Applied Mathematical Sciences

Volume 160

Jari Kaipio Erkki Somersalo

Statistical and Computational Inverse Problems

With 102 Figures

Springer

Jari P. Kaipio
Department of Applied Physics
University of Kuopio
70211 Kuopio
Finland
Jari.Kaipio@uku.fi

Erkki Somersalo
Institute of Mathematics
Helsinki University of Technology
02015 HUT
Finland
Erkki.Somersalo@hut.fi

Mathematics Subject Classification (2000): 15A29, 35R25, 35R30, 62F15, 62P30, 62P35, 62K05, 65C05, 65C40, 65J20, 65J22, 65F10, 65F22, 65N21, 65R30, 65R32, 78A70

Library of Congress Cataloging-in-Publication Data
Kaipio, Jari.
Statistical and computational inverse problems / Jari Kaipio and Erkki Somersalo.
p. cm. — (Applied mathematical sciences)
"April 1, 2004."
Includes bibliographical references and index.
ISBN 0-387-22073-9 (alk. paper)
1. Inverse problems (Differential equations) 2. Inverse problems (Differential equations)—Numerical solutions. I. Somersalo, Erkki. II. Title. III. Applied mathematical sciences (Springer-Verlag New York, Inc.)
QA377.K33 2004
515′.357—dc22 2004049181

ISBN 0-387-22073-9 Printed on acid-free paper.

Printed in the United States of America. (MP)

9 8 7 6 5 4 3 2 1 SPIN 10980149

springeronline.com

To Eerika, Paula, Maija and Noora

Between the idea
   And the reality
Between the motion
   And the act
      Falls the Shadow

T.S. Eliot

Preface

This book is aimed at postgraduate students in applied mathematics as well as at engineering and physics students with a firm background in mathematics. The first four chapters can be used as the material for a first course on inverse problems with a focus on computational and statistical aspects. On the other hand, Chapters 3 and 4, which discuss statistical and nonstationary inversion methods, can be used by students already having knowldege of classical inversion methods.

There is rich literature, including numerous textbooks, on the classical aspects of inverse problems. From the numerical point of view, these books concentrate on problems in which the measurement errors are either very small or in which the error properties are known exactly. In real-world problems, however, the errors are seldom very small and their properties in the deterministic sense are not well known. For example, in classical literature the error norm is usually assumed to be a known real number. In reality, the error norm is a random variable whose mean might be known.

Furthermore, the classical literature usually assumes that the operator equations that describe the observations are exactly known. Again, usually when computational solutions based on real-world measurements are required, one should take into account that the mathematical models are themselves only approximations of real-world phenomena. Moreover, for computational treatment of the problem, the models must be discretized, and this introduces additional errors. Thus, the discrepancy between the measurements and the predictions by the observation model are not only due to the "noise that has been added to the measurements." One of the central topics in this book is the statistical analysis of errors generated by modelling.

There is rich literature also in statistics, especially concerning Bayesian statistics, that is fully relevant in inverse problems. This literature has been fairly little known to the inverse problems community, and thus the main aim of this book is to introduce the statistical concepts to this community. As for statisticians, the book contains probably little new information regarding, for example, sampling methods. However, the development of realistic observation models based, for example, on partial differential equations and the analysis of the associated modelling errors might be useful.

As for citations, in Chapters 1–6 we mainly refer to books for further reading and do not discuss historical development of the topics. Chapter 7, which discusses our previous and some new research topics, also does not contain reviews of the applications. Here we refer mainly to the original publications as well as to sources that contain modifications and extensions which serve to illustrate the potential of the statistical approach.

Chapters 5–7, which form the second part of the book, focus on problems for which the models for measurement errors, errorless observations and the unknown are really taken as *models*, which themselves may contain uncertainties. For example, several observation models are based on partial differential equations and boundary value problems. It might be that part of the boundary value data are inherently unknown. We would then attempt to model these boundary data as random variables that could either be treated as secondary unknowns or taken as a further source of uncertainty and compute its contribution to the discrepancy between the observation model and the predictions given by the observation model.

In the examples, especially in Chapter 7 that discusses nontrivial problems, we concentrate on research that we have carried out earlier. However, we also treat topics that either have not yet been published or are discussed here with more rigor than in the original publications.

We have tried to enhance the readibility of the book by avoiding citations in the main text. Every chapter has a section called “Notes and Comments” where the citations and further reading, as well as brief comments on more advanced topics, are given.

We are grateful to our colleague and friend, Markku Lehtinen, who has advocated the statistical approach to inverse problems for decades and brought this topic to our attention. Much of the results in Chapter 7 have been done in collaboration with our present and former graduate students - as well as other scientists. We have been privileged to work with them and thank them all. We mention here only the people who have contributed directly to this book by making modifications to their computational implementations or otherwise: Dr. Ville Kolehmainen for Sections 7.2 and 7.9, Dr. Arto Voutilainen for Section 7.4, Mr. Aku Seppänen for Sections 7.5 and 7.7 and Ms. Jenni Heino for Section 7.8. We are also much obliged to Daniela Calvetti for carefully reading and commenting the whole manuscript and to the above-mentioned people for reading some parts of the book. For possible errors that remain we assume full responsibility.

This work was financially supported by the Academy of Finland and the Finnish Academy of Science and Letters (JPK) to whom thanks are due. Thanks are also due to the inverse problems group at the University of Kuopio and to vice head of the Applied Physics department, Dr. Ari Laaksonen, who saw to the other author’s duties during his leave. Thanks are also due to Dr. Geoff Nicholls and Dr. Colin Fox from the University of Auckland, NZ, where

much of the novel material in this book was conceived during the authors' visits there.

Helsinki and Kuopio
June 2004

Jari P. Kaipio
Erkki Somersalo

Contents

1

Inverse Problems and Interpretation of Measurements

Inverse problems are defined, as the term itself indicates, as the inverse of direct or forward problems. Clearly, such a definition is empty unless we define the concept of direct problems. Inverse problems are encountered typically in situations where one makes indirect observations of a quantity of interest. Let us consider an example: one is interested in the air temperature. Temperature itself is a quantity defined in statistical physics, and despite its usefulness and intuitive clarity it is not directly observable. A ubiquitous thermometer that gives us information of the air temperature relies on the fact that materials such as quicksilver expand in a very predictable way in normal conditions as the temperature increases. Here the forward model is the function relating the volume of the quicksilver as a function of the temperature. The inverse problem in this case is trivial, and therefore it is not usually considered as a separate inverse problem at all, namely the problem of determining the temperature from the volume measured. A more challenging inverse problem arises if we try to measure the temperature in a furnace. Due to the high temperature, the traditional thermometer is useless and we have to use more advanced methods. One possibility is to use ultrasound. The high temperature renders the gases in the furnace turbulent, thus changing their acoustic properties which in turn is reflected in the acoustic echoes. Now the forward model consists of the challenging problem of describing the turbulence as a function of temperature plus acoustic wave propagation in the medium, and its even more challenging inverse counterpart of determining the temperature from acoustic observations.

It is the legacy of Newton, Leibniz and others that laws of nature are often expressed as systems of differential equations. These equations are *local* in the sense that at a given point they express the dependence of the function and its derivatives on physical conditions at that location. Another typical feature of the laws is *causality*: later conditions depend on the previous ones. Locality and causality are features typically associated with direct models. Inverse problems on the other hand are most often *nonlocal* and/or *noncausal.* In our example concerning the furnace temperature measurement, the acoustic echo observed outside depends on the turbulence everywhere, and due to the finite signal speed, we can hope to reconstruct the temperature distribution in a time span prior to the measurement, i.e., computationally we try to go upstream in time.

The nonlocality and noncausality of inverse problems greatly contribute to their instability. To understand this, consider heat diffusion in materials. Small changes in the initial temperature distributions smear out in time, leaving the final temperature distribution practically unaltered. The forward problem is then stable as the result is little affected by changes in the initial data.

Going in the noncausal direction, if we try to estimate the initial temperature distribution based on the observed temperature distribution at the final time, we find that vastly different initial conditions may have produced the final condition, at least within the accuracy limit of our measurement. On the one hand, this is a serious problem that requires a careful analysis of the data; on the other hand we need to incorporate all possible information about the initial data that we may have had *prior* to the measurement. The *statistical inversion theory*, which is the main topic of this book, solves the inverse problems systematically in such a way that all the information available is properly incorporated in the model.

Statistical inversion theory reformulates inverse problems as problems of statistical inference by means of Bayesian statistics. In Bayesian statistics all quantities are modeled as random variables. The randomness, which reflects the observer's uncertainty concerning their values, is coded in the probability distributions of the quantities. From the perspective of statistical inversion theory, the solution to an inverse problem is the probability distribution of the quantity of interest when all information available has been incorporated in the model. This distribution, called the *posterior distribution*, describes the degree of confidence about the quantity after the measurement has been performed.

This book, unlike many of the inverse problems textbooks, is not concerned with analytic results such as questions of uniqueness of the solution of inverse problems or their a priori stability. This does not mean that we do not recognize the value of such results; to the contrary, we believe that uniqueness and stability results are very helpful when analyzing what complementary information is needed in addition to the actual measurement. In fact, designing methods that incorporate all prior information is one of the big challenges in statistical inversion theory.

There is another line of textbooks on inverse problems, which emphasize the numerical solution of *ill-posed problems* focusing on regularization techniques. Their point of view is likewise different from ours. Regularization techniques are typically aimed at producing *a reasonable estimate* of the quantities of interest based on the data available. In statistical inversion theory, the solution to an inverse problem is not a single estimate but a probability distribution that can be used to produce estimates. But it gives more than just a single estimate: it can produce very different estimates and evaluate their reliability. This book contains a chapter discussing the most commonly used regularization schemes, not only because they are useful tools for their own right but also since it is informative to interpret and analyze those methods from the Bayesian point of view. This, we believe, helps to reveal what sort of implicit assumptions these schemes are based on.

1.1 Introductory Examples

In this section, we illustrate the issues discussed above with characteristic examples. The first example concerns the problems arising from the noncausal nature of inverse problems.

Example 1: Assume that we have a rod of unit length and unit thermal conductivity with ends set at a fixed temperature, say 0. According to the standard model, the temperature distribution $u(x,t)$ satisfies the heat equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial u}{\partial t} = 0, \quad 0 < x < 1,\ t > 0,$$

with the boundary conditions

$$u(0,t) = u(1,t) = 0$$

and with given intial condition

$$u(x,0) = u_0(x).$$

The inverse problem that we consider is the following: Given the temperature distribution at time $T > 0$, what was the initial temperature distribution?

Let us write first the solution in terms of its Fourier components,

$$u(x,t) = \sum_{n=1}^{\infty} c_n e^{-(n\pi)^2 t} \sin n\pi x.$$

The coefficients c_n are the Fourier sine coefficients of the initial state u_0, i.e.,

$$u_0(x) = \sum_{n=1}^{\infty} c_n \sin n\pi x.$$

Thus, to determine u_0, one has only to find the coefficients c_n from the final data. Assume that we have two initial states $u_0^{(j)}$, $j = 1, 2$, that differ only by a single high-frequency component, i.e.,

$$u_0^{(1)}(x) - u_0^{(2)}(x) = c_N \sin N\pi x,$$

for N large. The corresponding solutions at the final time will differ by

$$u^{(1)}(x,T) - u^{(2)}(x,T) = c_N e^{-(N\pi)^2 T} \sin N\pi x,$$

i.e., the difference in the final data for the two initial states is exponentially small; thus any information about high-frequency components will be lost in the presence of measurement errors. ◇

Example 2: Consider the scattering of a time harmonic acoustic wave by an inhomogeneity. The acoustic pressure field u satisfies, within the framework of linear acoustic, the wave equation

$$\Delta u + \frac{\omega^2}{c^2} u = 0 \text{ in } \mathbb{R}^3, \tag{1.1}$$

where $\omega > 0$ is the angular frequency of the harmonic time dependence and $c = c(x)$ is the propagation speed. Assume that $c = c_0$=constant outside a bounded set $D \subset \mathbb{R}^3$. We shall denote

$$\frac{\omega^2}{c(x)^2} = k^2(1 + q(x))$$

where $k = \omega/c_0$ is the wave number and q is a compactly supported perturbation defined as

$$q(x) = \frac{c_0^2}{c(x)^2} - 1.$$

Assume that we send in a plane wave u_0 traveling in the direction $\omega \in S^2$. Then the total field is decomposed as

$$u(x) = u_0(x) + u_{\rm sc}(x) = e^{{\rm i}k\omega\cdot x} + u_{\rm sc}(x),$$

where the scattered field $u_{\rm sc}$ satisfies the Sommerfeld radiation condition at infinity,

$$\lim_{r\to\infty} r\left(\frac{\partial u_{\rm sc}}{\partial r} - {\rm i}k u_{\rm sc}\right) = 0, \quad r = |x|. \tag{1.2}$$

The field u satisfies the Lippmann–Schwinger integral equation

$$u(x) = u_0(x) - \frac{k^2}{4\pi}\int_D \frac{e^{{\rm i}k|x-y|}}{|x-y|} q(y)u(y)dy. \tag{1.3}$$

Expanding the integral kernel in Taylor series with respect to $1/r$, we find that asymptotically, the scattered part is of the form

$$u_{\rm sc} = \frac{e^{{\rm i}kr}}{4\pi r}\left(u_\infty(\hat{x}) + \mathcal{O}\left(\frac{1}{r}\right)\right), \quad \hat{x} = \frac{x}{r},$$

where the function u_∞, called the *far field pattern*, is obtained as

$$u_\infty(\hat{x}) = -k^2 \int_D e^{-{\rm i}\hat{x}\cdot y} q(y)u(y)dy. \tag{1.4}$$

The forward scattering problem is to determine the pressure field u when the wave speed c is known.

The inverse scattering problem wants to determine the unknown wave speed from the knowledge of the far field patterns with different incoming plane wave directions.

We observe the fundamental difference between the direct and inverse problem. The direct problem requiress the solution of one linear differential equation (1.1) with the radiation condition (1.2) or equivalently the Lippmann–Schwinger equation (1.3) which are linear problems. The inverse problem on the other hand is highly nonlinear since u in the formula (1.4) depends on q. Quite advanced techniques are needed to investigate the solvability of this problem as well as to implement a numerical solution. ◇

1.2 Inverse Crimes

Throughout this book, we shall use the term *inverse crime.*[1] By inverse crimes we mean that the numerical methods contain features that effectively render the inverse problem less ill-posed than it actually is, thus yielding unrealistically optimistic results. Inverse crimes can be summarized concisely by saying that the *model and the reality are identified*, i.e., the researcher believes that the computational model is exact. In practice, inverse crimes arise when

1. the numerically produced simulated data is produced by the same model that is used to invert the data, and
2. the discretization in the numerical simulation is the same as the one used in the inversion.

Throughout this book, these obvious versions of inverse crimes are avoided. Moreover, we show that the statistical inversion theory allows us to analyze the effects of modelling errors. We shall illustrate with examples what a difference the inverse crimes can make in simulated examples and, more importantly, how proper statistical error modelling effectively can remove problems related to discretization.

[1] To the knowledge of the authors, this concept was introduced by Rainer Kress in one of his survey talks on inverse problems.

2

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 $A : H_1 \to H_2$ 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 $y \in \mathrm{Ran}(A)$, and
2. the solution is *unique* if and only if $\mathrm{Ker}(A) = \{0\}$,

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

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

where ϕ 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

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

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

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

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 $\mathrm{Ran}(A)$ and the integral not to converge, thus making the inversion formula practically useless. ◇

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, $0 < s_1 < \cdots < s_n < \infty$, 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.,

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

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

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

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. ◇

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.

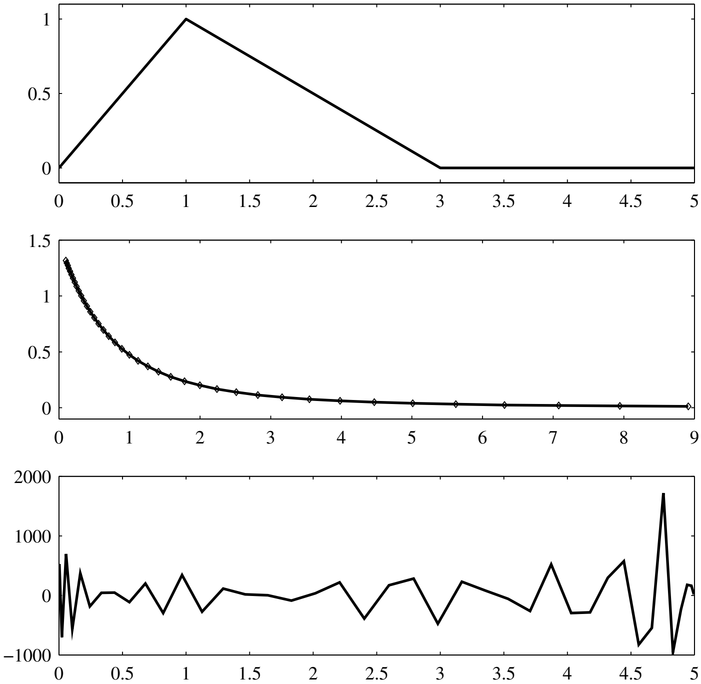

Figure 2.1. The original function (top), its Laplace transform (center) and the estimator obtained by solving the linear system (bottom).

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 $A : H_1 \to H_2$ 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*

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

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

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

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

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, λ_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 $x_0 \in \mathrm{Ker}(A)$ 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

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

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,

$$Ax = PAx = Py. \tag{2.3}$$

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

$$P_k : H_2 \to \operatorname{span}\{u_1, \ldots, u_k\}, \quad y \mapsto \sum_{n=1}^{k} \langle y, u_n \rangle u_n. \tag{2.4}$$

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

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

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

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

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 $A : H_1 \to H_2$ be a compact operator with the singular system (λ_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*

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

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*

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

as $k \to \infty$, 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 $\mathbb{R}^n \to \mathbb{R}^m$, and consider the matrix equation

$$Ax = y.$$

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

$$A = U\Lambda V^{\mathrm{T}},$$

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

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

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 $U = [u_1, u_2, \ldots, u_m]$ and $V = [v_1, v_2, \ldots, v_n]$ 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 $p = n$,

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

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

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

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

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

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

$$A^\dagger = V\Lambda^\dagger U^{\mathrm{T}},$$

where

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

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

$$\|y - y_0\| \simeq \varepsilon \tag{2.6}$$

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, $1 \le k \le m$ the largest index that satisfies

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

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 logarthms of its singular values is shown in Figure 2.2.

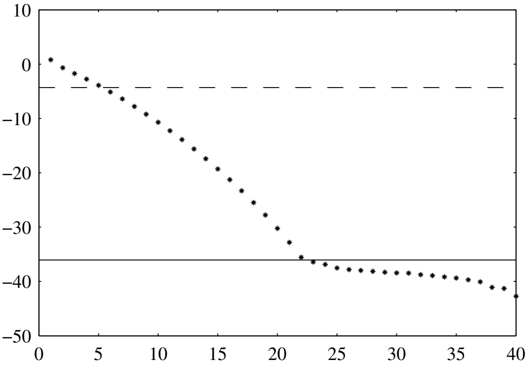

Figure 2.2. The singular values of the discretized Laplace transform on a logarithmic scale. The solid line indicates the level of the machine epsilon.

Let ε_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 ε_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 ε_0, the smallest ones above this level are quite close to ε_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) σ 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 σ.

When the standard deviation of the noise is given, it is not clear without further analysis how one should select the parameter ε in the dsicrepancy principle. In this example, expect somewhat arbitrarily the norm of the noise to be of the order of σ. 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. ◇

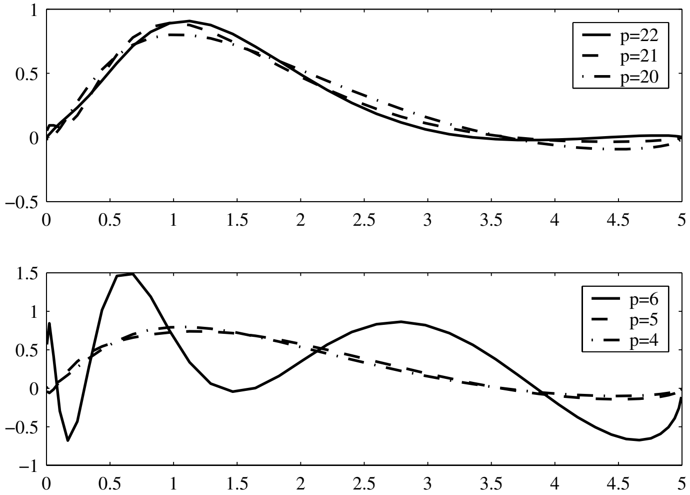

Figure 2.3. The inverse Laplace transform by using the singular value truncation. The top figure corresponds to no artificial noise in the data, the bottom one with 1% additive artificial noise.

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 $k \to \infty$. 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 $A : H_1 \to H_2$ be a compact operator with the singular system (λ_n, v_n, u_n). Then the Tikhonov regularized solution exists, is unique, and is given by the formula*

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

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

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

Hence, x_δ 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

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

where $P : H_1 \to \mathrm{Ker}(A)$ is the orthogonal projector. By projecting onto the eigenspaces $\mathrm{sp}\{v_n\}$, 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_δ minimizes the quadratic functional F_δ, 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_δ. The last term is nonnegative and vanishes only if $z = 0$. This proves the claim. □

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_δ 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 δ based on the noise level of the measurement y is a central issue in the literature discussing Tikhonov regularization. Several methods for choosing δ 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_δ be defined by (2.7), and

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

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

$$f(\delta) = \|Ax_\delta - y\| = \varepsilon, \tag{2.10}$$

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*

$$\|Py\| \leq f(\delta) \leq \|y\|, \tag{2.11}$$

where $P : H_2 \to \mathrm{Ker}(A^*) = \mathrm{Ran}(A)^\perp$ *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_δ, 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,

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

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

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

as claimed. □

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 $\lambda_1 \geq \cdots \geq \lambda_r$, one can employ e.g., Newton's method to find the unique zero of the function

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

The derivative of this function with respect to the parameter δ 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 ε 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