the second row at right.

### 2.3.1 Generalizations of the Tikhonov Regularization

The Tikhonov regularization method is sometimes applicable also when non-linear problems are considered, i.e., to find  $x \in H_1$  satisfying

$$y = A(x) + e,$$

where  $A : H_1 \rightarrow H_2$  is a nonlinear mapping and  $e$  is observation noise. If the mapping  $A$  is such that large changes in the vector  $x$  may produce small changes in  $A(x)$ , the problem is ill-posed and numerical methods, typically, iterative ones, may fail to find a satisfactory estimate of  $x$ . The nonlinear Tikhonov regularization scheme amounts to searching for an  $x$  that minimizes the functional

$$F_\delta(x) = \|A(x) - y\|^2 + \delta\|x\|^2.$$

As this functional is no longer a quadratic one, it is not clear whether a minimizer exists, it is unique or how to determine it. The most common method to search for a feasible solution is to use an iterative scheme based on successive linearizations of  $A$ .

**Definition 2.7.** *The operator  $A : H_1 \rightarrow H_2$  is Fréchet differentiable at a point  $x_0$  if it allows an expansion*

$$A(x_0 + z) = A(x_0) + R_{x_0}z + W(x_0, z).$$

Here  $R_{x_0} : H_1 \rightarrow H_2$  is a continuous linear operator and

$$\|W(x_0, z)\| \leq \|z\|\epsilon(x_0, z),$$

where the functional  $z \mapsto \epsilon(x_0, z)$  tends to zero as  $z \rightarrow 0$ .

Let  $A$  be Fréchet differentiable. The linearization of  $A$  around a given point  $x_0$  leads to the approximation of the functional  $F_\delta$ ,

$$\begin{aligned} F_\delta(x) &\approx \tilde{F}_\delta(x; x_0) = \|A(x_0) + R_{x_0}(x - x_0) - y\|^2 + \delta\|x\|^2 \\ &= \|R_{x_0}x - g(y, x_0)\|^2 + \delta\|x\|^2, \end{aligned}$$

where

$$g(y, x_0) = y - A(x_0) + R_{x_0}x_0.$$

From the previous section we know that the minimizer of the functional  $F_\delta(x; x_0)$  is

$$x = (R_{x_0}^* R_{x_0} + \delta I)^{-1} R_{x_0}^* g(y, x_0).$$

While a straightforward approach would suggest to choose the new approximate solution as the base point for a new linearization, it may happen that the solution of the linearized problem does not reflect adequately the nonlinearities of the original function. Therefore a better strategy is to implement some form of stepsize control. This leads us to the following iterative method.

1. Pick an initial guess  $x_0$  and set  $k = 0$ .
2. Calculate the Fréchet derivative  $R_{x_k}$ .
3. Determine

$$x = (R_{x_k}^* R_{x_k} + \delta I)^{-1} R_{x_k}^* g(y, x_k), \quad g(y, x_k) = y - A(x_k) + R_{x_k}x_k,$$

and define  $\delta x = x - x_k$ .

4. Find  $s > 0$  by minimizing the function

$$f(s) = \|A(x_k + s\delta x) - y\|^2 + \|x_k + s\delta x\|^2.$$

5. Set  $x_{k+1} = x_k + s\delta x$  and increase  $k \leftarrow k + 1$ .  
 6. Repeat steps 2.-5. until the method converges.

In the generalization of Tikhonov regularization that we just described, the linear operator  $A$  has been replaced by a nonlinear one. Another way of generalizing Tikhonov regularization method is concerned with the choice of the penalty term.<sup>1</sup> Indeed, we may consider the following minimization problem: Find  $x \in H_1$  that minimizes the functional

$$\|Ax - y\|^2 + \delta G(x),$$

where  $G : H_1 \rightarrow \mathbb{R}$  is a nonnegative functional. The existence and uniqueness of the solution of this problem depends on the choice of the functional  $G$ .

The most common version of this generalization sets

$$G(x) = \|L(x - x_0)\|^2, \quad (2.15)$$

where  $L : \mathcal{D}(L) \rightarrow H_2$ ,  $\mathcal{D}(L) \subset H_1$  is a linear operator and  $x_0 \in H_1$  is given. Typically, when  $H_1$  is a function space,  $L$  will be a differential operator. This choice forces the solutions of the corresponding minimization problem to be smooth.

In the finite-dimensional case, the operator  $L$  is a matrix in  $R^{k \times n}$ . The Tikhonov functional to be minimized can then be written as

$$\|Ax - y\|^2 + \delta \|L(x - x_0)\|^2 = \left\| \begin{bmatrix} A \\ \sqrt{\delta} L \end{bmatrix} x - \begin{bmatrix} y \\ \sqrt{\delta} L x_0 \end{bmatrix} \right\|^2.$$

The minimizer of this functional is

$$x_\delta = K^\dagger \begin{bmatrix} y \\ \sqrt{\delta} L x_0 \end{bmatrix}, \quad K = \begin{bmatrix} A \\ \sqrt{\delta} L \end{bmatrix},$$

provided that the singular values of  $K$  are all positive. If some of the singular values of  $K$  vanish, one may argue that the choice of  $L$  does not regularize the problem properly.

Finally, we may combine both the generalizations above and consider the problem of minimizing a functional of the type

$$F_\delta(x) = \|A(x) - y\|^2 + \delta G(x).$$

This problem leads to a general nonlinear optimization problem which is not discussed in detail here.

---

<sup>1</sup>This, in fact is the original form of Tikhonov regularization; see; “Notes and Comments.”

## 2.4 Regularization by Truncated Iterative Methods

Consider again the simple linear matrix equation (2.1),  $Ax = y$ . Numerical analysis offers a rich selection of various iterative solvers for this equation. It turns out that these solvers, albeit not originally designed for regularization purposes, can often be used as regularizers when the data  $y$  are corrupted by noise. In this section, we discuss three different iterative methods and their regularizing properties.

### 2.4.1 Landweber–Fridman Iteration

The first iterative scheme discussed here is a method based on *fixed point iteration*. We start recalling a few concepts. Let  $H$  be a Hilbert space and  $S \subset H$ . Consider a mapping, not necessarily linear,  $T : H \rightarrow H$ . We say that  $S$  is an *invariant set* for  $T$  if  $x \in S$  implies  $T(x) \in S$ , or briefly  $T(S) \subset S$ . The operator  $T$  is said to be a *contraction* on an invariant set  $S$  if there is  $\kappa \in \mathbb{R}$ ,  $0 \leq \kappa < 1$  such that for all  $x, z \in S$ ,

$$\|T(x) - T(z)\| < \kappa \|x - z\|.$$

A vector  $x \in H$  is called a *fixed point* of  $T$  if we have

$$T(x) = x.$$

The following elementary result, known as the *fixed point theorem*, is proved in Appendix A.

**Proposition 2.8.** *Let  $H$  be a Hilbert space and  $S \subset H$  a closed invariant set for the mapping  $T : H \rightarrow H$ . Assume further that  $T$  is a contraction in  $S$ . Then there is a unique  $x \in S$  such that  $T(x) = x$ . The fixed point  $x$  can be found by the fixed point iteration as*

$$x = \lim_{k \rightarrow \infty} x_k, \quad x_{k+1} = T(x_k),$$

where the initial value  $x_0 \in S$  is arbitrary.

Consider now the linear equation (2.1). By using the notation of Section 2.2, we write first the right-hand side  $y$  as

$$y = Py + (1 - P)y, \quad Py \in \overline{\text{Ran}(A)}, \quad (1 - P)y \in \text{Ker}(A^*).$$

Since there is no way of matching  $Ax$  with the vector  $(1 - P)y$  that is orthogonal to the range of  $A$ , we filter it out by applying  $A^*$  to both sides of the equation. This leads to the *normal equations*

$$A^*Ax = A^*Py + A^*(1 - P)y = A^*y. \quad (2.16)$$

We then seek to solve this normal equations by an iterative method. To this end, observe that when the normal equations are satisfied, we have

$$x = x + \beta(A^*y - A^*Ax) = T(x) \quad (2.17)$$

for all  $\beta \in \mathbb{R}$ . Therefore the solution  $x$  of the equation is a fixed point for the affine map  $T$ . Our aim is to solve this equation by fixed point iterations. Hence, let  $x_0 = 0$  and define

$$x_{k+1} = T(x_k).$$

In the following theorem, we assume that the dimension of  $\text{Ran}(A)$  is finite. For finite-dimensional matrix equations, this is always true. More generally, this assumption means that there are only finitely many nonzero singular values in the singular system of  $A$ , and we can write  $Ax$  as

$$Ax = \sum_{j=1}^N \lambda_j \langle v_j, x \rangle u_j.$$

We are now ready to prove the following result.

**Theorem 2.9.** *Let  $\dim(\text{Ran}(A)) = N < \infty$  and let  $0 < \beta < 2/\lambda_1^2$ , where  $\lambda_1$  is the largest singular value of  $A$ . Then the fixed point iteration sequence  $(x_k)$  converges to an  $x \in \text{Ker}(A)^\perp$  which satisfies the normal equations (2.16).*

*Proof:* Let  $S = \text{Ker}(A)^\perp = \overline{\text{Ran}(A^*)}$ . First we observe that  $S$  is an invariant set for the affine mapping  $T$  given by the formula (2.17), i.e.,  $T(S) \subset S$ . We show that the mapping  $T$  is a contraction on  $S$ . Indeed, if  $(v_n, u_n, \lambda_n)$  is the singular system of  $A$ , then for any  $x, z \in S = \text{sp}\{v_1, \dots, v_N\}$ , we have

$$\begin{aligned} \|T(x) - T(z)\|^2 &= \|(1 - \beta A^*A)(x - z)\|^2 \\ &= \sum_{j=1}^N (1 - \beta \lambda_j^2)^2 \langle v_j, x - z \rangle^2 \leq \kappa^2 \|x - z\|^2, \end{aligned}$$

where

$$\kappa^2 = \max_{1 \leq j \leq N} (1 - \beta \lambda_j^2)^2.$$

We observe that  $\kappa < 1$  provided that for all  $j$ ,  $1 \leq j \leq N$ ,

$$0 < \beta \lambda_j^2 < 2,$$

which holds true when  $0 < \beta < 2/\lambda_1^2$ .

Let  $x = \lim x_n$ . We have

$$0 = T(x) - x = \beta(A^*y - A^*Ax),$$

i.e., the limit satisfies the normal equations (2.16). □

In general, when  $\dim(\text{Ran}(A)) = \infty$  and  $A$  is compact, we cannot hope that the Landweber–Fridman iteration converges because the normal equations do not, in general, have a solution. This does not prevent us from using the iteration provided that we truncate it after finitely many steps.

To understand the regularization aspect of this iterative scheme, let us introduce

$$R = 1 - \beta A^* A : S \rightarrow S.$$

Inductively, we see that the  $k$ th iterate  $x_k$  can be written simply as

$$x_k = \sum_{j=0}^k R^j \beta A^* y,$$

and in particular,

$$\langle x_k, v_n \rangle = \sum_{j=0}^k \beta \lambda_n (1 - \beta \lambda_n^2)^j \langle y, u_n \rangle = \frac{1}{\lambda_n} (1 - (1 - \beta \lambda_n^2)^{k+1}) \langle y, u_n \rangle$$

by the geometric series sum formula. From this formula it is evident that when the singular value  $\lambda_n$  is small, the factor  $(1 - (1 - \beta \lambda_n^2)^{k+1}) < 1$  in the numerator is also small. Therefore, one can expect that the sum  $\sum \langle x_k, v_n \rangle$  is less sensitive to noise in  $y$  than the minimum norm solution.

When iterative methods are applied for regularization, the crucial issue is to equip the algorithm with a good stopping criterion. It should be pointed out that none of the criteria proposed in the literature has been proved to be failproof. Similar to the TSVD and Tikhonov regularization, one can try to apply also here the discrepancy principle and stop the iterations when

$$\|Ax_k - y\| = \varepsilon, \quad (2.18)$$

where  $\varepsilon$  is the estimated noise level.

We illustrate the behavior of this method with the stopping criterion just described in the following simple example.

**Example 6:** Consider the one-dimensional deconvolution problem of finding  $f(t)$ ,  $0 \leq t \leq 1$  from noisy observations of the function

$$g(s) = \int_0^1 \phi(s-t) f(t) dt, \quad 0 \leq s \leq 1,$$

where the convolution kernel is

$$\phi(t) = e^{-a|t|}, \quad a = 20.$$

As a test function, we use

$$f(t) = t(1-t).$$

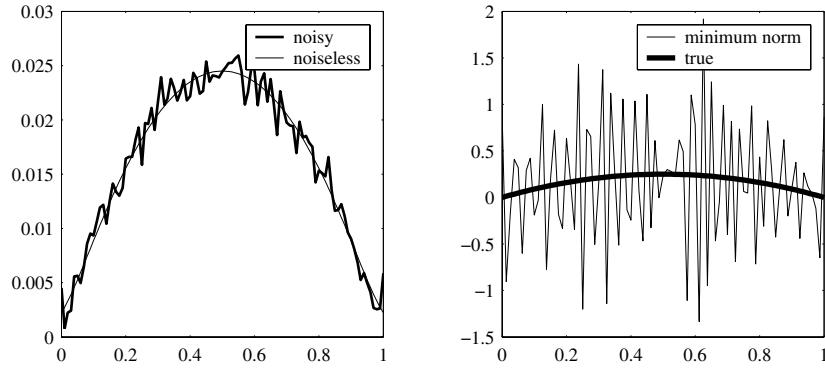
The function  $g$  can be computed analytically, yielding

$$g(s) = \frac{2}{a}s(1-s) + \frac{1}{a^2}(e^{-as} + e^{-a(1-s)}) + \frac{2}{a^3}(e^{-as} + e^{-a(1-s)} - 2).$$

The data is recorded on an even mesh with mesh size  $1/100$ . Random normally distributed zero mean noise is added to the exact data with independent components and standard deviation 5% of the maximum value of the noiseless data. The reconstruction mesh is also an equispaced mesh with mesh size  $1/80$ . The matrix  $A$  has entries

$$A_{ij} = \frac{1}{80}e^{-a|t_i - s_j|}, \quad t_i = \frac{i}{80}, \quad s_j = \frac{j}{100}, \quad 0 \leq i \leq 80, \quad 0 \leq j \leq 100.$$

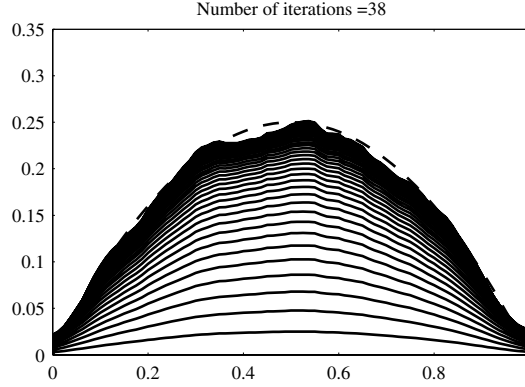
The condition number of the matrix  $A$  is  $\kappa(A) \approx 110$ , so it is possible to calculate directly the minimum norm solution  $f^\dagger = K^\dagger g$ . In Figure 2.8, the noiseless and noisy data are displayed, as well as the exact  $f$  and the minimum norm solution  $f^\dagger$ . The latter is essentially pure noise, showing that some form of regularization is required. We apply the Landweber–Fridman iteration. The relaxation parameter of the iterative scheme was chosen as  $\beta = 0.1\beta_{\max}$ , where  $\beta_{\max} = 2/\|A\|^2$ . We terminate the iteration according to the stopping criterion (2.18) with  $\varepsilon = \sqrt{81}\sigma$ ,  $\sigma$  being the standard deviation of the added noise. With the simulated data, the requested discrepancy level is attained after 38 iterations. In Figure 2.9, the final solution is displayed against the true solution (left). To get an idea of the convergence, we also display the iterated solutions  $f_n$  (right).  $\diamond$



**Figure 2.8.** Noisy and noiseless data and the minimum norm solution.

Usually, the Landweber–Fridman iteration progresses much slower than several other iterative methods. The slow convergence of the method is sometimes argued to be a positive feature of the algorithm, since a fast progress would bring us quickly close to the minimum norm solution that is usually nonsense, as in the previous example.





**Figure 2.9.** Iterated solutions. The final solution satisfies the discrepancy criterion.

### 2.4.2 Kaczmarz Iteration and ART

The idea of the Kaczmarz iteration to solve the matrix equation (2.1),  $Ax = y$ , is to partition the system rowwise, either into single rows or into blocks of rows. For the sake of definiteness, we consider first the single-row version. Writing

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad a_j \in \mathbb{R}^n,$$

where  $a_j^T \neq 0$  is the  $j$ th row of the matrix  $A$ , the equation  $Ax = y$  can be thought of as a system of equations

$$A_j x = a_j^T x = y_j, \quad 1 \leq j \leq m,$$

where  $A_j : \mathbb{R}^n \rightarrow \mathbb{R}$ . Each of these underdetermined equations define a hyperplane of dimension  $n - 1$ . The idea of the Kaczmarz iteration is to project the current approximate solution successively onto each one of these hyperplanes. It turns out that such a procedure converges to the solution of the system, provided that a solution exists.

More generally, we may write

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_\ell \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad A_j \in \mathbb{R}^{k_j \times n},$$

where  $k_1 + \dots + k_\ell = m$ . In this block decomposition of  $A$ , we must require that each row block  $A_j$  has full row rank and defines thus a surjective mapping.

In the following discussion, we generalize the setting slightly. Let us denote by  $H$  and  $H_j$ ,  $1 \leq j \leq m$  denote separable Hilbert spaces. We consider the system

$$A_j x = y_j, \quad 1 \leq j \leq m,$$

where the operators

$$A_j : H \rightarrow H_j, \quad 1 \leq j \leq m$$

are given linear bounded operators and  $y_j \in \text{Ran}(A_j)$ . Let

$$X_j = \{x \in H \mid A_j x = y_j\},$$

and denote by  $P_j : H \rightarrow X_j$  the orthogonal projectors onto these affine subspaces. Furthermore, we define the sequential projection

$$P = P_m P_{m-1} \cdots P_2 P_1.$$

The following definition essentially defines the Kaczmarz iteration.

**Definition 2.10.** *With the notations introduced above, we define the Kaczmarz sequence  $(x_k) \subset H$  recursively as*

$$x_{k+1} = P x_k, \quad x_0 = 0.$$

The following theorem is helpful in understanding the behavior of the Kaczmarz iteration.

**Theorem 2.11.** *Assume that  $X = \bigcap_{j=1}^m X_j \neq \emptyset$ . Then the Kaczmarz sequence converges to the minimum norm solution of the equation  $Ax = y$ , i.e.,*

$$\lim_{k \rightarrow \infty} x_k = x, \quad Ax = y, \quad x \perp \text{Ker}(A).$$

To sketch the main idea of the proof, let us denote by  $\mathcal{Q}$  the orthogonal projection

$$\mathcal{Q} : H \rightarrow \bigcap_{j=1}^m \text{Ker}(A_j) = \text{Ker}(A),$$

and let  $z \in X$  be arbitrary. We shall prove that

$$x_k \longrightarrow x = z - \mathcal{Q}z, \quad \text{as } k \rightarrow \infty.$$

Clearly, this limit  $x$  satisfies

$$A_j x = A_j z - A_j \mathcal{Q}z = y_j, \quad 1 \leq j \leq m,$$

and furthermore,  $x$  is by definition perpendicular to  $\text{Ker}(A)$ .

To relate the partial projections  $P_j$  to  $\mathcal{Q}$ , let us denote by  $Q_j$  the orthogonal projections

$$Q_j : H \rightarrow \text{Ker}(A_j), \quad 1 \leq j \leq m,$$

and by  $Q$  the sequential projection

$$Q = Q_m Q_{m-1} \cdots Q_2 Q_1. \tag{2.19}$$

For any  $z \in X$ , we have

$$P_j x = z + Q_j(x - z).$$

Indeed,

$$A_j P_j x = A_j z + A_j Q_j(x - z) = y_j,$$

and for arbitrary  $z_1, z_2 \in X_j$ , the difference  $\delta z = z_1 - z_2$  is in  $\text{Ker}(A_j)$ . Therefore, it follows that

$$\langle x - (z + Q_j(x - z)), \delta z \rangle = \langle (1 - Q_j)(x - z), \delta x \rangle = 0.$$

Now we may write the sequential projection  $P$  in terms of  $Q$  as follows. For  $z \in X$ , and  $x \in H$ , we have

$$P_2 P_1 x = z + Q_2(P_1 x - z) = z + Q_2 Q_1(x - z),$$

and inductively,

$$P x = z + Q(x - z).$$

Similarly, it holds that

$$P^2 x = z + Q(P x - z) = z + Q^2(x - z),$$

and again inductively,

$$P^k x = z + Q^k(x - z),$$

i.e., by the definition of the Kaczmarz sequence, we have

$$x_k = z - Q^k z.$$

Hence, it suffices to show that for any  $z \in H$ , we have

$$\lim_{k \rightarrow \infty} Q^k z = Qz.$$

This result is the consequence of the following three technical lemmas.

**Lemma 2.12.** *Let  $(x_k) \subset H$  be a sequence satisfying*

$$\|x_k\| \leq 1, \quad \lim_{k \rightarrow \infty} \|Qx_k\| = 1,$$

*where  $Q$  is given by (2.19). Then*

$$\lim_{k \rightarrow \infty} (1 - Q)x_k = 0.$$

*Proof:* The proof is given by induction on the number of the projections  $Q_j$ . For  $Q_1$ , the claim is immediate since, by orthogonality,

$$\|(1 - Q_1)x_k\|^2 = \|x_k\|^2 - \|Q_1 x_k\|^2 \leq 1 - \|Q_1 x_k\|^2 \rightarrow 0,$$

as  $k$  increases.

Next, assume that the claim holds for  $Q_j \cdots Q_1$ . We have

$$\|Q_{j+1}Q_j \cdots Q_1x_k\| \leq \|Q_j \cdots Q_1x_k\| \leq 1,$$

so  $\lim_{k \rightarrow \infty} \|Q_{j+1}Q_j \cdots Q_1x_k\| = 1$  implies  $\lim_{k \rightarrow \infty} \|Q_j \cdots Q_1x_k\| = 1$ , and the induction assumption implies that

$$(1 - Q_j \cdots Q_1)x_k \rightarrow 0.$$

We write

$$(1 - Q_{j+1}Q_j \cdots Q_1)x_k = (1 - Q_j \cdots Q_1)x_k + (1 - Q_{j+1})Q_j \cdots Q_1x_k.$$

Here, the first term on the right tends to zero as we have seen. Similarly, by denoting  $z_k = Q_j \cdots Q_1x_k$ , it holds that

$$\|z_k\| = \|Q_j \cdots Q_1x_k\| \leq 1$$

and

$$\|Q_{j+1}z_k\| = \|Q_{j+1}Q_j \cdots Q_1x_k\| \rightarrow 1,$$

proving that the second term tends also to zero.  $\square$

**Lemma 2.13.** *We have*

$$\text{Ker}(1 - Q) = \text{Ker}(1 - \mathcal{Q}) = \bigcap_{j=1}^m \text{Ker}(A_j).$$

*Proof:* Let  $x \in \text{Ker}(1 - \mathcal{Q})$ . Then  $x \in \text{Ker}(A_j)$  for all  $j$  and so  $x = Q_jx$ , implying that  $x = Q_m \cdots Q_1x = Qx$ , i.e.,  $x \in \text{Ker}(1 - Q)$ .

To prove the converse inclusion  $\text{Ker}(1 - Q) \subset \text{Ker}(1 - \mathcal{Q})$ , assume that  $x = Qx$ . We have

$$\|x\| = \|Q_m \cdots Q_2Q_1x\| \leq \|Q_1x\| \leq \|x\|,$$

i.e.,  $\|Q_1x\| = \|x\|$ . By the orthogonality,

$$\|(1 - Q_1)x\|^2 = \|x\|^2 - \|Q_1x\|^2 = 0,$$

i.e.,  $x = Q_1x$ . Hence,  $x = Q_m \cdots Q_2x$ . Inductively, we show that  $x = Q_jx$  for all  $j$ , i.e.,  $x \in \bigcap_{j=1}^m \text{Ker}(A_j) = \text{Ker}\mathcal{Q}$ .  $\square$

We have the following decomposition result.

**Lemma 2.14.** *Assume that  $Q : H \rightarrow H$  is linear and  $\|Q\| \leq 1$ . Then  $H$  can be decomposed into orthogonal subspaces,*

$$H = \text{Ker}(1 - Q) \oplus \overline{\text{Ran}(1 - Q)}.$$

*Proof:* Since the decomposition claim 1 of Proposition 2.1 is valid for all continuous linear operators, not just for compact ones (see Appendix A), it suffices to show that  $\text{Ker}(1-Q) = \text{Ker}(1-Q^*)$ . Assume therefore that  $Qx = x$ . It follows that

$$\begin{aligned}\|x - Q^*x\|^2 &= \|x\|^2 - 2\langle x, Q^*x \rangle + \|Q^*x\|^2 \\ &= \|x\|^2 - 2\langle Qx, x \rangle + \|Q^*x\|^2 \\ &= -\|x\|^2 + \|Q^*x\|^2 \leq -\|x\|^2 + \|x\|^2 = 0,\end{aligned}$$

implying that  $x = Q^*x$ .

The converse inclusion  $\text{Ker}(1-Q^*) \subset \text{Ker}(1-Q)$  follows similarly.  $\square$

Now we are ready to prove Theorem 2.11.

*Proof of Theorem 2.11:* As we saw, it suffices to prove that

$$\lim_{j \rightarrow \infty} Q^j x = Qx.$$

Since  $\|Q\| \leq 1$ , the decomposition result of the previous lemma holds. For any  $x \in H$ , it follows from Lemma 2.13 that  $Qx \in \text{Ker}(1-Q) = \text{Ker}(1-Q)$ , hence

$$x = Qx + (1-Q)x \in \text{Ker}(1-Q) \oplus \overline{\text{Ran}(1-Q)},$$

and furthermore,

$$Q^k x = Qx + Q^k z, \quad z = (1-Q)x \in \overline{\text{Ran}(1-Q)}.$$

Hence we need to show that  $Q^k z \rightarrow 0$  for every  $z \in \overline{\text{Ran}(1-Q)}$ . Assume first that  $z \in \text{Ran}(1-Q)$ , or  $z = (1-Q)y$  for some  $y \in H$ . Consider the sequence  $c_k = \|Q^k y\|$ . This sequence is decreasing and positive. Let  $c = \lim c_k$ . If  $c = 0$ , then

$$Q^k z = Q^k y - Q^{k+1} y \rightarrow 0,$$

as claimed. Assume next that  $c > 0$ , and define the sequence

$$y_k = \frac{Q^k y}{c_k},$$

having the properties

$$\|y_k\| = 1, \quad \lim \|Qy_k\| = 1.$$

By Lemma 2.12, we have  $\lim(1-Q)y_k = 0$ , or

$$Q^k z = Q^k y - Q^{k+1} y = c_k(1-Q)y_k \rightarrow 0.$$

This result extends also to the closure of  $\text{Ran}(1-Q)$ . If  $z \in \overline{\text{Ran}(1-Q)}$ , we choose  $z_0 \in \text{Ran}(1-Q)$  with  $\|z - z_0\| < \varepsilon$ , for arbitrary  $\varepsilon > 0$ . Then

$$\|Q^k z\| \leq \|Q^k(z - z_0)\| + \|Q^k z_0\| < \varepsilon + \|Q^k z_0\| \rightarrow \varepsilon,$$

i.e., we must have  $\lim_{k \rightarrow \infty} \|Qz\| = 0$ . This completes the proof.  $\square$

Finally, we discuss the implementation of the Kaczmarz iteration in finite-dimensional spaces. The iterative algorithm that we present is commonly used especially in tomographic applications. The following lemma gives the explicit form of the projections  $P_j$  appearing in the algorithm.

**Lemma 2.15.** *Let  $A_j \in \mathbb{R}^{k_j \times n}$  be a matrix such that the mapping  $A_j : \mathbb{R}^n \rightarrow \mathbb{R}^{k_j}$  is surjective. For  $y_j \in \mathbb{R}^{k_j}$ , the orthogonal projection  $P_j$  to the affine subspace  $X_j$  is given by the formula*

$$P_j x = x + A_j^T (A_j A_j^T)^{-1} (y_j - A_j x). \quad (2.20)$$

*Proof:* We observe first that the matrix  $A_j A_j^T$  is invertible. From the surjectivity of the mapping  $A_j$ ,

$$\mathbb{R}^{k_j} = \text{Ran}(A_j) = \text{Ker}(A_j^T)^\perp,$$

i.e., the mapping defined by the matrix  $A_j^T$  and consequently  $A_j A_j^T$  are injective. Furthermore, if  $z \perp \text{Ran}(A_j A_j^T)$ , we have in particular that

$$0 = z^T A_j A_j^T z = \|A_j^T z\|^2,$$

so  $z = 0$  by the injectivity of  $A_j^T$ .

As before, we may express  $P_j$  in terms of the projection  $Q_j$  as

$$P_j x = z_j + Q_j(x - z_j), \quad z_j \in X_j.$$

Since

$$x - P_j x = (1 - Q_j)(x - z_j) \in \text{Ker}(A_j)^\perp = \text{Ran}(A_j^T),$$

there is a  $u \in \mathbb{R}^{k_j}$  such that

$$x - P_j x = A_j^T u. \quad (2.21)$$

Multiplying both sides by  $A_j$ , we obtain

$$A_j A_j^T u = A_j x - y_j,$$

hence

$$u = (A_j A_j^T)^{-1} (A_j x - y_j).$$

Substituting this expression for  $u$  into formula (2.21) proves the claim.  $\square$

**Remark:** The Kaczmarz iteration allows a slightly more general form than the one given above. Instead of the projections  $P_j$ , one can use  $P_{j\omega} = (1 - \omega)I + \omega P_j$ , where  $\omega$  is a relaxation parameter,  $0 < \omega < 2$ . The proofs above hold also in this more general setting, too. The formula (2.20) takes the form

$$P_{j\omega} x = x + \omega A_j^T (A_j A_j^T)^{-1} (y_j - A_j x).$$

**Example 7:** Probably the most typical application of the Kaczmarz iteration to inverse problems is in X-ray tomography.<sup>2</sup> Here we consider the two-dimensional discretized problem. The tomography data consists of projections, or shadow images, of the image into given directions. These projections can be described in terms of a linear operator that is approximated by a matrix. Thus, let  $x \in \mathbb{R}^{N^2}$  be a vector containing the stacked pixel values of an  $N \times N$  image, and  $A \in \mathbb{R}^{M \times N^2}$  the sparse tomography matrix. We apply the Kaczmarz iteration row by row. Let

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_M^T \end{bmatrix} \in \mathbb{R}^{M \times N^2}, \quad a_j \in \mathbb{R}^{N^2}.$$

The iterative scheme to solve the equation  $Ax = y$ , known in the context of X-ray tomography as the *algebraic reconstruction technique*, or ART for short, proceeds as follows:

```

Set  $k = 0$ ,  $x_0 = 0$ ;
Repeat until convergence:
     $z_0 = x_k$ ;
    for  $j = 1 : M$  repeat
         $z_j = z_{j-1} + (1/\|a_j\|^2)(y_j - a_j^T z_{j-1})a_j$ ;
    end
     $x_{k+1} = z_M$ ;  $k \leftarrow k + 1$ ;
end

```

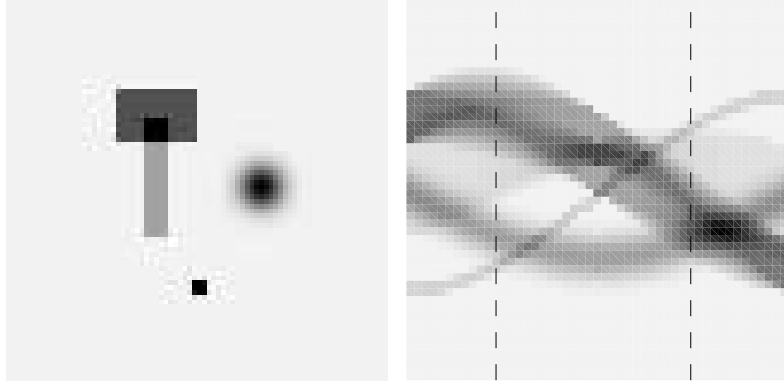
To illustrate this algorithm, we apply it to both *full angle data* and *limited angle data*. The data is best understood by considering Figure 2.10. The original image is pixelized into  $80 \times 80$  pixels. First, we compute the full angle data. The data consists of one-dimensional shadow images of the true image in different directions of projection. Let the projection angle vary over an interval of length  $\pi$ . We discretize the arc of a semicircle into 40 equal intervals and increase the look angle by  $\pi/40$  each step, yielding to discrete angles  $\phi_i$ ,  $1 \leq i \leq 40$ . The line of projection is divided into 41 intervals. Hence, the projection data has the size  $40 \times 41$ . This data is the full angle data.

In Figure 2.10 this data without added noise is plotted with the angular variable on the horizontal axis. This representation is called the *sinogram* for rather obvious reasons. Now we apply the ART algorithm. To avoid an inverse crime, we use a different grid for the reconstruction. We seek to find an image in a pixel map of size  $50 \times 50$ . We add noise to the sinogram data by adding to each data entry independent uniformly distributed noise drawn from the interval  $[0, \sigma]$ , where  $\sigma$  is 2% of the maximum value of the data matrix.

In Figure 2.11 we display three ART reconstructions: The first one is after one iteration round; the second is the one that satisfies the discrepancy condition

---

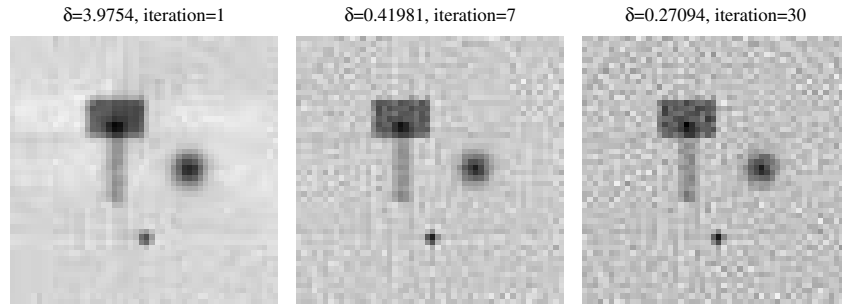
<sup>2</sup>The X-ray tomography is discussed further in Chapter 6.



**Figure 2.10.** Original image and the sinogram data, the abscissa being the illumination angle. The limited angle data is the part of the sinogram between the vertical dashed lines.

$$\|Ax_j - y\| \leq \varepsilon = 50\sigma,$$

where  $\sigma$  is the standard deviation of the additive noise, and the factor 50 comes from the image size (see Example 4). Finally, the third reconstruction corresponds to 30 iterations. Evidently, the full angle data is so good that already after one single iteration the reconstruction is visually rather satisfactory. In fact, one can see some slight artifacts in the 30 iterations image.

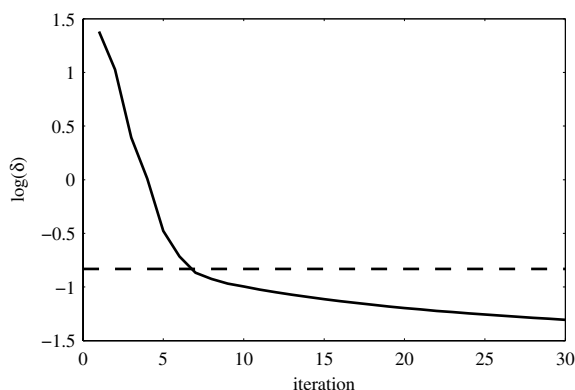


**Figure 2.11.** ART reconstruction from the full angle tomography data.

To get an idea of the convergence of the ART algorithm, we have also plotted the discrepancies in Figure 2.12

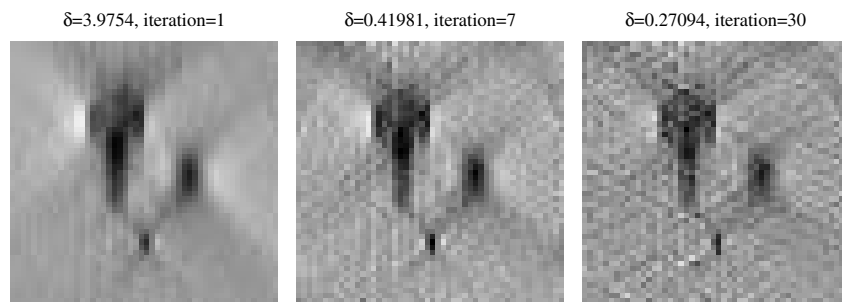
Now we repeat the computation starting with limited angle data. We assume that the look angle varies from  $-\pi/4$  to  $\pi/4$  around the vertical direction, i.e., the image is illuminated from below in an angle of  $\pi/2$  opening. We do this by discarding from the full angle data those illumination lines where





**Figure 2.12.** The full angle discrepancies. The estimated noise level is marked with a dashed line.

look angle differs from the vertical more than  $\pi/4$ , i.e., we use only the central part of the sinogram data. Figure 2.13 displays the ART reconstructions with limited angle data analogous to those ones with full angle data. The fact that no horizontal or close to horizontal integration lines are included is reflected in the reconstructions that show long shadows in directions close to vertical. These reconstructions demonstrate the limitation of ART (or in fact, any inversion scheme) when data is scarce and no additional or prior information is used in the reconstruction.  $\diamond$



**Figure 2.13.** ART reconstruction from the limited angle tomography data.

### 2.4.3 Krylov Subspace Methods

The Krylov subspace methods refer to a class of iterative solvers of large linear equations of the form  $Ax = y$ . Roughly, the idea is to produce a sequence of approximate solutions as linear combinations of vectors of the type

$u, Au, A^2u, \dots$ . The best known of these methods when the matrix  $A$  is symmetric and positive definite is the *conjugate gradient method* (CG). Here, we restrict the discussion to that method.

In the sequel, we assume that  $A \in \mathbb{R}^{n \times n}$  is a symmetric and strictly positive definite matrix, i.e.,

$$A^T = A, \quad u^T A u > 0 \text{ for } u \neq 0.$$

In particular, all the eigenvalues of  $A$  must be positive, and hence matrix  $A$  is invertible. The objective of the CG method is to find an approximating sequence  $(x_j)$  converging to the solution of the equation  $Ax = y$  by solving a sequence of minimization problems. Let us denote by

$$x_* = A^{-1}y$$

the exact solution, and denote by  $e$  and  $r$  the error and the residual of a given approximation  $x$ ,

$$e = x_* - x, \quad r = y - Ax = Ae.$$

Consider the quadratic functional

$$\phi(x) = e^T Ae = r^T A^{-1}r.$$

It is not possible to calculate the value of this functional for a given  $x$  without the knowledge of the exact solution  $x_*$  or, alternatively,  $A^{-1}$ . However, it is possible to consider the problem of minimizing this functional over a nested sequence of Krylov subspaces. First, let us observe that by the positive definiteness of  $A$ ,

$$\phi(x) = 0 = \min_{x \in \mathbb{R}^n} \phi(x) \text{ if and only if } x = x_*.$$

Assume that we have an initial guess  $x_1$  and an initial direction  $s_1$ , and we consider the problem of minimizing the function

$$\mathbb{R} \rightarrow \mathbb{R}, \quad \alpha \mapsto \phi(x_1 + \alpha s_1).$$

Interestingly, we can solve this minimization problem without knowing the value of  $\phi$ .

**Lemma 2.16.** *The function  $\alpha \mapsto \phi(x_1 + \alpha s_1)$  has a minimum at*

$$\alpha = \alpha_1 = \frac{s_1^T r_1}{s_1^T A s_1},$$

where  $r_1$  is the residual of the initial guess,

$$r_1 = y - Ax_1.$$

*Proof:* The residual corresponding to  $x = x_1 + \alpha s_1$  is

$$y - Ax = y - Ax_1 - \alpha As_1 = r_1 - \alpha As_1,$$

and so

$$\begin{aligned}\phi(x) &= (r_1 - \alpha As_1)^T A^{-1} (r_1 - \alpha As_1) \\ &= \alpha^2 s_1^T As_1 - 2\alpha s_1^T r_1 + r_1^T A^{-1} r_1.\end{aligned}$$

The claim follows immediately from this formula.  $\square$

Hence, given a sequence  $(s_k)$  of directions, we may produce a sequence  $(x_k)$  of approximate solutions by setting

$$x_{k+1} = x_k + \alpha_k s_k, \quad \alpha_k = \frac{s_k^T r_k}{s_k^T A s_k}, \quad (2.22)$$

where  $r_k$  is the residual of the previous iterate, i.e.,

$$r_k = y - Ax_k.$$

Note that the residuals in this scheme are updated according to the formula

$$r_{k+1} = y - A(x_k + \alpha_k s_k) = r_k - \alpha_k A s_k.$$

This procedure can be carried out with any choice of the search directions  $s_k$ . The conjugate gradient method is characterized by a particular choice of the search directions. We give the following definition.

**Definition 2.17.** We say that the linearly independent vectors  $\{s_1, \dots, s_k\}$  are  $A$ -conjugate, if

$$s_i^T A s_j = 0 \text{ for } i \neq j,$$

i.e., the vectors are orthogonal with respect to the inner product defined by the matrix  $A$ ,

$$\langle u, v \rangle_A = u^T A v.$$

Observe that if a given set of vectors  $\{u_1, \dots, u_k\}$  are linearly independent, it is always possible to find  $A$ -conjugate vectors  $v_j \in \text{sp}\{u_1, \dots, u_k\}$ ,  $1 \leq j \leq k$  so that  $\text{sp}\{u_1, \dots, u_k\} = \text{sp}\{v_1, \dots, v_k\}$ . This can be done, e.g., by the Gram–Schmidt orthogonalization process with respect to the inner product  $\langle \cdot, \cdot \rangle_A$ .

Introduce the matrix  $S_k = [v_1, \dots, v_k] \in \mathbb{R}^{n \times k}$ ; then the  $A$ -conjugacy of the vectors  $\{v_j\}$  is equivalent to

$$S_k^T A S_k = D_k = \text{diag}(d_1, \dots, d_k) \in \mathbb{R}^{k \times k}$$

with  $d_j \neq 0$ ,  $1 \leq j \leq k$ .

To understand the significance of using  $A$ -conjugate directions, consider the following *global* minimization problem: given the matrix  $S = [s_1, \dots, s_k]$  with linearly independent columns, find a minimum of the mapping

$$\mathbb{R}^k \rightarrow \mathbb{R}, \quad h \mapsto \phi(x_1 + S_k h),$$

i.e., seek to minimize  $\phi(x)$  not sequentially over each given directions but over the whole subspace in one single step. The following result is analogous to Lemma 2.16.

**Lemma 2.18.** *The function  $h \mapsto \phi(x_1 + S_k h)$  attains its minimum at*

$$h = (S_k^T A S_k)^{-1} S_k^T r_1. \quad (2.23)$$

*Proof:* We observe first that the matrix  $S_k^T A S_k$  is invertible. Indeed, if  $S_k^T A S_k x = 0$ , then also  $x^T S_k^T A S_k x = 0$ , and the positive definiteness of  $A$  and the linear independence of the columns of  $S_k$  imply that  $x = 0$ .

Since

$$r = y - A(x_1 + S_k h) = r_1 - A S_k h,$$

we have

$$\begin{aligned} \phi(x_1 + S_k h) &= (r_1 - A S_k h)^T A^{-1} (r_1 - A S_k h) \\ &= h^T S_k^T A S_k h - 2r_1^T S_k h + r_1^T A^{-1} r_1. \end{aligned}$$

The minimum of this quadratic functional satisfies

$$S_k^T A S_k h - S_k^T r_1 = 0,$$

so the claim follows.  $\square$

The computation of the minimizer  $h$  becomes trivial if the matrix  $D_k = S_k^T A S_k$  is diagonal. But this is not the only advantage of using  $A$ -conjugate directions. Assume that the sequential minimizers  $x_1, \dots, x_{k+1}$  have been calculated as given by (2.22). Since  $x_{k+1} \in x_1 + \text{sp}\{s_1, \dots, s_k\}$ , we have

$$\phi(x_{k+1}) \geq \phi(x_1 + S_k h),$$

with  $h \in \mathbb{R}$  given by (2.23). We are now ready to establish the following result.

**Theorem 2.19.** *Assume that the vectors  $\{s_1, \dots, s_k\}$  are linearly independent and  $A$ -conjugate. Then*

$$x_{k+1} = x_1 + S_k h,$$

*i.e., the  $(k+1)$ th sequential minimizer is also the minimizer over the subspace spanned by the directions  $s_j$ ,  $1 \leq j \leq k$ .*

*Proof:* Let  $a_j = [\alpha_1, \dots, \alpha_j]^T$ . With this notation, we have

$$x_j = x_1 + S_{j-1} a_{j-1},$$

and the corresponding residual is

$$r_j = y - A x_j = r_1 - A S_{j-1} a_{j-1}.$$

We observe that by the  $A$ -conjugacy,

$$s_j^T r_j = s_j^T r_1 - s_j^T A S_{j-1} a_{j-1} = s_j^T r_1.$$

Therefore,

$$\alpha_j = \frac{s_j^T r_j}{s_j^T A s_j} = \frac{s_j^T r_1}{s_j^T A s_j} = h_j,$$

i.e., we have  $a_k = h$ . □

As a corollary, we get also the following orthogonality result.

**Corollary 2.20.** *If the vectors  $\{s_1, \dots, s_k\}$  are  $A$ -conjugate and linearly independent, then*

$$r_{k+1} \perp \text{sp}\{s_1, \dots, s_k\}.$$

*Proof:* We have

$$r_{k+1} = y - A x_{k+1} = r_1 - A S_k h,$$

and so

$$r_{k+1}^T S_k = r_1^T S_k - h^T S_k^T A S_k = 0$$

by formula (2.23). □

The results above say that if we are able to choose the next search direction  $s_{k+1}$  to be  $A$ -conjugate with the previous ones, the search for the sequential minimum gives also the global minimum over the subspace. So the question is how to efficiently determine  $A$ -conjugate directions. It is well known that orthogonal polynomials satisfying a three-term recurrence relation could be used effectively to this end. However, it is possible to build an algorithm with quite elementary methods.

**Definition 2.21.** *Let  $r_1 = y - A x_1$ . The  $k$ th Krylov subspace of  $A$  with the initial vector  $r_1$  is defined as*

$$\mathcal{K}_k = \mathcal{K}_k(A, r_1) = \text{sp}\{r_1, A r_1, \dots, A^{k-1} r_1\}, \quad k \geq 1.$$

What is the dimension of  $\mathcal{K}_k$ ? Evidently, if  $r_1$  is an eigenvector of the matrix  $A$ , then  $\dim(\mathcal{K}_k) = 1$  for all  $k$ . More generally, if  $K \subset \mathbb{R}^n$  is an invariant subspace of  $A$  and  $\dim(K) = m$ , then  $r_1 \in K$  implies that  $\mathcal{K}_k \subset K$  and so  $\dim(\mathcal{K}_k) \leq m$ . The implications will be discussed later.

Our aim is to construct the sequence of the search directions inductively. Assume that  $r_1 \neq 0$ , since otherwise  $x_1 = x_*$  and we would be done. Then let  $s_1 = r_1$ .

We proceed by induction on  $k$ . Assume that for some  $k \geq 1$ , we have constructed an  $A$ -conjugate set  $\{s_1, \dots, s_k\}$  of linearly independent search directions such that

$$\text{sp}\{s_1, \dots, s_k\} = \text{sp}\{r_1, \dots, r_k\} = \mathcal{K}_k.$$

With our choice of  $s_1$ , this is evidently true for  $k = 1$ . The goal is to choose  $s_{k+1}$  so that the above conditions remain valid also for  $k + 1$ .

Let  $r_{k+1} = y - Ax_{k+1} = r_k - \alpha_k As_k$ . If  $r_{k+1} = 0$ , we have  $x_k = x_*$  and the search has converged. Assume therefore that  $r_{k+1} \neq 0$ . Since  $r_k, s_k \in \mathcal{K}_k$  by the induction assumption, it follows that  $r_{k+1} \in \mathcal{K}_{k+1}$ . On the other hand, by Corollary 2.20,  $r_{k+1} \perp s_j$  for all  $j$ ,  $1 \leq j \leq k$ , thus

$$\text{sp}\{s_1, \dots, s_k, r_{k+1}\} = \text{sp}\{r_1, \dots, r_{k+1}\} = \mathcal{K}_{k+1}.$$

To ensure that  $s_{k+1}$  is  $A$ -conjugate to the previous search direction, we express it in the form

$$s_{k+1} = r_{k+1} + S_k \beta \in \mathcal{K}_{k+1}, \quad \beta \in \mathbb{R}^k.$$

The coefficient vector  $\beta$  is determined by imposing the  $A$ -conjugacy condition

$$S_k^T As_{k+1} = 0,$$

that is,

$$D_k \beta = S_k^T As_k \beta = -S_k^T Ar_{k+1} = -(As_k)^T r_{k+1}.$$

Here, we have

$$AS_k = [AS_{k-1}, As_k].$$

The columns of the matrix  $AS_{k-1}$  belong all to  $A(\text{sp}\{s_1, \dots, s_{k-1}\}) = A(\mathcal{K}_{k-1}) \subset \mathcal{K}_k = \text{sp}\{s_1, \dots, s_k\}$ , and  $r_{k+1} \perp \text{sp}\{s_1, \dots, s_k\}$ . Therefore,

$$D_k \beta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -s_k^T Ar_{k+1} \end{bmatrix},$$

i.e.,  $\beta_1 = \dots = \beta_{k-1} = 0$ , and we have

$$s_{k+1} = r_{k+1} + \beta_k s_k, \quad \beta_k = -\frac{s_k^T Ar_{k+1}}{s_k^T As_k}.$$

Now we have all the necessary ingredients for the minimization algorithm. However, it is customary to make a small modification to the updating formulas that improves the computational stability of the algorithm. Since  $r_k \perp s_{k-1}$ , we have

$$s_k^T r_k = (r_k + \beta_{k-1} s_{k-1})^T r_k = \|r_k\|^2,$$

i.e., the formula (2.22) can be written as

$$\alpha_k = \frac{\|r_k\|^2}{s_k^T As_k}.$$

Furthermore, since  $r_k \in \text{sp}\{s_1, \dots, s_k\}$ , we have  $r_{k+1} \perp r_k$ , implying that

$$\begin{aligned}
\|r_{k+1}\|^2 &= r_{k+1}^T (r_k - \alpha_k A s_k) = -\frac{\|r_k\|^2}{s_k^T A s_k} r_{k+1}^T A s_k \\
&= \|r_k\|^2 \beta_k,
\end{aligned}$$

i.e., the expression for  $\beta_k$  simplifies to

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Now we are ready to state the CG algorithm.

Pick  $x_1$ . Set  $k = 1$ ,  $r_1 = y - A x_1$ ,  $s_1 = r_1$ ;

Repeat until convergence:

$$\alpha_k = \|r_k\|^2 / s_k^T A s_k;$$

$$x_{k+1} = x_k - \alpha_k s_k;$$

$$r_{k+1} = r_k - \alpha_k A s_k;$$

$$\beta_k = \|r_{k+1}\|^2 / \|r_k\|^2;$$

$$s_{k+1} = r_{k+1} + \beta_k s_k;$$

$$k \leftarrow k + 1;$$

end

Since the conjugate directions are linearly independent, the conjugate gradient algorithm needs at most  $n$  steps to converge. If the initial residual is in an invariant subspace  $K$  of  $A$  with  $\dim(K) = m < n$ , then the algorithm converges at most  $m$  steps. However, when using the conjugate gradient method to solve ill-posed inverse problems, one should not iterate until the residual is zero. Instead, the iterations are terminated e.g., as soon as the norm of the residual is smaller or equal to the estimated norm of the noise.

**Example 8:** We illustrate the use of the conjugate gradient method with the inversion of the Laplace transform. Let the data  $y$  and the matrix  $A$  be as in Examples 2 and 3 of this chapter with 1% normally distributed random noise  $e$  added to the data, i.e., we have

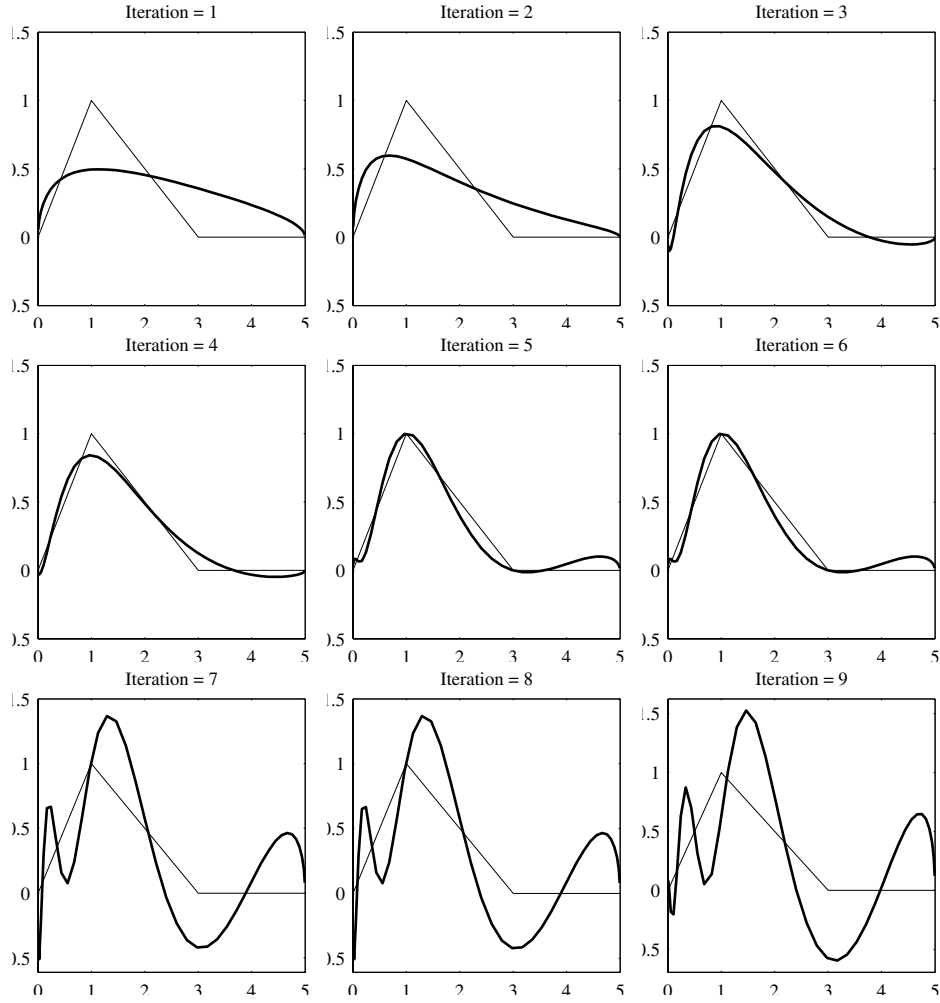
$$y = Ax + e.$$

To write this inverse problem in a form where the conjugate gradient method is applicable, we consider the normal equation,

$$A^T y = A^T A x + A^T e = Bx + \tilde{e}.$$

Observe that although the matrix  $B$  is numerically singular, this does not prevent us from using the method since the iteration process is terminated prior to convergence.

The approximate solutions computed by the conjugate gradient method in the iterations 1–9 from the noisy Laplace transform data are plotted in Figure 2.14. After the seventh iteration, the approximations get very rapidly



**Figure 2.14.** Conjugate gradient approximation after iterations 1–9.

out of control. By visual inspection, one can conclude that few iterations in this case give the best result. Observe that if we want to use the discrepancy principle for determining the stopping index, we can use the residual norm  $\|y - Ax_j\|$  of the original equation in the stopping criterion.  $\diamond$

## 2.5 Notes and Comments

The literature on regularization of inverse problems is extensive. We refer to the textbooks [12], [35], [50], [130] and [139].



The truncated singular value decomposition has been treated widely in the literature. We refer to the book [56] and references therein concerning this topic.

The pseudoinverse  $A^\dagger$  of a matrix  $A \in \mathbb{R}^{n \times m}$  has several properties that sometimes are useful. We mention here the *Moore–Penrose* equations,

$$\begin{aligned} A^\dagger A A^\dagger &= A^\dagger, & A A^\dagger A &= A, \\ (A^\dagger A)^\mathrm{T} &= A^\dagger A, & (A A^\dagger)^\mathrm{T} &= A A^\dagger. \end{aligned}$$

In fact, these equations characterize completely the pseudoinverse. The matrices  $A^\dagger A$  and  $A A^\dagger$  have a geometric interpretation as orthogonal projectors

$$\begin{aligned} A^\dagger A : \mathbb{R}^n &\rightarrow \mathrm{Ker}(A)^\perp, \\ A A^\dagger : \mathbb{R}^m &\rightarrow \mathrm{Ran}(A). \end{aligned}$$

In view of the original works of Tikhonov concerning ill-posed problems (see, e.g., [129]), the term Tikhonov regularization is used somewhat loosely. Tikhonov considered the regularization of Fredholm equations of the first kind by minimizing the functional

$$F(x) = \|Ax - y\|^2 + \alpha^2 \Omega(x)$$

in a function space  $H$ . The penalty functional  $\Omega$  was characterized by the property that the sets

$$\Omega_M = \{x \in H \mid \Omega(x) \leq M\}$$

are precompact in  $H$ . This condition guarantees the existence of the minimizer. In this sense, the Tikhonov regularization as defined here coincides with the original one only when  $H$  is finite-dimensional.

In large-scale inverse problems, the selection of the regularization parameter according to the discrepancy principle may be costly if one relies, e.g., on Newton's method. There are numerically effective methods to do the selection; see, e.g., [20].

In addition to Morozov's discrepancy principle, there are several other selection principles of the regularization parameters. We mention here the L-curve method (see [55]–[56]) and the generalized cross-validation (GCV) method ([39]).

The use of Kaczmarz iteration in tomographic problems has been discussed, e.g., in the book [96], which is a comprehensive representation of this topic in general.

At the end of Subsection 2.4.3, we considered the conjugate gradient iteration for nonsymmetric systems. Usually, when normal equations are considered, one avoids forming explicitly the matrix  $A^\mathrm{T} A$ . Since one works with the matrix  $A$  and its transpose, in comparison with the usual conjugate gradient, the algorithm requires one extra matrix-vector product per iteration. The

algorithm has several acronyms, such as *conjugate gradient normal residual* (CGNR), *conjugate gradient normal equation* (CGNE) or *conjugate gradient least squares* (CGLS) methods. For references, see, e.g., the books [1] or [54].

In addition, for nonsymmetric problems other iterative solvers are available, e.g., *generalized minimal residual* (GMRES) method ([19], [109]). The various method differ from each other in the memory requirements, among other things.

Statistical and Computational Inverse Problems

Kaipio, J.; Somersalo, E.

2005, XVI, 340 p., Hardcover

ISBN: 978-0-387-22073-4