Francisco Duarte Moura Neto
Antônio José da Silva Neto

An Introduction to Inverse Problems with Applications

Springer

An Introduction to Inverse Problems with Applications

Francisco Duarte Moura Neto
and Antônio José da Silva Neto

An Introduction to Inverse Problems with Applications

Springer

Authors

Francisco Duarte Moura Neto
Dept. de Modelagem Computational
Universidade do Estado do Rio de Janeiro
Rua Bonfim 25, parte UERJ
Nova Friburgo, RJ 28625-570
Brazil
E-mail: fmoura@iprj.uerj.br

Antônio José da Silva Neto
Dept. de Engenharia Mecânica e Energia
Universidade do Estado do Rio de Janeiro
Rua Bonfim 25, parte UERJ
Nova Friburgo, RJ 28625-570
Brazil
E-mail: ajsneto@iprj.uerj.br

ISBN 978-3-642-32556-4
e-ISBN 978-3-642-32557-1
DOI 10.1007/978-3-642-32557-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012945297

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to

my love Maysa, our beloved children Daniel and Cecilia,
and my dear parents Maria Luiz and Francisco *(in memoriam)*

fdmn

my beloved ones, Gilsineida, Lucas and Luísa,
and to my dear parents Antônio *(in memoriam)* and Jarleide

ajsn

Foreword

For the benefit of the general readership, it might be a good idea to first explain the difference between "inverse problems" versus "analysis problems" using simple general terms as defined by Professor Shiro Kubo of Osaka University in his book published in the early 1990s.

Analysis problems are well-posed problems concerned with finding distribution(s) of certain variable(s) in a domain of a given size and shape that can be multiply connected and time-dependent. Results of the analysis problems also depend on boundary and/or initial conditions given at every point of a boundary of the domain. Properties of the media filling the domain must also be given in addition to strengths and locations/distribution of any possible sources/sinks. Finally, equation(s) governing the field variable(s) must be given. If all of these pieces of information are given, then the problem of finding the field distribution of variable(s) is a well-posed problem that can be solved using appropriate numerical integration algorithms.

However, if at least one of these pieces of information is missing, such under-specified problems become inverse problems of determining the missing piece of information in addition to simultaneously solving the original analysis problem. To enable ourselves to accomplish this complex task, we must be given an additional piece of information (typically, a part of the solution of the corresponding analysis problem) which makes the inverse problem an over-specified or ill-posed problem.

For example, when given size and shape of an isotropic plate and either Neumann or Dirichlet boundary conditions at every point along the edges of the plate, steady state heat conduction in the plate will be governed by the LaPlace's equation for temperature. This would be a classical well-posed analysis problem. However, if boundary conditions are not given on one boundary of this plate, the problem of finding temperature distribution in the plate becomes under-specified and cannot be solved. This problem will become solvable if we provide both Dirichlet and Neuman boundary conditions simultaneously on at least some parts of the plate's boundaries which will make this an over-specified or ill-posed inverse problem of determining the missing boundary conditions and simultaneously determining the distribution of temperature throughout the plate.

Inverse problems have traditionally been considered mathematically challenging problems and have consequently been studied predominantly by mathematicians. Since there are many practical inverse problems in a variety of disciplines that require mathematical tools for their solution, it is scientists and engineers that have been developing many of these methods recently out of necessity to obtain practical results. Consequently, an initially wide gap between scientists and engineers versus applied mathematicians has been steadily narrowing as both communities have realized that they have many things to learn from each other.

This book is a welcome and unique publication that uses a common language to blend the rigour of the applied mathematics world and the reality of a research scientist's or an engineer's world. Thus, it should appeal to everyone who has the basic knowledge of differential equations and at least a rudimentary understanding of basic mathematical models used in field theory and general continuum mechanics and transport processes. Specifically, applied mathematicians will be able to find here physical relevance for some of their theoretical work and learn to appreciate the importance of developing understandable, easy-to-use, easy to adapt and reliable algorithms for the solution of different classes of inverse problems. At the same time, research scientists and engineers will be able to learn from this book that some of the methods and formulations that they have been using in the past are prone to problems of non-uniqueness of the results and that accuracy of many of the practical methods could easily become a real issue when solving inverse problems.

Actually, this book could be used not only as a valuable reference book, but also as a textbook for students in the fields of applied mathematics, engineering and exact sciences. Besides its simple language, this book is easy to comprehend also because it contains a number of illustrative examples and exercises demonstrating each of the major concepts and algorithms.

For example, basic concepts of regularization of ill-posed problems are very nicely explained and demonstrated so that even a complete novice to this field can understand and apply them. Formulations and applications in image processing and thermal fields presented in this book have direct practical applications and add significantly to the more complete understanding of the general problematics of inverse problems governed by elliptic and parabolic partial differential equations. Inverse scattering problems have not been covered in this book as this field can easily fill a separate book.

Many formulations for the solution of inverse problems used to be very discipline specific and even problem specific. Thus, despite their mathematical elegance and solution efficiency and accuracy, most of the classical inverse problems solution methods had severe limitations concerning their fields of applicability. Furthermore, most of these methods used to be highly mathematical, thus requiring highly mathematical education on the part of users.

Since industry requires fast and simple algorithms for the solution of a wide variety of inverse problems, this implies a growing need for users that do not have a very high degree of mathematical education. Consequently, many of the currently used general algorithms for the solution of inverse problems eventually result in some sort of a functional that needs to be minimized. This has been recognized by the authors of this book which have therefore included some of the most popular minimization algorithms in this text.

Hence, this book provides a closed loop on how to formulate an inverse problem, how to choose an appropriate algorithm for its solution, and how to perform the solution procedure.

I recommend this book highly to those that are learning about inverse problems as well as to those that think that they know everything about such problems. Both entities will be pleasantly surprised with the ease that concepts, formulations and solution algorithms are explained in this book.

George S. Dulikravich

Miami, Florida

June 2011

Preface

Archimedes, is this crown made of gold?
King Hiero II of Syracuse[a], *circa* 250 BCE.

(. . .) to find a shape of a bell by means of the sounds which it is capable of sending out.
Sir A. Schuster[b], 1882.

Can one hear the shape of a drum?
Marc Kac[c], 1966.

[a] Hiero II (308 BCE - 215 BCE).
[b] Sir A. Schuster (1851-1934).
[c] Marc Kac (1914-1984).

On inverse problems

Perhaps the most famous inverse problem for the mathematical community is: *Can one hear the shape of a drum?* [40, 65]. That is: *Is one able to figure out the shape of a drum based on the sound it emits?* The corresponding direct problem is to determine the sound emitted by a drum of known shape. The solution to the direct problem is long known, but the solution to the inverse problem eluded the scientific community for a long time. It was found to be negative: there are two drums, different in shape, that emit the same sound, see [36]. Several other mathematical aspects concerning the resolution of inverse problems have been investigated in recent years.

Parallel to that, a large number of significant inverse problem methodology applications were developed in engineering, medicine, geophysics and astrophysics, as well as in several other branches of science and technology.

Why? Because inverse problems is an interdisciplinary area that matches the mathematical model of a problem to its experimental data. Or, given a bunch of numbers, data, in a data driven research, looks for a mathematical model. It is an interface between theory and practice!

About this book

The general purpose of this book is to introduce certain key ideas on inverse problems and discuss some meaningful applications. With this approach, we hope to be able to stimulate the reader to study inverse problems and to use them in practical situations.

The book is divided, though not in sequence, in two main parts, one of a more mathematical nature, and the other more linked to applications. It adopts an elementary approach to the mathematical analysis of inverse problems and develops a general methodology for the solution of real inverse problems. Further, it discusses a series of applications of this methodology, ranging from image processing applied to medicine, to problems of radiation applied to tomography, and onto problems of conductive heat transfer used in the design of thermal devices. The choice of applications reflect the acquaintance of the authors.

In order to make the book of a manageable size and suitable for a larger audience, we opted to make the presentation of mathematical concepts in the context of linear, finite dimensional, inverse problems. In this setting, key issues can be handled with mathematical care, in a rather "pedestrian" and easy way because the problem is linear and finite dimensional, requiring only acquaintance with basic ideas from linear algebra. Geometrical ideas, which are a key to generalization, are emphasized.

Some of the applications considered, however, involve problems which are neither finite dimensional nor linear. The treatment then is numerical, and the ideas from the first part of the book are used as guidelines, through extrapolation. This is possible, in part, because, to simplify, this book deals, several times, with *least squares methods.* Although the subjects in this book are intricate, the chapters can be read, somewhat, independently of one another. This is because of the intended redundancy, employed for pedagogical reasons, like in an old teaching's tradition: attention, association and repetition. To make it easier to peruse the book, a description of the book's content, chapter by chapter, is given on pages 4-6.

The pre-requisites to read this book are calculus of several variables and linear algebra. Nonetheless, a few concepts and results from linear algebra and calculus are reviewed in the appendix, in order to make the book reasonably self-contained. Even though knowledge of differential equations is necessary to understand some parts, basic concepts on this subject are not supplied or reviewed. Knowledge of numerical methods might be useful for reading certain sections. We included exercises at the end of each chapter, many of them guided, to make it easier for readers to grasp, and extend the concepts presented.

We believe that this book can be read, with some interest and to their profit, by upper division undergraduate students and beginning graduate students in applied mathematics, physics, engineering, and biology. Since it also includes some thoughts on mathematical modeling, which are in the back of the minds of researchers, but are not usually spelled out, this book may interest them too.

Some remarks

Practical matters: we use *emphasised expressions* to signal a group of words with a specific meaning when they are being defined either explicit or implicitly. The end of an example, a proof, or an argument is indicated by a small black square. ■

The authors will be happy to receive any comments and/or suggestions on this book.

This text reflects the partnership between a mathematician and an engineer. It is the result of a cross fertilization, in full collaboration, that greatly enthuses us and that, we hope, encourages others to break the usual barriers and pursue similar partnerships.

Francisco Duarte Moura Neto
Antônio José da Silva Neto
Nova Friburgo, Brazil
July 2012

Acknowledgements

We are deeply indebted to Professors George S. Dulikravich from Florida International University, Haroldo Fraga Campos Velho from the brazilian National Space Research Institute (INPE), Ricardo Fabbri from Rio de Janeiro State University, and Robert E. White from North Carolina State University that read throughoutly the manuscript and pointed out innumerous ways to improve it. Professor Dulikravich encouraged us to have it published in English. Professor Campos Velho helped us to clarify our classification scheme of inverse problems leading to an improved Section 2.8. We also thank very strongly Professor Luiz Mariano Carvalho, from Rio de Janeiro State University, which read several parts of the book, including the Chapter 1 and Afterword, giving very pertinent suggestions, and encouraging us in several ways to pursue our best in writing this book. We thank them very emphatically for their time, patience, expertise, and friendship. Of course, entire responsability rests with the authors for the mistakes, typos and less than good explanations that remain in the book. Also, it is a great honor that Professor Dulikravich, founder and editor-in-chief of the journal *Inverse Problems in Science and Engineering*, accepted to write the foreword of this book.

A previous version of this book was published originally in Portuguese, but additional material has been included here. We recognize and appreciate the careful initial translation into English of the previous book by Mr. Rodrigo Morante.

During all these years we have been working on inverse problems, we benefited from discussions with several colleagues and students, which greatly helped us to understand better the issues involved on their solutions and we thank them all.

We want to acknowledge our colleagues and staff of Instituto Politécnico from the Rio de Janeiro State University, in Nova Friburgo, for constructing and mantaining such a challenging academic environment in the inner *Mata Atlântica*, *Serra do Mar*, a remaining part of the original rain forest in the State of Rio de Janeiro, a hundred miles northeast from the city of Rio de Janeiro.

The campus of Instituto Politécnico was severely affected by the torrential rains and devastating landslides that occurred on january 11, 2011 in the city of Nova Friburgo[1]. As a result, we were *campus-less* by more than a year and just recently, on march 2012, moved to a new campus. During this *annus horribilis* our Instituto was maintained 'alive' by brave colleagues and staff that take on the mission of education heartily. We again thank them all.

We also acknowledge the partial financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação Carlos Chagas Filho de

[1] NY Times http://www.nytimes.com/2011/01/14/world/americas/14brazil.html

Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj), which permited us to develop research projects, and supported our undergraduate and graduate students by granting scholarships to them.

We also thank the Brazilian Society of Applied and Computational Mathematics (SBMAC) that, by accepting our proposal to lecture on *inverse problems* during its annual meeting in 1999, stimulated us to write down some of our ideas and experiences on the subject. Thanks are also due to EDUERJ, the Rio de Janeiro State University Press, specially to its former Director, Professor Lúcia Bastos, which took on the mission to publish the earlier Portuguese version of this book.

We register special thanks to Professor Paulo Márcio Mello, former Director of the Centro de Produção (CEPUERJ) of the Rio de Janeiro State University. Partial support from CEPUERJ made it possible to publish the referred earlier version.

We register our thanks to Rui Ribeiro de Azevedo for helping us with the book cover and to Maysa Sacramento de Magalhães who took the photo in the book cover.

We thank the staff from Springer for such a friendly interaction.

fdmn

F. D. Moura Neto wants to acknowledge his luck, happiness and gratefullness to have been taught mathematics by Professor Carlos Tomei, from Pontiphical Catholic University of Rio de Janeiro, by the late Professor Ronald DiPerna, from the University of California, Berkeley, and by his thesis advisor, Professor John Charles Neu. Their lectures and scientific points of view were always very inspiring and set a high standard that he strives to attain. The fine institutions previously mentioned were his *Alma Mater* and he also wants to acknowledge their enormous contribution to his education.

He wants to thank and register the loving care of his late mother Maria Luiz who fought strongly for her children to have a good education. At this point he also wants to express, with love, a special gratitude to his wife, Maysa, and their children, Daniel and Cecília, for nurturing and sharing the most happy and adventurous moments of his life.

ajsn

A. J. Silva Neto wants to thank a number of professionals that were very important in his formal education, and also in his activities both in consulting engineering and the academia. He was very fortunate, since his early education at the undergraduate and graduate levels as an engineer, for being in touch with very intelligent and hard working professors, such as Raad Yahya Qassim (Nuclear Engineering Department Head), Nilson C. Roberty (undergraduate advisor - Mechanical/Nuclear Engineering) and Antônio C. Marques Alvim (M.Sc. advisor - Nuclear Engineering), all three from Federal University of Rio de Janeiro, Brazil, and Mehmet N. Özisik (Ph.D. Advisor - Major in Mechanical and Aerospace Engineering) and Robert E. White (Ph.D. Advisor - Minor in Computational Mathematics), both from North Carolina State University, USA. While working for Promon Engenharia, Brazil, as a consultant engineer, he had the privilege to work with, and learn from, Diogo Dominguez,

Raghavan P. Kesavan Nair, Leonardo Yamamoto, João Paulo Granato Lopes, Álvaro Bragança Jr., André Castello Branco, Silvio Zen, Eric Rosenthal, and Lázaro Brener, true leaders in their fields. Regarding his first steps on education, research and practice on inverse problems, the following professionals were fundamental in the establishment of the high quality standards to be followed: James V. Beck, Oleg M. Alifanov, George S. Dulikravich, Norman J. McCormick, K. A. Woodbury, Diego Murio, Eugene Artioukhine, César C. Santana, and Paulo A. Berquó de Sampaio. The following colleagues played an important role in the development of some of the results presented in this book: Geraldo A. G. Cidade (Chapter 4), fromFederal University of Rio de Janeiro, and Gil de Carvalho (Chapter 6), from Rio de Janeiro State University. As nothing is truly valuable unless being full of dedication and passion, thanks are also due to those magnificent human beings that have to endure so many hardships, besides my absence, in order to allow my dedication to the research, practice and teaching in engineering and applied mathematics, my dear wife Gilsineida, and our lovely children Lucas and Luísa.

Contents

Chapter 0
Road Map

We shall begin by presenting some areas of application of inverse problems and the content of book's chapters.

0.1 Inverse Problems in a Snapshot

Given a line, $y = 3x + 4$, to find a point that belongs to the line is a so-called *direct* problem. Given two points, $(-1,1)$ and $(0,4)$, an *inverse* problem is to find the equation of the line that they determine.

Let us make this a larger section. Given a line, $y = 3x + 4$, in order to find the y value when, say, $x = -1$ is a *direct* problem.

Given a class of lines, $y = ax + b$, and *theoretically measured* data

$$(x_{meas_1}, y_{-meas_1}) = (-1,1),\ (x_{meas_2}, y_{meas_2}) = (0,4)\,,$$

finding the specific line, that satisfies the data, i.e. finding that $a = 3$ and $b = 4$, is an inverse *identification* problem.

Given a line, $y = 3x + 4$, and *theoretically measured* data, $y_{meas} = 1$, finding the *source* $x = -1$, is an inverse *reconstruction* problem.

Reality enters the picture if, for instance, to solve an inverse identification problem, one is given a set of experimental data,

$$(x_{meas_1}, y_{meas_1}) = (-1,1.1), (x_{meas_2}, y_{meas_2}) = (1,6.85), (x_{meas_3}, y_{meas_3}) = (0,3.9)\,,$$

which does not allow an immediate matching with the model, i.e., there is no model in the class able to interpole the data.

These examples are quite simple and it is somewhat difficult to grasp their distinct nature. We shall return to these matters in Section 2.8.

0.2 Applied Inverse Problems

Due to the growing interest in inverse problems, several seminars and international congresses have taken place: *First International Conference on Inverse Problems in Engineering: Theory and Practice*, 1993 [8] (its seventh edition took place in Orlando, USA, in May 2011); *Inverse Problems in Diffusion Processes*, 1994 [31]; *Experimental and Numerical Methods for Solving Ill-Posed Inverse Problems: Medical and Nonmedical Applications*, 1995 [9] and the *Joint International Conference on Mathematical Methods and Supercomputing for Nuclear Applications*, 1997 [6], to mention but a few. The number of published results pertaining original research has also grown. Just as an example, in the 2011 edition of the International Congress on Industrial and Applied Mathematics, that took place in Vancouver, ICIAM 2011, one finds a large number of simposia on inverse problems, as well as on computational modeling, finances and applications to life sciences, in which inverse problems also play an important role.

As examples of inverse problems applications in everyday life we can mention the following *reconstruction* problems[1]: underground non-metallic materials detection by reflected radiation means; intensity and position estimation of luminous radiation from a biological source using experimental radiation measures; tomography (Computerized Tomography, or CT; Single Photon Emission Computerized Tomography, or SPECT; Near Infrared Optical Tomography, or NIROT); thermal sources intensity estimation with functional dependence in space and/or time in heat transfer problems, based in transient temperature measures, and initial condition estimation of transient problems in conductive heat transfer.

The estimation of properties used in mathematical models of physical systems is, in its own right, a special class of inverse problems, the so-called *identification* problems. It is common to estimate the thermal conductivity and diffusivity of materials, depending, or not, on temperature, in diffusive processes, and also the radiative properties — single scattering albedo, optical thickness and anisotropic scattering phase function — in heat transfer by thermal radiation in participating media. The study of combustion inside furnaces is one of the areas where this kind of thermal radiation inverse problems are applied.

Another important inverse problems class is image reconstruction (the aforementioned tomography problem can be included here too), which can be formulated as a mixed *identification and reconstruction* problem. Based on the estimation of medium properties, it is possible to identify defects in components — through non-destructive testing in industry, — or altered regions in biological tissues — applied in diagnosis and treatment in medicine.

A research and technological innovation area that has received growing demands is *new materials* development. Recent trends in the use of materials, points to the development of materials with specific properties, designed in advance to meet the demands of new applications, ranging from engineering to medicine.

The need to use adequate techniques for new materials characterization, that is, determination of their properties, is obvious. However, the degrees of freedom of an experimental device are usually restrained during its development and operational phases. This, in turn, frequently imposes practical limitations, which restricts the experiment's possibilities only to a fraction. In this case, inverse problems methodology has been used to determine those properties, and also in the design of experiments, allowing to tackle more complex and challenging problems.

[1] The notion of reconstruction problems is discussed in Sections 2.8 and 8.1. The problems mentioned here can be formulated as reconstruction problems but can also be formulated as identification problems or sometimes a mix of reconstruction and identification problems, depending on the mathematical model adopted and the role that the estimated quantity plays in it.

0.3 Crossing the Science-Technology Gap

We all have faced the resolution of inverse problems before. Perhaps not from a mathematical point of view, though. Kids love solving inverse problems. “ I’m the part of the bird that’s not in the sky. I can swim in the ocean and yet remain dry. What am I[2]?” They love the ambiguity inherent to inverse problems, one of which is that they, in general, have more than one answer. Mathematicians, due to this feature, classify inverse problems as *ill-posed* problems.

One of the greatest ideas in mathematics is giving a name to that which is not known: the *unknown* (x, perhaps?). In solving inverse problems this is done all the time. When something is named it may be talked about, even if one does not grasp it in its wholeness. After that, and persevering a little, enlightenment may be reached. The “unknown” will be limited or nailed down by assertions which it satisfies[3]. Step by step, the range of possible x’s is reduced. Many times, in the end, it remains just a unique x. That is when mathematicians utter in satisfaction “The solution *exists*, and is *unique*!”

Another great discovery was the notion of *function*[4]. From it the notion of *solution operator* is only a step away: the solution operator of a given problem is that abstruse mathematical concept that, from the data, *reveals* the solution. Oh, yes, this is a function because — do you remember? — the solution is unique, as it was shown.

The solution operator has to bear resemblance with reality, and it must be constructed from observations of the system or phenomenon under investigation. It must model reality. It must fit the data, as much as possible. Is this least squares?

Then comes the *solution operator* representation. What does it do with the data to produce the answer? Is it a linear function? A polynomial, perhaps? Is it the sum of the derivatives of order less than four of the data? There are lots of possibilities!

What we **want** is to know **the answer**. How about to transform that representation into an *algorithm*, a finite procedure that, in the end, will grant us the divine grace of knowing the answer to the problem?

[2] “I am a shadow”.

[3] Strictly speaking, this is not mathematics, it is modeling — physical modeling in a broad sense, which may include biological modeling, sociological modeling and so on.

[4] The notion of function, or almost it, has been reinvented under different names in so many sciences because it is a fundamental concept. When being taught mathematics, however, few really are able to grasp the concept in its fullness.

Just to make it clear, since we shall use it frequently, a *function* f, with *domain* A and *codomain* B, is a rule that for each element $x \in A$ attributes a unique element $f(x)$ belonging to B. The attribution is indicated by $\mapsto$, and the notation for f, with all its features is

$$\begin{aligned} f :& A \to B \\ & x \mapsto f(x) \end{aligned}$$

or, sometimes, a shorter version is usual: $A \ni x \mapsto f(x) \in B$.

Well, it is not exactly that way. The algorithm only tells us what should be done and the order thereof. In the end, we may feel much too lazy to do all the work. Or, perhaps, computations would take us so long that we may have to devote to it the rest of our lives! Now, instead, how about *programming* the algorithm?

And the computer does well what it is told to do. To do orderly? “So, they are talking about serial programming? So boring!” What then is the neat idea? Partial order? Parallel computing?

Finally, the numbers or a picture (worth a thousand words): the answer! Even so, a mathematical answer. And the problem ends. Ends? (The answer is yes, for this book. However...)

No! We still have to place it all inside a capsule, inside a wrapper, inside a product which is *intensely loaded with knowledge*.

And, through crossroads, the scientific and technological adventure continues.

0.4 Chapter by Chapter

The chapters in this book can be read more or less independently of one another. It is convenient to describe the contents of the book chapter by chapter, so the reader can locate more easily a specific content. We do it next.

Chapter 1 *Mathematical Modeling*. In this chapter we present some basic ideas on mathematical modeling, what *special* type of questions a mathematical modeler wants to answer (which a mathematical illiterate would not pose), and the role of inverse problems in this setting. The methodology used to illustrate the explanation is multivariate linear regression, and an application to flow in porous medium is considered. The notion of function, which plays a very prominent role throughout the book, is emphasized here.

Chapter 2 *Fundamental Concepts in Inverse Problems*. Here we present the classical analytical difficulties of inverse problems, the questions of existence, uniqueness and ill-posedeness. The discussion is elementary, and the examples come from algebraic equations. The condition on the evaluation of functions is touched upon and the question of stability of numerical algorithms is also pointed out by simple examples. General classifications of mathematical models and inverse problems are presented.

Chapter 3 *Spectral Analysis of an Inverse Problem*. This chapter introduces the regularization of a finite dimensional inverse problem, in order to get a meaningful approximation of its solution. Linear algebra ideas, in particular the spectral theory of symmetric matrices, are used to understand the workings of simple regularization schemes. The methods of Tikhonov, steepest descent, Landweber, and conjugate gradient are discussed in this context, and the discrepancy principle is presented.

Chapter 4 *Image Restoration*. This is the first chapter devoted to applications. De-blurring of images is considered, and examples from restoration of digital images, text, and biological images are handled. Some blurring mechanisms based on convolution are presented. These problems are formulated as nonlinear systems, using Tikhonov's regularization technique, with a family of regularization terms constructed with Bregman's divergences, and are solved by Newton-Raphson's root finder algorithm coupled with Gauss-Seidel iterative method. A schematic presentation of a restoration algorithm is included.

Chapter 5 *Radiative Transfer and Heat Conduction*. In this chapter inverse problems involving the linear Boltzmann equation of radiative transfer are presented. The interaction of radiation with heat conduction is also considered. The Levenberg-Marquardt method is discussed and used to solve inverse problems. Several parameters are determined: phase function expansion coefficients, single scattering albedo, optical thickness, thermal conductivity, and refraction index. Confidence intervals are obtained.

Chapter 6 *Thermal Characterization*. This chapter discusses thermal characterization of a polymeric material using data acquired with the hot-wire method. This amounts to determine the material's thermal conductivity and specific heat with a transient technique. The inverse problem methodology, which allows determination of more parameters than traditional experimental methods, gives better results. Again, we employ the Levenberg-Marquardt method. Confidence intervals for the results are determined.

Chapter 7 *Heat Conduction*. This chapter considers the reconstruction of a thermal source in a heat conduction problem. It is formulated as an optimization problem in an infinite dimensional space. The conjugate gradient method in a function space (Alifanov's iterative regularization method) is presented and applied to this problem. The use of discrepancy principle as a stopping criterion is illustrated.

Chapter 8 *A General Perspective*. This chapter proposes a reasonably general formulation of inverse problems and their types, while stressing the role of observation operators. Further, it discusses the gradient of objective functions, with some generality, leaving aside, however, mathematical aspects.

Afterword *Some Thoughts on Model Complexity and Knowledge*. In the last chapter we engage in a discussion of different levels of mathematical models, from *quantification of phenomena*, somewhat proposed by Pythagoras, to the revolutionary concept set in motion by Newton of a *dynamical model*, and the role of computation and inverse problems in that endeavor.

Appendix A *Spectral Theory and a Few Other Ideas from Mathematics*. In the appendix, we collect some concepts and results from linear algebra and calculus, that are used throughout the book: norms, distances, inner-products, orthogonal matrices, projections, spectral theorem of a real, symmetric matrix, singular value decomposition of a real matrix, and Taylor's formulae.

Chapter 1
Mathematical Modeling

In this chapter we present some of the aspects of the interface between mathematics and its applications, the so-called *mathematical modeling*. This is neither mathematics nor applied mathematics and is usually performed by scientists and engineers. We do this in the context of linear algebra, which allows easy comparison between direct and inverse problems. We also show some of the modeling stages and encourage the use of a least-squares method. An application to the study of flow in a porous medium is presented. Some key issues pertaining to the use of models in practical situations are discussed.

1.1 Models

We want to think about how we may come to understand a phenomenon, process or physical system. To simulate this endeavor, we make a *thought experiment*[1], under the assumption of conducting an investigation into the behavior of a hypothetical system, schematically represented by a black box.

Consider a *black box* whose inner working mechanisms are unknown and that, given a *stimulus* or *input*, answers with a *reaction* or *output*. See Fig. 1.1.

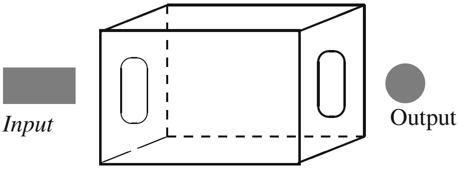


Fig. 1.1 Black box: the mental prototype

We aim to foretell the box behaviour in several distinct situations. In particular, we would like to tackle the following problems:

P_1: Given an arbitrary stimulus, tell what the corresponding reaction will be;

P_2: Given the reaction, tell what stimulus produced it.

However, we are not interested in generic predictions, but only in those based in scientific descriptions of the behaviour of the black box. To that end, we will

[1] Thought experiments have been used by several scientists, as for instance G. Galilei (1564-1642), R. Descartes (1596-1650), I. Newton (1643-1727), and A. Einstein (1879-1955), and it turns out to be a good artifact of investigation to keep in mind.

associate to the real situation a physical model and a corresponding mathematical model. Predictions will be made from the latter.

We shall call *physical model* any description of the phenomena involved using such concepts as, for example, number of items, mass, volume, energy, momentum, charge and their conservation, exchange, movement, or transference, etc.

By *mathematical model* we mean any kind of mathematical structure, such as, for example, a function, an equation, a set with an equivalence relationship, a vector space, a group, a graph, a probability distribution, a Markov chain, a neural network, a system of non-linear partial differential equations, an elliptic operator defined on a fiber space, etc.

Our guiding methodology when choosing a model, which we shall call *modeling*, is part of the *scientific method*, that we present in a somewhat *modern* language.

In practical situations we try to choose or develop a mathematical model that best describes the physical model, avoiding contradictions with reality and deviations from the phenomena of interest. We only keep the model as long as it is not refuted by experimental data[2].

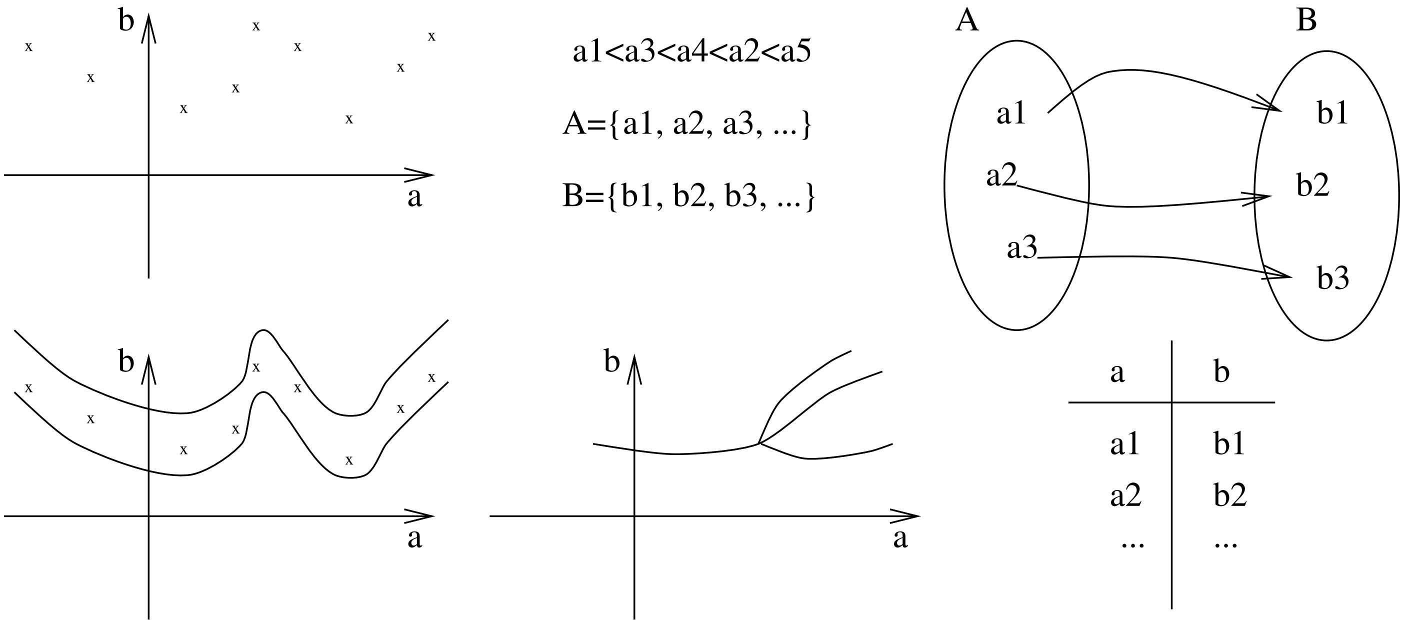
1.2 Observing Reality

The sleuth that is to solve a crime, or, in other words, who is to foretell the behaviour of a person or group of persons in a given situation, must create a careful profile of those persons and know, as best he can, the episode and the circumstances. Thus, he must bear in mind the various elements of the deed, the deeds before it and its consequences. To that end he must search for information where it is available, which, logically, includes the crime scene. He must observe and carry on the investigation, questioning persons, in any way, related to the crime, and analyzing their answers[3].

Our goal is to be able to guess the behaviour of the black box, Fig. 1.1. Analogously to the observation of a crime's trail, we begin the *observational* and *experimental phase* of our project. By applying different stimuli to the black box we observe and take note of the corresponding reactions.

[2] The mathematical model will be accepted as long as its predictions do not contradict experimental facts. That is a very strong requirement of the *scientific discourse*. If the model fails, one does not have any impediment to throw it away, to reject it, no matter for how long and for how many it has been used. In fact, one is obliged to reject it. This very strong stance, singles out the scientific method. This is precisely what Johannes Kepler (1571-1630) did. At a certain point, he believed that the orbits of the planets, around the sun, were circles. However, the data collected by Tycho Brahe (1546-1601) indicated that the orbit of Mars could not be fitted by a circle. This difficulty led him to discard the hypothesis that the orbits were circles, and embrace the more accurate model of elliptical orbits.

[3] Obviously, he can also consider the forensic analysis that makes it possible to track the trajectory of a projectile, a body, or a car, and any kind of materials involved. This constitutes, in foresight, a set of inverse problems, which, naturally, uses knowledge of physics and mathematics.



a1<a3<a4<a2<a5

A={a1, a2, a3, ...}

B={b1, b2, b3, ...}

a	b
a1	b1
a2	b2
...	...

Fig. 1.2 Organization of experimental data (from left to right, and from top to bottom): (a) Scatter-plot; (b) Set *A* represents a group of stimuli and set *B* a group of reactions; (c) Representing a function by a Venn diagram; (d) Functions that approximate and bound the experimental data; (e) Forking experimental data: it does not conform to a function but conforms to a level set of a function; (f) Table

The next step is to organize the experimental data in a way suitable for analysis. We thus create a *database of real* or *experimental data*. This is critical, since the way data is organized will emphasize certain aspects at the expense of others. Information can be presented, for example, as tables, diagrams, graphs and images. Some possible data structures are sketched in Fig. 1.2. Catalogues of possible stimuli and conceivable reactions can be created (Fig. 1.2b, first row, second column). Sometimes, these catalogues correspond to the *domain* and the *codomain* of appropriate functions.

Assume we want to answer problems formulated in Section 1.1, page 7, and some other questions that may arise out of curiosity or by chance. However, we would want to avoid resorting, everytime, to experimentation. With this is mind, but still wanting to do it scientifically, we build a mathematical model of the situation.

As a matter of fact, the experimental database allows one to answer some problems. Many people would be perfectly happy to solve their problems that way. Then, why should we build a mathematical model of the situation in the first place? Among the many reasons to do so, we mention two:

- Once a mathematical model has been devised, a natural environment is available in which several settings and hypothesis can be generated and tested, i.e. a number of scenarious may be constructed. In the example, it is possible, by standard deductive reasoning (logic implication), to propose candidates for reactions due to a wider range of stimuli values;

- A remarkable advantage of a mathematical model is the *compression* of information. Once the model is obtained, additional structures within the database become apparent. Those perceived structures are enough to render almost useless large parts of the database. This results in effective compression[4] which, in turn, leads to comprehension.

1.3 The Art of Idealization: Model Characterization

A class of models is chosen according to the given situation, technological and/or pratical measuring capabilities, purpose, and so on. Choosing a class of models is known as model *characterization*. We remark that this endeavour is not an application of mathematics in itself (in the sense that it is not solely the carrying out of an algorithm); indeed, it is an *intelligent* activity, a task generally highly complex, an art that demands an educated sensibility to execute it at its best. As a matter of fact,

[4] It is relevant to realize that *mathematical models* can effectively *compress data*. To illustrate this point, assume that we are dealing with data on the price of photocopies. If one has a database, then we need to have a table relating the number of copies and its price. We would have, for example, that one copy costs 10 cents, two copies cost 20 cents, three copies cost 30 cents and so on up to, say, a hundred copies that cost 10 monetary units (m.u.), a very long table. And it could be longer! By means of a mathematical model, using the notion of function, we say that the price of n copies is $n \times 0.10$ m.u., with $n \in \mathbb{N}$.

this is so much true that, at times, in novel situations, the stage of characterization may require the development of new mathematical structures[5].

To characterize the model, for example, we can perform an *exploratory data analysis*, in which we *contemplate* the data and afterwards may point out simple relationships between them.

Let's make it concrete by returning to the black box example. Assume that the stimulus and its corresponding reaction can be quantified/described each of them by three measurings that we collect in vectors[6], the

$$\textit{input signal},\quad \mathbf{x}=(x_1,x_2,x_3)^T\in\mathbb{R}^3\,,$$

and the

$$\textit{output signal},\quad \mathbf{y}=(y_1,y_2,y_3)^T\in\mathbb{R}^3\,.$$

The description of the input/output signals as elements of $\mathbb{R}^3$ is evidently part of the characterization stage[7],[8]. The summing up of the data in a table constitutes one of the simplest quantitative models, the *DB (database)* model.

An exploratory analysis of the available data may suggest as reasonable the following additional hypotheses on the black box workings:

H_1: **Reproducibility** — The repetition of the same input signal produces the same output signal[9];

[5] An illustration of this is the invention of complex numbers by Tartaglia (1500-1557) and Cardano (1501-1576) in 1545 to solve equations of the third degree. These equations were associated with practical problems, which indicated that they had three real roots. However, the available methods did not allow their solution because they required the square root of a negative number, which, at that time, had not yet been defined.

[6] Here, vectors are *vertical*, or column vectors, that is, $n\times 1$ matrices. Also, they are represented using **bold** letters. The entries of a $m\times n$ matrix, A, are denoted by A_{ij}, $i=1,\dots,m$, and $j=1,\dots,n$. For A, in particular for vertical vectors, A^T represents the *transpose* of A, that is, the matrix whose entries are given by $(A^T)_{ij}=A_{ji}$, for all $i=1,\dots,n$, $j=1,\dots,m$.

[7] We could have chosen $\mathbb{R}^4$, as the set of possible inputs if measures were related to three-dimensional space, and we also had to characterize time. Or, maybe, even higher order dimensional spaces could be necessary as is the case when, for example, we want to keep other information such as temperature, pressure, etc in the output signal.

[8] Note the non-sequential nature of modeling. Even before acquiring the experimental data of which we spoke in the previous section, we must start characterizing the model. See Section 2.8, page 44, and the Afterword, starting on page 177 for further considerations on modeling.

[9] It should be pointed out that *the same* here subtends *within* a certain margin of error associated with the model. The variability could be, for instance, characterized by a probabilistic random variable, or by an interval. This would require a more complete or detailed model. The characterization of a model always assumes some idealization. That is, even though the real phenomena do not strictly satisfy a certain assumption, we assume they satisfy in order to proceed with modeling. What is important is that conclusions from the model and from reality do not differ significantly. Sometimes this is a quantitative statement: the difference between numbers, coming from the model, and data, from reality, should be bounded by a certain predefined value. Modeling is an art that has to be practiced for one to have a good grasp on it.

H_2: **Proportionality** — If the input signal is amplified α times, the output signal is also amplified by that same factor;

H_3: **Superposition of effects** — If we add up two input signals, the output signal is the sum of the output signals corresponding to the individual input signals[10].

These hypotheses form an example of what we mean by a *physical model* of the real situation. Espousing these hypotheses allows us to obtain a model that is more complete than the DB model, a *descriptive model*[11]. This, in principle, contains the data set, though not necessarily in an explicit manner[12].

Mathematically, the first hypothesis means that

H_1: The attribution, *input* → *output*, is a function.

We shall denote that function by $\mathcal{F}$, and call it the *behaviour* function of the black box[13]. The second and third hypotheses essentially characterize $\mathcal{F}$ as a *linear function*,

H_2: $\mathcal{F}$'s evaluation commutes with scalar multiplication[14],

$$\mathcal{F}(\lambda\mathbf{u}) = \lambda\mathcal{F}(\mathbf{u}) \text{ for all } \lambda \in \mathbb{R},\ \mathbf{u} \in \mathbb{R}^3\ , \tag{1.2}$$

[10] In a sense, this means that inputs do not interact. Each one goes through the black box in its own way, ignoring the other.

[11] Descriptive models, roughly speaking, try to answer questions like: *"What happened?"* ("*What*") or *"When it happened?"* ("*When*"). Databases try to answer questions of the same nature. On the other hand, the *explanatory model* focuses in: *"How it happened?"* ("*How*"). Further discussion on the nature of models is presented in Afterword, page 177.

[12] What we mean here is that, for example, the function $\mathbb{R} \ni x \mapsto f(x) = x^2 \in \mathbb{R}$ contains the information of the following table,

x	-1	0	1	2
y=f(x)	1	0	1	4

We could say that f *encapsulates* the table.

[13] It is worth to remind that, for some functions, different inputs imply different outputs, whereas for some other functions, certain different inputs can give the same output. The former case is not the rule and deserves a special name, the function is called *injective* or a *one-to-one* function.

[14] $\mathcal{F}$ of a multiple is the multiple of $\mathcal{F}$. One may have some restrictions on this way of stating this property, and rightly so. Let us be more precise. First consider the function *multiplication by scalar* λ, given by

$$\begin{aligned} \Lambda : \mathbb{R}^3 &\rightarrow \mathbb{R}^3 \\ \mathbf{v} &\mapsto \Lambda(\mathbf{v}) = \lambda\mathbf{v}\ . \end{aligned}$$

Hypothesis H_2 can be written in terms of a commutation of a composition of functions,

$$\mathcal{F} \circ \Lambda = \Lambda \circ \mathcal{F}\ , \tag{1.1}$$

which justifies the assertion at the beginning of this footnote.

H_3: The evaluation of $\mathcal{F}$ commutes with vector summation[15],

$$\mathcal{F}(\mathbf{u}+\mathbf{v}) \quad = \quad \mathcal{F}(\mathbf{u})+\mathcal{F}(\mathbf{v}) \text{ for all } \mathbf{u},\ \mathbf{v} \in \mathbb{R}^3 \,. \tag{1.4}$$

Granting the validity of the hypothesis H_1, the act of organizing the data as stimuli or reactions, as suggested in Fig. 1.2b, page 9, would correspond, respectively, to the construction of the domain and the codomain of the behaviour — characterized by a function — of the black box.

In our example, the characterization stage reaches its end when we specify the model, by saying that the black box behaviour is described by a function, which is linear.

Non-linear models are relevant in many applications and will be dealt with in other sections within this book. We only considered the linear model so far, to facilitate the understanding of the concepts presented here.

1.4 Resolution of the Idealization: Mathematics

We are assuming that the mathematical model of the black box is a linear function, from $\mathbb{R}^3$ to $\mathbb{R}^3$. This characterizes the model. However, we do not know which specific linear function it is. One needs to determine that function, or, at least, to approximate it, in order to solve the problems posed in Section 1.1. This is the stage

[15] $\mathcal{F}$ of a sum is the sum of $\mathcal{F}$'s. Similarly to the previous footnote, we can rewrite hypothesis H_3 as commutation of composition of functions. The idea of preserving the notion of commutation is a neat one. However, to do that, one sometimes has to work a bit (to adapt notions). Let $\mathcal{S}$ be the function that adds two vectors,

$$\begin{aligned} \mathcal{S} : \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R}^3 \\ (\mathbf{u},\mathbf{v}) &\mapsto \mathcal{S}(\mathbf{u},\mathbf{v}) = \mathbf{u}+\mathbf{v} \,. \end{aligned}$$

Also, let $\bar{\mathcal{F}}$ be, essentially, two copies of $\mathcal{F}$, representing the function

$$\begin{aligned} \bar{\mathcal{F}} : \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R}^3 \times \mathbb{R}^3 \\ (\mathbf{u},\mathbf{v}) &\mapsto \bar{\mathcal{F}}(\mathbf{u},\mathbf{v}) = (\mathcal{F}(\mathbf{u}),\mathcal{F}(\mathbf{v})) \,. \end{aligned}$$

It may seem odd to define such a function, but that is exactly what we need here. In fact, Eq. (1.4) can be rewritten as

$$\mathcal{F} \circ \mathcal{S} \quad = \quad \mathcal{S} \circ \bar{\mathcal{F}} \,. \tag{1.3}$$

Note that if we stick to the *computer science* notion of *overloading* a function, we can still denote $\bar{\mathcal{F}}$ by $\mathcal{F}$. In this case, Eq. (1.3) could be written simply as $\mathcal{F} \circ \mathcal{S} = \mathcal{S} \circ \mathcal{F}$.

Simply put, we just want to say that, for a linear function, $\mathcal{F}$ of a sum is the sum of the $\mathcal{F}$'s, or that $\mathcal{F}$ commutes with summation. To make this statement precise, we need to do what we have just done. When you have to spell it out, before simplicity becomes simple, it may seem somewhat complex at times. Finally, we are not always blessed with commutation, so when we are, nothing more appropriate than to make an effort to recognize it.

Table 1.1 Data set: stimuli × reactions

stimuli	$\mathbf{x} \in \mathbb{R}^3$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$
reactions	$\mathcal{F}(\mathbf{x}) \in \mathbb{R}^3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$

in which the model is *determined*, where an inverse identification problem is solved to pinpoint a specific linear function.

Example 1.1. Model identification using an ideal database. In the example we are considering, it is very easy to determine the model. We will do it first in a particular case. Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, be the *canonical basis* of $\mathbb{R}^3$,

$$\mathbf{e}_1 = (1,0,0)^T\,, \quad \mathbf{e}_2 = (0,1,0)^T\,, \quad \mathbf{e}_3 = (0,0,1)^T\,.$$

Assume that we know the reaction of the black box when stimuli $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are applied. That is, let us say we have the information contained in the *ideal database*

$$\mathbf{e}_1 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_1) = \mathbf{a}_1\,, \quad \mathbf{e}_2 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_2) = \mathbf{a}_2\,, \quad \mathbf{e}_3 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_3) = \mathbf{a}_3\,, \tag{1.5}$$

where $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3 \in \mathbb{R}^3$. This dataset could also be represent by Table 1.1.

Now, construct the 3×3 matrix A whose columns are formed by the images of the vectors $\mathbf{e}_j$, $\mathcal{F}(\mathbf{e}_j)$, for $j = 1, 2, 3$, i.e.:

$$\begin{aligned} A &= \begin{pmatrix} | & | & | \\ \mathcal{F}(\mathbf{e}_1) & \mathcal{F}(\mathbf{e}_2) & \mathcal{F}(\mathbf{e}_3) \\ | & | & | \end{pmatrix} \\ &= \begin{pmatrix} | & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{pmatrix}\,. \end{aligned} \tag{1.6}$$

Then, if a generic stimulus is denoted by $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, the reaction $\mathcal{F}(\mathbf{x})$ will be given by the product $A\mathbf{x} \in \mathbb{R}^3$. In fact, by $\mathcal{F}$'s linearity,

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= \mathcal{F}(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3) \\ &= x_1\mathcal{F}(\mathbf{e}_1) + x_2\mathcal{F}(\mathbf{e}_2) + x_3\mathcal{F}(\mathbf{e}_3) \\ &= \begin{pmatrix} | & | & | \\ \mathcal{F}(\mathbf{e}_1) & \mathcal{F}(\mathbf{e}_2) & \mathcal{F}(\mathbf{e}_3) \\ | & | & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \end{aligned}$$

or, simply put, the reaction $\mathcal{F}(\mathbf{x})$ is given by the product $A\mathbf{x} \in \mathbb{R}^3$,

$$\mathcal{F}(\mathbf{x}) = A\mathbf{x}\,. \tag{1.7}$$

■

In the previous example, determining the model, that is, choosing a specific linear function $\mathcal{F}$, boils down to finding A, as can be seen immediately from Eq. (1.7).

We are now in a good position to define a third general problem. Although it is not as fundamental as the two presented in Section 1.1, even so it is equally important, and it is instrumental in the resolution of the former,

P_3: From a structured data set, determine matrix A.

If, as we assume, Eq. (1.5) is known to us, i.e., we have the information contained in that equation, or equivalently in Table 1.1, we can see that, based on Eq. (1.6), it is more than automatic to obtain A.

Now, let us consider the more general situation when we know $\mathcal{F}$ in some given basis of $\mathbb{R}^3$, not necessarily the canonical one.

Example 1.2. Identification using more general ideal data. Assume the values of $\mathcal{F}$ in a basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of $\mathbb{R}^3$ are known[16]. That is, let us say that the ideal database

$$\mathbf{u}_1 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_1) = \mathbf{v}_1 \,, \quad \mathbf{u}_2 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_2) = \mathbf{v}_2 \,, \quad \mathbf{u}_3 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_3) = \mathbf{v}_3 \,, \tag{1.8}$$

is available to us. Clearly, this information could be organized in a table similar to Table 1.1.

This is motivated by the fact that, for practical experimental[17], economical, or even ethical reasons, it is not always possible to apply the rather special choice of stimuli $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, but maybe it is possible to apply stimuli $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$.

Since the columns of matrix A in Eq. (1.6) are given by the images of the canonical vectors by $\mathcal{F}$, and since we have to determine them from the information in Eq. (1.8), we must, in the first place, write the canonical vectors in terms of the elements of the basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. We will use $\mathbf{e}_1$ as an example. Let c_1, c_2 and c_3 be the only scalars such that

$$\mathbf{e}_1 \quad = \quad c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 \,.$$

Let U be the matrix whose columns are the vectors $\mathbf{u}_i$, $i = 1,2,3$ and let V be the matrix such that its columns are $\mathbf{v}_i$, i.e.,

$$U = \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix}, \quad \text{and} \quad V = \begin{pmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{pmatrix}. \tag{1.9}$$

Also, let $\mathbf{c} = (c_1, c_2, c_3)^T$. With this notation,

$$\begin{aligned} \mathbf{e}_1 &= c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 \\ &= \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}, \end{aligned}$$

or, simply,

$$\mathbf{e}_1 = U\,\mathbf{c}\,. \tag{1.10}$$

[16] See page 189 in the Appendix A, to recall the definition of a basis of a vector space.

[17] This issue is exemplified in the last paragraph of Section 0.2, page 2.

Now, multiplying both sides of Eq. (1.10) by the inverse of matrix U, U^{-1}, and since $U^{-1}U = \mathcal{I}$ is the 3×3 identity matrix, we have

$$\mathbf{c} = U^{-1}\mathbf{e}_1 \,. \tag{1.11}$$

Notice that to obtain $\mathbf{c}$ explicitly corresponds to finding the solution of a system of linear equations (see Eq. (1.10), where the unknown is $\mathbf{c}$). Determination of the first column of A results now from the linearity of $\mathcal{F}$, and Eqs. (1.8), (1.9), and (1.11),

$$\begin{aligned}\mathcal{F}(\mathbf{e}_1) &= \mathcal{F}(c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3)\\ &= c_1\mathcal{F}(\mathbf{u}_1) + c_2\mathcal{F}(\mathbf{u}_2) + c_3\mathcal{F}(\mathbf{u}_3)\\ &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 \;=\; V\mathbf{c}\\ &= VU^{-1}\mathbf{e}_1 \,.\end{aligned}$$

If we do the same for $\mathbf{e}_2$ and $\mathbf{e}_3$, we can see that $\mathcal{F}(\mathbf{x}) = A\mathbf{x}$ with

$$A = VU^{-1} \,. \tag{1.12}$$

Thus, from Eq. (1.7) and the dataset, Eq. (1.8), we have just determined the model in this case,

$$\mathcal{F}(\mathbf{x}) = VU^{-1}\mathbf{x} \,.$$

■

It is worthwhile to pay attention to this result. Equation (1.12) justifies the following assertion[18]:

knowing a linear transformation in a basis[19] *leads to the full knowledge of it*[20].

[18] This is an interesting result. We have a function defined in $\mathbb{R}^3$. The set of all functions from $\mathbb{R}^3$ to $\mathbb{R}^3$ is a huge set. In principle, we have an enormous variety of possibilities to build a function whose domain and codomain are $\mathbb{R}^3$. Our task is simply this: we have to choose where to send through this function a very large number of points of the domain, having an equally large number of possibilities on the codomain from where to choose. Because it is a linear function, that is, because it satisfies a few simple algebraic rules, Eqs. (1.2) and (1.4), the number of possibilities to define the function is drastically reduced. In fact, knowing the value of the function in three randomly chosen points of $\mathbb{R}^3$ virtually determines the value of the function in all the other points. Of course, it is not just any three points that can be used, since, in particular, the origin cannot be chosen as one of those points. Technically, those three points must form a basis of $\mathbb{R}^3$. But, anyway, the limitation imposed by linearity is awesome: it reduces an uncountable set of information to a finite set with three data.

[19] In this example, to know a linear transformation on a basis is to have Eq. (1.8) or Eq. (1.9), (since both have the same structured information content).

[20] In the example, to know $\mathcal{F}$ is, by Eq. (1.7), to know Eq. (1.12).

1.5 Idealized Data Processing: Solution Operator

Notice that questions P_1 to P_3 are of different nature. Questions P_1 and P_2 are *natural* questions which can be posed by anyone who has devoted attention to the behaviour of the black box. However, question P_3 can only be asked by a *modeler*, that is, someone who tries to create a mathematical model to describe the behaviour of the black box.

After having characterized a model, problems P_1 to P_3 can be solved from the point of view of the mathematical model. This is very simple in this case, and that is what we will do next.

For the first one, given $\mathbf{x}$, an input signal, the solution is simply the product $A\mathbf{x}$. The second one is solved by means of the inverse of A. If $\mathbf{y}$ is the output, the input that generates it is $A^{-1}\mathbf{y}$.

Moreover, the third problem, determining A itself, can also be presented in mathematical terms. The determination of A was given in Eq. (1.12), $A = VU^{-1}$.

Schematically, we show in Table 1.2 the information required to answer each problem and the number crushing procedure that needs to be performed in these elementary linear algebra problems. It is customary to say that P_1 is a *direct problem* while P_2 and P_3 are *inverse problems*. This terminology is not in contradiction[21] with the kind of mathematical tasks involved to solve each one, see Table 1.2. Moreover, P_2 is an example of a so-called *reconstruction problem* and P_3 an *identification problem*.

It is now plain to see that, since the solutions to problems P_1 and P_2 depend on A, we need, in principle, to solve problem P_3 and only then deal with the other two.

On the other hand, as we will see in Section 1.8, knowing how to solve direct problems can be used for inverse problems resolution, if the notion of *solving* is made flexible.

In some applications, problems P_2 and P_3 can appear combined. One such example will be discussed in Chapter 4. There, one wants to recover a real image from a distorted one obtained from an experimental device. However, the distortion caused by the experimental technique is not known in advance.

1.6 Mind the Gap: Least Squares

We remark that, in the previous section, the solutions to problems P_1 to P_3 were constructed with, virtually, no contact with experiments, only in the realm of the mathematical model. In particular, it is assumed that the information in Eq. (1.8), which could come from an experiment, is ideally known, i.e., without errors.

In this section we will bring the mathematical model in contact with reality, using the physical model and the experimental data as a bridge. The criterion chosen to make such a contact possible is not a part of mathematics. However, its application is mathematical (or algorithmic).

[21] See, however, Exercise 1.3.

Table 1.2 Problem's nature and its data processing features

Problem	**Problem nature**	**Problem data**	**Solution-operator**	**Algorithm**
		Required data	Data processing task	
P_1	Direct	$(\mathbf{x}, A)$	$\mathbf{y} \hookleftarrow A\mathbf{x}$	matrix by vector multiplication
P_2	Inverse Reconstruction	$(\mathbf{y}, A)$	$\mathbf{x} \hookleftarrow A^{-1}\mathbf{y}$	matrix by vector multiplication
P_3	Inverse Identification	$U = (\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^3)$, $V = (\mathbf{v}^1,\mathbf{v}^2,\mathbf{v}^3)$	$A \hookleftarrow VU^{-1}$	matrix by matrix multiplication

The introduction of reality in the model begins with the answer to question P_3, basing it not on an ideal database, Eq. (1.8), but from a real database, made up of measurements with all the ambiguity thereof, that is, with all the contradictions between the experimental data and the hypotheses H_1 to H_3, listed on page 11, that may exist by chance (compare with Fig. 1.3, page 19).

In this case, the *estimation* of *m*odel parameters or the *identification of the model* corresponds to the determination of A, from experimental data[22].

Summing up, in the characterization stage, we choose a class of models. In the stage of estimation or identification, we pick, based on real data, one of the models within that class.

Example 1.3. Identification using experimental data. Let

$$\left\{(\mathbf{x}^k, \mathbf{y}^k) \in \mathbb{R}^3 \times \mathbb{R}^3, \text{ for } k = 1, \ldots, n\right\} \tag{1.13}$$

be the experimental data set, formed by ordered pairs (couples) of input and output signals. We say that the data is *perfectly representable* (or that it *can be interpolated*) by a linear function if there exists a matrix A such that

$$A\mathbf{x}^k = \mathbf{y}^k \quad \text{for all } k = 1, \ldots, n\,. \tag{1.14}$$

Otherwise we say that the data is not perfectly representable by a linear function. The one dimensional case is illustrated in Fig.1.3.

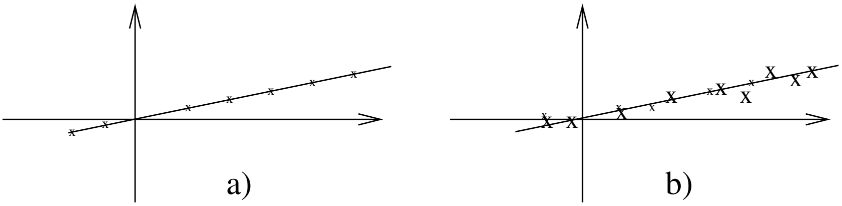


Fig. 1.3 Here are represented two data sets consisting of several pairs of real numbers, i.e., elements of $\mathbb{R}^2$. The first one may be viewed as related to an ideal data set and the second to a real, experimental, data set. (a) A set of input and output signals perfectly representable by a linear function. (b) The set of input and output signals displayed here is not perfectly representable by a linear function. However, we can choose a linear model to represent it, by means, for example, of the least squares method.

If the data is perfectly representable by a linear function we choose some input signals, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, that form a basis of $\mathbb{R}^3$ with the corresponding output signals, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, and Eqs. (1.9) and (1.12) define A.

[22] There are lots of ways to call this. Statisticians prefer, perhaps, *estimation*. You would *calibrate* the model, were you an engineer. Electrical engineers may prefer to speak of identification. Those which investigate artificial intelligence may *train* the model. Some call it *determination* of parameters, fitting the model to the data, or model selection.

If we want to keep the linear model even when the data is not perfectly representable by a linear function[23], we must relax the condition in Eq. (1.14). We consider now a way we can do it. Let

$$|\mathbf{v}| = (\sum_{j=1}^{3} v_j^2)^{\frac{1}{2}} = ((v_1)^2 + (v_2)^2 + (v_3)^2)^{\frac{1}{2}} ,$$

be the *Euclidean norm* of

$$\mathbf{v} = (v_1, v_2, v_3)^T \in \mathbb{R}^3 .$$

Also, let x_j^k be the j^{th} coordinate of $\mathbf{x}^k$, that is,

$$\mathbf{x}^k = (x_1^k, x_2^k, x_3^k)^T .$$

Denote by $M(3,3)$ the set of real 3×3 matrices. For $B \in M(3,3)$, we define half the *quadratic error function*[24],

$$\begin{aligned} E(B) &= \frac{1}{2} \sum_{k=1}^{n} |B\mathbf{x}^k - \mathbf{y}^k|^2 \\ &= \frac{1}{2} \sum_{k=1}^{n} \sum_{i=1}^{3} \left[\left(\sum_{j=1}^{3} B_{ij} x_j^k \right) - y_i^k \right]^2 . \end{aligned} \tag{1.15}$$

Note that the data is perfectly representable by A if and only if $E(A) = 0$.

We choose A, the *matrix defining the linear model*, as the one, within all $B \in M(3,3)$, that minimizes E,

$$A = \text{argmin}_{B \in M(3,3)} E(B) ,$$

that is,

$$E(A) = \min_{B \in M(3,3)} E(B) .$$

■

This criterion is known as the *least squares method*. It bridges the gap between the mathematical model and the experimental data. We stress that the determination of A satisfying the aforementioned criterion is, strictly speaking, a mathematical problem.

[23] One reason to want to keep the linear model is its simplicity which is a value that is worth to strive for.

[24] Here, and elsewhere in this book, we use 1 / 2 for a slight simplification in the critical point equation. There is no need for it.

1.7 Optimal Solution

To obtain the optimal solution, the minimum point of E in Eq. (1.15) must be determined. We search for this minimum by means of the *critical point equation*,

$$\frac{\partial E}{\partial B_{lm}} = 0\,, \text{ for } l,m = 1,2,3\,.$$

We have

$$\frac{\partial B_{ij}}{\partial B_{lm}} = \delta_{il}\delta_{jm}\,,$$

where δ_{il}, the *Krönecker's delta*, is a notation[25] for the elements of the identity matrix, $\mathcal{I}$, i.e.,

$$\delta_{il} = \begin{cases} 1, & \text{if } i=l \\ 0, & \text{if } i\neq l \end{cases}\,.$$

Thus,

$$\begin{aligned}
\frac{\partial E}{\partial B_{lm}} &= \sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3} B_{ij}x_j^k\right) - y_i^k\right]\left(\sum_{j=1}^{3}\frac{\partial B_{ij}}{\partial B_{lm}}x_j^k\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3} B_{ij}x_j^k\right) - y_i^k\right]\delta_{il}x_m^k \\
&= \sum_{k=1}^{n}\left[\left(\sum_{j=1}^{3} B_{lj}x_j^k\right) - y_l^k\right]x_m^k \\
&= \sum_{j=1}^{3} B_{lj}\left(\sum_{k=1}^{n} x_j^k x_m^k\right) - \sum_{k=1}^{n} y_l^k x_m^k \\
&= \left(\sum_{j=1}^{3} B_{lj}C_{jm}\right) - D_{lm}\,,
\end{aligned} \tag{1.16}$$

where

$$C_{jm} = \sum_{k=1}^{n} x_j^k x_m^k\,, \text{ and } D_{lm} = \sum_{k=1}^{n} y_l^k x_m^k\,. \tag{1.17}$$

The *critical point equation* for the critical point B (a matrix) is rewritten as

$$BC = D\,,$$

where C and D are known matrices, Eq.(1.17), determined from the data. The solution is then given by

$$B = DC^{-1}\,. \tag{1.18}$$

The expression for B looks like Eq. (1.12).

[25] Using this notation, the product of $\mathcal{I}$ by a matrix A, $\mathcal{I}A$ that, evidently, equals A, yields the strange looking expression: $\sum_{j=1}^{3}\delta_{ij}A_{jk} = A_{ik}$.

1.8 A Suboptimal Solution

A strategy that can be used at times when solving inverse problem P_3 involves solving the direct problem P_1 several times. One guesses or estimates values of the parameters (comprising the entries of matrix A in the example), a successive number of times, and solve P_1 for those values, selecting the value of A that best fits the data, in a previously established sense. In other words, the estimation (which corresponds to the minimization of E) can be performed by a *net search*, or *sampling*, that we describe briefly here.

Several values of the parameters, i.e., several matrices, will be chosen in succession which we shall denote by $A^1, A^2, \ldots, A^m$. They define a *grid* or *net* in $M(3,3)$. With them, the direct problem P_1 is solved successive times. That is, $A^j\,\mathbf{x}^k$ is computed, for $k = 1, \ldots, n,\ j = 1, \ldots, m$. We use that information to compute

$$E(A^j),\ j \quad = \quad 1, \ldots, m\,,$$

with E defined by Eq. (1.15).

The strategy to choose the next matrix in the sequence, A^{m+1}, or decide to stop the search, and to keep A^m, or any other of the previous matrices, $A^1, A^2, \ldots, A^{m-1}$, as solving the identification problem, is the defining step in several modern, non-gradient based, optimization methods. Nowadays, several methods employ different strategies in choosing the grid, as for instance in simulated annealing, [46, 50, 71], and others, so-called, metaheuristics, [69]. This is a *suboptimal* strategy, due to the fact that, since

$$\{A^1, \ldots, A^m\} \quad \subset \quad M\,,$$

we have

$$\min\nolimits_{B\in\{A^1,\ldots,A^m\}} E(B) \geq \min_{B\in M} E(B)\,. \tag{1.19}$$

Sure enough some limit theorem, when $m \to +\infty$, can be pursued. This is too far from our goal in this book.

1.9 Application to Darcy's Law

Here, we present an application of the model we have been discussing in this chapter. We trade the general black box model by the study of the flow of a fluid in a saturated *porous medium*. In such a medium, the flux, $\mathbf{u}$, and the gradient of the pressure, ∇p, satisfy *Darcy's law* ([28], [56]),

$$\mathbf{u} = -\frac{1}{\mu} K\nabla p\,,\quad \text{for } \mathbf{x} \in \Omega\,,$$

where K is the *permeability tensor* of the medium, $\mu > 0$ is the viscosity of the fluid, and $\Omega \subset \mathbb{R}^3$ is the porous region. Also, $p : \Omega \to \mathbb{R}$ and $\mathbf{u} : \Omega \to \mathbb{R}^3$. In general, the permeability $K = K(\mathbf{x})$ is a matrix-valued function,

$$K : \Omega \to M(3,3)\,,$$

where, we recall, $M(3,3)$ represents the set of 3×3 real matrices. In simple words, the permeability measures the *easiness* for the fluid to go through the porous medium. If K is a constant function, the porous medium is called *homogeneous*, otherwise it is a *heterogeneous porous medium*. If the medium is heterogeneous, the easiness of flow varies along the medium. If K is a scalar multiple of the identity matrix, $K = k\mathcal{I}$, then the medium is said to be *isotropic*. Otherwise, it is an *anisotropic* porous medium (in which case, the easiness of flow differs depending on the direction of the pressure gradient). Just to practice the terminology, we can have an isotropic medium, with k changing in space, in which case it is also heterogeneous.

Assume that we are investigating the permeability of a homogeneous anisotropic porous medium, and that we have a table containing several measurements of (vector) fluid flows, $\mathbf{u}^k$ for given applied pressure gradient, $\xi^k = -\nabla p$,

$$\xi^k \in \mathbb{R}^3 \, , \mathbf{u}^k \in \mathbb{R}^3 \, , \qquad k = 1, \ldots, n \ .$$

Assume also that the fluid viscosity, μ, is known. Then, following Eq. (1.15), we define half the quadratic error function,

$$\begin{aligned} E(K) &= \frac{1}{2}\sum_{k=1}^{n} |\frac{1}{\mu}K\xi^k + \mathbf{u}^k|^2 \\ &= \frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3}\frac{1}{\mu}K_{ij}\xi_j^k\right) + u_i^k\right]^2 \ . \end{aligned} \tag{1.20}$$

This function is to be minimized to obtain the permeability tensor of the porous medium, i.e., the permeability tensor K_* satisfies

$$K_* = \mathrm{argmin}_{K\in M(3,3)} E(K) \, .$$

Exercises

1.1. Show that: (a) Eq. (1.1) is equivalent to Eq. (1.2); (b) Eq. (1.3) is equivalent to Eq. (1.4).

1.2. Verify the validity of Eq. (1.12).

1.3. Assume that the mathematical model relating stimuli $\mathbf{x} \in \mathbb{R}^3$ and reactions $\mathbf{y} \in \mathbb{R}^3$ is given by $B\mathbf{y} = \mathbf{x}$, where B is a 3×3 matrix. Construct a table similar to Table 1.2 in this case. Read critically the paragraph where the footnote 21 is called, page 17.

1.4. (a) Verify the assertion in the sentence following Eq. (1.15). (b) Give conditions on the data, Eq. (1.13), such that there is only one solution to $E(A) = 0$.

1.5. (a) Compute $\sum_{j=1}^{3} \delta_{ij}A_{jk}$.

(b) Check the details on the derivation of Eq. (1.16).

(c) From data, Eq. (1.13), define matrices $X = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ and $Y = (\mathbf{y}^1, \ldots, \mathbf{y}^n)$. Show that $C = XX^T$ and $D = YX^T$, where C and D are defined in Eq. (1.17).

(d) If X is invertible (in particular, X must be a 3×3 matrix), show that the expression for B, Eq. (1.18), reduces to Eq. (1.12).

1.6. Determine the critical point equation for function $E = E(K)$ given by Eq. (1.20).

1.7. Simple regression model. Consider the *simple regression model*

$$y = a + bx \,, \tag{1.21}$$

where a and b are called *regression coefficients*, a is the *intercept* and b is the *slope*, x is called *regressor*, predictor or independent variable, and y is the *response* or dependent variable. Assume that you have pairs of measurements $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$, of regressor and response variables. The *residual* equation is

$$r_i = y_i - (a + bx_i) \,,$$

and half the *quadratic error* function is

$$E(a{,}b) = \frac{1}{2}\sum_{i=1}^{n} r_i^2 = \frac{1}{2}\sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \;.$$

(a) Obtain the *critical point* equation

$$\nabla E = \left(\frac{\partial E}{\partial a}, \frac{\partial E}{\partial b}\right) = (0, 0) \,. \tag{1.22}$$

(b) Denote the solution of Eq. (1.22) by $(\hat{a}, \hat{b})$, and show that

$$\hat{a} = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \,, \tag{1.23a}$$

$$\hat{b} = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \,. \tag{1.23b}$$

Hint. Write the equations in matrix form and make use of the expression for the inverse of a 2×2 matrix,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \frac{1}{AD - BC}\begin{pmatrix} D & -B \\ -C & A \end{pmatrix} .$$

(c) It is worthwhile to verify the dimensional correctness of Eq. (1.23). Check this.
Hint. Let the units of a variable x be denoted by $[x]$. Typical values of $[x]$ would be L for length, M for mass, and T for time. Then, if $[x] = X$ and $[y] = Y$, one can check from Eq. (1.23) that $[\hat{a}] = Y$ and $[\hat{b}] = Y/X$. This is compatible with Eq. (1.21) since b is multiplied by x to, partly, produce y, $[b][x] = [y]$, and then, $[b] = [y]/[x] = Y/X$.

(d) In the setting of statistics, the following notations are usual,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\,, \text{ and}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\,.$$

Show that

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2\,, \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})\,.$$

(e) Show that

$$\hat{a} = \bar{y} - \hat{b}\bar{x}\,,$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2} = \frac{S_{xy}}{S_{xx}}\,.$$

(f) Let $\mathcal{H}(E)$ denote the *Hessian* of E, that is, the matrix of the second order derivatives of E,

$$\mathcal{H}(E) = \begin{pmatrix} \frac{\partial^2 E}{\partial a^2} & \frac{\partial^2 E}{\partial a \partial b} \\ \frac{\partial^2 E}{\partial b \partial a} & \frac{\partial^2 E}{\partial b^2} \end{pmatrix}\,.$$

Compute $\mathcal{H}(E)\,|_{(\hat{a},\hat{b})}$.

(g) Determine an expression for the *eigenvalues*[26] of $\mathcal{H}(E)$.

(h) Show that the eigenvalues of $\mathcal{H}(E)$ are positive, whenever $n \geq 2$.
Hint. Show that $\alpha \pm \sqrt{\alpha^2 - \beta} > 0$ whenever $\alpha > 0$ and $\alpha^2 \geq \beta$.

(i) Use the spectral theorem to show that $(\hat{a}, \hat{b})$ is a minimum point of E since $\mathcal{H}(E)$ is a positive-definite matrix[27].

[26] For a reminder on how to compute eigenvalues in simple cases, see Example A.1, page 192.

[27] Further information about regression models can be seen in [55]. In particular, the modeling of residues as a probabilistic distribution is discussed. Usually, the normal distribution is used, which amounts to introducing another parameter, σ, in the model present in Eq. (1.21), corresponding to the standard deviation of the normal, leading to the model $y = a + bx + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$ stands for the normal distribution with zero mean and variance σ^2.

1.8. Consider Darcy's law for a homogeneous, isotropic porous medium, in one dimension. Use previous exercise with $a = 0$ in Eq. (1.21), to model the relationship between flux and pressure gradient. Determine an appropriate expression for an estimator of the scalar permeability, by mimicking the steps proposed in Exercise 1.7.

Chapter 2
Fundamental Concepts in Inverse Problems

The final answer to several problems can be reduced to evaluating a function—the solution function or the solution operator—and in the case of inverse problems it is not different. This is the point of view of a mathematician — insisting in the use of the notion of function. Not that one can always come about with the solution operator explicitly, but we can think abstractly on it and deduce its properties. This justifies the treatment that we present in this chapter of some of the aspects and complications that arise in the evaluation of functions. Next, we discuss some general aspects of mathematical models and inverse problems. A few classification schemes of inverse problems, illuminating different aspects, are presented. These classifications are used in subsequent chapters.

At times, the functions we are dealing with are quite complex, or are given in an extremely intricate way (for example, the function happens to be the solution of a partial differential equation [PDE]). In such a case it is almost always unavoidable to resort to a computer to produce a numerical value, i.e., a numerical solution. In this case the knowledge of the properties of the solution operator turns out to be very useful, even when we do not have the solution operator explicitly. In any case, the evaluation of the function (solution)—the final result—will be within a certain range of error, which varies mainly due to the following characteristics:

1. The problem is more complicated, or ill-behaved;
2. The function is more ill-behaved;
3. The way in which the function is evaluated in the computer (the algorithm) can be better or worse.

Case 1 is related to the need of regularizing the problem and we will present that concept in Chapter 3. Presently we deal with cases 2 and 3. Once a (small) error is introduced in a computation —through an experimental datum or due to round-off— it affects the final outcome. The error in the result can: (i) be reduced; (ii) remain small; (iii) be amplified.

There are two notions that can help us to understand *error dynamics* when evaluating a function: (a) the *condition* of the function being evaluated; and (b) the *stability*[1] of the algorithm used to evaluate it.

Condition will be dealt with in Sections 2.1 to 2.4 and algorithm stability in Section 2.5. Section 2.6 covers some questions related to existence and uniqueness. The notion of well-posed problem in the sense of Hadamard is considered in Section 2.7. In Section 2.8 a very simple classification of inverse problems is presented.

[1] The word stability can have several meanings. It is used even to name the notion of condition, in the sense that will be defined in this chapter, depending on authors. Caution is to be exercised as its meaning depends on the context.

2.1 Condition of Function Evaluation

Evaluation of a function at a given point can be *well* or *ill* conditioned. This is an intrinsic property of the function being evaluated and it does not depend on approximations.

Qualitatively it is said that the evaluation is *well-conditioned* if a small error in the point where the function is evaluated does not affect greatly the value of the function. If, however, a small error in the evaluation point leads to a large error in the value of the function, the evaluation is *ill-conditioned*.

It is possible to identify the notion of well-conditioned evaluation with continuity. Nevertheless, we desire a more restrictive notion, in the sense that well-conditioned implies continuity, but not the other way around. This shall be important when working with finite-precision arithmetic, and will allow us to distinguish different behaviours among continuous functions. Given a function, its qualitative behaviour can even depend on the region of the domain, having places where its evaluation is well-conditioned, and others where it is ill-conditioned. Let us see an example to illustrate this discussion.

Example 2.1. Evaluation of a rational function. Consider the evaluation of the function $f(x) = 1/(1-x)$. The computation of $f(x)$ is:

(a) *ill-conditioned*, if x lies near 1, (but, of course, x must be different from 1);

(b) *well-conditioned*, otherwise.

We shall treat these two cases next:

(a) When x is near 1.

Assume $x = 1.00049$ and that in the computation we use an approximate value, $x^* = 1.0005$. In this case, the absolute error of the evaluation is:

$$e_{\text{abs}} \quad = \quad f(x^*) - f(x) = -10^3/24.5\,.$$

We remark that an error of $10^{-5} = x^* - x$ in the data led to an evaluation error of $-10^3/24.5$. The error is magnified by the multiplication factor

$$m = \frac{\text{error in the result (of the evaluation)}}{\text{error in the point (of the domain)}} = -\frac{10^3/24.5}{10^{-5}}\,,$$

which, in absolute value satisfies $c = |m| > 10^6$.

(b) When x is far from 1.

When x is far from 1, the previous magnification phenomenon does not occur. Let $x = 1998$, and $x^* = 2000$ an approximation of x. Then the absolute error in the evaluation is

$$e_{abs} = \frac{1}{1-2000} - \frac{1}{1-1998} = \frac{2}{1999 \cdot 1997}\,.$$

Thus, the amplification factor of the error is $(1999 \cdot 1997)^{-1} < 10^{-6}$, *effectively* reducing the error. ■

2.2 Condition as a Derivative Bound

To study how a data error affects the evaluation of a function f, let x be the point of interest and let x^* be its approximation, and consider the quotient

$$m = \frac{\text{error in the result (evaluation)}}{\text{error in the datum (domain)}} = \frac{f(x^*) - f(x)}{x^* - x}\,. \tag{2.1}$$

Of course, Eq. (2.1) is *Newton's quotient* of f, a preliminary step in the definition of the derivative of a function. In the limit, $x^* \to x$, we have $m \to f'(x)$ and we define $f'(x)$ as the error *multiplication factor* of the evaluation error of f at x. This is a *local* quantity, coming from a local operator, $\frac{d}{dx}$.

Definition 2.1. Given[2] $f : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$ of class C^1, $c_f(x) = |f'(x)|$ is the *condition number* of the (*evaluation*) of f at x. We also say that the evaluation of f at x is *well-conditioned* if $c_f(x) \leq 1$ and *ill-conditioned* if $c_f(x) > 1$.

Thus, if the absolute value of the derivative of f is bounded by 1 at all the points of its domain of definition, i.e., if $|f'(x)| \leq 1$ for all $x \in \mathcal{D}$, then the evaluation of f is always well-conditioned. A simple example is the evaluation of $\sin x$, for any $x \in \mathbb{R}$.

2.3 Other Derivatives and Other Notions of Condition

It is not always convenient to access the sensitivity of the evaluation of a function by means of Eq. (2.1). As a matter of fact, Eq. (2.1) is written in terms of the quotient of two absolute errors: the evaluation absolute error and the data absolute error. At times, it is more interesting to consider relative errors. This leads to other possibilities to define the multiplication factor of the error in the evaluation. We present several alternatives in Table 2.1.

Table 2.1 Possible definitions for the multiplication factor of the error

numerator → denominator ↓	absolute error in the evaluation	relative error in the evaluation
data absolute error	(a) $\frac{f(x^*)-f(x)}{x^*-x} \sim f'(x)$	(b) $\frac{(f(x^*)-f(x))\,/f(x)}{x^*-x} \sim \frac{f'(x)}{f(x)}$
data relative error	(c) $\frac{f(x^*)-f(x)}{(x^*-x)/x} \sim x\,f'(x)$	(d) $\frac{(f(x^*)-f(x))\,/f(x)}{(x^*-x)/x} \sim \frac{xf'(x)}{f(x)}$

[2] We recall that $\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$ is called a function of *class* C^k if the derivatives of its component functions of order at least k exist and are continuous. A function of class C^0 is just a continuous function.

As done in the case of Eq. (2.1), we can propose the notion of multiplication factor by considering the limit $x^* \to x$, in the quotients of Table 2.1. Thus, we have the following *multiplication factor*[3], $m_f(x)$, at x:

(a) $f'(x)$: the (usual) derivative of f at x;

(b) $f'(x)/f(x)$: the *logarithmic derivative* of f at x (in fact, the derivative of $\ln f(x)$);

(c) $xf'(x)$: derivative (differential operator) without a special name;

(d) $xf'(x)/f(x)$: the *elasticity* of f at x (much used in economics).

Likewise, we define the *condition number* as the absolute value of the multiplication factor, $c_f(x) = |m_f(x)|$. In this case, we can talk about well or ill-conditioned evaluation for all of the condition numbers presented.

For all cases, from (a) to (d), the notion of well-conditioned evaluation is that the absolute value of the condition be less than or equal to one.

On the other hand, to be bounded by one can, sometimes, be considered too restrictive. We opt here for this criterion because in this case there is always a reduction of the error. That can be unnecessary, however. It must be pointed out that, in some applications, it is possible to work with values greater than one: a small amplification of the error in the data can be manageable. The transition value between well and ill-conditioned evaluations depends on the given problem and objectives.

2.4 Condition of Vector-Valued Functions of Several Variables

Consider now a vector-valued function of several variables,

$$\begin{aligned} \mathbf{f} : \mathbb{R}^n \supset \Omega &\quad\to\quad \mathbb{R}^m \\ \mathbf{x} = (x_1, \ldots, x_n)^T &\quad\mapsto\quad \mathbf{f}(\mathbf{x}) = (f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n))^T \end{aligned}$$

where Ω is a subset of $\mathbb{R}^n$. As done previously for functions of a single variable, we can define different *condition numbers* (of the evaluation) of function $\mathbf{f}$ at $\mathbf{x}$. The important thing to keep in mind is that the condition is the norm of the multiplier of the error in the data determining the error in the evaluation.

To begin, let us recall that *Taylor's formula*[4] gives

$$\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}) = \mathcal{J}\mathbf{f}_x \cdot (\mathbf{x}^* - \mathbf{x}) + \mathcal{O}\left(|\mathbf{x}^* - \mathbf{x}|^2\right), \text{ as } \mathbf{x}^* \to \mathbf{x}\,, \tag{2.2a}$$

[3] When one desires to be more specific, one may say, for example, that $f'(x)/f(x)$ is the multiplication factor (of the evaluation) of the *relative error* (in the result) with respect to the *absolute error* (in the datum).

[4] See precise statements of Taylor's formulae on page 204 of the Appendix A.

Here, $\mathcal{J}\mathbf{f}_x$ is the *Jacobian matrix* of $\mathbf{f}$, the $m \times n$ matrix of first-order derivatives of $\mathbf{f}$, evaluated at $\mathbf{x}$,

$$\mathcal{J}\mathbf{f}_x = \left.\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}\right|_x .$$

Also, $\mathcal{O}$ is the usual big-$\mathcal{O}$ order symbol, and the norm of a vector $\mathbf{v}$ is given by

$$|\mathbf{v}| = \left(v_1^2 + \ldots + v_n^2\right)^{\frac{1}{2}} .$$

Equation (2.2a) allows us to write

$$\frac{\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x})}{|\mathbf{f}(\mathbf{x})|} = \frac{1}{|\mathbf{f}(\mathbf{x})|}\mathcal{J}\mathbf{f}_x \cdot (\mathbf{x}^* - \mathbf{x}) + \mathcal{O}\left(\frac{|\mathbf{x}^* - \mathbf{x}|^2}{|\mathbf{f}(\mathbf{x})|}\right) , \tag{2.2b}$$

$$\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}) = |\mathbf{x}|\mathcal{J}\mathbf{f}_x \cdot \frac{(\mathbf{x}^* - \mathbf{x})}{|\mathbf{x}|} + \mathcal{O}\left(|\mathbf{x}^* - \mathbf{x}|^2\right) , \tag{2.2c}$$

$$\frac{\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x})}{|\mathbf{f}(\mathbf{x})|} = \frac{|\mathbf{x}|}{|\mathbf{f}(\mathbf{x})|}\mathcal{J}\mathbf{f}_x \cdot \frac{(\mathbf{x}^* - \mathbf{x})}{|\mathbf{x}|} + \mathcal{O}\left(\frac{|\mathbf{x}^* - \mathbf{x}|^2}{|\mathbf{f}(\mathbf{x})|}\right) , \tag{2.2d}$$

as $\mathbf{x}^* \to \mathbf{x}$, in all cases.

Due to this result, Eq. (2.2), we define the *multiplication matrices* (of the evaluation):

(a) absolute (error in the result) to absolute (error in the datum),

$$\mathcal{M}_{aa} = \mathcal{J}\mathbf{f}_x ; \tag{2.3a}$$

(b) relative (error in the result) to absolute (error in the datum),

$$\mathcal{M}_{ra} = \mathcal{J}\mathbf{f}_x / |\mathbf{f}(\mathbf{x})| ; \tag{2.3b}$$

(c) absolute (error in the result) to relative (error in the datum),

$$\mathcal{M}_{ar} = |\mathbf{x}|\mathcal{J}\mathbf{f}_x ; \tag{2.3c}$$

(d) relative (error in the result) to relative (error in the datum),

$$\mathcal{M}_{rr} = |\mathbf{x}|\mathcal{J}\mathbf{f}_x / |\mathbf{f}(\mathbf{x})| . \tag{2.3d}$$

It is worthwhile to compare these matrices with the one dimensional case, as shown in Table 2.1. These are generalizations, and therefore the notation is slightly more cumbersome, but the fundamental meaning remains.

Note that, from the definition of the norm of a matrix, Eq. (A5), and its properties, Eq. (A6), we have that

$$|\mathcal{J}_x\mathbf{f} \cdot (\mathbf{x}^* - \mathbf{x})| \leq |\mathcal{J}_x\mathbf{f}| \cdot |\mathbf{x}^* - \mathbf{x}| .$$

The condition number of the evaluation, in these cases, is the norm of the multiplication matrix,

$$c_f(\mathbf{x}) = |\mathcal{J}\mathbf{f}_x|\,, \text{ absolute to absolute} \tag{2.4a}$$

$$c_f(\mathbf{x}) = |\mathcal{J}\mathbf{f}_x|/|\mathbf{f}(\mathbf{x})|\,, \text{ relative to absolute} \tag{2.4b}$$

$$c_f(\mathbf{x}) = |\mathbf{x}|\cdot|\mathcal{J}\mathbf{f}_x|\,, \text{ absolute to relative} \tag{2.4c}$$

$$c_f(\mathbf{x}) = |\mathbf{x}|\cdot|\mathcal{J}\mathbf{f}_x/|\mathbf{f}(\mathbf{x})||\,, \text{ relative to relative} \tag{2.4d}$$

Example 2.2. Condition number of a matrix. Given a matrix A, we will analyze the condition of the function $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$.

For the linear function $\mathbf{f}$,

$$\mathcal{J}\mathbf{f}_x = A\,, \text{ for all } \mathbf{x}$$

The condition number, relative to relative, is given by (see Eq. (2.3d))

$$|\mathbf{x}|\,|A|/|A\mathbf{x}|\,.$$

Now, assume that A is invertible, and let $\mathbf{y} = A\mathbf{x}$. Therefore, $\mathbf{x} = A^{-1}\mathbf{y}$, and

$$\frac{|\mathbf{x}|\,|A|}{|A\mathbf{x}|} = \frac{|A^{-1}\mathbf{y}|}{|\mathbf{y}|}\,|A| \leq |A^{-1}|\,|A|\,,$$

due to Eq. (A6).

Thus, notwithstanding the point where the function $\mathbf{f}$ is being evaluated, the condition number is bounded by $|A^{-1}|\,|A|$. In this case, it is customary to say that

$$k(A) \quad = \quad |A^{-1}|\,|A|$$

is the *condition of matrix A*.

In general it is clear that the condition of the evaluation of $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ depends on the point $\mathbf{x}$ where such evaluation is to be done, and it can be less than $k(A)$. ■

Example 2.3. Condition in the *resolution* of a linear system of equations. Consider the linear problem

$$K\mathbf{x} \quad = \quad \mathbf{y}\,,$$

where $\mathbf{y} \in \mathbb{R}^n$ is a datum and $\mathbf{x} \in \mathbb{R}^n$ is the unknown. Assume that K is invertible so that $\mathbf{x} = K^{-1}\mathbf{y}$. We want to analyze the condition of the solution operator

$$\mathbf{y} \quad \mapsto \quad K^{-1}\mathbf{y}\,.$$

The condition number in this case, when considering the relation of the absolute error in the evaluation to the absolute error in the datum, Eq. (2.3a), is $|K^{-1}|$. ■

The previous example is used in Chapters 3 and 4.

Example 2.4. Difference between close numbers. It is known that the difference between two numbers that are close is *ill-behaved* when finite-precision arithmetic is used, as is usually the case with digital computers.

This fact is not directly related to finite-precision arithmetic, but to the intrinsic nature of the difference function, and to the fact that it is ill-conditioned for close numbers.

At the beginning, consider the function

$$\mathbb{R} \times \mathbb{R} \ni (x, y) \quad \mapsto \quad m(x, y) = x - y \in \mathbb{R}\,,$$

that calculates the difference between two numbers and let us compute its *elasticity* as defined in the following equation (compare with Eq. (2.3d) and verify their dissimilarity). We have

$$E\, m(x{,}y) = \left(\frac{x}{m} \frac{\partial m}{\partial x}\,,\ \frac{y}{m} \frac{\partial m}{\partial y} \right) = \left(\frac{x}{x-y}\,,\ -\frac{y}{x-y} \right)\,,$$

from whence

$$|E\, m(x, y)| = \sqrt{\frac{x^2 + y^2}{(x-y)^2}}\,. \tag{2.5}$$

Now, $|Em(x, y)|$ cannot be less than or equal to 1 (unless x and y have opposite signs, but then m would not be a difference, it would be a sum). Verify. Yet, it is possible to obtain regions of $\mathbb{R}^2$ in which the difference has a moderate elasticity, say, less than 2. Although there is an amplification of the error, it is a “small” amplification. On the other hand, if we choose y near x, $y = x + \epsilon$, with ϵ small, then the elasticity,

$$|E\, m(x, x + \epsilon)| = \sqrt{(x^2 + (x + \epsilon)^2)/\epsilon^2}\,,$$

can be arbitrarily large subject to ϵ being sufficiently small and $x \neq 0$.

Summing up, the problem of computing the difference between two numbers can be: (a) well-conditioned, if the numbers are far apart; (b) ill-conditioned if the numbers are close. Notice that this is not related to the use of finite-precision arithmetic, but will be observed in the presence of round-off errors when using such arithmetic.

Finally, it would be more correct to use the norm of the multiplication factor — the condition number of m — as obtained from Eq. (2.3d), instead of the norm of the elasticity, Eq. (2.4d). However, in this case, these two notions coincide. ■

2.5 Stability of Algorithms of Function Evaluation

As mentioned before, the notions of condition and stability are important in understanding how the errors alter the final outcome of a computation. The notion of condition of the evaluation of a function is intrinsic to the function whose outcome (image value) is to be computed, so it is unavoidable to deal with it as long as we work with that particular function. On the other hand, the notion of stability depends on the algorithm that is used to compute the value of the function.

Since there are several ways to evaluate a function, it is possible to select one that best fits our needs. An algorithm to compute the value of a function is *unstable* when the errors occurring through the intermediate steps (of the computation) are amplified. If this is not the case, the algorithm is *stable*.

An algorithm to compute the value of a function is constituted of several elementary steps. To study its stability every step must be analyzed separately, and the effect of the introduction of an error must be tracked.

Example 2.5. Evaluation of an algebraic function. The evaluation of $f(x) = \sqrt{x+1} - \sqrt{x}$ is well-conditioned, for all $x \geq 0$. We will show two algorithms for this evaluation: the more natural one is unstable, the other one is stable.

The evaluation of f is always well-conditioned, since the absolute value of its elasticity,

$$|Ef(x)| = \frac{1}{2}\left|\frac{x}{\sqrt{x+1}\,\sqrt{x}}\right| \leq \frac{1}{2}\,,$$

is always bounded by $\frac{1}{2}$.

We will consider two algorithms to evaluate f. The first one corresponds to the direct computation of

$$\sqrt{x+1} - \sqrt{x}\,,$$

and the second one is based on the algebraic expression,

$$\frac{1}{\sqrt{x+1} + \sqrt{x}}\,.$$

The two expressions are algebraically equivalent. In fact,

$$\sqrt{x+1} - \sqrt{x} \quad = \quad \frac{1}{\sqrt{x+1} + \sqrt{x}}\,,$$

for all $x > 0$.

To analyze the stability of the natural algorithms streaming from these two equivalent algebraic expressions, we depict in Fig. 2.1 the steps that make up each algorithm. The schemes in the figure help to prove that the second algorithm is preferable, as we shall see.

Each algorithm is assembled with the intermediate steps defined by the following functions,

$$\begin{aligned} p_1(x) &= \sqrt{x}\,, & p_2(x) &= x + 1\,, \\ p_3(x) &= \sqrt{x}\,, & p_4(y,z) &= z - y\,, \\ p_5(y,z) &= y + z\,, & p_6(x) &= 1/x\,. \end{aligned}$$

The steps of each one of the two algorithms correspond to some of these functions and the algorithms can be interpreted as compositions of these functions.

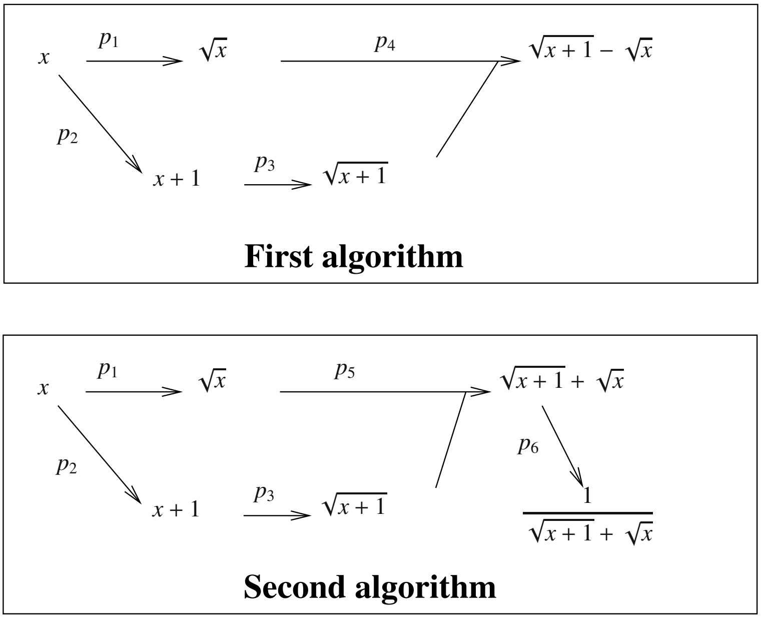


Fig. 2.1 Diagrams of the algorithms

Since the *elasticity of the composition* of functions is the product of the elasticity of the functions[5],

$$E\,(p_1 \circ p_2 \circ \ldots \circ p_n) = E\,(p_1) \times E\,(p_2) \times \ldots \times E\,(p_n)\ , \tag{2.6}$$

it suffices to analyze each step. We have that

$$|Ep_1(y)| = 1/2\,,\ |Ep_2(y)| = |y/(y+1)| \leq 1\,,\ Ep_3(y) = Ep_1(y)\,,$$

and, finally, we recall that the elasticity of the difference function, p_4, has already been analyzed in example 2.4.

Now, we see that step p_4 is crucial to decide if the first algorithm is stable or not. It will be stable if

$$y = \sqrt{x} \text{ and } z = \sqrt{x+1}$$

are not too close. However, due to the elasticity of p_4, we see that, if they are close, the multiplication factor will be large. Therefore, this algorithm will only work well

[5] We recall that the *composition* of two functions, p_1 and p_2, denoted by $p_1 \circ p_2$, is the function,

$$(p_1 \circ p_2)(x) = p_1(p_2(x))\,.$$

if x is very close to zero. For x far from zero, y and z will be close, and the algorithm will behave inadequately (it will be unstable).

Now let us check the second algorithm. Steps p_1, p_2 and p_3 are common to the first algorithm. Besides,

$$|Ep_5(y,z)| = \sqrt{(y^2+z^2)/(y+z)^2} \leq 1\,,$$

since we are dealing only with non-negative values of y and z. Also, $|Ep_6(y)| = 1$. Thus, we see that all the multiplication factors in the various steps of the second algorithm are less than or equal to 1, rendering it stable. ■

Example 2.6. Evaluation of a polynomial. Given the cubic polynomial

$$p(x) = (x - 99\pi)(x - 100\pi)(x - 101\pi)\,,$$

where π is given by the value of a pocket calculator, we will present two algorithms to evaluate p in $x_o = 314.15$. The first one will use directly the previous expression, the second one will be based in the power form of $p(x)$ and will use Horner's method.

Let us start by analyzing the condition of the evaluation. Since the *elasticity of the product* is the sum of the elasticities,

$$E(f \cdot g) = Ef + Eg\,, \tag{2.7}$$

we see that the elasticity of the evaluation of $p(x)$ is given by

$$Ep(x) = \frac{x}{x-99\pi} + \frac{x}{x-100\pi} + \frac{x}{x-101\pi}\,.$$

Thus, this evaluation is ill-conditioned at $x_o = 314.15$ due to the necessity of evaluating $(x - 100\pi)$ in this point. In fact,

$$\begin{aligned} Ep(314.15) &= \frac{314.15}{314.15-99\pi} + \frac{314.15}{314.15-100\pi} + \frac{314.15}{314.15-101\pi} \\ &\approx 100.2928 - 33905.86 - 99.7030 \approx -33905\,. \end{aligned}$$

Notwithstanding, the first algorithm is superior to the second one. In fact, in the first case, we obtain

$$p(x_o) = (x_o - 99\pi)(x_o - 100\pi)(x_o - 101\pi) = 0.091446\,,$$

while using *Horner's method* (starting the algorithm by the innermost expression, $x_0 - 300\pi$ and proceeding outwards), we obtain:

$$p(x_o) = ((x_o - 300\pi)x_o - 29999\pi^2)x_o - 999900\pi^3 = -0.229\,.$$

Evidently, this result contains numerical errors, since the answer should be positive. ■

The fact that a problem is ill-conditioned does not prevent us from computing its result. We must choose carefully the algorithm to be used, no matter if it is an ill or well-conditioned problem.

2.6 Questions on Existence and Uniqueness

Some of the difficulties in solving inverse problems are related to the available information: quantity (not sufficient data or seemingly overabundance thereof) and quality of information. We will illustrate these points by means of simple examples.

Let us suppose that the function that truly generates the phenomenon is

$$f(x) = 2x+1 \ .$$

In the inverse identification problem we assume that such function is unknown to us. We do assume, however, that we can determine to which class of functions the phenomenon, $f(x)=2x+1$, belongs, i.e., we characterize the *model*. The observation of the phenomenon allows us to characterize it as, say, belonging to the class of functions of the form

$$f_{a,b}(x) = ax + b \ ,$$

where a and b are arbitrary constants. From the available data we must then determine a and b, i.e., we must identify or select the model. We shall consider two situations, when one has exact data, or, otherwise, has real (noisy) data.

2.6.1 Exact Data

For the sake of the present analysis, we assume that the available data are exact. We then have three possibilities:

(a) **Not sufficient data.** The basic unit of information in this example corresponds to a point in the graph of the model. Assume known that the point $(1,3)$ belongs to the graph of f. It is obvious that this datum is not sufficient for us to determine a and b. As a matter of fact we only know that

$$f(1)=3 \text{ or } a+b=3 \ ,$$

It is then impossible to determine the model (find the values of a and b uniquely). ■

(b) **Sufficient data.** We know data $(1,3)$ and $(2,5)$. Thus, $a+b = 3$ and $2a+b = 5$, from whence we can determine that $a = 2$ and $b = 1$, and select the model $f(x) = 2x + 1$. ■

(c) **Too much data.** Assume now that it is known that the points $(0,1)$, $(1,3)$, and also $(2,5)$ belong to the graph of f. Then

$$a=2 \text{ and } b=1 \ .$$

It is plain to see that we have too much data, since we could determine a and b without knowing the three ordered pairs, being for that matter sufficient to know any two of such pairs. We point out that too much data does not cause problems in the determination of the model when exact data is used.

2.6.2 Real Data

In practice we do not have exact data, so it is best to have too much data, even if some repetition occurs. In spite of it, in many occasions, we only have a sufficient number of (non-repeated) data due, for example, to costs of acquiring data. Sometimes, real data will be called *noisy* data. We still have three possibilities, discussed below:

(a) **Insufficient data.** Datum $(1, 3.1)$ has an error—as we know, for our "phenomenon," $f(1) = 3$ and not 3.1. Moreover, this datum is insufficient, because we obtain only one relation between a and b,

$$a + b = 3.1\ ,$$

but cannot determine them individually, not even approximately. Another restriction must be imposed so the inverse problem has a unique solution. A concrete and unexpected example of how this can be achieved in real problems can be seen in Section 4.6. ■

(b) **Sufficient data.** Consider that we have the following approximate data: (1, 3.1) and (2, 4.9). Then, an approximation for a and b is obtained by substituting the data in a class of models,

$$\begin{cases} a + b = 3.1 \\ 2a + b = 4.9 \end{cases}, \tag{2.8}$$

which gives $a = 1.8$ and $b = 1.3$. ■

However, even with sufficient (but with errors, i.e., noisy) data, it is not always possible to estimate the parameters by imposing that the model fits or interpolates the data. Later, we will see on example 2.7 that clarifies this remark.

Alternatively, in these cases, we try to minimize a certain measure of discrepancy between the model and the data. In the example, for every proposed model within the characterized class, that is, for every pair (a, b), we would compute the difference between the data and what is established by the model and combine these differences in some way and minimize it.

For example, if the pair (a, b) is known, the model is given by $f_{a,b}$, and the point

$$(1, f_{a,b}(1))$$

should belong to the graph. This would be what the model establishes, the so-called *prediction*[6] of the model. The data however indicates that the point should be $(1, 3.1)$, so

$$f_{a,b}(1) - 3.1$$

[6] Here the word prediction means not only "foretelling the future" as is usual; it is foretelling in regions where data is not available. In models involving time, however, it carries the former meaning. In any case, we search for scientific predictions, as discussed in Chapter 1.

is a measure of the error due to the difference between what the model foretells and what actually happens. This must be done with all the experimental data. The results may be combined to create a *discrepancy measure* between what the model foretells with (a, b) and what the real data shows. Finally, we must find the value of the pair (a, b) that minimizes that measure.

As an example, let us consider the discrepancy (or error) measure given by

$$\begin{aligned} E(a,b) &= \frac{1}{2}\left[(f_{a,b}(1) - 3.1)^2 + (f_{a,b}(2) - 4.9)^2\right] \\ &= \frac{1}{2}\left[(a + b - 3.1)^2 + (2a + b - 4.9)^2\right] . \end{aligned}$$

The minimum point of E is given by its critical point, i.e., the point where the gradient of E is null,

$$\begin{aligned} 0 &= \frac{\partial E}{\partial a} = (a + b - 3.1) + 2(2a + b - 4.9) \\ 0 &= \frac{\partial E}{\partial b} = (a + b - 3.1) + (2a + b - 4.9) . \end{aligned}$$

We thus conclude that

$$a + b = 3.1 \quad \text{and} \quad 2a + b = 4.9 .$$

It is a coincidence that, due to the form of the function $f_{a,b}$, the system in Eq. (2.8), and the one just obtained, are the same. At times the problem obtained by interpolation as in Eq. (2.8) does not have a solution, while the one obtained by least squares, like the one we just deduced, is solvable. This is a way to reach a solution in real inverse problems. This subject is addressed in example 2.7.

(c) **Too much data.** Assume that

$$(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n),\ \text{with } n \geq 3 ,$$

are several experimental points associated with the "phenomenon" $f(x) = 2x + 1$.

It is unavoidable that these experimental data are contaminated by errors and imprecisions intrinsic to the measuring process. Thus, the data are usually *incompatible*, i.e., it is impossible to solve for a and b the system

$$\begin{aligned} y_1 - f_{a,b}(x_1) &= y_1 - (ax_1 + b) = 0 \\ y_2 - f_{a,b}(x_2) &= y_2 - (ax_2 + b) = 0 \\ &\vdots \\ y_n - f_{a,b}(x_n) &= y_n - (ax_n + b) = 0 . \end{aligned}$$

Usually we say that, since the system has n equations and only two unknown variables, it possibly has no solution.

In general we would say that there is no model in the characterized class of models that interpolates the data, i.e., the data cannot be fitted by the model.

We deal with this question from a geometrical point of view, in the context we are discussing. Note that the system can be rewritten as

$$a\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}+b\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}.$$

We introduce the notation $\mathbf{x}=(x_1,\ldots,x_n)^T$, $\mathbb{1}=(1,\ldots,1)^T$ and $\mathbf{y}=(y_1,\ldots,y_n)^T$. This notation allows the vector equation above to be written as

$$a\mathbf{x}+b\mathbb{1}=\mathbf{y}\,.$$

Thus, the solution of the system of equations can be rephrased as the problem of finding a linear combination of the vectors $\mathbf{x}$ and $\mathbb{1}$ to obtain $\mathbf{y}$. These vectors, $\mathbf{x}$ and $\mathbb{1}$, belong to $\mathbb{R}^n$, and they are only two. It is necessary n linearly independent vectors to represent an arbitrary vector of $\mathbb{R}^n$ as a linear combination of them. Therefore, it is very rare for one to be able to choose a and b in order that $\mathbf{y}=a\mathbf{x}+b\mathbb{1}$. That this is the case is easily visualized by means of a simple figure, see Fig. 2.2.

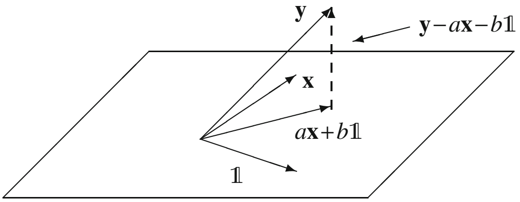

$\mathbf{y}$
$\mathbf{y}-a\mathbf{x}-b\mathbb{1}$
$\mathbf{x}$
$a\mathbf{x}+b\mathbb{1}$
$\mathbb{1}$

Fig. 2.2 Plane (2-dimensional subspace) in $\mathbb{R}^n$. Vector $\mathbf{y}$ does not belong to the plane spanned by $\mathbf{x}$ and $\mathbb{1}$, $\text{span}\,\{\mathbf{x},\mathbb{1}\}=\{a\mathbf{x}+b\mathbb{1}$, for all $a,b\in\mathbb{R}\}$.

Let us consider the method of least squares. Define the *error* or residual vector,

$$\mathbf{r}=\mathbf{y}-(a\mathbf{x}+b\mathbb{1})\,,$$

given as the vector of differences between the experimental measurements, $\mathbf{y}$, and the predictions of the model with coefficients a and b, $a\mathbf{x}+b\mathbb{1}$.

Effectively, what can be done is to choose a linear combination between $\mathbf{x}$ and $\mathbb{1}$ (i.e., choose a and b), in such a way that the functional error,

$$E(a,b)=\frac{1}{2}|\mathbf{r}|^2=\frac{1}{2}|\mathbf{y}-a\mathbf{x}-b\mathbb{1}|^2=\frac{1}{2}\sum_{i=1}^{n}(y_i-(ax_i+b))^2\ ,$$

is minimized. Since the sum stands for the error vector's squared norm, $|\mathbf{y}-a\mathbf{x}-b\mathbb{1}|^2$, a look at Fig. 2.2 suggests that it is equivalent to requiring that the error vector be orthogonal to the plane spanned by $\mathbf{x}$ and $\mathbb{1}$. Or, else,[7]

$$\langle \mathbf{x}, \mathbf{y} - a\mathbf{x} - b\mathbb{1} \rangle = 0 \quad \text{and} \quad \langle \mathbb{1}, \mathbf{y} - a\mathbf{x} - b\mathbb{1} \rangle = 0\,.$$

This can be written as

$$\mathbf{x}^T\left(\mathbf{y} - a\mathbf{x} - b\mathbb{1}\right) = 0\,, \quad \text{and} \quad \mathbb{1}^T\left(\mathbf{y} - a\mathbf{x} - b\mathbb{1}\right) = 0\,,$$

which leads to,

$$\begin{aligned} a\mathbf{x}^T\mathbf{x} + b\mathbf{x}^T\mathbb{1} &= \mathbf{x}^T\mathbf{y} \\ a\mathbb{1}^T\mathbf{x} + b\mathbb{1}^T\mathbb{1} &= \mathbb{1}^T\mathbf{y}\,, \end{aligned}$$

whence,

$$\begin{pmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbb{1} \\ \mathbb{1}^T\mathbf{x} & \mathbb{1}^T\mathbb{1} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{x}^T\mathbf{y} \\ \mathbb{1}^T\mathbf{y} \end{pmatrix}. \tag{2.9}$$

Defining $A = (\mathbf{x}, \mathbb{1})$, an $n \times 2$ matrix, Eq. (2.9) can be rewritten as

$$A^TA \begin{pmatrix} a \\ b \end{pmatrix} = A^T\mathbf{y} \tag{2.10}$$

which is usually called *normal equation*. Therefore, the determination of the model reduces to solving the linear system, Eq. (2.10). ■

We remark that, even though matrix A^TA may not be invertible, Eq. (2.10) will always have a solution. We shall treat this question later on. Assume, however, that A^TA is invertible. Then, the solution to the inverse problem can be represented by

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(A^TA\right)^{-1} A^T\mathbf{y}\,. \tag{2.11}$$

This is the solution of the inverse problem given by the *least squares method*, which corresponds to the evaluation of the function,

$$\mathbf{y} \quad \mapsto \quad \left(A^TA\right)^{-1} A^T\mathbf{y}\,.$$

It is pertinent here to recall the discussion of Section 2.5, regarding the stability of function evaluation algorithms. As a matter of fact, the algorithm suggested by the expression, $\left(A^TA\right)^{-1} A^T\mathbf{y}$ is not the best way to evaluate it; it can be *unstable* (depending on matrix A) and even inefficient from a computational point of view[8],

[7] The notation of inner product is recalled on page 189, Appendix A.

[8] The question of the inefficiency of the algorithms must be considered. Non-efficient methods can render impractical the use of a given algorithm. In the present case, depending on the size of A^TA, it may be very time consuming to find its inverse. Recall however that one wants to find $(a,b)^T$ satisfying Eq. (2.10).

because it presumes that the inverse of A^TA will be computed. The geometric interpretation of the problem at hand is the basis for the construction of alternative algorithms, stable and more efficient, see [35].

Assume now that A is invertible. Then A^T is also invertible, since

$$(A^T)^{-1} \quad = \quad (A^{-1})^T \, .$$

Therefore,

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad = \quad A^{-1}\left(A^T\right)^{-1} A^T\mathbf{y} = A^{-1}\mathbf{y} \, . \tag{2.12}$$

This result, Eq. (2.12), is valid whether the data are exact or not. ■

Example 2.7. Sufficient data and the least squares method. We consider now an example in which, although there are sufficient data, it is impossible to identify the model due to the existence of experimental errors.

Assume a phenomenon is given precisely by

$$g(x) = \frac{1}{1 + x^2/99} \ .$$

Thus, in particular, $g(1) = 0.99$. Again, assume that g is unknown and that, after the model has been characterized, the following class of models is obtained,

$$C = \left\{ g_c \text{ for all } c \in \mathbb{R} \text{ where } g_c(x) = \frac{1}{1 + (x - c)^2/99} \right\} \, .$$

Finally, let $(1, 1.1)$ be the only experimental datum.

Well the given datum is, in principle, sufficient to select the model, since only one parameter, c, is to be determined. However, if we try to determine c by interpolation, i.e., by means of the equation

$$g_c(1) = \frac{1}{1 + (1 - c)^2/99} = 1.1 \ ,$$

we see that it is impossible. In fact, for every value of c, $g_c(1)$ will be less than one. An adequate approach is to use the approximation of least squares previously presented.

Let

$$E(c) = \frac{1}{2}(1.1 - g_c(1))^2 = \frac{1}{2}\left(1.1 - \frac{1}{1 + \frac{(1-c)^2}{99}}\right)^2 \ .$$

The minimum of E is reached when $dE/dc = 0$, thus $c = 1$, (see Fig. 2.3). ■

It must be noted that, in the case of inexact data, it is certainly better to use “too much” data. In the example just discussed, if we have access to more data points, even if they contain errors, we may be able to estimate a model more symmetric with relation to the y-axis, like the phenomenon that is being modeled.

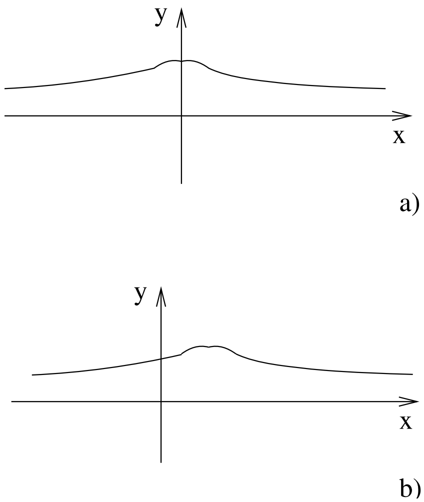


Fig. 2.3 a) The graph of a phenomenon: $g(x) = \frac{1}{1+x^2/99}$. b) Estimated graph: $g_1(x) = \frac{1}{1+(x-1)^2/99}$.

2.7 Well-Posed Problems

Hadamard defined the notion of a *well-posed problem* as being one that:

(i) has a solution (*existence*);

(ii) the solution is unique (*uniqueness*);

(iii) the solution depends "smoothly" on the data (*regularity*).

When any of these properties is not satisfied, we say that the problem is *ill-posed*. As we have already seen, inverse problems do not always satisfy properties (i)–(iii).

Sometimes, property (i) is not satisfied because it is impossible, once the class of models is characterized, to interpolate the data with any model within the class. This has been exemplified in the previous section. We may surpass this by relaxing the notion of solution — an approximation instead of an interpolation, for example in a least squares sense.

If property (ii) is not satisfied, additional restrictions must be found to force the solution to be unique. It is not possible to obtain a unique solution if information is lacking—there are no mathematical tricks to circumvent lack of knowledge. The difficulty here steams from the modelization of the phenomenon.

It is said implicitly that the problem involves data. In this case we can talk about the set of data and properties (i) and (ii) implies that the attribution

$$\text{data} \longrightarrow \text{solution}$$

is a function called *solution operator*, since for any data there is a solution and it is unique. Property (iii) asks for additional features of the solution operator.

Property (iii) is more complex from a technical point of view. The notion of smoothness has to be specified for each problem and, sometimes, can be translated as continuity or differentiability. It is common for inverse problems in infinite dimension spaces to be discontinuous. These problems must be rendered discrete to be solved by means of a computer. It is almost certain that a discrete problem, coming from the discretization of a differential equation model, is continuous. Even in this case it may be difficult to obtain the solution, since it can, still, be *ill conditioned*.

Thus, in practice, if "well-posed" is to mean a reasonable behaviour in the numerical solution to problems, property (iii) may be substituted by *well-conditioned* when we deal with finite dimension spaces. The goal of regularization techniques is to move around the difficulties associated with the lack of smooth dependence between the input data and the solution. In some texts this is called a stability problem.

2.8 Classification of Inverse Problems

We give a brief overview of general classes of mathematical models and a discussion of several ways to classify inverse problems.

2.8.1 Classes of Mathematical Models

When investigating a phenomenon or a physical system, one of the first things to do is to characterize a mathematical model, i.e., to select a class of models. This question, which is of the utmost importance, was considered very superficially in Chapter 1.

For several purposes, the characterization of the system leads to choosing a set of functions or equations (algebraic, ordinary and/or partial differential equations, integral equations, integro-differential, algebraic-differential equations, etc), containing certain unknown constant parameters and/or functions. Of course, other classes of models can be considered, expressing, nonetheless, basic relations satisfied by the system or phenomenon under investigation, but we shall only consider the previous ones.

Taking a somewhat more general point of view, we remark that models can be either discrete or continuous, either deterministic or stochastic, and either given by a function (kinematic) or by an equation (dynamic). Each pair of these concepts, although not exaustive, are exclusive.

Even though each kind of model set forth previously distinguishes itself by its own technical tools of the trade, we just focus in their conceptual differences, as far as modeling is concerned. We may also split most of the problems between linear and nonlinear types. So, when characterizing a model, one possible first thing to do is deciding that it will be, say, a linear, discrete, stochastic, dynamic model. See Table 2.2.

From a standpoint of knowledge level, the 'dynamic' or equation model is more fundamental than the 'kinematic' or function model. The distinction between

Table 2.2 Mathematical models

categories	values	
information content	full: deterministic	lacking: stochastic
nature of variables	discrete	continuous
nature of modelization	descriptive (kinematic)	explanatory (dynamic)
mathematical structure	function	equation
mathematical property	linear	non-linear

models, that are given by a function or by an equation, is best understood through examples. One such example is given by *summing integers*.

Example 2.8. Sum of integers. Consider the 'phenomenon' resulting from progressively adding integers,

$$1,\ 1+2=3,\ 1+2+3=6,\ 10,\ 15,\ 21,\ldots$$

This phenomenon can be modeled by the function $F : \mathbb{N} \to \mathbb{N}$ such that

$$F(n) = F_n = \frac{n(n+1)}{2}, \quad \text{for all } n \in \mathbb{N}\,.$$

This is a 'kinematic' or descriptive model. The corresponding *recurrence relation*

$$F_{n+1} = F_n + (n+1), \quad \text{for } n \geq 1\,,$$

together with initial condition $F_1 = 1$, is a 'dynamic' or explanatory model of the same 'phenomenon', the sum of the first n integers. Both are discrete, deterministic models. The kinematic model is *nonlinear* and the dynamic is (nonhomogeneous) linear. ■

In the same line a simple classical example from mechanics is, perhaps, the best. We shall consider it next.

Example 2.9. Uniform motion. For uniform motion of a pointwise particle in a straight line, the kinematic model (a function) is

$$[0,\infty[\ni t \mapsto x(t) = x_0 + v_0 t \in \mathbb{R}\,,$$

where $x(t)$ represents the position at a given time t, v_0 is the constant velocity, and x_0 is the initial position (position at time $t = 0$). Clearly, this is a continuous, deterministic, linear (better, affine function), kinematic model.

The corresponding dynamic model is given by Newton's second law (differential equation),

$$m\frac{d^2x}{dt^2} = F, \quad \text{for } t > 0, \text{ with } F = 0\,,$$

subjected to the following initial conditions,

$$x(0) = x_0, \quad \text{and} \quad \left.\frac{dx}{dt}\right|_{t=0} = v(0) = v_0\,.$$

Here m represents the mass of the particle, and F the resultant of forces acting on it. This is a continuous, deterministic, linear, dynamic model. ■

We shall not consider here the characterization of models any further. See, however, Afterword, page 177.

2.8.2 Classes of Inverse Problems

A classification scheme of inverse problems arises from the process point of view, represented by the *black box* (see Fig. 1.1 on page 7). There, the black box set-up could represent the interaction between an *external observer* (researcher) and a *physical system*, where the observer could interact with the system by stimulating it and collecting information about its behaviour or reaction[9].

In line with what was said in Chapter 1, we can consider three general classes of problems:

P_1: ***Direct problem*** — Given an arbitrary stimulus, tell what the corresponding reaction will be;

P_2: ***Inverse reconstruction problem*** — Given the reaction, tell what stimulus produced it;

P_3: ***Inverse identification problem*** — From a data set, determine the parameters and/or functions that specify the system.

This book emphasizes the solution of inverse problems, either reconstruction or identification of models, and we shall only consider problems related to classes of models represented by:

- linear or non-linear equations in spaces of finite dimension,

  $$A(\mathbf{x}) = \mathbf{y}\,,$$

  where A is a linear or non-linear function,

  $$\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto A(\mathbf{x}) \in \mathbb{R}^m$$

[9] Of course, it is not always possible to have such clear cut separation between observer and physical systems. However, we shall only consider these.

and $\mathbf{y} \in \mathbb{R}^m$;

- initial and/or boundary value problems for differential or integro-differential equations.

Once a particular class of models is selected, we *solve the inverse problem* by approximately determining the value of the unknown model's constants or functions, using experimental data.

When dealing with differential equation models, we classify the inverse problem with respect to the role, in the differential equation, of the object to be estimated, which can be:

(i) initial conditions;

(ii) boundary conditions;

(iii) forcing terms;

(iv) coefficients of the equation, ie. properties of the system[10].

Here, (i) to (iii) are *reconstruction* problems and (iv) is an *identification* problem.

Moreover, we have a natural splitting of inverse problems to be considered, either the models are of finite dimension (such as a system of n equations and m unknowns) or of infinite dimension (such as an initial value problem for a partial differential equation), and if the object being estimated is of finite dimension (some parameters or constants of the model) or of infinite dimension (such as a function, or an infinite number of parameters). Problems are then classified as belonging to one of the following types:

Type I: Estimation of a finite number of parameters in a model of finite dimension;

Type II: Estimation of an infinite number of parameters or a function in a model of finite dimension;

Type III: Estimation of a finite number of parameters in a model of infinite dimension;

Type IV: Estimation of an infinite number of parameters or a function in a model of infinite dimension.

[10] Just to mention a few, the coefficients could represent material or geometrical properties as the viscosity of a fluid, the thermal conductivity of a solid material, the permeability of a porous medium, and so on. If the coefficient varies throughout the space (it would be a function, not just a constant number) one says that the medium is *heterogeneous*, otherwise, if it is constant everywhere in space, one says that the medium is *homogeneous*. If the coefficients vary with respect to the direction in space, then the medium is said to be *anisotropic*, otherwise it is said *isotropic*.

Table 2.3 Classification of inverse problems according to the dimension of the model and of the object being estimated

Estimation of quantity → Dimension of the model ↓	Finite	Infinite
Finite	Type I	Type II
Infinite	Type III	Type IV

This is summed up in Table 2.3. Inverse problems of Type I are considered in Chapters 1–4; Chapters 5–8 deal mainly with inverse problems of Types III and IV. In this book we do not consider Type II inverse problems. Section 8.1 further elaborates on this classification.

Beck [11, 10] proposed the classification of inverse problems with respect to the type of the unknown of the inverse problem, either parameters or functions. Our classification is just an extension of his, in the sense that we split his classification taking into account the dimension of the model, essentially, either finite when, typically, the model is given by an algebraic function/equation, or infinite when, most of the times, the model is a differential/integral equation. Therefore, Beck's estimation of parameters corresponds to Type I or III problems, and Beck's estimation of functions corresponds to Type II or IV inverse problems. This further splitting is justified by the increased level of mathematical complexity of going from a model of finite dimension to one of infinite dimension.

Exercises

2.1. Let f be a real function of one variable, $f : \mathbb{R} \to \mathbb{R}$. It is worthwhile to see what is the consequence of the fact that the different multiplication factors are constant. Determine the functions such that:

(a) $f'(x) = c$. What can you say about its graph?

(b) $f'(x)/f(x) = c$;

(c) $xf'(x) = c$;

(d) $xf'(x)/f(x) = c$.

Verify that the graph of a function satisfying (b), when plotted in a *log-normal* scale, is a straight line. Analogously, verify that the graphs of a function satisfying (c), in a *normal-log* scale, and of a function satisfying (d), in a *log-log* scale, are straight lines.

2.2. For functions $f(x) = \ln x$, $g(x) = e^{\alpha x}$, $h(x) = x^{\beta}$, $l(x) = \frac{1}{x-a}$ discuss the regions of well and ill-conditioning, for the four types of condition numbers defined, in general, in Eq. (2.4), or on page 30, for a scalar function of one real variable.

2.3. (a) Since $\mathcal{I} = AA^{-1}$ use Eq. (A6) to show that

$$k(A) \geq 1\,,$$

that is, the condition of any matrix is greater than or equal to 1.

(b) Show that

$$k(A) \quad = \quad k(A^{-1})\,,$$

that is, the condition of a matrix and of its inverse are equal.

2.4. Let $p_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, N$, be differentiable functions that do not vanish at any point, i.e.,

$$p_j(x) \quad \neq \quad 0 \text{ for all } x \in \mathbb{R} \; j = 1, \ldots, N\,.$$

(a) Show that the *elasticity of the composition is the product of the elasticities*, that is,

$$E(p_1 \circ p_2) = Ep_1 \cdot Ep_2\,.$$

(b) Show, by induction, that

$$E(p_1 \circ p_2 \circ \ldots \circ p_N) = Ep_1 \cdot Ep_2 \cdot \ \ldots Ep_N\,.$$

(c) Show that the *elasticity of the product is the sum of the elasticities*, that is,

$$E(p_1 \cdot p_2) = Ep_1 + Ep_2\,.$$

(d) Show, therefore, that

$$E(p_1 \cdot p_2 \cdot \ \ldots \ \cdot p_N) = Ep_1 + Ep_2 + \ldots + Ep_N\,.$$

2.5. The *difference* function between non-negative numbers is given by

$$[0, +\infty[\times [0, +\infty[\ni (x,y) \mapsto m(x,y) \;=\; x - y \in \mathbb{R}\,.$$

(a) Show that

$$|\; Em(x,y) \;| \geq 1\,.$$

(b) Determine and sketch the region in $[0, +\infty[\times [0, +\infty[$ satisfying

$$|\; Em(x,y) \;| \leq 2\,.$$

Hint. Example A.1, on page 192, can be useful here.

(c) Do the same for

$$|\; Em(x,y) \;| \geq 4\,.$$

2.6. (a) Check the assertion on the last paragraph of Section 2.4 on page 33.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ and define the elasticity of f,

$$Ef(\mathbf{x}) = \left(\frac{x_1}{f}\frac{\partial f}{\partial x_1}, \frac{x_2}{f}\frac{\partial f}{\partial x_2}, \dots, \frac{x_n}{f}\frac{\partial f}{\partial x_n}\right) .$$

Let $c_f(\mathbf{x})$ denote the relative to relative condition number of f. Show that $|Ef(\mathbf{x})| \le c_f(\mathbf{x})$.

(c) Conclude that if $|Ef(\mathbf{x})| \le 1$ then f is well-conditioned with respect to the relative to relative condition number.

2.7. Let m be the difference function as in Exercise 2.5. Compute the condition number of m, for each notion of condition number set forth in Eq. (2.4).

2.8. We should have not used the elasticity in Example 2.5. Compute the relative to relative condition number of function p_5 in that example.

2.9. Let $c_f(\mathbf{x})$ denote either one of the condition numbers of $\mathbf{f}$ as defined by Eq. (2.4). Let $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$. Show that

$$c_h(\mathbf{x}) \le c_f(\mathbf{g}(\mathbf{x}))\, c_g(\mathbf{x}) .$$

Hint. Recall chain's rule, $\mathcal{J}\mathbf{h}_x = \mathcal{J}\mathbf{f}_{g(x)}\mathcal{J}\mathbf{g}_x$.

2.10. Write down the algorithms discussed in Example 2.5 as composition of functions.

2.11. Relate normal equation (2.10) and its solution, Eq. (2.11), with the results of Exercise 1.7. (Pay attention: the roles of constants a and b are interchanged.)

2.12. QR method for the solution of least squares. From Eq. (2.11) and recalling that $A = (\mathbf{x}, \mathbb{1})$ the solution of the least squares problem,

$$(\hat{a}, \hat{b})^T = \operatorname{argmin}_{(a,b)} E(a,b) = \operatorname{argmin}_{(a,b)} \frac{1}{2} |\mathbf{y} - a\mathbf{x} - \mathbf{b}|^2 ,$$

is given by

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left(A^T A\right)^{-1} A^T \mathbf{y} = \begin{pmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbb{1} \\ \mathbb{1}^T\mathbf{x} & \mathbb{1}^T\mathbb{1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}^T \\ \mathbb{1} \end{pmatrix} \mathbf{y} . \tag{2.13a}$$

Consider the function

$$\mathbb{R}^n \times \mathbb{R}^n \ni (\mathbf{x}, \mathbf{y}) \overset{G}{\mapsto} \begin{pmatrix} \hat{a}(\mathbf{x}, \mathbf{y}) \\ \hat{b}(\mathbf{x}, \mathbf{y}) \end{pmatrix} . \tag{2.13b}$$

The condition of the algorithm to compute $(\hat{a},\hat{b})^T$ suggested by Eq. (2.13) is very high since it depends on the computation of the triple product $\left(A^TA\right)^{-1}A^T$. This undermines the stability of the algorithm.

This exercise proposes an alternative algorithm based on the method QR, [35]. Consider the vector

$$\mathbf{v} = \hat{a}\mathbf{x} + \hat{b}\mathbb{1} = (\mathbf{x}\ \mathbb{1})\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = A\left(A^TA\right)^{-1}A^T\mathbf{y}$$

which represents the vector, in the plane generated by $\mathbf{x}$ and $\mathbb{1}$, closer to $\mathbf{y}$, representing its orthogonal projection[11].

(a) Consider the function

$$\mathbf{x} \mapsto \Lambda(\mathbf{x}) = A\left(A^TA\right)^{-1}A^T$$

which represents the projection matrix from $\mathbb{R}^n$ to the space generated by the vectors $\mathbf{x}$ and $\mathbb{1}$. Show that $\Lambda(\mathbf{x} - \lambda\mathbb{1}) = \Lambda(\mathbf{x})$, for all $\lambda \in \mathbb{R}$. (From a geometrical point of view this result is expected since the space generated by $\mathbf{x}$ and $\mathbb{1}$ is the same as the space generated by $\mathbf{x} - \lambda\mathbb{1}$ and $\mathbb{1}$. Of course, we are assuming that $\mathbf{x}$ and $\mathbb{1}$ are independent, i.e., that $\mathbf{x}$ is not a multiple of $\mathbb{1}$.)

(b) In particular, choose $\lambda_\star$ such that $\mathbf{x} - \lambda_\star\mathbb{1} \perp \mathbb{1}$. Check how the quadruple product present in $\Lambda(\mathbf{x} - \lambda_\star\mathbb{1})$ simplifies.
Hint. A^TA is diagonal, therefore easily invertible.

(c) In this case, obtain a simpler expression for $\mathbf{v}$.

(d) Determine α, β such that

$$\alpha(\mathbf{x} - \lambda_\star\mathbb{1}) + \beta\mathbb{1} = \mathbf{v}\,.$$

Hint. Use item (b) and Fourier-Pythagoras trick, page 209.

(e) From the result in item (d), determine an expression for $(\hat{a},\hat{b})$, where

$$\mathbf{v} = \hat{a}\mathbf{x} + \hat{b}\mathbb{1}\,,$$

obtaining a more stable algorithm for least squares.

2.13. Using the concepts introduced in Sections 2.1 and 2.5 study the evaluation and algorithms to compute

$$f(x) = \sin x - x\,.$$

Hint. You may consider using a truncated Taylor's series of $\sin x$.

[11] Further discussion on orthogonal projections can be seen in Section A.4.1.

2.14. Heat conduction problem. Consider a one-dimensional bar, isolated in its lateral surface, being heated in one extremity, and in contact with the ambient at the temperature T_{amb} in the other extremity. Assume also that heat is being transferred to its interior. Let $T = T(x,t)$ denote its temperature on position x, at time t . Then T satisfies the following initial and boundary value problem for a partial differential equation,

$$\rho c_p \frac{\partial T}{\partial t}(x,t) = \frac{\partial}{\partial x}\left(k(T,x)\frac{\partial T}{\partial x}\right) + g(x,t)\,,$$

for $x \in [0,L]$, $t > 0$,

$$-k\frac{\partial T}{\partial x}\bigg|_{x=0} = q''(t)\,, \quad \text{for } t > 0\,,$$

(left boundary: prescribed heat flux),

$$-k\frac{\partial T}{\partial x}\bigg|_{x=L} = h\,[T(L,t) - T_{amb}]\ , \quad \text{for } t > 0\,,$$

(right boundary: contact with ambient),

$$T(x,0) = T_0(x), \quad \text{for } t > 0\,,$$

(initial condition). Here, $\rho = \rho(\mathbf{x})$ is the specific mass of the material, $c_p = c_p(x)$ is the *specific heat* of the material, $k = k(T,x)$ is the *thermal conductivity*, $g = g(x,t)$ is an internal heat source/sink, q'' is the interfacial heat flux, h is the convection coefficient, and $T_0 = T_0(\mathbf{x})$ is the initial temperature of the bar.

Classify the following problems with respect to being direct, inverse identification, or inverse reconstruction problems:

(a) Given k, g, ρ, c_p, q'', h, and T_0, determine $T = T(x,t)$;

(b) Given k, g, ρ, c_p, h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine q'';

(c) Given k, ρ, c_p, q'', h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine g;

(d) Given k, ρ, g, c_p, q'', h, and measurements of temperature at certain times, on some locations of the bar, determine T_0;

(e) Given ρ, g, c_p, q'', h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine k;

(f) Given ρ, g, c_p, q'', T_0, and measurements of temperature at certain times, on some locations of the bar, determine k and h;

(g) Given ρ, g, c_p, h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine k and q''.

Chapter 3
Spectral Analysis of an Inverse Problem

In this chapter we treat the solution of systems of linear equations in finite dimension spaces. This can be seen as examples of inverse reconstruction problems[1] of Type I. We present a mathematical analysis of these linear inverse problems of finite dimension based on the spectral theorem. We study several aspects and behaviour of well-established methods for solving inverse problems. In particular, the concept of regularization, a very important notion in the area of inverse problems, will be dealt with. The analysis presented here is elementary — it depends on notions of linear algebra and convergence of numerical series. A similar study of regularization for problems in infinite dimensional spaces depends on functional analysis and is beyond the scope of this book. The interested reader is encouraged to consult [44, 29].

As seen in Chapter 2, when different algorithms are applied to the same problem, different properties and behaviours are to be expected. In particular, some are stable, while others are unstable. On the other hand, different formulations for the same problem (even when mathematically equivalent) lead to different algorithms. This has already been seen in example 2.5, page 34. We will develop this idea, presenting some algorithms, Tikhonov in Section 3.4, steepest descent in Section 3.7, Landweber in Section 3.8, and conjugate gradient in Section 3.10, to solve the same problem.

The mathematical analysis of some of the methods presented will be performed, in order to gain intuition in the behaviour of the algorithms and be able to chose and propose the best suited method to solve a specific inverse problem. In this analysis (Sections 3.4, 3.8, and 3.11), the spectral theorem and the singular value decomposition will be used. A derivation of the conjugate gradient method is performed in Section 3.10.

In Sections 3.3, 3.5, 3.6, and 3.9, we consider, respectively, regularization schemes, strategies and the discrepancy principle. We begin the discussion, in Section 3.1 with a numerical example to motivate the reader, and we analyse it, in general, in Section 3.2.

3.1 An Example

Given the matrix

$$K = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{1024} \end{pmatrix}, \tag{3.1}$$

[1] For classifications of inverse problems, see Section 2.8.2.

and the vector $\mathbf{y} = (1, 2^{-10})^T$, it is plain to see that $\mathbf{x} = (1,1)^T$ is the solution of the system

$$K\mathbf{x} = \mathbf{y} \,. \tag{3.2}$$

The solution to problems of the kind given by Eq. (3.2), where matrix K is known, $\mathbf{y}$ is fixed and where $\mathbf{x}$ is to be found, can be somewhat complex, since small variations in $\mathbf{y}$ may lead to large variations in $\mathbf{x}$. For instance, let us perturb $\mathbf{y}$ by $\mathbf{p} = (0, 2^{-10})^T$. We obtain a solution that differs from $\mathbf{x}$ by $\mathbf{r} = (0, 1)^T$, i.e.,[2]

$$\begin{aligned} K\left(\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right) &= K(\mathbf{x}+\mathbf{r}) = \mathbf{y}+\mathbf{p} \\ &= \begin{pmatrix}1\\2^{-10}\end{pmatrix} + \begin{pmatrix}0\\2^{-10}\end{pmatrix} \,. \end{aligned}$$

Thus, the input error, $\mathbf{p}$, is multiplied by a factor of $|\mathbf{r}|/|\mathbf{p}| = 1024$, and the problem is ill-conditioned.

This is, however, a radical behaviour that is not common to all perturbations, even if they are of the same magnitude. As a matter of fact, if $\mathbf{p} = (2^{-10}, 0)^T$, then $\mathbf{r} = (2^{-10}, 0)^T$, since

$$K\left(\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}2^{-10}\\0\end{pmatrix}\right) = \begin{pmatrix}1\\2^{-10}\end{pmatrix} + \begin{pmatrix}2^{-10}\\0\end{pmatrix} \,,$$

and thus the multiplication factor of the error is much smaller, $|\mathbf{r}|/|\mathbf{p}| = 1$.

We then realize that the computation of the evaluation of the inverse of K at points contaminated by small errors is more sensitive to certain perturbations than to others. Some of these perturbations result in excessive amplification.

This kind of situation can be avoided by subtly modifying the problem. Instead of solving Eq. (3.2), a slightly altered (perturbed) problem is solved,

$$K_\alpha \mathbf{x} = \mathbf{y}_\alpha \,, \quad \alpha > 0 \,, \tag{3.3}$$

in such a way that Eq. (3.3) behaves, as much as possible, as Eq. (3.2), but being better conditioned with relation to certain perturbations of $\mathbf{y}$. We say that the problem presented in Eq. (3.3) is a regularization of the original problem, Eq. (3.2). These notions will be defined precisely later.

Choose

$$K_\alpha = \begin{pmatrix}1 & 0\\ 0 & \frac{1}{(1-\alpha)^{10}}\frac{1}{2^{10}}\end{pmatrix} \,, \text{ and } \mathbf{y}_\alpha = \mathbf{y} \,, \quad 0 < \alpha < 1 \,, \tag{3.4}$$

and note that

$$K_\alpha \to K, \quad \text{as} \quad \alpha \to 0 \,.$$

Let $\mathbf{x}^\alpha$ be the solution of Eq. (3.3) for the choices of K_α and $\mathbf{y}_\alpha$ in Eq. (3.4). We let α assume a few values, $\alpha = 1/2^n$, $n = 1, \ldots, 9$, and determine the error and the

Table 3.1 Numerical results for the error multiplication factor

α	1/2	1/4	1/8	1/16	1/32	1/64	1/128	1/256	1/512
$\vert\mathbf{r}\vert$	0.998	0.887	0.474	0.049	0.456	0.709	0.849	0.923	0.961
$\vert\mathbf{r}\vert/\vert\mathbf{p}\vert$	1022	909	485	50	467	726	869	945	984

corresponding multiplication factor of the error, keeping the perturbation equal to $\mathbf{p} = (0{,}2^{-10})^T$. The results are presented in Table 3.1.

In this example, it is possible to obtain better results when the regularization parameter $\alpha = 1/16$. The error multiplication factor plummets from 1024, when $\alpha = 0$, to 50, when the perturbation in the data is $\mathbf{p} = (0, 2^{-10})^T$.

3.2 Uncertainty Multiplication Factor

Consider the linear problem

$$K\mathbf{x} = \mathbf{y}\,, \tag{3.5}$$

where K and $\mathbf{y}$ are given data, and $\mathbf{x}$ is the unknown. We assume that K is invertible. Also, consider a family of perturbations of $\mathbf{y}$ parameterized by ϵ, $\mathbf{y}^\epsilon$, such that $|\mathbf{y} - \mathbf{y}^\epsilon| \leq \epsilon$ and define $\mathbf{x}^\epsilon$ as the solution of the perturbed equation,

$$K\mathbf{x}^\epsilon = \mathbf{y}^\epsilon\,. \tag{3.6}$$

By subtracting the perturbed equation, Eq. (3.6), from the unperturbed equation, Eq. (3.5), we get

$$K(\mathbf{x}^\epsilon - \mathbf{x}) = \mathbf{y}^\epsilon - \mathbf{y} \quad \text{or} \quad \mathbf{x}^\epsilon - \mathbf{x} = K^{-1}(\mathbf{y}^\epsilon - \mathbf{y})\,.$$

Now,

$$|\mathbf{x}^\epsilon - \mathbf{x}| = |K^{-1}(\mathbf{y}^\epsilon - \mathbf{y})|\,,$$

and using Eq. (A6), we obtain,

$$|\mathbf{x}^\epsilon - \mathbf{x}| \leq |K^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}|\,. \tag{3.7}$$

That is, the error $\mathbf{y}^\epsilon - \mathbf{y}$ due to only knowing the approximation $\mathbf{y}^\epsilon$ of $\mathbf{y}$ —this one unknown— is, at most, amplified by $|K^{-1}|$, as can be deduced from Eq. (3.7). See example 2.3, page 32 where this has already been discussed.

The maximum amplification is reached for certain perturbations. As stated by Eq. (A5),

$$|K^{-1}| = \max_{|\mathbf{z}|=\epsilon} \frac{|K^{-1}\mathbf{z}|}{|\mathbf{z}|}\,.$$

[2] $K(\mathbf{x}+\mathbf{r}) = \mathbf{y}+\mathbf{p}$ that when subtracted from $K\mathbf{x}=\mathbf{y}$, yields $K\mathbf{r}=\mathbf{p}$.

Therefore, there exists at least one $\mathbf{z}_o$ with $|\mathbf{z}_o| = \epsilon$, such that

$$|K^{-1}| = |K^{-1}\mathbf{z}_o|/|\mathbf{z}_o|\,.$$

If we choose $\mathbf{y}^\epsilon$ in such a way that $\mathbf{y}^\epsilon - \mathbf{y} = \mathbf{z}_o$, we will have

$$|K^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}| = |K^{-1}(\mathbf{y}^\epsilon - \mathbf{y})|\,.$$

However,

$$K^{-1}\mathbf{y}^\epsilon = \mathbf{x}^\epsilon \quad \text{and} \quad K^{-1}\mathbf{y} = \mathbf{x},$$

due to the definitions of $\mathbf{x}$ and $\mathbf{x}^\epsilon$ (see Eqs. (3.5) and (3.6)). Therefore,

$$|K^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}| = |\mathbf{x}^\epsilon - \mathbf{x}|\,. \tag{3.8}$$

Summing up, we have, for every $\mathbf{y}$, at least a perturbation $\mathbf{y}^\epsilon = \mathbf{y} + \mathbf{z}_o$ (i.e., a point at a distance ϵ from $\mathbf{y}$), such that the error will be amplified by $|K^{-1}|$. Here, the solution is

$$\mathbf{x}^\epsilon = K^{-1}\mathbf{y}^\epsilon = K^{-1}(\mathbf{y} + \mathbf{z}_o) = \mathbf{x} + K^{-1}\mathbf{z}_o\,,$$

and the error's magnitude is

$$|K^{-1}\mathbf{z}_o| = |K^{-1}|\,\epsilon\,.$$

3.3 Regularization Scheme

As we saw, *uncertainties* in $\mathbf{y}$, in the right hand side of equation $K\mathbf{x} = \mathbf{y}$ are, at most, amplified by $|K^{-1}|$, see Eq. (3.7). This multiplication factor can be quite large. It only depends on the smallest *singular value* of matrix K, since

$$|K^{-1}| = \sigma_1(K^{-1}) = 1/\sigma_n(K)\,,$$

where, in general, $\sigma_1(A)$ represents the largest, and $\sigma_n(A)$, the smallest singular values of an $n \times n$ matrix A, (see page 195 of Section A.3).

Just to reinforce, if K has a singular value near zero, then $|K^{-1}|$ will be large. If the uncertainty in the data is mainly in the direction associated with the smallest singular values of K, the uncertainty in the solution of the equation can be greatly amplified. By regularizing the problem, as exemplified in Section 3.1, this situation can be avoided. We shall consider regularization procedures more systematically.

Let

$$K_\alpha : \mathbb{R}^n \to \mathbb{R}^n, \ \text{and} \ \mathbf{b}_\alpha \in \mathbb{R}^n, \quad \alpha > 0\,,$$

be, respectively, a family of invertible matrices (linear operators), and a family of vectors, parameterized by $\alpha > 0$. Also, assume that $\mathbf{y} \in \mathbb{R}^n$. Consider the problem given by

$$K_\alpha \mathbf{x} = \mathbf{y} + \mathbf{b}_\alpha, \quad \alpha > 0 \tag{3.9}$$

and let $\mathbf{x} = \mathbf{x}^\alpha$ be its solution, where the superscript α indicates that the solution depends on α.

Definition 3.1. The families K_α and $\mathbf{b}_\alpha$, $\alpha > 0$, constitute a *linear regularization scheme* for the linear problem $K\mathbf{x} = \mathbf{y}$, if the following conditions hold[3]:

$$K_\alpha \to K\,, \quad |\mathbf{b}_\alpha| \searrow 0 \;\text{ and }\; |K_\alpha^{-1}| \nearrow |K^{-1}|\,, \text{ as } \alpha \searrow 0\,. \tag{3.10}$$

Here, α is called *regularization parameter*. The perturbed problem, Eq. (3.9), is called the *regularized problem*, K_α is the *regularized matrix*, and $\mathbf{x}^\alpha$ is the *regularized solution*.

We remark that the solution operator of Eq. (3.9) is represented by the function

$$\mathbb{R}^n \ni \mathbf{y} \mapsto \mathbf{x}^\alpha = K_\alpha^{-1}(\mathbf{y} + \mathbf{b}_\alpha) \in \mathbb{R}^n\,. \tag{3.11}$$

Theorem 1. (a) If the pair K_α and $\mathbf{b}_\alpha$ constitute a linear regularization scheme, we have

$$\mathbf{x}^\alpha \to \mathbf{x}\,, \text{ as } \alpha \to 0\,. \tag{3.12}$$

(b) The evaluation's condition of the regularized problem solution, Eq. (3.11), is less than the evaluation's condition of the unperturbed problem

$$\mathbf{y} \mapsto \mathbf{x} = K^{-1}\mathbf{y}\,, \tag{3.13}$$

solution of the original problem, Eq. (3.5).

Proof

(a) By definition of $\mathbf{x}$ and $\mathbf{x}^\alpha$, Eqs. (3.11) and (3.13),

$$\mathbf{x}^\alpha - \mathbf{x} = K_\alpha^{-1}(\mathbf{y} + \mathbf{b}_\alpha) - K^{-1}\mathbf{y} = (K_\alpha^{-1} - K^{-1})\mathbf{y} + K_\alpha^{-1}\mathbf{b}_\alpha\,.$$

From triangle's inequality, Eq. (A27), Eq. (A6), and definition requirement on K_α, Eq. (3.10), we get,

$$\begin{aligned} |\mathbf{x}^\alpha - \mathbf{x}| &= |(K_\alpha^{-1} - K^{-1})\mathbf{y} - K_\alpha^{-1}\mathbf{b}_\alpha| \\ &\le |K_\alpha^{-1} - K^{-1}|\,|\mathbf{y}| + |K_\alpha^{-1}|\,|\mathbf{b}_\alpha| \\ &\le |K_\alpha^{-1} - K^{-1}|\,|\mathbf{y}| + |K^{-1}|\,|\mathbf{b}_\alpha|\,. \end{aligned} \tag{3.14}$$

By definition,

$$\lim_{\alpha\to 0} K_\alpha = K \text{ if and only if } \lim_{\alpha\to 0} |K_\alpha - K| = 0\,,$$

and, since the inversion of matrices is a continuous function,

$$\lim_{\alpha\to 0} K_\alpha^{-1} = K^{-1}\,, \text{ or } \lim_{\alpha\to 0} |K_\alpha^{-1} - K^{-1}| = 0\,.$$

Finally, using Eq. (3.14), Eq. (3.10) and previous result, we prove that $\mathbf{x}^\alpha \to \mathbf{x}$, as $\alpha \to 0$. ■

[3] Here, for instance, $a_\alpha \searrow 0$ as $\alpha \searrow 0$ means that a_α is an increasing function of $\alpha > 0$, and that $\lim_{\alpha\to 0} a_\alpha = 0$, is taken only from the right of zero.

(b) The condition of the evaluation of the solution of the regularized problem, Eq. (3.11), is less than the condition of the evaluation of $\mathbf{x} = K^{-1}\mathbf{y}$, solution of the original problem, Eq. (3.5), since $|K_\alpha^{-1}| < |K^{-1}|$, for $\alpha > 0$ as is required in the definition of a linear regularization scheme, Eq. (3.10). This is illustrated in Fig. 3.1. ■

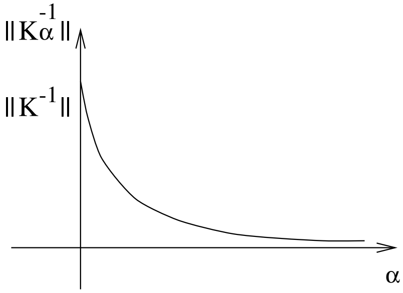


Fig. 3.1 Graph of the norm of K_α^{-1} as a function of α, illustrating some requirements for K_α to be a linear regularization

Let us say one more time that the requirement on the condition lays with the inverse, because it is this matrix that solves (theoretically) the inverse problem, and transforms it into a function evaluation problem. To restrain the growth of error in the data, matrices with smaller condition numbers are preferred, which justifies the last condition in Eq. (3.10).

3.4 Tikhonov's Regularization

In this section, we present and analyze a classical regularization scheme, the *Tikhonov's regularization scheme*, [88, 1, 2]. We apply this regularization to the problem given by Eq. (3.5). The analysis depends on the classical spectral theory of linear operators in finite dimension vector spaces (Sections A.2 and A.3).

We are going to see that Eq. (3.5) can be substituted by the equivalent problem of minimizing the functional

$$f(\mathbf{x}) = \frac{1}{2}|K\mathbf{x} - \mathbf{y}|^2 \,, \tag{3.15}$$

where we recall here that K and $\mathbf{y}$ are given.

Theorem 2. Let K be an invertible matrix. Then,

(i) $\mathbf{x}_\star$ is the minimum point of f if and only if $\mathbf{x}_\star$ is the solution of Eq. (3.5);

(ii) The critical point equation of f is $K^T K\mathbf{x} = K^T\mathbf{y}$;

(iii) The critical point equation of f is equivalent to Eq. (3.5).

Proof. (i) $\mathbf{x}_\star$ satisfies $K\mathbf{x}_\star = \mathbf{y}$ if and only if $f(\mathbf{x}_\star) = 0$. Now, $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. Therefore, $\mathbf{x}_\star$ is the only minimum point of f (since K is invertible).
(ii) Let us compute the critical point equation of f, $\nabla f(\mathbf{x}) = 0$. First, the *directional derivative* of f at $\mathbf{x}$, in the direction of $\mathbf{h}$, is given by

$$\begin{aligned}
df_x(\mathbf{h}) &= \lim_{\epsilon\to 0} \frac{f(\mathbf{x}+\epsilon\mathbf{h}) - f(\mathbf{x})}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{1}{2\epsilon}\left(|K(\mathbf{x}+\epsilon\mathbf{h}) - \mathbf{y}|^2 - |K\mathbf{x}-\mathbf{y}|^2\right) \\
&= \lim_{\epsilon\to 0} \frac{1}{2\epsilon}\left(\langle K(\mathbf{x}+\epsilon\mathbf{h}) - \mathbf{y}, K(\mathbf{x}+\epsilon\mathbf{h}) - \mathbf{y}\rangle\right. \\
&\quad \left. - \langle K\mathbf{x}-\mathbf{y}, K\mathbf{x}-\mathbf{y}\rangle\right) \\
&= \lim_{\epsilon\to 0} \frac{1}{2}\left(\langle K\mathbf{x}-\mathbf{y}, K\mathbf{h}\rangle + \langle K\mathbf{h}, K\mathbf{x}-\mathbf{y}\rangle + \epsilon\langle K\mathbf{h}, K\mathbf{h}\rangle\right) \\
&= \langle K\mathbf{x}-\mathbf{y}, K\mathbf{h}\rangle \\
&= \langle K^T(K\mathbf{x}-\mathbf{y}), \mathbf{h}\rangle\,, \qquad (3.16)
\end{aligned}$$

To get this result, we used the bilinearity of the inner product, its symmetry, and the way a matrix changes places in the inner product.

From the definition of *gradient* of f, ∇f — the vector that represents the derivative through the inner product — we obtain

$$df_x(\mathbf{h}) = \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle\,,$$

for all $\mathbf{h} \in \mathbb{R}^n$. We reach the conclusion that[4]

$$\nabla f(\mathbf{x}) = K^T(K\mathbf{x}-\mathbf{y})\,, \qquad (3.17)$$

which leads to the *critical point* equation

$$K^T K\mathbf{x} = K^T\mathbf{y}\,. \qquad (3.18)$$

(iii) We recall that a matrix K is invertible if and only if its transpose, K^T, is also invertible, and

$$(K^T)^{-1} \quad = \quad (K^{-1})^T\,. \qquad (3.19)$$

Then, by multiplying both sides of Eq. (3.18) on the left by $(K^T)^{-1}$, we see that $\mathbf{x}$ satisfies the critical point equation if and only if $\mathbf{x}$ satisfies Eq. (3.5). ■

Equation (3.18) is known as *normal equation.*

To avoid the amplification of the error in the solution of Eq. (3.5) it is common to penalize the distance from the solution to a reference value, or the norm of the

[4] This is an alternative deduction, more calculus and geometry oriented and using the algebraic structure of the function, than both the one presented in Section 1.6 —that employs index notation—, and the one presented in Section 2.6 —that employs orthogonal projections—, to obtain the critical point equation in similar situations.

solution vector (the distance with respect to the origin). By a reference value we mean a known approximate solution to the problem. We will denote a *reference value* by $\mathbf{x}_r$.

In the case of Tikhonov's method, this idea is implemented by modifying the regularization term, making it penalize the growth of the distance from the reference value. For the problem defined in Eq. (3.5), this consists in solving the critical point equation of the functional

$$f_\alpha(\mathbf{x}) = \frac{1}{2}|K\mathbf{x} - \mathbf{y}|^2 + \frac{\alpha}{2}|\mathbf{x} - \mathbf{x}_r|^2 \, , \tag{3.20}$$

with $\alpha > 0$, the so-called regularization parameter. The minimum point $\mathbf{x}^\alpha$ satisfies the critical point equation

$$\alpha(\mathbf{x}^\alpha - \mathbf{x}_r) + K^T K\mathbf{x}^\alpha = K^T\mathbf{y} \, , \tag{3.21}$$

obtained in a similar way to Eq. (3.18). We rewrite it as

$$\left(\alpha\mathcal{I} + K^T K\right)\mathbf{x}^\alpha = K^T\mathbf{y} + \alpha\mathbf{x}_r \, . \tag{3.22}$$

We will verify that Eq. (3.22) provides a *regularization scheme* for Eq. (3.18). Observe that, strictly speaking, the problem that is being regularized is Eq. (3.18) and not Eq. (3.5).

Let

$$A_\alpha = \alpha\mathcal{I} + K^T K \text{ and } \mathbf{b}_\alpha = \alpha\mathbf{x}_r \, , \quad \alpha > 0 \, . \tag{3.23}$$

Theorem 3. The families A_α, $\mathbf{b}_\alpha$, $\alpha > 0$, are a linear regularization scheme for equation $K^T K\mathbf{x} = K^T\mathbf{y}$.

Proof. Since $A_\alpha \to K^T K$, as $\alpha \to 0$ and $|\mathbf{b}_\alpha| \searrow 0$, the first two conditions in Eq. (3.10) are satisfied.

Now, following the notation in Appendix A, if

$$\lambda_1(K^T K) \geq \lambda_2(K^T K) \geq \ldots \geq \lambda_n(K^T K) > 0 \, ,$$

are the *eigenvalues* of $K^T K$ (the last inequality is strict, since $K^T K$ is invertible), then the eigenvalues of $(K^T K)^{-1}$ satisfy

$$\lambda_i((K^T K)^{-1}) = 1/\lambda_{n+1-i}(K^T K) \, ,$$

for $i = 1, \ldots, n$. It results that

$$\frac{1}{\lambda_n(K^T K)} \geq \frac{1}{\lambda_{n-1}(K^T K)} \geq \ldots \geq \frac{1}{\lambda_1(K^T K)} > 0 \, ,$$

Thus, due to remark A.1 (c), p. 199,

$$|(K^T K)^{-1}| = 1/\lambda_n(K^T K) \, .$$

Also, from Exercise A.8, p. 210,

$$\lambda_i(\alpha\mathcal{I} + K^T K) = \alpha + \lambda_i(K^T K)\,.$$

Therefore,

$$\begin{aligned} |A_\alpha^{-1}| &= |(\alpha\mathcal{I} + K^T K)^{-1}| \\ &= (\alpha + \lambda_n(K^T K))^{-1}\,. \end{aligned}$$

Hence, we see that the condition of the problem, when $\alpha > 0$, is inferior to that when $\alpha = 0$. Notice that $|A_\alpha^{-1}|$ is a decreasing function of α with limit $|(K^T K)^{-1}|$, as $\alpha \to 0$,

$$\lim_{\alpha\to 0} |A_\alpha^{-1}| \quad = \quad |(K^T K)^{-1}|$$

satisfying the third requirement in Eq. (3.10).

Thus, we reach the conclusion that A_α and $\mathbf{b}_\alpha$ determine a *linear regularization scheme* for the linear problem, Eq. (3.18). ■

When compared to the original problem, Eq. (3.18), we see that this regularization scheme shifts the eigenvalues of the matrix of the problem by α. Since they are all real and non-negative (see example A.2 on page 195), they are far from the origin at least by α, including the one with smallest absolute value, which determines the condition of $(K^T K)^{-1}$.

3.5 Regularization Strategy

The notion of regularization scheme has the goal of constraining, in the solution, the effect of the growth of the error coming from the data, by means of modifying the problem's condition. However, the knowledge of an error estimate or the specified level of uncertainty in the data is not taken into consideration. The notion of regularization strategy is important when this information is available.

Assume that the right-side of Eq. (3.5), $\mathbf{y}$, is known only approximately, with a maximum error of ϵ, i.e., $\mathbf{y}^\epsilon$ is known and not necessarily $\mathbf{y}$, in such a way that

$$|\mathbf{y} - \mathbf{y}^\epsilon| \le \epsilon\,. \tag{3.24}$$

Let the pair K_α and $\mathbf{b}_\alpha$ be a linear regularization scheme for Eq. (3.5), and denote by $\mathbf{x}^{\alpha,\epsilon}$ the approximation to $\mathbf{x}$ obtained by means of K_α when the datum is $\mathbf{y}^\epsilon$, i.e., $\mathbf{x}^{\alpha,\epsilon}$ is the solution of the *regularized equation with noisy or corrupted data*,

$$K_\alpha \mathbf{x}^{\alpha,\epsilon} = \mathbf{y}^\epsilon + \mathbf{b}_\alpha\,. \tag{3.25}$$

Then,

$$\mathbf{x}^{\alpha,\epsilon} = K_\alpha^{-1}(\mathbf{y}^\epsilon + \mathbf{b}_\alpha)\,.$$

Definition 3.2. Let K_α, $\mathbf{b}_\alpha$, for $\alpha > 0$, be a linear regularization scheme, and $\mathbf{y}^\epsilon$ an approximation of $\mathbf{y}$, with precision level ϵ. A *regularization strategy* corresponds to a choice of the value of the regularization parameter, α, to be used, as a function of the precision level of the data, i.e., it is a relation between α and ϵ, $\alpha = \alpha(\epsilon)$, such that $\alpha(\epsilon) \to 0$, as $\epsilon \to 0$.

In the presence of noisy data, this is the concept needed, as can be seen from the next simple theorem.

Theorem 4. Let K_α, $\mathbf{b}_\alpha$, $\alpha > 0$, be a linear regularization scheme for equation $K\mathbf{x} = \mathbf{y}$, and $\mathbf{y}^\epsilon$ be an approximation of $\mathbf{y}$ with precision level ϵ, i.e. satisfying Eq. (3.24). Also, let $\mathbf{x}^{\alpha,\epsilon}$ be the solution of the regularized problem with noisy data, Eq (3.25), and $\alpha = \alpha(\epsilon)$ a regularization strategy. Then

$$\lim_{\epsilon\to 0} \mathbf{x}^{\alpha(\epsilon),\epsilon} = \mathbf{x} \,.$$

Proof. From the solution of the regularized equation with noisy data we get,

$$\begin{aligned} \mathbf{x}^{\alpha,\epsilon} - \mathbf{x} &= K_\alpha^{-1}(\mathbf{y}^\epsilon + \mathbf{b}_\alpha) - K^{-1}\mathbf{y} \\ &= K_\alpha^{-1}(\mathbf{y}^\epsilon - \mathbf{y}) + (K_\alpha^{-1} - K^{-1})\mathbf{y} + K_\alpha^{-1}\mathbf{b}_\alpha \,. \end{aligned}$$

Therefore, using triangle inequality, Eq. (A2), and Eqs. (A6) and (3.10) we get,

$$|\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}| \leq |K_\alpha^{-1}|\epsilon + |K_\alpha^{-1} - K^{-1}|\,|\mathbf{y}| + |K^{-1}|\,|\mathbf{b}_\alpha| \,.$$

Now, letting $\alpha = \alpha(\epsilon)$ be a regularization strategy, and $\epsilon \to 0$ in the previous equation, we get the result. ■

A relevant magnitude to be taken into consideration when evaluating the quality of the regularization scheme with corrupted data is

$$|\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}| \;\;.$$

We now present a qualitative analysis that shows some of the problems likely to arise when we choose $\alpha = \alpha(\epsilon)$. We have,

$$\begin{aligned} |\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}| &\leq |\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}^\alpha| + |\mathbf{x}^\alpha - \mathbf{x}| \\ &\leq |K_\alpha^{-1}(\mathbf{y}^\epsilon + \mathbf{b}_\alpha) - K_\alpha^{-1}(\mathbf{y} + \mathbf{b}_\alpha)| + |K_\alpha^{-1}(K\mathbf{x} + \mathbf{b}_\alpha) - \mathbf{x}| \\ &\leq |K_\alpha^{-1}(\mathbf{y}^\epsilon - \mathbf{y})| + |(K_\alpha^{-1}K - \mathcal{I})\mathbf{x}| + |K_\alpha^{-1}\mathbf{b}_\alpha| \\ &\leq |K_\alpha^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}| + |K_\alpha^{-1}K - \mathcal{I}|\,|\mathbf{x}| + |K_\alpha^{-1}|\,|\mathbf{b}_\alpha| \,. \end{aligned} \tag{3.26}$$

We surmise the qualitative behaviour of the norm of the total error, $|\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}|$, by looking at the behaviour of its bound given by

$$E(\alpha) = E_1(\alpha) + E_2(\alpha) + E_3(\alpha) \,,$$

where

$$\begin{aligned} E_1(\alpha) &= |K_\alpha^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}|\,, \\ E_2(\alpha) &= |K_\alpha^{-1}K - I|\,|\mathbf{x}|\,, \text{ and} \\ E_3(\alpha) &= |K_\alpha^{-1}|\,|\mathbf{b}_\alpha|\,. \end{aligned}$$

From Eq. (3.10), $K_\alpha^{-1}K - I = K_\alpha^{-1}(K - K_\alpha)$, and Eqs. (A5) and (A6), we have,

$$\begin{aligned} E_2(\alpha) &\le |K_\alpha^{-1}|\,|K - K_\alpha|\,|\mathbf{x}| \\ &\le |K^{-1}|\,|K - K_\alpha|\,|\mathbf{x}|\,, \text{ and} \\ E_3(\alpha) &\le |K^{-1}|\,|\mathbf{b}_\alpha|\,. \end{aligned}$$

From this, $E_2(\alpha)$, $E_3(\alpha) \to 0$, as $\alpha \to 0$, and

$$\lim_{\alpha\to 0} E(\alpha) = \lim_{\alpha\to 0} E_1(\alpha) = |K^{-1}|\,|\mathbf{y}^\epsilon - \mathbf{y}|\,. \tag{3.27}$$

In practice, we cannot consider the limit as $\epsilon \to 0$; the value of $\epsilon \neq 0$ to be chosen should be a plausible or known level for the error in the data, due, for example, to the working precision of the measurement device. In general, once $\epsilon > 0$ is fixed, no advantage is gained by choosing α close to zero, since, as shown in Eq. (3.27), $E(\alpha)$ does not go to zero as $\alpha \to 0$. This is illustrated in Fig. 3.2a. The upper bound of the total error, $E(\alpha)$, may reach a minimum for a value of α, say α^*, far from zero. We must warn the reader, however, that strictly speaking this is not a conclusion about the behaviour of $|\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}|$.

Definition 3.3. We say that a regularization strategy, $\alpha = \alpha(\epsilon)$, is *optimal* when

$$|\mathbf{x}^{\alpha(\epsilon),\epsilon} - \mathbf{x}| \quad = \quad \inf_\alpha |\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}| \ .$$

3.6 Reference Value in Tikhonov's Method

Mimicking the derivation of Eq. (3.26), we can deduce that

$$\mathbf{x}^{\alpha,\epsilon} - \mathbf{x} = K_\alpha^{-1}(\mathbf{y}^\epsilon - \mathbf{y}) + (K_\alpha^{-1}K - I)\mathbf{x} + K_\alpha^{-1}\mathbf{b}_\alpha\,. \tag{3.28}$$

In Tikhonov's method, with *reference value* $\mathbf{x}_r$, we have

$$\left(\alpha\mathcal{I} + K^TK\right)\mathbf{x}^{\alpha,\epsilon} = K^T\mathbf{y}^\epsilon + \alpha\mathbf{x}_r\,,$$

and substituting, in Eq. (3.28),

$$\begin{aligned} &K_\alpha \text{ by } A_\alpha = \alpha\mathcal{I} + K^TK\,, \\ &\mathbf{b}_\alpha \text{ by } \alpha\mathbf{x}_r\,, \quad K \text{ by } K^TK\,, \\ &\ \mathbf{y} \text{ by } K^T\mathbf{y}\,, \quad \text{and } \ \mathbf{y}^\epsilon \ \text{ by } \ K^T\mathbf{y}^\epsilon \end{aligned}$$

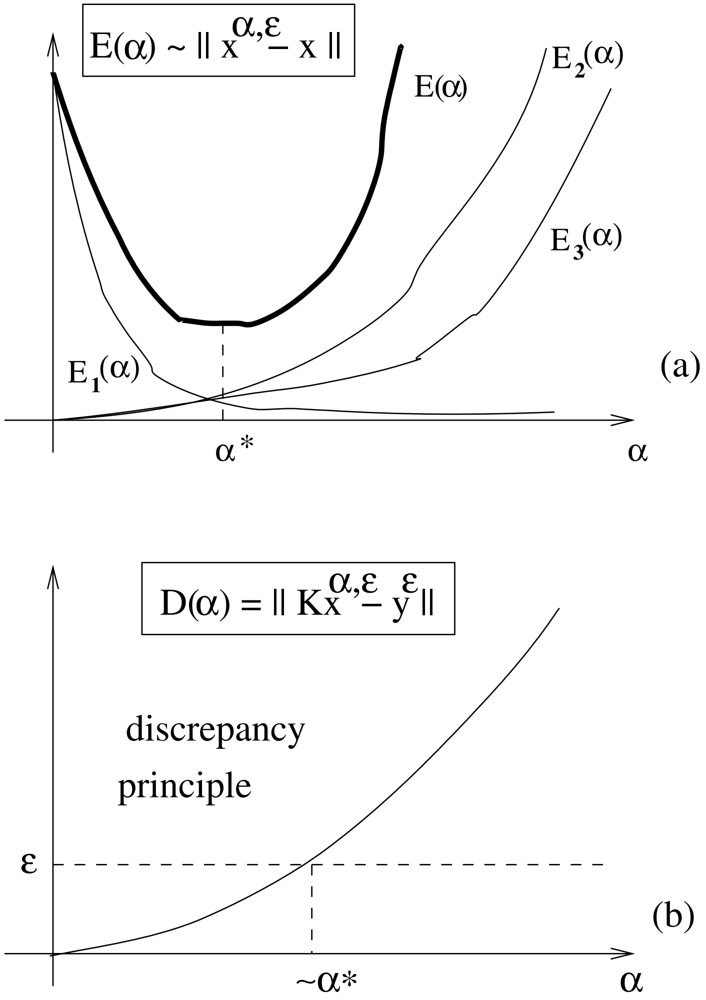


Fig. 3.2 Qualitative behaviour: (a) of the total error bound, $E(\alpha) = E_1(\alpha) + E_2(\alpha) + E_3(\alpha)$; (b) of the discrepancy, $D(\alpha) = \|Kx^{\alpha,\epsilon} - y^{\epsilon}\|$

we get

$$\begin{aligned}
\mathbf{x}^{\alpha,\epsilon} - \mathbf{x} &= A_\alpha^{-1} K^T (\mathbf{y}^\epsilon - \mathbf{y}) + (A_\alpha^{-1} K^T K - \mathcal{I})\mathbf{x} + \alpha A_\alpha^{-1} \mathbf{x}_r \\
&= A_\alpha^{-1} K^T (\mathbf{y}^\epsilon - \mathbf{y}) + A_\alpha^{-1} \Big((K^T K - A_\alpha)\mathbf{x} + \alpha \mathbf{x}_r \Big) \\
&= A_\alpha^{-1} K^T (\mathbf{y}^\epsilon - \mathbf{y}) - \alpha A_\alpha^{-1} (\mathbf{x} - \mathbf{x}_r) \ .
\end{aligned}$$

Therefore,

$$\|\mathbf{x}^{\alpha,\epsilon} - \mathbf{x}\| \le \|A_\alpha^{-1}\| \|K^T\| \|\mathbf{y}^\epsilon - \mathbf{y}\| + \alpha \|A_\alpha^{-1}\| \|\mathbf{x} - \mathbf{x}_r\| \ . \tag{3.29}$$

Assume that we will solve the inverse problem with *a priori* information — we know an approximate value of $\mathbf{x}$, $\mathbf{x}_r$, and a value $\delta > 0$ (as small as we are able to get), such that $|\mathbf{x} - \mathbf{x}_r| \leq \delta$, (i.e., $\mathbf{x}_r$ is a vector not much different from the solution, $\mathbf{x}$, we are looking for). In this case, the second term on the right side of the inequality given in Eq. (3.29), would be bounded by $\alpha\delta|(K^T K)^{-1}|$, which could be small, even if α is not. When $|\mathbf{x}|$ is large (in the absence of a reference value we may take $\mathbf{x}_r = 0$), it is advantageous to control the second term using some previous information. Thus, α is not forced to be small. That is convenient since it is preferable to choose α slightly different from zero, on the account of the first term. In fact, the first term, which depends on the multiplication factor of the data error, $|A_\alpha^{-1}|$, is not too large when α is not too small (see the graph of $E_1(\alpha)$ in Fig. 3.2a).

This analysis justifies the following assertion:

in inverse problems, good quality additional information contributes to a better determination of the solution, even when experimental errors are present.

3.7 Steepest Descent Method

Another regularization scheme that can be applied to the problem given by Eq. (3.5) consists in considering an *iterative minimization method* for functional f, defined by Eq. (3.15). Instead of directly solving the algebraic linear system, Eq. (3.5), we choose an indirect method, which generates a minimizing sequence for the functional f. This implies a change in the regularization strategy with respect to Tikhonov's method, as we shall see.

Assume that $\mathbf{x}_\infty$ is the minimum point of f. We say that $\mathbf{x}_l \in \mathbb{R}^n$, with $l \in \mathbb{N}$, is a *minimizing sequence* of f if

$$\mathbf{x}_l \to \mathbf{x}_\infty \text{ as } l \to +\infty \, .$$

In this case, $\mathbf{x}_l$ converges to the solution of Eq. (3.5), whenever K is invertible. Thus, if we generate a minimizing sequence for the functional defined in Eq. (3.15), we construct a sequence that converges to the solution of Eq. (3.5). If we stop the sequence in a finite number of steps, we may end up with only an approximate solution.

The *steepest descent method* generates a minimizing sequence for f in the following way. Given an initial approximation $\mathbf{x}_o$ of the minimum point of f, one computes $\mathbf{x}_1$, the minimum point of the restriction of f to the line that passes by $\mathbf{x}_o$, whose direction is the (local) steepest descent of function f at $\mathbf{x}_o$.

We recall here that the steepest descent direction of f at $\mathbf{x}_o$ is given by a vector in the direction opposite to the gradient, i.e., for example, in the direction of

$$-\nabla f(\mathbf{x}_o) = -K^T(K\mathbf{x}_o - \mathbf{y}) \, .$$

Therefore, the straight line along which we look for a minimum point of f is given parametrically by

$$\mathbb{R} \ni t \mapsto \mathbf{x}_o - t\nabla f(\mathbf{x}_o) \in \mathbb{R}^n \, .$$

The minimum point of f along the line can be written as

$$\mathbf{x}_1 = \mathbf{x}_o - t^* \nabla f(\mathbf{x}_o)\,, \tag{3.30}$$

where t^* is the minimum point of the real valued function of a real variable,

$$\mathbb{R} \ni t \mapsto h(t) = f\left(\mathbf{x}_o - t\nabla f(\mathbf{x}_o)\right) = \frac{1}{2}|K(\mathbf{x}_o - t\nabla f(\mathbf{x}_o)) - \mathbf{y}|^2 \in \mathbb{R}\,.$$

However,

$$\begin{aligned} h(t) &= \frac{1}{2}\langle K\mathbf{x}_o - \mathbf{y} - tK\nabla f(\mathbf{x}_o), K\mathbf{x}_o - \mathbf{y} - tK\nabla f(\mathbf{x}_o)\rangle \\ &= \frac{1}{2}\left(|K\mathbf{x}_o - \mathbf{y}|^2 - 2t\langle K\nabla f(\mathbf{x}_o), K\mathbf{x}_o - \mathbf{y}\rangle + t^2|K\nabla f(\mathbf{x}_o)|^2\right)\,. \end{aligned}$$

which is a second order polynomial in t. Therefore, the *critical point equation* for h is

$$0 = dh/dt = -\langle K\nabla f(\mathbf{x}_o), K\mathbf{x}_o - \mathbf{y}\rangle + t|K\nabla f(\mathbf{x}_o)|^2\,,$$

that, by using Eq. (3.17), yields

$$t^* = \frac{\langle K\nabla f(\mathbf{x}_o), K\mathbf{x}_o - \mathbf{y}\rangle}{|K\nabla f(\mathbf{x}_o)|^2} = \frac{|K^T(K\mathbf{x}_o - \mathbf{y})|^2}{|KK^T(K\mathbf{x}_o - \mathbf{y})|^2}\,.$$

Using again the expression for $\nabla f(\mathbf{x}_o)$, Eq. (3.17), and Eq. (3.30), we obtain

$$\mathbf{x}_1 = \mathbf{x}_o - \frac{|K^T(K\mathbf{x}_o - \mathbf{y})|^2}{|KK^T(K\mathbf{x}_o - \mathbf{y})|^2}\,K^T(K\mathbf{x}_o - \mathbf{y})\,.$$

The next step of the algorithm is obtained by applying the same idea, beginning now at $\mathbf{x}_1$. Thus, successively we obtain

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{|K^T(K\mathbf{x}_i - \mathbf{y})|^2}{|KK^T(K\mathbf{x}_i - \mathbf{y})|^2}\,K^T(K\mathbf{x}_i - \mathbf{y}) \quad i = 0, 1, 2, \ldots \tag{3.31}$$

This algorithm can be thought of as a *regularization scheme* if we identify the i-th step with parameter α, $\alpha = 1/i$, $i \geq 1$. This will be better seen in the next section when we consider the Landweber method. We note that we are searching for a solution in sets "much smaller" than the whole space, which may lead to a "regularity" in the solution. Even so, this scheme is not of the kind considered in definition 3.1. In particular, it does not have a linear structure and is more difficult to analyze.

3.8 Landweber Method

The *Landweber method* is an iterative minimization scheme obtained by modifying the steepest descent method.

Observing Eq. (3.31), we see that $-\nabla f(\mathbf{x}_i) = -K^T(K\mathbf{x}_i - \mathbf{y})$ is multiplied by the factor

$$\frac{|K^T(K\mathbf{x}_i - \mathbf{y})|^2}{|KK^T(K\mathbf{x}_i - \mathbf{y})|^2} \ .$$

This implies that $\mathbf{x}_{i+1}$ is the best solution through the line

$$t \quad \mapsto \quad \mathbf{x}_i - t\nabla f(\mathbf{x}_i) \ .$$

In the Landweber method this costly and complicated factor is substituted by a constant $\gamma > 0$, and then in every step only a *suboptimal solution* is obtained through the given line. Specified an initial guess $\mathbf{x}_0$, this method is then defined by the iteration

$$\begin{aligned}\mathbf{x}_{i+1} &= \mathbf{x}_i - \gamma K^T(K\mathbf{x}_i - \mathbf{y}) \\ &= \gamma K^T\mathbf{y} + \left(\mathcal{I} - \gamma K^T K\right)\mathbf{x}_i \ ,\end{aligned}$$

for $i = 0, 1, 2, \ldots$.

The solution of this recursion relation is given explicitly in terms of the initial condition $\mathbf{x}_0$ and the vector $\mathbf{y}$ as

$$\mathbf{x}_i = \gamma \left[\sum_{l=0}^{i-1} \left(\mathcal{I} - \gamma K^T K\right)^l\right] K^T\mathbf{y} + \left(\mathcal{I} - \gamma K^T K\right)^i \mathbf{x}_o \ , \quad \text{for } i = 1, 2, \ldots \tag{3.32}$$

We now analyze the behaviour of this iteration scheme to check when it is a linear regularization scheme. Let $D(\theta_1, \theta_2, \ldots, \theta_n)$ denote a diagonal matrix with diagonal entries $\theta_1, \theta_2, \ldots, \theta_n$. Let also $\lambda_j = \lambda_j(K^T K)$ be the j^{th} largest eigenvalue of $K^T K$, and $\mathbf{v}_j$ a corresponding eigenvector, in such a way that $P = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is an orthogonal matrix diagonalizing $K^T K$, i.e.,

$$K^T K = PD(\lambda_1, \lambda_2, \ldots, \lambda_n)P^T \ .$$

Since

$$\sum_{l=o}^{i-1} (1 - \gamma\lambda_j)^l = (1 - (1 - \gamma\lambda_j)^i)/\gamma\lambda_j \ ,$$

we have

$$\begin{aligned}\mathbf{x}_i &= \gamma P \left\{\sum_{l=0}^{i-1} \left[\mathcal{I} - \gamma D(\lambda_1, \ldots, \lambda_n)\right]^l\right\} P^T K^T \mathbf{y} \\ &\quad + P\left[\mathcal{I} - \gamma D(\lambda_1, \ldots, \lambda_n)\right]^i P^T \mathbf{x}_0 \\ &= PD\left(\frac{1 - (1 - \gamma\lambda_1)^i}{\lambda_1}, \ldots, \frac{1 - (1 - \gamma\lambda_n)^i}{\lambda_n}\right) P^T K^T \mathbf{y} \\ &\quad + PD\left((1 - \gamma\lambda_1)^i, \ldots, (1 - \gamma\lambda_n)^i\right) P^T \mathbf{x}_0 \ .\end{aligned} \tag{3.33}$$

Thus, identifying the i-th step with the *regularization parameter* α, $\alpha = 1/i$, $i \geq 1$, the i-th iterate, $\mathbf{x}_i$, is the solution of the following equation

$$A_{1/i}\mathbf{x}_{1/i} = K^T\mathbf{y} + \mathbf{b}_{1/i}\,, \tag{3.34a}$$

where

$$A_{1/i} = PD\left(\frac{\lambda_1}{1-(1-\gamma\lambda_1)^i}, \ldots, \frac{\lambda_n}{1-(1-\gamma\lambda_n)^i}\right)P^T\,, \tag{3.34b}$$

and

$$\mathbf{b}_{1/i} = PD\left(\frac{\lambda_1(1-\gamma\lambda_1)^i}{1-(1-\gamma\lambda_1)^i}, \ldots, \frac{\lambda_n(1-\gamma\lambda_n)^i}{1-(1-\gamma\lambda_n)^i}\right)P^T\mathbf{x}_0\,. \tag{3.34c}$$

Theorem 5. The family $A_{1/i}$, $\mathbf{b}_{1/i}$, $i = 1,2,\ldots$, defines a *linear regularization scheme* for the linear problem $K^TK\mathbf{x} = K^T\mathbf{y}$, when parameter $\gamma > 0$ is sufficiently small in such a way that

$$0 < \gamma\lambda_j < 1\,, \tag{3.35}$$

for $j=1,\ldots,n$.

Proof. As a matter of fact,

$$\begin{aligned} A_{1/i} &\to PDP^T = K^TK\,, \\ \mathbf{b}_{1/i} &\to 0\,, \end{aligned} \tag{3.36}$$

as $i \to +\infty$, and since $\left|1-\gamma\lambda_j\right| < 1$, we have that $\left(1-\gamma\lambda_j\right)^i \to 0$ as $i \to \infty$, and then

$$|A_{1/i}^{-1}| = \frac{1-(1-\gamma\lambda_n)^i}{\lambda_n} \nearrow \frac{1}{\lambda_n} = |(K^TK)^{-1}| \quad \text{as} \quad i \to +\infty\,.$$

■

Equation (3.33) tells us that the eigenvalues of the inverse, $1/\lambda_j$, which may be very large (when λ_j is very small, near zero), are mitigated by the regularization scheme by means of the factor

$$1-(1-\gamma\lambda_j)^i\,.$$

This factor is close to zero whenever λ_j is close to zero. We also remark that the initial condition $\mathbf{x}_0$ is progressively "forgotten" as it is multiplied by $(1-\gamma\lambda_j)^i$, which tends to zero as i grows, for all $j = 1,2\ldots,n$.

Note that Eq. (3.35) can be replaced by asking only that γ satisfies

$$0 < \gamma < \frac{1}{\lambda_1}\,. \tag{3.37}$$

3.9 Discrepancy Principle

In any regularization scheme such as steepest descent, Landweber iteration or conjugated gradient method, the *regularization parameter* α is related to the iteration counter i, $\alpha = 1/i$.

Applying a linear regularization scheme in a problem having at most a level ϵ of data error, $|\mathbf{y} - \mathbf{y}^\epsilon| \leq \epsilon$, we have

$$\begin{aligned} |K\mathbf{x}^{\alpha,\epsilon} - \mathbf{y}^\epsilon| &= |KK_\alpha^{-1}(\mathbf{y}^\epsilon + \mathbf{b}_\alpha) - \mathbf{y}^\epsilon| \\ &\leq |KK_\alpha^{-1} - \mathcal{I}|\, |\mathbf{y}^\epsilon| + |KK_\alpha^{-1}|\, |\mathbf{b}_\alpha| \,. \end{aligned}$$

It is obvious that, as the iteration number tends to infinity, $i \to +\infty$,

$$\alpha = 1/i \to 0 \quad \text{and} \quad |K\mathbf{x}^{\alpha,\epsilon} - \mathbf{y}^\epsilon| \to 0 \,.$$

This is illustrated in Fig. 3.2b.

Thus, we could think it is appropriate to pre-specify an arbitrary small tolerance, δ, in which case we would proceed with the iterations until

$$|K\mathbf{x}^{1/i,\epsilon} - \mathbf{y}^\epsilon| \text{ becomes smaller than } \delta.$$

We see that, if we adopt this *stopping criterion*, we risk choosing values of $\alpha = 1/i$ smaller than α^*, and, as shown in Fig. 3.2a, the error in

$$|\mathbf{x}^{1/i,\epsilon} - \mathbf{x}| \,,$$

which is to be minimized, but that cannot be directly accessed (when $\mathbf{x}$ is unknown), would end up increasing.

As a matter of fact,

$$|K\mathbf{x}^{1/i,\epsilon} - \mathbf{y}^\epsilon| \text{ can decrease,}$$

while, at the same time,

$$|\mathbf{x}^{1/i,\epsilon} - \mathbf{x}| \text{ can increase as } i \to +\infty \,.$$

This is an extremely important observation. In Fig. 3.2, it is immediately verifiable that this is the case when $\alpha < \alpha^*$.

Since the error in $\mathbf{y}^\epsilon$ is at most ϵ, it is reasonable to assume that the algorithm should stop when

$$|K\mathbf{x}^{1/i,\epsilon} - \mathbf{y}^\epsilon| \quad \leq \quad \epsilon \,.$$

That is, we are choosing $\delta = \epsilon$. This *stopping criterion* is known as *discrepancy principle*.

For a practical application and a visualization of the fact that this principle yields good results, while difficulties arise when it is not in use (i.e., whenever we use too small a δ) see Fig. 7.3 on page 151.

3.10 Conjugate Gradient

The conjugate gradient method is a very important regularization method that generates a minimizing sequence. In fact, when applied to the resolution of a linear system defined by a real symmetric matrix it arrives to the solution in a finite number of steps, at most the size of the matrix, considering the computations are carried out in infinite precision (exact) arithmetic.

For completeness, we present a derivation of the *conjugate gradient* method for finite dimension problems. Conjugate gradient has been introduced by Hestenes and Stiefel [37]. Rather comprehensive presentations are given by Shewchuck [67], and Golub and Van Loan [35]. A conjugate gradient method applied to a heat transfer problem is presented in Chapter 7.

3.10.1 Conjugate Gradient Algorithm

Let A be a real, symmetric, positive definite matrix[5], and consider the problem

$$A\mathbf{x} = \mathbf{b}\,, \tag{3.38}$$

where $\mathbf{b} \in \mathbb{R}^n$ is a known vector in $\mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$ is unknown. Denote the solution by $\mathbf{x}_\star$.

As in the case of the steepest descent method, the conjugate gradient method is an iterative method that searches for solution by minimizing a functional along certain directions. Here we consider the functional[6]

$$E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}\,. \tag{3.39}$$

It can be shown that

$$\nabla E = A\mathbf{x} - \mathbf{b}\,, \tag{3.40}$$

and, therefore, $\mathbf{x}_\star$ is a critical point of E if and only if it solves Eq. (3.38).

The derivation of the conjugate gradient method we present is based on two-step optimization, with very little geometry discussed. Geometrical aspects of the conjugate gradient would take us too long at this point, so we have spelled out the details in exercises for the interested reader in the Appendix A (see exercises beginning on Exercise A.40, page 219).

Before delving in its derivation, we present its algorithmic form. Conjugate gradient uses *search directions* $\mathbf{p}_i$, $i = 0,1,2,\dots,n-1$ defined iteratively. Given an initial guess $\mathbf{x}_0$ for the minimum point of $E = E(\mathbf{x})$, the first search direction is $\mathbf{p}_0 = \nabla E\,|_{x_0} = A\mathbf{x}_0 - \mathbf{b}$, and the method proceeds in the following way,

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \beta_i\mathbf{p}_i \tag{3.41a}$$

$$\mathbf{p}_{i+1} = \nabla E\,|_{x_{i+1}} + \gamma_{i+1}\mathbf{p}_i = A\mathbf{x}_{i+1} - \mathbf{b} + \gamma_{i+1}\mathbf{p}_i\,, \tag{3.41b}$$

[5] The definition of a positive definite matrix is given on page 195.

[6] See Exercise A.13, page 211, to verify that the level sets of E are ellipsoids.

for $i = 0,1,2,\ldots$ with the following choice of parameters,

$$\beta_i = \frac{(A\mathbf{x}_i - \mathbf{b})^T(A\mathbf{x}_i - \mathbf{b})}{\mathbf{p}_i^T A\mathbf{p}_i} \tag{3.41c}$$

$$\gamma_{i+1} = \frac{(A\mathbf{x}_{i+1} - \mathbf{b})^T(A\mathbf{x}_{i+1} - \mathbf{b})}{(A\mathbf{x}_i - \mathbf{b})^T(A\mathbf{x}_i - \mathbf{b})} \tag{3.41d}$$

For simplicity, but mainly for computational efficiency, we introduce the *residual*[7],

$$\mathbf{r}_i = \mathbf{b} - A\mathbf{x}_i \,,$$

and then, applying A on both sides of Eq. (3.41a), and subtracting $\mathbf{b}$, we get the following *recursion* relation for the residuals,

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \beta_i A\mathbf{p}_i \,. \tag{3.42}$$

Therefore, the *conjugate gradient* method can be rewritten in algorithmic form as

$$\beta_i = \frac{\mathbf{r}_i^T\mathbf{r}_i}{\mathbf{p}_i^T A\mathbf{p}_i} \tag{3.43a}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \beta_i\mathbf{p}_i \tag{3.43b}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \beta_i A\mathbf{p}_i \tag{3.43c}$$

$$\gamma_{i+1} = \frac{\mathbf{r}_{i+1}^T\mathbf{r}_{i+1}}{\mathbf{r}_i^T\mathbf{r}_i} \tag{3.43d}$$

$$\mathbf{p}_{i+1} = -\mathbf{r}_{i+1} + \gamma_{i+1}\mathbf{p}_i \tag{3.43e}$$

for $i = 0,1,2,\ldots$, with

$$\mathbf{p}_0 = A\mathbf{x}_0 - \mathbf{b} \,, \text{ and } \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 \,. \tag{3.43f}$$

3.10.2 Two-Step Optimization

We begin by presenting the first optimization step of the conjugate gradient method. Landweber, steepest descent, and conjugate gradient proceed, in each iteration step, along a prescribed direction

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \beta_i\mathbf{p}_i \,,\; i = 0,1,\ldots \tag{3.44}$$

where β_i is a parameter that specifies the size of the step to be taken in the *search* direction, $\mathbf{p}_i$.

Consider the line in $\mathbb{R}^n$ that goes through $\mathbf{x}_i$ and has direction $\mathbf{p}_i$, defined by the image of the function

$$\mathbb{R} \ni \beta \mapsto \mathbf{x}_i - \beta\mathbf{p}_i \in \mathbb{R}^n \,.$$

[7] Here, the residual is directly related to the gradient of E, $\mathbf{r}_i = -\nabla E\mid_{x_i}$.

For conjugate gradient, steepest descent (but not for Landweber), and other so-called *conjugate directions* methods one chooses β_i in such a way that the functional E, Eq. (3.39), restricted to the line, attains its minimum value. The restricted functional is given by

$$\mathbb{R} \ni \beta \mapsto l(\beta) = E(\mathbf{x}_i - \beta\mathbf{p}_i) \in \mathbb{R}\,,$$

where

$$l(\beta) = E(\mathbf{x}_i) - \beta\mathbf{p}_i^T(A\mathbf{x}_i - \mathbf{b}) + \frac{1}{2}\beta^2\mathbf{p}_i^T A\mathbf{p}_i\,. \tag{3.45}$$

It is worthwhile noticing that $l(\beta)$ is just a second degree polynomial in β, with funny looking coefficients. The critical point of l, i.e. the solution of $(dl/d\beta)\,|_{\beta_i} = 0$, is given by

$$\begin{aligned}\beta_i &= -\frac{\text{‘coefficient of }\beta\text{’}}{\text{‘twice the coef. of }\beta^2\text{’}}\\ &= \frac{\mathbf{p}_i^T(A\mathbf{x}_i - \mathbf{b})}{\mathbf{p}_i^T A\mathbf{p}_i} = -\frac{\mathbf{p}_i^T\mathbf{r}_i}{\mathbf{p}_i^T A\mathbf{p}_i}\,.\end{aligned} \tag{3.46}$$

Now, substituting the value of β_i in l, we get the minimum of E restricted to the search line,

$$\begin{aligned}l(\beta_i) = E(\mathbf{x}_i - \beta_i\mathbf{p}_i) &= E(\mathbf{x}_i) - \frac{1}{2}\frac{\left[\mathbf{p}_i^T(A\mathbf{x}_i - \mathbf{b})\right]^2}{\mathbf{p}_i^T A\mathbf{p}_i}\\ &= E(\mathbf{x}_i) - \frac{1}{2}\frac{\left(\mathbf{p}_i^T\mathbf{r}_i\right)^2}{\mathbf{p}_i^T A\mathbf{p}_i}\,.\end{aligned} \tag{3.47}$$

Clearly, the decrease in E,

$$\frac{1}{2}\frac{\left(\mathbf{p}_i^T\mathbf{r}_i\right)^2}{\mathbf{p}_i^T A\mathbf{p}_i}\,, \tag{3.48}$$

depends only on the direction, not on $\mathbf{p}_i$'s norm, as should be expected. In fact, the quantity in Eq. (3.48) is *homogeneous* of degree zero[8] in $\mathbf{p}_i$. In other words, if we change $\mathbf{p}_i$ by a non-null multiple of $\mathbf{p}_i$, $a\mathbf{p}_i$, with $a \neq 0$, the quantity in Eq. (3.48) does not change.

In the case of steepest descent, we would take $\gamma_i = 0$ for all i, in Eqs. (3.43d) and (3.43e), in which case the search *directions* are

$$\mathbf{p}_i = \nabla E\,|_{x_i} = A\mathbf{x}_i - \mathbf{b} = -\mathbf{r}_i,\ i = 0,1,2,\dots\,.$$

It should be remarked that if $\mathbf{x}_i$ is already the solution of the problem, $\mathbf{x}_i = \mathbf{x}_\star$, then $\mathbf{r}_i = 0$, the quantity in Eq. (3.48) is null and, in fact, the minimum value of $l(\beta)$, Eq. (3.45), would occur for $\beta = 0$.

[8] A function $\mathbb{R}^n\backslash\{0\} \ni \mathbf{x} \mapsto g(\mathbf{x}) \in \mathbb{R}^m$ is called *homogeneous* of degree σ if and only if $g(\lambda\mathbf{x}) = \lambda^\sigma g(\mathbf{x})$ for all $\lambda \in \mathbb{R}\backslash\{0\}$, and for all $\mathbf{x} \in \mathbb{R}^n\backslash\{0\}$.

Unless $\mathbf{x}_i$ is already the solution of the problem, the quantity in Eq. (3.48) will be strictly positive, representing an effective reduction in the value of E, which, in principle, is good since the goal is to find the minimum point corresponding to the minimum value of E.

We begin by presenting the second optimization step of the conjugate gradient method which is related to the choice of search direction. This choice has three ingredients. In the first step, it coincides with the steepest descent since the first search direction is $\mathbf{p}_0 = A\mathbf{x}_0 - \mathbf{b}$. The next ingredient is that it consists in a perturbation of the direction of the gradient of the functional E by the previous search direction,

$$\begin{aligned} \mathbf{p}_i &= \nabla E|_{x_i} + \gamma_i \mathbf{p}_{i-1} \\ &= A\mathbf{x}_i - \mathbf{b} + \gamma_i \mathbf{p}_{i-1} \\ &= -\mathbf{r}_i + \gamma_i \mathbf{p}_{i-1} \ , i = 1,2,\dots \end{aligned} \tag{3.49}$$

Finally, γ_i is specified by choosing it in order to maximize the decrease in E, Eq. (3.48). After substituting $\mathbf{p}_i$ given by Eq. (3.49), in Eq. (3.48), we have to maximize the function

$$\mathbb{R} \ni \gamma \mapsto h(\gamma) = \frac{\left[(-\mathbf{r}_i + \gamma \mathbf{p}_{i-1})^T \, \mathbf{r}_i\right]^2}{(-\mathbf{r}_i + \gamma \mathbf{p}_{i-1})^T A \, (-\mathbf{r}_i + \gamma \mathbf{p}_{i-1})} \, . \tag{3.50}$$

The maximum point of this function is given by

$$\gamma_i = -\frac{\mathbf{p}_{i-1}^T \mathbf{r}_i \mathbf{r}_i^T A \mathbf{r}_i - \mathbf{r}_i^T \mathbf{r}_i \mathbf{r}_i^T A \mathbf{p}_{i-1}}{\mathbf{p}_{i-1}^T \mathbf{r}_i \mathbf{r}_i^T A \mathbf{r}_i + \mathbf{r}_i^T \mathbf{r}_i \mathbf{p}_{i-1}^T A \mathbf{p}_{i-1}} \, . \tag{3.51}$$

Although at first sight it may seem daunting, it is a simple computation that we prefer to leave as a guided exercise for the reader, Exercise 3.16.

3.10.3 A Few Geometric Properties

What is left, now, in the derivation of conjugate gradient is to show that β_i and γ_i given, respectively, by Eqs. (3.46) and (3.51) correspond to the values presented in Eq. (3.43). This is a consequence of the geometric properties of the conjugate gradient method, presented next.

Theorem 6. Using the notation introduced previously, the following geometric assertions concerning the conjugate gradient method hold:

(a) (Orthogonality of residuals) $\mathbf{r}_{i+1}$ is orthogonal to $\mathbf{r}_i$, (and also to all the previous residuals $\mathbf{r}_{i-1}, \dots, \mathbf{r}_0$);

(b) (Orthogonality of residuals and search directions) $\mathbf{r}_{i+1}$ is orthogonal to $\mathbf{p}_i$ and to $\mathbf{p}_{i-1}$, (and all the way down until $\mathbf{p}_0$);

(c) ($A-$orthogonality[9] of $\mathbf{p}_i$'s) $\mathbf{p}_{i+2}$ is $A-$orthogonal to $\mathbf{p}_{i+1}$ and $\mathbf{p}_i$ (in fact, $\mathbf{p}_i \perp_A \mathbf{p}_j$, for all $i \neq j$), i.e., $\mathbf{p}_i^T A \mathbf{p}_j = 0$ for all $i \neq j$;

[9] For a discussion of A-orthogonality see Exercise A.5, page 208.

(d) $\mathbf{p}_i^T \mathbf{r}_i = -\mathbf{r}_i^T \mathbf{r}_i$, for all i.

■

Before proving this theorem, we show how it is used to simplify β_i and γ_i.

From Eq. (3.46) and using item (d) of the previous theorem we get,

$$\beta_i = -\frac{\mathbf{p}_i^T \mathbf{r}_i}{\mathbf{p}_i^T A\mathbf{p}_i} = -\frac{(-\mathbf{r}_i + \gamma_i \mathbf{p}_{i-1})^T \mathbf{r}_i}{\mathbf{p}_i^T A\mathbf{p}_i} = \frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{p}_i^T A\mathbf{p}_i} \, . \tag{3.52}$$

This concludes the derivation of Eq. (3.43a) for β_i.

As for Eq. (3.43d) for γ_i, a little more work has to be done. Start by multiplying both sides of Eq. (3.42) from the left by $\mathbf{r}_{i+1}^T$, and using item (a) above to get,

$$\mathbf{r}_{i+1}^T \mathbf{r}_{i+1} = \beta_i \mathbf{r}_{i+1}^T A\mathbf{p}_i \, ,$$

which yields

$$\beta_i = \frac{\mathbf{r}_{i+1}^T \mathbf{r}_{i+1}}{\mathbf{r}_{i+1}^T A\mathbf{p}_i} \, .$$

Equating the expression for β_i from Eq.(3.52) and from the previous equation, we get,

$$\frac{\mathbf{r}_{i+1}^T A\mathbf{p}_i}{\mathbf{p}_i^T A\mathbf{p}_i} = \frac{\mathbf{r}_{i+1}^T \mathbf{r}_{i+1}}{\mathbf{r}_i^T \mathbf{r}_i} \, .$$

Now, from Eq. (3.51) with $i+1$ in place of i, using item (b), $\mathbf{p}_{i-1}^T \mathbf{r}_i = 0$, and the previous equation, we get,

$$\gamma_{i+1} = \frac{\mathbf{r}_{i+1}^T \mathbf{r}_{i+1}}{\mathbf{r}_i^T \mathbf{r}_i} \, .$$

This concludes the derivation of the conjugate gradient algorithm, Eq. (3.43).

Next we prove assertions (a) to (d) of theorem 6.

Proof. (theorem 6). The proof of this theorem is done by induction. We sketch it by showing the first step when $i = 0$, i.e., we show that

(a_0) Orthogonality of residuals: $\mathbf{r}_1$ is orthogonal to $\mathbf{r}_0$;

(b_0) Orthogonality of residual and search direction: $\mathbf{r}_1$ is orthogonal to $\mathbf{p}_0$;

(c_0) $A-$orthogonality of search directions: $\mathbf{p}_1$ is $A-$orthogonal to $\mathbf{p}_0$;

(d_0) $\mathbf{p}_0^T \mathbf{r}_0 = -\mathbf{r}_0^T \mathbf{r}_0$,

and how to prove the second step, when $i = 1$, from the first, i.e., that

(a_1) Orthogonality of residuals: $\mathbf{r}_2$ is orthogonal to $\mathbf{r}_0$ and $\mathbf{r}_1$;

(b_1) Orthogonality of residual and search direction: $\mathbf{r}_2$ is orthogonal to $\mathbf{p}_0$ and $\mathbf{p}_1$;

(c_1) $A-$orthogonality of search directions: $\mathbf{p}_2$ is $A-$orthogonal to $\mathbf{p}_0$ and $\mathbf{p}_1$;

(d_1) $\mathbf{p}_1^T\mathbf{r}_1 = -\mathbf{r}_1^T\mathbf{r}_1$.

First step: $i = 0$. Now we tackle the first step.

(a_0) $\mathbf{r}_1$ is orthogonal to $\mathbf{r}_0$, i.e., $\mathbf{r}_1 \perp \mathbf{r}_0$. By multiplying both sides of Eq. (3.43c), with $i = 0$,

$$\mathbf{r}_1 = \mathbf{r}_0 + \beta_0 A\mathbf{p}_0 ,$$

on the left by $\mathbf{r}_0^T$, substituting the value of β_0 by expression on Eq. (3.43a), and recalling that $\mathbf{p}_0 = -\mathbf{r}_0$, we get

$$\begin{aligned}\mathbf{r}_0^T\mathbf{r}_1 &= \mathbf{r}_0^T\mathbf{r}_0 + \beta_0\mathbf{r}_0^T A\mathbf{p}_0\\ &= \mathbf{r}_0^T\mathbf{r}_0 + \frac{\mathbf{r}_0^T\mathbf{r}_0}{\mathbf{p}_0^T A\mathbf{p}_0}\mathbf{r}_0^T A\mathbf{p}_0 = 0 .\end{aligned}$$

(b_0) Since $\mathbf{p}_0 = -\mathbf{r}_0$, and $\mathbf{r}_0 \perp \mathbf{r}_1$, we conclude that $\mathbf{r}_1 \perp \mathbf{p}_0$.

(c_0) From Eq. (3.43c), with $i = 0$, and $\mathbf{r}_1 \perp \mathbf{r}_0$, we get

$$\begin{aligned}\mathbf{r}_1^T\mathbf{r}_1 &= \mathbf{r}_1^T\left(\mathbf{r}_0 + \beta_0 A\mathbf{p}_0\right)\\ &= \left(\mathbf{r}_0 + \beta_0 A\mathbf{p}_0\right)^T \beta_0 A\mathbf{p}_0\\ &= \beta_0\mathbf{r}_0^T A\mathbf{p}_0 + \beta_0^2\mathbf{p}_0^T A^2\mathbf{p}_0 .\end{aligned}$$

Now, from Eq. (3.43), and pre-multiplication by $\mathbf{p}_0^T A$ we get

$$\begin{aligned}\mathbf{p}_0^T A\mathbf{p}_1 &= \mathbf{p}_0^T A\left(-\mathbf{r}_1 + \gamma_1\mathbf{p}_0\right)\\ &= \mathbf{p}_0^T A\left(-\mathbf{r}_0 - \beta_0 A\mathbf{p}_0 + \frac{\mathbf{r}_1^T\mathbf{r}_1}{\mathbf{r}_0^T\mathbf{r}_0}\mathbf{p}_0\right)\\ &= -\mathbf{p}_0^T A\mathbf{r}_0 - \beta_0\mathbf{p}_0^T A^2\mathbf{p}_0 + \frac{\beta_0\mathbf{r}_0^T A\mathbf{p}_0 + \beta_0^2\mathbf{p}_0^T A^2\mathbf{p}_0}{\mathbf{r}_0^T\mathbf{r}_0}\mathbf{p}_0^T A\mathbf{p}_0 = 0 ,\end{aligned} \tag{3.53}$$

and we conclude that $\mathbf{p}_1$ is $A-$orthogonal to $\mathbf{p}_0$.

(d_0) Since $\mathbf{p}_0 = -\mathbf{r}_0$, item (d) follows for $i = 0$.

Second step: $i = 1$.

(a_1) Now, we consider the orthogonality of the second residual to the previous residuals. From Eq. (3.43c) with $i = 1$,

$$\mathbf{r}_2 = \mathbf{r}_1 + \beta_1 A\mathbf{p}_1 \,.$$

By multiplying the previous equation on the left by $\mathbf{r}_1$, and using Eq. (3.43a) with $i = 1$, we get

$$\begin{aligned}\mathbf{r}_1^T\mathbf{r}_2 &= \mathbf{r}_1^T\mathbf{r}_1 + \beta_1\mathbf{r}_1^T A\mathbf{p}_1\\ &= \mathbf{r}_1^T\mathbf{r}_1 + \frac{\mathbf{r}_1^T\mathbf{r}_1}{\mathbf{p}_1^T A\mathbf{p}_1}\mathbf{r}_1^T A\mathbf{p}_1 = 0\,,\end{aligned}$$

since, from $\mathbf{p}_1 = -\mathbf{r}_1 + \gamma_1\mathbf{p}_0$,

$$\begin{aligned}\mathbf{p}_1^T A\mathbf{p}_1 &= (-\mathbf{r}_1 + \gamma_1\mathbf{p}_0)^T A\mathbf{p}_1\\ &= -\mathbf{r}_1^T A\mathbf{p}_1 + \gamma_1 \overbrace{\mathbf{p}_0^T A\mathbf{p}_1}^{=0} \,.\end{aligned}$$

That is, $\mathbf{r}_2 \perp \mathbf{r}_1$. Also, $\mathbf{r}_2 \perp \mathbf{r}_0$

$$\begin{aligned}\mathbf{r}_0^T\mathbf{r}_2 &= \mathbf{r}_0^T(\mathbf{r}_1 + \beta_1 A\mathbf{p}_1)\\ &= \mathbf{r}_0^T\mathbf{r}_1 + \beta_1(-\mathbf{p}_0)^T A\mathbf{p}_1 = 0\,.\end{aligned}$$

(b_1) Since $\mathbf{r}_2 \perp \mathbf{r}_0$ and $\mathbf{r}_0 = -\mathbf{p}_0$, we have $\mathbf{r}_2 \perp \mathbf{p}_0$. Moreover,

$$\mathbf{r}_2^T\mathbf{p}_1 = \mathbf{r}_2^T(-\mathbf{r}_1 + \gamma_1\mathbf{p}_0) = 0\,.$$

(c_1) We show that $\mathbf{p}_2$ is A-orthogonal to $\mathbf{p}_0$ since

$$\begin{aligned}\mathbf{p}_0^T A\mathbf{p}_2 &= \mathbf{p}_0^T A(-\mathbf{r}_2 + \gamma_2\mathbf{p}_1)\\ &= -\mathbf{p}_0^T A\mathbf{r}_2 + \gamma_2 \overbrace{\mathbf{p}_0^T A\mathbf{p}_1}^{=0}\\ &= \frac{1}{\beta_0}(\mathbf{r}_0 - \mathbf{r}_1)^T\mathbf{r}_2 = 0\,.\end{aligned}$$

We remark that to show $\mathbf{p}_1^T A\mathbf{p}_2 = 0$ one only needs to change 1 to 2 and 0 to 1 in Eq. (3.53).

(d_1) Note that

$$\mathbf{p}_1^T\mathbf{r}_1 = (-\mathbf{r}_1 + \gamma_1\mathbf{p}_0)^T = -\mathbf{r}_1^T\mathbf{r}^1 + \gamma_1 \overbrace{\mathbf{r}_0^T\mathbf{r}^1}^{=0} = -\mathbf{r}_1^T\mathbf{r}^1\,.$$

To complete the proof of theorem 6, it is enough to use induction. ■

The 'dynamics' set forth by the conjugate gradient method, Eq. (3.43), 'preserves' the geometric structure stated in items (a) to (d). That is, at step i, the choice of the next search direction, that is of γ_{i+1}, is made in such a way that:

(a) the next residual, $\mathbf{r}_{i+1}$, is orthogonal to the previous ones;

(b) the next residual, $\mathbf{r}_{i+1}$, is orthogonal to the previous search directions;

(c) the search direction $\mathbf{p}_{i+1}$ is A-orthogonal to the previous search directions;

(d) and $\mathbf{p}_i^T\mathbf{r}_i = -\mathbf{r}_i^T\mathbf{r}_i$.

3.11 Spectral Analysis of Tikhonov's Regularization

We will see in detail the difference between solutions $\mathbf{x}$ and $\mathbf{x}^\alpha$, respectively, of Eqs. (3.5) and (3.22) and, in particular, we will analyze the *spectral behaviour of the regularization* of the problem.

Using the notation introduced in Appendix A, let

$$\sigma_1 \geq \ldots \geq \sigma_n > 0$$

be the *singular values* of K. The smallest singular value $\sigma_n = \sigma_n(K) > 0$, is non null because K is invertible.

The *eigenvalues* of K^TK,

$$\lambda_1 \geq \ldots \geq \lambda_n > 0\ ,$$

are related to the singular values of K by

$$\lambda_i = \sigma_i^2\ ,$$

as can be seen from observation A.1 (d) on page 199.

Let

$$D = D(\theta_1, \ldots, \theta_n)$$

be the diagonal matrix whose main diagonal is formed by $\theta_1, \ldots, \theta_n$. The singular value decomposition of K, theorem 8 on page 195, is rendered as

$$K = QD(\sigma_1, \ldots, \sigma_n)P^T\ ,$$

with $Q = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$, and $P = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, which are appropriate orthogonal matrices whose columns are denoted, respectively, by $\mathbf{w}_i$ and $\mathbf{v}_i$. Therefore,

$$K^TK = PD^2(\sigma_1, \ldots, \sigma_n)P^T = PD(\sigma_1^2, \ldots, \sigma_n^2)P^T\ .$$

Alternatively, we can write, emphasizing the use of matrices and vectors, that

$$K = QD(\sigma_1, \ldots, \sigma_n)P^T = \sum_{i=1}^{n} \sigma_i \mathbf{w}_i \mathbf{v}_i^T \ .$$

Therefore,

$$K^T K = PD(\lambda_1, \ldots, \lambda_n)P^T = PD(\sigma_1^2, \ldots, \sigma_n^2)P^T = \sum_{i=1}^{n} \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T \ . \tag{3.54}$$

Analogously, for the regularized matrix, we have,

$$\begin{aligned} A_\alpha &= \alpha \mathcal{I} + K^T K \\ &= \alpha PP^T + PD(\sigma_1^2, \ldots, \sigma_n^2)P^T \\ &= PD(\alpha + \sigma_1^2, \ldots, \alpha + \sigma_n^2)P^T \\ &= \sum_{i=1}^{n} (\alpha + \sigma_i^2) \mathbf{v}_i \mathbf{v}_i^T \ . \end{aligned} \tag{3.55}$$

Comparing the spectral representations of the matrices $K^T K$ and A_α, provided by Eqs. (3.54) and (3.55), we see how the regularization modifies the original problem, and the role of regularizing parameter.

From the previous results, we can conclude that,

$$K^T = PD(\sigma_1, \ldots, \sigma_n)Q^T = \sum_{i=1}^{n} \sigma_i \mathbf{v}_i \mathbf{w}_i^T \ ,$$

and

$$\begin{aligned} K^{-1} &= PD\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right) Q^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{w}_i^T \ , \\ (K^T K)^{-1} &= PD\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right) P^T = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \mathbf{v}_i \mathbf{v}_i^T \ , \\ (A_\alpha)^{-1} &= PD\left(\frac{1}{\alpha + \sigma_1^2}, \ldots, \frac{1}{\alpha + \sigma_n^2}\right) P^T = \sum_{i=1}^{n} \frac{1}{\alpha + \sigma_i^2} \mathbf{v}_i \mathbf{v}_i^T \ . \end{aligned}$$

Also,

$$A_\alpha^{-1} K^T = PD\left(\frac{\sigma_1}{\alpha + \sigma_1^2}, \ldots, \frac{\sigma_n}{\alpha + \sigma_n^2}\right) Q^T .$$

Finally, the solutions to Eqs. (3.5) and (3.22) are given respectively by

$$\mathbf{x} = K^{-1}\mathbf{y} = \sum_{i=1}^{n} \frac{1}{\sigma_i}\mathbf{v}_i\mathbf{w}_i^T\mathbf{y} = \sum_{i=1}^{n} \frac{1}{\sigma_i}\mathbf{v}_i\langle \mathbf{w}_i, \mathbf{y}\rangle, \tag{3.56}$$

$$\begin{aligned}
\mathbf{x}^{\alpha} &= A_{\alpha}^{-1}K^T\mathbf{y} + \alpha A_{\alpha}^{-1}\mathbf{x}_r \\
&= PD\left(\frac{\sigma_1}{\alpha+\sigma_1^2}, \dots, \frac{\sigma_n}{\alpha+\sigma_n^2}\right)Q^T\mathbf{y} \\
&\quad + PD\left(\frac{\alpha}{\alpha+\sigma_1^2}, \dots, \frac{\alpha}{\alpha+\sigma_n^2}\right)P^T\mathbf{x}_r \\
&= \sum_{i=1}^{n} \frac{\sigma_i}{\alpha+\sigma_i^2}\mathbf{v}_i\mathbf{w}_i^T\mathbf{y} + \sum_{i=1}^{n} \frac{\alpha}{\alpha+\sigma_i^2}\mathbf{v}_i\mathbf{v}_i^T\mathbf{x}_r \\
&= \sum_{i=1}^{n} \frac{\sigma_i}{\alpha+\sigma_i^2}\mathbf{v}_i\langle \mathbf{w}_i, \mathbf{y}\rangle + \sum_{i=1}^{n} \frac{\alpha}{\alpha+\sigma_i^2}\mathbf{v}_i\langle \mathbf{v}_i, \mathbf{x}_r\rangle\,.
\end{aligned} \tag{3.57}$$

From Eqs. (3.56) and (3.57), it is obvious that the regularized solution, $\mathbf{x}_\alpha$, converges to the exact solution, $\mathbf{x}$, as $\alpha \to 0$.

Exercises

3.1. Augment Table 3.1 by considering the values of $|\mathbf{r}|$ and $|\mathbf{r}|/|\mathbf{p}|$ where $K_\alpha\mathbf{r} = \mathbf{p}$, with $\mathbf{p} = (2^{-10},0)$, as α varies. Qualitatively, how the regularized problem behaves for these perturbations?

3.2. Check whether K_α, defined by Eq. (3.4), is a linear regularization scheme.

3.3. Show that Eq. (3.19) is valid.

3.4. Let K be an invertible matrix, and let $\mathbf{x}$, $\mathbf{y}$ be given such that $K\mathbf{x} = \mathbf{y}$. Let $\mathbf{p}$ be a perturbation on $\mathbf{y}$ and $\mathbf{r}$ the corresponding perturbation on $\mathbf{x}$, $K(\mathbf{x}+\mathbf{r}) = \mathbf{y}+\mathbf{p}$.

(a) Show that $\mathbf{r}$ depends only on $\mathbf{p}$ (and not on $\mathbf{y}$ nor $\mathbf{x}$) and show that $\mathbf{r} = \mathbf{r}(\mathbf{p}) = K^{-1}\mathbf{p}$.

(b) Consider the error multiplication factor

$$m_K(\mathbf{p}) = \frac{|\mathbf{r}(\mathbf{p})|}{|\mathbf{p}|}\,, \tag{3.58}$$

and show that it is a homogeneous function of zero degree. (See definition of homogeneous function in footnote on page 72.)

(c) Let $\lambda \neq 0$ be an eigenvalue associated with eigenvector $\mathbf{v}_\lambda$ of K, and compute $m_K(\mathbf{v}_\lambda)$.

(d) Let $\overline{m}_K = \max_{p\neq 0} m_K(\mathbf{p})$. Show that $\overline{m}_K = |K^{-1}|$. (See definition on Eq. (A5).)

3.5. For the family of perturbations on the right hand side, $\mathbf{y}$, of $K\mathbf{x} = \mathbf{y}$,

$$[0, 2\pi] \quad \mapsto \quad \mathbf{p}(\theta) = (\cos\theta, \sin\theta) ,$$

(a) determine the corresponding perturbation on $\mathbf{x}$, on the left hand side, for K given by Eq. (3.1);

(b) compute

$$g(\theta) = m_K(\mathbf{p}(\theta)) ,$$

and sketch its graph;

(c) do the same for the regularized matrix K_α, defined by Eq. (3.4).

3.6. Let

$$A \quad = \quad \begin{pmatrix} 1 & \beta \\ \beta & \gamma \end{pmatrix} .$$

(a) Verify that this class of matrices contains matrix K, from Eq. (3.1), and K_α, from Eq. (3.4).

(b) Determine the eigenvalues of A.

(c) Assume β and γ small. Use the approximation $\sqrt{1+x} \approx 1 + \frac{x}{2}$, for x near zero, to obtain approximations for the eigenvalues of A.

(d) Determine the eigenvectors of A.
Hint. Note that $ax + by = 0$ has solution $(x,y) = \xi(-b,a)$, for all $\xi \in \mathbb{R}$.

(e) Verify directly that the eigenvectors of A are orthogonal.

(f) Compute $m_A(\mathbf{p}(\theta))$ as defined in Exercise (3.5).

(g) Compute $\bar{m}_A = \max_{p\neq 0} m_A(\mathbf{p})$.

3.7. (a) Check the details on the derivation of Eq. (3.16).

(b) Let f_α be defined by Eq. (3.20). Show that its critical point equation is given by Eq. (3.21).

3.8. Let A be a real, square, symmetric matrix with non-negative eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$. A *Tikhonov regularization* of matrix A, is defined by

$$A_{_k} \quad = \quad A + k\mathcal{I} \text{ , for } k \in]0, +\infty[\, .$$

Show that

$$m_{A_k} \quad = \quad \frac{1}{k + \lambda_{_{min}}} ,$$

where m_A is defined in Exercise (3.4), and $\lambda_{_{min}} = \min_{i\in\{1,2,\ldots,n\}} \lambda_i$.

3.9. Consider the symmetric matrix

$$A \quad = \quad \begin{pmatrix} 100 & 98 \\ 98 & 100 \end{pmatrix},$$

or, more generally, let

$$A \quad = \quad \begin{pmatrix} n & n-2 \\ n-2 & n \end{pmatrix}.$$

(a) Compute the characteristic polynomial and find its roots (the eigenvalues of A). Denote the eigenvalues by $\lambda_1 > \lambda_2$.

(b) Determine eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ corresponding to each eigenvalue.

(c) Show that the eigenvectors corresponding to different eigenvalues are orthogonal.

(d) By dividing each eigenvector by its norm, find P such that $A = PDP^T$ where

$$D \quad = \quad \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

3.10. (a) Find the spectrum of

$$B \quad = \quad \begin{pmatrix} 1 & 1-\frac{2}{n} \\ 1-\frac{2}{n} & 1 \end{pmatrix}.$$

(b) Compute $\mathbf{v}^{1,n}$, an orthonormal eigenvector associated with eigenvalue λ_1, and $\mathbf{v}^{2,n}$, an orthonormal eigenvector associated with the other eigenvalue.

(c) Let

$$\mathbf{p}_n(\theta) \quad = \quad \cos\theta\,\mathbf{v}^{1,n} + \sin\theta\,\mathbf{v}^{2,n}\,.$$

Determine $\mathbf{r}_n(\theta)$ such that $B\mathbf{r}_n(\theta) = \mathbf{p}_n(\theta)$.

(d) Determine

$$m_B(\mathbf{p}_n(\theta)) \quad = \quad \frac{|\mathbf{r}_n(\theta)|}{|\mathbf{p}_n(\theta)|}\,.$$

(e) Sketch the graph of $m_B(\mathbf{p}_n(\theta))$, as a function of θ.

(f) Note that $B = A/n$, where A is the matrix of Exercise 3.9. Show that, for large n, $\lambda_1(B) > \lambda_2(B)$, satisfy

$$\lambda_1(B) \approx 2 \quad \text{and} \quad \lambda_2(B) \approx 0\,,$$

that is,

$$\lim_{n\to+\infty} \lambda_1(B) = 2\,, \quad \text{and} \quad \lim_{n\to+\infty} \lambda_2(B) = 0\,.$$

3.11. Let $B_\epsilon = B + \epsilon\mathcal{I}$, a *Tikhonov regularization* of matrix B from Exercise 3.10, and redo that exercise for it.

3.12. Check the details in the proof of theorem 4.

3.13. (a) Derive Eq. (3.28).

(b) Verify the validity of Eq. (3.32).

(c) Check the details on the derivation of Eq. (3.33).

(d) Check the details on the derivation of Eq. (3.34).

(e) Verify the equivalence between Eqs. (3.35) and (3.37).

3.14. We are in debt in Eq. (3.36) since Definition 3.1 requires $|\mathbf{b}_{1/i}| \searrow 0$, as $i \to \infty$. Show that this holds.
Hint. Let $0 < a < 1$ and show that $a^x/(1-a^x)$ is a decreasing function.

3.15. Descent method for positive definite matrices. Let A be a real, symmetric, positive definite matrix, and consider the solution of the linear system

$$A\mathbf{x} = \mathbf{b}\,, \tag{3.59}$$

where $\mathbf{b} \in \mathbb{R}^n$ is given. Let

$$E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}\,.$$

(a) Show that

$$\nabla E = A\mathbf{x} - \mathbf{b}$$

(b) Let $\mathbf{x}_\star$ be the solution of Eq. (3.59), $A\mathbf{x}_\star = \mathbf{b}$. For any $\mathbf{x} \neq \mathbf{x}_\star$, show that

$$E(\mathbf{x}) > E(\mathbf{x}_\star)$$

Hint. Let $\mathbf{x} = \mathbf{x}_\star + \mathbf{h}$ and expand $E(\mathbf{x}_\star + \mathbf{h})$, collecting the zero order terms in $\mathbf{h}$, the first order, and the second order terms. The zero and first order terms have simple expressions.

(c) Let $\mathbb{R} \ni t \mapsto \mathbf{x}_i - t\nabla E\mid_{x_i} \in \mathbb{R}^n$ be a line in $\mathbb{R}^n$, and let $l = l(t)$ be a function given by

$$\mathbb{R} \ni t \mapsto l(t) = E(\mathbf{x}_i - t\nabla E\mid_{x_i}) \in \mathbb{R}\,.$$

Show that its minimum point, t_i^*, is given by

$$t_i^* = \frac{(A\mathbf{x}_i - \mathbf{b})^T(A\mathbf{x}_i - \mathbf{b})}{(A\mathbf{x}_i - \mathbf{b})^T A(A\mathbf{x}_i - \mathbf{b})}\,.$$

(d) Define the *steepest descent* method by

$$\mathbf{x}_{i+1} = \mathbf{x}_i - t_i^*(A\mathbf{x}_i - \mathbf{b}) ,$$

and, likewise, define the *Landweber* method by,

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma(A\mathbf{x}_i - \mathbf{b}) , \tag{3.60}$$

where γ is a positive constant, $\gamma > 0$. Use the spectral theorem, the regularization parameter $\alpha = 1/i$, $i \geq 1$, and mimick the development in Section 3.8, to show that Eq. (3.60) leads to a linear *regularization* scheme associated with solving Eq. (3.59).

3.16. The aim of this exercise is to derive an expression for the maximum point, γ_i, Eq. (3.51), of function $h = h(\gamma)$, Eq. (3.50).

(a) Write $h(\gamma)$ as the quotient of two polynomials of second degree,

$$h(\gamma) = \frac{(a_1\gamma + a_2)^2}{b_1\gamma^2 + b_2\gamma + b_3} , \tag{3.61}$$

that is, determine appropriate constants a_1, a_2, b_1, b_2 and b_3 from Eq. (3.50).

(b) Using the simplified form of $h(\gamma)$ in Eq. (3.61), keeping the constants a_1, a_2, b_1, b_2 and b_3, differentiate it and write the numerator as the product of two polynomials of degree one.

(c) Find the two critical points of h written in terms of a_1, a_2, b_1, b_2 and b_3, and next, substitute their values obtained in item (a).

(d) Show that $h(\gamma) \geq 0$, for all $\gamma \in \mathbb{R}$. Compute $\lim_{\gamma\to\pm\infty} h(\gamma)$. Next, argue that one of the critical points obtained is already a root of h and corresponds to its minimum point, and the other critical point must be a maximum point.

3.17. Let ϵ_i denote the error in the $i-$th iteration of the conjugate gradient method, $\epsilon_i = \mathbf{x}_i - \mathbf{x}_\star$. Recall that $\mathbf{x}_{i+1} = \mathbf{x}_i - \beta_i\mathbf{p}_i$.

(a) From the previous equation, obtain a recursive equation for ϵ_{i+1} in terms of ϵ_i.

(b) If one requires that ϵ_{i+1} be $A-$*orthogonal* to $\mathbf{p}_i$, show that one gets an expression for β_i as in Eq. (3.46).

(c) From Eq. (3.49) for $\mathbf{p}_i$, and assuming that $\mathbf{p}_i$'s are $A-$orthogonal, $\mathbf{p}_i^T A\mathbf{p}_j = 0$ if $i \neq j$, show that

$$\gamma_i = \frac{\mathbf{p}_i^T A\mathbf{r}_i}{\mathbf{p}_{i-1}^T A\mathbf{p}_{i-1}} .$$

The request of $A-$orthogonality can replace the minimization procedures in the derivation of *conjugate gradient* method. However, one still needs to use the stated properties in theorem 6 to simplify parameters β_i and γ_i, as done in Section 3.10.

3.18. Use induction to prove the results stated in Theorem 6.

Chapter 4
Image Restoration

Assume we have access to a certain image obtained by means of a device that, in the process of *acquisition*, causes a degradation, such as *blurring*. The objective of this chapter is to show how to *restore* the image from the *degraded* image, considering specific examples[1].

In the image processing literature, the recovery of an original image of an object is called restoration or reconstruction, depending on the situation. We shall not discuss this distinction[2].

In any case, in the nomenclature of inverse problems we are using as setforth in Section 2.8, both of these cases, restoration or reconstruction, are suitably called *reconstruction* inverse problems.

We assume here that the image is in grayscale. Every shade of gray will be represented by a real number between 0 (pure black) and 1 (pure white). The original image is represented by a set of *pixels* (i, j), $i = 1, \dots, L$ and $j = 1, \dots, M$, which are small monochromatic squares in the plane that make up the image. Each pixel has an associated shade of grey, denoted by I_{ij} or $I(i, j)$, which constitutes an $L \times M$ matrix (or a vector of LM coordinates). A typical image can be made up of $256 \times 512 = 131\,072$ pixels. Analogously, we will denote by Y_{ij}, $i = 1, \dots, L$, $j = 1, \dots, M$ the grayscale of the blurring of the original image.

The inverse problems to be solved in this chapter deal with obtaining $\mathbf{I}$, the original image, from $\mathbf{Y}$, the blurred image. These *problems*[3] are of Type I, and, most of them, deal with inverse reconstruction problems, as presented in Section 2.8, which, properly translated, means that given a degraded image one wants to determine the original image, or, more realistically, to estimate it.

4.1 Degraded Images

We shall assume that the degraded image is obtained from the original image through a linear transformation. Let B be the transformation that maps the original image $\mathbf{I}$ to its degraded counterpart $\mathbf{Y}$, $\mathbf{Y} = B\,\mathbf{I}$, explicitly given by [27]

[1] The results presented in this chapter were obtained by G. A. G. Cidade [24].

[2] We just observe that reconstruction is associated with operators whose singular values are just 0 and 1 (or some other constant value), whereas when restoration is concerned the range of singular values is more extense. For understanding what are the consequences of having just 0 and 1 as singular values see Section A.4.2.

[3] See the classification in Table 2.3, page 48.

$$Y_{ij} = \sum_{i'=1}^{L} \sum_{j'=1}^{M} B_{ij}^{i'j'} I_{i'j'} \ , \quad i = 1, \ldots, L, \, j = 1, \ldots, M \, . \tag{4.1}$$

We call Eq. (4.1) the observation equation. We note that **Y** corresponds to the experimental data, and we call the linear operator B the *blurring matrix*[4] or, strictly speaking, blurring operator.

We assume that the blurring matrix has a specific structure

$$B_{ij}^{i+k\,j+l} = b_{kl} \, , \quad \text{for } -N \leq k \leq N \text{ and } -N \leq l \leq N \, , \tag{4.2a}$$

$$\text{and } B_{ij}^{mn} = 0, \text{ otherwise} \, . \tag{4.2b}$$

Here b_{kl}, for $-N \leq k,\ l \leq N$, is called the *matrix of blurring weights*, This means that blurring is equal for every position of the pixel (homogeneous blurring).

At each pixel (i, j), the blurring takes into consideration the shades of the neighbouring pixels so that the *domain of dependence* is the square neighbourhood centered in pixel (i,j) covering N pixels to the left, right, up and down (see Fig. 4.1) comprising $(2N+1)\times(2N+1)$ pixels whose values determine the value of the pixel (i,j). In particular, (b_{kl}) is a $(2N+1)\times(2N+1)$ matrix. Usually, one takes $N \ll L, M$. Moreover, the coefficients are chosen preferably in such a way that

$$\sum_{k=-N}^{N} \sum_{l=-N}^{N} b_{kl} = 1 \ , \tag{4.3}$$

with $b_{kl} \geq 0$ for $-N \leq k, l \leq N$.

Figure 4.1 illustrates the action of a blurring matrix with $N = 1$, on the pixel $(i, j) = (7, 3)$. This pixel belongs to the discrete grid where the image is defined. The square around it demarcates the *domain of dependence* of the shade of gray Y_{73}, of pixel (7,3) of the blurred image, in tones of gray of the pixels of the original image pixels,

$$I_{62},\ I_{72}, I_{82}, I_{63}, I_{73}, I_{83}, I_{64}, I_{74}, \ \text{ and } \ I_{84} \, .$$

We can assume that b_{kl} admits *separation of variables*, in the sense that $b_{kl} = f_k f_l$, for some f_k, $-N \leq k \leq N$. In this case, $f_k \geq 0$ for all k, and Eq. (4.3) implies that

$$\sum_{k=-N}^{N} f_k = 1 \, . \tag{4.4}$$

Here, f_k can be one of the profiles shown in Fig. 4.2: (a) truncated gaussian, (b) truncated parabolic, or (c) preferred direction. It could also be a combination of the previous profiles or some more general weighting function.

[4] Strictly speaking $B = (B_{ij}^{i'j'})$ is not a matrix. However, it defines a linear operator.

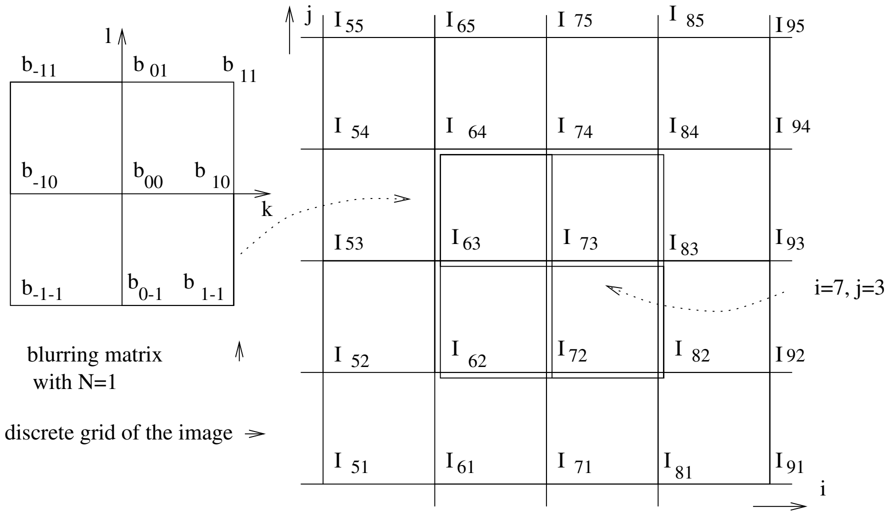


Fig. 4.1 The blurring matrix with $N = 1$ acts on the point $(i, j) = (7, 3)$ of the discrete grid of the image (*pixels*)

With respect to the situation depicted in Fig. 4.1, one possibility is to choose $f_{-1} = 1/4$, $f_0 = 1/2$ and $f_1 = 1/4$ which leads to the following matrix of blurring weights

$$\begin{pmatrix} b_{-1\,-1} & b_{-1\,0} & b_{-1\,1} \\ b_{0\,-1} & b_{0\,0} & b_{0\,1} \\ b_{1\,-1} & b_{1\,0} & b_{1\,1} \end{pmatrix} = \begin{pmatrix} f_{-1} \\ f_0 \\ f_1 \end{pmatrix} (f_{-1}\ f_0\ f_1)$$
$$= \begin{pmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{pmatrix} = \frac{1}{16} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} . \tag{4.5}$$

Therefore, in this case, for instance,

$$\begin{aligned} Y_{73} = \frac{1}{16} & (I_{62} + 2I_{72} + I_{82} \\ & +2I_{63} + 4I_{73} + 2I_{83} + I_{64} + 2I_{74} + I_{84}) . \end{aligned} \tag{4.6}$$

This scheme cannot be taken all the way to the boundaries of the image. At a boundary point we cannot find enough neighbour pixels to take the weighted average. we describe two possible approaches for such situations. One simple approach here is to consider that the required pixels lying outside the image have a constant value, for example zero or one (or some other intermediate constant value). Another approach is to add the weights of the pixels that lie outside of the image to the weights of the pixels that are in the image, in a symmetric fashion. This is equivalent to attributing the pixel outside of the image the shade of gray of its symmetric pixel across the boundary (border) of the image.

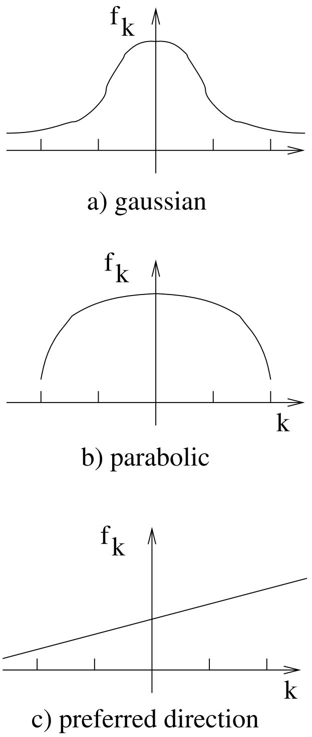


Fig. 4.2 The blurring matrix can represent several kinds of *tensorial* two-dimensional convolutions: gaussian, parabolic, with preferred direction, or a combination of them

For illustration purposes, say now that (7,3) is a pixel location on the right boundary of the image. The positions (8,4), (8,3), (8,2) are outside the image and do not correspond to any pixel. We let the weights 1/16, 2/16 and 1/16 of these positions to be added to the symmetric pixels, with respect to the boundary, respectively, (6,4), (6,3), and (6,2). The result is

$$Y_{73} = \frac{1}{16} \left(2I_{62} + 2I_{72} + 4I_{63} + 4I_{73} + 2I_{64} + 2I_{74} \right) \ . \tag{4.7}$$

4.2 Restoring Images

The inverse reconstruction problem to be considered here is to obtain the vector **I** when the matrix B and the experimental data **Y** are known, by solving Eq. (4.1) for **I**.

We can look at this problem as a finite dimension optimization problem. First, the norm of an image $\mathbf{I}$ is taken as the Euclidean norm of $\mathbf{I}$ thought of as a vector in $\mathbb{R}^{LM}$,

$$|\mathbf{I}| = \left(\sum_{i=1}^{L} \sum_{j=1}^{M} I_{ij}^2 \right)^{\frac{1}{2}} ,$$

not the norm in the set of $L \times M$ matrices, as defined in Eq. (A5). Next, consider the *discrepancy* (or residual) vector $\mathbf{Y} - B\,\mathbf{I}$, and the functional obtained from its norm

$$R(\mathbf{I}) = \frac{1}{2} |\mathbf{Y} - B\,\mathbf{I}|^2 , \tag{4.8}$$

to which we add a Tikhonov's regularization term, [81, 26, 27, 70, 86] getting

$$\begin{aligned} Q(\mathbf{I}) &= \frac{1}{2} |\mathbf{Y} - B\,\mathbf{I}|^2 + \alpha S(\mathbf{I}) \\ &= \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{M} \left(Y_{ij} - \sum_{k=-N}^{N} \sum_{l=-N}^{N} b_{kl} I_{i+k,j+l} \right)^2 + \alpha S(\mathbf{I}) , \end{aligned} \tag{4.9}$$

where S is the regularization term[5] and α is the *regularization parameter*, with $\alpha > 0$. Here, B and $\mathbf{Y}$ are given by the problem. Finally, the inverse problem is formulated as finding the minimum point of Q.

Several regularization terms can be used, and common terms are[6] [25, 26, 16]:

$$\textbf{Norm} \qquad S(\mathbf{I}) = \frac{1}{2} |\mathbf{I} - \bar{\mathbf{I}}|^2 = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{M} \left(I_{ij} - \bar{I}_{ij} \right)^2 \tag{4.10a}$$

$$\textbf{Entropy} \qquad S(\mathbf{I}) = - \sum_{i=1}^{L} \sum_{j=1}^{M} \left(I_{ij} - \bar{I}_{ij} - I_{ij} \ln \frac{I_{ij}}{\bar{I}_{ij}} \right) \tag{4.10b}$$

In both cases, $\bar{\mathbf{I}}$ is a *reference value* (an image as close as possible to the image that is to be restored), known *a priori*.

The notion of regularization and its properties are discussed in Chapter 3. The concept of reference value is introduced in Section 3.4, page 60 and the advantage of its use is explained in Section 3.6, page 63.

[5] To improve readability we insert a comma (,) between the subscripts of I whenever adequate: $I_{i+k,\,j+l}$.

[6] Some regularization terms can be interpreted as Bregman's divergences or distances [14, 81, 25, 42]. The use of Bregman's divergences as regularization terms in Tikhonov's functional was proposed by N. C. Roberty [27], from the Universidade Federal do Rio de Janeiro. Bregman's distance was introduced in [14] and it is not a metric in the usual sense. Exercise A.14 recalls the notion of metric spaces, while Exercise A.35 presents the definition of Bregman's divergences. Some other exercises in Appendix A elucidate the concept of Bregman's divergence.

In the case now under consideration, image restoration, the reference value can be the given blurred image $\bar{\mathbf{I}} = \mathbf{Y}$, or a gray image (i.e., with everywhere constant intensity),

$$\bar{I}_{ij} = c \ \ , \text{ for } i = 1, \dots, L \ j = 1, \dots, M \ \ ,$$

where c is a constant that can be chosen equal to the mean value of the intensities of the blurred image,

$$c = \frac{1}{LM}\left(\sum_{i=1}^{L}\sum_{j=1}^{M} Y_{ij}\right) \ .$$

4.3 Restoration Algorithm

Image restoration can be carried out by Tikhonov's method, with the regularization term given by the entropy functional, Eq. (4.10b). The regularized problem corresponds to the minimum point equation of the functional Q, and is a variant of the one analyzed in Chapter 3. The entropy functional chosen here renders the regularization as non-linear, differing from the one treated in the referred chapter.

Substituting the expression of $S(I)$ given in the right hand side of Eq. (4.10b) in Eq. (4.9), we obtain

$$\begin{aligned} Q(\mathbf{I}) \ &= \ \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{M}\left(Y_{ij} - \sum_{k=-N}^{N}\sum_{l=-N}^{N} b_{kl} I_{i+k,j+l}\right)^2 \\ &\quad -\alpha \sum_{i=1}^{L}\sum_{j=1}^{M}\left(I_{ij} - \bar{I}_{ij} - I_{ij}\ln\frac{I_{ij}}{\bar{I}_{ij}}\right) \ . \end{aligned} \tag{4.11}$$

To minimize this functional, the *critical point equation* is used

$$\frac{\partial Q}{\partial I_{rs}} = 0 \qquad \text{for all } r = 1, \dots, L,\ s = 1, \dots, M \ . \tag{4.12}$$

This is a non-linear system of LM equations and LM unknowns, I_{ij}, $i = 1, \dots, L$, $j = 1, \dots, M$. For notational convenience, let $F_{rs} = \partial Q/\partial I_{rs}$, and $\mathbf{F}$ the function

$$\mathbb{R}^{LM} \ni \mathbf{I} \mapsto \mathbf{F}(\mathbf{I}) \in \mathbb{R}^{LM} \ , \tag{4.13}$$

where, from Eq. (4.11), its (r, s)-th function[7] is given by F_{rs},

$$F_{rs} = -\sum_{i=1}^{L}\sum_{j=1}^{M}\left(Y_{ij} - \sum_{k=-N}^{N}\sum_{l=-N}^{N} b_{kl} I_{i+k,j+l}\right) b_{r-i,s-j} + \alpha \ln\frac{I_{rs}}{\bar{I}_{rs}} \ . \tag{4.14}$$

[7] When computing F_{rs}, we use that

$$\partial I_{ij}/\partial I_{rs} = \delta_{ir}\delta_{js} \qquad \text{and} \qquad \sum_{k=-N}^{N}\sum_{l=-N}^{N} b_{kl}\partial I_{i+k,j+l}/\partial I_{rs} = b_{r-i,s-j} \ .$$

Using this notation, the system of non-linear critical point equations of Q, Eq. (4.12), becomes

$$\mathbf{F}(\mathbf{I}) = 0\,. \tag{4.15}$$

Therefore, the inverse problem is reduced to solving Eq. (4.15). We will show how this non-linear system can be solved by the Newton's method.

4.3.1 Solution of a System of Non-linear Equations Using Newton's Method

Consider a system of equations like the one in Eq. (4.15). Newton's method is iterative and, under certain circumstances, converges to the solution. We sketch its derivation.

An initial estimate of the solution is needed: $\mathbf{I}^o$. Then, $\mathbf{I}^{p+1}$ is defined from $\mathbf{I}^p$, with $p = 0, 1, \ldots$, using a linearization of Eq. (4.15) by means of a Taylor's series expansion of function F around $\mathbf{I}^p$.

Due to the Taylor's formula[8] of $\mathbf{F}$, we have

$$\begin{aligned}\mathbf{F}(\tilde{\mathbf{I}}) = \mathbf{F}(\mathbf{I}^p) + \sum_{m=1}^{L}\sum_{n=1}^{M} \left.\frac{\partial \mathbf{F}}{\partial I_{mn}}\right|_{I^p} \left(\tilde{I}_{mn} - I^p_{mn}\right) \\ + \mathcal{O}\left(|\tilde{\mathbf{I}} - \mathbf{I}^p|^2\right)\,, \text{ as } \tilde{\mathbf{I}} \to \mathbf{I}^p\,.\end{aligned} \tag{4.16}$$

Here, we should be careful with the dimensions of the mathematical objects. As stated in Eq. (4.13), $\mathbf{F}$, evaluated at any point, is an element of $\mathbb{R}^{LM}$, i.e., the left side of Eq. (4.16) has LM elements. This also holds for every term $\partial\mathbf{F}/\partial I_{mn}$, for all $m = 1, \ldots, L$, $n = 1, \ldots, M$.

For Newton's method, $\mathbf{I}^{p+1}$ is defined by keeping only up to the first order term of Taylor's expansion of $\mathbf{F}$, right side of Eq. (4.16), setting the left side equal to zero (we are iteratively looking for a solution of equation $\mathbf{F} = 0$), and substituting $\tilde{\mathbf{I}}$ by $\mathbf{I}^{p+1}$. Thus, Newton's method for solution of Eq. (4.15) is

$$0 = \mathbf{F}(\mathbf{I}^p) + \sum_{m=1}^{L}\sum_{n=1}^{M} \left.\frac{\partial \mathbf{F}}{\partial I_{mn}}\right|_{I^p} \left(I^{p+1}_{mn} - I^p_{mn}\right)\,. \tag{4.17}$$

To determine $\mathbf{I}^{p+1}$, we assume that $\mathbf{I}^p$ is known. Therefore, we see that Eq. (4.17) is a system of linear equations for $\mathbf{I}^{p+1}$. This system has LM equations and LM unknowns, I^{p+1}_{mn}, where $m = 1, \ldots, L$, $n = 1, \ldots, M$.

Newton's method can be conveniently written in algorithmic form as below. Let the vector of corrections

$$\Delta\mathbf{I}^p = \mathbf{I}^{p+1} - \mathbf{I}^p\,,$$

with entries $(\Delta I^p)_{m,n}$, $m = 1, \ldots, L$, $j = 1, \ldots, M$. Choose an arbitrary tolerance (threshold) $\epsilon > 0$.

[8] Taylor's formula is recalled in Section A.5.

1. **Initialization**
   Choose an initial estimate[9] $\mathbf{I}^0$.

2. **Computation of the increment**
   For $p = 0, 1, \ldots$, determine $\Delta\mathbf{I}^p = (\Delta I^p)_{mn} \in \mathbb{R}^{M^2}$ such that[10]

$$\sum_{m=1}^{L}\sum_{n=1}^{M} \left.\frac{\partial \mathbf{F}}{\partial I_{mn}}\right|_{I^p} \Delta I^p_{mn} = -\mathbf{F}(\mathbf{I}^p)\,. \tag{4.18a}$$

3. **Computation of a new approximation**
   Compute[11]

$$\mathbf{I}^{p+1} = \mathbf{I}^p + \Delta\mathbf{I}^p\,. \tag{4.18b}$$

4. **Use of the stopping criterion**
   Compute $|\Delta\mathbf{I}^p| = |\mathbf{I}^{p+1} - \mathbf{I}^p|$. Stop if $|\Delta\mathbf{I}^p| < \epsilon$. Otherwise, let $p = p + 1$ and go to step 2.

4.3.2 Modified Newton's Method with Gain Factor

Newton's method, as presented in Eq. (4.18), does not always converge. It is convenient to introduce a modification by means of a *gain factor* γ, changing Eq. (4.18b) and substituting it by

$$\mathbf{I}^{p+1} = \mathbf{I}^p + \gamma\Delta\mathbf{I}^p\,, \tag{4.19}$$

that will lead to the convergence of the method to the solution of Eq. (4.15) in a wider range of cases, if the gain factor γ is adequately chosen.

4.3.3 Stopping Criterion

The iterative computations, by means of the modified Newton's method, defined by Eqs. (4.18a) and (4.19), is interrupted when at least one of the following conditions is satisfied

$$|\Delta\mathbf{I}^p| < \epsilon_1,\ |S(\mathbf{I}^{p+1}) - S(\mathbf{I}^p)| < \epsilon_2 \ \text{ or }\ |Q(\mathbf{I}^{p+1}) - Q(\mathbf{I}^p)| < \epsilon_3\,, \tag{4.20}$$

where ϵ_1, ϵ_2 and ϵ_3 are values sufficiently small, chosen *a priori*.

[9] For the problems we are aiming at, the initial estimate, $\mathbf{I}^0$, can be, for example, $\bar{\mathbf{I}}$, i.e., the blurred image, or a totally gray image.

[10] Compare with Eq. (4.17).

[11] Given $\mathbf{I}^p$, by choosing $\Delta\mathbf{I}^p$ and $\mathbf{I}^{p+1}$ as in Eq. (4.18), it follows that $\mathbf{I}^{p+1}$ satisfies Eq. (4.17).

4.3.4 Regularization Strategy

As mentioned in Chapter 3, the regularized problem differs from the original problem. It is only in the limit (as the regularization parameter approaches zero) that the solution of the regularized problem approaches the solution of the original problem, in special circumstances determined by a specific mathematical analysis. On the other hand, as was also mentioned in the same chapter, in practice the regularization parameter should not always approach zero, since, in the inevitable presence of measurement noise, the errors in the solution of the inverse problem can be minimized by correctly choosing the value of the regularization parameter. Thus, it is necessary to find the best regularization parameter, α^*, which lets the original problem be minimally altered, and yet, that the solution remains stable. The regularization presented here is non-linear, in contrast with that defined in Section 3.3.

It is possible to develop algorithms to determine the best regularization parameter, [84, 85]. However, they are computationally costly. A natural approach is to perform numerical experiments with the restoration algorithm, to determine a good approximation for the optimal regularization parameter.

4.3.5 Solving Sparse Linear Systems Using the Gauss-Seidel Method

We now present an iterative method suitable for solving the system of equations (4.18a).

From Eq. (4.14), we obtain

$$C^{rs}_{mn} = \frac{\partial F_{rs}}{\partial I_{mn}} = \sum_{k=-N}^{N} \sum_{l=-N}^{N} b_{kl} b_{r-m+k,s-n+l} + \frac{\alpha}{I_{rs}} \delta_{rm} \delta_{sn} \, , \tag{4.21}$$

for $r, m = 1, 2, \ldots, L$ and $s, n = 1, 2, \ldots, M$. Here, we also use the equations that can be found in the footnote on page 90.

The linear system of equations given by Eqs. (4.18a) and (4.21) is *banded*,[12] the length of it (distance from the non-zero elements to the diagonal) varying with the order of the blurring matrix represented in Eqs. (4.1) and (4.2). The *diagonal* of fourth order tensor C is given by the elements C^{rs}_{mn}, of C, such that $r = m$ and $s = n$, i.e., by elements C^{mn}_{mn}. For some types of blurring operator, it is guaranteed that the diagonal is dominant for the matrix C of the linear system of Eq. (4.18a).

Due to the large number of unknowns to be computed (for example, if $LM = 256 \times 512$), and the features of matrix C, that we just described, an iterative method

[12] Every point of the image is related only to its neighbours, within the reach of the blurring matrix.

is better suited to solve the system of Eq. (4.18a), and we choose the Gauss-Seidel method.[13]

Putting aside the term of the diagonal in Eq. (4.18a), we obtain the correction term of a Gauss-Seidel iteration,

$$\Delta I_{rs}^{p,q+1} = -\frac{1}{(\partial F_{rs}/\partial I_{rs})|_{I^{p,q}}}\left(F_{rs}|_{I^{p,q}} + \sum_{\substack{m=1\\ m\neq r}}^{L} \sum_{\substack{n=1\\ n\neq s}}^{M} \left.\frac{\partial F_{rs}}{\partial I_{mn}}\right|_{I^{p,q}} \Delta I_{mn}^{p,\tilde{q}} \right) . \tag{4.23}$$

Here q is the iteration counter of the Gauss-Seidel method and $\tilde{q}$ can be q or $q+1$. This is so because, depending on the form the elements of $\Delta\mathbf{I}^{p,q}$ are stored in the vector of unknowns $\Delta\mathbf{I}$, for every unknown of the system, characterized by specific r, s, it will use the previous value of the unknowns (i.e., the value computed in the previous iteration, $\Delta\mathbf{I}^{p,q}$) in some (m,n) positions, or the present values $\Delta\mathbf{I}^{p,q+1}$, in other (m,n) positions, computed in the current iteration, $q+1$.

For this problem, we can set the initial estimate to zero, $\Delta\mathbf{I}^{p,0} = 0$.

4.3.6 Restoration Methodology

We summarize here the methodology adopted throughout this chapter:

1. *Original problem.* Determine vector $\mathbf{I}$ that solves the equation
   $B\mathbf{I} = \mathbf{Y}$;

[13] We recall here the Gauss-Seidel method, [35]. Consider the system

$$A\mathbf{x} = \mathbf{b} \ . \tag{4.22}$$

Let D be the diagonal matrix whose elements in the diagonal coincide with the diagonal entries of A. Let L and U denote, respectively, the lower and upper-triangular matrices, formed by the elements of A. Then,

$$A = L + D + U \ ,$$

and the system can be rewritten as

$$(L+D)\mathbf{x} = -U\mathbf{x} + \mathbf{b} \ .$$

Denoting by $\mathbf{x}^q$ the q-th iteration (approximation of the solution), the *Gauss-Seidel method* is: given $\mathbf{x}^0 = \mathbf{x}_0$, arbitrarily chosen, let

$$(L+D)\mathbf{x}^{q+1} = -U\mathbf{x}^q + \mathbf{b} \ ,$$

for $q = 0,1,2\ldots$, until convergence is reached. Using index notation, we have

$$x_i^{q+1} = a_{ii}^{-1}\left\{ b_i - \left(\sum_{j<i} a_{ij}x_j^{q+1} + \sum_{j>i} a_{ij}x_j^{q} \right) \right\} \ .$$

It is expected that $\lim_{q\to+\infty} \mathbf{x}^q = \mathbf{x}$, where $\mathbf{x}$ denotes the solution of Eq. (4.22). This can be guaranteed under special circumstances.

2. *Alternative formulation.* Minimize the function
   $R(\mathbf{I}) = \frac{1}{2}|\mathbf{Y} - B\mathbf{I}|^2$;

3. *Regularized problem.* Minimize the function
   $Q(\mathbf{I}) = R(\mathbf{I}) + \alpha S(\mathbf{I})$;

4. *Critical point.* Determine the critical point equation
   $\nabla Q(\mathbf{I}) = 0$;

5. *Critical point equation solution.* use modified Newton's method, to solve the non-linear critical point system of equations;

6. *Linear system solution.* use Gauss-Seidel method to solve the linear system of equations which appears in Newton's method
   $C(\mathbf{I}^{p+1} - \mathbf{I}^p) = -F(\mathbf{I}^p)$, where C is from Eq. (4.21).

In the following sections, we will present three examples of the application of this methodology to the restoration of a photograph, a text and a biological image.

4.4 Photo Restoration

In Fig. 4.3b, it is shown the blurring of the original 256 × 256 pixels image, presented in Fig. 4.3a, due to a blurring matrix consisting on a Gaussian weight, with a dependence domain of 3 × 3 points, that is, b_{kl} satisfies Eq. (4.3) and

$$b_{kl} \quad \propto \quad \exp\left(-\frac{r_k^2 + r_l^2}{2\sigma^2}\right) \quad , \tag{4.24}$$

where σ is related[14] to the bandwidth and $r_k = |k|$.

The space of shades of gray, [0,1], is discretized and coded with 256 integer values, between 0 and 255, where 0 corresponds to black and 255 to white[15]. The histograms in Fig. 4.3, present the frequency distribution of occurrence of every (discrete) shade in the image. For example, if in 143 (horizontal axis) the frequency is 1003 (vertical axis), it means that there are 1003 pixels in the image with shade 143.

Figure 4.3c exhibits the photograph's restoration, done without regularization, stopped in the 200-th iteration of the Newton's method, and in Fig. 4.3d the regularized restoration ($\alpha = 0.06$).

The histogram of the image restored with regularization is, qualitatively, the one closer to the histogram of the original image. This corroborates the evident improvement in the image restored with regularization.

The behaviour of functionals Q, R and S defined in Eqs. (4.8)–(4.10) is recorded in Fig. 4.4. The minimization of functional Q is related to the maximization of the entropy functional, S.

[14] The symbol $\propto$ means that the quantities are proportional, that is, if $a \propto b$, then there is a constant c such that $a = cb$.

[15] The discretization of the shade space is known as *quantization*.

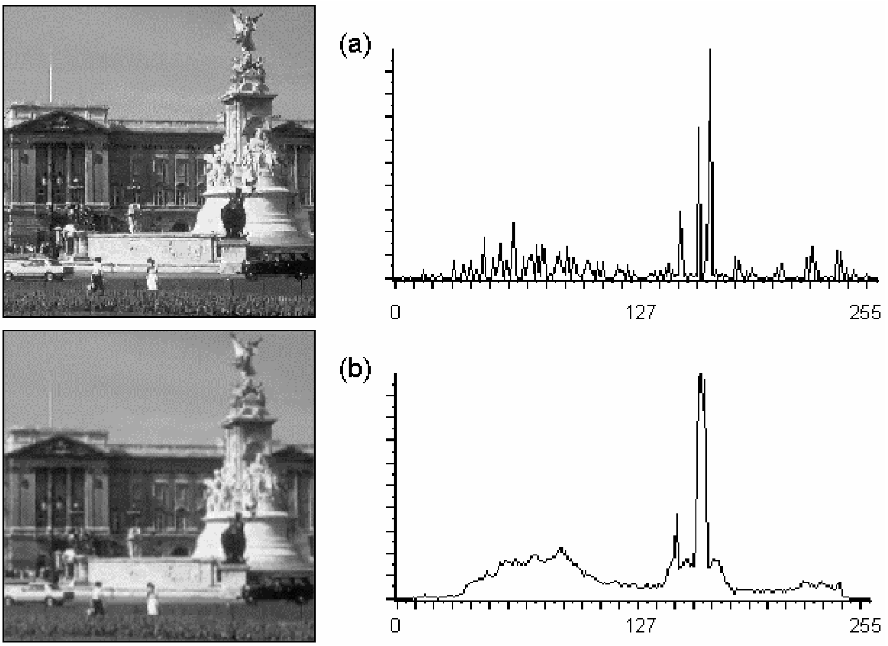


Fig. 4.3 Restoration of an artificially modified image, from a Gaussian with 3×3 points and $\sigma^2 = 1$. Images are on the left and their shade of gray histograms are on the right. (a) original image; (b) blurred image. (Author: G. A. G. Cidade from the Universidade Federal do Rio de Janeiro).

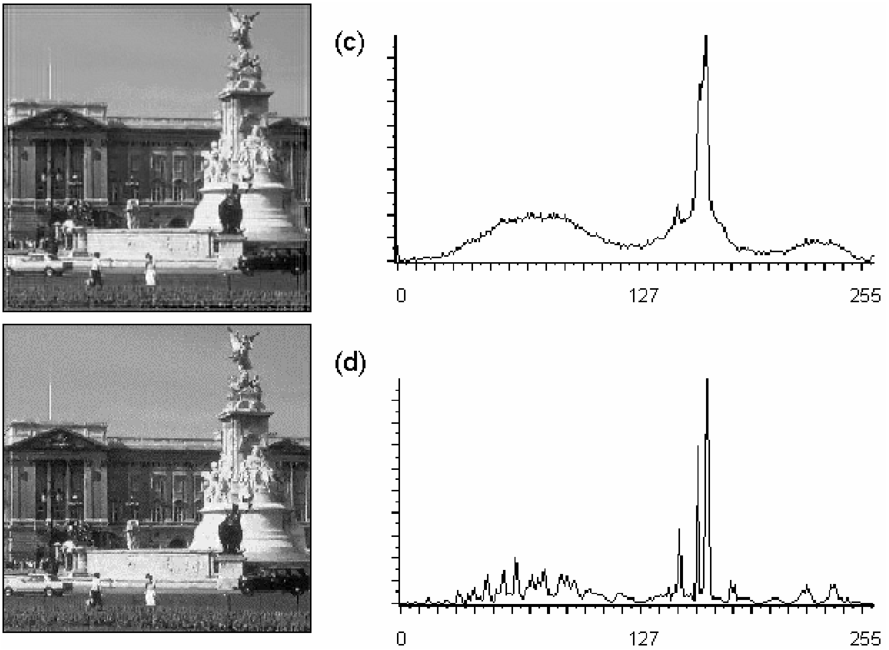


Fig. 4.3 (Cont.) Restoration of an artificially modified image, from a Gaussian with 3×3 points and $\sigma^2 = 1$. (c) restored image without regularization ($\alpha = 0,\ \ \gamma = 0.1$); (d) restored image with regularization ($\alpha = 0.06,\ \ \gamma = 0.1$). (Author: G. A. G. Cidade).

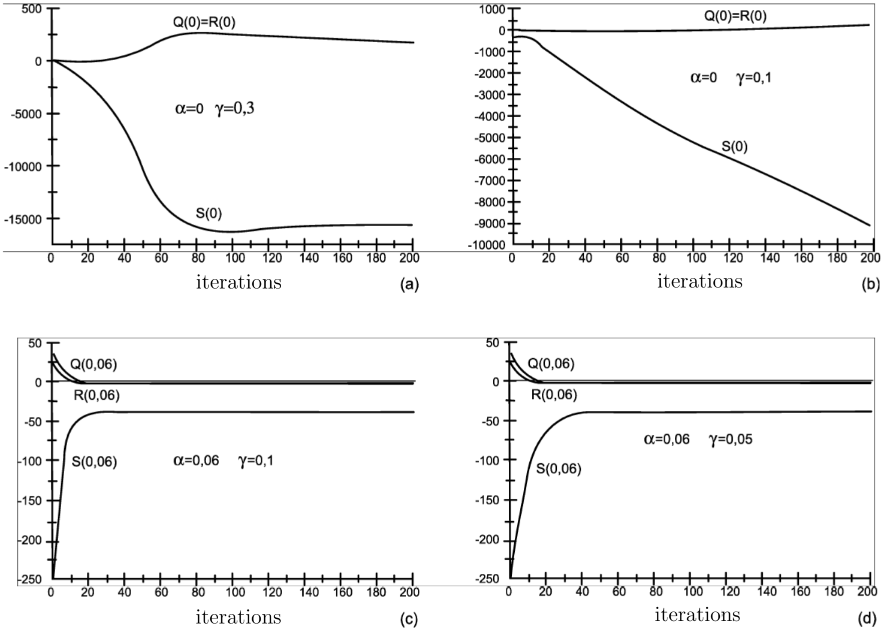


Fig. 4.4 Behaviour of the functionals Q, S and R during the iterative process. (a) $\alpha = 0$ and $\gamma = 0.3$; (b) $\alpha = 0$ and $\gamma = 0.1$; (c) $\alpha = 0.06$ and $\gamma = 0.1$; (d) $\alpha = 0.06$ and $\gamma = 0.05$. (Author: G. A. G. Cidade).

When the regularization parameter is not present, $\alpha = 0$, it is observed that the proposed algorithm diverges, even when one uses a gain factor in the corrections of the intensity in Newton's iterative procedure, Eq. (4.19), as can be seen in Figs. 4.4a,b.

Figures 4.4c,d show the convergence of the algorithm, that naturally is achieved faster for the largest gain factor, $\gamma = 0.1$. In this case, approximately 20 iterations are needed. On the other hand, 40 iterations will be necessary if $\gamma = 0.05$. Functional R, shown in these figures, corresponds to half the square of the norm of the *residuals*, defined as the difference between original blurred image, $\mathbf{Y}$, and restored blurred image, $B\mathbf{I}$, given by

$$R(\mathbf{I}) = \frac{1}{2}|\mathbf{Y} - B\mathbf{I}|^2 \, .$$

4.5 Text Restoration

Figures 4.5a and b present an original text (256 × 256 pixels) and the result of its blurring by means of a Gaussian blurring matrix with 5 × 5 points and $\sigma^2 = 10$.

Text restoration is shown in Fig. 4.5c. There are *border effects*, that is, structures along the border that are not present neither in the original text nor in the blurred image. These occur due to inadequate treatment of pixels near the border (boundary) of the image. This effect can be minimized by considering reflexive conditions at the borders, or simply by considering null the intensity of elements outside the image. See Exercises 4.2, 4.4.

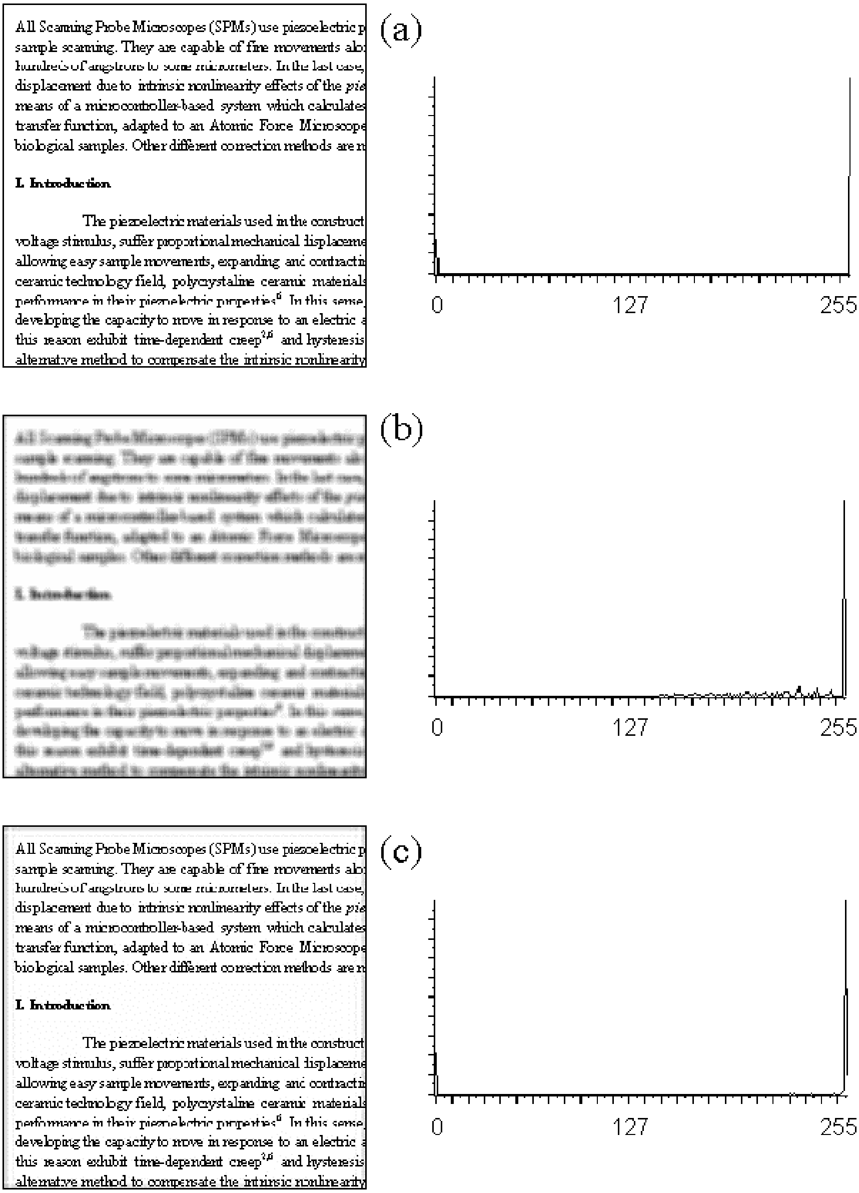
A simple text has, essentially, but two shades: black and white. This is reflected by the histograms of the original and restored texts (Fig. 4.5). However, the blurred text exhibits gray pixels, as shown by its histogram Fig. 4.5b. Notice that the original and restored texts can be easily read, unlike the blurred text.

4.6 Biological Image Restoration

In this section we consider a biological image restoration, which consists of an example of an inverse problem involving a combination of identification and reconstruction problems (problems P_2 and P_3, Section 2.8).

Results of applying the methodology described in this chapter to a real biological image of 600 nm × 600 nm are presented in Fig. 4.6. The image represents an erythroblast being formed, under a leukemic pathology. This image has been acquired by means of an atomic force microscope at the Institute of Biophysics Carlos Chagas Filho, of the Universidade Federal do Rio de Janeiro [25].

In the inverse problem presented here, the original image, $\mathbf{I}$, is being restored together with the choice of the blurring operator given by matrix B. This is an ill-posed problem to determine $\mathbf{I}$, since neither the blurring matrix B is known, in contrast with the problems treated in the previous sections, nor the original image is known.



All Scanning Probe Microscopes (SPMs) use piezoelectric p
sample scanning. They are capable of fine movements alo
hundreds of angstroms to some micrometers. In the last case,
displacement due to intrinsic nonlinearity effects of the *pie*
means of a microcontroller-based system which calculates
transfer function, adapted to an Atomic Force Microscope
biological samples. Other different correction methods are n

I. Introduction

The piezoelectric materials used in the construct
voltage stimulus, suffer proportional mechanical displaceme
allowing easy sample movements, expanding and contracti
ceramic technology field, polycrystaline ceramic materials
performance in their piezoelectric properties[6]. In this sense,
developing the capacity to move in response to an electric
this reason exhibit time-dependent creep[2,6] and hysteresis
alternative method to compensate the intrinsic nonlinearity

All Scanning Probe Microscopes (SPMs) use piezoelectric p
sample scanning. They are capable of fine movements alo
hundreds of angstroms to some micrometers. In the last case,
displacement due to intrinsic nonlinearity effects of the *pie*
means of a microcontroller-based system which calculates
transfer function, adapted to an Atomic Force Microscope
biological samples. Other different correction methods are n

I. Introduction

The piezoelectric materials used in the construct
voltage stimulus, suffer proportional mechanical displaceme
allowing easy sample movements, expanding and contracti
ceramic technology field, polycrystaline ceramic materials
performance in their piezoelectric properties[6]. In this sense,
developing the capacity to move in response to an electric
this reason exhibit time-dependent creep[2,6] and hysteresis
alternative method to compensate the intrinsic nonlinearity

Fig. 4.5 Restoration of a text blurred by means of a Gaussian. (a) original text; (b) blurred text; (c) restored text with regularization parameter $\alpha = 0.2$, and gain factor $\gamma = 0.1$. (Author: G. A. G. Cidade).

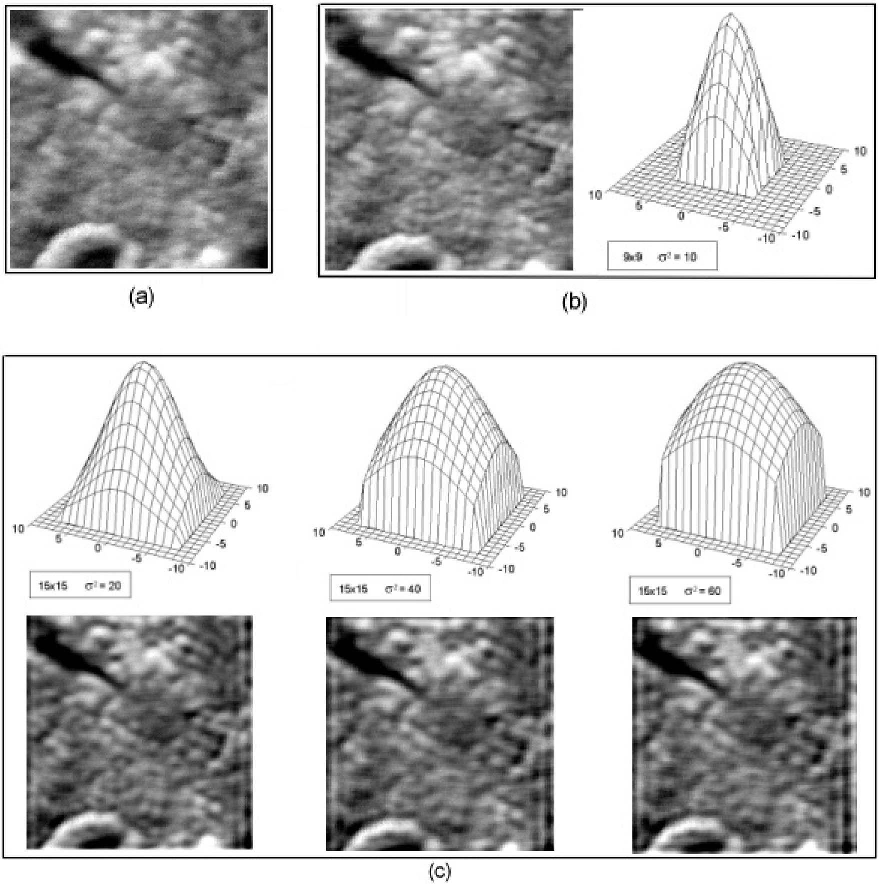


Fig. 4.6 Restorarion of an image obtained by means of an atomic force microscope, together with the blurring matrix identification . (a) original image; (b) restored image with $\alpha = 0.03$, $\gamma = 0.2$, Gaussian with 9×9 points and $\sigma^2 = 10$; (c) restored image with $\alpha = 0.03$, $\gamma = 0.2$, and Gaussians with 15×15 points and $\sigma^2 = 20$, 40 and 60. (Author: G. A. G. Cidade. Images acquired with an Atomic Force Microscope at the Instituto de Biofísica Carlos Chagas Filho of the Universidade Federal do Rio de Janeiro.)

If we knew the original image, or if we knew several images and their blurred counterparts, we could adapt the methodology employed to solve the model problem considered in Chapter 1 to *identify* the blurring matrix.

In the absence of a more quantitative criterion, the blurring matrix identification is performed qualitatively, taking into consideration the perception of medical *specialists*, in relation to the best informative content, due to different restorations, obtained from assumed blurring matrices.

To find the blurring matrix in a less conjectural way, one can use *Gaussian topographies* (see Fig. 4.6) to represent, as much as possible, the geometric aspect of the tip of the microscope, which is the main cause for the degradation that the image undergoes. In other words, we assume that the class of blurring operators is known, i.e., *we characterize the model*, being the specific model identified simultaneously with the restoration of the image.

We used blurring matrices of 9 × 9, 15 × 15 and 21 × 21 points, with different values for σ^2: 10, 20, 40 and 60.

From Fig. 4.6, it can be concluded that we gain more information with Gaussian topographies of 15 × 15 than with those of 9 × 9. For the tests with 21 × 21 points the results were not substantially better.

Figure 4.6c, resulting from the procedure, is considerably better than the blurred image. As in the previous section, the border effects here present are also due to inadequately considering the outside neighbour elements of the border of the image.

Exercises

4.1. Show that Eq. (4.4) is valid.

4.2. Assume that pixel (7,3) is located at the upper right corner of the image. Following the deduction of Eq. (4.7), and assuming symmetry, show that we should take

$$Y_{73} = \frac{1}{16}\left(4I_{62} + 4I_{72} + 4I_{63} + 4I_{73}\right) = \frac{1}{4}\left(I_{62} + I_{72} + I_{63} + I_{73}\right) .$$

4.3. Consider a 3 × 3 blurring weight matrix[16]

$$b = \begin{pmatrix} b_{-1\,-1} & b_{-1\,0} & b_{-1\,1} \\ b_{0\,-1} & b_{0\,0} & b_{0\,1} \\ b_{1\,-1} & b_{1\,0} & b_{1\,1} \end{pmatrix} .$$

Let **I** be an image and **Y** its blurred counterpart. Assume we use symmetric conditions on the boundary. Work out explicit formulae for the blurred image **Y** tone of grays, Y_{ij}, if

(a) pixel (i,j) is in the interior of the image;

[16] Working with indices is sometimes very cumbersome. Several of the following exercises proposes practicing a little ‘indices mechanics’...

(b) pixel (i,j) is in the image's lower boundary;

(c) pixel (i,j) is in the image's lower left corner.

4.4. Do a similar problem as Exercise (4.3), however, instead of using a symmetric condition at boundaries and corners, assume that outside the image, pixels have a uniform value, denoted by I_{ext}.

4.5. Given two pixels (i,j) and (i',j'), we define their *distance* by

$$d\left((i,j),(i',j')\right) = \max\{|i-i'|,|j-j'|\}\,,$$

where max of a finite set of real numbers denotes the largest one in the set.

(a) Give explicitly the pixels that comprise the *circle* centered at (i,j) and radius 1,

$$C_{(i,j)}(1) = \{(i',j') \mid d\left((i,j),(i',j')\right) = 1\}\ .$$

(b) Determine the *disk* $B_{(i,j)}(2)$, centered at (i,j) and radius 2,

$$B_{(i,j)}(2) = \{(i',j') \mid d\left((i,j),(i',j')\right) \le 2\}\ .$$

(c) Sketch the sets $C_{(i,j)}(1)$ and $B_{(i,j)}(2)$.

(d) Give, in the same way, the circle with center (i,j) and radius N.

4.6. (a) Let

$$\mathcal{P} = \{(i',j'),\ i',j' = 1,\ldots,M\} = \{i,\ i = 1,\ldots,M\}^2\ ,$$

be the set of pixels of a square image. Given a set of pixels $\mathcal{S} \subset \mathcal{P}$, define the distance of pixel (i,j) to set $\mathcal{S}$ by

$$d\left((i,j),\mathcal{S}\right) = \min\left\{d\left((i,j),(i',j')\right),(i',j') \in \mathcal{S}\right\}\ ,$$

where min of a finite set of real numbers denotes the smallest one in the set. The *right boundary* of the image is

$$R = \{(i',j') \in \mathcal{P},\ \mid i' = M\} = \{(M,j'),\ j' = 1,\ldots,M\}\ .$$

Determine $d\left((i,j),R\right)$.

(b) Likewise, define respectively, L, the left, U, the top, and D, the bottom image boundaries, and compute $d\left((i,j),L\right)$, $d\left((i,j),U\right)$, and $d\left((i,j),D\right)$.

(c) The boundary of the image is defined by $\mathcal{B} = L \cup T \cup R \cup D$. Compute $d\left((i,j),\mathcal{B}\right)$.
Hint. Use functions max and/or min to express your answer.

4.7. Assume that B is a blurring operator with form given by Eq. (4.2).

(a) Let (i,j) be a fixed *interior* pixel (not on the boundary of the image), and at a distance at least N from the boundary of the image. In particular, $B_{(i,j)}(N) \subset \mathcal{P}$. Show that

$$Y_{ij} = \sum_{k=-N}^{N} \sum_{l=-N}^{N} b_{kl} I_{i+k,j+l} \,. \tag{4.25}$$

Hint. Change, for instance, the index of summation i', by k, with $i' = i + k$, in Eq. (4.1), that is 'center' the summation around i.

(b) The structure of the blurring operator cannot be taken all the way to the boundary of the image. This is why in (a), the pixel (i,j) is restricted to the pixels of distance N or more from the boundary. Verify this. If (i,j) has distance less than N, we cannot compute Y_{ij} using Eq. (4.25).

4.8. Given a blurring operator B, as in Eq. (4.1), define the *domain of dependence* of pixel (i,j) as the set of pixels of the image $\mathbf{I}$ that contribute to the value of the pixel in the blurred image, Y_{ij}, that is,

$$\mathcal{D}_{(i,j)} = \left\{ (i',j') \text{ such that } B_{ij}^{i'j'} \neq 0 \right\} .$$

Assume that the blurring operator B has the structure specified in Eq. (4.2). Determine for which pixels $(i,j) \in \mathcal{P}$, one has

$$\mathcal{D}_{(i,j)} = B_{(i,j)}(N) \,.$$

4.9. The blurring operator B has the structure presented in Eq. (4.2). Let β be the $(2N + 1) \times (2N + 1)$ matrix given by

$$\beta = \begin{pmatrix} b_{-N,-N} & \cdots & \cdots & \cdots & b_{-N,N} \\ \vdots & \ddots & \vdots & & \vdots \\ b_{0,-N} & & b_{0,0} & & b_{0,N} \\ \vdots & & \vdots & \ddots & \vdots \\ b_{N,-N} & \cdots & \cdots & \cdots & b_{N,N} \end{pmatrix} .$$

(a) Give an expression for the entries of β, $\beta_{i'j'}$, in terms of b_{kl}. That is, determine $k(i')$ and $l(j')$ such that

$$\beta_{i'j'} = b_{k(i')} b_{l(j')}, \text{ for } i',j' = 1, \ldots, 2N + 1 \,.$$

(b) For pixel (i,j), define the $(2N + 1) \times (2N + 1)$ matrix $\mathcal{I} = \mathcal{I}^{(i,j)}$, given by

$$\mathcal{I} = \mathcal{I}^{(i,j)} = \begin{pmatrix} I_{i-N,j-N} & \cdots & I_{i-N,j} & \cdots & I_{i-N,j+N} \\ \vdots & \ddots & \vdots & & \vdots \\ I_{i,j-N} & & I_{i,j} & & I_{i,j+N} \\ \vdots & & \vdots & \ddots & \vdots \\ I_{i+N,j-N} & \cdots & I_{i+N,j} & \cdots & I_{i+N,j+N} \end{pmatrix} . \tag{4.26}$$

Give an expression for the entries of $\mathcal{I} = \mathcal{I}^{(i,j)}$, $\mathcal{I}^{(i,j)}_{i',j'}$, in terms of the entries I_{kl} of the image $\mathbf{I}$, similar to what was done in (a).

4.10. (a) Given $m \times n$ matrices, A and B, define the following *pointwise matrix product*, giving rise to a $m \times n$ matrix, $C = A : B$ where $C_{ij} = A_{ij} \cdot B_{ij}$. Compute $\beta : \mathbf{I}$, where β is defined in Exercise 4.9, and $\mathbf{I}$ is an image.

(b) Compute the pointwise matrix product, $\tilde{\beta} : \mathcal{I}^{(7,3)}$, between matrix given by Eq. (4.5), denoted here by $\tilde{\beta}$, and $\mathcal{I}^{(7,3)}$, defined in Eq. (4.26), when $N = 1$, and $(i,j) = (7,3)$.

(c) Let $S(A)$ be the sum of all elements of matrix A,

$$S(A) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} \, .$$

Compute $S(\tilde{\beta} : \mathcal{I}^{(7,3)})$ and compare with Eq. (4.6).

(d) Show that Eq. (4.25) can be written as

$$Y_{ij} = S\left(\tilde{\beta} : \mathcal{I}^{(i,j)}\right) \, .$$

4.11. The entropy regularization, Eq. (4.10b), makes use of the function $s(x) = x_0 - x + x \ln \frac{x}{x_0}$, with $x_0 = \bar{I}_{ij}$ and $x = I_{ij}$. Take, for concreteness, $x_0 = 1/2$.

(a) Compute and sketch s'.

(b) Compute and sketch s''. Show that $s''(x) > 0$ for all x. Conclude that s is strictly convex[17].

(c) Sketch s.

4.12. Let $\mathbf{f} = (f_p)$ be a vector in $\mathbb{R}^n$ with entries f_p, and m a constant in $\mathbb{R}$. Consider Csiszár measure, [42],

$$\Theta_q = \frac{1}{1+q} \sum_p f_p \frac{f_p^q - m^q}{q} \, , \tag{4.27}$$

and Bregman divergence, [14],

$$\mathcal{B}_{\Theta_q}\left(f, \overline{f}\right) = \Theta_q(f) - \Theta_q(\overline{f}) - \langle \nabla\Theta_q(\overline{f}), f - \overline{f} \rangle \, .$$

(a) Show that $\theta_q(x) = x\frac{x^q - m^q}{q}$, is strictly convex.

(b) Show that the family of Bregman divergence, parametrized by q, is

$$\mathcal{B}_{\Theta_q} = \frac{1}{q+1} \sum_p \left[f_p \frac{f_p^q - \overline{f}_p^q}{q} - \overline{f}_p^q (f_p - \overline{f}_p) \right] \, .$$

[17] The concept of convexity of functions is recalled in Exercise A.34, page 216.

(c) Derive the following family of regularization terms, parametrized by q

$$\begin{aligned} S(\mathbf{I}) &= \mathcal{B}_{\Theta_q}(\mathbf{I},\bar{\mathbf{I}}) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{M}\left[I_{ij}^q \frac{I_{ij}^q - \bar{I}_{ij}^q}{q} - \bar{I}_{ij}^q (I_{ij} - \bar{I}_{ij}) \right] . \end{aligned} \tag{4.28}$$

Here, each value of q yields different regularization terms such as the ones defined in Eq. (4.10). These regularization terms may be used in the last term of Eq. (4.9), [27].

Hint. Define $f_p = I_{ij}$ as the estimated value of the shade of gray for the image at the pixel (i,j), and $\bar{I}_{ij}$ its corresponding reference value. Observe that the term m^q in Eq. (4.27) cancels out in the derivation steps of Eq. (4.28).

(b) Setting $q = 1$, derive Eq. (4.10a), from the family of regularization terms given by Eq. (4.28).

(c) Considering the limit $q \to 0$, in Eq. (4.28), check that

$$\lim_{q\to 0} \mathcal{B}_{\Theta_q}(\mathbf{I},\bar{\mathbf{I}}) = S(\mathbf{I}) ,$$

where $S(\mathbf{I})$ is given by Eq. (4.10b).

Hint. Recall that $\lim_{q\to 0}(x^q - 1)/q = \ln x$.

4.13. Use the same notation as in Exercise 4.12.

(a) Show that $\theta_q(x) = \frac{x^q - m^q}{q}$, when $q > 1$, and $\theta_q(x) = -\frac{x^q - m^q}{q}$, when $0 < q < 1$, are strictly convex functions.

(b) Show that the family of Bregman divergence, parametrized by q, is

$$\mathcal{B}_{\Theta_q} = \sum_p \left[\frac{f_p^q - \overline{f}_p^q}{q} - \overline{f}_p^{q-1}(f_p - \overline{f}_p) \right] ,$$

when $q > 1$, and determine the corresponding expression when $0 < q < 1$.

(c) Derive the following family of regularization terms, parametrized by q

$$\begin{aligned} S(\mathbf{I}) &= \mathcal{B}_{\Theta_q}(\mathbf{I},\bar{\mathbf{I}}) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{M}\left[\frac{I_{ij}^q - \bar{I}_{ij}^q}{q} - \bar{I}_{ij}^{q-1}(I_{ij} - \bar{I}_{ij}) \right\} . \end{aligned} \tag{4.29}$$

when $q > 1$. Derive also the expression when $0 < q < 1$.

(b) Setting $q = 2$, derive Eq. (4.10a), from the family of regularization terms given by Eq. (4.29).

(c) Considering the limit $q \to 0$, in Eq. (4.29), check the relation between

$$\lim_{q\to 0} \mathcal{B}_{\Theta_q}(\mathbf{I},\bar{\mathbf{I}}) \ ,$$

and $S(\mathbf{I})$ given by Eq. (4.10b).

4.14. (a) Derive Eq. (4.14), from Eq. (4.11).

(b) Derive Eq. (4.21), from Eq. (4.14).

4.15. (a) Show that for a general value $q > 0$, Eq. (4.14) is written as

$$F_{rs} = -\sum_{i=1}^{M}\sum_{j=1}^{M}\left(Y_{ij} - \sum_{k=-N}^{N}\sum_{l=-N}^{N} b_{kl} I_{i+k,j+l}\right) b_{r-i,s-j} + \frac{\alpha}{q}\left(I_{rs}^{q} - \bar{I}_{rs}^{q}\right) \ . \tag{4.30}$$

(b) Considering the limit $q \to 0$, derive Eq. (4.14) from Eq. (4.30).

4.16. Show that for a general value $q > 0$ Eq. (4.21) is written as

$$C_{mn}^{rs} = \frac{\partial F_{rs}}{\partial I_{mn}} = \sum_{k=-N}^{N}\sum_{l=-N}^{N} b_{kl} b_{r-m+k,s-n+l} + \alpha I_{rs}^{q-1} \delta_{rm}\delta_{sn} \ .$$

4.17. Equation (4.24) states a proportionality that b_{kl} has to satisfy.

(a) Let c denote the constant of proportionality for a 3×3 blurring matrix. Determine it.

(b) Show that b_{kl} in Eq. (4.24) admits a separation of variables structure.

(c) Let c denote the constant of proportionality for a $N\times N$ blurring matrix. Obtain an expression for it.

Chapter 5
Radiative Transfer and Heat Conduction

We show in this chapter how to estimate the value of several material properties, such as single scattering albedo and thermal conductivity, present in the heating process of a material with thermal radiation.

The inverse problems considered here deal with the estimation of a finite number of parameters, related to an infinite dimensional model. These problems are solved as finite dimensional optimization problems, using the Levenberg-Marquadt method, which is a variant of Newton's method for non-linear systems of equations. In an intermediary step, this method includes a regularization, similar to Tikhonov's regularization (Section 3.4).

This chapter deals with inverse identification problems. According to Table 2.3, page 48, they are classified as Type III *inverse problems*.

5.1 Mathematical Description

Consider the example depicted in Fig. 5.1, of a material body subjected to heating due to thermal radiation. Assume we can regulate the intensity and the position of the heat sources and that we perform temperature measurements in some material body points, using temperature sensors[1].

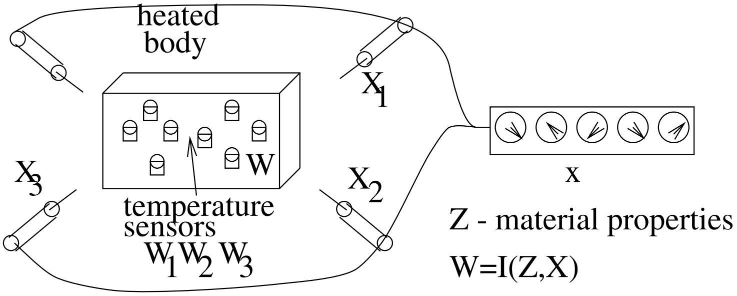


Fig. 5.1 Schematic representation of a participating medium (absorbent, emittant, and scatterer) being subject to radiation originated in external sources. Here **Z** represents material properties, **X** intensity of external sources of radiation and **W** = **I**(**Z**,**X**) measurements of temperature at several positions within the material body.

[1] In this example the measurements involve only temperature. Notwithstanding, problems involving thermal radiation measurements are discussed within this chapter, as well.

We denote by **W** a vector representing the output of this physical system, given by the resulting set of temperature measurements, comprising measurements in several specific points. By **X** we denote values representing levels and types of the regulation of the system (external sinks or sources), i.e. systems input, and by **Z** the physical/material properties of the body which influence the system's output.

The question is: given temperature measurements, **W**, and knowing the value of the regulation, **X**, obtain an estimate, **Z**, of the physical properties values.

In a general way, consider the question of obtaining an estimate, **Z**, for some constants, present in a mathematical model for a physical system's output. This output is represented by a physical magnitude, **W**, obtained from experimental measurements of the real output of the physical system.

The hypotheses used throughout this chapter are

$$\mathbf{Z} = (Z_1, \ldots, Z_N)^T \in \mathbb{R}^N \ , \ \ \mathbf{W} = (W_1, \ldots, W_M)^T \in \mathbb{R}^M \ ,$$

and **X** can be an element of $\mathbb{R}^K$, or even a function. We represent the dependence of **W** in **Z** and **X** in a functional form by $\mathbf{W} = \mathbf{I}(\mathbf{Z}, \mathbf{X})$.

The dependence of $\mathbf{I} = (I_1, \ldots, I_M)^T$ on **Z** can be explicit —as was the case considered in Chapter 1, when the relationship between the system inputs, **x**, and the system outputs, **y**, was given by a linear function, $\mathbb{R}^3 \ni \mathbf{x} \mapsto A\mathbf{x} \in \mathbb{R}^3$,— or implicit, if **W** and **Z** satisfy a given, possibly non-linear, system of equations, which would be written as

$$\mathbf{G}(\mathbf{W},\mathbf{Z},\mathbf{X}) = 0 \ ,$$

for some known function $\mathbf{G} : \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^K \to \mathbb{R}^M$.

Another possibility is when **Z** is a parameter of a differential or integro-differential equation, and **W** represents some *observation* of its solution (such as the value of the solution in some points of the domain, or an average of the solution for some part of the domain) [68, 81]. In this case, **X** can be the value of an initial or boundary condition, or a source. Unless we can find the solution explicitly, we would say that **I** is given implicitly (as the solution of the appropriate equation).

In the first two cases the inverse problem is of Type I, while in the last one the inverse problem is of Type III, in accordance with the classification given in Table 2.3, page 48. This is the type with which we will deal in this chapter.

More generally, the relation between **W**, **Z** and **X** implies a relation of cause-effect (stimulus-reaction), linear or non-linear, that can be explicit or implicit. Here **I** represents the *solution operator*—the abstract object that explicitly renders **W** as a function of **Z** and **X**. In practice, **I** may be impossible to obtain explicitly. Sometimes its existence, or even its uniqueness and smooth dependence on data, can be abstractly proven. That information may be insufficient for application purposes, and must be complemented by some numerical solution. Fortunately, the qualitative theoretical results can bring to light behaviour and properties of the algorithm for numerical solution.

Given material properties $\mathbf{Z}$ and parameters $\mathbf{X}$ the i-th *predicted* output of the system is $I_i(\mathbf{Z}, \mathbf{X})$. This is compared with the effectively measured quantity (experimental data), W_i, defining the *residual*,

$$R_i = R_i(\mathbf{Z},\mathbf{X}) = I_i(\mathbf{Z},\mathbf{X}) - W_i \,. \tag{5.1}$$

Here, $\mathbf{Z} = (Z_1, \ldots, Z_N)^T \in \mathbb{R}^N$ is the vector of unknowns of the problem.

The inverse problem is solved as an optimization problem, on a space of finite dimension, in which we pursue the minimization of the functional representing half the sum of the squared residuals,

$$\begin{aligned} Q = Q(\mathbf{Z}) &= \frac{1}{2}|\mathbf{R}|^2 = \frac{1}{2}\,\mathbf{R}^T\mathbf{R} = \frac{1}{2}|\mathbf{I}(\mathbf{Z}) - \mathbf{W}|^2 \\ &= \frac{1}{2}\sum_{i=1}^{M} [I_i(\mathbf{Z}) - W_i])^2 \,, \end{aligned} \tag{5.2}$$

where $\mathbf{R} = (R_1, \ldots, R_M)^T \in \mathbb{R}^M$ represents the residual between computed magnitudes $\mathbf{I}$ and measurements (experimental data) $\mathbf{W}$, and M is the total number of experimental data available.

This formulation is similar to the minimization problems presented in Sections 2.6 and 3.4, and is an instance of the *least squares method*.

5.2 Modified Newton's Method

The functional given by Eq. (5.2), is minimized by finding its critical point, $\nabla Q = 0$, that is,

$$\frac{\partial Q}{\partial Z_k} = 0 \quad \text{for } k = 1, 2, \ldots, N \,, \tag{5.3}$$

which constitutes a system of N non-linear equations and N unknowns, $\mathbf{Z} = (Z_1, \ldots, Z_N)$. From Eqs. (5.1) to (5.3), the *critical point equation* is rewritten as

$$\sum_{i=1}^{M} R_i \frac{\partial I_i}{\partial Z_k} = 0 \quad \text{for } k = 1, 2, \ldots, N \,. \tag{5.4}$$

We solve Eq. (5.4) by means of a modified *Newton's method*, that can be deduced as follows. Using a Taylor's expansion of $\mathbf{R}$ around $\mathbf{Z}^n$, where n will be the index of the iterations in the iterative procedure, and keeping only the zero and first order terms, we have

$$R_i^{n+1} = R_i^n + \sum_{j=1}^{N} \frac{\partial R_i^n}{\partial Z_j}\Delta Z_j^n \,, \quad \text{for } i = 1, 2, \ldots, M \,. \tag{5.5}$$

Here, $\mathbf{R}^n$ represents the evaluation of $\mathbf{R}$ in $\mathbf{Z}^n$,

$$\mathbf{R}^n \quad = \quad \mathbf{R}(\mathbf{Z}^n) \;=\; \mathbf{I}(\mathbf{Z}^n) - \mathbf{W} \,, \tag{5.6}$$

and,

$$\Delta\mathbf{Z}^n = \mathbf{Z}^{n+1} - \mathbf{Z}^n ,$$

or, in coordinates, $R_i^n = R_i(\mathbf{Z}^n)$ and $\Delta Z_j^n = Z_j^{n+1} - Z_j^n$.

Using Eq. (5.5) in the system of equations (5.4), and noticing that $\partial R_i/\partial Z_j = \partial I_i/\partial Z_j$, we obtain,

$$\sum_{i=1}^{M}\left(R_i^n + \sum_{j=1}^{N} \left.\frac{\partial I_i}{\partial Z_j}\right|_{Z=Z^n} \Delta Z_j^n\right) \left.\frac{\partial I_i}{\partial Z_k}\right|_{Z=Z^n} = 0 , \tag{5.7}$$

for $k = 1, 2, \ldots, N$.

We look at $\mathbf{I}$ as a function of $\mathbf{Z}$ only, making $\mathbf{X}$ constant. By definition, the *Jacobian matrix* of $\mathbf{I}$ with respect to $\mathbf{Z}$, $\mathcal{J} = \mathcal{J}\mathbf{I}|_Z$, has entries

$$\mathcal{J}_{ij} = \partial I_i/\partial Z_j ,$$

for $i = 1, \ldots, M$, and $j = 1, \ldots, N$, that is,

$$\mathcal{J} = \begin{pmatrix} \frac{\partial I_1}{\partial Z_1} & \cdots & \frac{\partial I_1}{\partial Z_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial I_M}{\partial Z_1} & \cdots & \frac{\partial I_M}{\partial Z_N} \end{pmatrix} .$$

The system of equations (5.7) can, therefore, be rewritten in the form known as *normal equation*

$$(\mathcal{J}^n)^T \mathcal{J}^n \Delta\mathbf{Z}^n = -(\mathcal{J}^n)^T \mathbf{R}^n , \tag{5.8}$$

where $\mathcal{J}^n$ represents $\mathcal{J}\mathbf{I}|_{Z=Z^n}$.

An iterative procedure can be constructed to determine the vector of unknowns $\mathbf{Z}$ that minimizes the functional Q, knowing the experimental data, $\mathbf{W}$, and computed values, $\mathbf{I}$, which depend on the unknowns to be determined, $\mathbf{Z}$.

Starting from an initial estimate, $\mathbf{Z}^0$ and measurements $\mathbf{W}$, residuals are computed from Eq. (5.6), and corrections are computed sequentially from Eq. (5.8), where n is the iteration counter. The algorithm can be written as[2],

$$\mathbf{R}^n = \mathbf{I}(\mathbf{Z}^n, \mathbf{X}) - \mathbf{W} \tag{5.9a}$$

$$(\mathcal{J}^n)^T \mathcal{J}^n \Delta\mathbf{Z}^n = -(\mathcal{J}^n)^T \mathbf{R}^n \tag{5.9b}$$

$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \Delta\mathbf{Z}^n . \tag{5.9c}$$

[2] The method described here is a modification of Newton's method (presented in Section 4.3, page 92). We remark that the goal is not Newton's method, but the solution of Eq. (5.4), that is obtained here by means of the modified Newton's method, given by Eq. (5.9). For the problem treated here Newton's method demands the computation of second order derivatives of $\mathbf{I}$, while the method based on Eq. (5.8) avoids that. See Exercise 5.1.

The iterative procedure is interrupted when a convergence criterion defined *a priori* is satisfied. For example,

$$\left|\Delta Z_j^n / Z_j^n\right| < \varepsilon\,, \text{ for all } j \text{ with } j = 1, 2, \ldots, N\,.$$

Here, ε is a sufficiently small value, say 10^{-5}. Another possibility is the use of the vector of corrections norm, $|\Delta \mathbf{Z}^n| < \epsilon$ as considered in Newton's method.

Observe that in every iteration the values $I_i^n = I_i(\mathbf{Z}^n)$, $i = 1, 2, \ldots, M$, are computed using the estimates for the unknowns, $\mathbf{Z}^n$. This involves, in the examples discussed later in this chapter, the solution of differential or integro-differential equations. This makes them implicit inverse problems.

Finally, we remark that the solution of Eq. (5.9b) can be written as

$$\Delta \mathbf{Z}^n = -\left[(\mathcal{J}^n)^T \mathcal{J}^n\right]^{-1} (\mathcal{J}^n)^T \mathbf{R}^n \tag{5.10}$$

Equation 5.10 represents explicitly the solution of Eq. (5.8), making use of the inverse[3] of $(\mathcal{J}^n)^T \mathcal{J}^n$.

5.3 Levenberg-Marquardt's Method

The methods presented in Section 4.3 and the algorithm presented in the previous section, are Newton-like, and can encounter convergence difficulties if the initial estimate for the vector of unknowns, $\mathbf{Z}^0$, is not adequately selected. Choosing an adequate initial estimate can prove extremely difficult.

In 1963, Marquardt [54] designed the algorithm that will be described presently, with the objective of reaching convergence with a wider range of initial estimates. One of the referees of his work noticed that, in 1944, Levenberg had made a similar proposal: adding a term in the diagonal of matrix $\mathcal{J}^T \mathcal{J}$. The method came to be known as the *Levenberg-Marquardt method*.

Based on Eq. (5.8), the Levenberg-Marquardt method considers the determination of the corrections $\Delta \mathbf{Z}^n$ by means of the following equation,

$$\left[(\mathcal{J}^n)^T \mathcal{J}^n + \lambda^n \mathcal{I}\right] \Delta \mathbf{Z}^n = -(\mathcal{J}^n)^T \mathbf{R}^n\,. \tag{5.11}$$

Here $\lambda = \lambda^n$ is the *damping factor* and $\mathcal{I}$ represents the identity matrix. Observe that this formulation is similar to the Tikhonov's regularization, Eq. (3.22).

Similar to the developments in the previous section, an iterative algorithm is built to determine the vector of unknowns $\mathbf{Z}$ that should minimize the functional Q. The procedure is based on Eq. (5.11). From an initial estimate, $\mathbf{Z}^0$, corrections are sequentially computed,

$$\Delta \mathbf{Z}^n = -\left[(\mathcal{J}^n)^T \mathcal{J}^n + \lambda^n \mathcal{I}\right]^{-1} (\mathcal{J}^n)^T\, \mathbf{R}^n\,, \text{ for } n = 0, 1, \ldots, \tag{5.12}$$

[3] In computations, one rarely inverts a matrix due to its high computational cost. It is preferable to solve the system of equations. Therefore, the corrections $\Delta \mathbf{Z}^n$ are computed by solving the linear algebraic system of equations (5.8). For theoretical considerations it is sometimes useful to have closed form solutions, that is, to have the solution written in terms of a solution operator.

where n is an iteration counter, and the new estimates for the vector of unknowns are computed by Eq. (5.9c). The iterations are interrupted when a convergence criterion established *a priori* is satisfied.

It should be noticed that the solution of the problem described by Eq. (5.11) differs from the one given by Eq. (5.8). On the other hand, our aim is to solve Eq. (5.4). To guarantee the convergence, at the beginning of the iterative process, a relatively high value is assigned to λ, λ^0, thus emphasizing the importance of the diagonal of matrix $(\mathcal{J}^T\mathcal{J} + \lambda\mathcal{I})$ relative to the information contained in the elements outside the diagonal. Through the iterative procedure, the value of the damping factor λ is to be reduced, in such a way that its value approaches zero as the procedure approaches its conclusion. In the light of the notion of regularization, described in Chapter 3, Eq. (5.11) is a kind of regularization of Eq. (5.8).

An algorithm to control the value of the damping factor will be described shortly, [54]. Let $c > 1$, $d > 1$ and $\xi > 1$. Let also $Q^n = Q(\mathbf{Z}^n)$. When $Q^{n+1} \leq Q^n/c$, the reduction $\lambda^{n+1} = \lambda^n/d$ is performed. Otherwise, $\lambda^n = \xi\lambda^{n-1}$ is taken, and a new estimate for the vector of unknowns $\mathbf{Z}^{n+1}$ is computed for the same value of the iteration counter n, using again the previous estimate $\mathbf{Z}^n$. Silva Neto and Özişik, [74, 75, 79], used $c = d = \xi = 2$ in the applications in heat transfer by means of thermal radiation.

5.4 Confidence Intervals for Parameter Estimates

Folowing Gallant [34, 38, 59], the *confidence intervals* of the estimates of the parameters $\mathbf{Z}$ are computed using the *sensitivity coefficients*,

$$\partial I_i/\partial Z_j\,, i = 1, \ldots, M \text{ and } j = 1, \ldots, N\,,$$

and the *standard deviation*, σ, of the error present in the experimental data.

Let $\nabla\mathbf{I}$ be the $M\times N$ matrix whose entries are given by the sensitivity coefficients,

$$(\nabla\mathbf{I})_{ij} = \left.\frac{\partial I_i}{\partial Z_j}\right|_Z \,.$$

In this case, the square of the *standard deviation* of the estimators of the parameters are given by

$$\sigma_Z^2 = \left(\sigma_{Z_1}^2, \ldots, \sigma_{Z_N}^2\right)^T = \sigma^2\left\{\text{diag}\left[(\nabla\mathbf{I})^T\,\nabla\mathbf{I}\right]^{-1}\right\} \tag{5.13a}$$

where diag(A) represents the vector whose elements are the elements of the diagonal of a matrix A.

Assuming a normal distribution for the experimental errors, with zero mean, the 99 % confidence intervals for the estimates Z_j are [33]

$$\left]Z_j - 2.576\sigma_{Z_j},\, Z_j + 2.576\sigma_{Z_j}\right[\qquad \text{for } j = 1, \ldots, N\,. \tag{5.13b}$$

In general, smaller confidence intervals are associated with larger sensitivity coefficients and smaller experimental errors, thus producing better estimates.

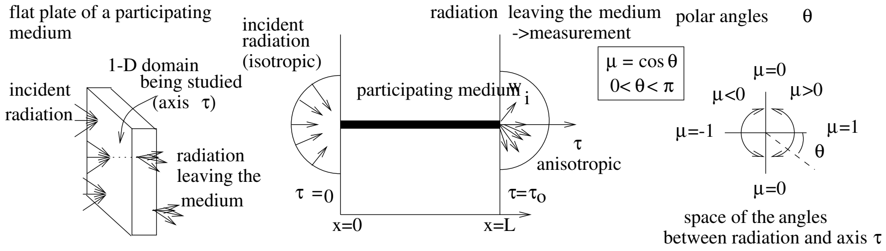

flat plate of a participating medium
incident radiation
1-D domain being studied (axis τ)
radiation leaving the medium
incident radiation (isotropic)
participating medium
W i
τ =0
τ=τo
x=0
x=L
radiation leaving the medium ->measurement
τ
anisotropic
μ = cos θ
0< θ< π
polar angles θ
μ=0
μ<0
μ>0
μ=-1
μ=1
θ
μ=0
space of the angles between radiation and axis τ

Fig. 5.2 Radiative transfer in a participating medium. Optical variable τ is related (sometimes linearly) to the spatial variable x.

5.5 Phase Function, Albedo and Optical Thickness

Different types of radiation, like neutral particles, gamma rays and photons have been used to identify objects in industry (through non-destructive tests) and also in medicine (for diagnosis and therapy).

Heat transfer by thermal radiation in a *participating medium*, that is, one that emits, absorbs and scatters radiation, schematically represented in Figs. 5.1 and 5.2, is modeled according to the linear version of the Boltzmann equation [60, 80, 69, 57].

It is worthwhile to mention that the physical phenomena relevant to neutron transport in nuclear reactors, or to tomography with scattering (NIROT—*Near Infrared Optical Tomography*) can also be represented mathematically by the linear Boltzmann equation.

Consider the situation depicted in Fig. 5.2 representing a flat plate made of a scattering anisotropic, gray material with transparent boundary surfaces, subject to external isotropic radiation on its left surface, in a permanent regimen (steady-state — it does not depend on time).

A material is *anisotropically* scattering when the scattering depends on angle and it is *gray* if the properties do not depend on the radiation's wavelength. A material has a *transparent* surface when this surface does not reflect radiation.

In this case, and also considering azimuthal symmetry and a cold medium (no emission), the linear *Boltzmann equation* is written as [60] (see Fig. 5.2)

$$\mu \frac{\partial I}{\partial \tau}(\tau,\mu) + I(\tau,\mu) = \frac{\omega}{2}\int_{-1}^{1} p(\mu,\mu')I(\tau,\mu')d\mu' \ , \tag{5.14a}$$

in $0<\tau<\tau_0$, $-1\leq\mu\leq 1$, and

$$I(0,\mu) = 1, \quad \mu > 0, \text{ and } I(\tau_0,\mu) = 0, \text{ for } \mu < 0 \ . \tag{5.14b}$$

In this equation, $I(\tau,\mu)$ is the radiation intensity in position τ, following direction represented by μ. Here, τ is the spatial optical variable, and μ is the cosine of the polar angle θ formed between the direction of the radiation beam and the τ axis. Also, ω is the single scattering *albedo* (the ratio between the scattering and the extinction coefficients, σ_s and β, respectively, with $\beta = \sigma_s + k_a$, and k_a is the absorption coefficient), τ_0 is the medium's *optical thickness* (related to the geometrical thickness of the medium) and $p(\mu,\mu')$ is the *anisotropic scattering phase* function.

We remark that $\frac{1}{2}p(\mu,\mu')$ represents the probability density of an incident beam with direction μ' to be scattered following the direction represented by μ. More explicitly, the probability that the scattered direction μ is between μ_1 and μ_2 given that the incident direction is μ' is given in terms of $p(\mu,\mu')$ by

$$P(\mu_1 \leq \mu \leq \mu_2 \mid \mu') = \frac{1}{2}\int_{\mu_1}^{\mu_2} p(\mu,\mu')\, d\mu \,.$$

The medium is *isotropic* when the scattering is uniform in all directions, i.e., when $p(\mu,\mu') = c$, for all $\mu,\mu' \in [-1,1]$, where c is a constant value. Since

$$P(-1 \leq \mu \leq 1 \mid \mu') = 1 \,, \tag{5.15}$$

it is then necessary that $c = 1$.

The second term in the left hand side of Eq. (5.14a) represents the absorption and scattering of radiation by the medium away from the direction represented by μ (out scattering) and the right hand side represents the way the radiation is scattered by the medium into such direction (in scattering). The emission of radiation by the medium can be neglected if compared to the incident radiation in $\tau = 0$. We recall that we are considering here a *steady state* problem (it does not depend on time).

When the operator, the medium's geometry (in this case, the optical thickness τ_0 for the plane-parallel medium), the material properties (here, ω and $p = p(\mu,\mu')$) and the boundary conditions (given in this example by Eq. (5.14b)) we say that the model is characterized. That is, it is modeled by steady state linear Boltzmann equation, with specific type Dirichlet boundary conditions and identified (all constants and auxiliary functions are given). In this case we deal with a direct problem, and the radiation intensity $I(\tau,\mu)$ can be computed in all of the spatial domain $0 \leq \tau \leq \tau_o$ and the angular domain $-1 \leq \mu \leq 1$.

Different analytic and numerical techniques have been developed to solve the linear radiative *transport* equation, Eq. (5.14). Wick [93] and Chandrasekhar [20, 21] created the *discrete-ordinates method*, by replacing the right-side term of Eq. (5.14a) by a Gaussian quadrature term. This leads to a system of ordinary differential equations with as many equations as points used in the quadrature. Silva Neto and Roberty [80] presented a comparison between spherical harmonics expansion methods, P_N, Galerkin, global base and discrete-ordinates (i.e., finite differences + Gaussian quadrature), S_N, for the case of isotropic scattering. Chalhout et al. [19] considered three variations of the discrete ordinates method and performed a comparison with the Monte Carlo method. Moura Neto and Silva Neto [57, 63] presented solutions using methods with integrating factor and operator splitting.

We use the customary representation of the scattering phase function by expansion in Legendre polynomials [60, 79, 47],

$$p(\mu,\mu') = \sum_{l=0}^{L} (2l+1)\, f_l P_l(\mu) P_l(\mu') \,, \tag{5.16}$$

where f_l, $l=0,1,\ldots,L$ are the coefficients of expansion with $f_0=1$.

In the example presented here, we consider the inverse problem of estimating simultaneously the medium's optical thickness, τ_0, the single scattering albedo, ω, and the scattering phase function, $p(\mu,\mu')$, by means of its coefficients, f_l, $l = 1, 2, \ldots, L$, in the expansion in Legendre polynomials, Eq. (5.16) [79, 47].

The vector of unknowns is made up of the following elements

$$\mathbf{Z} = (\tau_0, \omega, f_1, \ldots, f_L)^T \ ,$$

to be determined using the experimental measurements of the radiation that leaves the medium, W_i, $i = 1, 2, \ldots, M$, by minimizing the functional

$$\begin{aligned} Q &= Q(\tau_0, \omega, f_1, \ldots, f_L) \\ &= \frac{1}{2}|\mathbf{R}|^2 = \frac{1}{2}\sum_{i=1}^{M} [I_i[\mathbf{Z}] - W_i]^2 \ . \end{aligned}$$

Here $I_i[\mathbf{Z}]$, $i = 1,2,\ldots,M$ are the values of the radiation intensities computed by the solution of the direct problem described by Eq. (5.14) and evaluated in the same directions in which the radiation leaving the medium is measured, μ_i, $i = 1,2,\ldots,M$ which are pre-defined directions, assuming the parameters $\mathbf{Z} = (\tau_0, \omega, f_1, f_2, \ldots, f_L)$ are known,

$$I_i[\mathbf{Z}] = I[\tau_0, \omega, f_1, f_2, \ldots, f_L](\mu_i) \ , \ \text{for } i = 1, 2, \ldots, M \ .$$

The Levenberg-Marquardt method, described in Section 5.3, is used to solve this optimization problem in finite dimension.

Since we do not have experimental data on this problem, we use *synthetic* or artificial data. For that we mean data generated from solving the direct problem with known parameters and adding random values to simulate experimental errors. In the example we are considering, we assume that the parameters $\mathbf{Z} = (\tau_0, \omega, f_1, \ldots, f_L)^T$ are known in advance and we use them to solve the direct problem, Eq. (5.14). The synthetic data is then determined by

$$W_i = I_i[\tau_0, \omega, f_1, \ldots, f_L](\mu_i) + c\epsilon_i \ ,$$

where ϵ_i is a realization of a uniform random variable in the interval $[-1,1]$,

$$c = \gamma \max_i I(\tau_0,\mu_i), \ \text{for } i = 1, \ldots, M \ ,$$

and γ is a maximum percentage error parameter.

It should be remarked that the solution of inverse problems with synthetic data makes it possible to verify that the computational procedure is correct, before applying it to real problems.

The value of the phase function is related to the angle subtended between the directions of the incident radiation and the scattered radiation [22]. The results presented in Figs. 5.3 and 5.4 where computed assuming the incident radiation to be exclusively in the direction of vector $(1, 0)$. A representation in polar coordinates is

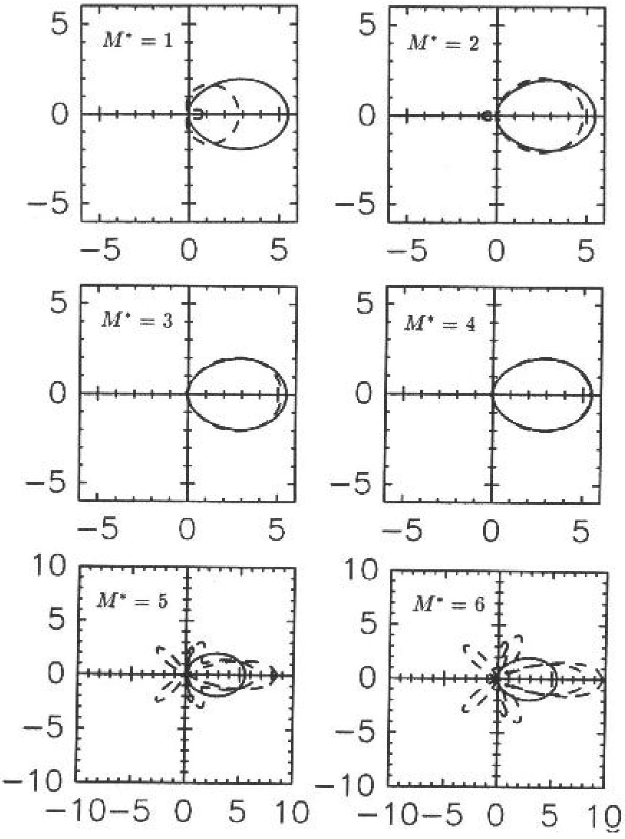


Fig. 5.3 Estimation of an anisotropic scattering phase function, with $L = 7$, characterizing forward scattering, $\omega = 0.5$ and $\tau_0 = 2.0$. For forward scattering, the phase function representation is restricted to the 1^{st} and 4^{th} quadrant. The experimental error reached 6.2 % of the highest intensity value that was measured. Here, M^* is the number of coefficients of the phase function that were considered in the estimations. Solid line represents the exact phase function, and dashed line represents the estimates for the phase function obtained with different values for M^*. The best result occurs with $M^* = 4$.

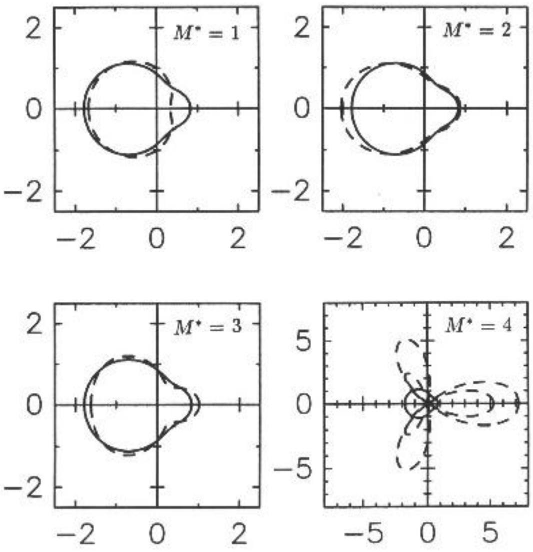


Fig. 5.4 Estimation of a phase function with $L = 5$, with preferential backward scattering, $\omega = 0.1$ and $\tau_0 = 10.0$. In a preferred backward scattering, one should have $\int_{-1}^{0} p(\mu,1)\,d\mu > \int_{0}^{1} p(\mu,1)\,d\mu$, which is the case if $p(\mu,1) < p(-\mu,1)$, for $\mu > 0$, as shown in the graph of the phase function. The experimental error reached 4.1 % of the highest intensity measured value. The number of coefficients of the phase function, M^*, were chosen as 1 to 4. Solid line represents the exact phase function, and dashed line represents the phase function obtained estimate with different values for M^*. Fairly good results were obtained with $M^* = 2$ and $M^* = 3$.

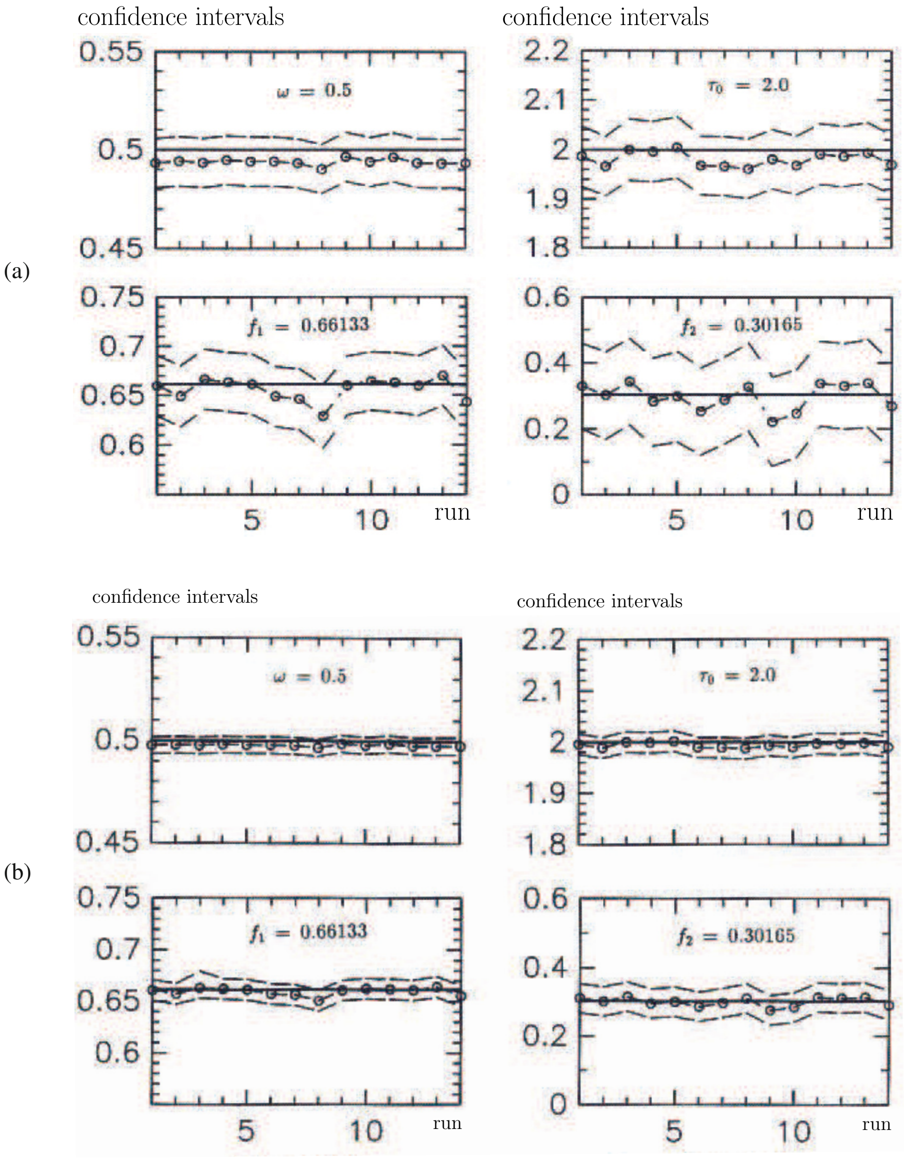


Fig. 5.5 Confidence intervals for the single scattering albedo (ω), optical thickness (τ_0), and first two coefficients of the anisotropic scattering phase function expansion (f_1 and f_2) estimates when the phase function has $L = 7$ terms.
– exact values, - - - confidence intervals, -•-•- estimates. Experimental error in the highest mean value of the measured intensities: (a) 6.2 %, and (b) 2.1 %.

given in which the distance from the point on the graph to the origin provides the probability density for the radiation to be scattered according to that direction.

Figure 5.3 represents the phase function estimation of an anisotropic scattering, with preferential forward scattering with $L = 7$, (see Eq. (5.16)). Due to the experimental error occurring in the measurements, **W**, of γ =6.2 % of the highest value of the measured intensity (where, 6.2 % is, therefore, the lowest percentage of experimental error in the measurements being considered), it is not possible to recover all the coefficients. The highest-order coefficients, with relatively low numerical values, are more rapidly affected, causing the estimation to deteriorate. However, the relevant information for the design of thermal equipments is the shape of the phase function, and not necessarily the value of each coefficient separately.

As a matter of fact, the number of coefficients, L, of the anisotropic scattering phase function represented in Eq. (5.16) is also an unknown of the problem. Silva Neto and Özişik [79] developed a criterion to choose the number of coefficients in the expansion M^* in such a way to obtain the best possible estimated phase function, considering the available experimental data.

In the test case represented in Fig. 5.3, $M^* = 3$ or 4 values would be chosen. It must be pointed out that, in the absence of experimental errors (an ideal situation that does not happen in practice) all seven coefficients, i.e., $M^* = L = 7$, were estimated within the precision established *a priori* in the stopping criterion.

Figures 5.5a and b show the estimates for ω, τ_0, and for the first two coefficients, f_1 and f_2, of the phase function represented in Fig. 5.3. The different executions of the computational code correspond to estimates due to different sets of experimental data. Figure 5.5a presents the results when the lowest experimental error reaches 6.2 % and, in Fig. 5.5b, the lowest experimental error reaches 2.1 %.

Figure 5.4 presents the results on the estimation of a phase function with $L = 5$, corresponding to a medium with preferential backward scattering. The lowest experimental error here considered is 4.1 %. In this example we would choose $M^* = 2$ or 3.

Using the expression for the confidence intervals, Eq. (5.13), we are able to compute them for the various parameters being estimated, and we represent them in graphical form in Fig. 5.5. As expected, the estimates, in the examples given here, have narrower confidence intervals, when the experimental data presents lower levels of noise.

5.6 Thermal Conductivity, Optical Thickness and Albedo

Silva Neto and Özişik [74] solved an inverse problem involving heat transfer due to conduction and radiation in a *participating medium*, considering only isotropic scattering (not depending on the polar angle). Lobato et al. [48] dealt with a similar problem, but using stochastic methods [69] for the minimization of the squared residues functional.

Consider the situation illustrated in Fig. 5.6. A plane-parallel, gray, isotropically scattering medium with transparent boundary surfaces, is subject to incident external

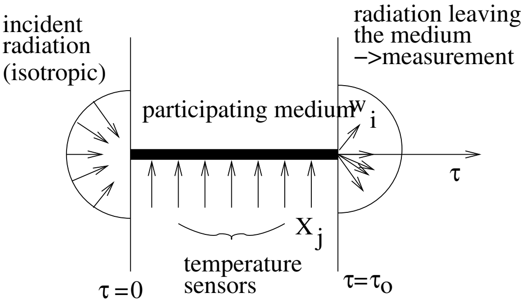


Fig. 5.6 Heat transfer by thermal conduction and radiation in a participating medium

isotropic radiation that reaches surface $\tau = 0$. The surfaces at $\tau = 0$ and $\tau = \tau_0$ are kept at constant temperatures T_1 and T_2, respectively.

The mathematical formulation of the heat transfer problem due to one dimensional, steady-state heat conduction, in which the coupling with the transfer due to radiation in the participating medium is achieved by the source term, is given, in a dimensionless formulation, by a boundary value problem of an ordinary differential equation (*Poisson's equation*, with a non-linear source term and Dirichlet boundary conditions) [60]

$$\frac{d^2\Theta}{d\tau^2} - \frac{(1-\omega)}{N}\left[\Theta^4(\tau) - G^*[I](\tau)\right] = 0\,, \quad \text{in } 0 < \tau < \tau_0\,, \tag{5.17a}$$

with boundary conditions,

$$\Theta(0) = 1 \text{ and } \Theta(\tau_0) = T_2/T_1\,. \tag{5.17b}$$

Here

$$G^*[I](\tau) = \frac{1}{2}\int_{-1}^{1} I(\tau,\mu)\,d\mu\,, \quad N = \frac{k\beta}{4\,n^2\,\overline{\sigma}\,T_1^3}\,, \quad \Theta = \frac{T}{T_1}\,, \tag{5.17c}$$

where Θ is the dimensionless temperature, N is the conduction-radiation parameter, ω is the simple scattering albedo, k is the thermal conductivity, β is the extinction coefficient (absorption + scattering), n is the medium's refractive index, and $\overline{\sigma}$ is the Stefan-Boltzmann's constant.

Modeling of the radiative transfer in the participating medium is achieved by means of the *linear* Boltzmann's equation [60],

$$\mu\,\frac{\partial I}{\partial\tau}(\tau,\mu) + I(\tau,\mu) = H(\Theta(\tau)) + \frac{\omega}{2}\int_{-1}^{1} I(\tau,\mu')\,d\mu'\,, \tag{5.17d}$$

in $0<\tau<\tau_0$ and $-1\leq\mu\leq 1$, and

$$I(0,\mu) = 1\,, \text{ for } \mu > 0, \text{ and } I(\tau_0,\mu) = 0\,, \text{ for } \mu < 0\,, \tag{5.17e}$$

where the source term, $H(\Theta)$, is related to the medium's temperature distribution,

$$H(\Theta) \quad = \quad (1-\omega)\,\Theta^4 \,, \tag{5.17f}$$

and the remaining symbols have already been defined. We observe that since the medium is an isotropic scatterer, the phase function of scattering is

$$\frac{1}{2}p(\mu,\mu') = \frac{1}{2} \text{ for all } \mu,\ \mu' \,.$$

Equation (5.17) provides a complete mathematical formulation for the one dimensional heat transfer problem, in steady state regime, by the combined mode of conduction and radiation. The problem of conduction, Eqs. (5.17a)–(5.17c), and the radiation problem, Eqs. (5.17d)–(5.17f), are coupled by means of the source terms, given respectively by

$$G^* = G^*[I] \text{ and } H = H(\Theta)\,.$$

To solve Eq. (5.17), we use an iterative procedure. Starting with a first estimate of I, we solve Eqs. (5.17a)–(5.17c) to obtain an estimate of Θ. From this estimate of Θ, we solve Eqs. (5.17d)–(5.17f) to obtain a new estimate of I. This is done until convergence is reached.

In the solution of the direct problem, Silva Neto and Özişik [74] used the iterative procedure described with the Galerkin method, global basis for the part of the problem related to heat transfer due to radiation in a participating medium, Eqs. (5.17d)–(5.17f), and the finite difference method for the part of the problem related to heat transfer by conduction, Eqs. (5.17a)–(5.17c).

In the inverse problem just presented, we consider the simultaneous estimation of the optical thickness, τ_0, the single scattering albedo, ω, and the conduction-radiation parameter, N. We use synthetic experimental measurements of the radiation, W_i, $i = 1, 2, \ldots, M$, and of temperature inside the medium, represented by X_j, $j = 1, 2, \ldots, K$. The vector of unknowns $\mathbf{Z} = (\tau_0, \omega, N)^T$ is determined — the model is identified — as the minimum point by minimization of the functional

$$\begin{aligned} Q = Q(\tau_0,\omega,N) &= \frac{1}{2}\sum_{i=1}^{M} \left[I_i\,(\tau_0,\omega,N) - W_i\right]^2 \\ &+ \frac{1}{2}\sum_{j=1}^{K} \left[\Theta_j\,(\tau_0,\omega,N) - X_j\right]^2 \,, \end{aligned} \tag{5.18}$$

where $I_i(\tau_0,\omega,N)$, $i = 1, 2, \ldots, M$, are the radiation intensities computed in the same surface and in the same directions in which the radiation is measured, W_i, $i = 1, 2, \ldots, M$. Here, $\Theta_j(\tau_0,\omega,N)$, $j = 1, 2, \ldots, K$, are temperatures computed in the same positions where the temperatures are measured, X_j, $j = 1, 2, \ldots, K$. The radiation and temperature intensities are computed by solving Eq. (5.17), following the procedure already described.

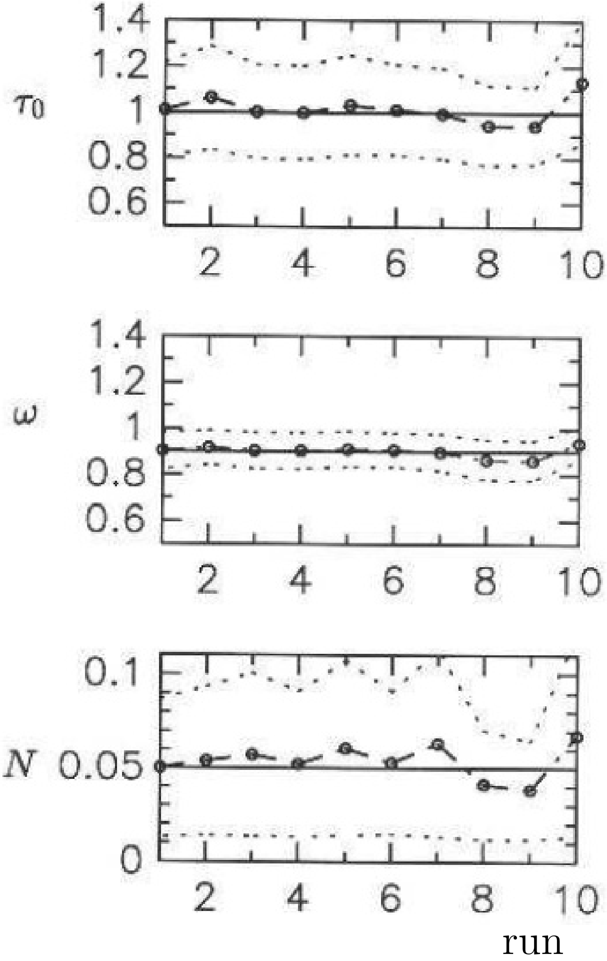


Fig. 5.7 Confidence intervals for the optical thickness (τ_0), single scattering albedo (ω), and conduction-radiation parameter (N) estimates in the combined conduction-radiation heat transfer model.
– exact values, - - - confidence intervals, -•-•- estimates. Experimental error of 4 % of the largest value of the magnitudes that were measured.

Figure 5.7 presents the results of the parameters (τ_0, ω, N) estimation, for a test case in which the exact values,

$$(\tau_0, \omega, N) = (1.0,\ 0.9,\ 0.05)\,,$$

were known *a priori*. The different results considered represent, as before, estimates, obtained with the execution of the computational code with different sets of synthetic experimental data. The artificially generated experimental error is set to 4 % of the value of the largest measured magnitude.

Conduction-radiation parameter, N, is relatively small, which indicates a dominance of the radiative heat transfer mechanism. This fact is proved by the relatively large size of confidence intervals[4] of the estimates of N (N depends on the medium's thermal conductivity, k), when compared to those obtained for the parameters τ_0 and ω.

5.7 Refractive Index and Optical Thickness

Consider a gray flat plate in *radiative equilibrium*, with two gray boundary surfaces, *opaque* —diffuse emittant and reflector (non-specular),— with emissivity ε and reflectivity ρ, which are kept at constant temperatures T_0 and T_L (see Fig. 5.8).

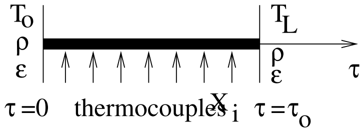


Fig. 5.8 Schematic representation of a one dimensional medium in radiative thermal equilibrium

The temperature distribution inside the medium, $T(\tau)$ satisfies [75],

$$\frac{T^4(\tau) - T_L^4}{T_0^4 - T_L^4} = \frac{\theta(\tau) + \left[\frac{1}{\varepsilon} - 1\right] S}{1 + 2\left[\frac{1}{\varepsilon} - 1\right] S}\,, \tag{5.19}$$

where the function $\theta(\tau)$ satisfies the *integral equation*

$$\theta(\tau) = \frac{1}{2}\left[E_2(\tau) + \int_0^{\tau_0} \theta(\tau')\, E_1(|\tau - \tau'|)\, d\tau'\right]. \tag{5.20}$$

Here, $E_m(\tau)$ represents the m-th *integral exponential function*, given by

$$E_m(\tau) = \int_0^1 \eta^{m-2}\, e^{-\frac{\tau}{\eta}}\, d\eta\,, \tag{5.21a}$$

[4] The confidence intervals are related to the sensitivity coefficients $\partial I/\partial\tau_0$, $\partial I/\partial\omega$, $\partial I/\partial N$, $\partial\Theta/\partial\tau_0$, $\partial\Theta/\partial\omega$, and $\partial\Theta/\partial N$.

and the constant S is given by

$$S = 1 - 2 \int_0^{\tau_0} \theta(\tau')\, E_2(\tau')\, d\tau' \,. \tag{5.21b}$$

Given the optical thickness of the medium, τ_0, function $\theta(\tau)$ is computed by Eq. (5.20). If $\theta(\tau)$ is known, parameter S is computed using Eq. (5.21b). The temperature distribution inside the medium is then computed by Eq. (5.19).

For opaque and gray surfaces, $\varepsilon = 1 - \rho$. When the refractive index of the medium, n_{m}, is higher than that of its environment, n_{e}, the reflectivity ρ is related to the relative refractive index $n = n_{\mathrm{m}}/n_{\mathrm{e}}$ as follows,

$$\begin{aligned}\rho(n) = 1 - \frac{1}{n^2}\Bigg\{\frac{1}{2} - \frac{(3n+1)(n-1)}{6(n+1)^2} - \frac{n^2(n^2-1)^2}{(n^2+1)^3}\ln\left(\frac{n-1}{n+1}\right) \\ + \frac{2n^3(n^2+2n-1)}{(n^2+1)(n^4-1)} - \frac{8n^4(n^4+1)}{(n^2+1)(n^4-1)^2}\ln(n)\Bigg\} \,.\end{aligned} \tag{5.22}$$

In the inverse problem presented here, the simultaneous estimation of the relative refractive index, n, and the medium's optical thickness, τ_0, is considered. The vector of unknowns $\mathbf{Z} = (n, \tau_0)^T$, is to be determined from the experimental measurements inside the medium, X_i, $i = 1, 2, \ldots, M$, by minimizing the functional

$$Q(n, \tau_0) = \frac{1}{2}|\mathbf{R}|^2 = \frac{1}{2}\sum_{i=1}^{M}\left[T_i\,(n, \tau_0) - X_i\right]^2 \,,$$

where $T_i(n, \tau_0)$, $i = 1, 2, \ldots, M$ are the temperatures computed in the same positions in which the experimental data are measured, X_i, using the problem described by Eqs. (5.19) through (5.22).

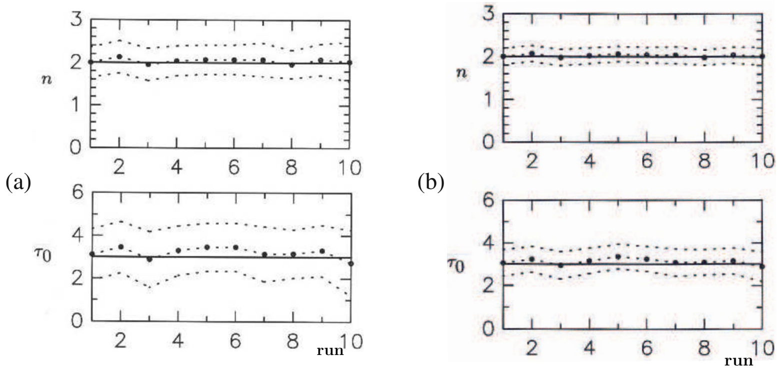


Fig. 5.9 Confidence intervals for the estimates.
– exact values, - - - confidence intervals, -•-•- estimates. Experimental error of the highest temperature measurement: (a) 5 %; (b) 2.5 %.

Results of the estimates of (n, τ_0) are presented in Fig. 5.9, in which the exact values for the test case, $(2.0, 3.0)$, were known *a priori*. The results correspond to estimates obtained by means of different sets of experimental synthetic data. Two levels of experimental errors corresponding to 5 % and 2.5 % are used, and the results are presented, respectively, in Figs. 5.9a and b.

As expected, these estimates are better (smaller confidence intervals for the same confidence level) when the experimental data present lower noise levels.

Exercises

5.1. We recall that Newton's method, presented in Section 4.3, for the system of equations $\mathbf{G}(\mathbf{Z}) = 0$, where $\mathbf{G} : \mathbb{R}^N \to \mathbb{R}^N$ is a vector-valued function, is given by the iterative scheme,

$$\begin{aligned} \mathcal{J}\mathbf{G}^n \Delta\mathbf{Z}^n &= -\mathbf{G}^n \\ \mathbf{Z}^{n+1} &= \mathbf{Z}^n + \Delta\mathbf{Z}^n \ , \end{aligned}$$

where $\mathcal{J}\mathbf{G}^n$ represents the Jacobian matrix of $\mathbf{G}$, evaluated at $\mathbf{Z}^n$

$$\mathcal{J}\mathbf{G}^n = \mathcal{J}\mathbf{G}|_{Z=Z^n} \ ,$$

and $\mathbf{G}^n = \mathbf{G}(\mathbf{Z}^n)$. We remark that $\Delta\mathbf{Z}^n$ satisfies a linear system of equations.

In the case of Eq. (5.4), the function $\mathbf{G}$ has the following structure,

$$G_k(\mathbf{Z}) = \sum_{i=1}^{M} R_i \ \partial I_i / \partial Z_k \ ,$$

where, for simplicity, we are omitting the dependency of $\mathbf{G}$ in $\mathbf{X}$.

Show that the equation for $\Delta\mathbf{Z}^n$ reads

$$[(\mathcal{J}^n)^T \mathcal{J}^n + A^n]\Delta\mathbf{Z}^n = -(\mathcal{J}^n)^T \mathbf{R}^n \ ,$$

where

$$A^n_{jk} = \sum_{i=1}^{n} R_i(\mathbf{Z}^n) \left(\partial^2 I_i / \partial Z_j \partial Z_k\right) \ .$$

(Compare this procedure with Eq. (5.8) and note that here, second order derivatives of $\mathbf{I}$ are needed in order to compute matrices A^n, which make this algorithm more expensive and more prone to numerical errors than the other.)

5.2. Deduce the Levenberg-Marquardt method considering the functional given by Eq. (3.20).

5.3. Assume that the participant medium is isotropic, i.e., $p(\mu,\mu') = c$ is a constant.

(a) Use Eq. (5.15) to determine the value of constant c.

(b) Obtain the simplified version of the Boltzmann equation, Eq. (5.14), for $I = I(\tau, \mu)$.

(c) Show that there is not a solution of the integro-differential equation for the isotropic medium, with the prescribed boundary conditions, depending only on τ, i.e, $I = I(\tau)$.

(d) Obtain a solution $I = I(\tau)$ if only the first requirement in Eq. (5.14b) is asked for.

5.4. An alternative approach for the solution of Eq. (5.14a) uses an integrating factor. Define a new dependent variable as

$$J(\tau,\mu) = e^{\frac{\tau}{\mu}} I(\tau,\mu)$$

and show that Eq. (5.14a) may be written as

$$\mu \frac{\partial J}{\partial \tau}(\tau,\mu) \quad = \quad \frac{\omega}{2} \int_{-1}^{1} p(\mu,\mu') e^{\tau\left(\frac{1}{\mu} - \frac{1}{\mu'}\right)} J(\tau,\mu')\, d\mu'$$

Write the boundary conditions given by Eq. (5.14b) in terms of the new dependent variable, [57].

Chapter 6
Thermal Characterization

The development of new materials is a research intensive area, which is fueled by ever-increasing technological demands. With relevant applications, both in engineering and medicine, there is an obvious need for using adequate techniques for the characterization of new materials to verify if they meet the physical and chemical properties specified in the design phase, such as viscosity, elasticity, density, etc. In this context, the meaning of characterization differs from the meaning we have established in Chapters 1 and 2. Here *characterization* means determination of the properties of the material, which we call inverse identification problem, and requires, most commonly, the conduction of laboratory tests. We recognize the ambiguity with our previous use of the word "characterization", but since in the area of materials the word characterization is used in the sense of identification we prefer to maintain it and be able to have the chosen chapter title.

During the development and operation of an experimental device, it is common to control different degrees of freedom to correctly estimate the properties under scrutiny. Frequently, with such a procedure, practical limitations are imposed that restrict the full use of the possibilities of the experiment.

Using a blend of theoretical and experimental approaches, determining unknown quantities by coupling the experiment with the solution of inverse problems, a greater number of degrees of freedom can be manipulated, involving even the simultaneous determination of new unknowns, included in the problem due to more elaborate physical, mathematical and computational models [18].

The *hot wire method*, for example, has been used successfully to determine the thermal conductivity of ceramic materials, even becoming the worldwide standard technique for values of up to 25 W/(m K). For polymers, the parallel hot wire technique is replaced by the cross-wire technique, where a junction of a *thermocouple*—a temperature sensor—is welded to the hot wire, which works as a thermal source at the core of the sample whose thermal properties are to be determined [17].

In this chapter we consider the determination of thermal properties of new polymeric materials by means of the solution of a heat transfer inverse problem, using experimental data obtained by the hot wire method. This consists of an identification inverse problem, being classified as a Type III *inverse problem* (see Section 2.8).

6.1 Experimental Device: Hot Wire Method

In this section we briefly describe the experimental device used to determine the thermal conductivity of new materials.

The hot wire method is a *transient technique*, i.e., it is based on the measurement of the time variation of temperature due to a linear heat source embedded in the material to be tested. The heat generated by the source is considered to be constant and uniform between both ends of the test body. The basic elements of the experimental device are sketched in Fig. 6.1. From the temperature variation, measured by the slope in Fig. 6.2a, in a known time interval, the thermal conductivity of the sample is computed. In practice, the linear thermal source is approximated by a thin electric resistor and the infinite solid is replaced by a finite size sample.

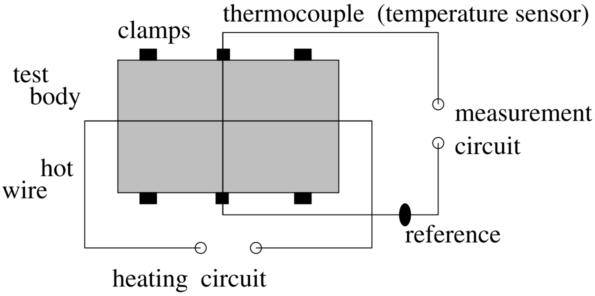


Fig. 6.1 Experimental apparatus for the standard hot wire technique

The experimental apparatus is made up of two test bodies. In the upper face of the first test body, two orthogonal incisions are *carved* to receive the measuring cross. The depth of these incisions corresponds to the diameter of the wires to be inserted within.

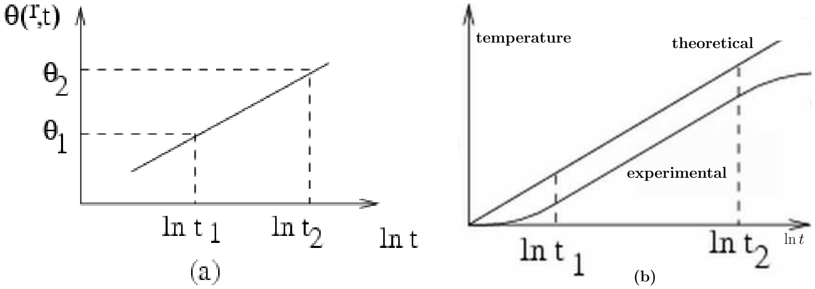


Fig. 6.2 Hot wire method. (a) Increase in temperature $\theta(r,t)$ as a function of time; (b) Theoretical (infinite sample size) and presumed experimental (finite sample size) graphs.

The measuring cross is formed by the hot wire (a resistor) and the thermocouple, whose junctions are welded perpendicular to the wire. After placing the measuring cross in the incisions, the second test body is placed upon it, wrapping the measuring cross. The contact surfaces of the two test bodies must be sufficiently flat to ensure good thermal contact. Clamps are used to fulfill this goal, pressing the two bodies together.

Some care should be taken when working with the hot wire method to ensure the reliability of the results: (i) a resistor must be used as similar as possible to the theoretical linear heat source; (ii) ensure the best possible contact between the sample and the hot wire; (iii) the initial part of the *temperature* × *time* graph should not be used for the computations —use only times in the range $t > t_1$, in Fig. 6.2b,— thus eliminating the effect of the thermal contact resistance between the electric resistor (the wire) and the sample material; (iv) limit the test time to ensure that the finite size of the sample does not affect the linearity of the measured temperatures ($t < t_2$ in Fig. 6.2b).

6.2 Traditional Experimental Approach

Consider a linear thermal source that starts releasing heat due to Joule's effect— a resistor, for example— at time $t = 0$, inside an infinite medium that is initially at temperature $T = T_0$. Let the linear thermal source be infinite in extension and located in the z axis. Due to the symmetry of the problem in the z direction, we have a solution that does not depend on z and this situation can be modeled as an initial value problem for the heat equation in two-dimensions,

$$\frac{\partial T}{\partial t} = k \triangle T + s(\mathbf{x},t)\,, \quad \mathbf{x} = (x,y) \in \mathbb{R}^2\,,\ t > 0\,, \tag{6.1a}$$

$$T(\mathbf{x},0) = T_0(\mathbf{x})\,, \quad \mathbf{x} \in \mathbb{R}^2\,. \tag{6.1b}$$

Here $T = T(\mathbf{x},t)$ is the temperature, $\triangle T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}$ is the laplacian of T with respect to the spatial variables, k is the medium's thermal conductivity, s is the thermal source term, and T_0 is the initial temperature. Under the previous hypothesis, T_0 is a constant, and s is a singular thermal source corresponding to a multiple of a *Dirac's delta* (generalized) function centered at the origin,

$$s(\mathbf{x},t) = q'\delta(\mathbf{x})\,, \tag{6.2}$$

where q' is the linear power density.

The solution of Eq. (6.1) can be written as the sum of a general solution of a homogeneous initial value problem, T^1, and a particular solution of a non-homogeneous initial value problem, T^2, that is, $T = T^1 + T^2$, where T^1 satisfies

$$\frac{\partial T^1}{\partial t} = k \triangle T^1\,, \quad \mathbf{x} \in \mathbb{R}^2, t > 0\,, \tag{6.3a}$$

$$T^1(\mathbf{x},0) = T_0\,, \quad \mathbf{x} \in \mathbb{R}^2\,. \tag{6.3b}$$

and T^2 satisfies

$$\frac{\partial T^2}{\partial t} = k \triangle T^2 + s(\mathbf{x},t)\,, \quad \mathbf{x} \in \mathbb{R}^2, t > 0\,, \tag{6.4a}$$

$$T^2(\mathbf{x},0) = 0\,, \quad \mathbf{x} \in \mathbb{R}^2\,. \tag{6.4b}$$

The solution of Eq. (6.3) relies on the fundamental solution of the heat equation [39], through a convolution with the initial condition,

$$T^1(\mathbf{x},t) = \frac{1}{4k\pi t}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4kt}}\, T_0(\mathbf{y})\, dy_1\, dy_2\ . \tag{6.5}$$

The solution of Eq. (6.4) is attained by *Duhamel's principle* [39, 61]. One looks for a solution in the form of *variation of parameters*,

$$T^2(\mathbf{x},t) = \int_0^t U(\mathbf{x},t,\tau)\, d\tau\ , \tag{6.6}$$

where, for each τ, $U(\,\cdot\,,\,\cdot\,,\tau)$ satisfies a homogeneous initial value problem, with initial time $t = \tau$,

$$\frac{\partial U}{\partial t} = k \bigtriangleup U\ ,\quad \mathbf{x}\in\mathbb{R}^2, t>\tau\ , \tag{6.7a}$$

$$U(\mathbf{x},\tau,\tau) = s(\mathbf{x},\tau)\ ,\quad \mathbf{x}\in\mathbb{R}^2\ . \tag{6.7b}$$

Since Eq. (6.7) is, in fact, a family of homogeneous problems, parametrized by τ, its solution is obtained by convolution with the fundamental solution of the heat equation,

$$U(\mathbf{x},t,\tau) = \frac{1}{4k\pi(t-\tau)}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4k(t-\tau)}}\, s(\mathbf{y},\tau)\, dy_1\, dy_2\ ,$$

and then, substituting this result in Eq. (6.6),

$$T^2(\mathbf{x},t) = \int_0^t \frac{1}{4k\pi(t-\tau)}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4k(t-\tau)}}\, s(\mathbf{y},\tau)\, dy_1\, dy_2\, d\tau\ .$$

Since

$$\frac{1}{4k\pi t}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4kt}}\, dy_1\, dy_2 = 1\ , \tag{6.8}$$

T_0 is a constant, and s is given by Eq. (6.2), we have

$$\begin{aligned} T(\mathbf{x},t) &= T_0 + \frac{q'}{4k\pi}\int_0^t \frac{1}{t-\tau} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4k(t-\tau)}}\, d\tau\ ,\\ &= T_0 + \frac{q'}{4k\pi}\int_{|x|^2/4kt}^{+\infty} \frac{e^{-u}}{u}\, du\ , \end{aligned} \tag{6.9}$$

where we have made the change of variables $u = |\mathbf{x}|^2/4k(t-\tau)$.

For times sufficiently greater than $t = 0$, and for radial distances, r, near the linear source, more precisely, when $|\mathbf{x}|^2/4kt \to 0$, the temperature increases in the following way, [12],

$$T(\mathbf{x},t) = T_0 + \frac{q'}{4\,\pi\, k}(\ln t - 2\ln|\mathbf{x}|) + \mathcal{O}(1),\quad \text{as } |\mathbf{x}|^2/4kt \to 0\ , \tag{6.10}$$

as can be seen from Eq. (6.9), and the following result,

$$\int \frac{e^{-u}}{u}\, du = \ln u\, e^{-u} + (u \ln u - u)\, e^{-u} + \int (u \ln u - u)\, e^{-u}\, du\,. \tag{6.11}$$

This dependence is represented in Fig. 6.2a.

Now, from Eq. (6.10) and Fig. 6.2, letting $\mathbf{x}_0 \neq 0$ be a certain fixed point of the medium, and denoting $\theta_1 = T(\mathbf{x}_0,t_1)$, and $\theta_2 = T(\mathbf{x}_0,t_2)$, we get

$$\text{'slope'} \approx \frac{\theta_2 - \theta_1}{\ln t_2 - \ln t_1} = \frac{q'}{4\pi k}\,,$$

$$\text{and then} \quad k = \frac{q'}{4\pi}\, \frac{\ln\left(\frac{t_2}{t_1}\right)}{(\theta_2 - \theta_1)}\,. \tag{6.12}$$

In the traditional experimental approach, temperatures are measured for different times, (t_l, T_l), for $l = 1,2,\ldots,L$, where L is the total number of experimental measurements, and, from the fitting of a line to the points

$$(\ln t_l, \theta_l)\,,\ \text{with}\ \theta_l = T_l - T_0\,,\ l = 1,2,\ldots,L\,,$$

by means of the least squares method, the slope of the line is obtained, and from it the thermal conductivity of the material by means of Eq. (6.12). A few more details can be found in Exercise 6.4.

This method was used to determine the thermal conductivity of a phenolic foam, with 25 % of its mass being of lignin.[1] The lignin used was obtained from sugarcane bagasse. This is an important by-product of the sugar and ethanol industry, and different applications are being sought for it, besides energy generation. The thermal conductivity was found as

$$k = (0.072 \pm 0.002)\,\mathrm{W/(m\,K)}\,. \tag{6.13}$$

The theoretical curve for an infinite medium and the expected curve, presumably obtainable in an experiment with a finite sample are presented in Fig. 6.2b. Observe that for time values relatively small ($t < t_1$) and relatively large ($t > t_2$), deviations from linearity occur. Therefore, experimental measurements in these situations are to be avoided. The deviation for $t < t_1$ is due to the thermal resistance between the hot wire and the sample. The deviation from linearity for $t > t_2$ occurs when heat reaches the sample's surface, thus starting the process of heat transfer by convection to the environment.

In a real experiment the sample's dimensions are finite. Moreover, for materials with high thermal diffusivity, $\alpha = k/\rho c_p$, where ρ is the specific mass and c_p is the specific heat at constant pressure per unit mass, the interval where linearity occurs can be very small. This feature renders experimentation unfeasible, within the required precision.

[1] The experimental data used here was obtained by Professor Gil de Carvalho from Rio de Janeiro State University [18].

6.3 Inverse Problem Approach

In this section, we present a more general approach to identifying the relevant parameters in the physical model, based on solving an inverse problem. First we present the model, next we set up an optimization problem to identify the model, present an algorithm to solve the minimization problem, and present the results on the determination of the thermal conductivity and specific heat of a phenolic foam.

6.3.1 Heat Equation

We shall consider the determination of the thermal conductivity of the medium using the point of view of applied inverse problems methodology. That is, we select a mathematical model of the phenomenon —heat transfer by conduction,— then formulate a least squares problem and set up an algorithm to solve it.

To deal with the inverse problem of heat transfer by conduction used here, consider a sample of cylindrical shape with radius R, with a linear heat source along its centerline, exchanging heat with the surrounding environment (ambient), and set initially at room temperature, T_{amb}. To keep the description as simple as possible, it will be considered that the cylinder is long enough, making the heat transfer depend only on the radial direction. The mathematical formulation of this problem is given by the *heat equation* and Robin's boundary conditions [12, 61],

$$\frac{1}{r}\frac{\partial}{\partial r}\left(k\, r\,\frac{\partial T}{\partial r}\right) + g(r,t)\,\delta(r) = \rho\, c_p\,\frac{\partial T(r,t)}{\partial t} \tag{6.14a}$$

in $0 \le r \le R$, for $t > 0$, and

$$-k\frac{\partial T}{\partial r}(R,t) = h\,(T(R,t) - T_{\mathrm{amb}})\,, \qquad \text{for } \ t > 0 \tag{6.14b}$$

$$T(r,0) = T_{\mathrm{amb}} \qquad \text{in } \ 0 \le r \le R\,, \tag{6.14c}$$

where $g(r,t)$ is the volumetric power density, h is the convection heat transfer coefficient, and the remaining symbols have already been defined.

When the geometry, material properties, boundary conditions, initial condition and source term are known, Eq. (6.14) can be solved, thus determining the medium's transient temperature distribution. This is a direct problem.

If some of these magnitudes, or a combination of them, are not known, but experimental measurements of the temperature inside or at the boundary of the medium are available, we deal with an inverse problem, which allows us to determine the unknown magnitudes, granted that the data holds enough information.

Most of the techniques developed to solve inverse problems rely on solving the direct problem with arbitrary values for the magnitudes that are to be determined. Usually, the procedures involved are iterative, so the direct problem has to be solved several times. It is thus desirable to have a method of solution of the direct problem capable of attaining a high precision. At the same time, it should not consume much computational time. In the example considered in Section 6.3.5, the finite difference method was used to solve the problem of heat transfer through the sample.

6.3.2 Parameter Estimation

Here we consider the formulation of the problem of simultaneously estimating the thermal conductivity and the specific heat of a material. These parameters are represented by

$$\mathbf{Z} = \left(k, c_p\right)^T .$$

Notice that other parameters could be estimated simultaneously with the thermal conductivity and the specific heat, such as the coefficient of heat transfer by convection from the sample to the environment, h. In this case, we should also perform measurements at times $t > t_2$.

Let $T_\mathrm{c}(r_m,t_l)$ be computed temperatures, and $T_\mathrm{e}(r_m,t_l)$ experimentally measured temperatures, at positions r_m, with $m = 1, 2, \ldots, M$, where M is the number of temperature sensors employed, at times t_l, with $l = 1, 2, \ldots, L$, and L denoting the number of measurements performed by each sensor. Consider the norm given by half the sum of the squares of the residues between computed and measured temperatures,

$$Q(\mathbf{Z}) = \frac{1}{2} \sum_{m=1}^{M} \sum_{l=1}^{L} \left[T_\mathrm{c}(r_m, t_l) - T_\mathrm{e}(r_m, t_l)\right]^2 , \qquad (6.15)$$

or, simply,

$$Q = \frac{1}{2} \sum_{i=1}^{I} (T_i - W_i)^2 = \frac{1}{2} \mathbf{R}^T \mathbf{R} .$$

Here, T_i and W_i, are compact notations, respectively for the calculated and measured temperature, referred to the same sensor and at the same time. Also, $R_i = T_i - W_i$ and $I = M \times L$.

The inverse problem considered here is solved as a finite dimension optimization problem, where the norm Q is to be minimized, and the parameters correspond to the minimum point of Q.

6.3.3 Levenberg-Marquardt

We describe here the *Levenberg-Marquardt method* [54], presented in section 5.3, to estimate the parameters.

The minimum point of Q, Eq. (6.15), is pursued by solving the critical point equation

$$\partial Q / \partial Z_j = 0 , \; j = 1, 2 .$$

Analogously to Section 5.3, an iterative procedure is built. Let n be the iteration counter. New estimates of parameters, $\mathbf{Z}^{n+1}$, of residuals, $\mathbf{R}^n$, and corrections, $\Delta\mathbf{Z}^n$, are computed sequentially,

$$\mathbf{R}^n = \mathbf{T}^n - \mathbf{W} \tag{6.16a}$$

$$\Delta\mathbf{Z}^n = -\left[(J^n)^T J^n + \lambda^n \mathcal{I}\right]^{-1} (J^n)^T \mathbf{R}^n \, , \tag{6.16b}$$

$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \Delta\mathbf{Z}^n \tag{6.16c}$$

for $n = 0, 1, 2, \ldots$, until the convergence criterion

$$\left|\Delta Z_j^n / Z_j^n\right| < \varepsilon \, , \; j = 1, 2$$

is satisfied. Here, ε is a small number, for example, 10^{-5}.

The elements of the $I \times 2$ Jacobian matrix,

$$J_{ij} = \partial T_i / \partial Z_j \, , \text{ for } i = 1, \ldots, I \, , \text{ and } j = 1, 2 \, ,$$

as well as the residuals, $\mathbf{R}^n$, are computed at every iteration, by the solution of the direct problem given by Eq. (6.14), using the estimates for the unknowns obtained in the previous iteration.

6.3.4 Confidence Intervals

As presented in Section 5.4, Walds's *confidence intervals* of the estimates $\mathbf{Z} = (k, c_p)^T$ are computed by [59], page 87, [34]. In this case, the square of the *standard deviation* is given by [38]

$$\sigma_Z^2 = \begin{pmatrix} \sigma_k^2 \\ \sigma_{c_p}^2 \end{pmatrix} = \sigma^2 \left\{ \mathrm{diag}\left[(\nabla\mathbf{T})^T \, \nabla\mathbf{T}\right]^{-1} \right\} \, . \tag{6.17}$$

where $\mathbf{T} = (T_1, \ldots, T_I)^T$, $\mathbf{T} = \mathbf{T}(\mathbf{Z}) = \mathbf{T}(k, c_p)$, and σ is the *standard deviation* of the experimental errors.

Assuming a normal distribution for the experimental errors and 99 % of confidence, the confidence intervals of the estimates of k and c_p are [33],

$$]k - 2.576\,\sigma_k \;\; , \;\; k + 2.576\,\sigma_k[\;\; ,$$

and

$$\left]c_p - 2.576\,\sigma_{c_p} \;\; , \;\; c_p + 2.576\,\sigma_{c_p}\right[\;\; .$$

6.3.5 Application to the Characterization of a Phenolic Foam with Lignin

This section presents the results obtained in the estimation of thermal conductivity and specific heat of a phenolic foam, with 25 % of its mass being of lignin. As mentioned at the beginning of this chapter, in materials's literature the word 'characterization' is used to mean what we call model identification, therefore we give credit to this usage by employing it in this section's title.

Recall that the traditional experimental approach, described in Section 6.1, was only able to determine the thermal conductivity. With the approach based on the inverse heat transfer problem, described in Section 6.3.2, we were able to obtain, from the same set of experimental data, not only the thermal conductivity, but also the sample's specific heat. The thermal conductivity was estimated as being

$$k = 0.07319\,\mathrm{W/(mK)}\,,$$

with the following 99 % confidence interval:

$$]0.07313, 0.07325[\,\mathrm{W/(m\,K)}\,.$$

This value excellently agrees with the one obtained by the traditional approach, Eq. (6.13).

For the specific heat, determined simultaneously with the thermal conductivity, the estimate obtained was $c_p = 1563.0\,\mathrm{J/(kg\,K)}$ and the following 99 % confidence interval was obtained

$$]1559.6, 1566.4[\,\mathrm{J/(kg\,K)}\,.$$

Vega [91] presents an expected value of 1590 J/(kg K) for the specific heat of phenolic resins, and this agrees very well with the value obtained by the solution of the inverse problem considered here. The traditional approach provides no means of estimating this property.

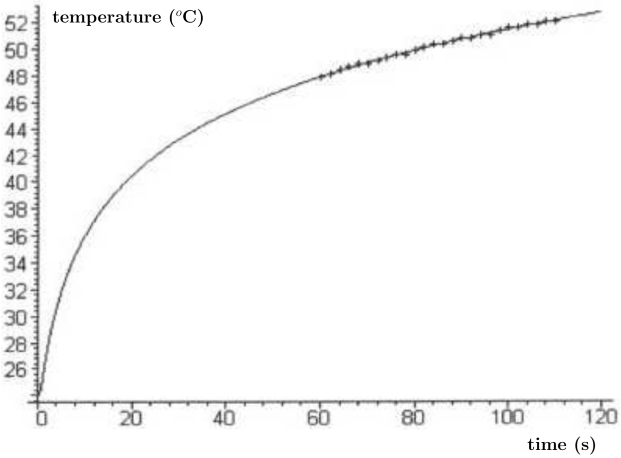


Fig. 6.3 Temperature profiles (– theoretical +++ experimental)

Figure 6.3 presents the *temperature*×*time* plot, which exhibits the computed temperatures with the estimated properties k and c_p. The values of the measured temperatures W_i, $i = 1,2,\ldots,I$, which were used in the solution of the inverse problem, are exhibited in the same graph. Notice the excellent agreement between experimental data and temperature determined by the computational simulation.

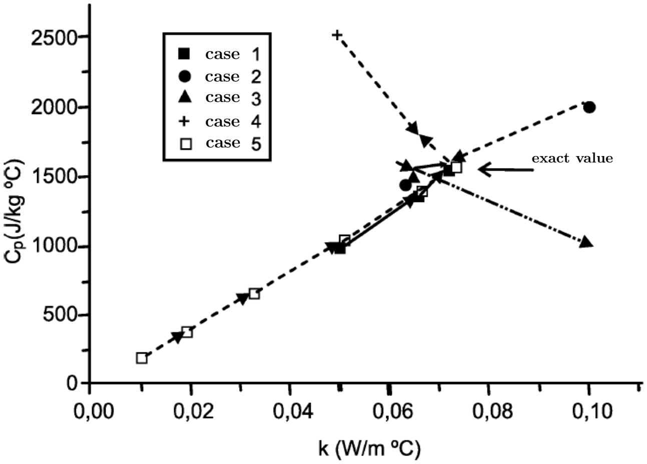


Fig. 6.4 Results of the simulations considering different initial estimates

Figure 6.4 shows that the minimization iterative process converges to the same solution, no matter which one of several initial estimates of the unknown magnitudes is used. This suggests that the global minimum of Q is reached.

6.4 Experiment Design

Using the mathematical model and the computational simulation presented here, it is possible to design experiments to obtain, precise and economically, physical parameters, determining *a priori* the best localization for the temperature sensors, as well as the best time intervals in which the experimental measurements are to be performed.

The concepts of "best" or "optimum" are necessarily bound to a judgment criterion that, in the situation described here, may, for example, consist in the minimization of the region contained in the confidence intervals that, as previously described in this chapter, and in Chapter 5, is related to larger values of the sensitivity coefficients.

Further details on inverse problems and experiment design for applications related to heat and mass transfer phenomena may be found in [50, 71, 45, 51, 89, 52, 49].

Exercises

6.1. Show that T^2 defined in Eq. (6.6) satisfies Eq. (6.4).
Hint. Use Bernoulli's formula

$$\frac{d}{dt}\int_a^t f(s,t)\,ds = f(t,t) + \int_a^t \frac{\partial f}{\partial t}(s,t)\,ds\,.$$

6.2. Show validity of Eq. (6.8).
Hint. Use polar coordinates in $\mathbb{R}^2$.

6.3. Use integration by parts to show Eq. (6.11).

6.4. From Eq. (6.10), $\theta = \alpha \ln t$, for $\alpha = q'/4\pi k$. (a) Given measurements (t_i,θ_i), obtain the least squares formulation to determine α; (b) obtain an expression for α in terms of the experimental data; (c) write an expression for k.

6.5. Let

$$\mathcal{D} = \sum_{l=1}^{L}\left(\frac{\partial T_l}{\partial k}\right)^2 \sum_{l=1}^{L}\left(\frac{\partial T_l}{\partial c_p}\right)^2 - \left(\sum_{l=1}^{L}\frac{\partial T_l}{\partial k}\frac{\partial T_l}{\partial c_p}\right)^2 .$$

Show that

$$\sigma_k^2 = \frac{\sigma^2}{\mathcal{D}}\sum_{l=1}^{L}\left(\frac{\partial T_l}{\partial c_p}\right)^2 \text{ , and}$$

$$\sigma_{c_p}^2 = \frac{\sigma^2}{\mathcal{D}}\sum_{l=1}^{L}\left(\frac{\partial T_l}{\partial k}\right)^2$$

6.6. Write the matrix $J^T J$, to be used in Eq. (6.16b), for the vector of unknowns $Z = (k,c_p,h)^T$, where k is the thermal conductivity, c_p is the specific heat, and h is the convection heat transfer coefficient, for the inverse heat conduction problem in which these three parameters are to be estimated simultaneously.

6.7. Why, for $t < t_2$, is the approximation of infinite medium, in the situation represented in Fig. (6.2), a good one?

6.8. The sensitivity coefficients, [11], are defined by

$$X_{z_j} = \frac{\partial T}{\partial Z_j}$$

where T represents the observable variable, that may be measured experimentally, and z_j one unknown to be determined with the solution of the inverse problem. Considering the situation represented in Fig. (6.2), is it possible to estimate the convection heat transfer coefficient, i.e. $Z_j = h$ considering the experimental data acquired at $t < t_2$? What is the link between this exercise and Exercise 6.7?

Chapter 7
Heat Conduction

We present in this chapter estimates of the intensity of thermal sources with spatial and temporal dependence in heat transfer by conduction. This is an *inverse* reconstruction problem, and it is classified as a Type IV *inverse problem*[1]. In other words, we consider a function estimation problem in an infinite dimensional model.

The regularization of the inverse problem is attained by changing it to the problem of optimizing a functional defined in an infinite dimension space by means of the conjugate gradient method. Thus, as described in Sections 3.7, 3.8 and 3.10, respectively for the steepest descent, Landweber, and conjugate gradient methods, regularization is achieved by means of an iterative minimization method. We employ here *Alifanov's iterative regularization method* [1, 92, 87].

7.1 Mathematical Formulation

Let Ω denote the set where the physical problem under investigation is defined —a space-time or a spatial region where the action takes place. See Fig. 7.1. Consider the problem of determining a function g, knowing the problem's input data, $\mathbf{f}$, experimental measurements, X, and how g influences the output of the physical system, represented by $T[g]$.

Let us represent by

$$\Gamma = \Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_M \ ,$$

the subset (subregion) of Ω where experimental measurements are taken, and by

$$\Omega \supset \Gamma \ni x \mapsto X(x) \in \mathbb{R}^l \ ,$$

the measurements. If Γ_i is just an isolated point, $\Gamma_i = \{x_i\}$, the measurements can be represented simply by $X_i = X(x_i)$. Let

$$\Lambda = \Lambda_1 \cup \ldots \cup \Lambda_L \ ,$$

be the subset of Ω where the unknown function g is defined,

$$\Omega \supset \Lambda \ni r \mapsto g(r) \in \mathbb{R}^k \ .$$

Here Λ could be a part of the boundary of Ω, or some other subset of Ω.

The physical system output, due to some source g, $T[g]$, is defined in Ω,

$$\Omega \ni \omega \mapsto T[g](\omega) \in \mathbb{R}^l \ .$$

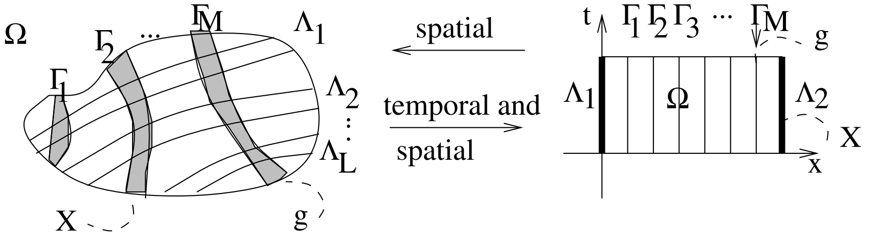


Fig. 7.1 Representation of the physical domain Ω, of subset Λ, where the unknown of the problem (function g) is defined, and of the region Γ, where the experimental measurements are performed

This quantity is evaluated at the points where the experimental measurements are performed, i.e., in Γ.

Given g, we can compute the difference between what the model predicts and the measurements. For each $x \in \Gamma$, we define the *residual*

$$T[g](x) - X(x)\,,$$

and, with this, by taking into account all the points where measurements are made, we define the functional representing half the sum — when Γ is a discrete set — or the integral — when Γ is a *continuum* set — of the squares of the residuals,

$$J[g] = \frac{1}{2}\int_{\Gamma} |T[g](x) - X(x)|^2\, dx\,. \tag{7.1}$$

The inverse problem of determining the function $g = g(r)$ is solved as a minimization problem in a space of infinite dimension, where one searches the minimum point of the functional J in a function space.

If $M > 1$, and $l = 1$, i.e., when several sensors are used and the measurements are scalar, it is customary to write the functional $J = J[g]$ in the form

$$J[g] = \frac{1}{2}\sum_{m}\int_{\Gamma_m} \{T_m[g](x) - X_m(x)\}^2\ dx\,,$$

where $T_m[g]$ and X_m represent the restrictions of $T[g]$ and X to the region Γ_m. If the measuring region is discrete —as is usually the case— and $\Gamma_m = \{x_1^m, \ldots, x_N^m\}$, the functional is rewritten as

$$J[g] = \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \{T_{m,n}[g] - X_{m,n}\}^2\ ,$$

where $T_{m,n}[g]$ and $X_{m,n}$ represent, respectively, the computed and measured quantities in $x_n^m \in \Gamma_m$. If Ω is a space-time subset, index n of x_n^m stands for successive time instants.

[1] See Table 2.3.

Observe that as mentioned in Section 5.1 the dependence of the system output T on the unknown g, represented here by $T[g]$, can be implicit in the sense that it is related, for example, to the solution of a differential equation, exactly as in the situation that we shall consider in this chapter. We remark that the dependence of the computed magnitude T in g implies a cause-effect relationship, which can be linear or non-linear, explicit or implicit.

To better understand the functional $J[g]$ described by Eq. (7.1), consider the situation depicted in Fig. 7.2a. This figure represents a thin plate with an internal thermal source, which can depend both on space and time, $g = g(x,t)$, and having its external surfaces insulated. The mechanism of heat transfer within the plate is purely by heat conduction. From transient temperature measurements, inside the medium, $X_m(t)$, $m = 1, 2, \ldots, M$ (see Fig. 7.2b), it is desired to estimate the intensity of the thermal source, $g(x, t)$.

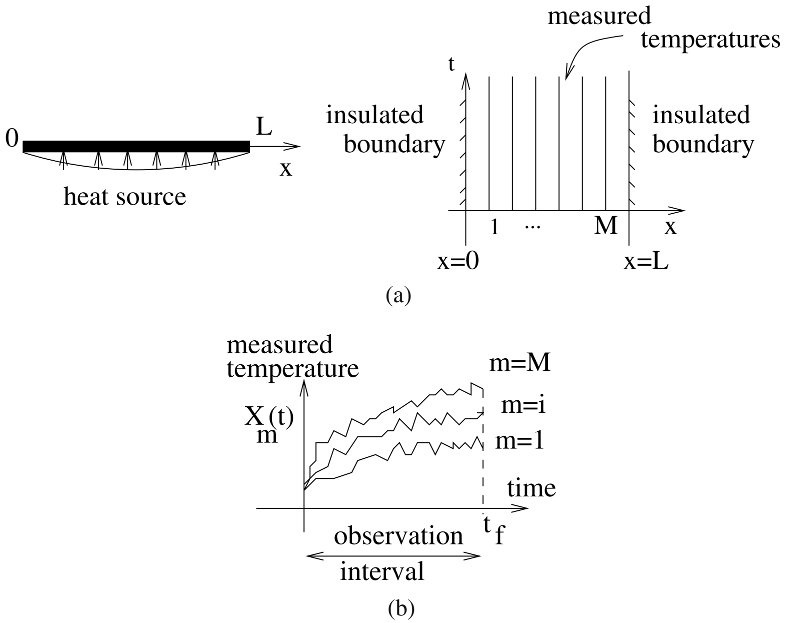


Fig. 7.2 (a) Distributed time-dependent thermal source. (b) Transient temperature measurements.

In this case, the functional J, Eq. (7.1), is written as

$$J[g] = \frac{1}{2} \sum_{m=1}^{M} \int_0^{t_f} \left[T_m[g](t) - X_m(t)\right]^2 \, dt \,, \tag{7.2}$$

where $[0,t_f]$ represents the observation time period in which experimental data is acquired.

To obtain a computational solution, a discretization of the independent variables (x and t in this example) is performed. Nonetheless, as will be seen in the next sections, the minimization procedure of the functional given generically by Eq. (7.1) does not depend on such discretization. The procedure is fully performed in a function space. The discretization is performed afterwards, and only to obtain a computational solution.

The parameter estimation problems presented in Chapters 5 and 6 are finite dimensional optimizations. The search for a solution is performed by solving the critical point equation of the residual functional,for example Eq. (5.2), in $\mathbb{R}^n$, to minimize it.

On the other hand, the conjugate gradient with the adjoint equation, Alifanov's regularization method, considers directly the functional minimization problem, by building a minimizing sequence. Conjugate gradient has been used successfully in solving inverse problems of function estimation. This method, which has been presented for real symmetric matrices in Section 3.10, will be presented in Section 7.3 for the problem of heat conduction. Its success is due, mainly, to two important features: (i) the regularization is embedded in it; and (ii) the method is computationally efficient.

7.2 Expansion in Terms of Known Functions

Before carrying on with the estimation of functions, assume known *a priori*, in the example presented in the previous section, that the intensity of the volumetric heat source $g = g(x,t)$ can be represented by *separation of variables*

$$g(x,t) = \left(\sum_{l=0}^{L} a_l P_l(x) \right) H(t) \,.$$

Here, $P_l(x)$ and $H(t)$ are known functions[2] and a_l, $l = 0, 1, \ldots, L$ are coefficients. The inverse problem of estimating a function $g(x,t)$ is, then, reduced to the estimation of a finite number of these coefficients,

$$\mathbf{Z} = (a_0, a_1, \ldots, a_L)^T \ .$$

In this case we would have a parameter estimation problem. Thus the technique described in Chapter 5 can be applied.

7.3 Conjugate Gradient Method

Consider the following iterative procedure for the *estimation* of the function $g(r)$ that minimizes the functional defined by Eq. (7.1) [64],

$$g^{n+1}(r) = g^n(r) - \beta^n P^n(r)\,, \qquad n = 0, 1, 2, \ldots\,, \tag{7.3}$$

where n is the iteration counter, β^n is a parameter that specifies the stepsize in the *search direction*, P^n, given by

$$P^n(r) = J'_{g^n}(r) + \gamma^n P^{n-1}(r)\,, \tag{7.4}$$

[2] Such as $P_l(x) = x^l$ and $H(t) = e^{-t}$.

where γ^n is the *conjugate coefficient*, with $\gamma^0 = 0$, and $J'_{g^n}(r)$ is the functional gradient[3] which, later on in this chapter, we shall show how to compute.

The case when $\gamma^n = 0$, $n = 1, 2, \ldots$, corresponds to the *steepest descent method*. This usually converges slowly.

The search stepsize, β^n, used in Eq. (7.3), is found by minimizing the functional

$$\mathbb{R} \ni \beta^n \mapsto J\left[g^{n+1}\right] = J\left[g^n - \beta^n P^n\right]$$

with relation to β^n, that is,

$$J\left[g^n - \beta^n P^n\right] = \min_{\beta} \frac{1}{2} \sum_{m=1}^{M} \int_{\Gamma_m} \left\{T_m\left[g^n - \beta P^n\right](x) - X_m(x)\right\}^2 \, dx \,. \tag{7.5}$$

The stepsize β^n is the solution of the critical point equation of functional J, restricted to a line passing through $g^n = g^n(r)$ in the direction defined by $P^n = P^n(r)$, i.e., β^n is the critical point of

$$\mathbb{R} \ni \beta \mapsto J[g^n - \beta P^n] \in \mathbb{R} \,.$$

In this case, β^n satisfies Eq. (7.5).

We will later show that in initial and boundary value problems for linear partial differential equations, homogeneous or not, β^n is given by

$$\beta^n = \frac{\displaystyle\sum_{m=1}^{M} \int_{\Gamma_m} \left(T_m[g^n](x) - X_m(x)\right) \Delta T_m[P^n](x) \, dx}{\displaystyle\sum_{m=1}^{M} \int_{\Gamma_m} \left(\Delta T_m[P^n](x)\right)^2 \, dx} \,, \tag{7.6}$$

where $\Delta T = \Delta T[P]$ represents the linear operator that solves the sensitivity problem with input value P and, likewise, ΔT_m is the evaluation of $\Delta T[P]$ in Γ_m.

The conjugate coefficient can be computed by

$$\gamma^n = \int_{\Lambda} \left[J'_{g^n}(r)\right]^2 \, dr \Bigg/ \int_{\Lambda} \left[J'_{g^{n-1}}(r)\right]^2 \, dr \,, \tag{7.7}$$

which minimizes $J[g^n - \beta^n P^n]$ with respect to the possible choices of γ^n in the definition of P^n, Eq. (7.4). This is in the same vein of what was done to show Eq (3.43d).

To use the iterative procedure described, it is necessary to know the gradient $J'_g(r)$ and the variation $\Delta T_m = \Delta T_m[P^n]$. When obtaining the gradient, an adjoint equation is used, and the variation ΔT_m is obtained as the solution to the sensitivity problem. In Section 7.4, we show how to obtain the adjoint equation, and consequently the gradient, as well as how to construct the sensitivity problem for the heat conduction problem.

[3] The general definition of the gradient of a functional is presented in Chapter 8.

7.4 Thermal Source in Heat Transfer by Conduction

In this section, we present basic problems — the sensitivity problem and the adjoint equation — which are instrumental in constructing an algorithm to determine thermal sources in heat conduction, from temperature measurements.

Obtaining the sensitivity problem, the adjoint equation and the gradient equation is directly related to the operator of the direct problem, that connects the computed magnitude, $T[g]$, to the function to be estimated, $g(r)$. These questions will be discussed in a more general way, in Chapter 8.

We discuss here an example considering the estimation of the spatial and temporal functional dependence of a thermal source, in a one dimensional medium, without any previous knowledge of the functional dependence, [78].

7.4.1 Heat Transfer by Conduction

Consider a one dimensional plate of thickness L, under the action of heat sources distributed within the material body, and time-dependent, represented by function $g = g(x,t)$. The plate is, initially, at temperature T_0, and its two boundary surfaces are thermally insulated at all times. The mathematical formulation of the problem of heat transfer by conduction in the medium, considering constant thermal properties (homogeneous, isotropic medium), is given by

$$k\frac{\partial^2 T}{\partial x^2}(x,t) + g(x,t) = \rho c_p \frac{\partial T}{\partial t}(x,t)\,, \quad 0 < x < L\,,\ t > 0 \tag{7.8a}$$

$$\frac{\partial T}{\partial x}(0,t) = 0\,, \quad \frac{\partial T}{\partial x}(L,t) = 0\,, \qquad \text{for } t > 0 \text{ and} \tag{7.8b}$$

$$T(x,0) = T_0\,, \qquad \text{in } 0 \leq x \leq L\,, \tag{7.8c}$$

where k is the material's thermal conductivity, ρ is its specific mass and c_p is the specific heat.

When the geometry, the material properties, the source term, the initial and boundary conditions are known, we have a direct problem, whose solution provides the knowledge of the temperature field in the full spatial and temporal domain. When some of these characteristics (or a combination thereof) is unknown, but experimental measurements of temperature inside the medium or at its boundary are available, we deal with an inverse problem, from which it is possible to estimate the unknown quantities.

Here, we will consider the inverse problem for the estimation of $g(x,t) = g_1(x)g_2(t)$ from the experimental transient measurements of temperature, $X_m(t)$, $m = 1,2,\ldots,M$, on the boundaries and inside the medium, [78].

The temperature field necessary to compute step 2 of the algorithm presented in Section 7.5.1 is obtained by solving Eq. (7.8), using as source term the estimate obtained in a previous step of the iterative procedure.

7.4.2 Sensitivity Problem

In general, the sensitivity problem corresponds to a *linearization* of the original problem. In the present case, the sensitivity problem is obtained by perturbing the heat source, $g \to g{+}\Delta g$, causing a variation in the temperature field, $T \to T{+}\Delta T$. The sensitivity problem corresponds to the problem satisfied by ΔT.

The problem given by Eq. (7.8) is then written as

$$\begin{aligned}
&k\frac{\partial^2(T+\Delta T)}{\partial x^2} + (g+\Delta g) = \rho c_p \frac{(\partial T+\Delta T)}{\partial t}\ ,\ 0<x<L\ ,\ t>0\ ,\\
&\frac{\partial(T+\Delta T)}{\partial x} = 0\ ,\ \text{ at } x=0 \text{ and } x=L\ ,\ t>0\ ,\ \text{and}\\
&T+\Delta T = T_0 \text{ for } t=0\ ,\ \text{ in } 0\le x\le L\ .
\end{aligned}$$

Since the problem is linear, when the perturbed equations are subtracted from the original problem, the *sensitivity problem* for ΔT is obtained,

$$k\frac{\partial^2 \Delta T}{\partial x^2}(x,t) + \Delta g(x{,}t) = \rho c_p \frac{\partial \Delta T}{\partial t}(x,t)\ ,\quad 0<x<L\ ,\quad t>0\ , \tag{7.9a}$$

$$\frac{\partial \Delta T}{\partial x}(x,t) = 0\ ,\ \text{ at } x=0 \text{ and } x=L\ ,\ t>0\ , \tag{7.9b}$$

$$\text{and } \Delta T(x,t) = 0 \text{ for } t=0\ ,\ \text{ and } 0\le x\le L\ . \tag{7.9c}$$

Notice that this problem is linear in Δg, but the problem in Eq. (7.8) is not linear in g, due to the non-homogeneity of the initial condition.

A perturbation in the source term, $g \to g+\Delta g$, causes a perturbation in the temperature distribution $T \to T+\Delta T$, that satisfies $T[g+\Delta g] = T[g]+\Delta T[\Delta g]$, where $\Delta T = \Delta T[\Delta g]$ represents the solution of Eq. (7.9) and is linear in Δg.

7.4.3 Adjoint Problem and Gradient Equation

To obtain the adjoint problem, we first consider the minimization problem for $J[g]$ given by Eq. (7.2), repeated here for convenience,

$$J[g] = \frac{1}{2}\sum_{m=1}^{M}\int_0^{t_f} \left[T[g](x_m,t) - X_m(t)\right]^2\, dt \tag{7.10}$$

subjected to the restriction that T satisfies Eq. (7.8). Here, x_m, $m=1,2,\ldots,M$, represent the positions of the temperature sensors, and t_f is the final time of observation (aquisition of experimental data).

This minimization with restriction may be turned into a minimization problem with no restrictions (unconstrained optimization). Multiply the restriction, Eq. (7.8a), properly equated to zero by letting the term in the right hand side be moved to the left hand side, by a *Lagrange multiplier*, $\lambda = \lambda(x{,}t)$, integrate it and add to $J[g]$ to construct a *Lagrangian* functional,

$$\mathcal{L}[g] = \frac{1}{2}\sum_{m=1}^{M}\int_0^{t_f} \left[T[g](x_m,t) - X_m(t)\right]^2\, dt + \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\left[k\frac{\partial^2 T}{\partial x^2}(x,t) + g(x,t) - \rho c_p \frac{\partial T}{\partial t}(x,t)\right] dx\, dt\,. \tag{7.11}$$

Here, $\lambda = \lambda(x,t)$ is a so-called *adjoint function*. The minimization of $\mathcal{L}$ is equivalent to the minimization of J restricted to Eq. (7.8a). Further, we have to minimize $\mathcal{L}$ restricted to satisfying Eqs. (7.8b,7.8c).

We shall compute the derivative[4] of $\mathcal{L}$ in g, and denote it by $d\mathcal{L}_g$. We have

$$\mathcal{L}[g+\Delta g] = \frac{1}{2}\sum_{m=1}^{M}\int_0^{t_f}\left[T_m[g](t) + \Delta T_m[\Delta g](t) - X_m(t)\right]^2\, dt \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\left[k\frac{\partial^2 (T+\Delta T)}{\partial x^2} + g(x,t)\right. \\ \left. + \Delta g(x,t) - \rho c_p \frac{\partial (T+\Delta T)}{\partial t}\right] dx\, dt\,. \tag{7.12}$$

Here, $T_m[g](t) = T[g](x_m,t)$. Subtracting Eq. (7.11) from Eq. (7.12) results

$$\mathcal{L}[g+\Delta g] - \mathcal{L}[g] \\ = \frac{1}{2}\int_0^{t_f}\sum_{m=1}^{M}\left[2\left(T_m[g](t) - X_m(t)\right) + \Delta T_m[\Delta g](t)\right]\Delta T_m[\Delta g](t)\, dt + \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\left[k\frac{\partial^2 \Delta T}{\partial x^2}(x,t) - \rho c_p\frac{\partial \Delta T}{\partial t}(x,t)\right] dx\, dt \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\Delta g(x,t)\, dx\, dt\,.$$

The derivative is obtained from the previous expression by dropping the second order terms in Δg due to the linearity of the considered problems. Here, it is only necessary to think that Δg is small, an *infinitesimal*, that $\Delta T_m[\Delta g] = O(\Delta g)$, and that second order terms, $(\Delta g)^2$, are even smaller and can be despised. Therefore,

$$d\mathcal{L}_g[\Delta g] = \int_0^{t_f}\sum_{m=1}^{M}\left[T_m[g](t) - X_m(t)\right]\Delta T_m[\Delta g](t)\, dt \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\left[k\frac{\partial^2 \Delta T}{\partial x^2}(x,t) - \rho c_p\frac{\partial \Delta T}{\partial t}(x,t)\right] dx\, dt \\ + \int_0^{t_f}\int_0^L \lambda(x,t)\Delta g(x,t)\, dx\, dt\,. \tag{7.13}$$

[4] The definition of the derivative of a functional, in an abstract context, is presented in Chapter 8.

Now, λ must be chosen in a way to cancel the sum of the first two integrals on the right side of the equation, or, in other words, in such a way that $d\mathcal{L}_g[\Delta g]$ is given by the last integral on the right hand side of Eq. (7.13). This is an operational rule that allows to determine the gradient of $\mathcal{L}$, $\mathcal{L}'_g$, and its justification is presented in Section 8.2.4.

Integrating by parts the second integral and using the boundary conditions that ΔT satisfies, we see that these integrals will cancel each other if λ satisfies the following *adjoint problem*

$$k\frac{\partial^2\lambda}{\partial x^2}(x,t) + \sum_{m=1}^{M}\left[T_m(x,t) - X(x,t)\right]\delta(x - x_m) = -\rho c_p \frac{\partial\lambda}{\partial t}(x,t) \tag{7.14a}$$

$$\text{for } \ 0 < x < L\,, \quad t > 0 \tag{7.14b}$$

$$\frac{\partial\lambda}{\partial x}(x,t) = 0 \ \text{ at } \ x = 0 \ \text{ and } \ x = L\,, \ t > 0 \tag{7.14c}$$

$$\text{and } \ \lambda(x,t_f) = 0, \ \ \text{for } 0 \le x \le L\,. \tag{7.14d}$$

Therefore, Eq. (7.13) is reduced to

$$d\mathcal{L}_g[\Delta g] = \int_0^{t_f}\int_0^L \lambda(x,t)\,\Delta g(x,t)\,dx\,dt\,. \tag{7.15}$$

Due to the definition, the *gradient* $\mathcal{L}'_g(x,t)$ is the function that represents the derivative in the inner product, i.e., it is such that

$$d\mathcal{L}_g[\Delta g] = \int_0^{t_f}\int_0^L \mathcal{L}'_g(x,t)\,\Delta g(x,t)\,dx\,dt\,. \tag{7.16}$$

Comparing Eqs. (7.15) and (7.16) we get that

$$\mathcal{L}'_g(x,t) = \lambda(x,t)\,.$$

It is obvious that

$$J[g] = \mathcal{L}[g,\lambda], \text{ for every } g \text{ and } \lambda\,,$$

because T satisfies the direct problem, and the term multiplied by λ in $\mathcal{L}$ is null. Then, both derivatives coincide, $dJ_g = d\mathcal{L}_g$ and the same is true for the gradients, $J'_g = \mathcal{L}'_g$. We conclude that

$$J'_g(x,t) = \lambda(x,t)\,,$$

i.e., the gradient of J is given by the solution of the adjoint problem.

Notice that the adjoint problem is a final value problem. Defining a new variable

$$t^* = t_f - t\,, \ \text{for } 0 \le t \le t_f\,, \tag{7.17}$$

it is transformed into an initial value problem, [78, 92, 87].

The direct problem, Eq. (7.8), the sensitivity problem, Eq. (7.9), and the adjoint problem, Eq. (7.14), after the change of variable mentioned previously, differ only in the source term. Therefore, the same computational routine can be used to solve the three problems, except that the solution to the adjoint equation has to take into account the change of variables (it should be integrated backwards in time).

7.4.4 Computation of the Critical Point

We present a derivation of Eq. (7.6). Initially notice that,

$$J[g+\gamma P] = \frac{1}{2}\sum_{m=1}^{M}\int_{\Gamma_m} \left(T_m[g](x) - X_m(x) + \gamma \Delta T_m[P](x)\right)^2 \, dx \,.$$

From this expression it can be deduced that

$$\frac{d}{d\gamma}J[g+\gamma P] = \sum_{m=1}^{M}\int_{\Gamma_m} (T_m[g](x) - X_m(x) \\ + \gamma \Delta T_m[P](x))\, \Delta T_m[P](x)\, dx \,. \qquad (7.18)$$

Having this derivative set equal to zero, replacing γ by $-\beta^n$, g by g^n, and P by P^n on the right side of the previous equation, and solving with respect to β^n, we obtain Eq. (7.6).

7.5 Minimization with the Conjugate Gradient Method

The iterative procedure that defines the conjugate gradient method in the infinite dimensional setting written to be applied to a heat conduction problem is presented here. The results are discussed in Section 7.6.

7.5.1 Conjugate Gradient Algorithm

The iterative procedure that defines the conjugate gradient method can be summarized as follows.

1. Let $n = 0$. Choose an initial estimate, $g^0(r)$, for example, $g^0(r) =$ ‘constant’;
2. Compute $T_m[g^n]$, $m = 1, 2, \ldots, M$, by solving the direct problem[5], Eq. (7.8);
3. Since $T_m[g^n]$ and the experimental measurements $X_m(x)$, $m = 1, 2, \ldots, M$, for $x \in \Gamma_m$, are known, solve the adjoint problem, Eq. (7.14), which determines the gradient[6], $J'_{g^n}(r)$;
4. Compute the conjugate coefficient, γ^n, with Eq. (7.7);
5. Compute the search direction, $P^n(r)$, with Eq. (7.4);
6. Solve the sensitivity problem with input data $\Delta g = P^n$, and obtain $\Delta T[P^n]$, Eq. (7.9);

[5] As mentioned in Section 7.1, this computation can be related to the solution of a differential equation, an integro-differential equation, or of an algebraic system of equations.

[6] See Section 7.4, for the appropriate gradient in a heat conduction problem.

7. Compute the stepsize in the search direction, β^n, with Eq. (7.6);

8. Compute the new estimate $g^{n+1}(r)$ with Eq. (7.3);

9. Interrupt the iterative procedure if the stopping criterion is satisfied. Otherwise, set $n = n + 1$ and return to Step 2.

7.5.2 Stopping Criterion and Discrepancy Principle

We now present a brief discussion on the stopping criterion. Real experimental data, $X_m(x)$, $m = 1, 2, \ldots, M$, for $x \in \Gamma_m$, is always contaminated by noise. In this case, the usual stopping criterion $J[g^{n+1}] < \delta$, where δ is a small value fixed *a priori*, without any further reasoning, may prove inadequate. In fact, the high frequencies of the experimental noise could be incorporated in the estimate $g^n(r)$, possibly rendering it useless. This effect can be appreciated in Fig. 7.3a. The original function, $g(r)$, is made up of an ascending slope and a descending slope. The estimate obtained presents high frequency oscillations.

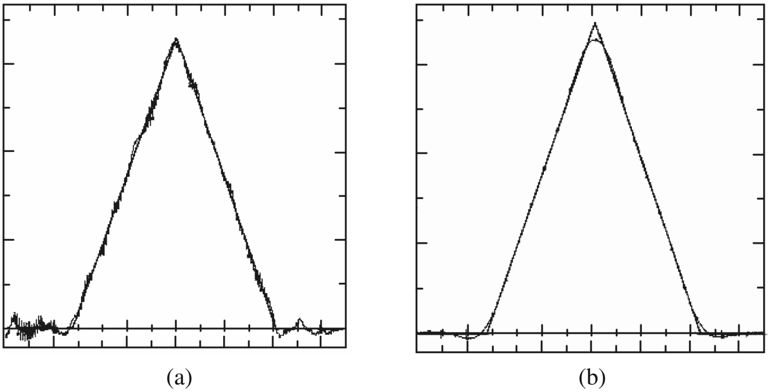


Fig. 7.3 Estimate of a function $g = g(r)$. The exact value of g is represented by a graph of triangular shape. (a) one can observe that the estimate is contaminated by high frequency experimental noise; (b) The estimate is performed by using the *discrepancy principle* as the stopping criterion for the iterative procedure, resulting in a smoother estimate, almost free from experimental noise.

Following the concepts presented in Section 3.9, let us analyze further this situation. We want to solve a problem of the form

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

where $\mathbf{x}$ is the unknown. However, instead of $\mathbf{y}$, we have a noisy experimental data $\mathbf{y}^\epsilon$, in such a way that

$$|\mathbf{y} - \mathbf{y}^\epsilon| < \epsilon\,.$$

Let $\mathbf{x}^{\alpha,\epsilon}$ be the estimate of $\mathbf{x}$ determined using the noisy data $\mathbf{y}^{\epsilon}$, with a *regularization scheme* (where α is the *regularization parameter*). It does not make sense, therefore, to demand that the residual, $|A\mathbf{x}^{\alpha,\epsilon} - \mathbf{y}^{\epsilon}|$, be much smaller than ϵ, [30]. The best we can expect is that

$$\text{`residual'} \approx \epsilon, \quad \text{or then,} \quad J = |\mathbf{R}|^2/2 \approx \epsilon^2/2 \,. \tag{7.19}$$

Let us assume that the standard deviation of the experimental errors, σ, is the same for all the sensors and measurements,

$$|T_m[g](t) - X_m(t)| \cong \sigma \,.$$

Replacing this approximation in Eq. (7.2), we have

$$J \approx M\sigma^2\, t_f/2 \,.$$

Let

$$\eta^2 = M\sigma^2 t_f/2 \,.$$

The stopping criterion to be followed then is the *discrepancy principle*, [1, 2], in which the iterative procedure is interrupted when

$$J[g^{n+1}] < \eta^2 \,.$$

7.6 Estimation Results

Recall that Fig. 7.3a presents the estimation of g when the stopping criteria does not take in consideration the discrepancy principle.

Figure 7.3b presents the estimate of the same function considered in Fig. 7.3a, using, however, the discrepancy principle as the stopping criterion for the iterative procedure. It is plain to see in this case that the interruption of the iterative procedure occurred previously to the beginning of the degradation of the estimates. Thus, the high frequency oscillations were not incorporated into the estimate. We can see with this example the role of the discrepancy principle.

A further example of the employment of the methodology concerns the estimation of the strength of a heat source of the form,

$$g(x,t) = g_1(x)\, g_2(t) \,.$$

The algorithm described in Section 7.5.1 is used, where, in Step 3, the adjoint function is computed using Eq. (7.14). The gradient is obtained and, in Step 6, the variation ΔT_m is computed using the sensitivity problem, given by Eq. (7.9), with the source term $\Delta g(x,t) = P^n(x,t)$, where $P^n(x,t)$ has been computed in Step 5.

Figures 7.4 and 7.5 show such an example of the estimation of the strength of time and space dependent heat sources for two different test cases. In Fig. 7.4 we consider a heat source with a gaussian dependence in both space and time. In the situation depicted in Fig. 7.5, a piecewise-linear periodic sawtooth function in space is considered. In these figures, the dimensionless time variable $\tau = \alpha t/L^2$, where $\alpha = k/\rho c_p$ is the material thermal diffusivity, and dimensionless space variable $X = x/L$, are used. Solid lines correspond to exact values for the strength of the heat source. Estimates are presented for the dimensionless time instants

$$\tau = 0.1\tau_f;\ 0.3\tau_f \text{ and } 0.5\tau_f\ .$$

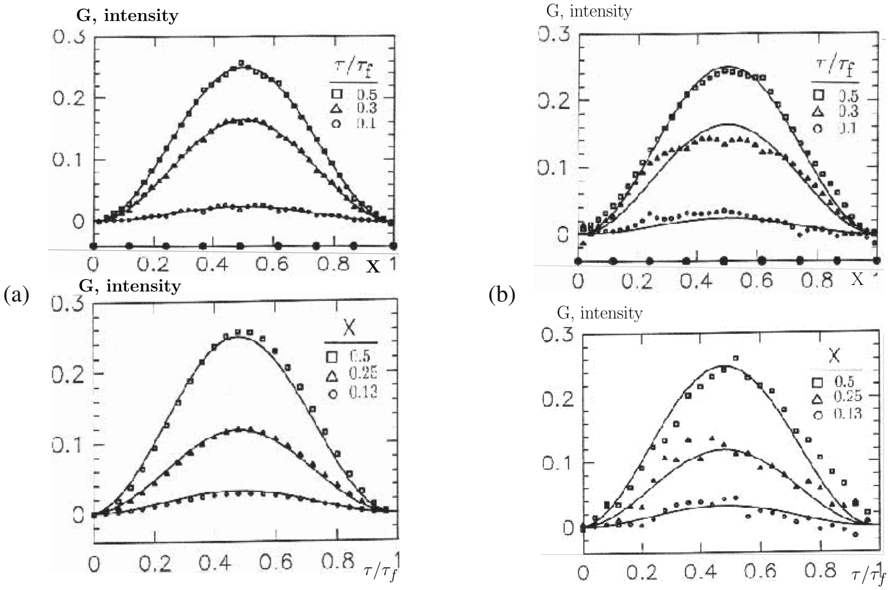


Fig. 7.4 Estimation of a space and time dependent volumetric heat source using nine temperature sensors (sensors positions are marked on X axis by bullets): (a) without experimental error – spatial variation and temporal variation. (b) with 13 % of experimental error – spatial variation and temporal variation.

Figure 7.4 present estimates for the strength of the heat source, in the dimensionless positions

$$X = 0.13;\ 0.25 \text{ and } 0.5\ ,$$

with and without experimental error. It is possible to observe the degradation of the estimates when a relatively large error level is considered in the experimental data used for the solution of the inverse heat conduction problem. Figure 7.5 exhibits results for

$$X = 0.06;\ 0.12 \text{ and } 0.17\ .$$

In both figures, $G = g/g_{\text{ref}}$, where g_{ref} was adjusted in such a way that the maximum measured value for the dimensionless temperature was unity.

In Figs. 7.5 estimates are presented for a heat source with sudden variations in the spatial component. The goal is to show the effect of the position of the temperature sensors. In both cases, experimental errors are not considered. Solid lines correspond to exact values for heat source strength. In Fig. 7.5a the temperature sensors are not in the same position where the sudden temperature variations occur. When better positions are chosen, the estimates improve significantly, as shown in Fig.7.5b. We surmise that in a general way the quality of the estimates obtained with the solution of an inverse problem improves when adequate information is introduced in the problem.

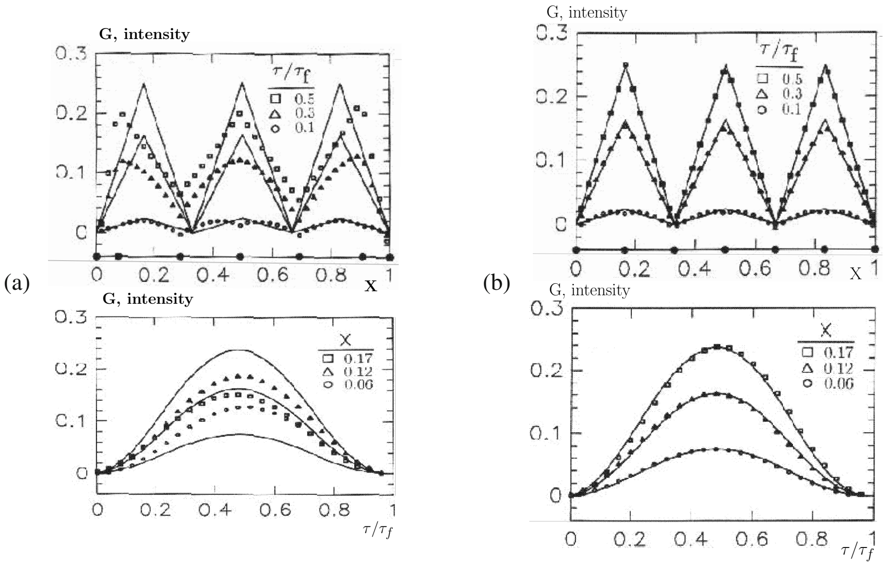


Fig. 7.5 Estimation of a space and time dependent volumetric heat source using seven temperature sensors (sensors positions are indicated by black bullets along X axis): (a) badly positioned – spatial variation and temporal variation; (b) well positioned – spatial variation and temporal variation

Exercises

7.1. An alternative way to derive Eq. (7.6) that is popular among engineers is by using Taylor's formula. Do it.

Hint. On the right hand side of Eq. (7.5), use the following Taylor's expansions (keeping only the first order terms),

$$T_m\left[g^n - \beta^n P^n\right] = T_m\left[g^n\right] - \frac{\partial T_m}{\partial g_n}\beta^n P^n \text{ , and}$$

$$T_m\left[g^n + P^n\right] = T_m\left[g^n\right] + \frac{\partial T_m}{\partial g_n} P^n \text{ .}$$

Use also $\Delta T_m\left[P^n\right] = T_m\left[g^n + P^n\right] - T_m\left[g^n\right]$.

7.2. Mimick the proof that Eq. (3.41d) can be written as Eq. (3.43d) to prove Eq. (7.7).
Hint. Lot's of work here!

7.3. (a) Show that the problem defined by Eq. (7.9) is linear in ΔT. That is, let $\Delta T[\delta g_i]$, $i = 1{,}2$, be the solution of Eq. (7.9) when the source is Δg_i. Let $k \in \mathbb{R}$. Show that

$$\Delta T[k\Delta g_1] = k\Delta T[\Delta g_1]\,, \text{ and}$$
$$\Delta T[\Delta g_1 + \Delta g_2] = \Delta T[\Delta g_1] + \Delta T[\Delta g_2]\,.$$

(b) Show that the solution of Eq. (7.8) is not linear with respect to g unless $T_0 = 0$. In this case, show that it is linear.

7.4. Check the details in the derivation of Eq. (7.14).
Hint. Show by integration by parts that

$$\int_0^{t_f} \lambda \frac{\partial \Delta T}{\partial t}\, dt = \lambda\ \Delta T|_{t=t_f} - \int_0^{t_f} \frac{\partial \lambda}{\partial t} \Delta T\, dt\,,$$
$$\int_0^{L} \lambda \frac{\partial^2 \Delta T}{\partial x^2}\, dx = \left(-\Delta T \frac{\partial \lambda}{\partial x}\right)\Bigg|_0^L + \int_0^L \frac{\partial^2 \lambda}{\partial x^2} \Delta T\, dx\,,$$

and the boundary and initial conditions of the sensitivity problem, Eq. (7.9) are used in the derivation. Show that, for appropriate choices of final values and boundary values for λ,

$$\int_0^{t_f}\int_0^L \lambda(x,t)\left[k\frac{\partial^2 \Delta T}{\partial x^2} - \rho c_p \frac{\partial \Delta T}{\partial t}\right] dx\, dt$$
$$= \int_0^{t_f}\int_0^L \left[k\frac{\partial^2 \lambda}{\partial x^2} + \rho c_p \frac{\partial \lambda}{\partial t}\right] \Delta T\, dx\, dt\,.$$

Also, assume that Dirac's delta 'function', $\delta = \delta(x)$ has the property that

$$\int_{-a}^{a} h(x)\delta(x)\, dx = h(0)\,,$$

for $a > 0$ and $h = h(x)$ a continuous function. Show that

$$\int_0^{t_f} \sum_{m=1}^{M} [T_m[g](t) - X_m(t)]\, \Delta T_m[\Delta g](t)\, dt$$
$$= \sum_{m=1}^{M} [T_m(x,t) - X(x,t)]\, \delta(x - x_m)$$

7.5. Perform the change of variables, Eq. (7.17), and determine the problem satisfied by

$$\bar{\lambda}(t) = \lambda(t_f - t)\,,$$

where λ satisfies Eq. (7.14).

7.6. Verify the derivation of Eq. (7.18) and show that it leads to Eq. (7.6).

7.7. In the heat conduction in a one-dimensional medium, with constant thermal properties, and insulated boundaries, the strength of two plane heat sources, $g_1 = g_1(t)$ and $g_2 = g_2(t)$, can be estimated simultaneously, from measurements made of the temperature at a certain number of interior points, [76]. Derive the sensitivity problem, the adjoint problem, and the gradients J'_{g_1} and J'_{g_2} for Alifanov's iterative regularization method, considering the perturbations $g_1 \to g_1 + \Delta g_1$ and $g_2 \to g_2 + \Delta g_2$, which leads to the perturbation $T \to T + \Delta T$.
Hint. Consider the conjugate coefficients

$$\gamma_{g_1}^n = \frac{\int_\Lambda \left[J'^{\,n}_{g_1}\right]^2 dr}{\int_\Lambda \left[J'^{\,n-1}_{g_1}\right]^2 dr} \quad \text{and} \quad \gamma_{g_2}^n = \frac{\int_\Lambda \left[J'^{\,n}_{g_2}\right]^2 dr}{\int_\Lambda \left[J'^{\,n-1}_{g_2}\right]^2 dr}$$

Chapter 8
A General Perspective

This chapter introduces a more general perspective of direct and inverse problems.

At the outset, we must observe that, from a mathematical standpoint, there is no reason to justify the distinction between direct and inverse problems. As pointed out by Keller [43], it is adequate to say that two problems are *mutually inverse* if the formulation of one involves partly or completely the solution of the other. Our presentation emphasises this. However, in applying the methodology of inverse problems to any particular physical problem, we can say that the model represents a *cause-effect* relationship and, in this case, finding that relationship is an inverse problem, determining the effect produced by a cause is a direct problem,while determining the cause by the effect it provokes is again an inverse problem. Also, when considering initial and boundary value problems for evolution equations — as those involving ordinary and partial differential equations or recurrence relations, — there is a standard distinction between cause and effect which leads to a classification of problems as direct and inverse problems. Therefore, to facilitate the discussion, we will continue calling some of them direct and others inverse. A more detailed discussion of the classification of inverse problems is presented.

We also consider the question of equivalence between gradient computation, based on its definition, and the more common computation by means of an operational rule using a Lagrangian function. This operational rule was used in Section 7.4. The theoretical derivation presented here of the operational rule serves as a proof in the case of Type I problems, but only as a guide for Types II to IV problems. For the latter, more sophisticated mathematical considerations have to be carried out because they involve infinite dimensional spaces [44, 29, 66].

8.1 Inverse Problems and Their Types

We consider a somewhat more general presentation of inverse problems. It has the ingredients already discussed in Chapter 1, and in other chapters. Here, we just mention the most basic mathematical structures needed. To treat the problems in a rigorous mathematical fashion, more complex concepts are called for, which, typically, are problem-dependent. We do not discuss that. We proceed with some examples, to illustrate the distinct pieces of the formulation and their applicability.

8.1.1 Classification of Problems

Let $\mathcal{U}$, $\mathcal{V}$, $\mathcal{T}$, and $\mathcal{Y}$ be normed vector spaces with inner product[1]. We think of these sets as

[1] These notions are recalled in the exercises of Appendix A.

$\mathcal{U}$ — ‘space of fields’;

$\mathcal{V}$ — ‘space of sources’;

$\mathcal{T}$ — ‘space of physical properties of a class of mathematical models’;

$\mathcal{Y}$ — ‘auxiliary space’.

All these spaces are ‘glued’ together by a function

$$\mathcal{G} : \mathcal{U} \times \mathcal{V} \times \mathcal{T} \to \mathcal{Y} ,$$

constituting a model of a cause-effect relationship. We define[2,3]

P_1: ***Direct problem*** Given a source $b \in \mathcal{V}$, and a model specified by $p \in \mathcal{T}$, determine the field $x \in \mathcal{U}$ such that

$$\mathcal{G}(x,b,p) = 0 . \tag{8.1}$$

The solution of the direct problem, Eq. (8.1), is denoted by $\mathcal{X}^p(b)$, and $\mathcal{X} : \mathcal{V} \times \mathcal{T} \to \mathcal{U}$, with $\mathcal{X}(b,p) = \mathcal{X}^p(b)$. It is the *solution operator* of the direct problem. We can, therefore, write

$$\mathcal{G}(\mathcal{X}^p(b),b,p) = 0 . \tag{8.2}$$

With this notation, we can represent the *black box*, Fig. 1.1, as

$$\text{input: } b \quad \xrightarrow{\mathcal{X}^p} \quad \text{output: } x = \mathcal{X}^p(b) .$$

Note that the direct problem is problem P_1, described in Section 2.8: given stimuli (sources), determine reactions (fields).

P_2: ***Inverse reconstruction problem*** Having a specific model by knowing $p \in \mathcal{T}$, and given a field $x \in \mathcal{U}$, determine a source $b \in \mathcal{V}$, satisfying Eq. (8.1).

This is problem P_2 described in Section 2.8: given reactions, determine stimuli.

P_3: ***Inverse identification problem*** Given a field $x \in \mathcal{U}$, and source $b \in \mathcal{V}$, determine the physical properties of the system $p \in \mathcal{T}$, still satisfying Eq. (8.1).

This is problem P_3 described in Section 2.8: given stimulus and corresponding reactions, pinpoint the properties of the model. We remark that, in fact, this general formulation makes virtually no real distinction between direct and inverse problems. All of them refer to the same equation, Eq. (8.1).

[2] We think of the elements of $\mathcal{V}$ as causing an effect belonging to $\mathcal{U}$ which is the set of effects. A specific cause-effect relationship is established by $\mathcal{G}$ and a particular element of $\mathcal{T}$.

[3] We use the same notation, P_1 to P_3, already introduced in Chapters 1 and 2, since the concepts presented here are just generalizations of the ones presented there.

The issue of characterizing a model appropriate for a particular physical situation is completed when the sets $\mathcal{U}$, $\mathcal{V}$, $\mathcal{T}$, and $\mathcal{Y}$, and a function $\mathcal{G}$ relating them (explicit or implicitly defined), are chosen.

Most of the times we are thinking about direct problems consisting of initial and boundary value problems for partial differential equations. This is why we say that b is the *source*, or the *forcing term*, and it could also indicate an initial and/or a boundary condition, p represents physical properties of the system, and x is the *field*. We remark, though, that the analysis presented here does not limit itself to this case.

Example 8.1. Heat equation. We illustrate these problems by the *initial value problem* for the *heat equation* with *Robin's boundary condition* in a material domain $\Omega \subset \mathbb{R}^n$,

$$\rho c_p \frac{\partial u}{\partial t} = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(k_{ij}\frac{\partial u}{\partial x_j}\right) + f\ , \text{ for } \mathbf{x} \in \Omega, t > 0\ ,$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x})\ , \text{ for } \mathbf{x} \in \Omega \text{ (initial condition)}\ ,$$

$$\alpha u + \beta\frac{\partial u}{\partial n} = g, \text{ for } \mathbf{x} \in \partial\Omega \text{ and } t > 0 \text{ (Robin's boundary condition)}\ .$$

Here,

$$\Omega \times [0, +\infty[\ni (\mathbf{x},t) \quad \mapsto \quad u(\mathbf{x},t) \in \mathbb{R}\ ,$$

is the *temperature distribution*, $\rho = \rho(\mathbf{x})$ is the specific mass of the material, $c_p = c_p(\mathbf{x})$ is the *specific heat* of the material, $K = K(\mathbf{x}) = (k_{ij})_{i,j=1,\ldots,n}$ is the *thermal conductivity tensor*, $f = f(\mathbf{x},t)$ is an internal volumetric heat source/sink, α and β are material properties, such as the convection heat transfer coefficient, defined at the boundary, $\mathbf{n}$ is the exterior normal vector to the boundary of Ω, $\frac{\partial}{\partial n} = \mathbf{n} \cdot \nabla$ is the exterior normal derivative, and g is a boundary heat source/sink.

One possibility is to write Robin's boundary condition as

$$\alpha(u - u_{amb}) + \beta\frac{\partial u}{\partial n} = 0\ ,$$

that is, $g = \alpha u_{amb}$, where u_{amb} refers to the external temperature of the body.

Recall the *black box* mental prototype discussed in Chapter 1. In the present example, the *input* (source term), b, has three components, the initial condition u_0, and the heat sources/sinks, f and g, respectively, in the interior of the body, and at its boundary. The *output* (field), x, is given by the temperature distribution u. The class of models, $\mathcal{F} = \mathcal{F}_p$, can be parametrized by p, which corresponds to the *properties of the model*, and is given by various mathematical objects, Ω, ρ, c_p, K, α and β.

More explicitly, in this heat equation problem, the source term, b, consists of

$$b = (u_0,f,g) \in \mathcal{V}\ ,$$

where

$$\mathcal{V} = C^0(\Omega, \mathbb{R}) \times C^0(\Omega\times]0, +\infty[, \mathbb{R}) \times C^0(\partial\Omega\times]0, +\infty[,\mathbb{R})\ ,$$

and, for instance,

$$C^0(\partial\Omega\times]0,+\infty[,\mathbb{R})\,,$$

represents the set of continuous functions defined on $\partial\Omega\times]0,+\infty[$, with values in $\mathbb{R}$. Likewise, C^2 represents functions with continuous second order derivatives.

The field (generically represented by x) would be u,

$$u\in\mathcal{U}=C^2(\Omega\times]0,+\infty[,\mathbb{R})\cap C^0(\overline{\Omega}\times[0,+\infty[,\mathbb{R})\,,$$

where $\overline{\Omega}$ stands for the closure of Ω, $\overline{\Omega}=\Omega\cup\partial\Omega$.

Next, the physical properties are represented by

$$p=(\Omega,\rho,c_p,K,\alpha,\beta)\in\mathcal{T}\,,$$

where,

$$\begin{aligned}\rho&\in C^0(\Omega\times]0,+\infty[,\mathbb{R})\,,\\ c_p&\in C^0(\Omega\times]0,+\infty[,\mathbb{R})\,,\\ K&\in C^0(\Omega\times]0,+\infty[,M_{3\times3})\,,\end{aligned}$$

and $M_{3\times3}$ is the set of real 3×3 matrices. Also,

$$\begin{aligned}\alpha&\in C^0(\partial\Omega\times]0,+\infty[,\mathbb{R})\,,\text{ and}\\ \beta&\in C^0(\partial\Omega\times]0,+\infty[,\mathbb{R}^3)\,.\end{aligned}$$

One way to handle the set of regions Ω as a subset of an inner-product space, as required by, for instance, the space of physical properties, $\mathcal{T}$ in Eq. (8.1), is to consider each subset Ω as represented by is *characteristic* or indicator function,

$$\chi_\Omega(x)=\begin{cases}1, & \text{if } x\in\Omega\\ 0, & \text{otherwise}\end{cases}\,.$$

Now, the operator $\mathcal{G}:\mathcal{U}\times\mathcal{V}\times\mathcal{T}\to\mathcal{Y}$ would have the form $\mathcal{G}=(\mathcal{G}_1,\mathcal{G}_2,\mathcal{G}_3)$, with

$$\begin{aligned}\mathcal{G}_1&=\rho c_p\frac{\partial u}{\partial t}-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(k_{ij}\frac{\partial u}{\partial x_j}\right)-f\in C^0(\Omega\times]0,+\infty[,\mathbb{R})\,,\\ \mathcal{G}_2&=u(\mathbf{x},0)-u_0(\mathbf{x})\in C^0(\Omega,\mathbb{R})\,,\\ \mathcal{G}_3&=\alpha u+\beta\frac{\partial u}{\partial n}-g\in C^0(\partial\Omega\times]0,+\infty[,\mathbb{R})\,.\end{aligned}$$

Finally, the equation would be written as

$$\mathcal{G}(u,b,p)=(\mathcal{G}_1(u,b,p),\mathcal{G}_2(u,b,p),\mathcal{G}_3(u,b,p))=0\,,$$

where the zero on the right hand side is a special zero,

$$0\in C^0(\Omega\times]0,+\infty[,\mathbb{R})\times C^0(\Omega,\mathbb{R})\times C^0(\partial\Omega\times]0,+\infty[,\mathbb{R})\,.$$

■

8.1.2 Observation Operators

One major use of inverse problems is to model data coming from real experiments. In order to understand this better, it is convenient to model the data acquisiton process, and we shall do this right away.

Let $C : \mathcal{U} \to \mathcal{W}$ be a function where $\mathcal{W}$ is a vector space with inner product $\langle \cdot , \cdot \rangle$, and $\mathcal{U}$ is the space of fields. Given $x \in \mathcal{U}$, we say that $z = C(x)$ is an *observation* of x, and that C is an *observation operator*. Notice that we can have, as a particular case, $\mathcal{W} = \mathcal{U}$ and C the identity function.

Example 8.2. One or several measurements. Let $T = T(a,t)$, $a \in \mathbb{R}$, $t \in [0,\infty)$, be the temperature distribution in a one dimensional, isolated, bar. As time advances, the temperature at a fixed point, say $a = 2$, represented by the function $\tau(t) = T(2,t)$, is an observation of T. The observation operator corresponding to the situation we just described represents the observations depending on the possible temperature fields, and is given by function Γ,

$$\begin{array}{rcl} C^0(\mathbb{R}\times[0,\infty),\mathbb{R}) & \xrightarrow{\Gamma} & C^0([0,\infty),\mathbb{R}) \\ T & \mapsto & \tau = \Gamma[T]\,, \end{array}$$

with

$$\begin{array}{rcl} [0,\infty) & \xrightarrow{\Gamma[T]} & \mathbb{R} \\ t & \mapsto & \tau(t) = \Gamma[T](t) = T(2,t)\,. \end{array}$$

We remark that the notation is quite flexible and can be worked out to accommodate one or several measuring points. If one measures the temperature at two points, say $a = 2$, and $a = 5$, it is sufficient to let $\Gamma[T] \in C^0([0,\infty),\mathbb{R}^2)$ and $\Gamma[T](t) = (T(2,t), T(5,t))$,

$$[0,\infty[\ni t \mapsto \tau(t) = \Gamma[T](t) = (T(2,t), T(5,t)) \in \mathbb{R}^2\,.$$

■

Example 8.3. Discrete time observation operator. With $T = T(a,t)$ as in the previous example, the sequence

$$\tau_n \quad = \quad T(2,n)\ ,\ \text{for } n \in \mathbb{N}\,,$$

is an observation of T. It means that one is measuring the temperature at position $x = 2$ and at times 1, 2, 3, ... The observation operator $\mathcal{O}$ is given by

$$\begin{array}{rcl} C^0(\mathbb{R}\times[0,\infty),\mathbb{R}) & \xrightarrow{\mathcal{O}} & s(\mathbb{N}) \\ T & \mapsto & \tau = \mathcal{O}[T]\,, \end{array}$$

where $s(\mathbb{N})$ is the set of sequences of real numbers,

$$s(\mathbb{N}) = \{(x_n)_{n\in\mathbb{N}},\ \text{such that } x_n \in \mathbb{R},\ \text{for all } n \in \mathbb{N}\}\,,$$

and

$$\begin{array}{rcl} \mathbb{N} & \xrightarrow{\mathcal{O}[\mathcal{T}]} & \mathbb{R} \\ n & \mapsto & \tau_n = \mathcal{O}[T](n) = T(2,n)\,. \end{array}$$

■

Example 8.4. Blurring operator. Another example of an observation operator would be the *blurring* operator, defined in Eq. (4.1), p. 86.

Let $M_{L\times M}$ be the set of real $L \times M$ matrices, and

$$\begin{array}{rcl} M_{L\times M} & \xrightarrow{\beta} & M_{L\times M} \\ \mathbf{I} & \rightarrow & \beta(\mathbf{I}) = \mathbf{Y} \end{array}$$

where

$$Y_{ij} = \sum_{i'=1}^{L} \sum_{j'=1}^{M} B_{ij}^{i'j'} I_{i'j'}\ , \quad i = 1,\dots,L,\ j = 1,\dots,M\,.$$

■

8.1.3 Inverse Problems in Practice

There is a slight difference between the theoretical formulation of inverse problems, presented in Section 8.1.1, and the way it has to be formulated to handle practical problems. This leads to another formulation of inverse problems, though still presented in mathematical terms. In practice we have to be more flexible or to settle for weaker goals.

Due to the necessity of solving problems taking into account the behaviour of the system under investigation, represented by the field x, and since it is captured by observations, the formulation has to include observation operators.

For instance, the theoretical inverse reconstruction problem to be solved can, therefore, be rephrased as: if the physical properties of the system, p, are known, and a field observation, $z = C(x)$, is known (but x is unknown), determine the source b. That is,

$$\begin{gathered} \text{given } \ z = C(x), \text{ and } p, \ \text{ determine } \ b \ \text{ such that} \\ \mathcal{X}^p(b) = x\,. \end{gathered} \tag{8.3}$$

Notice that since x is unknown, we cannot check whether Eq. (8.3) is satisfied or not. That is, unless C is an invertible function, it is impossible, in general, to solve this problem, or at least it is impossible to verify if it has been solved.

By applying the observation operator on both sides of Eq. (8.3), and since $z = C(x)$, we get the *null residual equation*,

$$C(\mathcal{X}^p(b)) = z\,. \tag{8.4}$$

This is an equation that can be verified.

It is worthwhile to emphasize that Eq. (8.4) is a consequence of Eq. (8.3). However, they are not equivalent. Having found b satisfying Eq. (8.4), it does not imply that Eq. (8.3) is satisfied. Nonetheless, we *ellect* the problem of knowing p and z, and wanting to determine b satisfying Eq. (8.4).

Given a source b and physical properties p, the *predicted observation* — due to the source b — is defined by $C(\mathcal{X}^p(b))$. We remark that, to compute the predicted observation, first the direct problem is solved for the source b, determining the field $\mathcal{X}^p(b)$, and then, to observe it, the observation operator C is applied to the field $\mathcal{X}^p(b)$, and $C(\mathcal{X}^p(b))$ is obtained.

The *theoretical residual* is defined by

$$r = C(\mathcal{X}^p(b)) - z\,,$$

and, for each b, it is a measure of how much b does not satisfy Eq. (8.4), if $r \neq 0$.

In practical applications, an inverse problem must be solved not with the knowledge of an observation, z, of the solution of the direct problem, but from experimental data, Z_{meas}. That is, in practice z is not known. Even so, the data is modeled as if it was obtained by means of an observation operator, when in fact it was not.

The *practical residual* is defined as

$$R = C(\mathcal{X}^p(b)) - Z_{meas}\,.$$

We have, therefore, two reasons to change what is understood as solution of the inverse problem:

(i) instead of knowing x, only $z = C(x)$ is known;

(ii) as a matter of fact, not even $z = C(x)$ is known, only Z_{meas} is, which, supposedly, is a measurement (or several), or an approximation, of z.

Thus, it is no longer possible to *interpolate* the data with the model to solve the inverse problem, as required by Eq. (8.3). There one needs the explicit knowledge of x, which is unavailable.

We define practical inverse problems:

P_2^*: **Practical inverse reconstruction problem** Given a measured observation of field x, denoted by Z_{meas}, and physical properties p, determine the source b^*, that minimizes half the *quadratic error function*, (or half the squared residuals)

$$\mathcal{V} \ni b \mapsto E[b] = \frac{1}{2}|R|^2 = \frac{1}{2}|C(\mathcal{X}^p(b)) - Z_{\text{meas}}|^2\,, \tag{8.5}$$

i.e., determine b^* such that

$$E[b^*] = \min_{b\in\mathcal{V}} E[b]\,.$$

Notice that the size of $C(\mathcal{X}^p(b)) - Z_{\text{meas}}$ is to be minimized, i.e., b^* must be chosen to minimize the difference between the predicted observation when the source b is assumed known, which is computed by $C(\mathcal{X}^p(b))$, and the experimental measurement, Z_{meas}, which is obtained experimentally.

P_3^*: **Practical inverse identification problem** Given a measured observation of field x, Z_{meas}, and a known source b, determining the physical properties p^*, that minimizes half the quadratic error function

$$\mathcal{T} \ni p \mapsto E[p] = \frac{1}{2}|C(\mathcal{X}^p(b)) - Z_{\text{meas}}|^2 \,. \tag{8.6}$$

This is problem P_3 described in Section 2.8: given measurements of stimuli and reactions, determine their relationship.

8.1.4 Domain of Dependence

We intend to distinguish different types of operators by using the concept of *domain of dependence*, and for that we begin by discussing three simple examples.

Example 8.5. Let $\mathcal{B}^0(\mathbb{R},\mathbb{R})$ be the set of continuous and bounded functions defined on $\mathbb{R}$ with values on $\mathbb{R}$. That is, $f \in \mathcal{B}^0(\mathbb{R},\mathbb{R})$ if and only if $f \in C^0(\mathbb{R},\mathbb{R})$ and there is $K \in \mathbb{R}$ such that $|f(x)| \leq K$, for all $x \in \mathbb{R}$.

Consider a function $\mathcal{H}$ from $\mathcal{B}^0(\mathbb{R},\mathbb{R})$ into itself,

$$\begin{aligned} \mathcal{H} : \mathcal{B}^0(\mathbb{R},\mathbb{R}) &\to \mathcal{B}^0(\mathbb{R},\mathbb{R}) \\ f &\mapsto \mathcal{H}[f] = g \,, \end{aligned}$$

defined by

$$\mathcal{H}[f](x) = g(x) = \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f(s)\, ds \,.$$

Roughly, we want to say that the domain of dependence of the image of f at x, that is, the domain of dependence of $\mathcal{H}[f](x) = g(x)$, denoted by $\mathcal{D}_g(x)$, is the subset of the domain of f where the values of f influence the values of $g(x)$. In this case,

$$\begin{aligned} \mathcal{D}_{\mathcal{H}[f]}(x) &= \{s \in \mathbb{R} \text{ such that the value of } f(s) \text{ affects } \mathcal{H}[f](x)\} \\ &= \left[x - \frac{1}{2}, x + \frac{1}{2}\right] . \end{aligned}$$

■

The previous example has the particularity of f and $g = \mathcal{H}[f]$ sharing the same domain. We overthrow this special case by presenting a finite dimension example. For this, we need to interpret $\mathbb{R}^n$ as the function space $\mathcal{F}_n$.

Denote by $\mathcal{F}_n$ the set of functions defined on the first n integers with values in $\mathbb{R}$, that is, $\mathcal{F}_n = \{f \mid f : \{1,2,\ldots,n\} \to \mathbb{R}\}$. As we shall see, $\mathcal{F}_n$ is essentially the same as $\mathbb{R}^n$. Consider the function

$$\begin{aligned} \mathcal{L} : \mathcal{F}_n &\to \mathbb{R}^n \\ f &\mapsto \mathcal{L}(f) = (f(1),f(2),\ldots,f(n)) \,. \end{aligned} \tag{8.7}$$

Note that $\mathcal{F}_n$ and $\mathbb{R}^n$ are vector spaces and that $\mathcal{L}$ is a *linear* function between them, that is,

$$\mathcal{L}(f+g) = \mathcal{L}(f) + \mathcal{L}(g)\,,\ \text{for all } f,g \in \mathcal{F}_n \tag{8.8a}$$

and

$$\mathcal{L}(kf) = k\mathcal{L}(f)\,,\ \text{for all } k \in \mathbb{R}, f \in \mathcal{F}_n\,. \tag{8.8b}$$

Moreover, $\mathcal{L}$ has an inverse,

$$\begin{aligned} \mathcal{M} : \mathbb{R}^n &\to \mathcal{F}_n \\ \mathbf{x} &\mapsto \mathcal{M}[\mathbf{x}] = f\,, \end{aligned} \tag{8.9a}$$

with

$$f(i) = x_i\,,\ \text{for } i = 1,2,\ldots,n\,. \tag{8.9b}$$

One can show that $\mathcal{L} \circ \mathcal{M}(\mathbf{x}) = \mathbf{x}$ and $\mathcal{M} \circ \mathcal{L}(f) = f$. We denote the inverse by $\mathcal{M} = \mathcal{L}^{-1}$. The function $\mathcal{L}$ is called an *isomorphism* between vector spaces since it is invertible, Eq. (8.9), and is linear, Eq. (8.8). In a coloquial way, we interpret $\mathbb{R}^n$ as a set of real valued functions defined on the finite set of the first n integers, $\{1,2,\ldots,n\}$.

Example 8.6. Let $h : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by

$$h(x,y,z) = \begin{pmatrix} 3x+2y+z \\ x+z \end{pmatrix}.$$

We want to discuss an appropriate notion of domain of dependence. For that, we reinterpret this function as a function between function spaces

$$\begin{aligned} \tilde{h} : \mathcal{F}_3 &\to \mathcal{F}_2 \\ f &\mapsto \tilde{h}[f] = g \end{aligned}$$

where

$$\begin{aligned} g(1) &= \tilde{h}[f](1) = 3f(1) + 2f(2) + f(3)\,, \\ g(2) &= \tilde{h}[f](2) = f(1) + f(3)\,. \end{aligned}$$

Now, it seems plausible to define

$$\mathcal{D}_{\tilde{h}[f]}(1) = \mathcal{D}_g(1) = \{1,2,3\}\,, \text{and } \mathcal{D}_{\tilde{h}[f]}(2) = \mathcal{D}_g(2) = \{1,3\}\,.$$

■

Note that in the previous example, f and $g = h[f]$ do not share the same domain. The same kind of affairs can show up in an infinite dimensional setting.

Example 8.7. Wave equation. Let $\mathcal{B}^0(\mathbb{R}\times[0,+\infty[,\mathbb{R})$ be the set of real valued, bounded, continuous functions defined on $\mathbb{R}\times[0,+\infty[$. Consider the function $\mathcal{W}$,

$$\begin{array}{rcl}\mathcal{W}:\mathcal{B}^0(\mathbb{R},\mathbb{R}) & \to & \mathcal{B}^0(\mathbb{R}\times[0,+\infty[,\mathbb{R})\\ v & \mapsto & \mathcal{W}[v]=u\end{array}$$

defined by

$$u(x,t)=\mathcal{W}[v](x,t)=\frac{1}{2}\int_{x-ct}^{x+ct}v(s)\,ds\,. \tag{8.10}$$

The domain of dependence, in this case, is given by

$$\mathcal{D}_{\mathcal{W}[v]}(x,t)=\mathcal{D}_u(x,t)=[x-ct,\;x+ct]\;.$$

It just happens that u, defined by Eq. (8.10), is the solution of the wave equation, $u_{tt}=c^2u_{xx}$, satisfying the initial conditions, $u(x,0)=0$ and $u_t(x,0)=v(x)$. ■

In order to be able to express more precisely what is to be meant by domain of dependence of the answer to a problem, we need to introduce some notation.

Let Ω_d be either $\{1,2,\ldots,k\}$ or an open subset[4] of $\mathbb{R}^k$. Likewise, let also Ω_a be such kind of set.

To make things not too technical, let $U=C_0^\infty(\Omega_d,\mathbb{R}^m)$ and $V=C_0^\infty(\Omega_a,\mathbb{R}^n)$, where C_0^∞ stands for functions infinitely differentiable with compact support[5]. The sets U and V are function spaces. If $\Omega_d=\{1,2,\ldots,k\}$ then U is just k copies of $\mathbb{R}^m$, $U=\mathbb{R}^m\times\mathbb{R}^m\times\ldots\times\mathbb{R}^m=(\mathbb{R}^m)^k$.

Let τ be an operator between these spaces,

$$\tau:U\to V\,,$$

which shall play the role of the solution operator of a certain problem.

We want to grasp what influences an answer to a problem, the domain of dependence. That is the main, very roughly prescribed, goal. We shall do that by first determining which places do not interfere with the answer.

Let A be an open subset of Ω_d. Then[6], $C_0^\infty(A,\mathbb{R}^m)\subset C_0^\infty(\Omega_d,\mathbb{R}^m)$.

[4] We recall that a set $\Omega\subset\mathbb{R}^m$ is *open* if for every point $\mathbf{a}\in\Omega$, there is an open ball centered around $\mathbf{a}$ fully contained in Ω, i.e. there is $r>0$ such that if $|\mathbf{x}-\mathbf{a}|<r$ then $\mathbf{x}\in\Omega$. A set $F\subset\mathbb{R}^m$ is *closed* if its complement, $\mathbb{R}^m\backslash F$, is open. A set $K\subset\mathbb{R}^m$ is called *compact* if it is closed and bounded, and a set K is called *bounded* if there is a number $l\in\mathbb{R}$ such that $|x|\leq l$, for all $x\in K$. For further discussion see [53].

[5] We recall that the *support* of a function f is the smallest closed set containing all the points x where $f(x)\neq 0$. Equivalently, it is the complement of the open set defined by the union of all open sets where f is not null in any point — the complement of the largest set where f is different from zero. (Qualitatively, the support is to be the set where all the 'action' occurs.)

[6] Strictly speaking, given $\alpha\in C_0^\infty(A,\mathbb{R}^m)$, it is not true that $\alpha\in C_0^\infty(\Omega_d,\mathbb{R}^m)$ because the domain of definition of α is A and not Ω_d. Nontheless, given α, we can construct an extension of α defined on Ω_d, which we denote by $\tilde\alpha\in C_0^\infty(\Omega_d,\mathbb{R}^m)$, given by $\tilde\alpha(x)=\alpha(x)$ for all $x\in A$, and $\tilde\alpha(x)=0$ for all $x\in\Omega_d\backslash A$. By abuse of notation, instead of using $\tilde\alpha$, we say that $\alpha\in C_0^\infty(\Omega_d,\mathbb{R}^m)$.

Let $u \in U$, $y \in \Omega_a$, and $v = \tau[u]$. Assume $A \subset \Omega_d$ be an open set such that

$$v(y) = \tau[u + \phi](y)\,,$$

for all $\phi \in C_0^\infty(A, \mathbb{R}^m) \subset C_0^\infty(\Omega_d, \mathbb{R}^m)$. We say that the values of u on A do not affect the value of $\tau[u]$ on $y \in \Omega_a$. Let Λ_y be the union of all such open sets $A \subset \Omega_d$, that is, Λ_y is the largest such open set. We define the *domain of dependence* of $v = \tau[u]$ at $y \in \Omega_a$ as the closed set given by

$$\mathcal{D}_{\tau[u]}(y) = \mathcal{D}_v(y) = \Omega_d \backslash \Lambda_y\,.$$

This definition is compatible with the examples already presented, and work as well, in a simple way, when Ω_d is a finite set. Nonetheless, it is unable to grasp some fundamental information, as we shall see by example.

Consider Exercises 8.3 and 8.4. In both cases, the domain of dependence is $\{0\} \subset \mathbb{R}$, a finite set. However, in order to be able to compute $\beta[f] = f'(0)$, in Exercise 8.4, one needs the information of f in a neighbourhood of 0, no matter how small, and not just the value of f in 0. This means that we need an infinite amount of information.

When the operator τ is such that $\mathcal{D}_{\tau[u]}(y) \subset \Omega_d$ is an infinite set for at least a $y \in \Omega_a$, we say that τ is a *global* operator. When the domain of dependence $\mathcal{D}_{\tau[u]}(y) \subset \Omega_d$ is a finite set for all $y \in \Omega_a$, we have two possibilities. If we need to have the knowledge of u in a small neighbourhood of at least a point in Ω_d, then we say that τ is a *local* operator. Otherwise, it is a pointwise operator.

We rewrite these ideas. Using the notation previously introduced, let $v = \tau[u]$ and $y \in \Omega_a$. We say that τ is

(a) *pointwise operator* if $v(y)$ depends on the values of u in a finite number of points;

(b) *local operator* if $v(y)$ depends on the values of u in a neighborhood of a finite numbers of points in its domain;

(c) *global operator* if $v(y)$ depends on the values of u in a way that cannot be classified in items (a) or (b).

Example 8.8. The operator Λ given by

$$\begin{aligned} \Lambda : C_0^1(\mathbb{R}, \mathbb{R}) &\to C_0^1(\mathbb{R}, \mathbb{R}) \\ f &\mapsto \Lambda[f] = g \end{aligned}$$

where

$$\Lambda[f](x) = g(x) = \frac{f(x) + f(-x)}{2}\,,$$

is a pointwise operator. ■

8.1.5 Types of Inverse Problems

In the framework described in this chapter, the classification presented in Section 2.8, based on the dimension of the model and of the estimated quantity can be clarified a little more.

In the case of the dimension of the estimated quantity, we have

- In the inverse reconstruction problem, the dimension of the estimated properties clearly is the dimension of the input space (space of sources), $\mathsf{dim}\,(\mathcal{V})$;
- In the inverse identification problem, the dimension of the estimated properties clearly is the dimension of the property space, $\mathsf{dim}\,(\mathcal{T})$.

We say that the model is of *infinite* dimension if the problem involves global or local operators, and we say that it is of *finite* dimension if it involves pointwise operators.

For convenience, we recall here Table 2.3 of inverse problems type:

Type I Estimation of a finite number of parameters in a model of finite dimension;

Type II Estimation of a infinite number of parameters or a function in a model of finite dimension;

Type III Estimation of an finite number of parameters in a model of infinite dimension;

Type IV Estimation of an infinite number of parameters or a function in a model of infinite dimension.

Example 8.9. Dimension of the model

(a) Let $x = x(t)$, $p = p(t)$, and $b \in \mathbb{R}$ and consider the problem

$$\begin{aligned} x' &= px\,, \text{ for } t > 0\,, \\ x(0) &= b\,, \end{aligned}$$

If $p = p(t)$ and measurements of x are known, determining b is a type III inverse reconstruction problem. Whereas, if b and measurements of x are known, to determine $p = p(t)$ is a type IV inverse identification problem, whereas if p is a constant, its determination is a type III inverse identification problem.

(b) Let $px = b$, with $p, x, b \in \mathbb{R}$. Assume b is known, and measurements of x are also known. The determination of p is a type I inverse identification problem.

(c) Let $p(t)x(t) = b(t)$, for $t \in \mathbb{R}$, where p, x, and b are real functions of a real variable. Assume $b(t)$ is known, and measurements of $x(t)$ are also known. The determination of $p = p(t)$ is a type II inverse identification problem.

■

8.2 Gradient Calculation

In this section we begin with some considerations on the computation of the gradient, and then establish an operational rule, in general, which has been used in Section 7.4.

8.2.1 Reformulation of the Direct Problem

It is necessary, for the analysis we developed to establish the operational rule, that there is a function, $\mathcal{A}$,

$$\mathcal{A} : \mathcal{U} \times \mathcal{T} \quad \longrightarrow \quad \mathcal{V} ,$$

relating $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{T}$, in such a way that

$$\mathcal{G}(x,b,p) = b - \mathcal{A}(x,p) .$$

The direct problem[7] is therefore, rewritten as: given source $b \in \mathcal{V}$ and model specified by $p \in \mathcal{T}$, determine the field $x \in \mathcal{U}$ such that[8]

$$\mathcal{A}(x, p) = b . \tag{8.11}$$

Since $\mathcal{X}^p(b)$ is the solution of the direct problem, Eq. (8.11), is rewritten as

$$\mathcal{A}(X^p(b), p) = b .$$

We find it convenient to write $\mathcal{A}(x,p) = \mathcal{A}^p(x)$.

In the case of the heat equation problem, Example 8.1, it is fairly simple to define $\mathcal{A}$.

8.2.2 Gradient Calculation by Its Definition

Here, we obtain an expression for calculating the gradient of functional E, Eq. (8.5), from the definition of gradient [53]. Recall that the gradient of E is an element of $\mathcal{V}$. Initially we will compute the directional derivative (Gâteaux derivative) of functional E. Let b^ϵ be a curve in $\mathcal{V}$, parameterized by ϵ, with

$$b^0 = b \text{ , and } \left(d/d\epsilon\right)|_{\epsilon=0}\, b^\epsilon = \tilde{b} ,$$

as illustrated in Fig. 8.1.

For small values of ϵ, we may think of b^ϵ as a small perturbation of b, that is,

$$b^\epsilon \sim b + \epsilon\tilde{b} \text{ , when } \epsilon \ll 1 .$$

[7] The reader is advised to pay attention because $\mathcal{A}$ plays a different role from matrix A in Section 1.4. In the way the direct problem is formulated here, $\mathcal{A}$ corresponds to matrix A^{-1}.

[8] Notice that b is multiplied by 1. This is the form for the equivalence between the operational rule and the result using the definition of the gradient. The equation of heat transfer by conduction, Eq. (7.8a), was written to meet this condition.

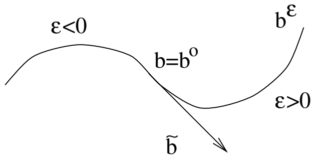


Fig. 8.1 Curve parameterized by parameter ϵ. It's tangent vector at b is $\tilde{b}$.

If we denote by $dE_b[\tilde{b}]$ the *directional derivative* of E at b in the direction $\tilde{b}$, then by the chain rule

$$dE_b[\tilde{b}] = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} E[b^\epsilon] = \langle C(\mathcal{X}^p(b)) - Z_{\text{meas}}, dC_{\mathcal{X}^p(b)}\, d\mathcal{X}^p(b)\, \tilde{b}\rangle\,. \tag{8.12}$$

Here, $dC_\mathcal{X}$ and $d\mathcal{X}^p(b)$ denote the derivatives of C and $\mathcal{X}$, evaluated, respectively, at $\mathcal{X}$ and b. The inner product $\langle\,,\,\rangle$ is defined in $\mathcal{W}$.

Recall that the *gradient* of functional E, evaluated at b, E'_b, is the element of $\mathcal{V}$ that represents the *Fréchet derivative* of E, dE_b, with respect to the inner product, [53], i.e., such that

$$dE_b[\tilde{b}] = \langle E'_b, \tilde{b}\rangle\,,\ \text{for all } \tilde{b} \in \mathcal{V}\,. \tag{8.13}$$

From Eqs. (8.12) and (8.13) we see that, while in Eq. (8.13) $\tilde{b}$ is alone in the second entry of the inner product, in Eq. (8.12) it is being acted upon, successively, by linear operators $d\mathcal{X}^p(b)$ and $dC_{\mathcal{X}^p(b)}$.

Given a linear operator

$$\mathcal{O} : W_1 \to W_2\,,$$

we denote by $\mathcal{O}^*$ the *adjoint operator*[9] of $\mathcal{O}$, i.e., the operator that switches places with $\mathcal{O}$ in the inner product, or, explicitly

$$\langle \mathcal{O}w_1, w_2\rangle = \langle w_1, \mathcal{O}^* w_2\rangle\,, \qquad \text{for all } \ w_1 \in W_1,\ w_2 \in W_2\,.$$

Using this concept we have, successively, from Eq. (8.12),

$$\begin{aligned} dE_b[\tilde{b}] &= \langle (dC_{\mathcal{X}^p(b)})^* \,[C(\mathcal{X}^p(b)) - Z_{\text{meas}}],\ d\mathcal{X}^p(b)\tilde{b}\rangle \\ &= \langle (d\mathcal{X}^p(b))^*\,(dC_{\mathcal{X}^p(b)})^*\,[C(\mathcal{X}^p(b)) - Z_{\text{meas}}],\ \tilde{b}\rangle\,. \end{aligned} \tag{8.14}$$

[9] The adjoint operator can be quite complex, [66], but for matrices it is simple. Given a real $n \times m$ matrix, M, the associated linear operator is

$$\mathbb{R}^m \ni \mathbf{x} \mapsto M\mathbf{x} \in \mathbb{R}^n\,.$$

Let $\langle\,,\,\rangle$ be the usual real scalar product. Then, the adjoint operator, M^*, corresponds, simply, to the transpose of M.

Comparing Eqs. (8.14) and (8.13), we obtain the following expression for the gradient of E at b

$$E'_b = (d\mathcal{X}^p(b))^* \left(dC_{\mathcal{X}^p(b)}\right)^* \left[C(\mathcal{X}^p(b)) - Z_{\text{meas}}\right] . \tag{8.15}$$

Equation (8.15) for the gradient of E at b is quite complicate and we shall interpret it next.

8.2.3 Interpretation of the Gradient: Sensitivity and Adjoint Problems

Denote $d\mathcal{X}^p(b)\tilde{b}$ by $\tilde{\mathcal{X}}$. The sensitivity problem, associated with the direct problem, Eq. (8.11), corresponds to the linearization of this problem, that is, it is the problem that $\tilde{\mathcal{X}}$ satisfies.

To obtain the linearized problem, it is enough to substitute in Eq. (8.11), b by b^ϵ and x by $\mathcal{X}(b^\epsilon)$, obtaining $\mathcal{A}^p(\mathcal{X}(b^\epsilon)) = b^\epsilon$, and to differentiate with respect to ϵ, in $\epsilon = 0$. We then have

$$(d/d\epsilon)|_{\epsilon=0}\, \mathcal{A}^p(\mathcal{X}^p(b^\epsilon)) = (d/d\epsilon)|_{\epsilon=0}\, b^\epsilon \ ,$$

where, by the chain rule, gives

$$d\mathcal{A}^p_{\mathcal{X}^p(b)} d\mathcal{X}^p(b)\tilde{b} = \tilde{b} \ . \tag{8.16}$$

Therefore, the problem that $\tilde{\mathcal{X}}$ satisfies, the sensitivity problem, is

$$d\mathcal{A}^p_{\mathcal{X}^p(b)}\tilde{\mathcal{X}} = \tilde{b} \ . \tag{8.17}$$

Given b, the solution of the direct problem $\mathcal{X}^p(b)$ is to be computed, and then $d\mathcal{A}^p_{\mathcal{X}^p(b)}$ is evaluated. The *sensitivity problem* is then: given $\tilde{b}$, determine $\tilde{\mathcal{X}}$, such that Eq. (8.17) is satisfied. Since

$$\tilde{\mathcal{X}} = d\mathcal{X}^p(b)\tilde{b} \ ,$$

it can be concluded that $d\mathcal{X}^p(b)$ represents the *solution operator* of the sensitivity problem, i.e., it is the inverse operator of $d\mathcal{A}^p_{\mathcal{X}^p(b)}$,

$$\left[d\mathcal{A}^p_{\mathcal{X}^p(b)}\right]^{-1} = d\mathcal{X}^p(b) \ .$$

This could also have been seen from Eq. (8.16).

Now, we note that the adjoint operator of the solution operator of the sensitivity problem, $(d\mathcal{X}^p(b))^*$, is used in Eq. (8.15). Moreover, it should be equal to the solution operator of the adjoint problem of the sensitivity problem[10], $\left\{\left[d\mathcal{A}^p_{\mathcal{X}^p(b)}\right]^*\right\}^{-1}$, [58, 73]. In fact,

$$[d\mathcal{X}^p(b)]^* = \left\{\left[d\mathcal{A}^p_{\mathcal{X}^p(b)}\right]^{-1}\right\}^* = \left\{\left[d\mathcal{A}^p_{\mathcal{X}^p(b)}\right]^*\right\}^{-1} . \tag{8.18}$$

In particular, the adjoint problem of a linear problem is the linear problem that corresponds to the adjoint operator of the linear operator that defines the linear problem.

Summarizing, from Eq. (8.18), to obtain the gradient, Eq. (8.15), it is sufficient to solve the following problem for y,

$$(d\mathcal{A}^p_{\mathcal{X}^p(b)})^* y = (dC_{\mathcal{X}^p(b)})^* [C(\mathcal{X}^p(b)) - Z_{\text{meas}}] , \tag{8.19}$$

from which we get the gradient, $E'_b = y$.

We remark that the source for problem defined by Eq. (8.19) is

$$(dC_{\mathcal{X}^p(b)})^* [C(X^p(b)) - Z_{\text{meas}}] ,$$

that depends on the difference between what was foretold and what was measured —discrepancy between the model and the real data.

To emphasize, we repeat that to determine the gradient of E in b, these steps are to be followed:

1.a Determine the solution of the direct problem, $\mathcal{X}^p(b)$;

1.b Determine the linear operator that defines the sensitivity problem, $d\mathcal{A}^p_{\mathcal{X}^p(b)}$;

1.c Determine the adjoint operator of the operator of the sensitivity problem, $(d\mathcal{A}^p_{\mathcal{X}^p(b)})^*$;

2.a Determine the observation, $C(\mathcal{X}^p(b))$ (prediction);

2.b Determine the residual, the difference between the predicted and measured values, $C(\mathcal{X}^p(b)) - Z_{\text{meas}}$;

3.a Determine the derivative of the observation operator at $\mathcal{X}^p(b)$, $dC_{\mathcal{X}^p(b)}$;

[10] If P denotes the operator of a linear problem, P^{-1} denotes the solution operator. The assertion of the paragraph can be rewritten as

$$(P^{-1})^* = (P^*)^{-1} .$$

Formally, this is shown as follows:

$$\begin{aligned}\langle u, (P^{-1})^* v\rangle &= \langle P^{-1}u, v\rangle = \langle P^{-1}u, P^*(P^*)^{-1}v\rangle \\ &= \langle PP^{-1}u, (P^*)^{-1}v\rangle = \langle u, (P^*)^{-1}v\rangle ,\end{aligned}$$

for all u and v. This derivation is correct for bounded operators in Hilbert spaces. Nonetheless, mathematical rigorousness is lacking if unbounded operators are involved [66], since the mathematics involved is much more refined.

3.b Determine the adjoint of the operator $dC_{\mathcal{X}^p(b)}$, $(dC_{\mathcal{X}^p(b)})^*$;

4.a Determine the source term for Eq. (8.19), $(dC_{\mathcal{X}^p(b)})^*[C(\mathcal{X}^p(b)) - Z_{\text{meas}}]$;

4.b Solve, for y, the adjoint problem, Eq. (8.19).

In essence, to determine the gradient, it suffices to solve the adjoint problem of the sensitivity problem, with a source that, essentially, depends on how different the prediction is from the measurement.

8.2.4 Operational Rule for the Computation of the Gradient

Using the notation previously established, we consider the minimization of

$$E[b] = \frac{1}{2}\left|C(\mathcal{X}^p(b)) - Z_{\text{meas}}\right|^2 ,$$

subjected to the constraint

$$\mathcal{G}(\mathcal{X}^p(b),b) = 0 .$$

Here, differently from Eq. (8.1), we let $\mathcal{G}$ ba a function defined only in $\mathcal{U} \times \mathcal{V}$, where, we recall, $\mathcal{U}$ is the space of fields, and $\mathcal{V}$ is the space of sources.

We reformulate this problem as an unconstrained minimization problem by means of a *Lagrange multiplier*, denoted by λ, to handle the constraint. We define the following *Lagrangian* function

$$\begin{aligned} \mathcal{L} : \mathcal{V} \times \mathcal{V} \quad &\to \quad \mathbb{R} \\ (b, \lambda) \quad &\mapsto \quad \mathcal{L}[b,\lambda] = \frac{1}{2}\left|C(\mathcal{X}^p(b)) - Z_{\text{meas}}\right|^2 + \langle \lambda, \mathcal{G}(\mathcal{X}^p(b), b)\rangle . \end{aligned}$$

Then, for the curve b^ϵ,

$$\begin{aligned} \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \mathcal{L}[b^\epsilon, \lambda] &= \langle C(\mathcal{X}^p(b)) - Z_{\text{meas}}, dC_{\mathcal{X}^p(b)} d\mathcal{X}^p(b)\tilde{b}\rangle \\ &\quad + \langle \lambda, \tilde{b} - d\mathcal{A}^p_{\mathcal{X}^p(b)} d\mathcal{X}^p(b)\tilde{b}\rangle \\ &= \left\langle (dC_{\mathcal{X}^p(b)})^*[C(\mathcal{X}^p(b)) - Z_{\text{meas}}] - (d\mathcal{A}^p_{\mathcal{X}^p(b)})^*\lambda \, , d\mathcal{X}^p(b)\tilde{b}\right\rangle + \\ &\quad + \langle \lambda, \tilde{b}\rangle . \end{aligned} \tag{8.20}$$

On the other hand, by definition, $\mathcal{X}(b^\epsilon)$ satisfies the direct equation

$$\mathcal{A}^p(\mathcal{X}(b^\epsilon)) = b^\epsilon .$$

Hence, it can be concluded that $\mathcal{G}(\mathcal{X}(b^\epsilon)), b^\epsilon) = 0$. Thus, $E[b^\epsilon] = \mathcal{L}[b^\epsilon, \lambda]$, regardless the value of λ. It follows that

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \mathcal{L}[b^\epsilon, \lambda] = \langle E'_b, \tilde{b}\rangle . \tag{8.21}$$

It seems useless to consider the function $\mathcal{L}$. However, it suggests an operational rule to determine the equation that the gradient of E satisfies. Note that λ is arbitrary. If we choose λ to nullify the first factor of the first inner product on the right side of Eq. (8.20), i.e., if we impose that λ satisfies the following equation,

$$(d\mathcal{A}^p_{\mathcal{X}^p(b)})^*\lambda = (dC_{\mathcal{X}^p(b)})^*[C(\mathcal{X}^p(b)) - Z_{\text{meas}}]\,, \tag{8.22}$$

and if we compare Eqs. (8.21) and (8.20), we reach the conclusion that $\lambda = E'_b$. This is no news since Eq. (8.22) is the equation for the gradient, as already seen in Eq. (8.19).

One can wonder what is the advantage of this approach, and the answer is 'the operational rule'. To obtain Eq. (8.22), that tells us what is the problem to be solved to determine the gradient of E, i.e., the adjoint equation, we compute, as indicated in Eq. (8.20),

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \mathcal{L}[b^\epsilon,\lambda]\,,$$

and arrange the resulting terms in two groups. The first one is $\langle\lambda,\tilde{b}\rangle$. The second one comprising the rest, should be made null. When we impose that the rest is to be null, we obtain Eq. (8.22), without consciously considering all the stages that lead to the derivation of Eq. (8.19), and which were detailed in the previous section. This operational rule was used in Section 7.4.

Exercises

8.1. Show that $\mathcal{L}$ in Eq. (8.7) is an isomorphism, i.e, verify that it is linear, Eq. (8.8) and that $\mathcal{M}$ in Eq. (8.9) is its inverse, i.e., $\mathcal{L}\circ\mathcal{M}(\mathbf{x}) = \mathbf{x}$ and $\mathcal{M}\circ\mathcal{L}(f) = f$.

8.2. Heat equation. Consider the operator

$$\begin{aligned} \mathcal{H} : \mathcal{B}^0(\mathbb{R},\mathbb{R}) &\rightarrow \mathcal{B}^0(\mathbb{R}\times]0,+\infty[,\mathbb{R}) \\ u_0 &\mapsto u = \mathcal{H}[u_0] \end{aligned}$$

where

$$u(x,t) = \mathcal{H}[u_0](x,t) = \frac{1}{\sqrt{4k\pi t}}\int_{-\infty}^{+\infty} e^{-\frac{(x-y)^2}{4kt}}\,u_0(y)\,dy\,.$$

Here u is the solution of the heat equation $u_t = ku_{xx}$, with initial condition $u(x,0) = u_0(x)$. Show that the domain of dependence is $\mathcal{D}_u(x,t) = \mathbb{R}$.

8.3. *Dirac's delta function.* Let

$$\begin{aligned} \delta : \mathcal{B}^0(\mathbb{R},\mathbb{R}) &\rightarrow \mathbb{R} \\ u &\mapsto v = \delta[u] = u(0) \end{aligned}$$

Set $\mathbb{R} = \mathcal{F}_1$. Show that the domain of dependence is $\mathcal{D}_{\delta[u]}(1) = \{0\}$.

8.4. Derivative. Let

$$\begin{aligned}\beta : C^1(\mathbb{R},\mathbb{R}) &\rightarrow \mathbb{R}\\ u &\mapsto v = \beta[u] = u'(0)\end{aligned}$$

Show that the domain of dependece is $\{0\} \subset \mathbb{R}$.

8.5. Dirac's delta function on a circle. Let

$$\begin{aligned}C : \mathcal{B}^0(\mathbb{R}^2,\mathbb{R}) &\rightarrow \mathbb{R}\\ u &\mapsto C[u]\end{aligned}$$

where

$$C[u] = \int_0^{2\pi} u(\cos\theta, \sin\theta)\, d\theta\,.$$

Determine the domain of dependence.

8.6. Pointwise, local and global operators. Consider the set of functions $W_1 = C_0^1(\mathbb{R},\mathbb{R})$, consisting of functions of class C^1 that are not null at most on a bounded interval. Let also $k > 0$ be a constant, and $m \in C^1(\mathbb{R},\mathbb{R})$, be a bounded function. Show that

(a) The operator L_1 given by

$$\begin{aligned}L_1 : C_0^1(\mathbb{R},\mathbb{R}) &\rightarrow \mathbb{R}\\ f \mapsto L_1[f] &= \int_{-\infty}^{+\infty} f(y)\, dy\,,\end{aligned}$$

is a global operator.

(b) The operator L_2 given by

$$\begin{aligned}L_2 : C_0^1(\mathbb{R},\mathbb{R}) &\rightarrow \mathbb{R}\\ f \mapsto L_2[f] &= f(0)\,,\end{aligned}$$

is a pointwise operator.

(c) The operator L_3 given by

$$\begin{aligned}L_3 : C_0^1(\mathbb{R},\mathbb{R}) &\rightarrow C_0^0(\mathbb{R},\mathbb{R})\\ f \mapsto L_3[f] &= \frac{df}{dx}\end{aligned}$$

is a local operator.

(d) The operator L_4 given by

$$\begin{aligned} L_4 : C_0^1(\mathbb{R}, \mathbb{R}) &\rightarrow C_0^1(\mathbb{R}, \mathbb{R}) \\ f &\mapsto L_4[f] = g \end{aligned}$$

where

$$L_4[f](x) = g(x) = m(x)g(x) \, ,$$

is a pointwise operator.

(e) The operator L_5 given by

$$\begin{aligned} L_5 : C_0^1(\mathbb{R}, \mathbb{R}) &\rightarrow C_0^1(\mathbb{R} \times]0, +\infty[, \mathbb{R}) \\ f &\mapsto L_5[f] = u \end{aligned}$$

where

$$L_5[f](x,t) = u(x,t) = \frac{1}{\sqrt{4k\pi t}} \int_{-\infty}^{+\infty} e^{-\frac{(x-y)^2}{4kt}} \, dy \, ,$$

is a global operator.

Afterword: Some Thoughts on Model Complexity and Knowledge

O objetivo final de uma aula deveria ser formar futuros pesquisadores, e não decoradores de matéria.[a]
Stephen Kanitz, 2003.

We have an incapacity for proving anything which no amount of dogmatism can overcome. We have an idea of truth which no amount of skepticism can overcome.
Blaise Pascal (1623-1662).

... a verdadeira e inexpugnável glória de Deus começa onde termina a linguagem...[b]
Luis Fernando Veríssimo, 2008.

[a] Translation from portuguese: "The main purpose of a class should be to train researchers, and not memorizers", [41].

[b] Translation from portuguese: "...the true and inexpugnable God's glory begins where language ends...", [90].

What are we doing when we use mathematical models to understand physical (chemical, biological, social, environmental,...) phenomena? Why it works? When it works? These are very deep questions. E. Wigner's "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," [94] and R. Feynman's "The Character of Physical Law" [32] treat some aspects of these questions. Due to the nature of inverse problems, these questions are always on the back of our minds. Here we take a look at some of these issues. We shall present the 'how' questions. How models get more fundamental? More fundamental means more knowledge? These are implicit questions we try to shed light on. These considerations just intend to say what is being done, not how to do it.

Computational Modeling

Computational science and engineering (CSE) is a modern way to confront problems. It builds its strength by bringing together ideas coming from computing, mathematics, and engineering to handle contemporary challenges. Strang's book

[83] presents his very interesting account of the basic technical skills the CSE practitioner should have.

Here we will foccus on model building activity, its historical emergence, and its levels. This a major skill in CSE and deserves to be seen from several points of view. In a very sketchy picture, we could say that engineers and physicists are well trained at building models, whereas mathematicians are good at drawing logical conclusions from them. Computer scientists are good at the art of constructing fast algorithms to get numbers out of the models. We shall use the term *computational modeling* for this whole set of activities. Roughly speaking, it corresponds to modeling a physical phenomenon, process or system, and constructing an automatic information system which provides information about its behaviour or its properties by means of a computational device.

Computational modeling is then an area of scientific research inherently interdisciplinary (or at least multidisciplinary, [62]) since it covers at least two assumptions for the study of a particular subject. This field of research and their problems should be able to be described using mathematical language and models, and these have to be solved with the aid of a computer. Moreover, the computer may also serve to capture data, in an automated manner, in real situations, and to present results and scenarios, or even, to command an intervention in an environment. This, in particular, leads to a sequence of steps[11] for solving a problem or to enhance its understanding[12]:

> observation → intuition about the phenomenon → data → physical model[13] → mathematical model → solution algorithm[14] → computational implementation[15] → simulation scenarios (results) → viewing and making the results available[16] → analysis of the results → inference about the phenomenon → comparison → validation → intervention .

Consequently, computational modeling is not mathematics, because it needs modern technology to crunch numbers, it is not physics, even in its broadest sense, because it requires the analysis of the behaviour of discretization processes, and it is not

[11] Clearly, this scheme has several important internal feedbacks. Among them, there is the computational implementation. Frequently, this forces a return to the algorithm and to the mathematical model, given the unsurpassable computational difficulties. This is just one example of the many possible feedbacks. It exhibits the *non-sequential* nature of computational modeling, and demonstrates that it is not an easy task.

[12] It goes without saying that all these steps involve diverse branchs of science and several scientists, which work with specific tools through ingenious arts and crafts. Nevertheless, the time is ripe for a computational modeler, a professional with a comprehensive view of the whole process.

[13] *Physical model*: By this we mean a specific linguistic jargon adequate to describe the main features of the phenomenon under investigation, not necessarily restricted to Physics.

[14] Solution algorithms, implemented in computers, usually involve discretization.

[15] Use of technology. Here, we use the word ‘technology’ in a broader sense, to include, for instance, object oriented programming techniques, debuggers, code evaluators, and so on.

[16] Again, the use of technology.

computing, because it needs the construction of mathematical models of physical phenomena, and draw conclusions from them.

Computational modeling has a lot of *mathematical structures* (numbers, shapes, and so on) associated with *technologies* for the capture, representation, processing and provision of these structures, and specific *'physical' understanding* of a problem. The need for the knowledge from various disciplines is the foundation of the interdisciplinary nature of computational modeling. It is clearly a way to approach a problem, at least those that can be described by mathematical objects, combining theory and practice through technology[17].

Historical and Conceptual Perspective

Science, and more generally the evolution of human knowledge, is part of history of mankind, and therefore its leaps forward, its main ideas, and its unequivocal advances are difficult to pinpoint in one time, in one place.

In this whirlwind, the roots of what is called computational modeling did not appear yesterday at the dawn of computing and computers. The interplay between science and technology, one of the ways of computational modeling, has been one of the pillars of human society for millenia. Could we find a landmark that would indicate its beginning? Certainly this would include mathematics and would take us to its origins.

Pythagoras (about 570 to 497 BC), centuries before the Christian era, has insisted that "*everything is numbers*". This assertion is, of course, a highly distorted snapshot of his thoughts. Probably other persons and peoples before him may have had similar perceptions and understandings. Anyway, what we want is to understand this as "*it is a good idea to quantify the phenomena of nature!*", and properly acknowledge that it has been around for a long time.

At that time, there was already technology to handle numbers or mathematical concepts. In fact, at a certain point in history, one realizes that two lines can occupy particular positions and deserve to be called *perpendicular* lines. This is due to the symmetry that we can observe that this configuration presents. There is, however, the issue of constructing an artefact (structure) with that property. A method — a technology known by the Egyptians was to make knots equally spaced on a string, and construct a triangle with sides 3, 4 and 5. The right angle would be constructed in this way, technologically speaking! It is relevant to remember here the saying of V. I. Arnold,*"...Mathematics is the part of physics where experiments are cheap"*[18].

Rooted on the brilliant contributions of the geniuses of antiquity, it is possible to justify that a major step forward in the genesis of computational modeling was performed at the time when Newton was *standing on the shoulders of giants*, as

[17] Due to the large scope of tasks and knowledge required, this type of action cannot be done nor grasped by an individual. It requires multidisciplinary teams.

[18] "Mathematics is a part of physics. Physics is an experimental science, a part of natural science. Mathematics is the part of physics where experiments are cheap...", [7]

he himself said[19]. It is a time when mathematical models, capable of capturing the essence of various phenomena, were devised.

Often, different groups of scientists end up constructing their own jargon for certain common and known notions, which, due to the difficulties of communication between specialities, have not percolated through the scientific community. Although the set of fundamental ideas is not that big, we would say that it usually falls victim of the *Babel's tower effect* and its *diversification force*.

We shall try to overthrow, at least a little, this communication barrier. Or, at least, we shall present our view why this is the best time in history to weaken such barrier. We will reflect on the concept of computational modeling, and try to understand its essence and to uncover its origin. For this, we introduce a little nomenclature, and illustrate it with some elementary examples.

We need to make some effort in order to see how these concepts apply to more complex situations such as those presented in other chapters of this book. We may seize the opportunity, however, because we are living a historic moment in science, vibrant, conducive to a weakening of that communication barrier due to the ability of integrating mathematical models with information technology.

Let us begin by thinking about as sets are, usually, defined. Sets can be provided by *extension*

$$A = \{6, 7, 8, 9, 10\}\,, \tag{1}$$

or by *comprehension*

$$A = \{y \in \mathbb{N},\ \text{such that}\ (y-8)^2 \leq 4\}\,.$$

Here we denote the set of natural numbers by $\mathbb{N}$. One may think that defining a set by extension is almost the same as providing a function that generates it. That is, the set that we want to have a hold on is given as the image of a function. Consider the function f given by

$$\begin{aligned} f : \{1,2,3,4,5\} &\rightarrow \mathbb{N} \\ x &\mapsto f(x) = x+5\,. \end{aligned}$$

Then, the image of f,

$$\begin{aligned} \mathsf{Im}\,(f) &= \{y = f(x),\ \text{for}\ x = 1,2,\ldots,5\} \\ &= \{f(1) = 6,\ f(2) = 7, \ldots, f(5) = 10\} \\ &= \{6,7,8,9,10\}\,. \end{aligned} \tag{2}$$

coincides with the set A. In symbols, $\mathsf{Im}\,(f) = A$. Presenting a set in this way, as the image of a function, we may say that it is given *encapsulated*.

Table 8.1 sketches what we have just said, and relates to the way a computer scientist deals with the same concepts. The second part of the table also includes other ideas on the issue of models and modeling, discussed later on.

[19] Quote of Isaac Newton (1642-1727) "If I have seen further queries [than certain other men] it is by standing upon the shoulders of giants." in a letter to Robert Hooke (1635-1703), on February 5, 1675, referring to the work of Galileo Galilei (1564-1642) and Johannes Kepler (1571-1630) in physics and astronomy, [5].

Table 8.1 Complexity of the way of defining sets or of the structure models: it increases from left to right

Model	Scientist	Extension	Encapsulated	Comprehension
Set specification		database, table	function	equation
	Mathematician	$\{6, 7, 8, 9, 10\}$	$\{1,2,\ldots,5\} \ni x \mapsto$ $f(x) = x + 5 \in \mathbb{N}$	$\{y \in \mathbb{N}, \mid (y-8)^2 \leq 4\}$
	Computer Scientist	data.txt	Loop (FOR)	Conditional (IF)
Model complexity		database (DB)	descriptive model	explanatory model
	Engineer/Physicist	experimental data	kinematics	dynamics
Questions		what, where, when?	what, where, when?	how?
Theoretical meaning		raw data	information	knowledge
Range of validity		specific	restrict	universal

Let us reflect about the construction of models of physical phenomena. In a practical situation, following the ‘advice of Pythagoras’, we associate numbers to phenomena, that is, we construct a database (DB) of the situation. In the next stage, we perform the creative act of obtaining, first, the function that represents the data, and then the equation that has the function as one of its solution. Finally, we can say that we have modeled the situation mathematically to its full extent.

Next, we check the database. For the sake of concreteness, let us assume that the set A is given by Eq. (1). After analysing it, it is reasonable to guess that the class of models for data, C, is the set whose elements are

$$C_{a,b} = \{y \in \mathbb{N},\ \text{such that } (y-a)^2 \leq b\}, \quad \text{for all } a, b \in \mathbb{R}\,,$$

that is,

$$C = \{C_{a,b},\ \text{for all } a,b \in \mathbb{R}\}\ .$$

This highly creative stage, we call the *characterization of the model*. At this point, it should be emphasized that the values of the parameters a and b are unknown. Therefore, the next step is to determine a and b. This must be done in such a way as to select, among the elements of the class of models C, the one that best represents, in some sense, the set A, according to experimental data. This is the *inverse identification problem*. This is also called model selection, model identification, model estimation, optimization, machine learning, calibration etc., depending on the class of scientists who are tackling this problem[20].

We can classify the models as *descriptive*, when, in a sense, it is a ‘picture’ of the state of things, or *explanatory* if it establishes the relations of interdependence. The intellectual sophistication of models grows from descriptive models to explanatory models, going from specifics to fundamentals. The information contained in each of the models is about the same, but is being compressed, by focusing on fundamental principles, in the direction presented in the diagram below:

$$DB \rightarrow \quad function \rightarrow \quad equation\ . \tag{3}$$

Most theories, at least those that adopt mathematical models, start with a database, then proceed to represent the data by means of a function and then to get an equation[21] for which the function is one of the solutions[22].

This is already too much confusion! The beginner finds him/herself in a very difficult position, and if he/she gives it a little thought, tends to get muddled. Is it not so? The scholar, in possession of an equation, works to get a function (the solution

[20] Most likely, this *Babel tower effect* might be mitigated as the *modus operandi* of computational modeling becomes more widely used in getting results.

[21] Here, usually, these are differential equations. By no means it is universal and, in no way, it is essential for the ideas presented here, that the models are differential equations.

[22] A statistician would say he/she has a sample of a random variable, then a random variable and finally a stochastic equation. If time is involved, maybe he/she can have a time series — a *realization* of a stochastic process— followed by a stochastic process and finally a stochastic equation.

Table 8.2 Position of a fallen body as a function of time

When	time (s)	t	0	1	2	3	4
Where	space (m)	x	100	95.1	81.4	55.9	21.6

to the problem) and, in fact, still has more interest in the database generated from the model (because he/she is usually interested in the interpretation of the solution)[23].

But if what you want is data, why would you build the equation? A reminder is that you still have to have data anyway in order to construct a meaningful equation. Can only be a strange thing to do! Understanding (or is it compressing data?) is the key.

Let us consider a physical situation, for example, the drop of a solid body due to gravity. To study and investigate the fall of a body, one can make use of a *descriptive* model or even of an *explanatory*. In the case of the motion of point particles (point masses), the descriptive model is known as the kinematics of the particle, and the explanatory model is obtained from Newton's 2^{nd} law.

Typically, descriptive models are useful in answering to questions like: "what, where and when? ". It also handles questions like "how many, when?", "how much, where?", "how many, how much?" or combinations thereof.

In the fall of a solid, *what* is the very solid, and *where and when* can be given by a worksheet, as Table 8.2, that is a DB. After careful analysis, it is possible to give a descriptive model (or kinematic model) in parametric form (function)

$$x(t) = x_0 + v_0 t - \frac{1}{2} g t^2 \, . \tag{4}$$

Here, g denotes the acceleration of gravity, x_0, the initial position and v_0 the initial velocity. We let the x-axis to point out of the earth's surface. In the example, we assume that $g = 9.8 m/s^2$, $x_0 = 100m$, and $v_0 = 0m/s$. Therefore,

$$x(t) \quad = \quad 100 - 4.9t^2 \, . \tag{5}$$

We want to comment on this model. When we write Eq (4) to address the status of the fall of a point particle, what we have done is to *characterize* a class of descriptive models. The determination of the constants present there is the *identification* of the model. This is done, for example, using the experimental measurements present in Table 8.2

In contrast, explanatory models answer the question: "how?". In the case of a body, how it falls to the earth. It falls satisfying Newton's second law. This expresses the change in velocity (the acceleration) by the action of a force,

[23] Let us be more explicit. Consider the design of an airplane. Assume that one has modeled the problem using differential equations that tells how to balance forces and other physical quantities. What one wants, for example, is to estimate the value of the pressure at a few critical points on the wing, to see if it will endure flying conditions. In short: one wants specific numbers. These would constitute a small 'database'. That is, we want to follow Eq. (3) in the direction opposite to the arrows.

$$m\frac{dv}{dt} = F\,, \tag{6}$$

where m is the mass of the body, v is its velocity, and F is the force acting on the body. Since $v = dx/dt$, and, in the fall of a body, the gravitational attraction force is equal to mg, then

$$\frac{d^2x}{dt^2} = -g\,. \tag{7}$$

Similarly to the case of the descriptive model, here you can also employ the concept of characterization and identification of the model. When you write Eq. (6), what you have done is to *characterize* the model. Once you obtained Eq. (7) you, essentially, *identified* the model.

This is the explanatory (or dynamic) model. Integrating twice Eq. (7) with respect to time, we obtain

$$\frac{dx}{dt} = v_0 - gt\,,$$
$$x(t) = x_0 + v_0 t - \frac{1}{2}gt^2\,,$$

which turns out to be the descriptive model discussed previously.

What we are calling descriptive and explanatory models, engineers or physicists would call, respectively, kinematic and dynamic models. We could say that the explanatory model contains enough information for one to recover the descriptive model. It is advantageous at this time, to review Table 8.1, where we present, in summary form, the notions of model, its purpose, and its meaning.

All these concepts should be used with a *grain of salt*, they help you organize your thoughts, but have their limitations[24].

Another natural question to ask yourself is 'why?'. For example, why the body is attracted to the Earth? The question *"Why it happened?"* ("*Why*") was deemed by Newton as a deterrent to scientific understanding, and he urged scientists to deviate their attention away from it. There was no need to assume or ask anything else. As Occam[25] would say, 'let us use my 'razor', and forget the whys!'.

The development of scientific knowledge can be, somewhat, understood from the example just being presented. Typically, at the beginning of the study or investigation of a phenomenon, one gathers data. Usually, by means of observation and controlled experiments, one quantifies the actual situation and constructs a DB, as Table 8.2. Next, one theorizes a little bit and gets a descriptive (parametric, functional) model as given in Eq. (4). Finally, one looks for explanatory models, which, in the example considered, is given by Newton's second law, Eq. (6).

[24] Roughly, it can be said that mathematicians keep the direct problems and statisticians work with the inverse problems. Would that fit what one calls theoretical and experimental work? Scientific or technologically oriented job? Almost any generalization of this nature tends to be imprecise because of its simplicity.

[25] William of Occam (1285-1347), an English theologian, used a logical implement, the so-called Occam's razor, to cut absurd arguments. Occam's maxim is that the simpler the explanation is, the better it is.

After all, when did first emerged the structuring of a chain of models of increasing complexity, as depicted by Eq. (3)? Who built the database, who discovered the descriptive model (kinematic) and who unveiled the explanatory model (dynamic)? And in what context? A historical example is called for and we shall duly interpret it.

One of the earliest examples of this type of method, constructing the first models that quantify the most basic situation — making observations carefully — and proceeding to more sophisticated models was performed in studying the orbits of the planets and other celestial bodies. In this case, the DB was produced by Tycho Brahe (1546-1601), who built an observatory (Uranienborg) in an island of Denmark, to do so. The descriptive model is due to Johannes Kepler — Kepler's laws. The explanatory model was developed by Galileo Galilei — the law of inertia — and more strongly by Isaac Newton with his second law and the law of gravitational attraction (universal gravitation theory)[26]. Gauss, with his *least squares method*, enters the picture, opens the door to future computations, and to the solution of practical inverse problems.

There is a simple relationship between the classes of models and mathematical structures:

$$\begin{aligned}\text{explanatory models} &\longleftrightarrow \text{equations}\\ \text{descriptive models} &\longleftrightarrow \text{functions}\\ \text{database (tables)} &\longleftrightarrow \text{function evaluation}\end{aligned}$$

From the foregoing it is clear that explanatory models are conceptually more sophisticated than descriptive ones, which in turn are more sophisticated than tables (DB). Needless to say, it is not our intention to devalue nor valuing the skills needed to obtain any of the types of models discussed. It is blatantly clear that the advancement of scientific knowledge depends crucially in DB's, descriptive, and explanatory models, each equipped with its own set of difficulties. Table 8.1 deserves to be revisited at this time since it presents a summary of these ideas.

The advancement of knowledge follows, in many sciences, this path

$$\text{DB} \quad \overset{induction:art}{\Longrightarrow} \quad \text{descriptive model} \quad \overset{induction:art}{\Longrightarrow} \quad \text{explanatory model}$$

whereas[27] applications run the other way around

$$\text{explanatory model} \quad \overset{deduction:math}{\Longrightarrow} \quad \text{descriptive model} \quad \overset{deduction:math}{\Longrightarrow} \quad \text{DB}$$

If this is so old, dating back at least to the sixteenth and seventeenth centuries, what else is new? And how it all fits together?

Computational modeling — with its science, technique and technology — comes exactly at the transition from explanatory to descriptive model and from descriptive to database. Computer and its peripherals correspond, in fact, to the technological novelty in the previous diagrams. Moreover, the technology that allows experiments was and remains extremely relevant in the first of the above diagrams, especially in

[26] For an exciting presentation of these events in the history of science, see Feynman [32].

[27] Here *art* refers to model building skills, which in fact is a refined art.

building database. It also happens that computational tools, together with electronics, have been developed for automatic data aquisition, through the widest range of sensors. In this way its technological strength enters the former diagram.

Our goal as scientists is to predict the behaviour of phenomena, processes and systems in different situations. Thus, and using the drop of a point particle as an example, we would like to answer the following problems:

P_1 : Given any specific time, say the position of the point mass;

P_2 : Given a position, determine the time instant when the point mass passes through this position;

P_3 : How is it that a point mass falls?

We are not, however, interested in answers/predictions of a general nature, but only those based on scientific descriptions of the fall of the point mass. To do that ... (**goto** Chapter 1).

Exercise

Descriptive model. We recall Kepler's laws:

1^{st} law the orbit of the planet forms an ellipse, with the sun as one of the focus of the ellipse;

2^{nd} law equal areas are swept in equal time intervals. The area that is swept by a planet in the interval of time between time t and time $t + \triangle t$ is defined in the following way. At time, t, draw the line joining the Sun's position and the planet's position, the so-called *radius vector*, and do the same at a given time interval, $\triangle t$, later. The area swept is defined as the area enclosed by the radius vectors and the ellipse;

3^{rd} law the time to go round the ellipse, the period T, varies as (it is proportional to) the size of the orbit raised to the power of three halves, where the size of the orbit is the biggest diameter across the ellipse.

Assume that the orbit of a certain planet is given by the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 , \tag{8}$$

with $a > b$, and that the planet travels anti-clockwise.

(a) Assume that the Sun has position (x_s,y_s), with $x_s > 0$ and $y_s = 0$. Give an expression for x_s in terms of a and b.

(b) Assume that the constant of proportionality in the third law is represented by α. Give an expression for T.

(c) Let $A = A(t)$ denote the area swept by the planet, when starting from the position $(x(0), y(0)) = (a,0)$. Obtain an expression for $\frac{dA}{dt}$.

(d) Let the position of the planet be represented by polar coordinates,

$$(x(t), y(t)) = (r(t)\cos\theta(t), r(t)\sin\theta(t))\ .$$

Obtain an expression for $r(t)$ as a function of $\theta(t)$.

(e) Write, using polar coordinates, an integral expression for the area, $S = S(t)$, formed by the x-axis, the ellipse, and the position of the planet vector at time t, $(x(t),y(t))$.

(f) Relate $A(t)$ with $S(t)$, and their derivatives with respect to time.

(g) Compute $A(t)$, $t \leq T$ when $\theta(t) = \pi/2$, and when $\theta(t) = \pi$.

(h) Let

$$\begin{aligned} v_m^1 &= \text{‘mean velocity of the planet going from } (a,0) \text{ to } (0,b)\text{’}\ , \\ v_m^2 &= \text{‘mean velocity of the planet going from } (0,b) \text{ to } (-a,0)\text{’}\ . \end{aligned}$$

Compute them. Which one is biggest? Compute v_m^2/v_m^1.

Hint. Recall that the area of the ellipse defined by Eq. (8) is πab and its perimeter is $2\pi\sqrt{\frac{a^2+b^2}{2}}$.

Appendix A
Spectral Theory and a Few Other Ideas From Mathematics

In this appendix, we collect some notations, as well as linear algebra and calculus results that are used throughout the book, particularly the spectral theorem, singular value decomposition, and Taylor's formula which are the basic tools of analysis to understand the behaviour of some methods for solving inverse problems. Several concepts are also recalled in step-by-step exercises.

A.1 Norms, Inner Products and Tensor Products

A set of vectors $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k \in \mathbb{R}^n$ is said to *generate* $\mathbb{R}^n$ if and only if any element $\mathbf{x} \in \mathbb{R}^n$ can be written as a *linear combination* of them, i.e.,

$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_k\mathbf{v}_k \,,$$

for appropriate choices of the coefficients $c_1,c_2,\ldots,c_k$. The set is called *linearly independent* if the only way to represent the zero vector by a linear combination

$$0 = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_k\mathbf{v}_k \,,$$

is when all the coefficients $c_1,c_2,\ldots,c_k$ are null. (Otherwise, it is said that the vectors are *linearly dependent*.) A set of vectors is a *basis* of $\mathbb{R}^n$ if and only if they are linearly independent and generate $\mathbb{R}^n$.

In $\mathbb{R}^n$, we denote by

$$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \,, \tag{A1}$$

the *Euclidean norm* of vector $\mathbf{x} = (x_1,\ldots,x_n)^T$, (we always think of vectors as $n \times 1$ matrices, i.e., "column vectors"). If $|\mathbf{x}| = 1$, we say that the vector is *normalized* or a *unit* vector. For any $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, the *triangle inequality* is written as

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}| \,. \tag{A2}$$

For any vector $\mathbf{x} \neq 0$, $\mathbf{x}/|\mathbf{x}|$ is a unit vector.

Given $\mathbf{y} \in \mathbb{R}^n$, the *inner product* between $\mathbf{x}$ and $\mathbf{y}$ is denoted by

$$\langle \mathbf{x},\mathbf{y} \rangle = x_1y_1 + \ldots + x_ny_n \;=\; \mathbf{x}^T\mathbf{y} \,. \tag{A3}$$

Two vectors are said *orthogonal* if and only if $\langle \mathbf{x},\mathbf{y} \rangle = 0$, which we denote by $\mathbf{x}\perp\mathbf{y}$.

Given a real $m \times n$ matrix A, we denote by A^* the *adjoint matrix* or *adjoint operator* of A, i.e., the matrix that changes places within the inner product in such a way that

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A^*\mathbf{y}\rangle \,, \ \forall \mathbf{x} \in \mathbb{R}^n, \ \forall \mathbf{y} \in \mathbb{R}^m \,. \tag{A4}$$

Notice that the inner product computed on the left hand side of Eq. (A4) is computed in $\mathbb{R}^m$, while the one on the right hand side is computed in $\mathbb{R}^n$. In particular, A^* is $n \times m$. Since

$$\langle A\mathbf{x},\mathbf{y}\rangle = (A\mathbf{x})^T\mathbf{y} = \mathbf{x}^T A^T\mathbf{y} = \langle \mathbf{x}, A^T\mathbf{y}\rangle$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, due to the properties of matrix transposition, we conclude that the adjoint operator of A is simply its transpose, $A^* = A^T$.

Recall that the *induced norm* of a matrix A is given by

$$|A| = \max_{x\neq 0} \frac{|A\mathbf{x}|}{|\mathbf{x}|} = \max_{\substack{|x|=r\\ r\neq 0}} \frac{|A\mathbf{x}|}{|\mathbf{x}|} = \max_{|x|=1} |A\mathbf{x}| \,. \tag{A5}$$

If B is a matrix and $\mathbf{x}$ a vector, both with appropriate dimensions, then

$$|AB| \leq |A|\,|B| \quad \text{and} \quad |A\mathbf{x}| \leq |A|\,|\mathbf{x}| \,. \tag{A6}$$

Given a square matrix A, $n \times n$, we say that $\lambda \in \mathbb{C}$ is an *eigenvalue* of A if and only if there exists a non zero vector $\mathbf{x}$, possibly in $\mathbb{C}^n$, such that

$$A\mathbf{x} = \lambda\mathbf{x} \,.$$

The set of all eigenvalues of a matrix,

$$\sigma_A = \{\lambda \in \mathbb{C} \text{ such that there exists } \mathbf{x} \neq 0 \text{ satisfying } A\mathbf{x} = \lambda\mathbf{x}\} \,,$$

is called the *spectrum* of A.

We recall that the eigenvalues of a matrix A are the roots of its *characteristic polynomial*,

$$p_c(\lambda) = \det(A - \lambda\mathcal{I}) \,.$$

It is possible to show that, for real matrices A,

$$|A| = (\text{largest eigenvalue of } A^T A)^{\frac{1}{2}} \,. \tag{A7}$$

Given two vectors $\mathbf{u}$ and $\mathbf{v}$ of $\mathbb{R}^n$, the *tensor product* $\mathbf{u} \otimes \mathbf{v}$ is defined as the $n \times n$ matrix whose ij element is given by $u_i v_j$, i.e.,

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T \,.$$

Notice the similarities and differences between the definitions of tensor product and inner product, Eq. (A3).

We denote by $L^2(\Omega)$ the set of *square-integrable functions*, that is,

$$f \in L^2(\Omega) \text{ if and only if } \int_\Omega f^2(x)\,dx < +\infty\,,$$

and by

$$\langle f, g\rangle = \int_\Omega f(x)g(x)\,dx\,,$$

the *inner product* defined in $L^2(\Omega)$.

A.2 Spectral Theorem

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of $\mathbb{R}^n$. We say the basis is *orthonormal* if the vectors are mutually orthogonal and normalized, i.e., if they satisfy the following *orthogonality relations*,

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = \delta_{ij} = \begin{cases} 1, & \text{if } i = j\,, \\ 0, & \text{if } i \neq j \end{cases} \quad \text{for } i, j = 1, \ldots, n\,.$$

Let V be the matrix whose i-th column is the vector $\mathbf{v}_i$, $i = 1, \ldots, n$. Then, the basis is orthonormal if and only if

$$V^T V = \mathcal{I}\,,$$

where $\mathcal{I}$ is the identity matrix. In this case, matrix V is said to be *orthogonal*.

Here, we recall the spectral theorem [82].

Theorem 7. Spectral theorem. Let A be a real, square, *symmetric* matrix, that is such that $A^T = A$. Then, its eigenvalues are real, $\lambda_1 \geq \ldots \geq \lambda_n$, and there exists an orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$, formed by eigenvectors of A,

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \text{for } i = 1, \ldots, n\,. \tag{A8}$$

■

If we denote

$$D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} | & \cdots & | \\ \mathbf{v}_1 & \ddots & \mathbf{v}_n \\ | & \cdots & | \end{pmatrix},$$

we can collect all equations in (A8) in a single matrix equation

$$AP = PD\,, \tag{A9}$$

or, considering that P is an orthogonal matrix (in particular, an invertible matrix),

$$A = PDP^T\,. \tag{A10}$$

It is sometimes useful to denote D by $diag(\lambda_1,\lambda_2,\ldots,\lambda_n)$. Also, we denote the *eigenvalues* of real symmetric matrices, A, by $\lambda_i(A)$, $i=1,\ldots,n$, in a decreasing order of magnitude,

$$\lambda_1(A)\geq\lambda_2(A)\geq\ldots\geq\lambda_n(A)\,.$$

Example A.1. As a simple, useful and nice application of the spectral theorem, we show how to sketch the region $\mathcal{H}$ constituted by $(x,y)\in\mathbb{R}^2$ satisfying

$$x^2+y^2-4xy \quad\geq\quad 1\,. \tag{A11}$$

Solution. First we rewrite Eq. (A11) by means of the product of matrices, choosing symmetric the matrix in the middle, as

$$\begin{pmatrix} x & y \end{pmatrix}\overbrace{\begin{pmatrix} 1 & -2\\ -2 & 1\end{pmatrix}}^{A}\begin{pmatrix} x\\ y\end{pmatrix} \quad\geq\quad 1\,. \tag{A12}$$

Next, we construct matrices P and D for A as guaranteed by the spectral theorem. The eigenvalues of A, roots of the characteristic polynomial,

$$p_c(\lambda)=\det\begin{pmatrix} 1-\lambda & -2\\ -2 & 1-\lambda\end{pmatrix}=(1-\lambda)^2-4\,,$$

are $\lambda_1=-1$ and $\lambda_2=3$.

For the eigenvalue λ_1, we solve the system

$$\overbrace{\begin{pmatrix} 1-(-1) & -2\\ -2 & 1-(-1)\end{pmatrix}}^{A-\lambda_1 I}\begin{pmatrix} x\\ y\end{pmatrix} \quad=\quad \begin{pmatrix} 0\\ 0\end{pmatrix},$$

and get $(x,y)^T=(1,1)$ as an eigenvector for $\lambda_1=-1$. Analogously, for $\lambda_2=3$, we solve the system

$$\overbrace{\begin{pmatrix} -2 & -2\\ -2 & -2\end{pmatrix}}^{A-3I}\begin{pmatrix} x\\ y\end{pmatrix} \quad=\quad \begin{pmatrix} 0\\ 0\end{pmatrix},$$

and get $(x,y)^T=(1,-1)$ as an eigenvector for $\lambda_2=3$.

The eigenvectors are orthogonal and, dividing each one by its norm, give orthonormal vectors. Therefore, $A=PDP^T$, where

$$P=\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix} \quad\text{and}\quad D=\begin{pmatrix} -1 & 0\\ 0 & 3\end{pmatrix}.$$

Next, we consider a change of variables,

$$\begin{pmatrix} u\\ v\end{pmatrix} \quad=\quad P^T\begin{pmatrix} x\\ y\end{pmatrix}=\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix} x\\ y\end{pmatrix}=\begin{pmatrix} \frac{x+y}{\sqrt{2}}\\ \frac{x-y}{\sqrt{2}}\end{pmatrix}.$$

Equation (A12) then becomes $(u\ v)D(u\ v)^T \geq 1$ or

$$-\frac{1}{2}(x+y)^2 + \frac{3}{2}(x-y)^2 \geq 1\,. \tag{A13}$$

Notice that $-\frac{1}{2}(x+y)^2 + \frac{3}{2}(x-y)^2 = 1$ is a hyperbola and its *line asymptotes* are given by

$$-\frac{1}{2}(x+y)^2 + \frac{3}{2}(x-y)^2 = 0\,,$$

that is,

$$(\sqrt{3}-1)x = (1+\sqrt{3})y \text{ or } (\sqrt{3}+1)x = (\sqrt{3}-1)y\,,$$

and its *vertices* are given by solving the system (that comes from Eq. (A13), by making it an equality, and equating the first term to zero)

$$\begin{cases} x+y &= 0 \\ \frac{3}{2}(x-y)^2 &= 1 \end{cases},$$

yielding

$$\left(\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}\right) \text{ and } \left(-\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6}\right).$$

From these calculations the sketch of region $\mathcal{H}$ is easily done. ■

The conclusion of the spectral theorem can be represented by the *commutative diagram* in Fig. A.1. Notice that depending on the use of this result, it can be convenient to write A, Eq. (A8), in the form

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T\,, \tag{A14a}$$

or, when multiplied by $\mathbf{x}$,

$$A\mathbf{x} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{x} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \langle \mathbf{v}_i, \mathbf{x}\rangle\,. \tag{A14b}$$

$$\begin{tikzcd}
R^n \arrow[r, "A"] \arrow[r, bend right] \arrow[d, "P^T"'] & R^n \\
R^n \arrow[r, "D"] & R^n \arrow[u, "P"']
\end{tikzcd}$$

Fig. A.1 Commutative diagram

In particular, Eq. (A14b) is in the form of *separation of variables*, where

$$\langle \mathbf{v}_i, \mathbf{x} \rangle ,$$

is the *Fourier coefficient* of $\mathbf{x}$, in the direction of vector $\mathbf{v}_i$, and

$$\lambda_i \langle \mathbf{v}_i, \mathbf{x} \rangle ,$$

is the corresponding Fourier coefficient of its image, $A\mathbf{x}$.

The analysis and synthesis of a real symmetric matrix provided by the spectral theorem are very convenient for the evaluation of analytic matrix functions. Note that

$$A^2 = AA = PDP^T PDP^T = PDDP^T = PD^2 P^T ,$$

For the polynomial

$$q(x) = a_0 + a_1 x + \ldots + a_k x^k ,$$

we denote

$$q(A) = a_0 \mathcal{I} + a_1 A + \ldots + a_k A^k .$$

It can be shown that

$$q(A) = Pq(D)P^T . \tag{A15}$$

For every analytic function

$$f(x) = \sum_{j=0}^{\infty} a_j x^j ,$$

defined in a disk of radius less than $|A|$, we let

$$f(A) = \sum_{j=0}^{\infty} a_j A^j ,$$

and then,

$$f(A) = Pf(D)P^T = P\Gamma P^T ,$$

where Γ is a diagonal matrix whose elements are $f(\lambda_i)$, $i = 1, \ldots, n$. It is simple to verify that the inverse of A is given by

$$A^{-1} = PD^{-1}P^T . \tag{A16}$$

Example A.2. Matrices of the form $K^T K$. Let K be a real $m \times n$ matrix. Matrix $A = K^T K$ is symmetric and its eigenvalues are all non-negative. The same is true for matrices of the form KK^T.

Solution. Let λ be an eigenvalue and $\mathbf{v}$ the corresponding eigenvector ($\mathbf{v} \neq 0$). Then, $A\mathbf{v} = \lambda\mathbf{v}$ and

$$\lambda\langle \mathbf{v}, \mathbf{v}\rangle = \langle \lambda\mathbf{v}, \mathbf{v}\rangle = \langle A\mathbf{v}, \mathbf{v}\rangle = \langle K^T K\mathbf{v}, \mathbf{v}\rangle = \langle K\mathbf{v}, K\mathbf{v}\rangle \, .$$

From this it can be concluded that

$$\lambda = \frac{\langle K\mathbf{v}, K\mathbf{v}\rangle}{\langle \mathbf{v}, \mathbf{v}\rangle} = \frac{|K\mathbf{v}|^2}{|\mathbf{v}|^2} \geq 0 \, .$$

■

Definition A.1. A real, symmetric matrix A is *positive definite* if and only if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$, $\mathbf{x} \in \mathbb{R}^n$. If $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ the matrix is *positive semi-definite*. ■

A.3 Singular Value Decomposition

We begin by enunciating the singular value decomposition theorem.

Theorem 8. Singular value decomposition — SVD. Let K be a real $m \times n$ matrix and $p = \min\{m, n\}$. Then there exist orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ of $\mathbb{R}^n$, $\mathbf{v}_1$,$\mathbf{v}_2, \ldots, \mathbf{v}_m$ of $\mathbb{R}^m$, and scalars $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$, such that,

$$K = V\Lambda U^T \, , \tag{A17}$$

with Λ, a $m \times n$ matrix given by

$$\Lambda = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \\ 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix} , \quad \text{if } m > n = p \, ,$$

$$\text{or,} \quad \Lambda = \begin{pmatrix} \sigma_1 & \cdots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m & 0 & \ldots & 0 \end{pmatrix} , \quad \text{if } p = m < n \, ,$$

or $\Lambda = diag(\sigma_1$,$\sigma_2, \ldots, \sigma_p)$, if $m = n = p$, and $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$, an orthogonal $m \times m$ matrix, $U = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ an orthogonal $n \times n$ matrix. ■

Before giving a proof of the theorem we present some conclusions streaming from it. The *singular values* of K are the numbers σ_i, that we denote by $\sigma_1(K) \geq \sigma_2(K) \geq$

$$\begin{tikzcd}
\mathrm{R}^{n} \arrow[r, "K"] \arrow[d, "U^{T}"'] \arrow[r, bend right] & \mathrm{R}^{m} \\
\mathrm{R}^{n} \arrow[r, "\Lambda"'] & \mathrm{R}^{m} \arrow[u, "V"']
\end{tikzcd}$$

Fig. A.2 Commutative diagram

$\ldots \geq \sigma_p(K) \geq 0$, explictly mentioning matrix K. As the index i grows, the corresponding singular value diminishes.

Equation (A17) can be written in different ways. We have

$$K = \sum_{i=1}^{p} \sigma_i \mathbf{v}_i \otimes \mathbf{u}_i = \sum_{i=1}^{p} \sigma_i \mathbf{v}_i \mathbf{u}_i^T \ ,$$

or, what is the same, the commutative diagram in Fig. A.2 is valid.

The evaluation of $K\mathbf{x}$, that can be thought of as a generalization of the Fourier method of separation of variables, is written as

$$K\mathbf{x} = \sum_{i=1}^{p} \sigma_i \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{x} \rangle \ .$$

The singular value decomposition does not allow the generality that the spectral theorem does with relation to function evaluation, but the representation of the inverse is still provided, as seen immediately.

If K is invertible ($n = m = p$), the singular values of the inverse are

$$\sigma_1(K^{-1}) = 1/\sigma_n(K) \ , \ \sigma_2(K^{-1}) = 1/\sigma_{n-1}(K) \ , \ \text{and so on.}$$

In particular, the largest singular value of K originates the smallest singular value of K^{-1} and vice versa. The inverse is given by

$$K^{-1} = U\Lambda^{-1}V^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{u}_i \otimes \mathbf{v}_i = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{u}_i \mathbf{v}_i^T \ , \tag{A18}$$

where Λ^{-1} is a diagonal matrix and the elements of the diagonal are σ_i^{-1},

$$\Lambda^{-1} = diag\left(\frac{1}{\sigma_n}, \frac{1}{\sigma_{n-1}}, \ldots, \frac{1}{\sigma_1} \right) \ .$$

The evaluation in $\mathbf{y}$ is given by

$$K^{-1}\mathbf{y} = \sum_{i=1}^{n} \sigma_i^{-1} \mathbf{u}_i \langle \mathbf{v}_i, \mathbf{y} \rangle \ .$$

Now, we present a proof of theorem 8, since this result is not commonly discussed in elementary courses on linear algebra.

Proof of theorem 8. We start by building a basis of $\mathbb{R}^n$. Define $A = K^T K$ (we could have choosen $A = KK^T$). A is a real symmetric matrix and the spectral theorem can be applied to it. Therefore, let

$$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \,,$$

be an orthonormal basis of $\mathbb{R}^n$, that diagonalizes A. Let

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \,,$$

be the eigenvalues of A,

$$A\mathbf{u}_i = \lambda_i \mathbf{u}_i \,.$$

Now, we construct a basis of $\mathbb{R}^m$. Considering n_o as the index of the smallest non-zero eigenvalue of A, we note that the vectors

$$\mathbf{w}_i = K\mathbf{u}_i \text{ , for } i = 1, \ldots, n_o \leq p \,,$$

are non-zero since

$$K^T \mathbf{w}_i = K^T K\mathbf{u}_i = A\mathbf{u}_i = \lambda_i \mathbf{u}_i \neq 0 \,, \quad i = 1, \ldots, n_o \,.$$

Also, the vectors $\mathbf{w}_i$ are mutually orthogonal since

$$\begin{aligned} \langle \mathbf{w}_i, \mathbf{w}_j \rangle &= \langle K\mathbf{u}_i, K\mathbf{u}_j \rangle = \langle \mathbf{u}_i, K^T K\mathbf{u}_j \rangle \\ &= \langle \mathbf{u}_i, A\mathbf{u}_j \rangle = \langle \mathbf{u}_i, \lambda_j \mathbf{u}_j \rangle = \lambda_j \delta_{ij} \,, \end{aligned}$$

for $i,j = 1, \ldots, n_o$. We define

$$\mathbf{v}_i = \frac{\mathbf{w}_i}{|\mathbf{w}_i|} = \frac{K\mathbf{u}_i}{|K\mathbf{u}_i|} \,, \quad i = 1, \ldots, n_o \tag{A19a}$$

and

$$\sigma_i = |\mathbf{w}_i| = |K\mathbf{u}_i| \neq 0 \,. \tag{A19b}$$

Thus, $\mathbf{v}_1, \ldots, \mathbf{v}_{n_o}$ are orthonormal and

$$K\mathbf{u}_i = \sigma_i \mathbf{v}_i \,,$$

as well as

$$K^T \mathbf{v}_i = K^T \frac{\mathbf{w}_i}{\sigma_i} = \frac{\lambda_i}{\sigma_i} \mathbf{u}_i \,.$$

Also,

$$\begin{aligned}\lambda_i &= \lambda_i\langle \mathbf{u}_i, \mathbf{u}_i\rangle = \langle \lambda_i\mathbf{u}_i, \mathbf{u}_i\rangle = \langle A\mathbf{u}_i, \mathbf{u}_i\rangle \\ &= \langle K^T K\mathbf{u}_i, \mathbf{u}_i\rangle = \langle K\mathbf{u}_i, K\mathbf{u}_i\rangle \\ &= \langle \sigma_i\mathbf{v}_i, \sigma_i\mathbf{v}_i\rangle = \sigma_i^2\langle \mathbf{v}_i, \mathbf{v}_i\rangle = \sigma_i^2\,,\end{aligned}$$

whence $\lambda_i = \sigma_i^2$ for $i = 1,2,\dots,n_o$.

We define $\sigma_i = 0$, for $n_o < i \le p$, if there is such i, and likewise let $\mathbf{v}_i$, for $n_o < i \le m$, be orthonormal vectors that together with $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_{n_o}$, form an orthonormal basis of $\mathbb{R}^m$. Then, defining Λ as one of the possibilities in the statement of the theorem, accordingly to the order relation between m and n, letting $U = (\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_n)$, and $V = (\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_m)$, the stated properties are satisfied. ■

Example A.3. Determine the singular value decomposition of

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Solution First compute

$$A^T A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

(We choose $A^T A$ instead of AA^T because the order of the former is smaller. Nonetheless, we could have used AA^T.) The positive square root of non-zero eigenvalues of $A^T A$ are singular values of A. All other singular values of A, if any, are zero. The characteristic polynomial of $A^T A$ is

$$p_c(\lambda) = \det(A^T A - \lambda I) = \det\begin{pmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{pmatrix} = (2-\lambda)(6-\lambda)\,.$$

Therefore, $\sqrt{6}$ and $\sqrt{2}$ are the singular values of A. An orthonormal eigenvector associated with eigenvalue 6 is $\mathbf{v}_1 = (\sqrt{2}/2, \sqrt{2}/2)^T$, and one associated with 2 is $\mathbf{v}_2 = (\sqrt{2}/2, -\sqrt{2}/2)^T$. Now,

$$AA^T = \begin{pmatrix} 2 & 2 & 2 & 0 \\ 2 & 2 & 2 & 0 \\ 2 & 2 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

A unit eigenvector associated with $\lambda = 6$ is

$$\mathbf{u}_1 = \left(\sqrt{3}/3,\ \sqrt{3}/3, \sqrt{3}/3, 0\right)^T\,,$$

and one associated with the other eigenvalue, $\lambda = 2$, is

$$\mathbf{u}_2 = \mathbf{e}_4 = (0,0,0,1)^T \ .$$

We use a trick to complete $\mathbf{u}_1$ and $\mathbf{u}_2$ to an orthonormal basis of $\mathbb{R}^4$. The plane of equation $x+y+z=0$ is the set of vectors, in $\mathbb{R}^3$ that are orthogonal to $\mathbf{n} = (1,1,1)^T$. One such vector is $\mathbf{o}_1 = (1,-1,0)^T$. Another is determined by the cross product[1] $\mathbf{o}_2 = \mathbf{o}_1 \times \mathbf{n} = (-1,-1,2)^T$. Normalizing and imbedding these vectors in $\mathbb{R}^4$, with the 4^{th}-entry equal to zero, we get an orthonormal basis of $\mathbb{R}^4$. Finally, the SVD of A is $A = U\Sigma V^T$, with,

$$U = \begin{pmatrix} \frac{\sqrt{3}}{3} & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & 0 & 0 & 2\frac{\sqrt{6}}{6} \\ 0 & 1 & 0 & 1 \end{pmatrix} ,$$

$$\Sigma = \begin{pmatrix} \sqrt{6} & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 \end{pmatrix}^T , \text{ and}$$

$$V = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} .$$

■

Remark A.1. Eigenvalues and singular values.

(a) Given any matrix K, we have that

$$|K| = \sigma_1(K) ,$$

and if K is invertible,

$$|K^{-1}| = [\sigma_n(K)]^{-1} .$$

(b) If K is a real symmetric matrix and $\lambda_1(K) \geq \lambda_2(K) \geq \ldots \geq \lambda_n(K)$ are its eigenvalues, then the set of its singular values is

$$\{|\lambda_1(K)|, |\lambda_2(K)|, \ldots, |\lambda_n(K)|\} .$$

(c) Let A be a real symmetric matrix. Then $|A| = \max_i |\lambda_i(A)|$.

(*d*) If $A = K^T K$, the eigenvalues of A, $\lambda_i(K^T K)$, are related to the singular values of K, $\sigma_i(K)$, by $\lambda_i(K^T K) = \sigma_i^2(K)$.

[1] For the sake of completeness, we recall the definition of cross product of two vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^3$, given by

$$\mathbf{x} \times \mathbf{y} = (x_2 y_3 - x_3 y_2, x_3 y_1 - x_1 y_3, x_1 y_2 - x_2 y_1)^T \ .$$

A.4 Special Classes of Matrices

In this section we shall present the class of projection matrices and another one that generalizes them, representing projection functions. For that we need to introduce a few more concepts from linear algebra.

A.4.1 Projection Matrices

In $\mathbb{R}^n$, a set U is called a *vector subspace* if it is

(a) *closed* under sums of its elements,

$$\text{for all } \mathbf{u} \in U, \text{ and all } \mathbf{v} \in U \text{ we must have } \mathbf{u} + \mathbf{v} \in U \ ,$$

(b) and it is *closed* under multiplication by scalars,

$$\text{for all } r \in \mathbb{R}, \text{ and all } \mathbf{u} \in U \text{ we need that } r\mathbf{u} \in U \ .$$

Given two vector subspaces U and V in $\mathbb{R}^n$, with their intersection containing only the null vector, $U \cap V = \{0\} \subset \mathbb{R}^n$, we define their *direct sum*,

$$U \oplus V = \{\mathbf{u} + \mathbf{v} \in \mathbb{R}^n, \text{ for all } \mathbf{u} \in U, \mathbf{v} \in V\} \ .$$

As an example, if U is the x-axis in $\mathbb{R}^3$, $U = \mathbb{R}\times\{(0,0)\} \subset \mathbb{R}^3$, and V is the yz-plane, $V = \{0\} \times \mathbb{R}^2$, then $U \oplus V = \mathbb{R}^3$.

The notion of a direct sum of two vector subspaces can be extended to the direct sum of three or more vector subspaces. However, since we do not use it, we leave it to the reader to pursue it.

Two vector subspaces, U and $V \subset \mathbb{R}^n$, are said to be *orthogonal* if for all $\mathbf{u} \in U$ and $\mathbf{v} \in V$, we have $\mathbf{u}\perp\mathbf{v}$. For this definition, the xy-plane in $\mathbb{R}^3$ and the yz-plane are not orthogonal subspaces, while the x-axis and the yz-plane are.

Given two orthogonal subspaces, it is clear that their intersection is just the null vector, and we denote their direct sum by $U \oplus^{\perp} V$, just to reinforce, through the notation, that the vector subspaces are orthogonal.

Now we introduce the notion of projection matrices and present its geometrical interpretation.

Definition A.2. A real, square, $n \times n$ matrix P is called a *projection matrix* if

$$P^2 = P \ . \tag{A20}$$

Moreover, if it is symmetric, $P^T = P$, it is called an *orthogonal projection*. Otherwise it is called an *oblique projection*.

Requirement present in Eq. (A20) does not correspond immediatly to what one would expect a projection should be. It is, however, easy to verify. That this definition lives up to the expected geometrical meaning depends on further notions, that we present next.

Given an $m \times n$ matrix A, define the vector subspaces

$$\text{null space of } A\text{: } N(A) = \{\mathbf{x} \in \mathbb{R}^n \text{ such that } A\mathbf{x} = 0\}\,, \text{ and}$$
$$\text{image of } A\text{: } \mathrm{Im}(A) = \{A\mathbf{x} \in \mathbb{R}^m, \text{ for all } \mathbf{x} \in \mathbb{R}^n\}\,.$$

When P is a projection, we call $\mathcal{S} = \mathrm{Im}(P)$ the *screen* of the projection P, and call $\mathcal{V} = N(P)$ the space of *directions of projection*.

Given a projection P, any $\mathbf{x} \in \mathbb{R}^n$ can be uniquely decomposed as a sum of two vectors, one in the screen and the other in the directions of projection, $\mathbf{x}_1 \in \mathcal{S}$ and $\mathbf{x}_2 \in \mathcal{V}$, with $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$. Clearly,

$$\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}\,.$$

Now, take $\mathbf{x}_1 = P\mathbf{x}$ and $\mathbf{x}_2 = (I - P)\mathbf{x}$. One can check that

$$\mathbf{x}_1 = P\mathbf{x} \in \mathcal{S} \text{ and } \mathbf{x}_2 = (I - P)\mathbf{x} \in \mathcal{V}\,. \tag{A21}$$

Moreover, this decomposition is unique. If $\mathbf{x} = \tilde{\mathbf{x}}_1 + \tilde{\mathbf{x}}_2$, with $\tilde{\mathbf{x}}_1 \in \mathcal{S}$, and $\tilde{\mathbf{x}}_2 \in \mathcal{V}$, then, $\mathcal{S} \ni \tilde{\mathbf{x}}_1 - \mathbf{x}_1 = \mathbf{x}_2 - \tilde{\mathbf{x}}_2 \in \mathcal{V}$. Therefore, since $\tilde{\mathbf{x}}_1 - \mathbf{x}_1$ belongs to the screen, there exists $\mathbf{h} \in \mathbb{R}^n$ such that $P\mathbf{h} = \tilde{\mathbf{x}}_1 - \mathbf{x}_1$. Now,

$$\begin{aligned}\tilde{\mathbf{x}}_1 - \mathbf{x}_1 &= P\mathbf{h} = P^2\mathbf{h} = P(P(\mathbf{h}))\\ &= P(\tilde{\mathbf{x}}_1 - \mathbf{x}_1) = P(\mathbf{x}_2 - \tilde{\mathbf{x}}_2) = 0\,.\end{aligned}$$

Therefore, $\tilde{\mathbf{x}}_1 = \mathbf{x}_1$ and $\tilde{\mathbf{x}}_2 = \mathbf{x}_2$, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are unique.

We can write that

$$\mathbb{R}^n = \mathcal{S} \oplus \mathcal{V}\,, \tag{A22}$$

that is, $\mathbb{R}^n$ is the direct sum of $\mathcal{S}$ and $\mathcal{V}$. Moreover, when P is symmetric, $\mathcal{S}$ and $\mathcal{V}$ are orthogonal, and we write,

$$\mathbb{R}^n = \mathcal{S} \oplus^{\perp} \mathcal{V}\,.$$

Given $\mathbf{x} \in \mathcal{S} \oplus^{\perp} \mathcal{V}$ and its unique decomposition $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, $\mathbf{x}_1 \in \mathcal{S}$ and $\mathbf{x}_2 \in \mathcal{V}$, we can interpret $\mathbf{x}_1$ and $\mathbf{x}_2$ as the sides of a right triangle with its hypothenuse being $\mathbf{x}$, and therefore, $|\mathbf{x}|^2 = |\mathbf{x}_1|^2 + |\mathbf{x}_2|^2$

Now, P restricted to the screen, $P|_{\mathcal{S}}$, is the identity, and restricted to the directions of projection, $P|_{\mathcal{V}}$, is the null operator. In fact, if $\mathbf{u}_1 \in \mathcal{S} = \mathrm{Im}(P)$, there exists $\mathbf{h}_1$ such that $P\mathbf{h}_1 = \mathbf{u}_1$, and then,

$$P|_{\mathcal{S}}(\mathbf{u}_1) = P\mathbf{u}_1 = P(P(\mathbf{h}_1)) = P^2\mathbf{h}_1 = P\mathbf{h}_1 = \mathbf{u}_1\,.$$

Since it is the identity in $\mathrm{Im}(P)$, it is an isometry.

This gives a simple explanation of a projection. From Eq. (A22), given any $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{u}$ can be written as $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, with $\mathbf{u}_1 \in \mathcal{S}$ and $\mathbf{u}_2 \in \mathcal{V}$, and $P(\mathbf{u}) = \mathbf{u}_1$.

The spectrum of a projection matrix has only 0's and 1's,

$$\sigma_P \subset \{0,1\}\,.$$

If P is a projection, but it is not the zero matrix nor the identity, then

$$\sigma_P = \{0,1\}\,.$$

A.4.2 Forgetting Matrices

Here we introduce a generalization of projection matrices that has interesting properties, in a fashion similar to projection matrices, preserving certain subspaces.

Definition A.3. A real $m \times n$ matrix A is called a *forgetting matrix* if and only if $A^T A$ is a projection matrix.

Theorem 9. A matrix is forgetting if and only if its singular values are 0's and 1's.
Proof. Let A be a forgetting $m \times n$ matrix and let its singular value decomposition be given by

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

where U is a $m \times m$ orthogonal matrix, V is a $n \times n$ orthogonal matrix and Σ is a possibly retangular $m \times n$ matrix with the singular values in the main diagonal and null outside the diagonal. Therefore,

$$A^T A = V\Sigma^T \overbrace{U^T U}^{=I} \Sigma V^T = V\Sigma^T \Sigma V^T$$

Note that $\Sigma^T \Sigma$ is a diagonal matrix, with the form

$$\begin{pmatrix} \sigma_1^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma_2^2 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_k^2 \end{pmatrix} \text{ or } \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & \vdots & & \vdots \\ \vdots & & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & \sigma_k^2 & & 0 \\ \vdots & & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} ,$$

where $k = \min\{m,n\}$. The n eigenvalues of $A^T A$, $\lambda_1, \dots, \lambda_n$ are related to the singular values of A. Since, by hypothesis, $A^T A$ is a projection matrix, its eigenvalues can be 0's or 1's. And so, the singular values of A are also 0's and 1's.

The converse is that if $\sigma_A \subset \{0,1\}$, then it is a forgetting matrix. In fact, since $A^T A$ is symmetric, the spectral theorem guarantees the existence of an orthogonal matrix V and a diagonal matrix Γ with 0's and 1's in the diagonal, such that $A^T A = V\Gamma V^T$. Therefore,

$$(A^T A)^2 = A^T A A^T A = V\Gamma V^T V\Gamma V^T = V\Gamma^2 V^T = V\Gamma V^T = A^T A \ ,$$

which implies that $A^T A$ is a projection matrix. Therefore, A is a forgetting matrix. ■

Let F be an $m \times n$ forgetting matrix. Likewise to the projection matrix, define

(a) the *screen* of F, as the set $\mathcal{S} = \text{Im}(F) \subset \mathbb{R}^m$;

(b) the *forgetting directions* of F, as the set $\mathcal{V} = N(F) \subset \mathbb{R}^n$.

In $\mathbb{R}^n$, given a vector subspace, U, we denote by $U^\perp$, (reads U 'perp') the subspace of all vectors orthogonal to each vector of U,

$$U^\perp = \{\mathbf{x} \in \mathbb{R}^n \text{ such that } \langle \mathbf{x}, \mathbf{u} \rangle = 0, \text{ for al } \mathbf{u} \in U\}$$

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be vectors in $\mathbb{R}^n$. By the vector *subspace generated* by those vectors we mean the set of all *linear combinations* of them,

$$\begin{aligned}&\mathsf{span}\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}\\ =&\{a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots + a_k\mathbf{v}_k, \text{ for all } a_i \in \mathbb{R}, i = 1,2,\ldots,k\}\,.\end{aligned}$$

Theorem 10. Let F be a forgetting matrix, and let

$$\mathcal{Y} = \mathcal{V}^\perp$$

where $\mathcal{V}$ is the forgetting directions subspace of F. Then, F restricted to $\mathcal{Y}$ is an isometry.

Proof. In fact, consider the singular value decomposition of F, $F = U\Sigma V^T$. Here,

$$U = \begin{pmatrix} | & \cdots & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_m \\ | & \cdots & | \end{pmatrix} \quad \text{and} \quad V^T = \begin{pmatrix} — & \mathbf{v}_1^T & — \\ \cdots & \cdots & \cdots \\ — & \mathbf{v}_n^T & — \end{pmatrix},$$

and Σ is the matrix whose possibly non-zero entries are in the main diagonal and represent the singular values of F. Let k be the number of singular values of F equal to 1. Matrix F can be written as

$$F = \mathbf{u}_1\mathbf{v}_1^T + \mathbf{u}_2\mathbf{v}_2^T + \ldots + \mathbf{u}_k\mathbf{v}_k^T\,.$$

With the notation set forth,

$$\begin{aligned}\mathcal{Y} &= \mathsf{span}\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}\,,\\ \mathcal{V} &= \mathsf{span}\,\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}\,,\\ \mathcal{S} &= \mathsf{span}\,\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}\,.\end{aligned} \tag{A23}$$

Any $\mathbf{y} \in \mathcal{Y} = \mathcal{V}^\perp$ can be written as

$$\mathbf{y} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots + a_k\mathbf{v}_k\,,$$

for appropriate values of the constants $a_1, a_2, \ldots, a_k$. Now,

$$\begin{aligned}|\mathbf{y}|^2 &= \langle \sum_{i=1}^k a_i\mathbf{v}_i, \sum_{j=1}^k a_j\mathbf{v}_j \rangle\\ &= \sum_i \sum_j a_i a_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_i \sum_j a_i a_j \delta_{ij} = \sum_{i=1}^k a_i^2\end{aligned}$$

Moreover,

$$\begin{aligned} F(\mathbf{y}) &= \left(\sum_{j=1}^{k} \mathbf{u}_j \mathbf{v}_j^T\right)\left(\sum_{i=1}^{k} a_i \mathbf{v}_i\right) \\ &= \sum_{j=1}^{k}\sum_{i=1}^{k} a_i \mathbf{u}_i \mathbf{v}_j^T \mathbf{v}_i \\ &= \sum_{j=1}^{k}\sum_{i=1}^{k} a_i \mathbf{u}_j \delta_{ji} = \sum_{j=1}^{k} a_j \mathbf{u}_j \,, \end{aligned}$$

and then,

$$|F(\mathbf{y})| = \left(\sum_{i=1}^{k} a_i^2\right)^{\frac{1}{2}} \,. \tag{A24}$$

Therefore, $|F(\mathbf{y})| = |\mathbf{y}|$ and this implies that F, restricted to $\mathcal{V}^{\perp}$, is an isometry. ■

A.5 Taylor's Formulae

We state here a few of *Taylor's* formulae. For a presentation with proofs in the general context of normed vector spaces see [23].

First of all, we recall the meaning of big-$\mathcal{O}$ and little-o, the order symbols. We write that

$$\mathbf{h}(\mathbf{y}) = \mathcal{O}(\mathbf{g}(\mathbf{y}))\,, \text{ as } \mathbf{y} \to \mathbf{x}\,,$$

if and only if, there is a constant $M > 0$ such that

$$\frac{|\mathbf{h}(\mathbf{y})|}{|\mathbf{g}(\mathbf{y})|} \le M\,,$$

for all $\mathbf{y} \neq \mathbf{x}$ in a neighbourhood of $\mathbf{x}$. Also, $\mathbf{h}(\mathbf{y}) = \mathbf{f}(\mathbf{y}) + \mathcal{O}(\mathbf{g}(\mathbf{y}))$, as $\mathbf{y} \to \mathbf{x}$, if and only if $\mathbf{h}(\mathbf{y}) - \mathbf{f}(\mathbf{y}) = \mathcal{O}(\mathbf{g}(\mathbf{y}))$, as $\mathbf{y} \to \mathbf{x}$.

We say that

$$\mathbf{h}(\mathbf{y}) = o(\mathbf{g}(\mathbf{y}))\,, \text{ as } \mathbf{y} \to \mathbf{x}\,,$$

if and only if,

$$\lim_{y \to x} \frac{|\mathbf{h}(\mathbf{y})|}{|\mathbf{g}(\mathbf{y})|} = 0\,.$$

Note that for $\alpha > 0$ (for instance, $\alpha = 1$)

$$\begin{aligned} &\mathbf{f}(\mathbf{y}) = \mathcal{O}\left(|\mathbf{y}-\mathbf{x}|^{k+\alpha}\right)\,, \text{ as } \mathbf{y} \to \mathbf{x}\,, \text{ implies that} \\ &\mathbf{f}(\mathbf{y}) = o\left(|\mathbf{y}-\mathbf{x}|^{k}\right)\,, \text{ as } \mathbf{y} \to \mathbf{x}\,. \end{aligned} \tag{A25}$$

Let **f** be a vector valued function of several variables,

$$\begin{aligned} \mathbb{R}^n \supset \Omega &\to \mathbb{R}^m \\ \mathbf{x} &\mapsto \mathbf{f}(\mathbf{x}) \,, \end{aligned}$$

where

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{pmatrix} .$$

We collect the first order derivatives of **f**, whenever they exist, in the *Jacobian* matrix,

$$\mathcal{J}\mathbf{f}_x = \left.\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}\right|_{(x_1, x_2, \ldots, x_n)} .$$

When $m = 1$, the Jacobian matrix is represented by the gradient,

$$\nabla f_x = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right) .$$

Consider now the second order derivatives of the k^{th} entry of **f**, f_k, collected in the *Hessian* matrix

$$\mathcal{H}(f_k)|_x = \left.\begin{pmatrix} \frac{\partial^2 f_k}{\partial x_1^2} & \frac{\partial^2 f_k}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f_k}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f_k}{\partial x_2 \partial x_1} & \frac{\partial^2 f_k}{\partial x_2^2} & \cdots & \frac{\partial^2 f_k}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_k}{\partial x_n \partial x_1} & \frac{\partial^2 f_k}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f_k}{\partial x_n^2} \end{pmatrix}\right|_{(x_1, x_2, \ldots, x_n)} .$$

Given $\mathbf{h} \in \mathbb{R}^n$, we have

$$\mathbf{h}^T \mathcal{H}(f_k)|_x \mathbf{h} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_i \frac{\partial^2 f_k}{\partial x_i \partial x_j} h_j .$$

Denote the second order derivative bilinear operator of **f** at **x** applied to vector **h** by

$$\mathbb{R}^n \ni \mathbf{h} \mapsto \mathbf{h}^T \mathcal{H}(\mathbf{f})|_x \mathbf{h} \in \mathbb{R}^m ,$$

where

$$\mathbf{h}^T \mathcal{H}(\mathbf{f})|_x \mathbf{h} = \left(\mathbf{h}^T \mathcal{H}(f_1)|_x \mathbf{h}, \mathbf{h}^T \mathcal{H}(f_2)|_x \mathbf{h}, \ldots \mathbf{h}^T \mathcal{H}(f_m)|_x \mathbf{h} \right)^T .$$

A function $\mathbb{R}^n \supset \Omega \xrightarrow{\mathbf{f}} \mathbb{R}^m$ is called *differentiable* at $\mathbf{x} \in \Omega$ if and only if

$$\mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x}) + \mathcal{J}\mathbf{f}_x(\mathbf{y} - \mathbf{x}) + o(|\mathbf{y} - \mathbf{x}|), \ \text{ as } \mathbf{y} \to \mathbf{x} .$$

Theorem 11. Taylor's formulae. Let $\mathbf{f}$ be a vector valued function of several variables,

$$\begin{aligned} \mathbb{R}^n \supset \Omega &\to \mathbb{R}^m \\ \mathbf{x} &\mapsto \mathbf{f}(\mathbf{x}) \end{aligned}$$

(a) (mean value inequality — first derivative) Assume that $\mathbf{f}$ is of class C^2, that is, its derivatives of first and second order exist and are continuous. Then,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| \leq \sup_{0<t<1} |\mathcal{J}\mathbf{f}_{(1-t)x+ty}| \cdot |\mathbf{y} - \mathbf{x}| \,.$$

(b) (mean value inequality — second derivative) Assume that $\mathbf{f}$ is of class C^3, that is, its derivatives of first, second and third order exist and are continuous. Then,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) - \mathcal{J}\mathbf{f}_x(\mathbf{y} - \mathbf{x})| \leq \frac{1}{2} \sup_{0<t<1} |\mathcal{H}(\mathbf{f})|_{(1-t)x+ty}| \, |\mathbf{y} - \mathbf{x}|^2 \,.$$

(c) (mean value theorem — first derivative) Assume[2] that $m = 1$ and f is of class C^2, $\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}$. Then, there is $t \in]0,1[$, such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f_{(1-t)x+ty} \cdot (\mathbf{y} - \mathbf{x}) \;.$$

(d) (mean value theorem — second derivative) Assume again that $m = 1$, and that f is of class C^2, $\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}$. Then, there is $0 < t < 1$, such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f_x \cdot (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^T \, \mathcal{H}(f)|_{(1-t)x+ty} \, (\mathbf{y} - \mathbf{x}) \;.$$

■

We remark that the important point in order to get equalities and not inequalities, in Taylor's formulae, is the dimension of the codomain, which for equality must be one, and not the domain. See Exercise A.32.

Taylor's formulae give precise results. Sometimes, a simpler, less informative version, might be useful. This can be accomplished by use of big-O notation. In fact, we have the following results:

(a) Assume that $\mathbf{f}$ is of class C^2, then,

$$\mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x}) + \mathcal{J}\mathbf{f}_x \cdot (\mathbf{y} - \mathbf{x}) + O\left(|\mathbf{y} - \mathbf{x}|^2\right), \text{ when } \mathbf{y} \to \mathbf{x} \,. \tag{A26a}$$

(b) Assume that $\mathbf{f}$ is of class C^3, then,

$$\begin{aligned} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x}) + \mathcal{J}\mathbf{f}_x \cdot (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^T \, \mathcal{H}(\mathbf{f})|_x \, (\mathbf{y} - \mathbf{x}) \\ + O\left(|\mathbf{y} - \mathbf{x}|^3\right), \text{ when } \mathbf{y} \to \mathbf{x} \,. \end{aligned} \tag{A26b}$$

[2] When the function assumes real values, $m = 1$, items (c) and (d) can be improved to an equality. Also, in this case, $\mathcal{J} f_x$ is just the gradient and the second derivative is given by the Hessian matrix, $\mathcal{H}(f)$.

Exercises

A.1. Transpose. Denote by $\mathbf{e}_i \in \mathbb{R}^n$ the i^{th} *canonical vector*, that is, all entries of $\mathbf{e}_i$ are zero except the i^{th} entry which equals to 1. Let $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$, $i,j = 1,2,\ldots,n$ in

$$\langle B\mathbf{x},\mathbf{y}\rangle = \langle \mathbf{x},C\mathbf{y}\rangle\,,$$

and, from it, show that the way a matrix changes its position in the inner product is by its transpose, $C = B^T$, therefore giving another proof of this result.

A.2. Cauchy-Schwartz inequality. This exercise is a guide to prove that $|\langle \mathbf{x},\mathbf{y}\rangle| \leq |\mathbf{x}|\,|\mathbf{y}|$.

(a) Given two vectors $\mathbf{a},\mathbf{b} \in \mathbb{R}^n$, orthogonal, $\mathbf{a} \perp \mathbf{b}$, then $\mathbf{a}$ and $\mathbf{b}$ are the sides of a right triangle, the *catheti*, with $\mathbf{c} = \mathbf{a} + \mathbf{b}$, its *hypothenuse*. Show that

$$|\mathbf{a}|^2 + |\mathbf{b}|^2 \quad = \quad |\mathbf{c}|^2\,.$$

Hint. Develop $\langle \mathbf{a}+\mathbf{b},\mathbf{a}+\mathbf{b}\rangle$.

(b) Given vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, such that $\mathbf{x} \neq 0$, show that $\mathbf{y}$ is the sum of a vector in the direction of $\mathbf{x}$, $\mathbf{v}_1$, and another one, $\mathbf{v}_2$, orthogonal to $\mathbf{x}$, $\mathbf{v}_1 \perp \mathbf{v}_2$, explicitly,

$$\mathbf{y} = \overbrace{\frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle}\mathbf{x}}^{\mathbf{v}_1} + \overbrace{\mathbf{y} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle}\mathbf{x}}^{\mathbf{v}_2}\;.$$

(c) Using (a) and (b), show that

$$|\mathbf{y}|^2 = \frac{|\langle \mathbf{x},\mathbf{y}\rangle|^2}{\langle \mathbf{x},\mathbf{x}\rangle} + |\mathbf{v}_2|^2\,,$$

and, therefore, since $|\mathbf{v}_2|^2 \geq 0$, show that

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leq |\mathbf{x}|\,|\mathbf{y}|\,.$$

A.3. Let $\mathbf{c}(t) = a(t)\mathbf{v}_1 + b(t)\mathbf{v}_2$, where $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1,2$ are orthogonal. Show that $|\mathbf{c}(t)|^2 = a^2(t) + b^2(t)$.

A.4. Norm. In general, a *norm* in $\mathbb{R}^n$ is a function with the ‘strange’ looking symbol $|\cdot|$,

$$\begin{aligned} \mathbb{R}^n &\to \mathbb{R} \\ \mathbf{x} &\mapsto |\mathbf{x}|\,, \end{aligned}$$

which satisfies the following properties,

(i) (positivity) $|\mathbf{x}| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and $|\mathbf{x}| = 0$ if and only if $\mathbf{x} = 0$.

(ii) (homogeneity of degree 1) For all $\lambda \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$, $|\lambda\mathbf{x}| = |\lambda|\,|\mathbf{x}|$.

(iii) (triangle inequality) For all pairs of vectors, $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$,

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}| \,. \tag{A27}$$

(a) Show that

$$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \,,$$

is a norm in $\mathbb{R}^n$.

(b) Show that, for a general norm, $|\mathbf{a} - \mathbf{b}| \leq |\mathbf{a} - \mathbf{c}| + |\mathbf{c} - \mathbf{b}|$

Hint. Here is a guide to prove the triangle inequality for norm defined by Eq. (A1).

(i) Write explicitly $|\mathbf{x} + \mathbf{y}|^2$.

(ii) Do the same for $(|\mathbf{x}| + |\mathbf{y}|)^2$.

(iii) Show the triangle inequality.

A.5. Inner product. An *inner product* in $\mathbb{R}^n$ is a function with a funny looking symbol, $\langle \cdot \,, \cdot \rangle$,

$$\begin{aligned} \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ (\mathbf{x},\mathbf{y}) &\mapsto \langle \mathbf{x},\mathbf{y} \rangle \end{aligned}$$

satisfying the following conditions

(i) (positivity) For all vectors $\mathbf{x} \in \mathbb{R}^n$,

$$\langle \mathbf{x},\mathbf{x} \rangle \geq 0 \,, \text{ and } \langle \mathbf{x},\mathbf{x} \rangle = 0 \text{ if and only if } \mathbf{x} = 0 \,. \tag{A28a}$$

(ii) (bilinearity) For all vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$, and all scalars $a_1,a_2,b_1,b_2 \in \mathbb{R}$,

$$\begin{aligned} \langle a_1\mathbf{x}_1 + a_2\mathbf{x}_2, b_1\mathbf{y}_1 + b_2\mathbf{y}_2 \rangle &= a_1b_1\langle \mathbf{x}_1,\mathbf{y}_1 \rangle \\ &+ a_1b_2\langle \mathbf{x}_1,\mathbf{y}_2 \rangle + a_2b_1\langle \mathbf{x}_2,\mathbf{y}_1 \rangle + a_2b_2\langle \mathbf{x}_2,\mathbf{y}_2 \rangle \end{aligned} \tag{A28b}$$

(iii) (symmetry) For all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$,

$$\langle \mathbf{x},\mathbf{y} \rangle = \langle \mathbf{y},\mathbf{x} \rangle \tag{A28c}$$

(a) Show that, in general, the bilinearity, Eq. (A28b), follows from symmetry, Eq. (A28c), and linearity in the first slot in the inner product, i.e.,

$$\begin{aligned} \langle a\mathbf{x},\mathbf{y} \rangle &= a\langle \mathbf{x},\mathbf{y} \rangle \,, \\ \langle \mathbf{x}_1 + \mathbf{x}_2,\mathbf{y} \rangle &= \langle \mathbf{x}_1,\mathbf{y} \rangle + \langle \mathbf{x}_2,\mathbf{y} \rangle \,. \end{aligned}$$

(b) Show Cauchy-Schwartz inequality,

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leq |\langle \mathbf{x},\mathbf{x}\rangle|^{\frac{1}{2}} |\langle \mathbf{y},\mathbf{y}\rangle|^{\frac{1}{2}} .$$

(c) (*Fourier-Pythagoras trick*) For a general inner product, we say that $\mathbf{x}$ is *orthogonal* to $\mathbf{y}$, and denote this by $\mathbf{x} \perp \mathbf{y}$ if and only if $\langle \mathbf{x},\mathbf{y}\rangle = 0$. Given a set of non-null orthogonal vectors $\mathbf{p}_1,\mathbf{p}_2,\dots,\mathbf{p}_n$, and a vector $\mathbf{x}$,

$$\mathbf{x} = a_1\mathbf{p}_1 + a_2\mathbf{p}_2 + \dots + a_n\mathbf{p}_n ,$$

find a simple expression for the constants a_i, $i = 1,2,\dots,n$, by making the inner product of $\mathbf{p}_i$ with both sides of the previous equation.

(d) Given a real, symmetric, positive definite, $n \times n$ matrix A, show that

$$\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}^T A\mathbf{y}$$

defines an inner product.

In the case that the inner product is defined by matrix A, we denote the orthogonality between $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x} \perp_A \mathbf{y}$, and say that $\mathbf{x}$ is $A-$orthogonal to $\mathbf{y}$.

(e) Given an inner product, show that

$$|\mathbf{x}| = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} ,$$

is a norm in the sense defined in Exercise A.4.

(f) Redo Exercise A.1 for the inner product defined in item (d).

A.6. Norm of a matrix. Given a real matrix A, $m \times n$, define the function

$$\mathbb{R}^n\backslash\{0\} \ni \mathbf{x} \mapsto q(\mathbf{x}) = \frac{|A\mathbf{x}|}{|\mathbf{x}|} .$$

(a) Show that q is a homogeneous function of degree zero.

(b) Show that

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

(c) Let D be a $n \times n$ diagonal matrix and d_i, $i = 1,2,\dots,n$ be its diagonal entries. Show that

$$|D| = \max_{i=1,2,\dots,n} |d_i| .$$

A.7. Show the validity of Eq. (A15).

A.8. (a) Given a square matrix A, show that the spectrum is *translated* when A is *translated*, that is,

$$\sigma_{A+k\mathcal{I}} \quad = \quad k + \sigma_A \ .$$

Here, if S is a subset of $\mathbb{C}$, $k + S = \{s + k.$ for all $s \in S\}$, is the *translation* of set S by k.

(b) Show that eigenvectors remain the same, that is, if $\mathbf{v} \neq 0$ is an eigenvector associated with eigenvalue λ of A, then $\mathbf{v}$ still remains an eigenvector of $A+k\mathcal{I}$.

(c) Given an invertible matrix, and $\sigma_A = \{\lambda_1,\lambda_2,\ldots,\lambda_l\}$, show that

$$\sigma_{A^{-1}} \quad = \quad \left\{\frac{1}{\lambda_1},\frac{1}{\lambda_2},\ldots,\frac{1}{\lambda_l}\right\} .$$

(d) Discuss the eigenvectors of A^{-1}.

A.9. Gerschgorin's circles theorem. Let A be a square matrix with real or complex entries, $A = (a_{ij})_{i,j=1,\ldots,n}$. Every eigenvalue of a matrix A, $n\times n$, be it real or complex, is inside (belongs to) the union of *Gerschgorin's* $C_1, C_2, \ldots, C_n$, where C_i is the circle, on the complex plane centered in a_{ii}, the i^{th} element of the main diagonal of matrix A, and radius given by

$$r_i = \sum_{j=1;j\neq i}^{n} |a_{ij}| \ .$$

which is the sum of the absolute values of the remaining entries in the i^{th} line. In symbols,

$$C_i = \{z = x + iy \in \mathbb{C}, \text{ such that } |z - a_{ii}| \leq r_i\} \, ,$$

and,

$$\sigma_A \subset \cup_{i=1}^{n} C_i \ .$$

(a) Given the matrix,

$$D = \begin{pmatrix} 4 & 2 & 1 \\ 1 & 5 & 3 \\ 2 & 4 & 7 \end{pmatrix} ,$$

determine centers and radii of Gerschgorin's circles.

(b) A matrix is *diagonally dominant* when each diagonal entry exceeds the sum of the absolute values of the remaining entries in that line. Is matrix D diagonally dominant?

(c) Recall that a matrix is invertible if and only if 0 is not one of its eigenvalues. Show that a diagonally dominant matrix is invertible.

(d) Given any square matrix A, show that there exists $k_0 \in \mathbb{R}$ such that $A + k\mathcal{I}$ is invertible for $k \geq k_0$.
Hint. Show that k_0 can be chosen such that $A + k_0\mathcal{I}$ is diagonally dominant.

(e) Show Gerschgorin's circles theorem.
Hint. Given an eigenvector associated with an eigenvalue, λ, at least one component, say the i^{th} component, is non-null. Dividing the eigenvector by the i^{th} entry, get another eigenvector, $\mathbf{v}$. Write the i^{th} equation from $A\mathbf{v} = \lambda\mathbf{v}$, and note that the absolute value of all entries of $\mathbf{v}$ are all less than or equal to 1.

A.10. Let A be a real matrix.

(a) Show that A^TA and AA^T are symmetric matrices.

(b) Show that $|A\mathbf{x}|^2 = \langle A\mathbf{x}, A\mathbf{x}\rangle = \langle A^TA\mathbf{x}, \mathbf{x}\rangle$.

(c) Given a real $m \times n$ matrix A, show that all eigenvalues of A^TA are real, non-negative.
Hint. Use spectral theorem and item (b).

(d) Use the spectral theorem to prove Eq. (A7).

(e) Assume that A is, furthermore, a symmetric matrix. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be its eigenvalues, and

$$\lambda_{abs} = \max_{i=1,2\ldots,n} |\lambda_i| \,.$$

Use the spectral theorem to show that

$$|A| = \lambda_{abs} \,.$$

A.11. (a) Verify that Eq. (A12) can be written as Eq. (A13).

(b) Sketch the region defined by Eq. (A11).

A.12. Use the spectral theorem to show that a symmetric matrix is positive definite if and only if all eigenvalues are strictly positive.

A.13. Ellipsoids. An hyper-ellipsoide in $\mathbb{R}^n$ is a hypersurface (*dimension* $n-1$, can be locally parametrized by $n - 1$ parameters) which can be transformed by a rigid motion to the *canonical* form[3]

$$\left(\frac{y_1}{a_1}\right)^2 + \left(\frac{y_2}{a_2}\right)^2 + \ldots + \left(\frac{y_n}{a_n}\right)^2 = 0 \,.$$

[3] The usual definition in two dimensions, that the sum of the distances of a point in the ellipse to two points, the *focci*, is constant, is too restrictive in $\mathbb{R}^n$, when $n > 2$.

Given a real, $n \times n$, symmetric, positive-definite matrix A and $\mathbf{b} \in \mathbb{R}^n$, show that the level sets of

$$E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}\,,$$

are either the empty set, a point (the solution of $A\mathbf{x} = \mathbf{b}$), or an ellipsoid.
Hint. By the spectral theorem, there exists an orthogonal matrix, P, and a diagonal matrix, D, such that $A = PDP^T$. Consider the change of variables $\mathbf{x} = P\mathbf{y}$, giving the function $F(\mathbf{y}) = E(P\mathbf{y})$, complete squares, and translate.

A.14. Metric spaces. A *metric* space in $\mathbb{R}^n$ is defined when there is a function,

$$d : \mathbb{R}^n \times \mathbb{R}^n \quad \to \quad \mathbb{R}$$

called a *metric* or *distance* function, such that, for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathbb{R}^n$,

(i) (positivity) $d(\mathbf{x},\mathbf{y}) \geq 0$, and $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$;

(ii) (symmetry) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$;

(iii) (triangle's inequality) $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$.

(a) In $\mathbb{R}^n$, we use the metric coming from a norm,

$$d(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}\,.$$

Show that it is a metric.

(b) Show that the function

$$d(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}|\,,$$

defined by means of a norm, is a distance in $\mathbb{R}^n$, (a so-called metric *induced* by the norm).

(c) The function

$$\rho(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2\,,$$

is not a metric. Show that it satisfies (i), and (ii), but fails to satisfy (iii).

(d) Given a symmetric, positive definite matrix M, show that

$$d(\mathbf{x},\mathbf{y}) = \left[(\mathbf{x} - \mathbf{y})^T M(\mathbf{x} - \mathbf{y})\right]^{\frac{1}{2}}\,,$$

is a metric.

A.15. Given an orthogonal matrix, P, show that it preserves distances and angles in the following way:

(a) Show that it preserves the inner-product

$$\langle P\mathbf{x}, P\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle \text{ for all } \mathbf{x}, \mathbf{y} .$$

(b) Show that it preserves the norm $|\mathbf{x}| = \langle \mathbf{x}, \mathbf{x}\rangle^{\frac{1}{2}}$, i.e.,

$$|P\mathbf{x}| = |\mathbf{x}| .$$

(c) Recall that the angle between two vectors, $\theta(\mathbf{x}, \mathbf{y})$, $0 \le \theta \le \pi$, is defined from the equation

$$\cos\theta(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{|\mathbf{x}||\mathbf{y}|} .$$

Show that $\theta(P\mathbf{x},P\mathbf{y}) = \theta(\mathbf{x},\mathbf{y})$.

A.16. Rigid motions. A function $G : U \to V$ is an *isometry* between metric spaces (sets with a distance) if and only if

$$d(G(\mathbf{x}),G(\mathbf{y})) = d(\mathbf{x}, \mathbf{y}) , \text{ for all } \mathbf{x},\mathbf{y} \in U . \tag{A29}$$

Let $G : \mathbb{R}^n \to \mathbb{R}^n$. Show that G is an isometry if and only if $G(\mathbf{x}) = U\mathbf{x} + \mathbf{b}$, where, U is an orthogonal matrix, and $\mathbf{b} \in \mathbb{R}^n$, that is, an isometry in $\mathbb{R}^n$ is an orthogonal transformation followed by a translation, the so-called *rigid motions*.

Hint. A general idea here is that if something is conserved, then derive it and at a certain point it will give zero. Something should come out. Differentiate Eq. (A29) with respect to x_i and y_j. A partial differential system of equations for G should come up. Its solution yields the result.

A.17. Find the singular value decomposition of

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & -1 \end{pmatrix} .$$

A.18. Let A be an $n \times 2$ matrix given by $A = (\mathbb{1}, \mathbb{1} - 2\mathbf{e}_n)$. Determine its singular value decomposition.

A.19. Show that $\mathbf{v}_1,\mathbf{v}_2, \ldots, \mathbf{v}_{n_o}$ defined in Eq. (A19a) are orthonormal.

A.20. (a) Show that P is a projection if and only if $I-P$ is a projection. Likewise, P is an orthogonal projection, if and only if $I - P$ is.

(b) If P is an orthogonal projection, show that its screen and its directions of projection spaces are orhogonal subspaces, i.e., show that $\mathcal{S} \perp \mathcal{V}$.

A.21. (a) Show that the xy-plane in $\mathbb{R}^3$ and the yz-plane are not orthogonal subspaces. (b) Show that the x-axis and the yz-plane are orthogonal subspaces.

A.22. Show the validity of Eq. (A21).

A.23. (a) Given a matrix A, show that $A(A^TA)^{-1}A^T$ is an orthogonal projection, whenever A^TA is invertible.

(b) Determine its screen and its directions of projection space.

A.24. (a) Given a projection matrix P, show that $\sigma_P \subset \{0,1\}$.

(b) Given an orthogonal matrix, show that its eigenvalues are complex numbers with absolute value 1.

A.25. Show that projection matrices and orthogonal matrices are forgetting matrices. (As with the identity matrix, which is a projection matrix, but does not 'project' anything, orthogonal matrices are forgetting matrices, but, do not 'forget' anything.)

A.26. Let Λ be a subset of $\{1,2,\ldots,n\}$ and $k = \#\Lambda=$'number of elements of Λ'. Define

$$\begin{aligned} \pi_\Lambda : \mathbb{R}^n &\to \mathbb{R}^k \\ \mathbf{x} &\mapsto \pi_\Lambda(\mathbf{x}) = (x_i)_{i\in\Lambda} \end{aligned}$$

For example, let $n = 5$ and $\Lambda = \{1,3,4\}$. Then

$$\pi_{\{1,3,4\}}(\mathbf{x}) = (x_1,x_3,x_4)^T \ .$$

(a) Determine the matrix F that represents $\pi_{\{1,3,4\}}$, i.e.,

$$\pi_{\{1,3,4\}}(\mathbf{x}) = F \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_3 \\ x_4 \end{pmatrix} .$$

(b) Show that F is a forgetting matrix.

(c) In general, let $\Lambda = \{i_1, i_2, \ldots, i_k\}$, with $i_1 < i_2 < \ldots < i_k$. Determine F such that

$$\pi_\Lambda(\mathbf{x}) = F\mathbf{x} \, .$$

(d) Show that F is a forgetting matrix.

A.27. Averaging is forgetting

(a) Show that $M = \frac{1}{n}\mathbb{1}\mathbb{1}^T$ is a forgetting matrix, and interpret the product $M\mathbf{x}$, for $\mathbf{x} \in \mathbb{R}^n$.

(b) Show that $K = \frac{1}{\sqrt{n}}\mathbb{1}^T$ is a forgetting matrix.

A.28. Let U be a subspace of $\mathbb{R}^n$. Show

$$\mathbb{R}^n \quad = \quad U \oplus^{\perp} U^{\perp} \tag{A30}$$

A.29. (a) Show the validity of Eq. (A23). (b) Show the validity of Eq. (A24).

A.30. Convex sets. Given two points in $\mathbb{R}^n$, $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, the *line segment* joining them is the set

$$[\mathbf{x},\mathbf{y}] \quad = \quad \{(1-t)\mathbf{x} + t\mathbf{y} \in \mathbb{R}^n,\ \text{for all } t \in [0,1]\}\ .$$

Note that if $\mathbf{x} = \mathbf{y}$, then $[\mathbf{x},\mathbf{y}] = \{\mathbf{x}\}$. A set $\mathcal{K} \subset \mathbb{R}^n$ is called *convex* set if and only if, for all $\mathbf{x},\mathbf{y} \in \mathcal{K}$, the line segment is contained in $\mathcal{K}$. In symbols,

$$[\mathbf{x},\mathbf{y}] \quad \subset \quad \mathcal{K}$$

(a) Consider the function

$$[0,1] \ni t \quad \mapsto \quad \mathbf{c}(t) = (1-t)\mathbf{x} + t\mathbf{y} \in \mathbb{R}^n\ ,$$

and compute $\mathbf{c}(0)$, $\mathbf{c}(1/2)$, $\mathbf{c}(1)$ and $\frac{d\mathbf{c}}{dt}$. What is the relation between the function $\mathbf{c}$ and the set $[\mathbf{x},\mathbf{y}]$?

(b) Let $\mathcal{H}^n = [0,+\infty[^n$. Show that $\mathcal{H}^n \subset \mathbb{R}^n$ is a convex set.

(c) Let

$$\begin{aligned} B_a(r) &= \text{'open ball of center } \mathbf{a} \text{ and radius } r \text{ in } \mathbb{R}^n\text{'} \\ &= \{\mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x}-\mathbf{a}| \le r\}\ . \end{aligned}$$

Show that $B_0(r)$ is a convex set.

A.31. Show the validity of Eq. (A25).

A.32. Most people are used to Taylor's formula for scalar functions, i.e., when the function has values in $\mathbb{R}$, $\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}$, in which case, given two points $\mathbf{x}$, $\mathbf{y}$, there exists a point in-between $\xi = (1-t)\mathbf{x} + t\mathbf{y}$, $0 < t < 1$ such that $f(\mathbf{y}) = f(\mathbf{x}) + \nabla f_\xi \cdot (\mathbf{y}-\mathbf{x})$. That this cannot be the case, in general, for vector valued functions with values in $\mathbb{R}^m$, with $m \ge 2$, can be seen in the following example. Given

$$\mathbb{R} \ni \theta \mapsto \mathbf{f}(\theta) = (\cos\theta, \sin\theta)\ ,$$

let $\theta_0 = 0$ and $\theta_1 = \pi$. Show that there is no ξ, $0 < \xi < \pi$, such that

$$\mathbf{f}(\pi) = \mathbf{f}(0) + \left.\frac{d\mathbf{f}}{d\theta}\right|_{\xi} \cdot (\pi - 0)\ .$$

We remark that the important thing here is the dimension of the codomain and not of the domain.

A.33. Use Theorem 11 to prove Eq. (A26).

A.34. Convex functions. Let $\mathcal{K} \to \mathbb{R}$ be a twice-continuously differentiable function defined on a convex set. Here, we can think of $\mathcal{K}$ as either $\mathbb{R}^n$ or $\mathcal{H}^n$.

A function F is called *convex* if and only if

$$F((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)F(\mathbf{x}) + tF(\mathbf{y})\,, \tag{A31}$$

for all $t \in [0,1]$, and all $\mathbf{x},\mathbf{y} \in \mathcal{K}$.

(a) The *graph* of a function $F : \mathcal{K} \to \mathbb{R}$ is the set

$$\mathsf{graph}(F) = \{(\mathbf{x}, F(\mathbf{x})) \in \mathcal{K} \times \mathbb{R} \mid \mathbf{x} \in \mathcal{K}\}\ ,$$

and the *epigraph* of a function is the set

$$\begin{aligned}\mathsf{epigraph}(F) &= \text{‘set of points \textit{above} the graph’}\\ &= \{(\mathbf{x}, y) \in \mathcal{K} \times \mathbb{R} \mid \mathbf{x} \in \mathcal{K} \text{ and } y \geq F(\mathbf{x})\}\ ,\end{aligned}$$

Show that F is a convex function if and only if $\mathsf{epigraph}(F)$ is convex.

(b) Let F be a convex function. Rearranging appropriately Eq. (A31) (taking all dependence on t to the left hand side of the equation), and letting $t \to 0$, show that

$$\langle \nabla F\,|_x, \mathbf{y} - \mathbf{x}\rangle \quad \leq \quad F(\mathbf{y}) - F(\mathbf{x})\,.$$

(c) Use the previous result and Taylor's formula to show that the Hessian of a convex function F,

$$H(F)\,|_x \quad = \quad \begin{pmatrix} \frac{\partial^2 F}{\partial x_1^2} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 F}{\partial x_n \partial x_n} \end{pmatrix},$$

is a *positive semidefinite* matrix, i.e., for all $\mathbf{z} \in \mathbb{R}^n$,

$$\mathbf{z}^T H(F)\,|_x\,\mathbf{z} \quad \geq \quad 0\,.$$

Hint. For $\mathbf{y}$ sufficiently close to $\mathbf{x}$, ($\mathbf{z} = \mathbf{y} - \mathbf{x}$), the second order term in Taylor's formula, dominates the higher order terms.

(d) A function F is called *strictly convex* if and only if

$$F((1-t)\mathbf{x} + t\mathbf{y}) \quad < \quad (1-t)F(\mathbf{x}) + tF(\mathbf{y})\,,$$

for all $t \in]0,1[$, and all $\mathbf{x},\mathbf{y} \in \mathcal{K}$, $\mathbf{x} \neq \mathbf{y}$. Show that if $H(F)|_y$ is positive definite for all $\mathbf{y} \in \mathcal{K}$, then F is strictly convex.

A.35. Bregmann divergence. Given a strictly convex differentiable function[4], $F : \mathcal{K} \to \mathbb{R}$, with $\mathcal{K} \subset \mathbb{R}^n$, a *Bregman* distance or *divergence* is a function defined by,

$$\begin{aligned} \mathcal{B}_F : \mathcal{K} \times \mathcal{K} &\to \mathbb{R} \\ (\mathbf{x},\mathbf{y}) &\mapsto \mathcal{B}_F(\mathbf{x},\mathbf{y}) = F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F \,|_y, \mathbf{x} - \mathbf{y} \rangle \,. \end{aligned}$$

(Here, divergence is used to mean a quantification of the difference of points, $\mathbf{x}$, with respect to another, $\mathbf{y}$, fixed point, as measured[5] by a function F.)

(a) Give an expression for $F(\mathbf{x})$, expanding it by Taylor's formula with a second order error term, around $\mathbf{y}$, and relate it with $\mathcal{B}_F(\mathbf{x},\mathbf{y})$.

(b) Show that Bregman divergence is non-negative, i.e., $\mathcal{B}_F(\mathbf{x},\mathbf{y}) \geq 0$ for all $\mathbf{x},\mathbf{y} \in \mathcal{K}$ and it is equal to zero if and only if $\mathbf{x} = \mathbf{y}$.

(c) Show that $S(\mathbf{x}) = |\mathbf{x}|^2$ is strictly convex.

(d) Let $\mathcal{B}_S$ denote the Bregman divergence corresponding to S. Show that $\mathcal{B}_S(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2$. Verify that $\mathcal{B}_S$ is not a distance (see Exercise A.14).

(e) Let M be a positive definite matrix, and consider the *inner product* defined by

$$\langle \mathbf{x},\mathbf{y} \rangle_M = \mathbf{x}^T M \mathbf{y} \,,$$

the related norm,

$$|\mathbf{x}|_M = \sqrt{\mathbf{x}^T M \mathbf{x}} \,.$$

and metric (distance),

$$d_M(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|_M = \sqrt{(\mathbf{x} - \mathbf{y})^T M (\mathbf{x} - \mathbf{y})} \,.$$

When M is the inverse of the covariance matrix of some probability distribution, d_M is called a Mahalanobis distance. Let $G(\mathbf{x}) = \mathbf{x}^T M \mathbf{x}$. Show that it generates a Bregman divergence, $\mathcal{B}_G$, and relate it to Mahalanobis distance.

A.36. (a) For the function below, show that it is strictly convex, and construct the corresponding Bregman divergence,

$$F(\mathbf{x}) = \sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} x_i \,,$$

$$B_F(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i} - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \,.$$

[4] We remark that a sufficient condition for a function to be strictly convex is given in Exercise A.34.

[5] For fixed $\mathbf{y}$, consider the hyperplane in $\mathbb{R}^n \times \mathbb{R}$, through $(\mathbf{y}, F(\mathbf{y}))$ and orthogonal to ∇F_y, defined by,

$$z = F(\mathbf{y}) + \langle \nabla F \,|_y, \mathbf{x} - \mathbf{y} \rangle \,,$$

for $\mathbf{x} \in \mathbb{R}^n$ and $z \in \mathbb{R}$. Clearly, z can be seen as a function of $\mathbf{x}$, $z = z(\mathbf{x})$. Note that the Bregman divergence, $\mathcal{B}_F(\mathbf{x},\mathbf{y})$ is just the height difference between the function F and the hyperplane, i.e., $\mathcal{B}_F(\mathbf{x},\mathbf{y}) = F(\mathbf{x}) - z(\mathbf{x})$.

(b) Do the same for the functions below,

$$G(\mathbf{x}) = -\sum_{i=1}^{n} \log x_i \,,$$

$$B_G(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \left(\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right) .$$

A.37. Given $x, x_0 \in]0, +\infty[$, in order to compare x relative to x_0, we consider the ratio x/x_0. Now, let $f = f(y)$ be a function of class C^2 such that

$$f(1) = 0,\ f'(y) < 0, \text{ for } y < 1, \text{ and } f'(y) > 0, \text{ for } y > 1\,. \tag{A32}$$

(a) Show that $u(y) = (y-1)^2$ satisfies Eq. (A32).

(b) Show that $v(y) = 1 - y + y \ln y$ satisfies Eq. (A32). Sketch the graph of v.

(c) Show that a function satisfying Eq. (A32) is a convex function.

(d) Show that it also satisfies $f'(1) = 0$.

(e) Show that $g(x) = f(\frac{x}{x_0})$ has the interesting property that it increases as x departs[6] away from x_0.

(f) Show that the same result is true for the function $h(x) = x_0 f(\frac{x}{x_0})$.

(g) Compute $s(x,y) = yv\left(\frac{x}{y}\right)$.

(h) Compute Bregman's divergence for v, $\mathcal{B}_v(x,y)$, and compare the result with $s(x,y)$.

(i) Do the same for u, i.e., compare $s(x,y) = yu\left(\frac{x}{y}\right)$ and $\mathcal{B}_u(x,y)$.

(j) Consider the following generalization of Bregman's divergence. Let $f = f(x,y)$ be a real-valued, strictly convex function in the first entry, defined in a subset of $\mathbb{R}^2$. Define

$$\mathcal{B}_f(x,y) = f(x,y) - f(y,y) - \left.\frac{\partial f}{\partial x}\right|_{(y,y)} (x-y)\,.$$

Let $f(x,y) = (x/y)^2$. Show that the corresponding Bregman's divergence is the square of the relative distance between x and y relative to y, i.e., $\mathcal{B}_f(x,y) = (\frac{x-y}{y})^2$.

A.38. Let $f = f(x)$ be a convex, real valued function defined on $\mathbb{R}$.

[6] This is a very simple property that one could require in order to compare x and x_0. It is also homogeneous of degree zero with respect to multiplying both x and x_0 by some constant value.

(a) Show that

$$F(\mathbf{x}) = f(x_1) + f(x_2) + \ldots + f(x_n)\,,$$

is also convex.

(b) Show that if $f''(x) > 0$ then $JF|_x$ is positive definite, and therefore $F = F(\mathbf{x})$ is strictly convex.

(c) Show that

$$\begin{aligned}\mathcal{B}_F(\mathbf{x},\mathbf{y}) &= f(x_1) + f(x_2) + \ldots + f(x_n) - [f(y_1) + f(y_2) + \ldots + f(y_n)]\\ &+ f'(y_1)(x_1 - y_1) + f'(y_2)(x_2 - y_2) + \ldots + f'(y_n)(x_n - y_n)\\ &= \sum_{i=1}^{n} [f(x_i) - f(y_i) - f'(y_i)(x_i - y_i)]\ .\end{aligned}$$

(d) Show that $F(\mathbf{x}) = x_1^2 + x_2^2 + \ldots + x_n^2$ is strictly convex.

(e) Show that $F(\mathbf{x}) = \sum_{i=1}^n [x_i \ln x_i + (1 - x_i)]$ is strictly convex.

(f) Show that $F(\mathbf{x}) = -\sum_{i=1}^n \ln x_i$ is strictly convex.

(g) Compute Bregman's divergence for functions in items (d), (e) and (f).

A.39. Consider the following generalization of Bregman's divergence. Let $f = f(\mathbf{x},\mathbf{y})$ be a strictly convex function in the first entry, real valued function defined on a subset of $\mathbb{R}^n \times \mathbb{R}^n$. Define

$$\mathcal{B}_f(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) - f(\mathbf{y},\mathbf{y}) - \nabla_x f|_{(y,y)} \cdot (\mathbf{x} - \mathbf{y})\,.$$

(a) Let $f(\mathbf{x},\mathbf{y}) = g(x_1,y_1) + g(x_2,y_2) + \ldots + g(x_n,y_n)$ and obtain an expression for $\mathcal{B}_f(\mathbf{x},\mathbf{y})$.

(b) For $g(x,y) = (x/y)^2$, determine $\mathcal{B}_f(\mathbf{x},\mathbf{y})$.

(c) Let $f(\mathbf{x},\mathbf{y}) = (|\mathbf{x}|/|\mathbf{y}|)^2$. Compute $\mathcal{B}_f(\mathbf{x},\mathbf{y})$.

Exercises: The Conjugate Gradient Method

The *conjugate gradient* method is, somewhat, geometrically simple. However, before simplicity is achieved, some previous work has to be done. Next we present a collection of exercises to pave the road of simplicity in its understanding.

A.40. Intersection of a linear space with an ellipsoide is an ellipsoide. Let A be a positive definite symmetric matrix. From Exercise A.13, we know that

$$\frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} = \text{`constant'}$$

is an ellipsoide.

(a) Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be two linearly independent vectors. Consider the restriction of the ellipsoide to the plane τ, consisting of points

$$\mathbf{x} = y_1\mathbf{p}_1 + y_2\mathbf{p}_2 = \overbrace{\begin{pmatrix} | & | \\ \mathbf{p}_1 & \mathbf{p}_2 \\ | & | \end{pmatrix}}^{P} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = P\mathbf{y}\,,$$

for all $y_1, y_2 \in \mathbb{R}$, that is the set of solutions $\mathbf{y} \in \mathbb{R}^2$ of the equation

$$\frac{1}{2}\mathbf{y}^T P^T A P\mathbf{y} - \mathbf{y}^T P^T \mathbf{b} = \text{`constant'}\,.$$

Show that this set is an ellipse, a single point, or the empty set.

(b) Propose and show a similar result for the restriction of the elipsoide to a k–dimensional subspace $\tau = \mathsf{span}\,\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k\}$ generated by k linearly independent vectors.

A.41. How to change places in an inner product. Let A be a symmetric, $n \times n$, positive definite matrix, and S an $n \times m$ matrix. Consider the inner product $\langle \mathbf{y}, \mathbf{z} \rangle_A = \mathbf{y}^T A\mathbf{z}$, for all $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$. Show that if

$$\langle S\mathbf{x}, \mathbf{y} \rangle_A = \langle \mathbf{x}, S^*\mathbf{y} \rangle_A\,,$$

for all $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, then $S^* = A^{-1}S^TA$.

A.42. Translation: completing squares

(a) By completing squares show that

$$\frac{a}{2}x^2 - bx = \frac{a}{2}\left(x - \frac{b}{a}\right)^2 - \frac{b^2}{2a}\,.$$

(b) Let $f(x) = \frac{a}{2}x^2 - bx$. Verify that the minimum point of f is b/a, and find the minimum of f.

(c) Obtain the critical point equation of f and its solution.

(d) Let $D = diag(a_1, a_2, \ldots, a_n)$ be a diagonal matrix with $a_1 \geq a_2 \geq \ldots \geq a_n > 0$. Let $\mathbf{b} = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$, and $f_i(x) = \frac{a_i}{2}x^2 - b_ix$, for $i = 1, 2, \ldots, n$. Complete the squares for the function

$$\begin{aligned} E(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^T D\mathbf{x} - \mathbf{x}^T\mathbf{b} \\ &= f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n)\,. \end{aligned}$$

(e) From the result in the previous item, determine the minimum point and the minimum of E.

(f) Write the critical point equation of E, as a system of linear equations, using matrix D.

(g) Assume that A is a positive definite symmetric matrix, and let P, orthogonal, and D, diagonal matrices, as given by the spectral theorem, $A = PDP^T$. Redo appropriately items (d) to (f) with A in the place of D.
Hint. Use the change of variables $\mathbf{y} = P^T\mathbf{x}$ twice.

A.43. Decoupling. Let A, P and D as in the spectral theorem, $A = PDP^T$, and consider the system of equations $A\mathbf{x} = \mathbf{b}$. Show that the equation satisfied by $\mathbf{y} = P^T\mathbf{x}$, coming from a change of variables, is *decoupled*, i.e., the equation for each y_i, does not depend on the other variables

A.44. Spheres and ellipsoides. Let A be a symmetric positive definite matrix. As in Exercise A.5 denote by $\langle\ ,\ \rangle_A$ the inner product $\langle \mathbf{x}\,,\mathbf{y}\rangle_A = \mathbf{x}^T A\mathbf{y}$. Consider the norm $|\mathbf{x}|_A = \langle \mathbf{x}\,,\mathbf{x}\rangle_A^{\frac{1}{2}}$, the distance $d_A(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|_A$ and the corresponding *sphere*, centered at $\mathbf{c}$, with radius r,

$$S_c^{n-1}(r) = \{\mathbf{x} \in \mathbb{R}^n \mid d_A(\mathbf{x},\mathbf{y})\}\ .$$

(a) Let

$$D = diag\left(\frac{1}{a^2},\frac{1}{b^2}\right) = \begin{pmatrix} a^{-2} & 0 \\ 0 & b^{-2} \end{pmatrix}\ .$$

Determine the analytical expression for the one-dimensional sphere, $S^1_{(x_0,y_0)}(1)$ i.e., the ‘circle’ of radius 1 and center (x_0,y_0) for the distance d_D, and represent it in the cartesian plane. What is it usually called?

(b) In $\mathbb{R}^3$, let $D = diag\left(\frac{1}{a^2},\frac{1}{b^2},\frac{1}{c^2}\right)$. Represent $S^2_{(x_0,y_0,z_0)}(1)$.

A.45. Rotated ellipse

(a) Consider the equation

$$(x-y)^2 + \frac{(x+y)^2}{36} = 1\ .$$

Represent it. **Hint.** Let $x - y = 0$ and find the *extrema* points of the ellipse on that line. Do the same for $x + y = 0$.

(b) Do the same for the equation

$$(x-y+1)^2 + \frac{(x+y-5)^2}{36} = 4\ . \qquad \text{(A33)}$$

(c) Equation (A33) is the level 4 set of a function F of the form

$$F(x,y) = \frac{1}{2}(x\,y)A\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}(x\,y)\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + c\,,$$

where A is a positive definite, symmetric matrix, $\mathbf{b} = (b_1,b_2)^T \in \mathbb{R}^2$ and $c \in \mathbb{R}$. Determine A, $\mathbf{b}$ and c.

(d) Show that the minimum point of F is the same as the minimum point of $E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}$.

A.46. Different gradients and the steepest descent method. Let $\mathbf{p} \in \mathbb{R}^n$ be a search direction, $\mathbf{x}_0 \in \mathbb{R}^n$ and $E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}$, with A a symmetric, positive definite matrix, and $\mathbf{b} \in \mathbb{R}^n$.

(a) Show that the minimum point of E along the line

$$\mathbb{R} \ni t \mapsto \mathbf{x}_0 + t\mathbf{p} \in \mathbb{R}^n \,,$$

is $\mathbf{x}_0 + t^*\mathbf{p}$ where

$$t^* = \frac{\mathbf{p}^T(\mathbf{b} - A\mathbf{x}_0)}{\mathbf{p}^T A\mathbf{p}} \,.$$

(b) The derivative of E at $\mathbf{x}_0$ applied to vector $\mathbf{p}$, $dE_{x_0}(\mathbf{p})$, i.e., the *directional derivative*, is given by

$$dE_{x_0}(\mathbf{p}) = \lim_{t\to 0} \frac{E(\mathbf{x}_0 + t\mathbf{p}) - E(\mathbf{x}_0)}{t} \,,$$

and the gradient of E at $\mathbf{x}_0$, ∇E_{x_0}, is the vector that represents the derivative, with respect to an inner product,

$$dE_{x_0}(\mathbf{p}) = \langle \nabla E_{x_0}, \mathbf{p}\rangle \,.$$

Show that

$$\nabla E_{x_0} = A\mathbf{x}_0 - \mathbf{b} \,,$$

when the inner product is $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}$, and

$$\nabla E_{x_0} = A^{-1}(A\mathbf{x}_0 - \mathbf{b}) \,,$$

when the inner product is $\langle \mathbf{x}, \mathbf{y}\rangle_A = \mathbf{x}^T A\mathbf{y}$.

(c) The steepest descent is a method that choses minus the gradient as the search direction. Use the previous items to show that applying the steepest descent method to E, we arrive at the solution of the system $A\mathbf{x} = \mathbf{b}$ in one step, no matter which is the initial guess, if we use the gradient coming from the inner product $\langle\ ,\ \rangle_A$. That is, show that

$$\mathbf{x}_0 + t^*\mathbf{p} = \mathbf{x}^* \,, \text{ where } \mathbf{x}_* = A^{-1}\mathbf{b} \,,$$

if $\mathbf{p} = -\nabla E_{x_0} = -A^{-1}(A\mathbf{x}_0 - \mathbf{b})$.

Remark. This seems too good to be true, and the catch is this: In order to find the search solution, $\mathbf{p} = -A^{-1}(A\mathbf{x}_0 - \mathbf{b}) = -\mathbf{x}_0 + A^{-1}\mathbf{b}$, we already have to find the solution $A^{-1}\mathbf{b}$. One can also check that in this case $t^* = 1$. This works because of the following three reasons:

(i) The level sets of E, as seen through the 'eyes' of the inner product $\langle\ ,\ \rangle_A$, are 'evenly' rounded spheres, centered at the solution of equation $A\mathbf{x} = \mathbf{b}$ (this fact is seen in item (e), Exercise A.54);

(ii) Given any initial guess $\mathbf{x}_0$, it will be in a level set of E;

(iii) The gradient of E at $\mathbf{x}_0$ is A-orthogonal to the level sets of E and its negative is directed towards the center of the sphere.

(d) Given a point $\mathbf{x}_0$, in a level set of E, which is a sphere, and a set of n A-orthogonal search directions, $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$, one can find a linear combination of the search directions,

$$c_1\mathbf{p}_1 + c_2\mathbf{p}_2 + \ldots + c_n\mathbf{p}_n$$

such that the center of the sphere $\mathbf{x}^*$ can be represented as

$$\mathbf{x}^* = \mathbf{x}_0 + c_1\mathbf{p}_1 + c_2\mathbf{p}_2 + \ldots + c_n\mathbf{p}_n\ ,$$

and constants c_i are determined independently of one another, in particular, can be determined in whatever order one wishes. Show that

$$\begin{aligned} F(\mathbf{c}) &= E(\mathbf{x}_0 + c_1\mathbf{p}_1 + c_2\mathbf{p}_2 + \ldots + c_n\mathbf{p}_n) \\ &= \alpha_0 + \sum_{i=1}^{n} \frac{1}{2} c_i^2 \langle \mathbf{p}_i, \mathbf{p}_i \rangle_A + \alpha_i c_i\ , \end{aligned}$$

with

$$\begin{aligned} \alpha_0 &= \frac{1}{2}\langle \mathbf{x}_0, \mathbf{x}_0\rangle_A - \mathbf{x}_0^T\mathbf{b} = \frac{1}{2}\mathbf{x}_0^T A\mathbf{x}_0 - \mathbf{x}_0^T\mathbf{b}\ , \\ \alpha_i &= \langle \mathbf{x}_0, \mathbf{p}_i\rangle_A - \mathbf{p}_i^T\mathbf{b} = \mathbf{p}_i^T(A\mathbf{x}_0 - \mathbf{b})\ . \end{aligned}$$

(e) Determine the constants c_i by finding the critical point equation of F.

A.47. Optimization of a *decoupled* function. Let $f_i : \mathbb{R} \to \mathbb{R}$, for $i = 1,2,\ldots,n$, be functions of class C^∞, and define

$$F(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n)\ .$$

Assume that F has a global point of minimum, which then is determined as the solution of the critical point equation. Show that the critical point of F can be obtained by considering, separately, the critical point of each f_i.

A.48. Drawing a method. In $\mathbb{R}^2$, let (u,v) and $(-v,u)$ be two orthogonal vectors. Let also $\mathbf{x}^* = (x^*,y^*)^T$ be the center of concentric circles (level sets). Let $\mathbf{x}_0 = (x_0,y_0)^T$ be an initial guess for $\mathbf{x}^*$. (This exercise is to solve $\mathcal{I}\mathbf{x} = \mathbf{x}^*$, i.e., finding $\mathbf{x} = \mathbf{x}^*$ by minimization along search directions).

(a) Minimize $E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathcal{I}\mathbf{x} - \mathbf{x}^T\mathbf{x}^*$, along the line $t \mapsto \mathbf{x}_0 + t(u,v)^T$. Let the minimum point be $\mathbf{x}_1 = \mathbf{x}_0 + t_1(u,v)^T$. Show that

$$t_1 = \frac{(u,v)(\mathbf{x}^* - \mathbf{x}_0)}{(u,v)(u,v)^T} \,.$$

(b) Show that $\mathbf{x}_1 - \mathbf{x}^* \perp \mathbf{x}_1 - \mathbf{x}_0$.

(c) Draw a picture of this, representing at least $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}^*$, (u,v), $(-v,u)$, the level sets $E(\mathbf{x}_0)$ and $E(\mathbf{x}_1)$, and the lines through $\mathbf{x}_0$ and $\mathbf{x}_1$ and through $\mathbf{x}_1$ and $\mathbf{x}^*$.

A.49. Curvature, Gauss map of a curve & all that. Let $\mathbb{R} \supset]a,b[\ni t \mapsto (c_1(t),c_2(t)) \in \mathbb{R}^2$ be a parametrization of a curve C in $\mathbb{R}^2$. Then $(c_1'(t),c_2'(t))^T$ is a tangent vector to the curve C at $(c_1(t),c_2(t))$, and

$$T = \frac{(c_1'(t),c_2'(t))^T}{\sqrt{(c_1')^2 + (c_2')^2}}$$

is a unit tangent vector. The vector field

$$N(t) = \frac{(-c_2'(t),c_1'(t))^T}{\sqrt{(c_1')^2 + (c_2')^2}}$$

is a unit normal vector field to the curve, with the added feature that, together with T, in the order (T,N), be a positively oriented basis of $\mathbb{R}^2$, that is, $\det(T\ N) > 0$. Recall that $S_0^1(1) = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$. Let $(x,y) \in C$, then there is $t \in]a,b[$ such that $x = c_1(t)$ and $y = c_2(t)$. Define

$$N : C \to S_0^1 \subset \mathbb{R}^2$$

in the following way, $N(x,y) = \frac{(-c_2'(t),c_1'(t))}{\sqrt{(c_1')^2+(c_2')^2}}$. Strictly speaking, we are using N in two different ways, but we hope this does not confuse the reader. Function N is called the *Gauss map* for the curve. It records the way the curve *bends*. Its rate of variation per unit length at point (x,y), in the direction of the tangent,

$$k = \frac{1}{\sqrt{(c_1')^2 + (c_2')^2}}\langle\frac{dN}{dt},T\rangle\,,$$

is called the signed *curvature* of the curve. The curvature is just $|k|$. The definition of the Gauss map, N, of a curve, and of its signed curvature, k, depends on the parametrization of the curve, differing, however, at most, by a minus sign.

(a) Show that the signed curvature is given by

$$k(t) = \frac{c_1'c_2'' - c_1''c_2'}{((c_1')^2 + (c_2')^2)^{3/2}} \,.$$

(b) Consider the line $ax + by = c$. Show that N is either

$$N = \frac{-(a,b)^T}{\sqrt{a^2+b^2}}\ , \text{ or } \frac{(a,b)^T}{\sqrt{a^2+b^2}}\ ,$$

and compute k.

(c) Given the circle of radius R and center (a,b),

$$S^1_{(a,b)}(R) = \left\{(x,y) \in \mathbb{R}^2 \mid (x-a)^2 + (y-b)^2 = R^2\right\}\ ,$$

and the usual parametrization given by

$$\begin{aligned} x(\theta) &= a + R\cos\theta \\ y(\theta) &= b + R\sin\theta \end{aligned}$$

determine T, N and k. Note that k is negative.

(d) Consider the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\ ,$$

with $a > b$, and its parametrization given by

$$\begin{aligned} x(t) &= a\cos t \\ y(t) &= b\sin t\ . \end{aligned}$$

Determine T, N and k. Also, determine the maximum and the mininum of its curvature, and where they occur.

A.50. Gauss map for surfaces. A smooth surface S in $\mathbb{R}^3$ is called *orientable* if it has a smooth unit normal vector field. The function

$$N : S \to S^2_0(1) = \left\{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\right\}$$

where, for each $\mathbf{x} \in S$, $N(\mathbf{x})$ is the unit normal vector to the surface S at $\mathbf{x}$, is called *Gauss map*. It records the way the surface 'bends' and 'twists' ('writhes').

If the surface S is given by an equation of the form

$$F(x,y,z) = d$$

where F is a smooth function, $F : \mathbb{R}^3 \to \mathbb{R}$, then S is just the level set of F, corresponding to level d. In this case, we recall from vector calculus, that the gradient

$$\nabla F = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}\right)\ ,$$

is normal to the level sets, pointing in the direction of local steepest ascent of F, and we can take $N = \nabla F / |\nabla F|$.

Let

$$\mathbb{R}^2 \supset U \ni (u,v) \overset{\mathbf{x}}{\mapsto} \mathbf{X}(u,v) = (x(u,v),y(u,v),z(u,v)) \in S \subset \mathbb{R}^3$$

be a parametrization (coordinates) of a piece of S. Given $(u_0,v_0) \in]a,b[\times]c,d[\subset U$, consider the *coordinate curves*,

$$\mathbb{R} \supset]a,b[\ni u \mapsto \mathbf{c}_1(u) = (x(u,v_0),y(u,v_0),z(u,v_0))^T \in S \subset \mathbb{R}^3\,, \text{ and}$$
$$\mathbb{R} \supset]c,d[\ni v \mapsto \mathbf{c}_2(v) = (x(u_0,v),y(u_0,v),z(u_0,v))^T \in S \subset \mathbb{R}^3\,,$$

and, in order to avoid the ambiguity of the previous definition, we define the Gauss map at $\mathbf{X}_0 = (x(u_0,v_0), y(u_0,v_0),z(u_0,v_0)) \in S$, $\mathbf{N}(\mathbf{X}_0)$, by the unit vector in the direction of $\mathbf{c}_1'(u_0) \times \mathbf{c}_2'(v_0)$. Since

$$\det(\mathbf{c}_1'(u_0)\, \mathbf{c}_2'(v_0)\, \mathbf{N}(\mathbf{X}_0))$$

is positive, we say that those vectors, in that order, constitute a positively oriented basis for $\mathbb{R}^3$, and that $(\mathbf{c}_1'(u_0)\, \mathbf{c}_2'(v_0))$ constitute a positively oriented basis for the tangent space of S at $\mathbf{X}_0 = (x(u_0,v_0), y(u_0,v_0),z(u_0,v_0))$.

(a) A special case of this occurs when S is parametrized as the graph of a function,

$$(x,y) \mapsto f(x,y) \in \mathbb{R}\,.$$

In this case, the parametrization is $\mathbf{X}(x,y) = (x,y,f(x,y))$. Show that

$$\mathbf{N} = \frac{(-\partial f/\partial x, -\partial f/\partial y, 1)}{\sqrt{1 + (\partial f/\partial x)^2 + (\partial f/\partial y)^2}}\,.$$

(b) (Plane) Compute $\mathbf{N}$ when S is a plane, say,

$$S = \left\{(x,y,z) \in \mathbb{R}^3 \mid ax + by + cz = d\right\}\,.$$

In this case, $\mathbf{N}$ is a constant (vector) function, given, up to a sign, by

$$\mathbf{N} = \frac{(a,b,c)}{\sqrt{a^2 + b^2 + c^2}}\,.$$

(c) (Sphere) Let S be a sphere,

$$S = S_0^2(R) = \left\{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = R^2\right\}$$

Compute $\mathbf{N}$ in three different ways by considering S as the level set of a function, by considering it as the graph of a function, and by means of the usual spherical coordinates parametrization,

$$]0,2\pi[\times]0,\pi[\ni (\theta,\phi) \mapsto (R\cos\theta\sin\phi, R\sin\theta\sin\phi, R\cos\phi) \in S \subset \mathbb{R}^3\,.$$

(d) (Cylinder) Consider the surface of a cylinder,

$$C = \left\{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 = r^2\right\} ,$$

and determine its Gauss map.

(e) (Ellipsoide) Determine the Gauss map of the ellipsoide,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 .$$

A.51. Metric tensor Given a surface S in $\mathbb{R}^3$, to go from one point of the surface to another over the surface cannot, in general, be accomplished by a straight line, due to the surface's curvature. Assume we are given a parametrization of a piece of the surface,

$$\mathbb{R}^2 \supset U \ni (u,v) \overset{\mathbf{X}}{\mapsto} \mathbf{X}(u,v) = (x(u,v),y(u,v),z(u,v)) \in S \subset \mathbb{R}^3 .$$

Consider the tangent vectors to the surface in $\mathbf{X}(u,v)$ given by the partial derivatives of $\mathbf{X}$,

$$\mathbf{t}_1 = \mathbf{X}_u = \frac{\partial \mathbf{X}}{\partial u} = \begin{pmatrix} \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial u} \end{pmatrix} , \quad \text{and} \quad \mathbf{t}_2 = \mathbf{X}_v = \frac{\partial \mathbf{X}}{\partial v} = \begin{pmatrix} \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial v} \end{pmatrix} .$$

Furthermore, assume that $\{\mathbf{X}_u,\mathbf{X}_v\}$ form a basis of the tangent plane to the surface S at the point $\mathbf{X}(u,v)$, denoted by $T_{X(u,v)}S$. Let $\alpha, \beta \in T_{X(u,v)}S$. Then, there are scalars a_1, a_2, b_1 and b_2 such that

$$\alpha = a_1\mathbf{t}_1 + a_2\mathbf{t}_2 , \quad \text{and} \quad \beta = b_1\mathbf{t}_1 + b_2\mathbf{t}_2 .$$

Show that:

(a) the inner product of α and β is

$$\alpha^T\beta = \langle\alpha, \beta\rangle = \mathbf{a}^T M\mathbf{b} = \langle\mathbf{a}, \mathbf{b}\rangle_M ,$$

where $\mathbf{a}^T = (a_1, a_2)$, $\mathbf{b}^T = (b_1, b_2)$,

$$M = (J\mathbf{X})^T J\mathbf{X} = \begin{pmatrix} \langle\mathbf{X}_u, \mathbf{X}_u\rangle & \langle\mathbf{X}_u, \mathbf{X}_v\rangle \\ \langle\mathbf{X}_v, \mathbf{X}_u\rangle & \langle\mathbf{X}_v, \mathbf{X}_v\rangle \end{pmatrix} , \tag{A34}$$

and

$$J\mathbf{X} = (\mathbf{X}_u\ \mathbf{X}_v) = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{pmatrix} .$$

Matrix M is called the *metric tensor* (and the Gram matrix of the basis $\{\mathbf{X}_u, \mathbf{X}_v\}$) and it is related to the first fundamental form of the surface [15].

(b) Assume a piece of surface $S \subset \mathbb{R}^3$ is parametrized as the graph of a function,

$$\mathbb{R}^2 \supset U \ni (x,y) \quad \mapsto \quad (x,y,z(x,y)) \in S \subset \mathbb{R}^3 \, .$$

Determine an expression for the metric tensor.

(c) Given a plane $ax + by + cz = d$ let $c \neq 0$ and parametrize it as a graph of a function, $z = z(x,y)$. Determine its metric tensor.

(d) Let a surface be given as the level set of a function,

$$G(x,y,z) = 0 \, .$$

Assume that z can be written implicitly as a function of x and y, $z = z(x,y)$. Show that

$$M = \frac{1}{(\partial G/\partial z)^2} \begin{pmatrix} (\partial G/\partial x)^2 + (\partial G/\partial z)^2 & (\partial G/\partial x)\,(\partial G/\partial y) \\ (\partial G/\partial x)\,(\partial G/\partial y) & (\partial G/\partial y)^2 + (\partial G/\partial z)^2 \end{pmatrix} .$$

(e) Given the sphere

$$S_0(r) = \left\{ (x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = r^2 \right\} ,$$

consider the standard spherical coordinates,

$$]0,2\pi[\times]0,\pi[\ni (\theta,\phi) \quad \mapsto \quad (x(\theta,\phi),y(\theta,\phi),z(\theta,\phi)) \in \mathbb{R}^3 \tag{A35a}$$

with

$$x(\theta,\phi) = r\cos\theta\sin\phi \tag{A35b}$$

$$y(\theta,\phi) = r\,\mathrm{sen}\,\theta\sin\phi \tag{A35c}$$

$$z(\theta,\phi) = r\cos\phi \tag{A35d}$$

and determine the metric tensor. Do the same using the result in (d).

(f) Given the ellipsoide

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

determine the metric tensor.

A.52. Curvatures & all that: how to measure variations on the Gauss map of a surface. Given a local parametrization of a piece of a surface,

$$(u,v) \quad \mapsto \quad \mathbf{X}(u,v) = (x(u,v),y(u,v),z(u,v)) \, ,$$

denote by $\mathbf{N} = \mathbf{N}(u,v)$ the Gauss map. Denote the derivatives of $\mathbf{N}$ by

$$\mathbf{N}_u = \frac{\partial \mathbf{N}}{\partial u} \quad \text{and} \quad \mathbf{N}_v = \frac{\partial \mathbf{N}}{\partial v} \, .$$

From the standpoint of the twists and bendings of the surface, the relevant quantities are the projections of $\mathbf{N}_u$ and $\mathbf{N}_v$ onto the tangent plane T_XS. Since $\mathbf{N}$ is a unit vector, the orthogonal projection of $\mathbf{N}_u$ onto the direction of $\mathbf{N}$ is $\langle \mathbf{N}_u, \mathbf{N} \rangle \mathbf{N}$ and its projection onto T_XS is, therefore,

$$\mathbf{V}_u = \mathbf{N}_u - \langle \mathbf{N}_u, \mathbf{N} \rangle \mathbf{N} \,.$$

Likewise, the projection of $\mathbf{N}_v$ over T_XS is

$$\mathbf{V}_v = \mathbf{N}_v - \langle \mathbf{N}_v, \mathbf{N} \rangle \mathbf{N} \,.$$

(a) Since $\mathbf{V}_u$ and $\mathbf{V}_v$ are vectors in the tangent plane T_XS, they can be written as linear combinations of the basis $\{\mathbf{X}_u, \mathbf{X}_v\}$ of T_XS,

$$\begin{aligned} \mathbf{V}_u &= k_{11}\mathbf{X}_u + k_{12}\mathbf{X}_v \\ \mathbf{V}_u &= k_{21}\mathbf{X}_u + k_{22}\mathbf{X}_v \,. \end{aligned}$$

By making inner products of $\mathbf{V}_u$ and $\mathbf{V}_v$ with $\mathbf{X}_u$ and $\mathbf{X}_v$, show that

$$K = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} = M^{-1}H \,, \tag{A36}$$

where

$$M = \begin{pmatrix} \langle \mathbf{X}_u, \mathbf{X}_u \rangle & \langle \mathbf{X}_v, \mathbf{X}_u \rangle \\ \langle \mathbf{X}_u, \mathbf{X}_v \rangle & \langle \mathbf{X}_v, \mathbf{X}_v \rangle \end{pmatrix} ,$$

and

$$H = \begin{pmatrix} \langle \mathbf{V}_u, \mathbf{X}_u \rangle & \langle \mathbf{V}_v, \mathbf{X}_u \rangle \\ \langle \mathbf{V}_u, \mathbf{X}_v \rangle & \langle \mathbf{V}_v, \mathbf{X}_v \rangle \end{pmatrix} .$$

Matriz K that records the variations of the Gauss map, $\mathbf{N}$, on the surface, with respect to the basis $\{\mathbf{X}_u, \mathbf{X}_v\}$ of the tangent space T_XS, is a symmetric matrix and captures the curvature of S. Its eigenvalues are called *principal curvatures* of S at $\mathbf{X}$, and their eigenvectors are the principal directions of S at $\mathbf{X}$. The determinant of K is called the *Gaussian curvature*, and half the trace of K, $(k_{11} + k_{22})/2$, is called the *mean curvature* of S at $\mathbf{X}(u,v)$.

(b) Consider the spherical coordinates of $S_0(r)$, given by Eq. (A35), and determine matriz K. Also, determine the principal curvatures of $S_0(r)$.

A.53. Let A be a symmetric, positive definite matrix. Given vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$, show that $A^{-1}(\mathbf{a} \times \mathbf{b})$ is A-orthogonal to $\mathbf{a}$ and $\mathbf{b}$.

A.54. How is it that an ellipse has constant curvature? Given a scalar function

$$\mathbb{R}^3 \supset U \ni (x,y,z)^T \mapsto G(x,y,z) \in \mathbb{R} \,,$$

let S be a surface defined as a level set of G,

$$S = \left\{(x,y,z) \in \mathbb{R}^3 \mid G(x,y,z) = c\right\} .$$

Assume that z can be written implicitly as a function of x and y, $z = z(x,y) \in \mathbb{R}$, such that,

$$G(x,y,z(x,y)) = c . \tag{A37}$$

Consider the local parametrization of S given by

$$(x,y) \mapsto \mathbf{X}(x,y) = (x,y,z(x,y)) .$$

Let the inner product of the ambient space, $\mathbb{R}^3$, be given by

$$\langle \mathbf{a},\, \mathbf{b}\rangle_A = \mathbf{a}^T A\mathbf{b} ,$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$.

(a) Differentiate Eq. (A37) with respect to x. Let

$$\nabla G = (\partial G/\partial x, \partial G/\partial y, \partial G/\partial z)$$

and obtain a relation between $\mathbf{X}_x$ and ∇G. Do the same for $\mathbf{X}_y$.

(b) Use (a) to show that $A^{-1}(\nabla G)^T$ is A-orthogonal to $\mathbf{X}_x$ and $\mathbf{X}_y$. Note that Gauss map can be written as

$$\mathbf{N} = \frac{A^{-1}(\nabla G)^T}{|A^{-1}(\nabla G)^T|_A} ,$$

where $|\cdot|_A = \sqrt{\langle\cdot\,,\,\cdot\rangle_A}$.

(c) Let

$$\mathbf{V}_x = \mathbf{N}_x - \langle \mathbf{N}_x, \mathbf{N}\rangle_A \mathbf{N}$$
$$\mathbf{V}_y = \mathbf{N}_y - \langle \mathbf{N}_y, \mathbf{N}\rangle_A \mathbf{N} ,$$

and, similarly to what was done in Exercise A.52, write

$$\mathbf{V}_x = k_{11}\mathbf{X}_x + k_{12}\mathbf{X}_y$$
$$\mathbf{V}_y = k_{21}\mathbf{X}_x + k_{22}\mathbf{X}_y .$$

Show that $K = M_A^{-1} H_A$ where

$$M_A = \begin{pmatrix} \langle \mathbf{X}_u, \mathbf{X}_u\rangle_A & \langle \mathbf{X}_v, \mathbf{X}_u\rangle_A \\ \langle \mathbf{X}_u, \mathbf{X}_v\rangle_A & \langle \mathbf{X}_v, \mathbf{X}_v\rangle_A \end{pmatrix} ,$$

and

$$H_A = \begin{pmatrix} \langle \mathbf{V}_u, \mathbf{X}_u\rangle_A & \langle \mathbf{V}_v, \mathbf{X}_u\rangle_A \\ \langle \mathbf{V}_u, \mathbf{X}_v\rangle_A & \langle \mathbf{V}_v, \mathbf{X}_v\rangle_A \end{pmatrix} .$$

(d) From Exercise A.13, we know that S defined as

$$1 = G(x,y,z) = (x\ y\ z)A\begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

is an ellipsoide. Show that

$$(\nabla G)^T = 2A\begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

(e) Show that the curvature matrix K of S, is the identity. That is, the ellipsoide, as seen from the metric defined by A, is isotropically round, i.e., it is a sphere (of radius 1).

References

[1] Alifanov, O.M.: Solution of an inverse problem of heat conduction by iteration methods. J. of Engineering Physics 26, 471–476 (1974)

[2] Alifanov, O.M., Mikhailov, V.V.: Solution of the nonlinear inverse thermal conductivity problem by iteration method. J. of Engineering Physics 35, 1501–1506 (1978)

[3] Alvarez Acevedo, N.I., Roberty, N.C., Silva Neto, A.J.: An explicit formulation for the inverse transport problem using only external detectors. Part I: Computational modelling. Computational & Applied Mathematics 29(3), 343–358 (2010)

[4] Alvarez Acevedo, N.I., Roberty, N.C., Silva Neto, A.J.: An explicit formulation for the inverse transport problem using only external detectors. Part II: Application to one-dimensional homogeneous and heterogeneous participating media. Computational & Applied Mathematics 29(3), 359–374 (2010)

[5] Andrews, R., Seidel, M., Biggs, M. (eds.): The Columbia World of Quotations. Columbia University Press, New York (1996)

[6] ANS. Joint International Conference on Mathematical Methods and Supercomputing for Nuclear Applications – special technical session on inverse and ill-conditioned problems. American Nuclear Society (1997)

[7] Arnold, V.I.: On teaching mathematics. Extended text of the address at the discussion on teaching of mathematics in Palais de Découverte in Paris (March 7, 1997), http://pauli.uni-muenster.de/~munsteg/arnold.html (accessed on June 6, 2011)

[8] ASME. First International Conference on Inverse Problems in Engineering: Theory and Practice (1993)

[9] Barbour, R.L., Carvlin, M.J., Fiddy, M.A.: Experimental and numerical methods for solving ill-posed inverse problems: medical and nonmedical applications. SPIE – The International Society for Optical Engineering (1995)

[10] Beck, J.V.: Combined parameter and function estimation in heat transfer with applications to contact conductance. Journal of Heat Transfer 110(4b), 1046–1058 (1988)

[11] Beck, J.V., Blackwell, B., Clair Jr., C.R.S.: Inverse heat conduction — Ill-posed problems. John Wiley & Sons (1985)

[12] Bejan, A.: Heat Transfer. John Wiley & Sons (1993)

[13] Bennett, A.F.: Inverse methods in physical oceanography. Cambridge Monographs on Mechanics and Applied Mathematics, Cambridge (1992)

[14] Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7, 200–217 (1967)

[15] do Carmo, M.P.: Differential geometry of curves and surfaces. Prentice-Hall Inc. (1976)

[16] Carita Montero, R.F., Roberty, N.C., Silva Neto, A.J.: Absorption coefficient estimation in heterogeneous media using a domain partition consistent with divergent beams. Inverse Problems in Engineering 9, 587–617 (2001)

[17] Carvalho, G., Frollini, E., Santos, W.N.: Thermal conductivity of polymers by hot-wire method. Journal of Applied Polymer Science 62, 2281–2285 (1996)

[18] Carvalho, G., Silva Neto, A.J.: An inverse analysis for polymers thermal properties estimation. In: Proc. 3rd International Conference on Inverse Problems in Engineering: Theory and Practice (1999)

[19] Chalhoub, E.S., Campos Velho, H.F., Silva Neto, A.J.: A comparison of the one-dimensional radiative transfer problem solution obtained with the Monte Carlo method and three variations of the discrete ordinates method. In: Proc. 19th International Congress of Mechanical Engineering. ABCM, Brasília (2007)

[20] Chandrasekhar, S.: On the radiative equilibrium of a stellar atmosphere, II. Astrophys. J. 100, 76–86 (1944)

[21] Chandrasekhar, S.: Radiative transfer. Dover (1960)

[22] Tables of angular distribution coefficients for light scattering by spheres. University of Michigan Press, Ann Arbor (1957)

[23] Ciarlet, P.G.: Introduction to numerical linear algebra and optimization. Cambridge University Press, Cambridge (1989)

[24] Cidade, G.A.G.: Development of methodologies for biological images acquisition and processing in atomic force microscope (AFM). D. Sc. Thesis, Institute of Biophysics Carlos Chagas Filho. Federal University of Rio de Janeiro (2000) (in Portuguese)

[25] Cidade, G.A.G., Anteneodo, C., Roberty, N.C., Silva Neto, A.J.: A generalized approach for atomic force microscopy image restoration with Bregman distances as Tikhonov regularization terms. Inverse Problems in Engineering 8, 457–472 (2000)

[26] Cidade, G.A.G., Costa, L.T., Weissmüller, G., Silva Neto, A.J., Roberty, N.C., Moraes, M.B., Prazeres, G.M.P., Hill, C.E.M., Ribeiro, S.J.M., Souza, G.G.B., Teixeira, L.S.P., Monçores, M.C., Bisch, P.M.: Atomic force microscopy as a tool for biomedical and biotechnological studies. Artificial Organs 27(5), 447–451 (2003)

[27] Cidade, G.A.G., Silva Neto, A.J., Roberty, N.C.: Image restoration with applications in biology and engineering — Inverse problems in nanoscience and nanotechnology, São Carlos, Brazil. SBMAC (Brazilian Society of Computational and Applied Mathematics). Notes in Applied Mathematics, vol. 1 (2003) (in Portuguese)

[28] Darcy, H.: Les Fontaines Publiques de la Ville de Dijon. Dalmont, Paris (1856)

[29] Denisov, A.M.: Elements of the theory of inverse problems. VSP BV, The Nederlands (1999)

[30] Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Kluwer Academic Publishers (1996)

[31] Engl, H.W., Rundell, W.: Inverse problems in diffusion processes. GAMM-SIAM (Gesellschaft für Angewandte Mathematik und Mechanik Regensburg – US Society for Industrial and Applied Mathematics) (1994)

[32] Feynmann, R.: The character of physical law. MIT Press, Cambridge (1964)

[33] Flach, G.P., Özisik, M.N.: Inverse heat conduction problem of simultaneously estimating spatially varying thermal conductivity and heat capacity per unit volume. Numer. Heat Transfer 16, 249–266 (1989)

[34] Gallant, A.R.: Nonlinear statistical models. Wiley (1987)

[35] Golub, G.H., Van Loan, C.F.: Matrix computations. The Johns Hopkins University Press (1989)

[36] Gordon, C., Webb, D.L., Wolpert, S.: One cannot hear the shape of a drum. Bull. Amer. Math. Soc. 27(1), 134–138 (1992)

[37] Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. Journal of Research of the National Bureau of Standards 49(6), 409–436 (1952)

[38] Ho, C.H., Özisik, M.N.: An inverse radiation problem. International Journal of Heat and Mass Transfer 32(2), 335–341 (1989)

[39] John, F.: Partial Differential Equations. Springer (1982)

[40] Kac, M.: Can one hear the shape of a drum? Am. Math. Monthly 73(4, pt. II), 1–23 (1966)

[41] Kanitz, S.: Stimullating creativity Veja, Editora Abril, São Paulo 1826 (October 29, 2003) (in Portuguese), http://veja.abril.com.br/acervodigital/home.aspx

[42] Kapur, J.N., Kesavan, H.K.: Entropy Optimization Principles with Applications. Academic Press, Inc. (1992)

[43] Keller, J.: Inverse problems. Am. Math. Mon. 73, 1–23 (1966)

[44] Kirsch, A.: An introduction to the mathematical theory of inverse problems. Springer (1996)

[45] Knupp, D.C., Silva Neto, A.J., Sacco, W.F.: Radiative properties estimation with the particle collision algorithm based on a sensitivity analysis. High Temperatures — High Pressures 38(2), 137–151 (2009)

[46] Lobato, F.S., Steffen Jr., V., Silva Neto, A.J.: A comparative study of the application of differential evolution and simulated annealing in inverse radiative transfer problems. Journal of the Brazilian Society of Mechanical Sciences and Engineering XXXII(5), 518–526 (2010)

[47] Lobato, F.S., Steffen Jr., V., Silva Neto, A.J.: Self-adaptive differential evolution based on the concept of population diversity applied to simultaneous estimation of anisotropic scattering phase function, albedo and optical thickness. Computer Modeling in Engineering & Sciences 69(1), 1–18 (2010)

[48] Lobato, F.S., Steffen Jr., V., Silva Neto, A.J.: Resolution of inverse problems in diffusive processes and radiative transfer using the differential evolution algorithm. In: Lopes, H.S., Takahashi, R.H.C. (eds.) Evolutionary Computation in Engineering Problems, ch. 9, pp. 173–195. Omnipax Ed. Ltda, Curitiba (2011)

[49] Lugon Jr., J., Silva Neto, A.J.: Solution of porous media inverse drying problems using a combination of stochastic and deterministic methods. Journal of the Brazilian Society of Mechanical Sciences and Engineering XXXIII(4), 400–407 (2011)

[50] Lugon Jr., J., Silva Neto, A.J., Biondi Neto, L., Soeiro, F.J.C.P., Santana, C.C., Campos Velho, H.F.: Application of artificial neural networks and hybrid methods in the solution of inverse problems. In: Artificial Neural Networks - Application, ch. 26, pp. 541–566. InTech (2011)

[51] Lugon Jr., J., Silva Neto, A.J., Santana, C.C.: A hybrid approach with artificial neural networks, Levenberg-Marquardt and simulated annealing methods for the solution of gas-liquid adsorption inverse problems. Inverse Problems in Science and Engineering 11(5), 391–408 (2003)

[52] Lugon Jr., J., Silva Neto, A.J., Rodrigues, P.P.G.W.: Assessment of dispersion mechanisms in rivers by means of an inverse problem approach. Inverse Problems in Science and Engineering 16(8), 967–979

[53] Marsden, J.E., Hoffman, M.: Elementary Classical Analysis, 2nd edn. W.H. Freeman (1993)

[54] Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Industr. Appl. Math. 11, 431–441 (1963)

[55] Montgomery, D.C., Runger, G.C.: Applied statistics and probability for engineers. John Wiley & Sons (1994)

[56] Moura Neto, F.D., Melo, S.T.R.: Darcy's law for a heterogeneous porous medium. Journal of Porous Media 4(2), 165–178 (2001)

[57] Moura Neto, F.D., Silva Neto, A.J.: Solution of the one-dimensional transport equation with operator splitting and integrating factor. In: Proc. 15th Brazilian Congress of Mechanical Engineering. ABCM, Águas de Lindóia (1999) (in Portuguese)

[58] Moura Neto, F.D., Silva Neto, A.J.: Two equivalent approaches to obtain the gradient in algorithms for function estimation. In: 34th National Heat Transfer Conference, Pittsburgh (2000)

[59] Myers, R.H., Montgomery, D.C., Vining, G.G.: Generalized Linear Models with Applications in Engineering and the Sciences. John Wiley & Sons, New York (2002)

[60] Özisik, M.N.: Radiation transfer and interactions with conduction and convection. John Wiley (1973)

[61] Özisik, M.N.: Boundary Value Problems of Heat Conduction. International Textbook Co., Scranton (1968)

[62] Philippi Jr., A., Silva Neto, A.J. (eds.): Interdisciplinarity in science, technology and innovation. Manole, São Paulo (2011) (in Portuguese)

[63] Pinheiro, R.P.F., Silva Neto, A.J., Moura Neto, F.D.: Operator splitting techniques for radiative transfer in a participating medium. In: Proc. 5th World Congress on Computational Mechanics, Vienna, Austria (2002)

[64] Polak, E.: Computational methods in optimization. Academic Press (1971)

[65] Protter, M.H.: Can one hear the shape of a drum? Revisited. SIAM Rev. 29(2), 185–197 (1987)

[66] Reed, M., Simon, B.: US Methods of modern mathematical physics, vol. 1. Academic Press (1972)

[67] Shewchuck, J.R.: An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, http://www.cs.cmu.edu/ quake-papers/painless-conjugate-gradient.pdf (accessed June 6, 2011)

[68] Silva Neto, A.J.: Explicit and implicit formulations for inverse radiative transfer problems. In: Proc. 5th World Congress on Computational Mechanics, Mini-Symposium MS-125-Computational Treatment of Inverse Problems in Mechanics (2002)

[69] Silva Neto, A.J., Becceneri, J.C. (eds.): Nature inspired computational intelligence techniques — application in inverse radiative transfer problems, São Carlos, Brazil. SBMAC (Brazilian Society of Computational and Applied Mathematics). Notes in Applied Mathematics, vol. 41 (2009) (in Portuguese)

[70] Silva Neto, A.J., Cella, N.: A regularized solution with weighted Bregman distances for the inverse problem of photoacoustic spectroscopy. Computational & Applied Mathematics 25, 139–165 (2006)

[71] Silva Neto, A.J., Lugon Jr., J., Soeiro, F.J.C.P., Biondi Neto, L., Santana, C.C., Lobato, F.S., Steffen Jr., V.: Application of simulated annealing and hybrid methods in the solution of inverse heat and mass transfer problems. In: Simulated Annealing, Theory with Applications, ch. 2, pp. 17–50. Sciyo (2010)

[72] Silva Neto, A.J., McCormick, N.J.: An explicit formulation based on the moments of the exit radiation intensity for the one-dimensional inverse radiative transfer problem. In: Proc. 4th International Conference on Inverse Problems in Engineering: Theory and Practice, Angra dos Reis, Brazil, vol. II, pp. 347–354 (2002)

[73] Silva Neto, A.J., Moura Neto, F.D., Su, J.: Function estimation with the conjugate gradient method in linear and nonlinear heat conduction problems. In: Proc. 15th Brazilian Congress of Mechanical Engineering. ABCM, Águas de Lindóia (1999)

[74] Silva Neto, A.J., Özisik, M.N.: An inverse problem of estimating thermal conductivity, optical thickness, and single scattering albedo of a semi-transparent medium. In: First International Conference on Inverse Problems in Engineering: Theory and Practice, pp. 267–273. ASME (1993)

[75] Silva Neto, A.J., Özisik, M.N.: Estimation of refractive index and optical thickness for a semi-transparent medium in radiative equilibrium. In: Proc. 12th Brazilian Congress of Mechanical Engineering, vol. II, pp. 665–668 (1993)

[76] Silva Neto, A.J., Özisik, M.N.: Inverse problem of simultaneously estimating the timewise-varying strength of two plane heat sources. Journal of Applied Physics 73(5), 2132–2137 (1993)

[77] Silva Neto, A.J., Özisik, M.N.: Simultaneous estimation of location and timewise varying strength of a plane heat source. Numer. Heat Transfer, Part A: Applications 24, 467–477 (1993)

[78] Silva Neto, A.J., Özisik, M.N.: The estimation of space and time dependent strength of a volumetric heat source in a one-dimensional plate. Int. J. Heat and Mass Transfer 37(6), 909–915 (1994)

[79] Silva Neto, A.J., Özisik, M.N.: An inverse problem of simultaneous estimation of radiation phase function, albedo and optical thickness. J. Quant. Spectrosc. Radiat. Transfer 53(4), 397–409 (1995)

[80] Silva Neto, A.J., Roberty, N.C.: A comparison of the discrete ordinates method with other techniques for the solution of the one-dimensional transport equation. In: Proc. 7th Brazilian Congress of Engineering and Thermal Sciences, vol. II, pp. 878–883. Rio de Janeiro, Brazil (1998) (in Portuguese)

[81] Silva Neto, A.J., Roberty, N.C., Pinheiro, R.P.F., Alvarez Acevedo, N.I.: Inverse problems explicit and implicit formulations with applications in engineering, biophysics and biotechnology. Inverse Problems in Science and Engineering 15(4), 373–411 (2007)

[82] Strang, G.: Linear algebra and its applications, 3rd edn. Academic Press (1988)

[83] Strang, G.: Computational Science and Engineering. Wellesley Cambridge Press (2007)

[84] Stutz, D.: Parallel computation strategies for image restoration with Tikhonov's regularization functional. D.Sc. Thesis. Computational Modelling Program. Polytechnique Institute. Rio de Janeiro State University (2009) (in Portuguese)

[85] Stutz, D., Silva Neto, A.J.: Parallel computing fundamentals for atomic force microscopy image restoration, São Carlos, Brazil. SBMAC (Brazilian Society of Computational and Applied Mathematics). Notes in Applied Methematics, vol. 56 (2011) (in Portuguese)

[86] Stutz, D., Silva Neto, A.J., Cidade, G.A.G.: Parallel computation approach for the restoration of AFM images based on the Tikhonov regularization method. Microscopy & Microanalysis 11(suppl. S03), 22–25 (2005)

[87] Su, J., Silva Neto, A.J.: Two-dimensional inverse heat conduction problem of source strength estimation in cylindrical rods. Applied Mathematical Modelling 25, 861–872 (2001)

[88] Tikhonov, A.N., Arsenin, V.Y.: Solutions of ill-posed problems. V. H. Winston & Sons, Washington, D.C (1977)

[89] Vasconcellos, J.F.V., Silva Neto, A.J., Santana, C.C.: An inverse mass transfer problem in solid-liquid adsorption systems. Inverse Problems in Engineering 11(5), 391–408 (2003)

[90] Veríssimo, L.F.: Glories. O Estado de São Paulo (January 5, 2008), http://www.estadao.com.br/noticias/impresso,glorias,104819,0.html (in Portuguese)

[91] Vega, J.A.: Ingeneria de la producción y transformación de polímeros. Instituto de Plásticos y Caucho (1984)

[92] Wang, J., Silva Neto, A.J., Moura Neto, F.D., Su, J.: Function Estimation with Alifanov's Iterative Regularization Method in Linear and Nonlinear Heat Conduction Problems. Applied Mathematical Modelling 26(11), 1093–1111 (2002)

[93] Wick, G.C.: Über ebene Diffusionsprobleme. Z. Phys. 121(11-12), 702–718 (1943)

[94] Wigner, E.: The Unreasonable Effectiveness of Mathematics in the Natural Sciences. Communications in Pure and Applied Mathematics 13(I) (1960)

Index